Poster: THREATKG - A System for Automated Cyber Threat Knowledge Gathering and Management

Zhengjie Ji*†, Xiaoyuan Liu*‡, Edward Choi‡, Sibo Ma‡, Xinyu Yang†, Dawn Song‡, Peng Gao†
†Virginia Tech ‡UC Berkeley
†{zhengjie,xinyuyang,penggao}@vt.edu ‡{xiaoyuanliu,edwardc1028,siboma,dawnsong}@berkeley.edu

***Abstract*—Open-source cyber threat intelligence (OSCTI) serves as a crucial resource for understanding cyber threats. However, little effort has been made to harvest knowledge from unstructured OSCTI reports from publicly available sources (e.g., technical reports, security blogs, and news articles). These reports provide comprehensive threat knowledge in various entities (e.g., IOCs, threat actors, TTPs) and relations (e.g., usage, indication, mitigation). However, these entities and relations are hard to gather due to diverse report formats, large report volumes, and complex structures and nuances in the natural language report text. To bridge the gap, we propose THREATKG, a system for automated open-source cyber threat knowledge gathering and management. THREATKG autonomously collects OSCTI reports from various sources, extracts high-fidelity threat knowledge, constructs a large threat knowledge graph, and continuously updates the graph by continuously ingesting new knowledge.**

I. INTRODUCTION

Sophisticated cyber attacks have plagued many high-profile businesses [1]. To remain aware of the fast-evolving cyber threat landscape and gain insights into the most dangerous threats, security researchers and practitioners actively gather knowledge about cyber threats from past incidents, and share the knowledge through public sources like security websites and blogs. Such open-source cyber threat intelligence (OSCTI) [2] has received growing attention from the community.

Despite the pressing need for high-quality threat knowledge to empower defenses, existing OSCTI gathering and management systems [3]–[5], however, have primarily focused on structured Indicator of Compromise (IOC) feeds, which are forensic artifacts of intrusions such as hashes of malware samples, names of malicious files/processes, and IP addresses of botnets. Though useful in capturing fragmented views of threats, these IOCs are low-level and disconnected, and thus they lack the capability to uncover the complete threat scenario as to how the threat unfolds into multiple steps, which is typically observed in most sophisticated attacks these days [1]. Consequently, defensive measures that rely on these low-level, fragmented indicators are easy to bypass when the attacker repurposes the tools and changes their signatures [2].

In contrast, a large number of unstructured OSCTI reports have been significantly overlooked (e.g., security blogs and news [6], threat encyclopedia pages [7]), which contain more comprehensive knowledge about threats in natural language text. Besides low-level IOC entities, OSCTI reports contain various (1) *higher-level threat knowledge entities* (e.g., threat actors, adversary tactics, techniques, and procedures (TTPs) [8]), and (2) *semantic relationships* between entities that indicate their interactions. Such high-level and connected knowledge is tied to the attacker's goals and thus more difficult to change, which is critical for uncovering the complete multi-step threat scenario and building more robust defenses. As the volume of OSCTI reports increases day by day, it becomes increasingly challenging for threat analysts to manually maneuver through and correlate the myriad of sources to gain useful knowledge. Unfortunately, prior approaches do not provide an automated and principled way to gather such knowledge from OSCTI reports and manage the knowledge.

In this work, we seek to design and build a system that (1) automatically gathers high-fidelity cyber threat knowledge from a large number of OSCTI reports, and (2) manages such knowledge in a unified knowledge base to provide comprehensive views of various threats. We identify four *major challenges*. First, OSCTI reports contain various types of entities and relations (e.g., IOCs, threat actors, tools) that capture threat behaviors. Second, OSCTI reports collected from different sources have diverse formats. Third, accurately extracting threat knowledge from natural language text is non-trivial due to the massive nuances (e.g., dots, underscores in IOCs) in the security context. Fourth, new OSCTI reports are being published every day that contain fresh knowledge about the latest threats. The system also needs to be scalable (to handle the large report volume) and extensible (to generalize to new reports with unseen formats).

To bridge the gap, we propose THREATKG (∼26K LOC), an AI-powered system for automated open-source cyber threat knowledge gathering and management. THREATKG automatically collects a large number of OSCTI reports from a wide range of sources, uses a combination of ML and NLP techniques to extract high-fidelity threat knowledge, constructs a *threat knowledge graph*, and updates the knowledge graph by continuously ingesting new knowledge. For more information, please refer to our full paper on [9].

II. DESIGN OF THREATKG

A. Automated OSCTI Report Collection

We built a robust multi-threaded crawler framework that manages crawlers to collect OSCTI reports from 40 major security websites [6], [7], including threat encyclopedias, enterprise security blogs, influential personal security blogs, security news, etc. These websites provide a large number

*Equal contribution

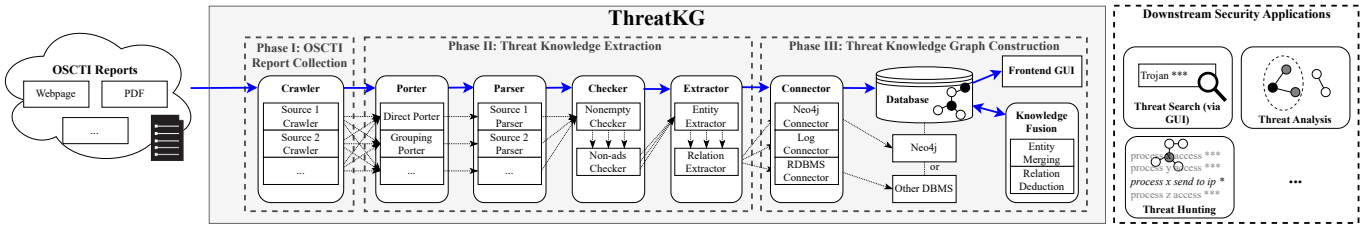

Fig. 1. The architecture of THREATKG. Arrows between system components represent data flows.

of OSCTI reports (in the form of webpages) that cover various types of threats (e.g., malware, vulnerabilities, attack campaigns), making them a valuable source of threat knowledge. The crawler framework schedules periodic execution and reboots after failure for individual crawlers in a robust manner.

B. Hierarchical Threat Knowledge Ontology

To comprehensively model the threats, we construct a hierarchical threat knowledge ontology that covers various threat knowledge entities and relations for capturing both low-level threat behaviors and high-level threat contexts. It consists of three layers: the report context layer, the threat behavior layer, and the threat context layer. The report context layer of the ontology contains report-level knowledge (e.g., report URLs, OSCTI vendors). The threat behavior layer of the ontology contains knowledge of low-level threat behaviors (e.g., filename, IP). The threat context layer of the ontology provides high-level contexts for threats in addition to detailed threat behavior steps (e.g., vulnerabilities, threat actors). The three layers of ontology collectively model the threats from multiple dimensions and in different granularities.

C. Threat Knowledge Extraction

THREATKG employs a specialized NLP pipeline that targets the unique problem of extracting knowledge from OSCTI text.

1) Threat Knowledge Entity Extraction: THREATKG incorporates a set of regex rules in a rule-based entity extractor for extracting IOCs. For other types of entities (e.g., malware, threat actors, tools) that are hard to specify using rules, THREATKG employs a bidirectional LSTM-CRF model [10] to perform named entity recognition (NER) over OSCTI text.

2) Threat Knowledge Relation Extraction: THREATKG employs a dependency parsing-based relation extractor to extract interaction verbs between two entities and a Piecewise Convolutional Neural Networks (PCNN) model [11] to extract relations that are not explicitly associated with words in the text (e.g., the use relation between CozyDuke and Office Monkeys (Short Flash Movie).exe).

3) Data Programming: To reduce the cost of obtaining supervision, we leverage data programming [12], which programmatically synthesizes annotations via unsupervised modeling of sources of weak supervision.

D. Scalable and Extensible System Architecture

THREATKG constructs the threat knowledge graph from the extracted threat knowledge and stores it in the backend database for persistence. We parallelize the system components in the same processing step (e.g., multiple parsers) for scalability and allow multiple system components to work together with the same input/output interface for extensibility. THREATKG is fully automated and continuously running to gather and integrate knowledge from the latest OSCTI reports.

E. Downstream Security Applications

THREATKG can empower many downstream security applications. Here, we provide two examples.

1) Threat Search and Knowledge Graph Exploration: We constructed a web GUI using React and Elasticsearch. The GUI interacts with the database and provides various types of interactivity. A demo video showcasing our GUI can be found at [13]. THREATKG can also be incorporated into our previous cyber threat hunting system, as demonstrated by [14].

2) Threat Question Answering: To enable flexible and intuitive knowledge acquisition via natural language, we built a system using transformer models [15] for knowledge graph question answering [16]. A demo video is available at [17].

Acknowledgment. This work was supported in part by Cisco, the Commonwealth Cyber Initiative, Amazon, and Microsoft.

REFERENCES

[1] "Biggest Data Breaches in US History," 2023, https://www.upguard.com/blog/biggest-data-breaches-us.
[2] V. G. Li, M. Dunn, P. Pearce, D. McCoy, G. M. Voelker, and S. Savage, "Reading the tea leaves: A comparative analysis of threat intelligence," in *USENIX Security*, 2019.
[3] "AlienVault OTX," 2022, https://otx.alienvault.com/.
[4] "MISP - Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing," 2022, https://www.misp-project.org/.
[5] "OpenCTI," 2022, https://www.opencti.io/en/.
[6] "Symantec Threat Intelligence," 2017, https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence.
[7] "Kaspersky Threat Encyclopedia," 2022, https://threats.kaspersky.com/.
[8] "Mitre att&ck," n.d., https://attack.mitre.org.
[9] P. Gao, X. Liu, E. Choi, S. Ma, X. Yang, Z. Ji, Z. Zhang, and D. Song, "Threatkg: A threat knowledge graph for automated open-source cyber threat intelligence gathering and management," 2022, arXiv:2212.10388.
[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *NAACL*, 2016.
[11] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," *EMNLP*, 2015.
[12] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. Ré, "Data programming: Creating large training sets, quickly," in *NeurIPS*, 2016.
[13] "Demo of threatkg," 2022, https://www.youtube.com/watch?v=wR4TdK7uc_U.
[14] P. Gao, F. Shao, X. Liu, X. Xiao, Z. Qin, F. Xu, P. Mittal, S. R. Kulkarni, and D. Song, "Enabling efficient cyber threat hunting with cyber threat intelligence," in *ICDE*, 2021.
[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
[16] Z. Ji, E. Choi, and P. Gao, "A knowledge base question answering system for cyber threat knowledge acquisition," in *ICDE Demo*, 2022.
[17] "Demo of threatqa," 2022, https://https://www.youtube.com/watch?v=lNAZYgjm7xE.