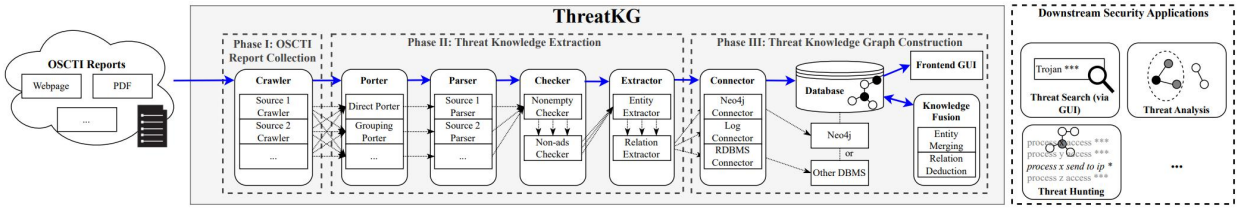
ThreatKG: A System for Automated Cyber Threat Knowledge Gathering and Management

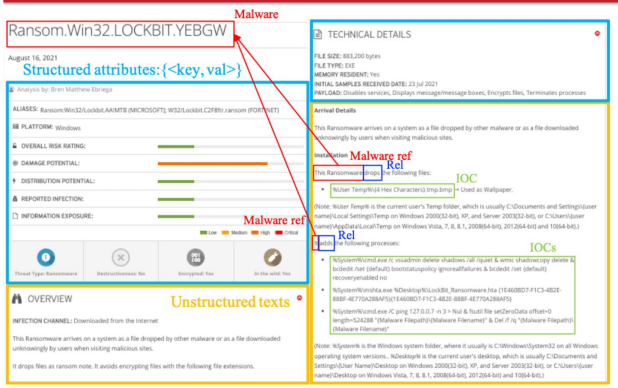
Zhengjie Ji[1*], Xiaoyuan Liu[2*] (*equal contribution), Edward Choi[2], Sibo Ma[2], Xinyu Yang[1], Dawn Song[2], Peng Gao[1]

[1] VIRGINIA TECH [2] Berkeley UNIVERSITY OF CALIFORNIA

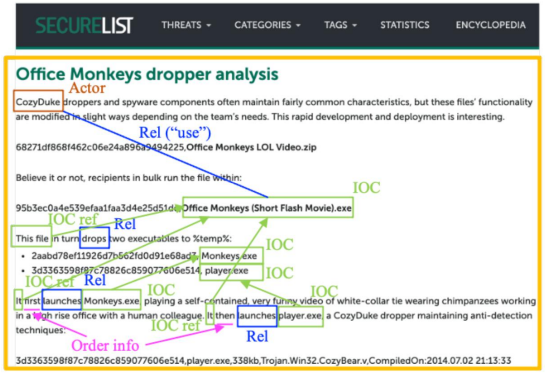
Design of ThreatKG



OSCTI Reports Contain Rich Threat Knowledge



Semi-structured report snippet from Trend Micro threat encyclopedia



Unstructured report snippet from Securelist

Challenge: **open-source cyber threat intelligence (OSCTI) reports** collected from different sources have **diverse formats** and new reports are **being published every day**.

Solution: a system that automatically gathers high-fidelity cyber threat knowledge from a large number of OSCTI reports. We built a robust multi-threaded crawler framework that manages crawlers to collect OSCTI reports from **40 major security websites**, including threat encyclopedias, enterprise security blogs, security news, etc.

Threat Knowledge Extraction

Challenge: OSCTI reports contain **various** other types of entities and relations that capture threat behaviors. However, existing OSCTI gathering and management systems **ignore many entity and relation types**.

Solution: ThreatKG employs a **specialized NLP pipeline** for extracting knowledge from OSCTI text.

- **Threat Knowledge Entity Extraction**: **regex rules** for extracting IOCs, **Bidirectional LSTM-CRF** model for performing named entity recognition on other types of entities.
- **Threat Knowledge Relation Extraction**: **dependency parsing-based relation extractor** to extract interaction verbs between two entities, **Piecewise Convolutional Neural Networks (PCNN) model** to extract relations that are not explicitly associated with words in the text.

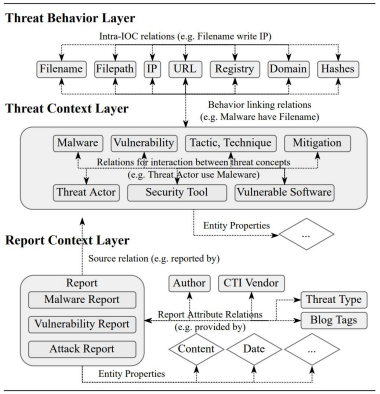
	Precision	Recall	F1
Seen Sources	99.98%	99.98%	99.98%
Unseen Sources	99.83%	99.83%	99.83%

Entity extraction performance

	Precision	Recall	Accuracy	F1
W/O Data Programming	80%	78%	78%	79%
W/ Data Programming	85%	85%	85%	85%

Relation extraction performance

Hierarchical Threat Knowledge Ontology



To comprehensively **model the threats**, we construct a hierarchical threat knowledge ontology that includes various threat knowledge entities and relations for capturing both low-level threat behaviors and high-level threat contexts.

- **Report context layer** contains **report-level knowledge**.
- **Threat behavior layer** contains **knowledge of low-level threat behaviors**.
- **Threat context layer** provides **high-level contexts** for threats in addition to detailed threat behavior steps.

The three layers of ontology collectively model the threats **from multiple dimensions** and **in different granularities**.

Scalable and Extensible System Architecture

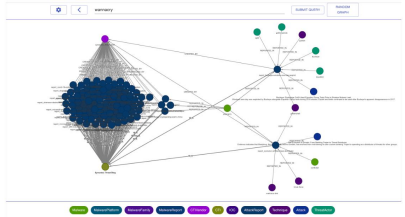
ThreatKG constructs the **threat knowledge graph** from the extracted threat knowledge and stores it in the backend database for persistence.

- For **scalability**, we parallelize the system components for the processing steps.
- For **extensibility**, we allow multiple system components in the same processing step to work together with the same input/output interface.

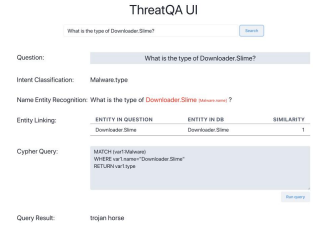
ThreatKG is fully automated and continuously running to provide the latest threat knowledge timely.

Stage	Total Processing Time (h)	Percentage		
Porter	0.54			0.6%
Checker	0.03			0.0%
Parser	1.45			1.7%
Extractor	85.26	97.7%	Content relevance analysis	2.1%
			Dependency parsing for IOC relation extraction	83.1%
			BiLSTM CRF entity extraction recognition	6.0%
			Potential relation marking	0.9%
			PCNN-ATT relation extraction	5.7%

Downstream Security Applications



Threat Search and Exploration: to facilitate **threat search** and **knowledge graph exploration**, we built a web GUI using React and Elasticsearch. The GUI interacts with the database and provides various interactivity.



Threat Question Answering: to enable **flexible** and **intuitive threat knowledge acquisition**, we build a question-answering system using the Transformer model.