Theory and Practice of Cybersecurity Knowledge Graphs and Further Steps

GRASEC 2024 Workshop Keynote

RNDr. Martin Husák, Ph.D.
husakm@ics.muni.cz

Institute of Computer Science, Masaryk University

August 1, 2024

Presenter’s Biography

RDNr. Martin Husák, Ph.D.

- Researcher at Institute of Computer Science, Masaryk University, Czech Republic
- Member of Masaryk University’s incident response team CSIRT-MU (`https://csirt.muni.cz/`)
- Recently also a visiting researcher at The Cyber Center for Security and Analytics, The University of Texas at San Antonio, USA.
- This research was supported by OP JAK “MSCAfellow5_MUNI” (No. CZ.02.01.01/00/22_010/0003229).

Co-funded by
the European Union

MŠMT
MINISTRY OF EDUCATION,
YOUTH AND SPORTS

Outline

Introduction to Knowledge Graphs

Predecessors: Graphs for Cyber Situational Awareness

Current State: Knowledge Graphs in Cybersecurity

Future Prospects and Conclusions

Section 1

Introduction to Knowledge Graphs

What is a knowledge graph?

Storing data in Graphs

- Verteces and Edges representing Entities and Relationships
- Graph databases (e.g, Neo4j) – popular approach to NoSQL

Knowledge Graphs (KG)

- Many, often biased or contradicting, definitions[1]
- Originated in Google in 2012 for contextualized answers to searches ("knowledge panel")

[1] https://neo4j.com/blog/what-is-knowledge-graph/

Semantics

Organizing Principle

- schema of organizing nodes and relationships according to fundamental concepts
- simple (product line -> product category -> product taxonomy) or complex
- multiple organizing principles can be used simultaneously

Ontology

- formal specification of the concepts and the relationships between them
- needs to be created for a given subject area
- defining an ontology is a hard task – research problem

Knowledge Graphs in Cybersecurity

Is this a new trend? Yes, just check the number of surveys:

- Li, H., et al. (2024). Cybersecurity knowledge graphs construction and quality assessment. Complex & Intelligent Systems.
- Sikos, L. F. (2023). Cybersecurity knowledge graphs. Knowledge and Information Systems.
- Zhao, X., et al. (2023). A survey on cybersecurity knowledge graph construction. Computers & Security.
- Ma, Y., et al.. (2023). The Advancement of Knowledge Graphs in Cybersecurity: A Comprehensive Overview. In ICCES.
- Bolton, J., et al. (2023). An Overview of Cybersecurity Knowledge Graphs Mapped to the MITRE ATT&CK Framework Domains. In 2023 IEEE ISI.
- Wang, Z., Zhu, H., Liu, P., & Sun, L. (2021). Social engineering in cybersecurity: a domain ontology and knowledge graph application examples. In Cybersecurity.

Is this a new thing?

- Well, not really...

Section 2

Predecessors: Graphs for Cyber Situational Awareness

Provisioning CSA via Complex Networks

Cyber situational awareness (CSA)

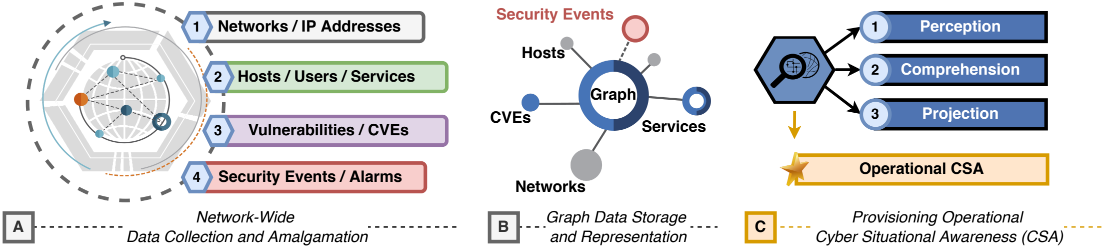
- Perception of the elements in the environment
- Comprehension of the situation
- Projection of future state and events

Proposed tools and models

- CAULDRON (George Mason University), CyGraph (MITRE)
- CRUSOE (Masaryk University)
- VirtualTerrain, CAMUS, M2D2, ...

Simple graphs are becoming complex networks

Provisioning CSA via graph-based analytics



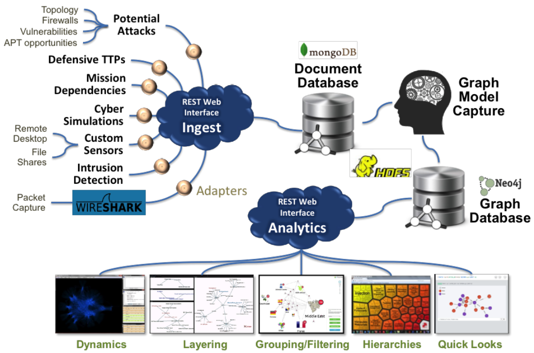
1 Networks / IP Addresses
2 Hosts / Users / Services
3 Vulnerabilities / CVEs
4 Security Events / Alarms
A
Network-Wide Data Collection and Amalgamation
Security Events
Hosts
Graph
CVEs
Services
Networks
B
Graph Data Storage and Representation
1 Perception
2 Comprehension
3 Projection
Operational CSA
C
Provisioning Operational Cyber Situational Awareness (CSA)

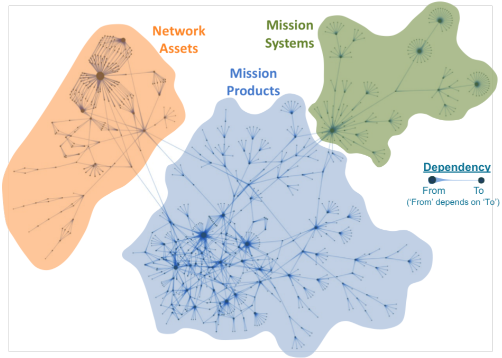
- *(A)* collect and amalgamate network-wide data using heterogeneous tools for computer network monitoring and reconnaissance,
- *(B)* leverage graph-based analytics to store, visualize, and query the data,
- *(C)* leverage this data to provision operational CSA for defensive measures, incident responses, and network forensics.

Wait, isn't this a knowledge graph of a network?

Example – CyGraph (MITRE)

- Graph-based data model for cyber situational awareness
- Detailed representation of the network and security posture

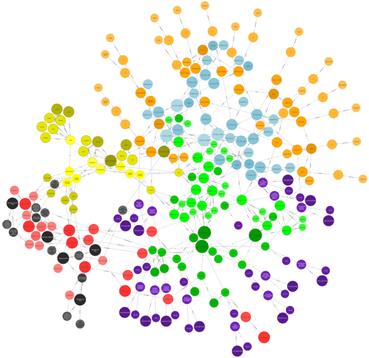


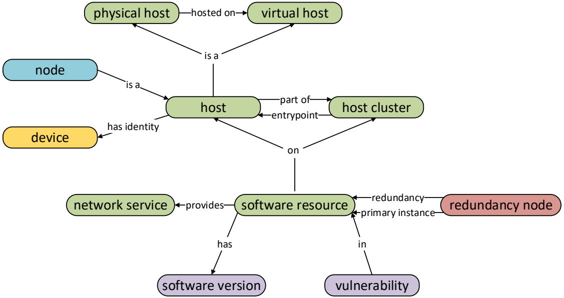


Noel, S., Harley, E., Tam, K. H., Limiero, M., & Share, M. (2016). CyGraph: graph-based analytics and visualization for cybersecurity. In Handbook of Statistics (Vol. 35, pp. 117-167). Elsevier.

Example – CRUSOE (Masaryk University)

- Inspired by CyGraph, more lightweight and focused on automation and orchestration
- `https://github.com/CSIRT-MU/CRUSOE`





physical host
hosted on
virtual host
is a
node
is a
host
part of
entrypoint
host cluster
device
has identity
on
network service
provides
software resource
redundancy
primary instance
redundancy node
has
in
software version
vulnerability

Husák, M., Sadlek, L., Špaček, S., Laštovička, M., Javorník, M., & Komárková, J. (2022). CRUSOE: A toolset for cyber situational awareness and decision support in incident handling. Computers & Security, 115, 102609.

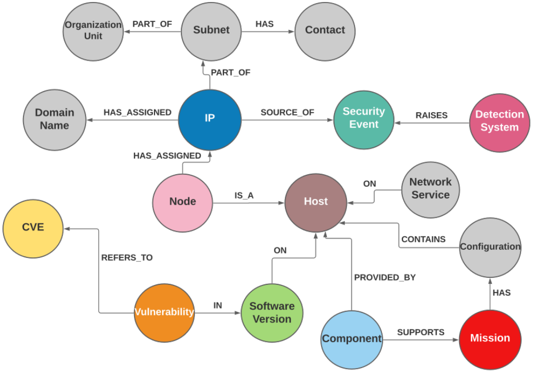
Sample Queries

How many hosts in my network are vulnerable?

- Count all hosts (node type "Host")
- Count hosts for which there exists any path like "Host – ON – Software Version – ON – Vulnerability – REFERS-TO – CVE"

Or for a specific vulnerability

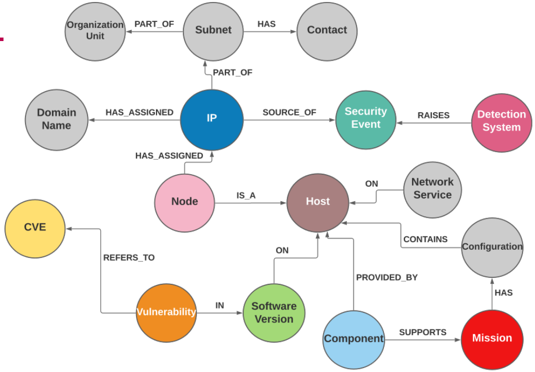
- Find the node of type "CVE", where CVE = <input>
- List all nodes of type "Host", for which there exists a path... see above



Sample Queries

Which admins are not securing their systems?

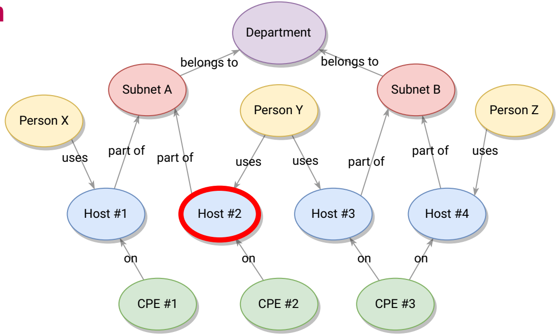
- List all vulnerable hosts or hosts with obsolete SW version
- List all their neighbours of type "User" or "Administrator"
- Aggregate the users and admins
- Fetch their contnact emails to send out notification



Sample Queries

Host #2 is infected, where can the infection spread?

- If it’s a worm, it can scan and exploit hosts in the same subnet
- If it’s a phishing ransomware, it can spread to other hosts controlled by the same user
- If it exploits certain vulnerability, let’s look for similat devices



Husák, M. (2021, November). Towards a data-driven recommender system for handling ransomware and similar incidents. In 2021 IEEE International Conference on Intelligence and Security Informatics (ISI).

Section 3

Current State: Knowledge Graphs in Cybersecurity

Supporting Technologies – Databases

Graph Database

- Neo4j – popular and widely used (e.g., CRUSOE, CyGraph), technologically mature
- Alternatives - OrientDB, JanusGraph, Dgraph, ...

Query Languages

- Cypher – native to Neo4j
- GraphQL – versatile choice
- SPARQL – query language for RDF databases
- any other – still requires learning a new paradigm (for a newbie)

Plenty of tools, no issues

Existing Ontologies

Several existing ontologies

- UCO – Unified Cybersecurity Ontology
- SEPSES
- BRON
- CRUSOE

Other standards and knowledge bases

- CVE, CWE, CVSS, CPE...
- MITRE ATT&CK
- CAPEC
- DFAX – Digital Forensics Analysis eXpression

Ontology Standardization

CyberOnto Initiative

- proposed by Stéphane Gagnon, Université du Québec en Outaouais
- informal gathering aiming at unifying cybersecurity ontologies
- active in 2022-2023, now inactive, but raised the research questions

MITRE D3FEND

- knowledge graph of cybersecurity countermeasures
- MITRE is a strong player, providing *de facto* standards
- hosts CyberOnto initiative in D3FEND Slack channel

Data Collection

Automated

- Network scanning and monitoring – asset discovery, fingerprinting, topology
- Vulnerability scanning, fetching vulnerability information
- Export from asset management, configuration databases, and other sources
- Automated dependency and criticality detection (partially)

Manual

- Identification of critical infrastructures and dependencies
- Reflecting the organization structure (people, offices, departments)
- Business missions and their dependencies

Issues – Data Quality

Automated

- Network scanning and vulnearibility detection has its limits – but agents cannot be installed everywhere (e.g., in large, heterogeneous networks)
- IT domain is typically OK, but what about IoT and OT? Lack of details
- Problematic dependencies – not everyone has asset inventory, configuration database, identity management system, ...

Manual

- Often time-consuming to gather and insert
- Some details may get obsolete fast, some changes may have vast consequences

Section 4

Future Prospects and Conclusions

Future Trends – Research

Autonomous Cybersecurity Operations

- Replacing incident handler with a machine/tool
- Knowledge graphs as an enabler
- Risky interactions with the environment

Leveraging AI

- Open, active, and interesting area or research
- Learning from the graphs, inferring further (hidden) knowledge

Integration with LLMs

- Asking question in natural language
- Generating responses or reports
- May be worth it in some cases, not reliable in others

Future Trends – Operations

Operational Aspects

- Potential duplicity with asset management, configuration databases, etc.
- Frequent automated scanning of assets is a common practice
- *"Why should we use another tool mostly duplicates what we already have?"*

Finding the Killer Use Case

- What is the key feature of knowledge graph that would make people use KGs?
- Queries and visualizations may not be enough, full automation may be too much
- Decision support, recommender systems, error prevention, ...
- *"Are you sure you want to do this action?"*

Conclusions

Knowledge Graphs

- Well-known in other fields, rising popularity in cybersecurity as well
- All prerequisities are met – storage, ontologies, data collection, ...
- Emerging use cases, analyses, and visualizations
- Current issues – standardization, data quality, commercialization

Future Research and Development

- Research in KG&AI in cybersecurity is booming
- Everything "cyber" and "AI" is cool, but can we actually use it in practice?
- Find the killer use case and stick to it!

MASARYK
UNIVERSITY