Next-Generation Security Entity Linkage: Harnessing the Power of Knowledge Graphs and Large Language Models

Daniel Alfasi, Tal Shapira, Anat Bremler-Barr
Reichman University and Tel Aviv University

Abstract

- The number of reported Common Vulnerabilities and Exposures (CVEs) continues to rise annually
- There are over 211,371 CVEs and 25,093 new vulnerabilities were disclosed in 2022 alone
- Security vulnerability analysis methods rely on manual and time-consuming linking of CVE, Common Weakness Enumeration (CWEs), and Common Attack Pattern Enumeration and Classification (CAPECs)
- Main Objectives:
  - Automated and efficient entity linking
  - Improve predictions accuracy compared to prior works
  - Handling unseen entities and dealing with incomplete data

Datasets

- 4,096 Linux CVEs from 1999-2020, obtained from the MITRE CVE database
- 16,044 CVEs from 1999-2023 acquired from the RedHat Security Database
- We provided the datasets and models for results reproduction

CVE Classification

CVE-2020-12662 and its CWE classification

A network amplification vulnerability was found in Unbound, in the way it processes delegation messages from one authoritative zone to another. This flaw allows an attacker to cause a denial of service or be part of an attack against another DNS server when Unbound is deployed as a recursive resolver or authoritative name server.

CWE-406->CWE-400: Insufficient Control of Network Message Volume (Network Amplification) leads to Uncontrolled Resource Consumption

Method and Results

Methodology

Collecting entities from the CVE, CWE, and CAPEC databases

⇩

Embedding entities description into continuous vector space using OpenAI's Ada Large Language Model

⇩

Generating textual descriptions for entity relations using ChatGPT and embedding them with Ada

⇩

Initializing RotatE with the acquired embeddings

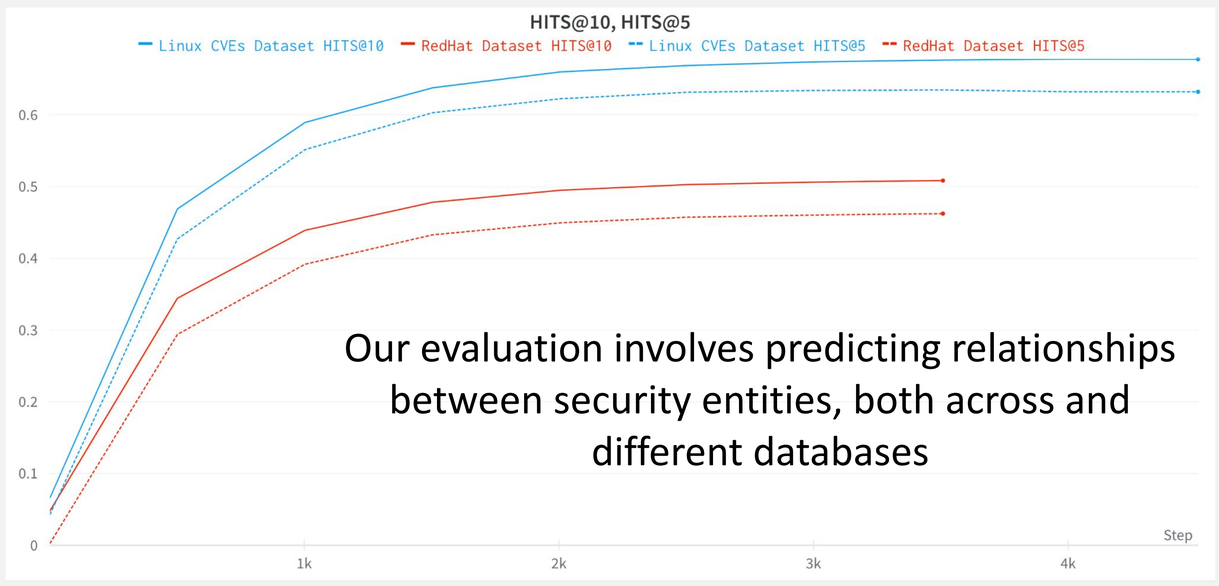
⇩

Evaluating model with inductive link prediction for performance on unseen entities

Model Initialization

Best result achieved with RotatE initialized with Ada embedding

Results



Our evaluation involves predicting relationships between security entities, both across and different databases

Model	MRR	Hit@10	Hit@5	Hit@1
Han et al. [1]	0.25	0.46	0.41	0.13
Xing et al. [2]	0.49	0.65	0.59	0.4
Our approach	0.6	0.7	0.66	0.55

Table 1: Evaluation on Linux CVEs dataset.

Previous Works

- [1] Han et al. "Embedding and Predicting Software Security Entity Relationships: A Knowledge Graph Based Approach"
- [2] Xing et al. "Predicting Entity Relations across Different Security Databases by Using Graph Attention Network"

Our Contribution

- **Using large language model to Improve knowledge graph embedding**
- **Introducing an innovative approach for initializing relations embeddings**
- **Providing comprehensive datasets and models for future benchmarking**

daniel.alfasi@post.runi.ac.il
talshapirala@gmail.com
anatbr@tauex.tau.ac.il

Future Work

In the future we will:

- Extended research of entities and relations representation to improve results
- Prediction of CVEs and CWEs relations with Fine-grained CWEs hierarchy
- Broaden our research by predicting relations between new types of entities, such as Internet Protocols

Scan for additional info