Plan Overview

A Data Management Plan created using DMPonline

Title: Improving the Detection of Cyber Threats using Large Language Models for Network Protocols

Creator: Fidel Cacheda

Principal Investigator: Víctor Carneiro Díaz, Fidel Cacheda Seijo

Data Manager: Víctor Carneiro Díaz, Fidel Cacheda Seijo

Project Administrator: Víctor Carneiro Díaz, Fidel Cacheda Seijo

Affiliation: Other

Template: DCC Template

ORCID iD: 0000-0002-7536-9422

ORCID iD: 0000-0002-6438-1661

Project abstract:

Security Information and Event Management (SIEM) constitutes the state-of-the-art for handling heterogeneous data sources for security analysis. It operates by collecting and combining data from event sources across an organization’s IT and security framework, including host systems, networks, firewalls and antivirus security devices.

The evolution and nature of cyber-attacks, increasingly complex and coordinated, render traditional tools insufficient due to the volume and heterogeneity of events to be monitored. SIEM systems, which enable the collection, analysis, aggregation, and correlation of diverse event sources, have become valuable tools in identifying, tracking, and responding to threats. However, the growing volume and complexity of security data pose challenges for SIEMs to stay up to date, often leading to a high number of false positives, resulting in the phenomenon known as "alert fatigue" in security teams.

On the other side, in recent years, the field of NLP has witnessed a transformative shift with the advent of Large Language Models. These models have demonstrated unparalleled capabilities in understanding, generating, and manipulating human language. The fundamental concept underlying a language model lies in its capacity to anticipate the subsequent word or sub-word (named tokens) based on the text it has observed so far.

The starting hypothesis of this research project is the ability to improve the performance and accuracy of anomaly detection and network attacks on a SIEM by applying and integrating LLM-based models for network traffic.

From a computer communication point of view, networking protocols behave similarly to languages, but in a smaller syntactic and semantic scope. In this sense, we claim that defining specific LLMs for network protocols and traffic, based on predicting the next token, can help improving the detection of cyber anomalies and attacks on a typical digital communication environment.

Expected contributions include: development of specific LLMs based on network protocols for

network-based threat detection; integration of LLM-based techniques on SIEM framework to incorporate integrated analysis of protocols and network traffic.

Design and development of specific early detection metrics for absolute time aware evaluation of cyber-threats; improvement of state-of-the-art accuracy for detection and time aware detection of cyber-threats using LLMs based on networking protocols; and improvement in the effectiveness and efficiency of current SIEMs by using LLM-based techniques and incorporating integrated analysis of protocols and network traffic.

The novelty of our proposal lies on the use of LLMs for network protocols and traffic to improve cyber-threats detection.

ID: 163810

Start date: 01-09-2024

End date: 31-08-2027

Last modified: 14-11-2024

Grant number / URL: PID2023-150794OB-I00

Improving the Detection of Cyber Threats using Large Language Models for Network Protocols

Data Collection

What data will you collect or create?

The data collected will be focused on public datasets used for intrusion detection or computer network attacks or data traffic in general to reuse and exploit existing data. Datasets considered include: CIC-IDS-2017, CIC-IDS-2018 or KDD99.
Also, for the code development a GitLab repository will be used.

How will the data be collected or created?

Datasets will be downloaded from public repositories and stored on a local NAS that all members of the research group can access. Each dataset will be stored in its own folder named as the dataset for easy access. Dataset files will be stored unchanged.
The source code will be stored in a GitLab repository locally availabe.

Documentation and Metadata

What documentation and metadata will accompany the data?

For the datasets, a simple text file will identify the basic information for each dataset available locally including: name, site URL and local folder.
The documentation for the source code will be provided by a Readme.

Ethics and Legal Compliance

How will you manage any ethical issues?

To the best of our knowledge, no ethical issues are associated with the datasets utilized nor the data generated from the experiments in the research project.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The IPR of data generated along the research project is owned by the research team.

Storage and Backup

How will the data be stored and backed up during the research?

The data will be stored on a local NAS managed by the research team. Regular back ups are programmed.
The GitLab repository is part of the CITIC research institute facilities and regular back ups are automatically programmed.

How will you manage access and security?

The local NAS will include accounts for each research team member to provide access control.
In case collaborators need to access some data, a local account with limited permissions (i.e. reading) and access will be provided.
Regarding the GitLab repository, the research team member are provided with an account with the corresponding permits to access or modify the code.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The results of the research project will be retained, at least, for the duration of the project plus another 3 years for any potential future exploitation or use.

What is the long-term preservation plan for the dataset?

The long-term preservation for the data generated in the research project will be done using the resource available by the research team (i.e. NAS) and CITIC research institute (i.e. GitLab repository).

Data Sharing

How will you share the data?

The datasets collected are already publicly available.
Regarding the source code, relevant parts of code will be made publicly available for the research community to replicate, reproduce and reuse our research.

Are any restrictions on data sharing required?

To the best of our knowledge, no restrictions will be included when sharing the source code from the research project. Only a citation to our work will be required.

Responsibilities and Resources

Who will be responsible for data management?

The persons responsible for data management will be the principal investigators of the research project.

What resources will you require to deliver your plan?

To the best of our knowledge, no additional resources are required to implement this data management plan.