Presentation Transcript:

Security Event Log Anonymization and Synthetic Attack Generation for Realistic Security Analyst Training and Intrusion Detection Research

S1:

While the title is lengthy, “Security Event Log Anonymization and Synthetic Attack Generation for Realistic Security Analyst Training and Intrusion Detection Research” summarizes both the project’s primary technology development goals as well as two key use cases.

S2:

The project overview confirms it will include tangible deliverables such as a literature review, software with testable requirements and operational guides as well as extensive analysis of the solution’s effectiveness and ethical concerns related to cyber security event data. The anticipated project duration is approximately 6 months.

S3:

The MSc Cyber Security degree granted by the University of Essex requires completion of an independent project meeting the British Computer Society's technical requirements, demonstrating in-depth knowledge in a specific area and academically defendable research.

To maintain BCS accreditation, all computer science projects must include a practical technology implementation, not just a literature review or industry survey. In addition to a technical artefact such as software, the project must also include a report showing in-depth analysis of the problem being researched, outcomes of experiments or validation testing and identify where the project met, or failed to meet its objectives, weaknesses within the proposed solution and any changes to the project scope that may have occurred.

The MSc Cyber Security project also requires the research topic explores one or more Cyber Security Body of Knowledge (CYBOK) knowledge areas (Essex, n.d.).

S4:

Postgraduate level projects must demonstrate a deeper subject matter analysis and understanding and primarily be completed in a self-directed fashion. This table summarizes these expectations and how the project is intended to meet them.

Advanced understanding of the chosen topic will be achieved through the literature review conducted during the project’s initial phases.

Strong research questions, examined later in this proposal, also help validate problem understanding and identification of gaps in existing research.

knowledge application originality *is taking a less frequently explored path for generating large amounts of security event test data applicable to a specific network environment.*

Most research identified during the preliminary project design use various types of artificial intelligence (AI) to generate anonymous data with limited success. Sections of Sommer and Paxson's seminal 2010 paper (Sommer & Paxson, 2010), echoed in CYBOK section 8.3 (Rashid et al., 2021), outline significant challenges with anomalous detection yet this continues to be a common solution approach.Conversely, in most medium and large organizational networks, millions of real events are generated daily that would be suitable for training and experimenting, but practical anonymization solutions do not appear to be widely researched.

Technical proficiency will be demonstrated by developing a data processing pipeline that performs extraction, transformation, and loading (ETL) operations, then using this pipeline for processing data for experiments and testing.

Initial reflections on potential risks with the project have highlighted the assessment of deanonymization attack vectors and feasibility. This also creates an excellent opportunity to ***demonstrate critical thinking and communication***. Although the Cambridge Analytica scandal is a credible worst-case scenario, (Manokha, 2018), organizational stakeholders must consider the likelihood and impact of internal information leakage against the benefits of cyber security analysts developing intrusion detection skills with representative data.

Continuing with the anonymization attack theme**;** the project will validate which data elements within security events need anonymization to ensure regulatory compliance with common standards like GDPR, PCI DSS and NIST. It will also analyze what types of inferences could be made about an organization simply through metadata analysis of the anonymized datasets.

CyBOK knowledge area 8, Security Operations and Incident management utilizes security event log data extensively, creating the primary link between this project and CyBOK. Sections 5.1 and 6.4 are also relevant to this research project, maintaining a reasonable level of privacy once the data is outside the organization's control is a critical success factor, and improving training for detection of malicious behaviour is the primary driver for the project.

S5:

A literature review completed for this module, researching deep learning for cyber-crime detection revealed artificial intelligence researchers were often hampered by access to large amounts of truly representative data. Although the problems with applying machine learning to cyber security event detection were well understood, (Sommer & Paxson, 2010), a decade later the problem is still not resolved, (Bresniker et al, 2019).

Academic institutions attempting to address the cybersecurity skills gap (ICS2, 2021) are challenged to provide more than theoretical intrusion detection training without access to realistic event data and organizations requiring cybersecurity event analysts are understandably restricting disclosure of internal security event data.

In addition to the lack of realistic benign events, independent intrusion detection research is also hampered by the ability to seed data sets with known intrusions in a controlled manner (Trizna, 2020) since actual cyber attack events represent a very small percentage of overall security event data.

S6:

The first research question summarizes the primary goal of the project to develop a technical solution. One that can accept a large security event log dataset export from an organization, apply anonymization processes to reduce confidentiality concerns and create a corresponding output dataset suitable for training cyber security analysts on realistic looking data or training artificial intelligence models.

The next two questions are intended to drive research into the details of the anonymization processes since generating inferences from large amounts of obfuscated or anonymized data is a legitimate privacy concern, as discussed in Cybok section 5.1 (Rashid et al., 2021).

Question four will evaluate technical approaches to generating log event data that aligns with a publicly documented cyber attacks (Red Canary, 2022) without the need to recreate the attack in a lab environment (Splunk, 2021).

S7:

The proposed project incorporates elements of both information system design and software development which can be conceptually visualized as a UML activity diagram.

Although a little small for reading in this slide, this UML diagram illustrates the main inputs and checkpoints for the extraction, transformation and load (ETL) data processing pipeline which is explained in more detail in the following aims and objectives section.

S8:

The project will develop software that can accept security event log data selected by an organization as input, but from an ethical perspective the organization must be aware of the limitations of anonymization and assess the risk. It is assumed that not all data sources may be acceptable candidates and each organization considering sharing their data will have a unique risk tolerance.

S9:

Although the data owner will be responsible for generating the initial data extraction, the transformation software can be run by either the data owner or the party planning to work with the anonymized data since the software will have an opensource license and be publicly available.

The aim of the transformation process is to output anonymized or obfuscated versions of each record while retaining the time series relationships and log message elements critical to most intrusion analysis. The output itself should retain each record's original log message format rather than an arbitrary, normalized output for two reasons:

- Retaining original record formats allows existing security event log collection solutions to be used, enabling organizations to deploy training environments more easily.

- Additionally, researchers developing feature extraction processes for artificial intelligence experiments can have high confidence the processes will work similarly with new training data.

S10:

Loading large amounts of log data into an investigation platform will typically result in a percentage of records failing to load. Logically, if the original and anonymized datasets generate the same number of errors, the anonymization process has potentially retained all key characteristics.

As a secondary validation, output data can be searched for known input values that were to be anonymized. A negative result would indicate anonymization for the specified data was successful, random sampling could be used to avoid searching for all input values.

S11:

The project requires a comprehensive literature review to further understanding of the problem space and identify current approaches to resolution. While additional research terms will be discovered during the review, the key words in this table form the starting point for the research as well as indicating what aspects of the research will be specifically applicable to the project.

At time of writing there is good awareness that realistic log replication of adversarial actions poses unique challenges due time series restrictions needed to simulate the context of an attack scenario (Yichiet et al., 2022).

Log management problems, exist with synthetic generation and traditional collection methods due to the lack of a common schema for labeling the features within logged events (Elastic, 2022a) requiring cybersecurity subject matter expertise (LaFerrera, 2022) to map various log source syntaxes to common labels suitable for either human or machine-based analysis (Yichiet et al., 2022) (Elastic, 2022)

Since the projects primary goal is to support better training of cybersecurity analysts and the tools they use, the literature review must capture the state of the art which will include identifying widespread commercial solutions since they represent realistic log sources an analyst is likely to encounter.

Data anonymization concerns have been discussed at length in previous sections, inclusion here acknowledges this is an area requiring solid research.

S12:

The literature review was addressed in the previous slide,

The project will begin with data from the researcher's own lab environment and publicly available data sets to develop the initial transformation prototypes. In parallel industry connections will be approached to gauge willingness to provide other data sets that may not necessarily be publicly available but representative of current commercial data sources.

The transformation, anonymization and realism development is anticipated to be an iterative process. Testing at both the anonymization or realism steps could result in revisiting the transformation process as signified by the dashed arrows.

An additional testing output from validation by external parties is included in the project but omitted from the diagram for readability.

Results from both the anonymization testing and security analyst experience testing will be analyzed and findings shared with industry stakeholders to gauge data sharing comfort levels.

S13:

Although organizations are supporting the greater good though their data sharing, it is also in their self interest to help resolve the shortage of skilled cybersecurity analysts in the marketplace. (ISC2, 2021)

That said, these same organizations must have assurances that anonymization measures remove all reasonable likelihood that datasets could be reverse engineered to reveal sensitive system or employee information since that creates a version of the information loss scenario they are actively trying to prevent.

Subject to further research, the current presumption is there will be a trade-off between the usability of anonymized logs and resistance to deanonymization attacks. Organizational stakeholders providing data for anonymization must be fully informed of the trade-offs and limitations.

S14:

A project of this magnitude and scope would likely prove daunting to someone new to the cyber security field. This researcher is in mid-career and has access to the identified resources needed to complete the work within the allotted time frame.

The two biggest risks identified with the project are writing the transformation software components and developing sufficiently robust tests to detect deanonymization issues. To address these risks both tasks have been allocated a great deal of time and are initiated early in the project.

Other tasks such as report writing can begin earlier and proceed in parallel allowing some flexibility with the start dates for external party testing and industry stakeholder interviews.

S15:

An Excel based project planner will be used to track tasks both in table and visual format. Once the module commences, additional milestones that are currently unknown can be added and any related timelines adjusted.