A scalable and practical privacy-preserving framework

ENCRYPT

Cyber Threat Intelligence (CTI) Use-Case

Angelos Papoutsis, apapoutsis@iti.gr, CERTH

Athens, 17/06/2025

encrypt

CTI Use Case-General Information

- **Increasing sophistication of cyber threats:** Threat actors are constantly developing more advanced techniques, which creates a need for improved cybersecurity measures
- **Role of Cyber Threat Intelligence (CTI):** CTI involves gathering knowledge on cyber-threats, which enhances situational awareness and helps to fortify defense strategies
- **Limitations of internal logs:** Organizations typically rely on internal logs that record attacks against their systems, which may not provide sufficient visibility into broader cyber threats
- **Benefits of sharing CTI:** Sharing CTI and security artifacts across organizations can lead to a deeper and broader understanding of threats, potentially improving cybersecurity awareness and resilience
- **Concerns over information sharing:** Organizations often hesitate to share CTI due to the risk of exposing sensitive or confidential data
- **Privacy Preserving Techniques (PETs):** Implementing PETs can help maintain the confidentiality of shared data, alleviating concerns over privacy and encouraging information sharing

CTI Use Case- Technologies

ENCRYPT Supportive Technologies	
Technology Name	**Actions Performed and Description**
Advanced Data Pre-Processing Tool	The tool preprocesses the holders' datasets and identifies Private Identifiable Information (PII).
Knowledge Graphs (KGs)	KGs amplify the PIIs identification procedure

CTI Use Case- Technologies

ENCRYPT Supportive Technologies	
Technology Name	**Actions Performed and Description**
Methodological Framework for the Assessment of the PP Computation Services	The tool helps data owners to: a) assess the organization's privacy, b) perform risk assessment and cost-effective analysis in a potential security breach.
AI- Recommendation System	Cybersecurity domain data holders of the ENCRYPT platform have access to the AI recommendation system via the ENCRYPT's UI, advising them on the trade-offs of the PP they will opt to choose.
Front/Back-end System	The Front-end of the ENCRYPT platform used by the Cybersecurity domain data holders when interacting with the ENCRYPT system and its functionalities.

CTI Use Case- Technologies

ENCRYPT Privacy Preservation Technologies	
Technology Name	**Actions Performed and Description**
Trusted Execution Environment (TEE)	TEE is used for secure computation of sensitive data. Data analysis/pattern identification is performed directly in the TEE environment.
Fully Homomorphic Encryption (FHE)	FHE used by the cybersecurity domain data holders to protect their data. Correlation activities are performed between data gathered and stored from the external world (via MISP channels) and data coming from the TEE, with the view to extracting advanced CTI.

CTI Use Case- Scenario

- Data holders send their data to a TEE
  - ENCRYPTs' Data providers/end-users (DPC, 8BELLS, EXUS)

- Within the TEE, data analysis and pattern identification are performed without exposing the underlying data, identifying potential malicious actions like SQL injection attacks
  - Grouping techniques are applied to data such as IP addresses to summarize actions and detect anomalies

- The TEE enhances CTI by exporting homomorphically encrypted data, ensuring PII remains secure during analysis

- MISP uses the encrypted data to enrich CTI by correlating it with external data sources, such as blacklists

CTI Use Case-Advanced Data Pre-Processing and Preparation

Aim:

- ✓ Develop the data preprocessing and preparation techniques (e.g., data clearing, sampling) required to format the data in the proper way for the application of the PP technologies on them and for facilitating any post-analysis (e.g., Machine Learning)
  - ✓ The fewer the dimensions, the better the implementation of Homomorphic encryption

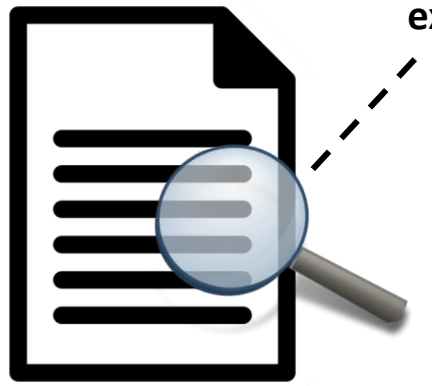
- ✓ Develop a PII extraction module
  - ✓ The PIIs identified on the user's data helps the Recommendation Engine (RE) in the selection of the best privacy technique

CTI Use Case-Advanced Data Pre-Processing and Preparation

- MAC Address
- Full Name
- Gender
- Age
- Weight
- Smoking
- Language
- Nationality
- Marital Status

PIIs Indentification

PII identification using regular expression patterns



- Tax ID
- Social Security Number
- Birthday
- Country
- Deceased
- Deceased Date
- Address
- Email

- Supported Formats: JSON, CSV, Text
- Three domains of ENCRYPT project

CTI Use Case-Advanced Data Pre-Processing and Preparation

- **Regular Expressions (Regex)**
  - Used for pattern-based extraction of standard PII types (e.g., emails, Ips, phone numbers, credit cards, ZIP codes)
  - Applied to structured/tabular data for rule-based detection

- **NERPII Python Library**
  - Performs Named Entity Recognition (NER) on structured datasets
  - Integrates the Presidio framework to detect various sensitive entities such as: CREDIT_CARD, CRYPTO, DATE_TIME, EMAIL_ADDRESS, IBAN_CODE, IP_ADDRESS, LOCATION, PERSON, PHONE_NUMBER MEDICAL_LICENSE, URL

- **Manual Entity Assignment**
  - Custom rules for specific entity types not handled by default models (e.g., ZIPCODE, custom credit card numbers)

Knowledge Graphs' role in the CTI Use Case

- **Input:**
  - A ZIP file from the UI, inside it there are either JSON, TXT, XLSX/CSV files, even more than one
  - PIIs from Preprocessing Tool

- **Activities:**
  - Unzip and identify the data for the exact use case (for example, Threat-Intelligence)
  - Process the various files we get and then export TTL and PIIs

- **Output:**
  - Additional PIIs and send them to the Kafka Broker

CTI Use Case-Data Preparations

- **Input**: A file containing multiple entries

- **Need for cleansing-Aim**: Minimization of the resource overhead of CTI creation, optimizing efficiency and enhancing data integrity for encryption support
  - ✓ Transforming existing timestamps to the STIX format
  - ✓ Removing duplicate entries to optimize resource usage during preprocessing
  - ✓ Fixing entries with merged fields or extra fields (cause: lack or extra commas) – in 8BELLs

    Mar/07/2023 10:38:03 firewall, prerouting: out connection-state:new src-mac 1:1:1:1:, proto ICMP (type 8, code 0), 1.2.3.4->5.6.7.8, len 3
  - ✓ Removing extra characters in the beginning or end of email addresses – DPC: <, >, <- etc.
  - ✓ Fixing MAC address format – EXUS: Mac addresses use “-” as a separator instead of “:”
  - ✓ Minimization of features to support encryption

- **Output**: Cleansed file with multiple entries

Data preprocessing: Automated feature extraction

- **8bells**: Automation is achieved based on the following assumptions:
  - ✓ Source IP: First IP address
  - ✓ Destination IP: Second IP address
  - ✓ NAT IP: Last IP address
  - ✓ Port: Follows each IP with a colon (:)
  - ✓ Start Time: First timestamp
  - ✓ Source MAC: First MAC address
  - ✓ The communication protocol is after the indicator “proto” and before the “(” or “ “” or before “,”. (e.g., “proto **TCP** (ACK)”)

Utilization of regex

- **DPC**: feature extraction is based on the provided CSV format. No further automation can be achieved since the number of email addresses varies – i.e., sender and recipients cannot be identified in the row. (see example below)

- **Exus**: feature extraction is based on the provided JSON format – field/value extraction

Trusted Execution Environment (TEE)

- **Receive Datasets** over a TLS-secured channel terminating inside the enclave
  - ✓ ensures secure in-enclave file storage.

- **Processes the code/data** through multiple steps in a hardware-enforced secure area
  - ✓ Cleans input files
  - ✓ Attempts to process each received file with multiple extractors (8BELLs, DPC, EXUS)
  - ✓ Runs pattern analysis on the extracted data

- Cleanup operation and return results

CTI Creation

- **Input**: A cleaned file

- **Need for CTI creation**:
  - ✓ Each entry is utilized to create a CTI in STIX format

- **Process** (automatic):
  1. Extract features of each entry of the file to a list
  2. Create necessary STIX objects using the features of the list
     - Select specific features to create a STIX object
       - IP addresses are used to create a “ip-addr” STIX objects
       - Indicators STIX objects based on the network traffic
  3. Add the created STIX objects into a STIX Bundle object

- **Output**: Multiple CTI in STIX format. One for each entry of the cleaned file.

CTI Creation

- **Input**: A cleaned file
- **Process**:
  - ✓ Objective: Analyzing entries to uncover patterns of cyber threats
  - ✓ Methodology: Identifying the IP and email address targeted a data provider
  - ✓ Focus: 1) Pinpointing the private IP or email address attacked and timestamp of each incident (statistics). 2) Identification of botnet attacks, and 3) Identification of IP addresses that targeted more than one data providers (EXUS and 8Bells)
- **Output**:
  - ✓ Insights: The IP (EXUS and 8Bells) and email (DPC) addresses that have launched an attack on a particular data provider
  - ✓ Correlation insights: Identification of the IP addresses executed attacks on more than one data provider (EXUS and 8Bells)
  - ✓ Details: Presentation of the entries related to each attack, offering a comprehensive view
  - ✓ Bot-related detection:
    - Each cleansed file is analysed to extract the private infrastructure (e.g., private IP and Email addresses). To identify the private infrastructure, the target of an attack (IP or email address) of each entry is stored in a file:
      - 8Bells: IP addresses are extracted and stored in a file.
      - DPC: Email addresses are extracted and stored in a file.
      - EXUS: addresses are extracted and stored in a file.

CTI Enrichment

- Extracted CTI is analysed to identify Indicators of Compromise (IoCs) of interest
  - ✓ Email address
  - ✓ IPv4 address

- IoCs of interest are analysed by querying the values against well-known analysis platforms
  - ✓ IPQualityScore – Emails, IPs
  - ✓ VirusTotal - IPs
  - ✓ AbuseIPDB - IPs
  - ✓ Hybrid Analysis - IPs

Fully Homomorphic Encryption (FHE)

- **Functionality:** This workflow enables privacy-preserving correlation between attacker IP addresses or emails and a blacklist, ensuring sensitive data is never exposed. The process involves four main tools:

- **CTIEncrypt-K:**
  Extracts and encrypts attacker and victim IP addresses from the generated CTI
- **BLEncrypt-K:**
  Encrypts a blacklist of suspicious IP addresses using the same encryption scheme
- **Server Tool:**
  Performs secure, privacy-preserving correlation to check if the attacker or victim IPs appear in the blacklist, without revealing the actual data
- **Decrypt-K:**
  Decrypts the final correlation results, showing whether each IP was found in the blacklist (O,1), while keeping all other information confidential

- **Key Benefit:**
  Sensitive cyber threat intelligence data (such as attacker/victim IPs) and blacklists are never revealed during processing—only the final yes/no result is disclosed, supporting compliance and privacy requirements in collaborative threat intelligence scenarios.

Thank you!

Angelos Papoutsis, apapoutsis@iti.gr

https://encrypt-project.eu/

encrypt-project

@encrypt_project

encrypt