Data Mining in Incident Response

Challenges and Opportunities

CIRCL
Computer Incident
Response Center
Luxembourg

Alexandre Dulaunoy -
TLP:WHITE

Information Security Education Day

CIRCL

CIRCL
Computer Incident
Response Center
Luxembourg

- The Computer Incident Response Center Luxembourg (CIRCL) is a government-driven initiative designed to provide a systematic response facility to computer security threats and incidents.
- CIRCL is the CERT for the private sector, communes and non-governmental entities in Luxembourg.

Figures at CIRCL

- **1.4GB** of compressed malware sample in a day.
- An average of **2-4TB** per evidence acquisition (disk, memory, ...) including analysis artefacts or duplicate analysis information.
- **1.2GB** of compressed network capture from the operational honeypot network (HoneyBot).
- **10-20 million** records added or updated in the Passive DNS in a day.
- **500 million** of X.509 certificates in the Passive SSL.

Do we have an issue with such volume of data?

- Storage price goes down and it will probably follow this trend.
- Storing huge amount of data is still practical and CSIRT can usually handle it.
- **Write-speed on disk** is still the main limitation (e.g. wire speed increased faster than the I/O).

Where are the real challenges in a day-to-day CSIRT operation?

- **12000 requests per second** to lookup records in the Passive DNS.
- Collections (network, disk, memory) by CSIRTs are often **unstructured**,
- sources of data are **uncontrolled and unstrusted**
- and **incomplete**.

Homogeneous data versus heterogeneous data

- 45TB of **normalized and homogeneous network capture** is fundamentaly different than 45TB of **black-hole network capture**.
- Discarding is easy in normalized traffic.
- In incident response, **protocol errors or incomplete packets are part of the potential attacks**.
- Parser errors and exceptions are more common on an untrusted and uncontrolled data sources.
- Data mining capabilities highly depend of the **data structuration** (e.g. exfiltration channels are rarely respecting the network layers).
- If the structuration is close to zero, more human pre-analysis is required.

What are the key factors in incident response?

- **Reduce workload** for the analysis (e.g. a full file-system forensic analysis of a standard system can take up to 10 days).
- Allow **fast lookup** in the data collected and processed.
  - Easier the access of correlation is, faster is the exclusion or inclusion of data.
  - Dynamic feedback on the data from the users (what are the most queried records?).
- Reduce false positive but **false negative reduction is more important** (e.g. can you miss an evidence in a critical case?).

How do we try to improve?

- Data-structure allowing **fast lookup** and fast update/counting.
  - Bitindex, Bloom filters, HyperLogLog...
  - Space efficient in-memory key/value store.
- **Parallel processing** of large datasets introduces challenges in checkpointing and updates (e.g. a crash of a parser is not uncommon from untrusted datasets).
  - Simple "parallel processing" frameworks versus complex frameworks (e.g. "limiting the cost of bootstrapping", memory usage and overhead of a framework).

Improving with the feedback loop

- The greatest benefit for data mining is to introduce **human feedback early**.
- Analysts discover outliers, errors or even missing data.
- Feedback can be used to improve algorithms, data structuration (e.g. 4th iteration of the CIRCL Passive SSL data structure) or query interfaces.

How to get analysts feedback?

- Integrate lookup services in the tools used by the analysts.
- Provide multiple UI to promote the reuse of the datasets.
- Support the classification of the results (e.g. a source of classified dataset).
- $\rightarrow$ MISP, malware information and threat sharing platform, is developed to support this.

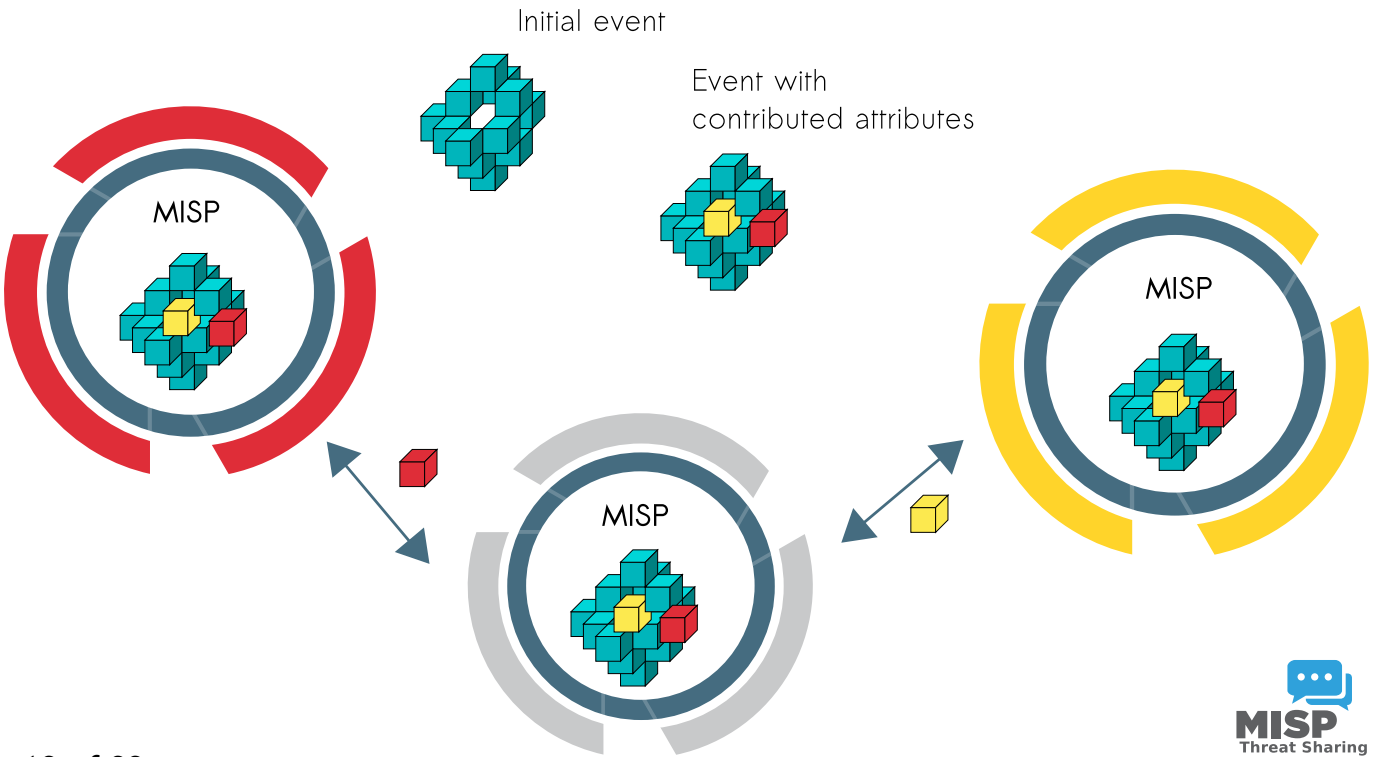
Quick MISP introduction

- MISP[1] is an IOC and threat indicators sharing free software.
- MISP has **many functionalities** e.g. flexible sharing groups, automatic **correlation**, **expansion and enrichment modules**, free-text import helper, event distribution and collaboration.
- CIRCL operates multiple MISP instances with a significant user base (around 400 organizations and more than 1000 users).
- After some years of trial-and-error, we explain the background behind current and new **MISP features**.

[1] https://github.com/MISP/MISP

MISP core distributed sharing functionality

- MISP's core functionality is sharing where everyone can be a consumer and/or a contributor/producer.
- Quick benefit without the obligation to contribute.
- Low barrier access to get acquainted to the system.



Development based on practical user feedback

- There are many different types of users of an information sharing platform like MISP:
  - **Malware reversers** willing to share indicators of analysis with respective colleagues.
  - **Security analysts** searching, validating and using indicators in operational security.
  - **Intelligence analysts** gathering information about specific adversary groups.
  - **Law-enforcement** relying on indicators to support or bootstrap their DFIR cases.
  - **Risk analysis teams** willing to know about the new threats, likelyhood and occurences.
  - **Fraud analysts** willing to share financial indicators to detect financial frauds.

Events and Attributes in MISP

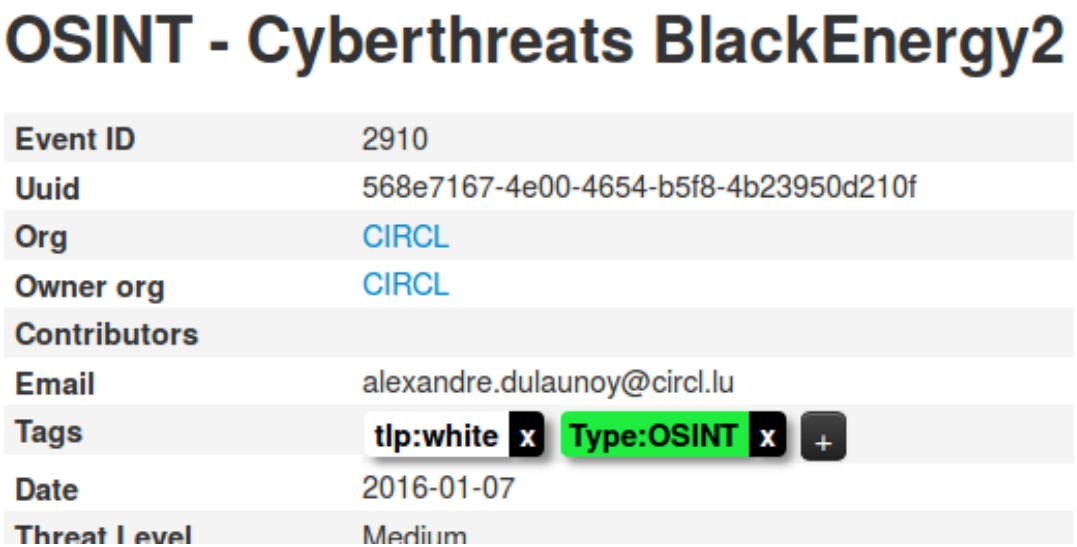
- MISP attributes[2] initially started with a standard set of "cyber security" indicators.
- MISP attributes are purely **based on usage** (what people and organizations use daily).
- Evolution of MISP attributes is based on practical usage and users (e.g. recent addition of the **financial indicators** in 2.4).
- In version 3.0, MISP objects will be added to give the freedom to the **community to create new and combined attributes** and share them.

[2] attributes can be anything that helps describe the intent of the event package from indicators, vulnerabilities or any relevant information

Helping Contributors in MISP

- Contributors can use the UI, API or using the freetext import to add events and attributes.
  - Modules existing in Viper (a binary framework for malware reverser) to populate and use MISP from the vty or via your IDA.
- Contribution can be direct by creating an event but **users can propose attributes updates** to the event owner.
- **Users should not be forced to use a single interface to contribute**.

From Tagging to Flexible Taxonomies



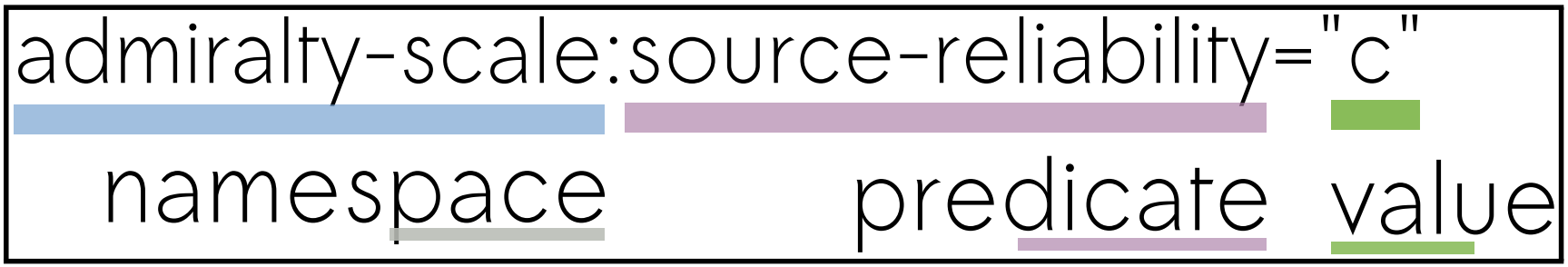
OSINT - Cyberthreats BlackEnergy2

Event ID	2910
Uuid	568e7167-4e00-4654-b5f8-4b23950d210f
Org	CIRCL
Owner org	CIRCL
Contributors	
Email	alexandre.dulaunoy@circl.lu
Tags	tlp:white x Type:OSINT x +
Date	2016-01-07
Threat Level	Medium

- Tagging is a simple way to attach a classification to an event.
- In the early version of MISP, tagging was local to an instance.
- After evaluating different solutions of classification, we build a new scheme using the concept of machine tags.

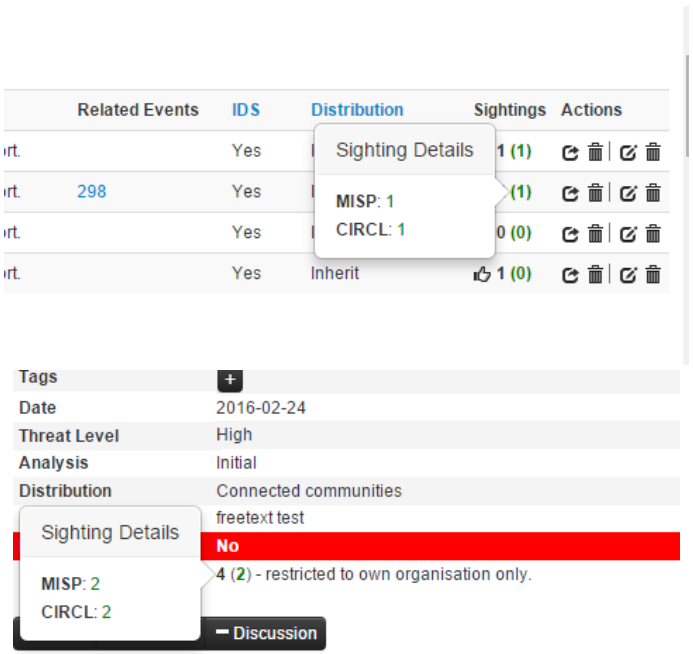
Machine Tags

- Triple tag or machine tag was introduced in 2004 to extend geotagging on images.



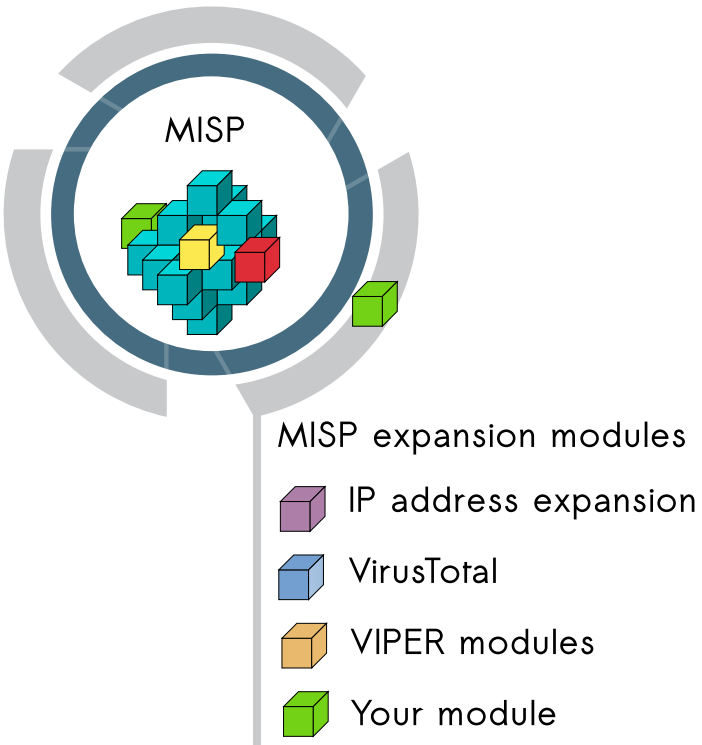
- A machine tag is just a tag expressed in way that allows systems to parse and interpret it.
- Still have a human-readable version:
  - admiralty-scale:Source Reliability="Fairly reliable"

Sightings support



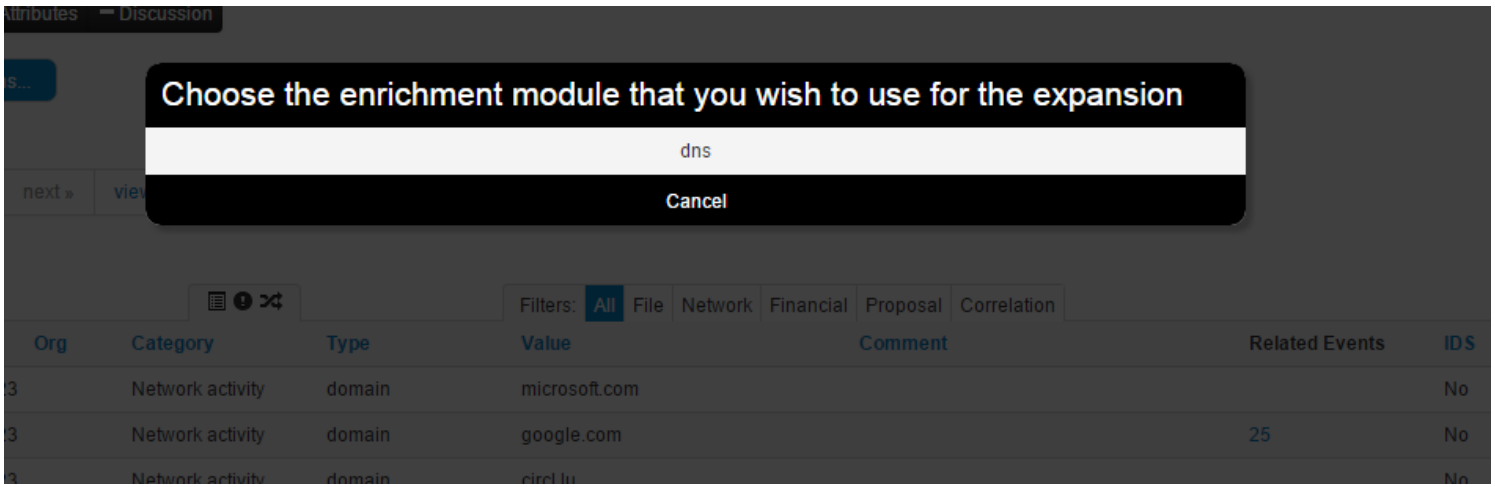
- Sightings allow users to notify the community about the activities related to an indicator.
- Refresh time-to-live of an indicator.
- Sightings can be performed via API, and UI including import of STIX sighting documents.
- Many research opportunities in scoring indicators based on users sighting.

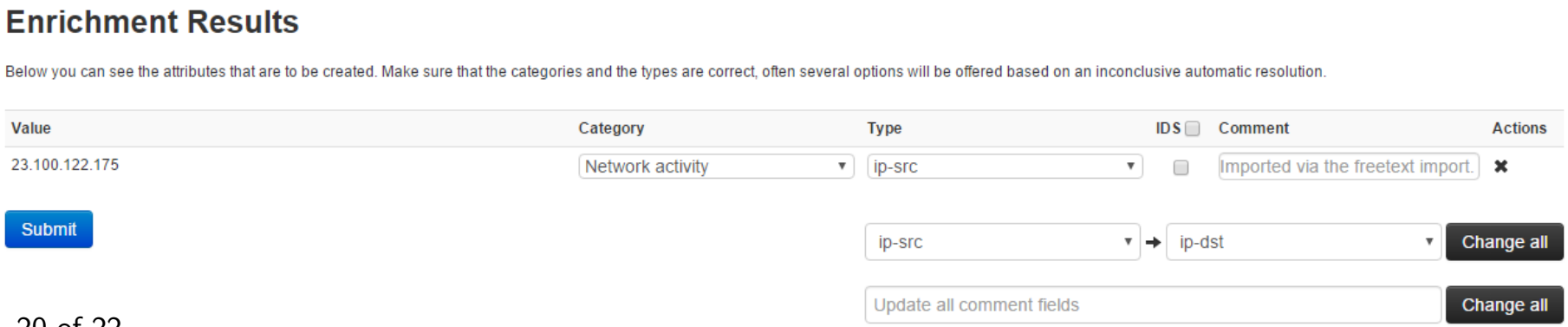
MISP modules - extending MISP with Python scripts



- Extending MISP with expansion modules with zero customization in MISP.
- A simple ReST API between the modules and MISP allowing auto-discovery of new modules with their features.
- Benefit from existing Python modules in Viper or any other tools.
- Current modules include: Passive Total, Passive SSL/DNS (CIRCL), CVE expansion...

MISP modules - How it's integrated in the UI?





Conclusion

- Data mining is a core activity in incident response and forensic analysis.
- Many challenges remain in order to reduce the **time-to-process** and improve the accessibility of the data-sets.
- Information sharing is a way to couple the **cross-validation of datasets**, improving usage and finally improving the data mining processes (collection, filtering, aggregation and query).
- Ongoing research ideas on the operational MISP platforms like:
  - Gamification of information sharing (more people contribute...).
  - Sharing of privacy-aware data structure.
  - Scoring models of information correlation.

Q&A

- https://github.com/CIRCL/
- https://github.com/MISP/ - https://www.circl.lu/
- info@circl.lu - research projects and partnerships
- PGP key fingerprint: CA57 2205 C002 4E06 BA70 BE89 EAAD CFFC 22BD 4CD5