CIRCL
Computer Incident
Response Center
Luxembourg

AIL Project - Practical and Efficient Data-Mining of Chats, Suspicious Websites, Forums and Tor Hidden-Services

CIRCL - Virtual Summer School 2025

https://ail-project.org/

Alexandre Dulaunoy - alexandre.dulaunoy@circl.lu
Aurelien Thirion - aurelien.thirion@circl.lu

July 17, 2025

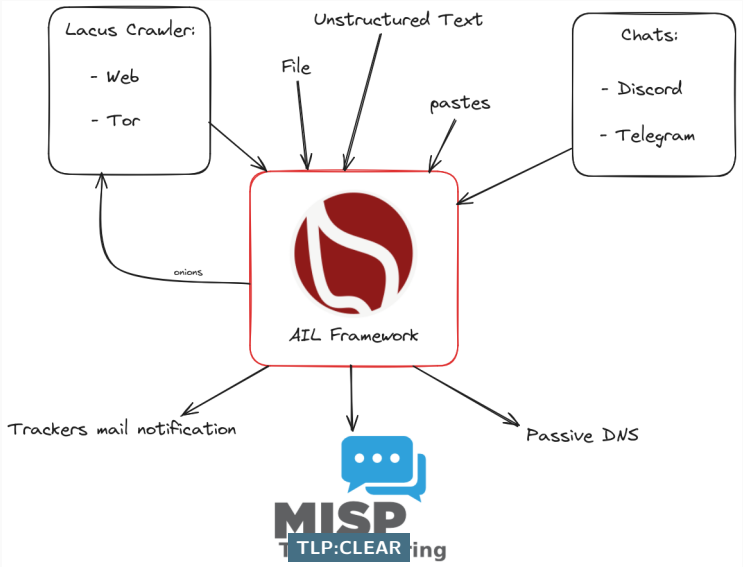
CIRCL https://www.circl.lu

Background

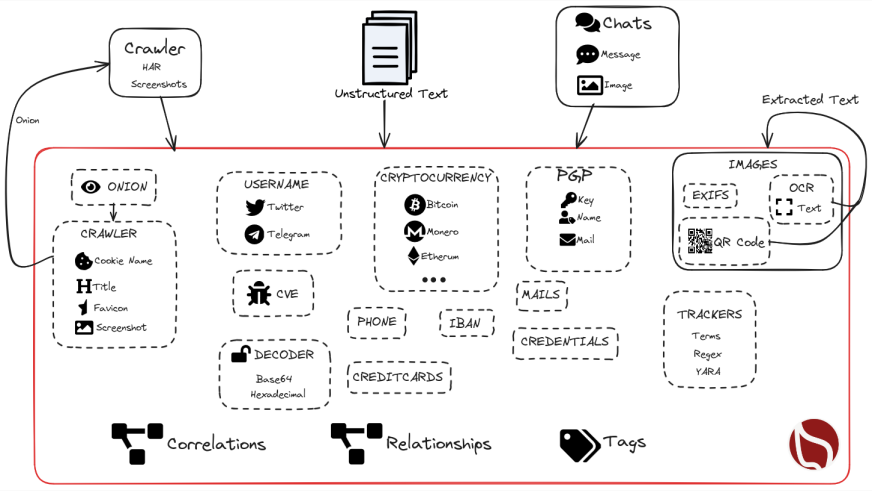
- Over the past years, CIRCL has developed the AIL project[1] to fulfill our needs at CIRCL in intelligence gathering and analysis.
- AIL features an extensible Python-based framework for the **analysis of unstructured information**, collected either through an advanced crawler manager or from various feeders, including social networks and custom feeders.
- The AIL Project is an **open-source** framework[2] comprising various modules designed for the **collection, crawling, digging, and analysis of unstructured data**.

[1] https://www.ail-project.org/
[2] https://github.com/ail-project

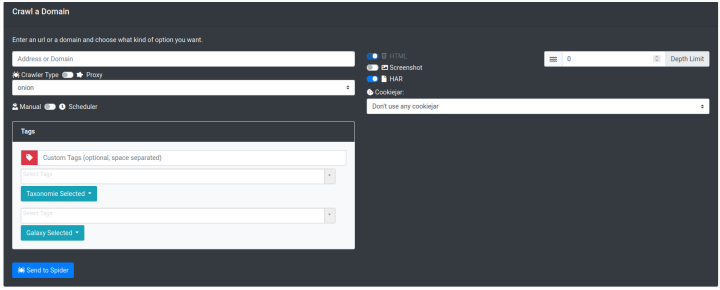
High Level Overview



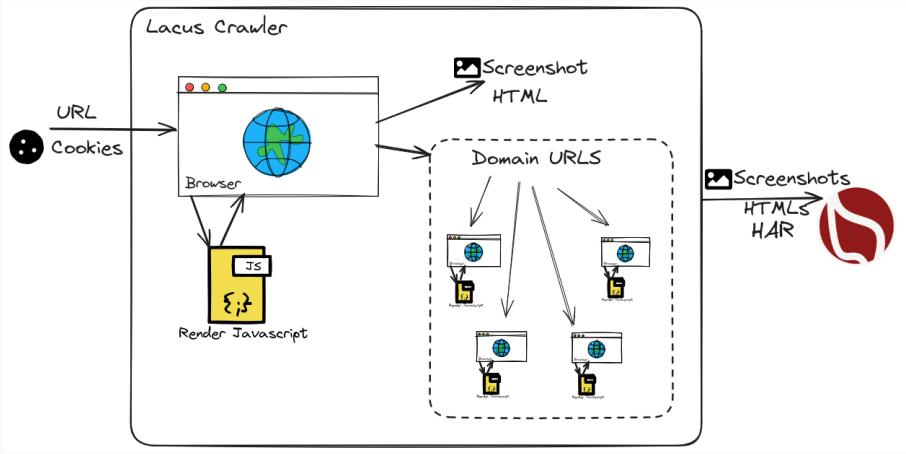
Analysis of unstructured information



- Crawling can be a challenging task, for example, gathering all the blog posts from ransomware groups[3], which can be demanding for an analyst.
- AIL offers a crawling feature that can **initiate regular crawls using a standard spawned browser**.



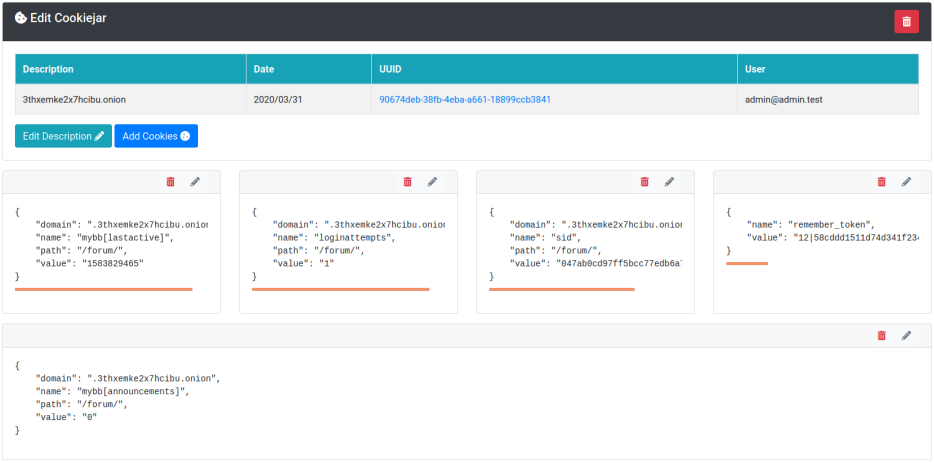
[3] https://www.ransomlook.io/



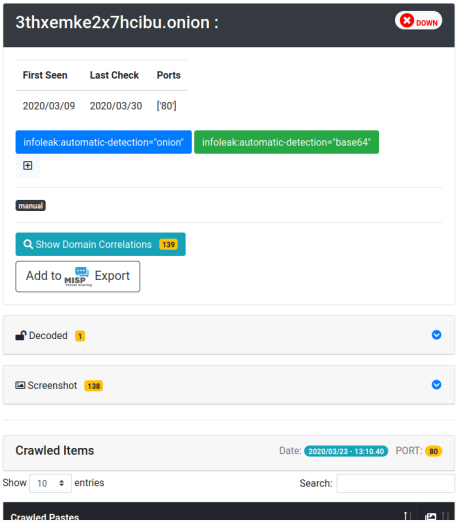
[4] https://github.com/ail-project/lacus

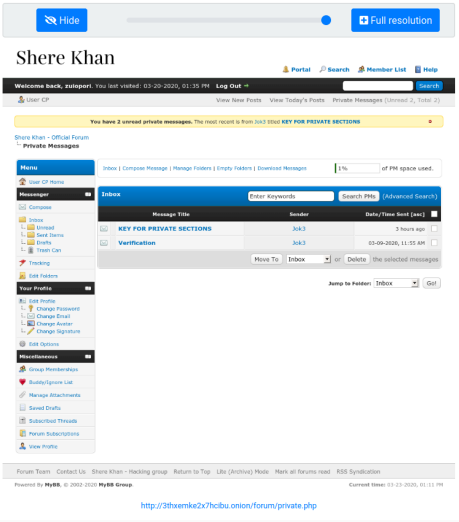
Crawler: Cookiejar

Use your cookies to login and bypass captcha



Crawler: Cookiejar

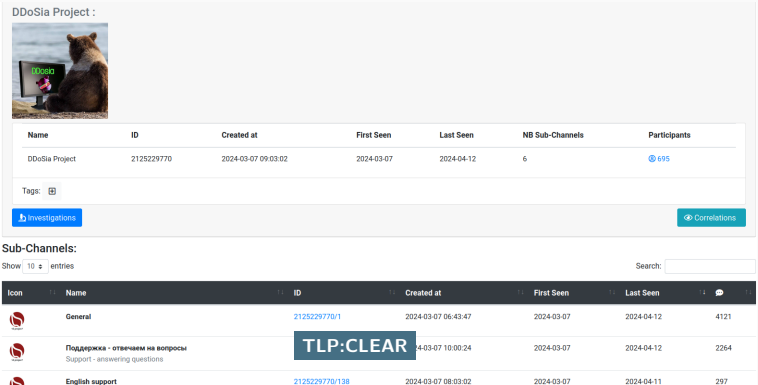




TLP:CLEAR

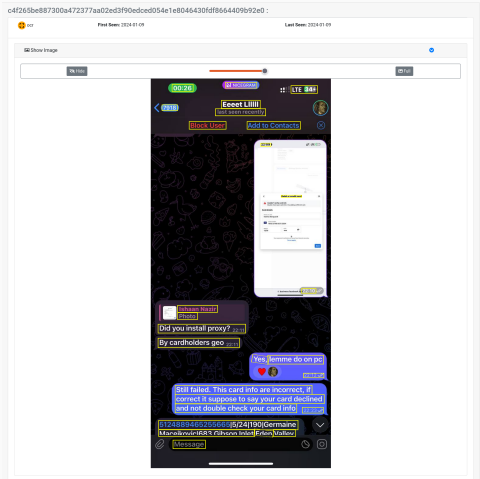
Collection - Automate Collection

- Collecting data from various chat sources can be a **tedious task for analysts**.
- AIL offers a set of feeders (e.g., Telegram, Discord) that can be used to subscribe to chat channels.
- All the **collected messages are then processed and analyzed** within AIL's *processing* and *analysis* stages.



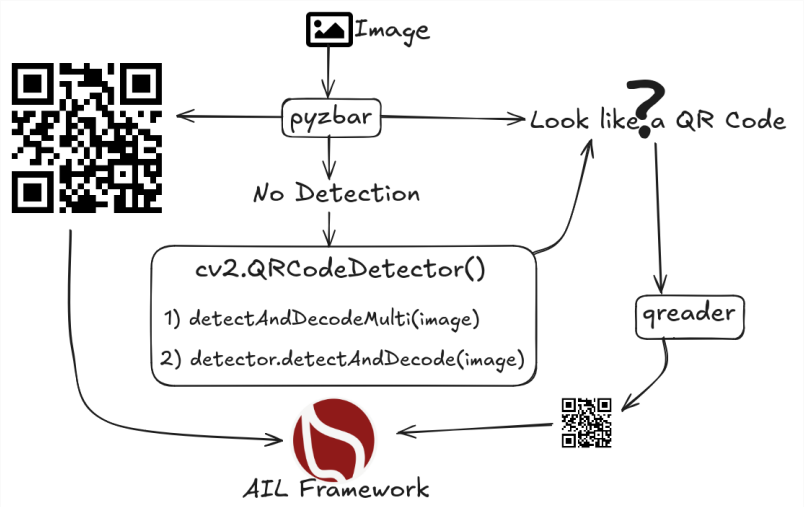
TLP:CLEAR

OCR: Optical Character Recognition

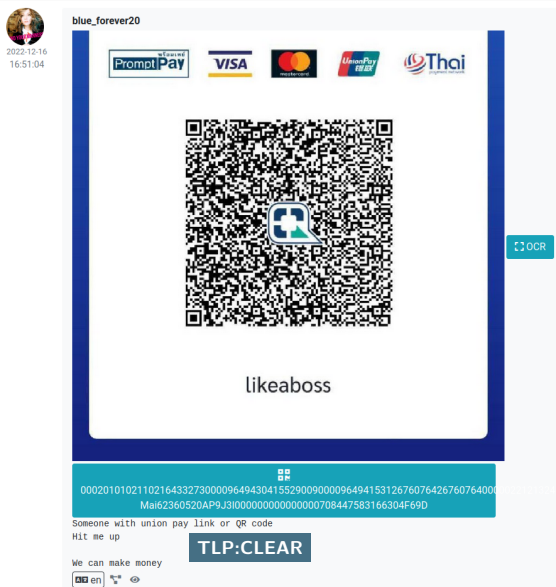


- Threat actors are often verbose and frequently **share extensive details in private channels**.
- Many messages contain screenshots and images.
- Text detection and extraction are performed across 80+ languages using a CRNN (Convolutional Recurrent Neural Network).
- **Enables keyword-based matching and detection**.

QR Code Extractor



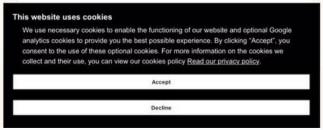
QR Codes



TLP:CLEAR

Bar Codes

OCR 8745860537300061927671300649 13



Cool man its done as always , I appreciate for all you do beloved❤️ .
You have brought a lot of ppl to riches and I will forever be loyal, God bless you ..I wish I knew this earlier...🏦

en

AIL Framework: Features

- Extracting **credit card numbers, credentials, phone numbers, . . .**
- Extracting and validating potential **hostnames**
- Submission to threat sharing and incident response platforms (**MISP** and **FlowIntel**[5])
- **Tagging**[6] for classification and searches
- Terms, sets, regex, and YARA **tracking, occurrences, and history**
- Archives, files, and raw **submission** from the UI
- Correlation engine based on PGP ID, cryptocurrencies, decoded (Base64, . . .), usernames, cookie names, and many selectors to find relationships
- And many more

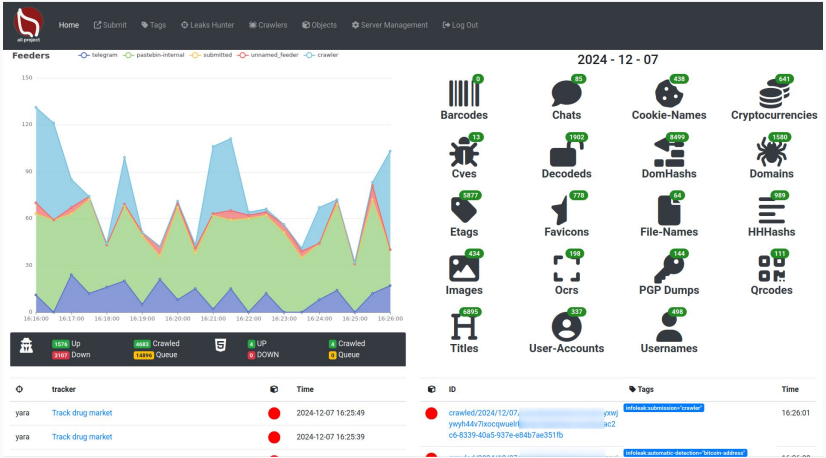
[5] https://github.com/flowintel/flowintel
[6] Relying on MISP taxonomies and galaxy

Live Trackers & Retro Hunt

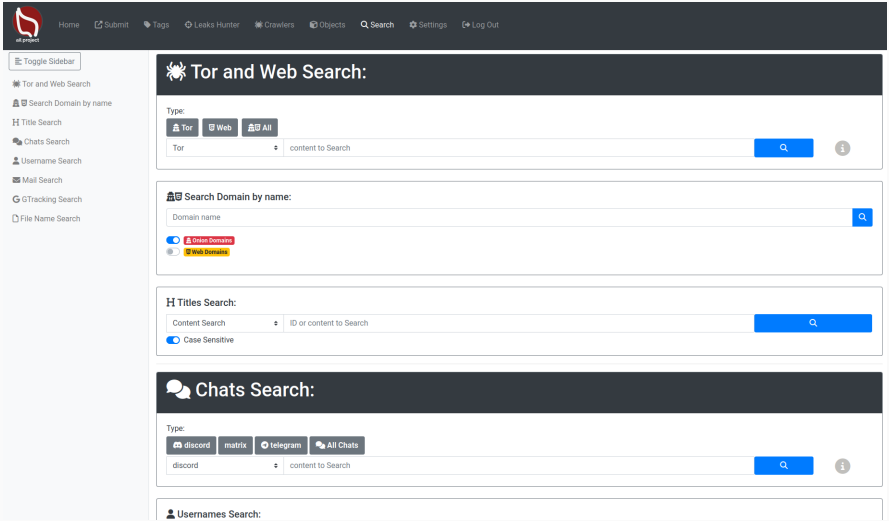
- Search and monitor specific keywords/patterns
  - Automatic tagging
  - Email/webhook notifications
- Track Word
  - ddos
- Track Set
  - booter, ddos, stresser; 2
- Track Regex
  - circl\lu
- **YARA rules**
  - `https://github.com/ail-project/ail-yara-rules`

Live Demo

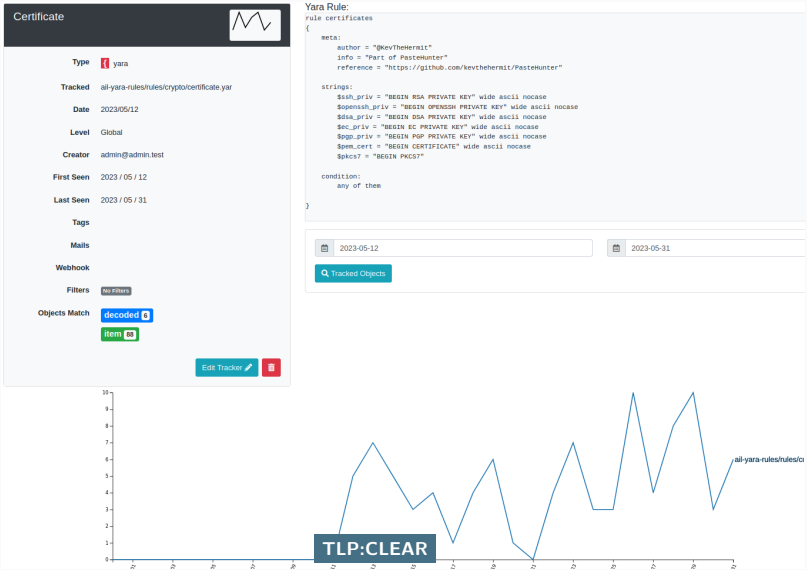
Dashboard


Home
Submit
Tags
Leaks Hunter
Crawlers
Objects
Server Management
Log Out
Feeders
telegram
pastebin-internal
submitted
unnamed_feeder
crawler
2024 - 12 - 07
Barcodes
Chats
Cookie-Names
Cryptocurrencies
Cves
Decodeds
DomHashs
Domains
Etags
Favicons
File-Names
HHHashs
Images
Ocrs
PGP Dumps
Qrcodes
Titles
User-Accounts
Usernames
Up
Down
Crawled
Queue
UP
DOWN
Crawled
Queue
tracker
Time
yara
Track drug market
2024-12-07 16:25:49
yara
Track drug market
2024-12-07 16:25:39
ID
Tags
Time
16:26:01

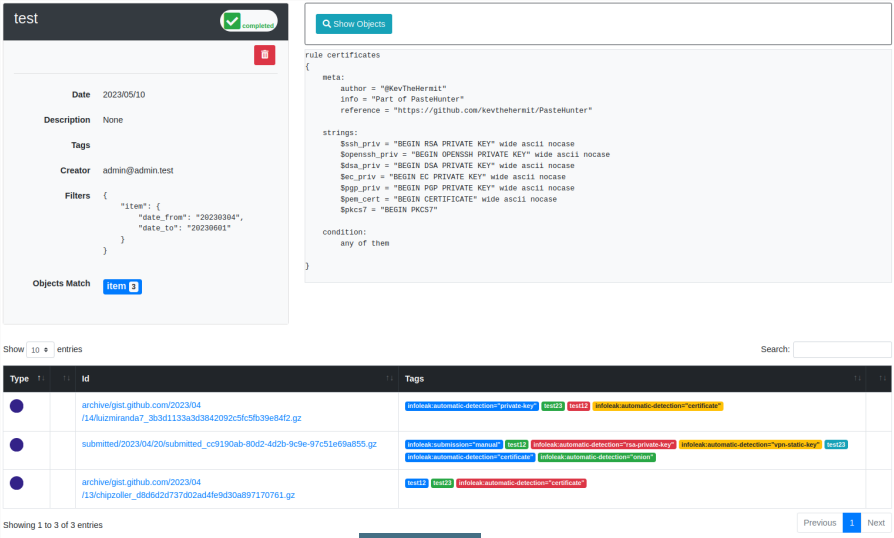
Search


Home
Submit
Tags
Leaks Hunter
Crawlers
Objects
Search
Settings
Log Out
Toggle Sidebar
Tor and Web Search
Search Domain by name
Title Search
Chats Search
Username Search
Mail Search
GTracking Search
File Name Search
Tor and Web Search:
Type:
Tor
Web
All
Tor
content to Search
Search Domain by name:
Domain name
Onion Domains
Web Domains
Titles Search:
Content Search
ID or content to Search
Case Sensitive
Chats Search:
Type:
discord
matrix
telegram
All Chats
discord
content to Search
Usernames Search:

YARA Tracker

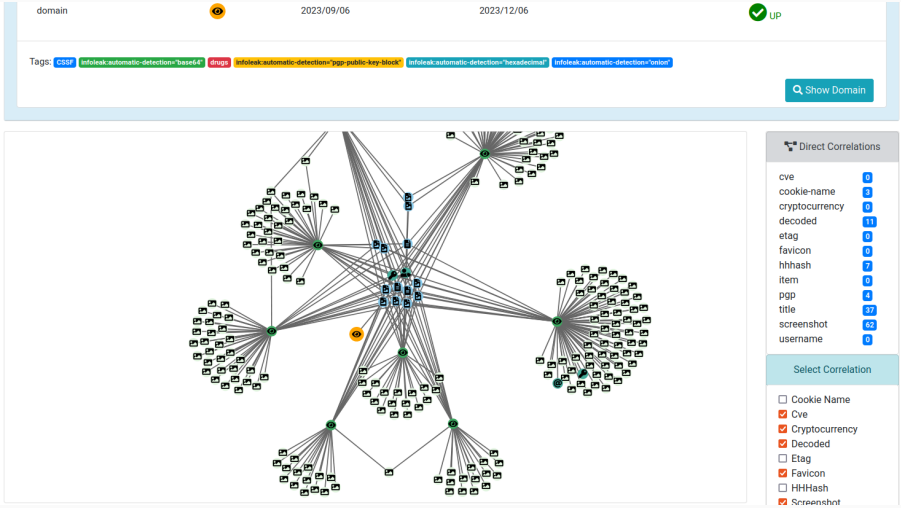


Retro Hunt



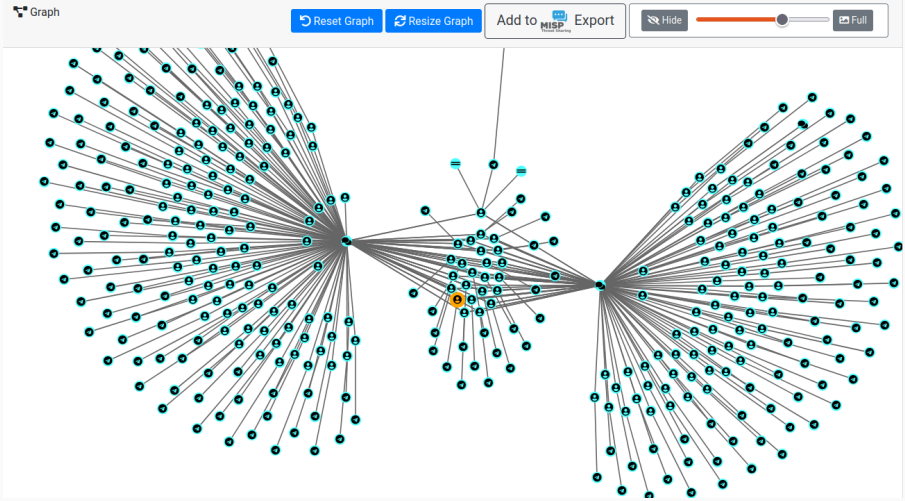
TLP:CLEAR

Correlations and relationship



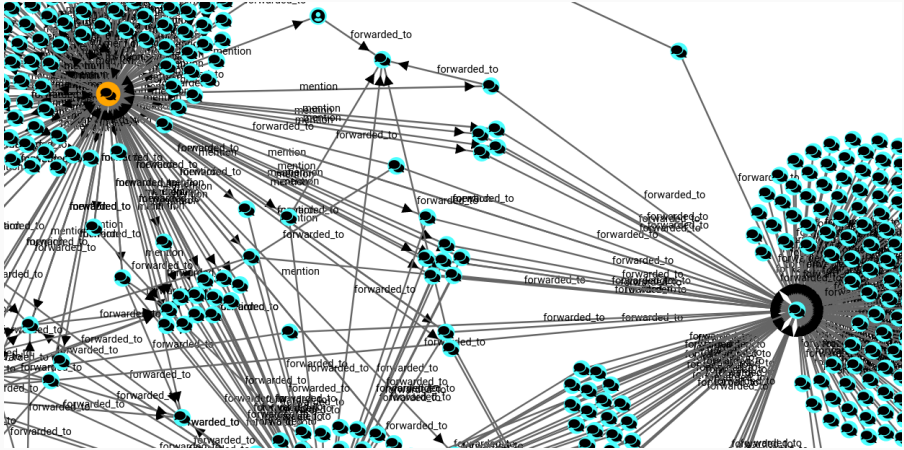
TLP:CLEAR

User Correlation - Common Chats

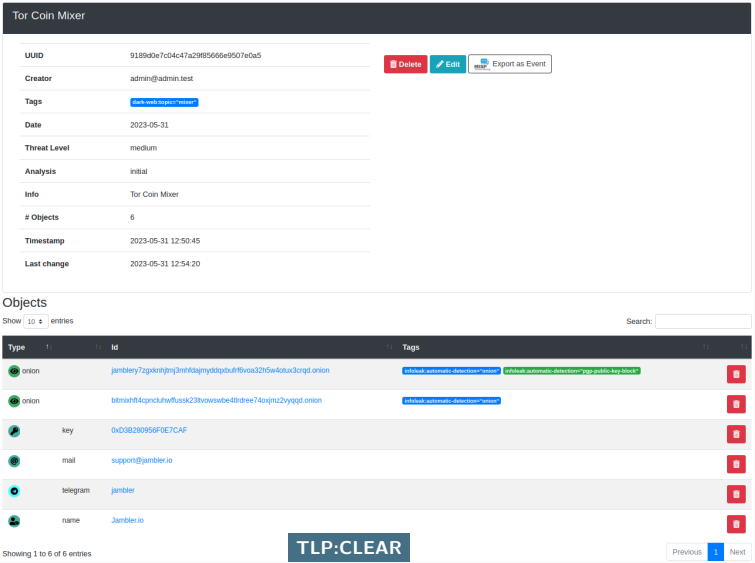


TLP:CLEAR

What are the Relationships Between Chats Group?



Investigations



TLP:CLEAR

AIL Framework: Extensible Capabilities

- Extending AIL to add a new **analysis module** can be done in 50 lines of Python.
- The framework **supports multi-processors/cores by default**. Any analysis module can be started multiple times to support faster processing during peak times or bulk import.
- **Multiple** concurrent **data inputs**.
- Tor Crawler (handles cookies and authentication).
- Feeders: Discord, Telegram, ...

Ongoing Developments

- **Mail Search – Next Release**
- **Lacus crawler improvements:** proxy selection, local cache, and cookie state transfer
- **Translate search:** support for searching in other languages
- **Advanced video processing and extraction**
- **MISP export with new correlation types**
- **Automatic geolocation**

Links

- AIL project website `https://ail-project.org`
- AIL project open source framework `https://github.com/ail-project`
- Training materials `https://github.com/ail-project/ail-training`
- Online chat `https://gitter.im/ail-project/community`

Thank you for your attention

- AIL project[7] : `https://github.com/ail-project/ail-framework`
- For questions, contact: info@circl.lu

[7] All techniques and indicators mentioned in these slides are implemented in the AIL project, using an instance backed by a three-year dataset collected from Tor hidden services and various social networks.