AIL Framework for Analysis of Information Leaks

Improving semantic analysis of dark web for operational and strategic CTI

CIRCL
Computer Incident
Response Center
Luxembourg

Alexandre Dulaunoy
alexandre.dulaunoy@circl.lu
Aurelien Thirion
aurelien.thirion@circl.lu
Jean-Louis Huynen
jean-louis.huynen@circl.lu

info@circl.lu

May 20, 2022

AIL Project

- AIL Project[1] is an open source framework to **collect**, **crawl**, **dig** and **analyse unstructured data**.
- The framework can be used to find information leaks, intelligence, insights and much more.

[1] `https://www.ail-project.org/`

AIL Project History

- AIL initially started as an **internship project** (2014) to evaluate the feasibility to automate the analysis of (un)structured information to find leaks.
- In 2019, AIL framework is an **open source software** in Python. The software is actively used (and maintained) by CIRCL and many organisations.
- In 2020, AIL framework is now a complete project called **ail project**[2].

[2] `https://github.com/ail-project/`

Common and Original Use-Cases

- **Check** if mail/password/other sensitive information (terms tracked) leaked;
- **Detect** reconnaissance of your infrastructure;
- **Search** for leaks inside an archive;
- **Monitor** and crawl websites or Tor hidden services;

Support CERT and Law Enforcement activities

- Proactive investigation: leaks detection;
    - List of emails and passwords;
    - Leaked Databases;
    - AWS Keys;
    - Credit-cards;
    - PGP private keys;
    - Certificate private keys;
- Feed Passive DNS or any passive collection system;
- Discover CVE and PoC of vulnerabilities most used by attackers;

AIL Framework - Current Capabilities

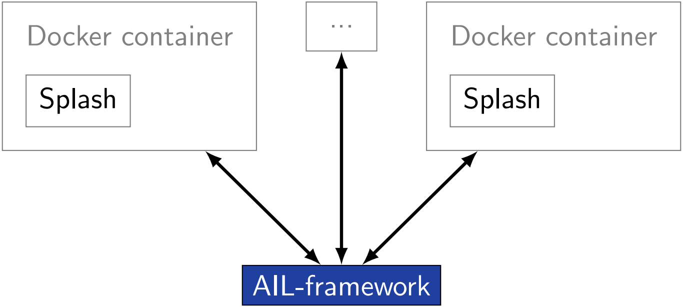
- Extending AIL to add a new **analysis module** can be done in 50 lines of Python;
- The framework **supports multi-processors/cores by default**. Any analysis module can be started multiple times to support faster processing during peak times or bulk import;
- **Multiple** concurrent **data input**;
- Tor Crawler (handle cookies authentication);
- **Feeder model** to extend crawling or collection (From Telegram, Discord, CT-logs to GitHub commits);

AIL Framework: Current features

- Extracting **credit cards numbers, credentials, phone numbers, ...**;
- Extracting and validating potential **hostnames**;
- Keeps track of **duplicates**;
- Submission to threat sharing and incident response platform (**MISP** and **TheHive**);
- **Full-text indexer** to index unstructured information;
- **Tagging** for classification and searches;
- Terms, sets, regex and YARA **tracking and occurences**;
- Archives, files and raw **submission** from the UI or API;
- PGP, Cryptocurrency, Decoded (Base64,...) and username correlation;
- And many more...

Crawler capabilities

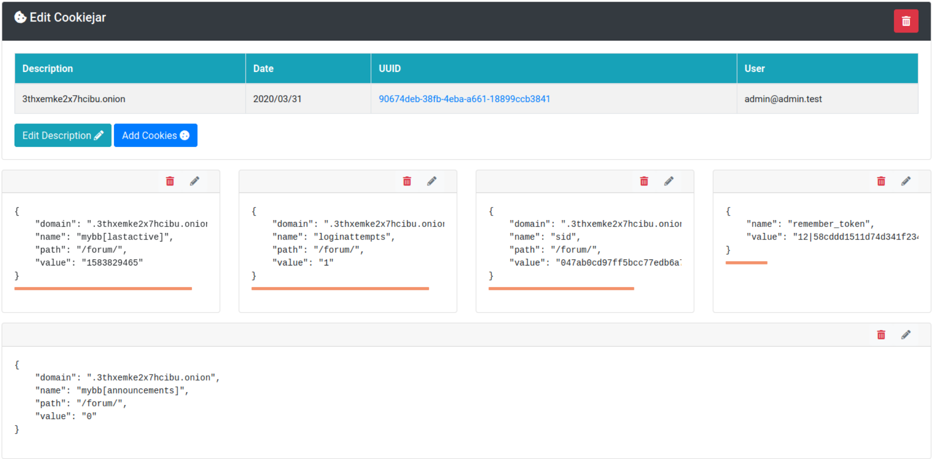
- Crawlers are used to navigate on regular website as well as .onion addresses (via automatic extraction of urls or manual submission);
- Splash[3] (”scriptable” browser) is rendering the pages (including javascript) and produce screenshots (HAR archive too);



[3] Next version of AIL → https://github.com/microsoft/playwright

Crawler: Cookiejar

Use your cookies to login to bypass CAPTCHA



Edit Cookiejar

Description	Date	UUID	User
3thxemke2x7hcibu.onion	2020/03/31	90674deb-38fb-4eba-a661-18899ccb3841	admin@admin.test

```
{
    "domain": ".3thxemke2x7hcibu.onio
    "name": "mybb[lastactive]",
    "path": "/forum/",
    "value": "1583829465"
}
```

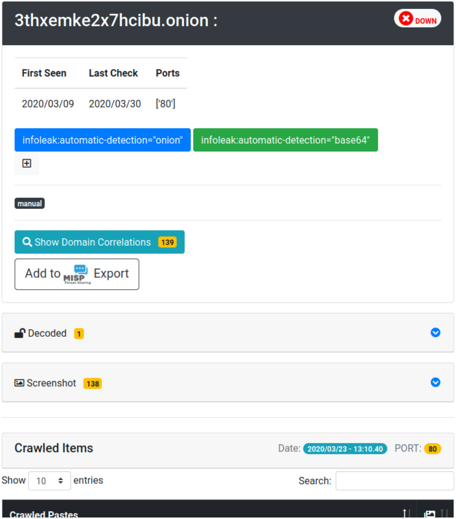
```
{
    "domain": ".3thxemke2x7hcibu.onio
    "name": "loginattempts",
    "path": "/forum/",
    "value": "1"
}
```

```
{
    "domain": ".3thxemke2x7hcibu.onio
    "name": "sid",
    "path": "/forum/",
    "value": "047ab0cd97ff5bcc77edb6a
}
```

```
{
    "name": "remember_token",
    "value": "12|58cddd1511d74d341f23
}
```

```
{
    "domain": ".3thxemke2x7hcibu.onion",
    "name": "mybb[announcements]",
    "path": "/forum/",
    "value": "0"
}
```

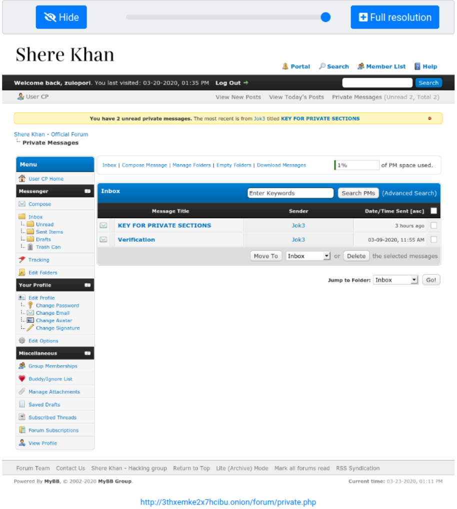
Crawler: Cookiejar



3thxemke2x7hcibu.onion :

First Seen	Last Check	Ports
2020/03/09	2020/03/30	['80']

DOWN
infoleak:automatic-detection="onion"
infoleak:automatic-detection="base64"
manual
Show Domain Correlations 139
Add to MISP Export
Decoded 1
Screenshot 138
Crawled Items
Date: 2020/03/23 - 13:10.40
PORT: 80
Show 10 entries
Search:
Crawled Pastes



Shere Khan

Welcome back, zulopori. You last visited: 03-20-2020, 01:35 PM Log Out

Private Messages (Unread 2, Total 2)

You have 2 unread private messages. The most recent is from Jok3 titled KEY FOR PRIVATE SECTIONS

Shere Khan - Official Forum
Private Messages

1% of PM space used.

Inbox

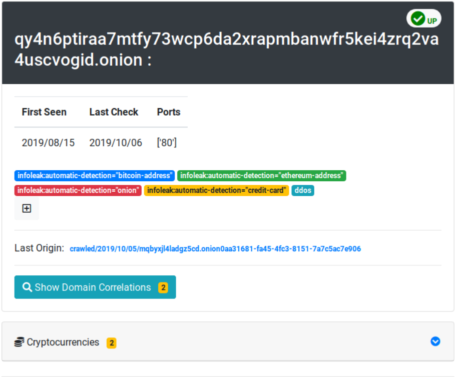
Message Title	Sender	Date/Time Sent [asc]
KEY FOR PRIVATE SECTIONS	Jok3	3 hours ago
Verification	Jok3	03-09-2020, 11:55 AM

Powered By MyBB, © 2002-2020 MyBB Group.

Current time: 03-23-2020, 01:11 PM

http://3thxemke2x7hcibu.onion/forum/private.php

Crawler: Example DDoS Booter

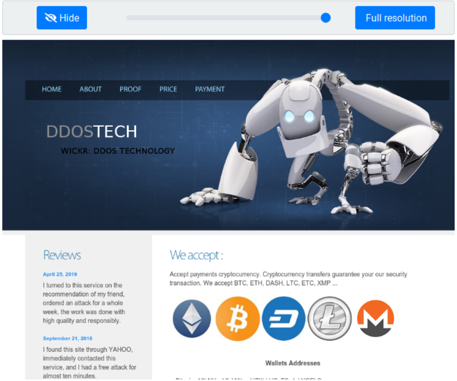


qy4n6ptiraa7mtfy73wcp6da2xrapmbanwfr5kei4zrq2va4uscvogid.onion :

First Seen	Last Check	Ports
2019/08/15	2019/10/06	['80']

infoleak:automatic-detection="bitcoin-address" infoleak:automatic-detection="ethereum-address" infoleak:automatic-detection="onion" infoleak:automatic-detection="credit-card" ddos

Last Origin: crawled/2019/10/05/mqbyxjl4ladgz5cd.onion0aa31681-fa45-4fc3-8151-7a7c5ac7e906



Reviews

April 25, 2019

I turned to this service on the recommendation of my friend, ordered an attack for a whole week, the work was done with high quality and responsibly.

September 21, 2018

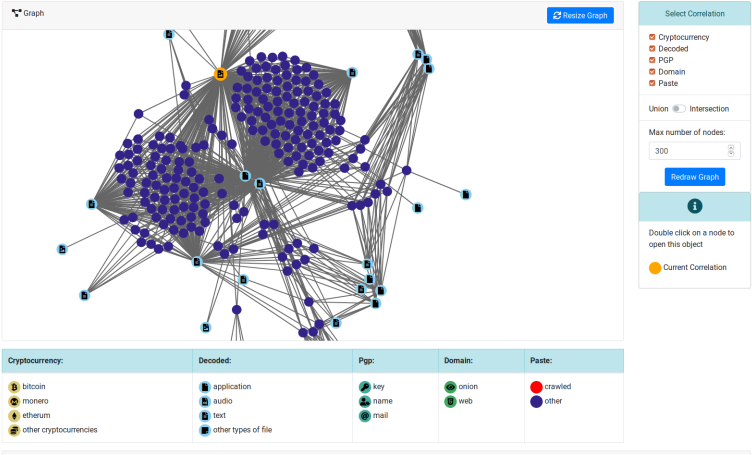
I found this site through YAHOO, immediately contacted this service, and I had a free attack for almost ten minutes.

We accept :

Accept payments cryptocurrency. Cryptocurrency transfers guarantee your our security transaction. We accept BTC, ETH, DASH, LTC, ETC, XMP ...

Wallets Addresses

Correlations and Relationships



☑ Cryptocurrency
☑ Decoded
☑ PGP
☑ Domain
☑ Paste

Union ○ Intersection

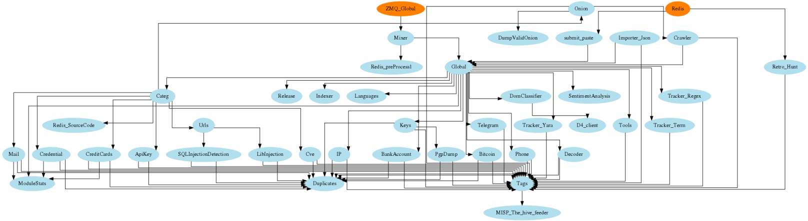
Max number of nodes: 300

Double click on a node to open this object

● Current Correlation

Cryptocurrency:	Decoded:	Pgp:	Domain:	Paste:
bitcoin	application	key	onion	🟥 crawled
monero	audio	name	web	● other
etherum	text	mail		
other cryptocurrencies	other types of file			

AIL Project - a Modular Framework



JTAN and Future of the AIL Project

Joint Threat Analysis Network (JTAN) is an European CEF-funded project to provide a data-stream exchange network. The following activities will performed:

- Improve AIL framework to produce **operational and strategic CTI**.
- Integration of AIL with JTAN and other communities.
- Interconnecting JTAN (**AIL CTI exchange**) with data streams at pan-european level.

Problem Statement and Current Limitation of AIL framework

- **Significant analysts' fatigue**.
- Data retention and lifetime of data.
- Lack of topic-based detection, location-based detection and classification (versus current pattern matching).
- Unstructured information and difficulty to produce operational threat intelligence (e.g. **actionable**) and strategic (e.g. **readable**) insight.

Implementation Steps

- **Gradual changes** in AIL to add required functionalities to support the objectives.
- **Time-memory trade-off** can be challenging to ensure a functional framework.
- Evaluation and integration of new modules in AIL based on time-memory comparisons.
- **Semantic** aspects are challenging due to the diverse data sources, unstructured data and languages seen.

Short-term Developments

- AIL synchronisation and interconnection (first version with AIL sync published version 4.0) → Evolution foreseen for JTAN Task 8.
- Data retention, lifetime and **decay of items** in AIL (ongoing development).
- "Stars project"/collection of **labelled items** created by analysts (active development) → required for the semantic analysis task.
- Extending the crawling/collecting in AIL to support other overlay networks such as I2P, Nym or other mixnet.

Semantic Analysis

- Review of state-of-the-art and design-implementation papers concerning topic-based classification.
- Review and testing of related **open source packages supporting natural and non-natural language processing**.
- Testing corpus creation on specific category in cybersecurity (e.g. ransomware related activities).
- Collecting existing corpus of documents and/or labels from JTAN partners.

Long-term Developments

- **Integrating semantic modules** analysis in AIL framework based on positive results and tests.
- Support of binary analysis[4] to extract meta-data in addition to textual item analysis.
- Label items based on guessed or deduced geographic location from textual and binary items.

[4] The focus is only to ensure the extraction of textual representative of an item. It's not full-fledged binary analysis framework.

Collaboration

- If you have **specific corpus** of documents[5] and/or classification labels to share, this will help the semantic topic-based work/implenentation.
- Recent version of AIL[6] now works like the MISP project[7], synchronisation can be created to **build community of sharing**.
- **Join our JTAN hackathon**[8] - 2nd and 3rd June 2022 in Luxembourg and/or remote.

[5] public or non-public

[6] `https://www.ail-project.org/`

[7] `https://www.misp-project.org/`

[8] `https://twitter.com/circl_lu/status/1527307514651369474`