

# Darknet and Social Network Monitoring

Introduction to Challenges, Concepts and Data Mining of the Deep Web



**CIRCL**

Computer Incident  
Response Center  
Luxembourg

Alexandre Dulaunoy

[alexandre.dulaunoy@circl.lu](mailto:alexandre.dulaunoy@circl.lu)

Aurelien Thirion

[aurelien.thirion@circl.lu](mailto:aurelien.thirion@circl.lu)

Jean-Louis Huynen

[jean-louis.huynen@circl.lu](mailto:jean-louis.huynen@circl.lu)

[info@circl.lu](mailto:info@circl.lu)

October 15, 2021



**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.

## CIRCL Missions 1/2

---

- Provide a **systematic response** facility to ICT-incidents;
- Support national ICT users to **recover quickly and efficiently** from security incidents;
- **Minimize ICT incident-based losses**, theft of information and disruption of services for the private sector;

## CIRCL Missions 2/2

---

- Gather information related to incident handling to better **prepare future incidents** management and provide optimized protection for systems and data;
- **Coordinate communication** among national and international incident response teams during security emergencies and to help prevent future incidents;
- Provide a security related **alert and warning system** for organisations in Luxembourg and abroad;
- Foster **knowledge and information** exchange in cybersecurity **lead the development of MISP project**;

## Links

---

- AIL project: <https://github.com/ail-project>
- AIL framework:  
<https://github.com/ail-project/ail-framework>
- Training materials:  
<https://github.com/ail-project/ail-training>
- Online chat: <https://gitter.im/ail-project/community>

# Ethics in Information Security and Cybersecurity

---

- The materials and tools presented can open a significant numbers of questions regarding ethics;
- Our researches and tools are there for education, supporting the public good and improve incident response;
- We ask all users and participants to **follow ethical principles and act professionally**<sup>1</sup>.

---

<sup>1</sup><https://www.acm.org/code-of-ethics>

<https://www.first.org/global/sigs/ethics/ethics-first>

## Privacy, AIL and GDPR (PII) - An Ethical Question

---

- Many modules in AIL can process personal data and even special categories of data as defined in GDPR (Art. 9).
- The data controller is often the operator of the AIL framework (limited to the organisation) and has to define **legal grounds for processing personal data**.
- To help users of AIL framework, a document is available which describe points of AIL in regards to the regulation<sup>2</sup>.

---

<sup>2</sup>[https:](https://www.circl.lu/assets/files/information-leaks-analysis-and-gdpr.pdf)

[//www.circl.lu/assets/files/information-leaks-analysis-and-gdpr.pdf](https://www.circl.lu/assets/files/information-leaks-analysis-and-gdpr.pdf)

# Objectives



## Our objectives

---

- Provide a quick overview to darknet, deep web, collection and cyber threat intelligence lifecycle;
- Review different **collection mechanisms** and **sources**;
- Some practical examples of criminal activities and their use of modern technologies;
- Show the benefits of developing open source tools to monitor web pages, pastes, forums and hidden services;
- Quick introduction to the open source AIL project;

# Introduction

## Concepts - Deep Web

---

- **Deep Web** is the part of World Wide Web not indexed or directly accessible by standard web search-engines;
- This can be content hidden from **crawlers** by requiring a specific access and this can includes private social media, password-protected forums or content protected by different measures such as paywalls or specific security interface to access the information;
- A large portion of content accessible via Internet is part of the deep web<sup>3</sup>.

---

<sup>3</sup>also called invisible web, hidden web or non-indexed web

## Concepts - darknet

---

- **Darknet** is an overlay network running on top of Internet requiring specific software to access the network and its services;
- Tor, I2P and Freenet are the most commonly used ones. Many are used for hidden services access and some for proxy access to the Internet;
- There are **legitimate use-cases** for such network but also many **illegal or criminal usage**.

# Collection and Sources

---

- Collection (mainly OSINT<sup>4</sup> or covert/ clandestine sources) is the act of gathering manually or automatically data from different sources;
- **Determining and maintaining the sources:**
  - Hidden services (on Tor) such as forums, market places, chatrooms, public site<sup>5</sup>...
  - Social network (e.g. Twitter) from Twitter<sup>6</sup> to Instagram.
  - Chat and discussion forum from Discord, Telegram<sup>7</sup> and private hidden one on Tor or other overlay networks.
  - News and security reports<sup>8</sup>.

---

<sup>4</sup>public or open source sources

<sup>5</sup><https://github.com/ail-project/ail-splash-manager>

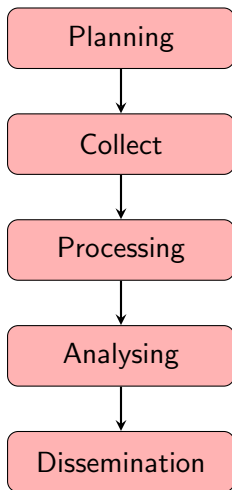
<sup>6</sup><https://github.com/ail-project/ail-feeder-twitter>

<sup>7</sup><https://github.com/ail-project/ail-feeder-telegram>

<sup>8</sup><https://github.com/ail-project/ail-feeder-atom-rss>

## Lifecycle of collection and analysis

---



# Collecting, processing and analysing content - web pages

---

- Building a search engine on the web is a challenging task because:
  - it has to crawl webpages,
  - it has to make sense of **unstructured data**,
  - it has to **index** these data,
  - it has to provide a way to retrieve data and structure data (e.g. correlation).
- Doing so on Tor is even more challenging because:
  - services don't always want to be found,
  - parts of the dataset have to be discarded.
- in each case, it requires a lot of bandwidth, storage and computing power.

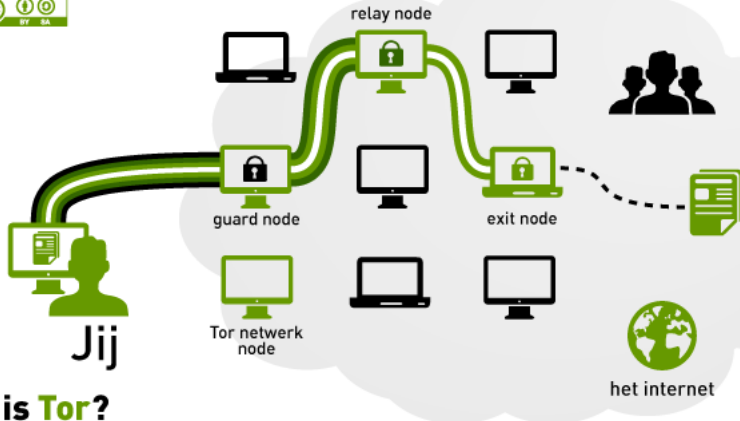
## Demo of analysed content



Tor - a detailed overview of an overlay network

# Concepts - tor<sup>9</sup>

- tor makes use of **onion routing** to obfuscate user identify,

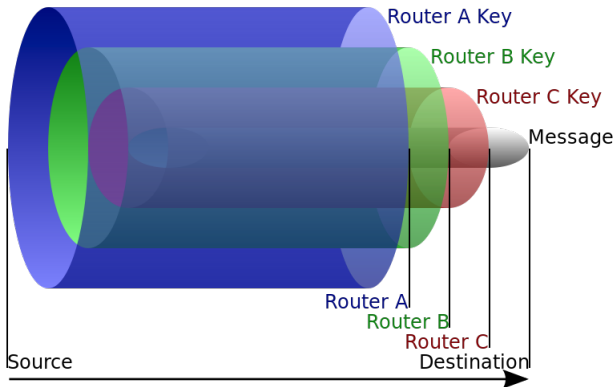


## wat is Tor?

<sup>9</sup>licence: Bits of Freedom CC BY-SA 3.0

# Concepts - tor<sup>10</sup>

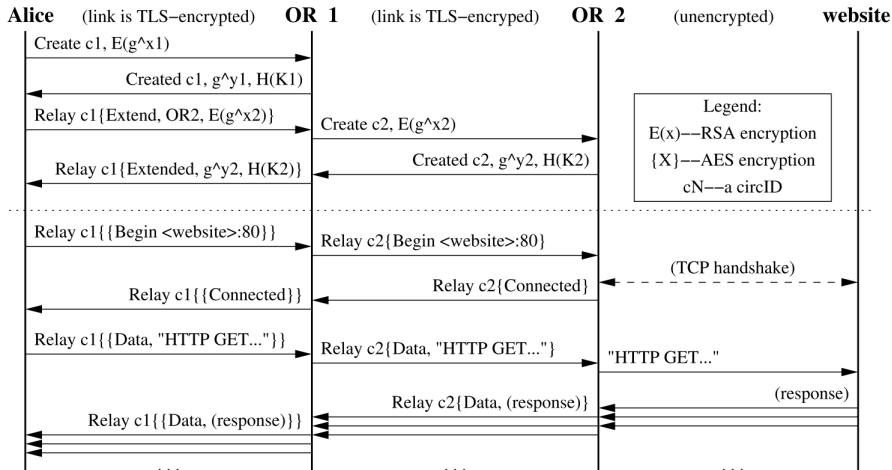
---



---

<sup>10</sup>licence: Harrison Neal CC BY-SA 3.0

# Concepts - tor<sup>11</sup>



<sup>11</sup>licence: Roger Dingledine, Nick Mathewson, Paul Syverson CC BY 3.0

## Concepts - tor

---

- tor provide hidden services: addresses in **.onion**,
- one can only reach such service when one knows its address,
- hidden services' information are stored in a **Distributed Hash Table**,
- these are really interesting for attackers as:
  - these are anonymous,
  - these can be provided through NATs,
  - these can be moved easily.