The Evolution of Cybercrime Through the Lens of Cryptocurrencies

Ioannis Arakas

Thesis submitted in partial fulfillment of the requirements for the

Masters' of Science degree in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, 700 13 Heraklion, Crete, Greece

Thesis Advisor: Prof. *Evangelos P. Markatos*

This work has been performed at the University of Crete, School of Sciences and Engineering, Computer Science Department.

The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science (ICS).

University of Crete
Computer Science Department

The Evolution of Cybercrime Through the Lens of Cryptocurrencies

Thesis submitted by
Ioannis Arakas
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

THESIS APPROVAL

Author: _____
Ioannis Arakas

Committee approvals: _____
Evangelos P. Markatos
Professor, Thesis Supervisor

_____
Ioannis Tzitzikas
Professor, Committee Member

_____
Kostas Magoutis
Associate Professor, Committee Member

Departmental approval: _____
Polivios Pratikakis
Associate Professor, Director of Graduate Studies

Heraklion, February 2024

The Evolution of Cybercrime Through the Lens of Cryptocurrencies

Abstract

In the face of escalating cyber threats such as ransomware attacks, internet scams, and email extortions on a global scale, accurately estimating the overall damage remains a daunting challenge. The diversity in the forms and currencies involved in these cybercrimes complicates efforts to comprehend the full extent of the inflicted harm. Existing reporting mechanisms primarily shed light on incidents disclosed by victims, leaving a substantial number of attacks and pilfered funds unaccounted for.

The cybersecurity landscape hosts numerous Computer Security Incident Response Teams (CSIRTs) and Blockchain Threat Intelligence Platforms, yet their isolated operations contribute to a fragmented data landscape. This fragmentation hinders a holistic understanding of potential threats, exacerbated by major platforms like Chainalysis withholding data from public access. The resultant lack of integration and limited accessibility to data from key players pose significant barriers to accurately assessing the scope and severity of security incidents.

Remarkably, a notable proportion of these cybercrimes unfolds within blockchain environments. Paradoxically, criminals' attempts to conceal identities often unveil critical information. By tracing attackers' wallet addresses, a comprehensive timeline of the crime emerges, from inception to the dispersal of ill-gotten funds. This methodology facilitates quantifying the scale of the crime in cryptocurrencies, such as Bitcoin, with the potential for conversion to conventional currencies. Crucially, it enables the comprehensive tracking of all funds amassed by attackers, regardless of official reporting.

Motivated by these challenges and opportunities, we have developed a system that systematically collects, processes, and visualizes public datasets. This approach enhances the understanding and assessment of the impact of cybercrimes, particularly within blockchain realms, addressing the current limitations in estimating the overall damage caused by these sophisticated threats.

Η εξέλιξη του Κυβερνοεγκλήματος μέσω της ανάλυσης κρυπτονομισμάτων

Περίληψη

Μπροστά στην κλιμάκωση των απειλών στον κυβερνοχώρο, όπως οι επιθέσεις ransomware, οι διαδικτυακές απάτες και οι εκβιασμοί μέσω ηλεκτρονικού ταχυδρομείου σε παγκόσμια κλίμακα, η ακριβής εκτίμηση της συνολικής ζημίας παραμένει μια τρομακτική πρόκληση. Η ποικιλομορφία των μορφών και των νομισμάτων που εμπλέκονται σε αυτά τα εγκλήματα στον κυβερνοχώρο περιπλέκει την κατανόηση της συνολικής ζημίας που προκαλείται. Τα υπάρχοντα εργαλεία εστιάζουν στις καταγγελίες από τα θύματα, αφήνοντας έναν σημαντικό αριθμό επιθέσεων και κλεμμένων κεφαλαίων χωρίς να καταγράφονται.

Το τοπίο της κυβερνοασφάλειας φιλοξενεί πολυάριθμες ομάδες αντιμετώπισης περιστατικών ασφάλειας υπολογιστών και πλατφόρμες πληροφοριών για απειλές που συσχετίζονται με το Blockchain, ωστόσο οι απομονωμένες δραστηριότητές τους συμβάλλουν σε ένα κατακερματισμένο τοπίο δεδομένων. Αυτός ο κατακερματισμός εμποδίζει την κατανόηση των πιθανών απειλών, γεγονός που επιδεινώνεται από μεγάλες πλατφόρμες όπως η Chainalysis που αποκρύπτουν δεδομένα από τη δημόσια πρόσβαση. Όλα αυτά θέτουν σημαντικά εμπόδια στην ακριβή εκτίμηση του μεγέθους και της σοβαρότητας αυτών των περιστατικών.

Είναι αξιοσημείωτο ότι ένα σεβαστό ποσοστό αυτών των εγκλημάτων στον κυβερνοχώρο εκτυλίσσεται σε περιβάλλοντα blockchain. Παραδόξως, οι προσπάθειες των εγκληματιών να αποκρύψουν τις ταυτότητες τους συχνά αποκαλύπτουν κρίσιμες πληροφορίες. Με την ανίχνευση των διευθύνσεων πορτοφολιού των επιτιθέμενων, αναδύεται ένα ολοκληρωμένο χρονοδιάγραμμα του εγκλήματος, από την έναρξη έως τη διασπορά των παράνομα αποκτηθέντων κεφαλαίων. Αυτή η μεθοδολογία διευκολύνει την ποσοτικοποίηση της κλίμακας του εγκλήματος σε κρυπτονομίσματα, όπως το Bitcoin, με δυνατότητα μετατροπής σε συμβατικά νομίσματα. Αξιοσημείωτο είναι ότι επιτρέπει την ολοκληρωμένη παρακολούθηση όλων των χρημάτων που συσσωρεύονται από τους επιτιθέμενους, ανεξάρτητα από τις επίσημες αναφορές.

Με κίνητρο αυτές τις προκλήσεις και τις ευκαιρίες, αναπτύξαμε ένα σύστημα που συλλέγει, επεξεργάζεται και οπτικοποιεί τα δεδομένα που συλλέξαμε κατά την διάρκεια της έρευνας μας. Η προσέγγιση αυτή ενισχύει την κατανόηση και την αξιολόγηση του αντίκτυπου των εγκλημάτων στον κυβερνοχώρο, ιδίως στο πεδίο του Bitcoin.

Acknowledgements

I would like to express my sincere appreciation and gratitude to all those who have contributed to the completion of this master thesis and the research project. First and foremost, I would like to extend my deepest gratitude to my supervisor Prof. Evangelos Markatos for his guidance, expertise and support throughout both my Bachelor's and Master's studies. Furthermore, I would like to express my gratitude to my friends and colleagues Dionisis Kalochristianakis, Manos Papadogiannakis, Thomas Marchioro, Alexandros Karagiannis, and Kostantinos Spiridakis. I am deeply grateful to them and the rest of the DiSCS Lab of FORTH, who enriched my development experience by discussing and sharing their insights to the breadth of the project. I am deeply grateful to my family especially to my grandfather Nikos Skarlis and my mother Evangelia Skarli for their unwavering love, encouragement, and support. Being there for any ups and downs during this work, and most important, giving me the ability to attend to University of Crete and complete my Master's Degree. Lastly, I want to thank my friends for supporting me in every step.

στους γονείς μου

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Ransomware attacks, internet scams, email extortions, and various other forms of cybercrimes occur daily across the globe. These cybercrimes occur in diverse forms and currencies, making it challenging to accurately estimate the extent of the damage inflicted [10] [14] Reporting such incidents provides insight only into the specific damages suffered by the victims who come forward, leaving a vast number of victim attacks and stolen money undocumented. Calculating the overall damage under these circumstances is challenging and important.

In the realm of cybersecurity, numerous Computer Security Incident Response Teams (CSIRTs) and Blockchain Threat Intelligence Platforms exist, yet they often operate in isolation from one another. This fragmentation leads to a plethora of smaller, fragmented data sources, obscuring the true extent of potential threats and hindering a comprehensive understanding of their impact. Furthermore, prominent platforms, such as Chainalysis, withhold their data from public access, preventing both utilization and evaluation. As a consequence, the lack of integration and limited accessibility to data from major players contribute to challenges in accurately assessing the scope and severity of security incidents in the field.

Interestingly, a significant percentage of these cybercrimes are happening within the realm of blockchains. Ironically, the attempts of criminals to conceal their identities can actually reveal crucial information about them. By obtaining the attackers' wallet addresses, we gain the ability to track the timeline of the crime, from its occurrence to the moment the attacker disperses the ill-gotten funds to collaborators. Moreover, we can quantify the scale of the crime in cryptocurrencies such as Bitcoin, which can be easily converted to conventional currency like the Euro or Dollar. Most notably, this approach allows us to comprehensively trace all the funds ever amassed by the attacker, whether officially reported or undisclosed. This led us to implement a system, that collects processes and visualizes public datasets.

1.1 Abuse Classification

An abuse classification refers to the systematic categorization of various malicious activities based on their characteristics and methods. In this context, the classification encompasses distinct types of cyber abuses, such as ransomware, blackmail scams, Bitcoin tumbling, darknet markets, and sanctions. Each category represents a specific form of malicious behavior, aiding in the understanding and analysis of diverse threats for the purpose of cybersecurity and investigative efforts.

1.2 CSIRT

A Computer Security Incident Response Team (CSIRT) is a group of cybersecurity professionals responsible for responding to and managing computer security incidents. CSIRTs play a crucial role in identifying, analyzing, and mitigating security threats and incidents within an organization or a community. They may also provide guidance on improving cybersecurity practices to prevent future incidents.

1.3 BlockChain Threat Intelligence Platform

A Blockchain Threat Intelligence Platform is a system or service that focuses on collecting, analyzing, and disseminating threat intelligence specific to blockchain environments. This platform is designed to identify and report on security incidents, vulnerabilities, and emerging threats within the blockchain space. It may aggregate information from various sources, including crowd-sourced reports, to enhance the overall security of blockchain networks and ecosystems.

1.4 Threat Intelligence

In this study, data was exclusively sourced from publicly available platforms, encompassing information such as crime-related wallet details, crime types (e.g., ransomware, blackmail), and additional details like the incident date, attack type, and comments. Evaluating the reliability of each source required researching its nature, publication consistency, and data usability. These sources were curated and disseminated by specific groups. The selection process for trustworthy sources is initiated with the scrutiny of their publicly accessible feeds. After monitoring the overall reputation of blacklisted Bitcoin wallets for several weeks, a decision was made on whether to incorporate or exclude the source based on its credibility.

Given the limited extent of previous work in this domain, evaluating these sources and conducting tests demanded specific analyses for each, rendering this process one of the most challenging aspects of our work.

As a component of this study, we present a dataset comprising accessible sources, with a preview provided in Table 1.1. For each assessed data source,

we compile, among other details, its name, the number of wallet transactions, and both the value in Bitcoin and Euro.

1.5 Software Development

Initially, our research primarily focused on Bitcoin; however, we expanded our investigation to include other blockchains such as Ethereum and Monero. Unfortunately, the results gathered from these alternative blockchains were either significantly smaller (Ethereum) than those observed in Bitcoin or we were unable to to do any research due to the functionality of those blockchains (Monero).

To gather a large dataset, our approach involves the systematic collection of malicious addresses (wallets), Table 1.1 illustrates our datasets. Subsequently, we extract all transactions associated with each of these identified wallets, enabling us to calculate the balance in the currency of our preference and present relevant statistics. The process of assembling an extensive dataset of malicious wallets involves aggregating data from reputable open-source sites, datasets obtained from scholarly papers and researchers, and other credible sources.

For the extraction of transaction data related to the wallets of interest, we leverage the blockchain.com API [3]. This API facilitates the retrieval of detailed transaction information crucial for our analysis. Finally, to convert Bitcoin values to Euro equivalents on the respective transaction dates, we utilize another API. This approach ensures the completeness and accuracy of our dataset, allowing for a robust exploration of malicious activities across the bitcoin blockchain.

Source	Contribute Wallets
Ofac	374
Alienvault	2.167
Ransomwhere	10.444
ChainAbuse	2.151
Ransomlook	10.275
CryptoScamDB	4.478
SophosLab	3.488
Irvine	20.849
Tessii	20.849
Kaggle	2.549
Behas	7.223
EtherScamDb	2.357
Boulevard	1.072
Cryptoexchangescam	182
KillingTheBear	61
Traceer	120

Table 1.1: List of datasets aggregated by our platform.

1.5.1 Challenges

Managing a vast array of datasets presented several challenges that required tailored solutions. Among these challenges, scaling issues emerged as a prominent obstacle, necessitating the development of efficient methods to handle large volumes of data. Additionally, distinguishing reliable sources from less credible ones posed another hurdle in ensuring the accuracy and validity of our datasets. Storage of extensive data sets also demanded a robust solution to optimize efficiency.

While addressing these primary concerns, we encountered secondary challenges that, while less critical, demanded attention. Crafting a website encompassed the creation of both backend and frontend components, aiming to site balance between aesthetic appeal and providing insightful information on the subject matter. Simultaneously, we faced the task of developing numerous scripts to generate statistics for the website. Ensuring the seamless functionality of the website, became crucial due to the sheer size of the data involved.

In overcoming these challenges, our efforts were directed toward the optimization of data management, source reliability, and the seamless operation of the website.

Chapter 2

Threat Intelligence Sources

2.1 Overview

Our primary data sources include Threat Intelligence Platforms (e.g., CryptoScamDB, Alienvault [1] ChainAbuse), government agencies (OFAC [4]), and researchers (e.g., Ransomwhere [6], Behas). These sources present information in diverse formats, including malicious wallet data, abuse classification, timestamps, comments, ransomware family details [15], and more. The data collection process centers around identifying reputable organizations or teams that maintain accurate datasets. Source evaluation involved constructing a multidimensional array containing information about each source.

However, the challenge arises from the disparate formats and publication mechanisms used by each source. Some employ APIs, others use datasets, and some rely solely on websites. Despite these variations, all blacklisting sources collected for this study are available at https://crypto-abuses.ics.forth.gr/. Our published dataset includes vital details such as distinct names, short descriptions, and additional information for each wallet, encompassing received amounts in Bitcoin and Euros, along with the sources that contain the wallet and its abuse classification.

For a more comprehensive overview, our website offers additional information, including general statistics such as annually stolen funds, funds per source, wallet transactions, and more.

2.2 Datasets

To identify suitable platforms and datasets for our research, we employed diverse techniques, including SimilarWeb, academic papers, Google searches, GitHub repositories, and custom scripts, enabling the extensive collection of top results. Then we conducted a thorough analysis of these sources to assess their legitimacy. While we identified numerous sources beyond those mentioned here, a combination of factors precluded their use. Many either did not publish their data altogether, demanded a substantial fee, or lacked trustworthiness due to an abundance of false

crowd-sourcing, making all these sources challenging to verify.

Ultimately, our analysis drew from a curated selection of 16 sources. These sources, including APIs, datasets, and platforms, underwent meticulous scrutiny to confirm their legitimacy and public availability. The processes of downloading, crawling, and parsing public data through various APIs can present varying levels of difficulty because of several factors.

- **Documentation:** The availability and quality of API documentation significantly influence the ease of parsing public data.
- **API Design:** The structure of the API itself plays a crucial role in the parsing process. APIs with clean and consistent data structures, standardized formats (such as JSON or XML), and intuitive naming conventions are generally easier to parse.
- **Authentication and Access:** While some APIs are free to use, certain ones require authentication mechanisms like API keys, tokens, or OAuth for data access.
- **Data Complexity:** The inherent complexity of the data can impact parsing difficulty.
- **Data Volume and Pagination:** Extensive and paginated public data introduces complexity. Handling pagination involves managing links, orchestrating multiple API requests, and merging or appending data from different pages.
- **Error Handling:** Robust parsing necessitates addressing various error scenarios, such as rate-limiting errors, authentication failures, or malformed data responses, to ensure reliability and fault tolerance.
- **Limited Access:** The rate at which websites can serve data affects parsing feasibility.
- **Downtime:** The availability of websites throughout the day influences parsing reliability.
- **Continual Changes:** Data may transition between labels, for example, from malicious to benign, over time.

Considering these factors, parsing public data becomes a considerable challenge. To address the diverse parsing requirements, we developed an adaptive parsing model.

2.3 Data Source Structure

Throughout this study, we encountered various types of public datasets, each demanding distinct extraction mechanisms and techniques. Our approach commenced with the simplest yet widely used dataset type—raw files. Many reporters opt for simplicity by publishing their data in files accessible to the public through their original websites. In our research, we diligently collected and processed datasets in CSV, Text, JSON, and XML formats.

Subsequently, we leveraged API services provided by select organizations through specific endpoints. Notable examples include AlienVault and https://cryptoscamdb.org/. These services expose data through designated endpoints, necessitating adherence to guidelines and policies outlined in their respective documentation. This often involves compliance with token authentication, maximum downloads per period, and rate limits.

The final category of dataset parsing involves web crawling. Extracting data through this method requires the creation of scripts tailored to each source, in addition to ongoing maintenance feature needs.

2.4 Abuse Categories

We extract the informations provided of each source and we create some distinct categories to classify each cyber crime those are:

- **Ransomware:** is malicious software designed to block access to a computer system or files until a ransom, is paid.
- **Blackmail Scam:** is scheme where perpetrators threaten to disclose sensitive information or images unless a victim pays money or provides certain services.
- **Bitcoin Tumbler:** is a service that enhances the privacy of Bitcoin transactions by mixing and obfuscating the source of funds, making them more challenging to trace.
- **Darknet Market:** An online marketplace operating on the dark web, often facilitating the trade of illegal goods and services using cryptocurrency.
- **Sanctions:** contain wallets that government agencies have concluded that this wallets are dangerous and you should not interact with those.

Chapter 3

System Overview & Architecture

Our objective is to develop a user-friendly tool that effectively illustrates the impact of cybercrime on the blockchain. We aim to present our research insights in an easily understandable manner, incorporating diagrams depicting the progression of stolen funds over time and illustrating overlaps in datasets.

The primary objective is to develop a user-friendly tool that effectively communicates the impact of cybercrime on the blockchain. The tool emphasizes collecting and visualizing illicit gains from various cybercrime types such as ransomware, scams, extortion, tumblers, and more. This involves extracting information from multiple websites, APIs, and datasets, retrieving transaction details for every wallet identified in the data sources, and analyzing transactions to calculate the amount of money stolen in both Bitcoin and euros for each wallet. The data is then grouped according to analysis criteria. The statistical analysis includes calculating the overall stolen money, the overall money in each crime classification, and determining the stolen money annually, both in aggregate and by classification. Further, we break down the contribution of each data source in stolen money for each category. The final step involves serving and displaying the generated statistics on our website, ensuring accessibility and user-friendliness. This structured approach aims to highlight the overall impact of cybercrime on the blockchain while providing a nuanced understanding of the specific dynamics within different crime classifications.

3.1 Extracting Data

To enrich our tool with data, we utilize publicly available APIs, websites, and datasets. These datasets include sources from inactive GitHub repositories and data from older research papers, as well as actively maintained APIs, websites, and datasets. To systematically extract and update a large quantity of data, we've implemented a parsing system.

3.1.1 Exchange Rates

To process transactions for each wallet and determine the euro value, we require the corresponding Bitcoin-to-euro exchange rate for the relevant time period. Following research, we have opted for a sampling approach, conducting calculations daily as shown in the figure below. This means that we will determine the Bitcoin value in euros based on the exchange rate at the beginning of each month in which a transaction occurs. This strategy allows us to capture the fluctuations in exchange rates over time, ensuring a more accurate representation of the monetary value associated with each transaction.

```
{
    "2024-01-06": 40511.21290946505,
    "2024-01-07": 40744.78193701536,
    "2024-01-08": 43239.39491129363,
    "2024-01-09": 43875.97621921745,
    "2024-01-10": 43653.84340879144,
    "2024-01-11": 44851.901239447885,
    "2024-01-12": 42555.36425127867,
    "2024-01-13": 39634.70210445877,
    "2024-01-14": 39403.451144290746
}
```

Listing 3.1: Bitcoin exchage rates.

3.1.2 Inactive Datasets

As previously mentioned, a portion of our data originates from inactive sources, including datasets associated with research papers and GitHub repositories from research centers or labs. One such dataset is the "Bitcoin Heist Ransomware Address Dataset" from UC Irvine [2]. Another notable example is the GitHub repository maintained by Sophos Labs. This repository contains datasets sourced from an open threat intelligence platform affiliated with Sophos Lab.

3.1.3 Parsing System

For the active sources, we created a parsing system to handle the data extraction. The parsing system is designed in both Python and JavaScript to manage various scripts in these languages. A cronjob triggers the execution of the two parsing systems (Python and JavaScript), which, in turn, invoke the controller for each data source. When the source provides an API, the controller requests the latest data that has not been previously extracted. If the source offers data in a downloadable format (e.g., CSV, JSON, XML), the script downloads and appends the new data to the existing dataset. In cases where a source lacks an API or downloadable dataset, web parsers are employed to extract the necessary information.

This systematic approach ensures the consistent and periodic enrichment of our tool's dataset.

CryptoScamDb An example of source that provides a free API through their website. Therefore, a Python script was developed to make requests once a week and update the existing data. Another source that we used the apis was uses an source is CryptoScamDb or Alienvault. CryptoScamDb.

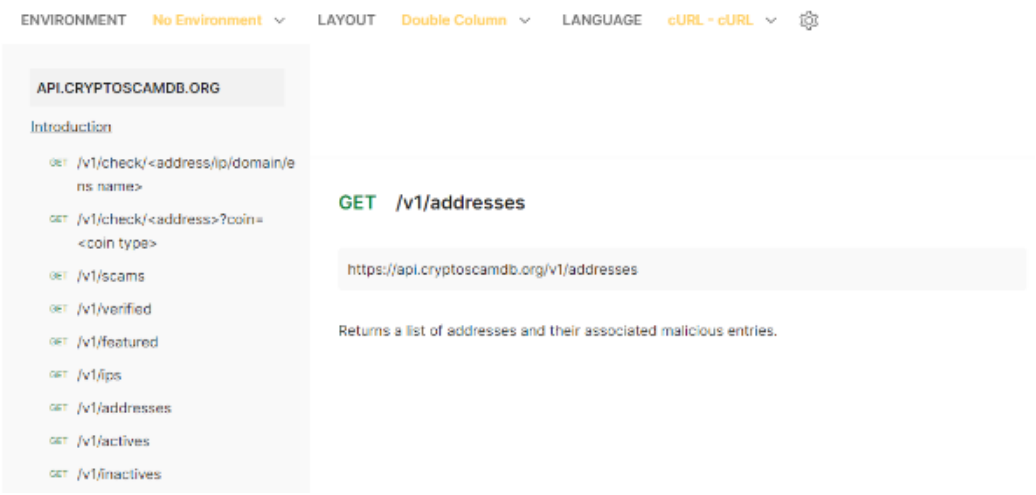

Figure 3.1: CRYPTOSCAMDB API.

```
def CrawlCryptoScamDB():

    response = requests.get("https://api.cryptoscamdb.org/...")

    dirname = os.path.dirname(__file__)
    csdb = os.path.join(dirname, "__path__/cryptoscamdb.json")

    f = open(csdb, "w" , encoding="utf-8")
    f.write(response.text)
    f.close()
```

Listing 3.2: Cralwer that gathers CryptoScamDB data.

ChainAbuse although free, lacks an official API, necessitating web parsing. To address this limitation, a JavaScript script utilizing Puppeteer was created. This script extracts essential information from each report page on the website, including the wallets implicated in cybercrime, the crime classification, and the report status (trusted contributor, checked, or unverified). Given the potential for multiple reports from various users, a dictionary was established. Wallet IDs serve as unique keys, and the corresponding values consist of arrays containing

the report status and crime type. This approach accommodates situations where multiple users report various crimes or crime types associated with the same wallet. Other sources that do not offer any api were Ransomlook and Security Boulevard.

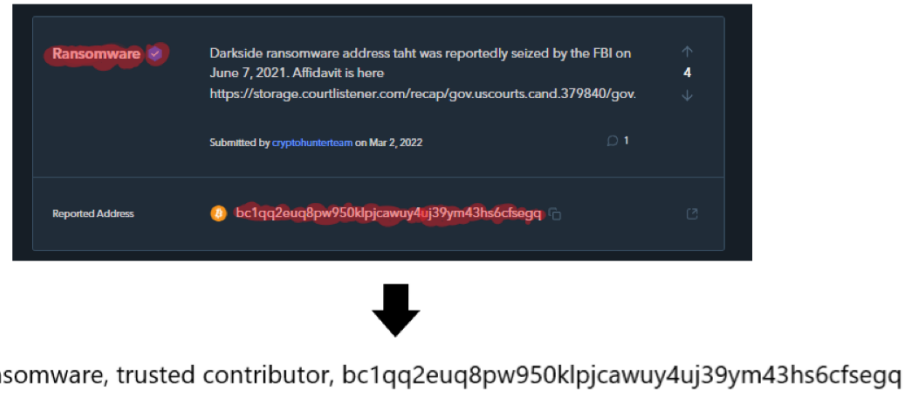

Figure 3.2: ChainAbuse extracted data

Ransomwhe.re [6] provides both an API and a downloadable JSON file containing a comprehensive dataset of their findings. For the sake of simplicity and stability, we opted to download the entire dataset in JSON format. Consequently, a script was developed to request the file on a weekly basis, ensuring the most up-to-date information. This approach is also applied to other sources such as Kaggle, where a similar script is employed for regular data retrieval.

3.2 Blockchain Transactional data

To facilitate comprehensive wallet analysis on a weekly basis, we require the complete transaction history of the collected wallets. To achieve this efficiently without the need to download the entire blockchain node and to maintain adaptability to any changes, we have opted to leverage an API. Our choice for this purpose is the blockchain.com API. This decision was influenced by the request limit and the additional information provided by the API. While alternatives like Blockchair offer an even higher request limit and comparable stability, they lack a timestamp for each transaction. This absence posed a limitation, as it would hinder our ability to translate the Bitcoin value to the Euro value during specific periods, thereby compromising the precision of our results.

3.2.1 Transaction System

Limitations and Issues During the development of this transaction storage system, we had some major limitations and issues. We encountered significant limitations and challenges. The primary issue was efficiently managing all the wallet

information. To ensure swift address and transaction lookups during statistical analysis without any performance degradation, we aimed for constant time complexity, denoted as O(1). Given the size of our data, we opted to store each wallet in a JSON file. For instance, the wallet `"13AM4VW2dhxYgXeQepoHkHSQuy6NgaEb94"` would be saved in a file named `"13AM4VW2dhxYgXeQepoHkHSQuy6NgaEb94.json."`

After resolving the initial problem, a subsequent challenge arose when dealing with a substantial number of JSON files within the same directory, leading to performance issues during file lookups. To address this concern, a strategy was implemented where the first three characters of each wallet were utilized to distribute the data across multiple folders. This organizational approach successfully alleviated the performance impact on lookups. As an illustration, consider the wallet `"13AM4VW2dhxYgXeQepoHkHSQuy6NgaEb94"`, which was positioned at `"data/transactions/bitcoin/12A/13AM4VW2dhxYgXeQepoHkHSQuy6NgaEb94.json"`.

Having decided on the data structure the next issue was the API and provides data through pagination, which is a common approach for APIs. The blockchain.com API, for instance, has a page limit of 100 transactions per call. Assuming a wallet has 550 transactions, we would need to make 6 requests to the API. In each call, it's necessary to redefine the offset to ensure we retrieve the correct transactions. To address this limitation, a simple Python script was created. This script checks if we have already stored the wallet and then recursively requests the remaining transactions, starting from the offset of the last transaction we have.

Figure 3.3: Endpoint from blockcain.com API that serves transaction data

```
def getBitcoinTransactionsRecursively(address):

    limit = 100
    offset = 0
    prev_current_tx = -1
    current_tx = getExistingTransactions(address)

    if(current_tx == -1):
        current_tx = 0;

    total_tx = current_tx+100
    while(True):
        if(total_tx <=  current_tx ) or total_tx > 5000 :
            break

        if(current_tx == prev_current_tx):

            break


        if(  total_tx  - current_tx <= 100):
            limit =  total_tx  - current_tx

        url = f'https://blockchain.info/multiaddr?active={address}&limit=
        {limit}&offset={offset}'

        responseObject =  downloadTransactions(url)

        if(responseObject["status"] == True):
            data  = responseObject["data"]

            if("addresses" in data):
                if(len(data["addresses"])==1):
                    summary_data = data["addresses"][0]
                    total_tx = summary_data["n_tx"]
                    updateSummary(address , summary_data)

            if("txs" in data):

                updateTransactions(address, data)


            prev_current_tx = current_tx
```

```
        current_tx+= len(data["txs"])
        offset+= len(data["txs"])
    else:
        break
sleep()
```

Listing 3.3: Bitcoin transactions system.

As shown in the code above, a script is executed to create or update a wallet with any missing transactions. This script is invoked for each wallet individually. To automate this process, a cronjob was set up to run weekly. This cronjob triggers the controller responsible for calling the script, ensuring regular updates for all wallets.

Furthermore, the complexity of the program led to significant problems such as missing/duplicate transactions and inaccurate wallet summaries. To tackle the issue of false summaries, we opted to discontinue their use. Although this decision slowed down our data updating process by a factor of 10, it significantly enhanced the reliability of our data over time.

To address missing transactions, a script was developed to iterate through all the wallets and identify those with missing transactions. Regarding duplicate transactions, a mechanism was implemented. The script temporarily stored transactions in a hashmap, using the transaction hash as the key and the transaction itself as the value. By storing all transactions in a hashmap, any duplicate transactions were automatically removed. Subsequently, using the hashmap, we updated the information for each wallet, resolving the issue of duplicate transactions.

3.3 Data Processing

With access to both our transaction data exchange rates and the source data, we can efficiently process the information to generate statistics for our dashboard. In our work, we chose to store the data of our sources in the most convenient way for each source. So before moving with processing, we create a parse for each source. Each parser reads the source data, parses them, and returns a dictionary with keys the abuse classifications and as value a set of wallet that this source has categorize with this abuse classification. To calculate these statistics in the most effective manner, we begin by consolidating all our sources into a unified set. Subsequently, for each Bitcoin wallet, we compute various wallet statistics, including total Bitcoin/Euro received, annual Bitcoin/Euro received, the number of transactions, and more. All this data is organized and stored in a comprehensive dictionary, utilizing the wallet ID as the key.

Before advancing to the next stage, we have the option to apply filtering, such as excluding wallets with zero transactions from our statistics. Following this preprocessing step, we can proceed to calculate more generalized statistics, such as overall crime rates or crime rates per classification or source. This is achieved by simply accessing the dictionary that contains the preprocessed data.

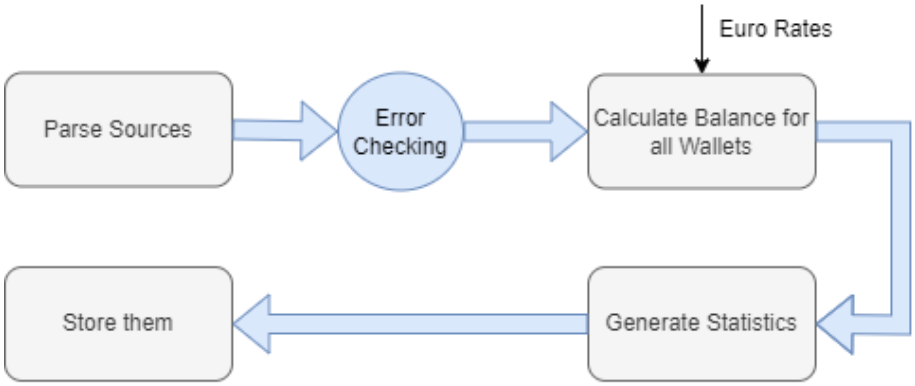

Figure 3.4: Open threat intelligence Parsing System

3.4 Dashboard

Given the dynamic nature of our tool, a backend is essential to serve the data, while a frontend is necessary to display this information to the user.

3.4.1 Backend

The backend played a crucial role in reading all the data generated during the data processing phase and subsequently serving it. Developed using Node.js, the backend comprises of four routes housing a total of 12 endpoints. Those endpoints can be broken down into four API routes address, rates, sources, and statistics. The route address is responsible for returning all the data related to one address such as transaction number bitcoin stolen and the sources that report this wallet as well the category. The rates route return data related to exchange rates from Bitcoin or ethereum to euro to dollars for a specific date or time period. The source statistics is responsible for giving data such as the sources and information for the sources external links for those and more. Finally, the route statistics provide data and labels to fill all of our plot diagrams and graphs in the website.

```
routes
├── /address
│   └── /info
├── /sources
│   ├── /getSources
│   └── /:addr
├── /rates
│   ├── /bitcoin
│   └── /bitcoin/monthly
└── /statistics
    ├── /
    ├── /statistics-site
    ├── /abuses/:abuseType
    ├── /bitcoin/getStatistics
    ├── /bitcoin/getAllStatistics
    ├── /bitcoin/getLabels
    └── /bitcoin/getSummary
```

3.4.2 FrontEnd

The frontend was crafted using React.js. When calling an endpoint of our backend, the frontend retrieves data/statistics. The front end is composed of four key pages: the homepage, individual pages for each abuse classification, pages dedicated to each wallet, and a summary page aggregating information from our sources.

Homepage

The homepage provides a comprehensive overview, ensuring a broad understanding of our findings without delving into specific details. The key components include:

Summary of Findings: This section offers a quick snapshot of our research, presenting the total number of wallets and transactions, along with the overall stolen bitcoin and its equivalent value in euros. This summary provides a high-level understanding of the scale of cybercriminal activities.

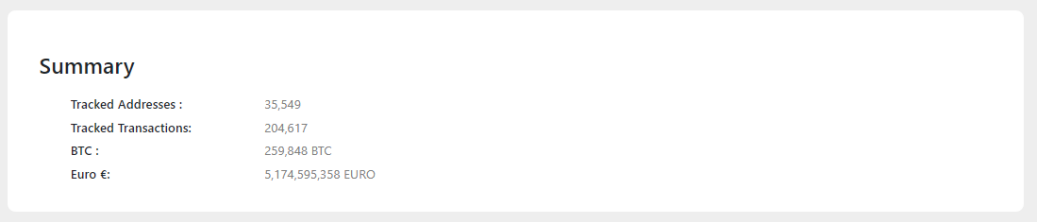

Figure 3.5: Overall revenue of cyber criminals in Billions of euros

Annual Summary: Displaying cybercriminal earnings annually from 2017 to the present, this component provides an overarching perspective on financial trends

over time. The annual summary contributes to a longitudinal understanding of the evolving landscape of cybercrime.

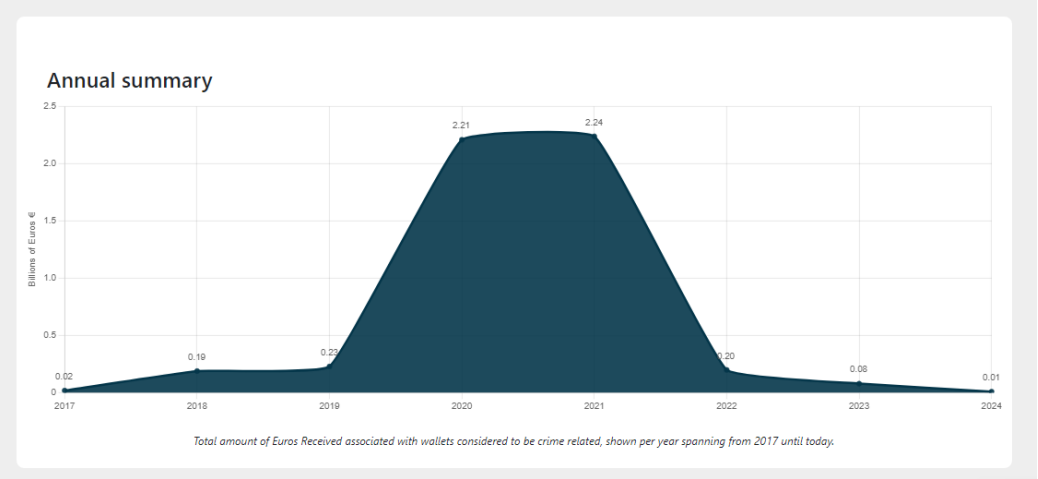

Figure 3.6: Anual revenue of cyber criminals in Billions of euros

Stolen Funds per Crime Classification: This section further breaks down cybercriminal earnings, offering both an overall total 4.2 and an annual breakdown per crime category depicted in figure 3.7. This nuanced approach allows users to grasp not only the overall impact but also the specific contributions of different cybercrime classifications.

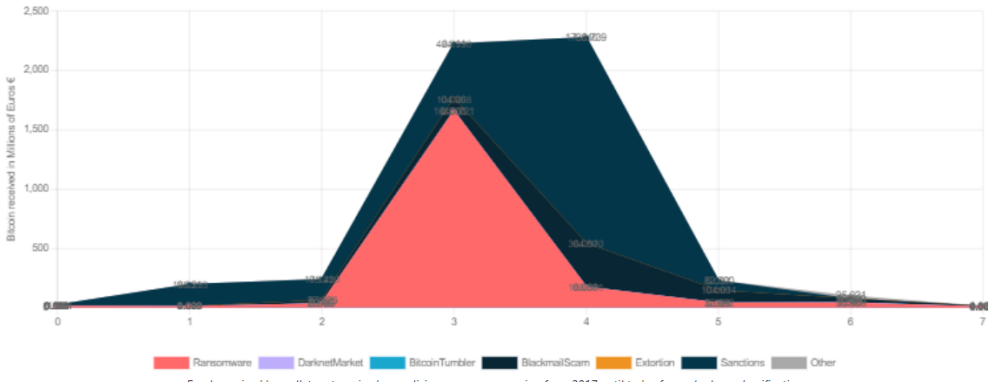

Figure 3.7: Anual revenue of cyber criminals in Billions of euros for each crime category

Bitcoin Price Evolution: Figure 3.8 illustrates the fluctuation in Bitcoin prices over the years, providing additional context to understand the economic backdrop in which cybercriminal activities occurred.

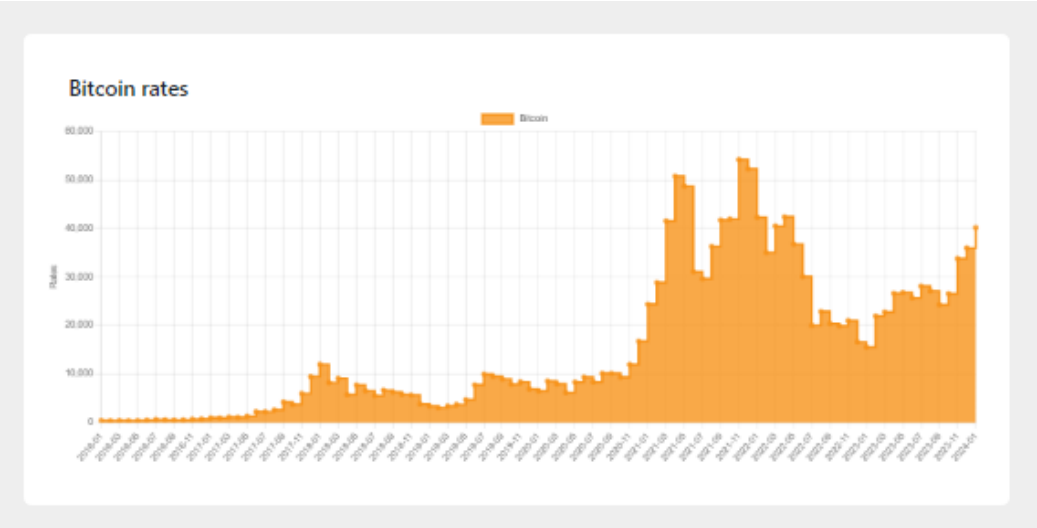

Figure 3.8: Exchange rates of Bitcoin to Euro

Abuse Classification Page

On this page we focus on abuse classifications, we present valuable insights into each crime category through three main components: Annual Summary, Overall Summary, and Intersections.

Annual Summary: In this component, a chart illustrates the earnings of the crime category for each year, providing a clear visual representation of the financial trends over time.

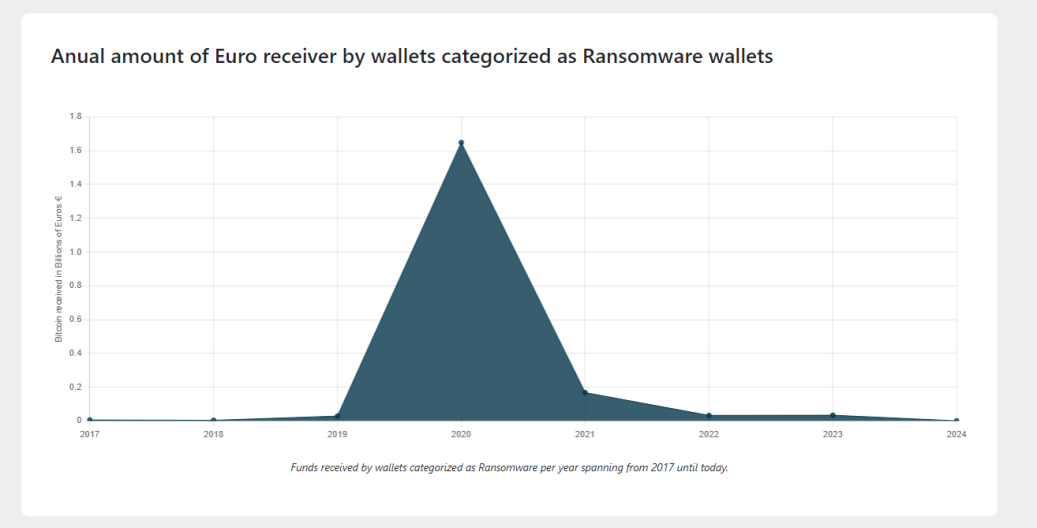

Figure 3.9: Stolen money by Ransomware yearly

Overall Summary: The second component, represented by Figure 3.13 offers a

comprehensive overview of the contribution of each source to the specified crime category. It includes essential details such as the source name, reported wallets for the cybercrime, and the associated stolen bitcoins. Additionally, a pie chart visually depicts the proportional contribution of each source, facilitating a quick comparison with the other sources.

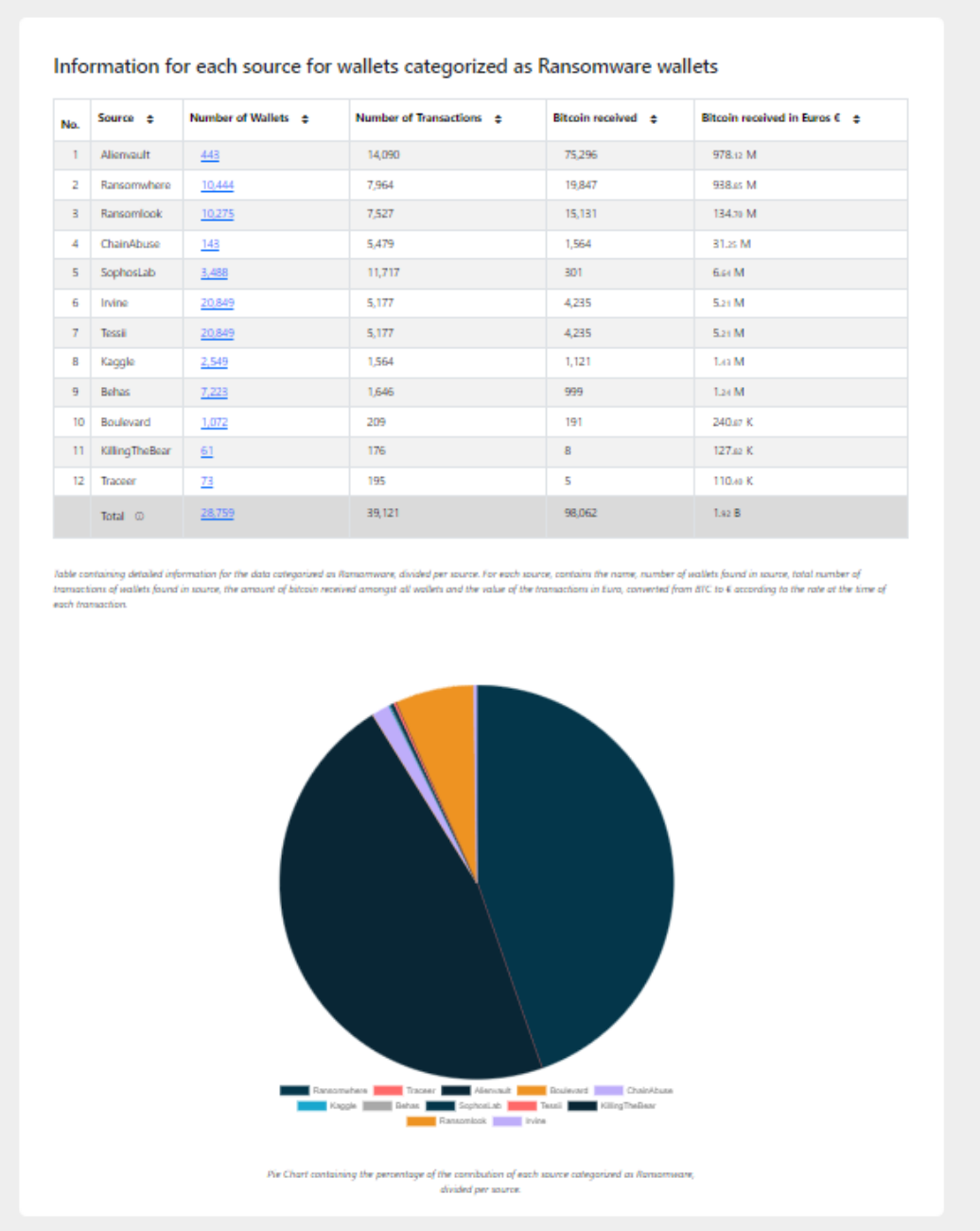

Figure 3.10: Exchange rates of Bitcoin to Euro

Intersections: The final component, shown in Figure 3.13, presents a heatmap displaying the intersections of each source with others. Calculating the Jaccard similarity for the sources on both the x and y axes provides a unique understanding of collaborations and perspectives among sources. This heatmap enhances our comprehension of the relationships and shared data points between different sources, contributing to a more nuanced analysis of the crime category.

Figure 3.11: Source correlation based on Jaccard similarity

This structured presentation ensures that users can gain in-depth insights into the annual financial trends, overall source contributions, and collaborative intersections within the specified crime category.

Wallet Page

On this page, we concentrate solely on crucial details about individual wallets. While our initial emphasis was on the aggregate amount of stolen funds, this section delves into specific wallets. We believe this focus is pivotal for revealing the extent of funds pilfered by each individual. Moreover, by highlighting the sources that include a particular wallet in their dataset, we aim to demonstrate the reliability of our tool. Specifically, our initial component includes an external link directing users to view transactions in a Bitcoin Explorer. Alongside this link, we

present key information such as the number of transactions, the amount in Bitcoin, and its equivalent value in Euros. In the next component, we outline all the sources that reference this particular wallet, along with the classification assigned by each source.

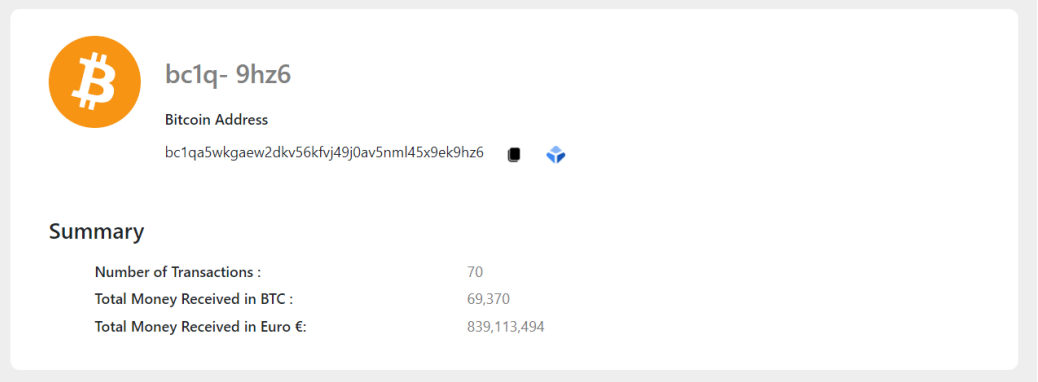

Figure 3.12: Information provided by our app for a specific wallet

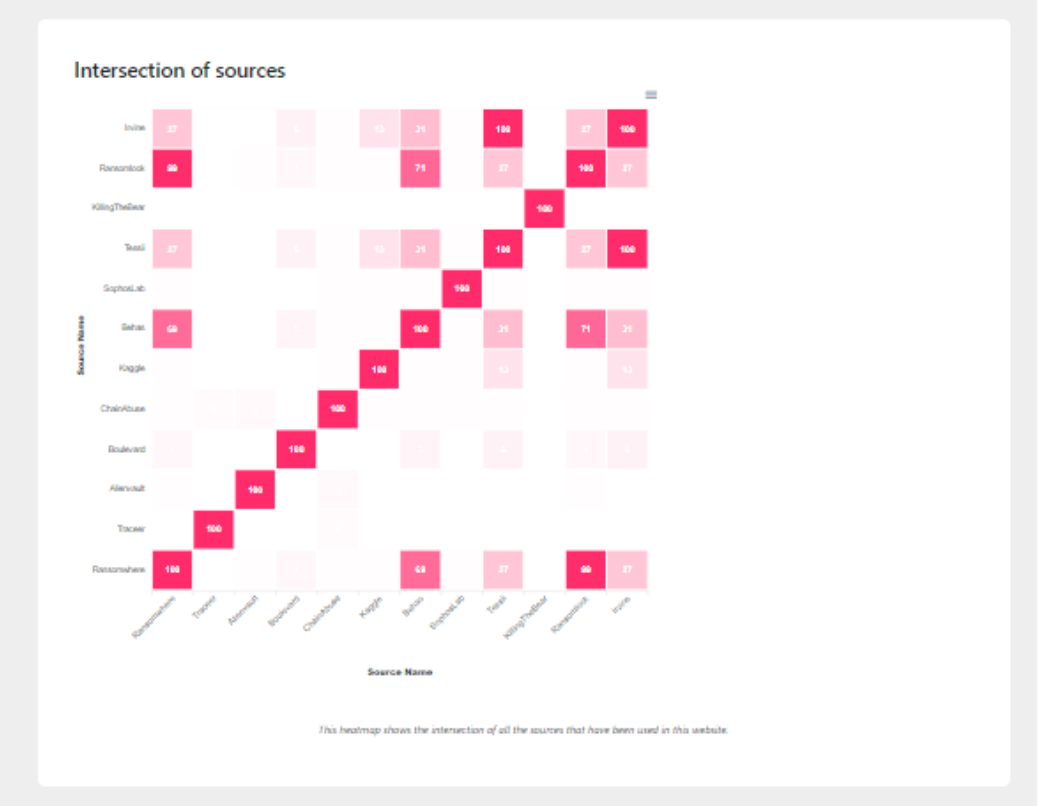

Figure 3.13: Source correlation based on jaccard similarity

Sources Page

Here, we present a list of the sources incorporated into our tool. Each source is accompanied by its name, a corresponding link for reference, and a brief description. This transparent approach reflects our confidence in the data underpinning our tool. Additionally, a heatmap is included on this page to visually represent the correlations among the sources. This chosen visual aid illustrates potential interactions between our sources or instances where certain sources diverge in data collection.

Sources

No.	Website	Source Type	Info
1	Ransomwhere	Open Threat Intelligence Platform	Ransomwhere is the open, crowdsourced ransomware payment tracker. Browse and download ransomware payment data or help build our dataset by reporting ransomware demands you have received.
2	Traceer	Open Threat Intelligence Platform	Traceer AML / KYT Compliance is a powerful Bitcoin verification tool for business. In-depth real-time bitcoin verification of compliance with anti-money laundering legislation. A flexible approach that meets the needs of crypto companies, financial institutions and investigators.
3	Alienvault	Open Threat Intelligence Platform	The World's First Truly Open Threat Intelligence Community
4	Cryptoscamdb	Open Threat Intelligence Platform	CryptoScamDB's is open-source dataset tracks malicious URLs and their associated addresses to make this entire ecosystem safer for you. It is designed to keep track of malicious URLs and their associated addresses that have the intent of deceiving people for financial gains.
5	Security boulevard	Open Threat Intelligence Platform	Security Boulevard is a division of Techstrong Group, Inc., producers of leading technology communities like DevOps.com, Container Journal and Techstrong TV. Security Boulevard's mission is to serve the security and related communities by providing a single destination for information, education and discourse on the leading topics and issues facing the security, as well as the larger IT community today.
6	Chainabuse	Open Threat Intelligence Platform	Chainabuse is an industry-led initiative, backed by leading crypto businesses, protocols, and foundations with an interest in making crypto safe and trusted for the next billion users
7	Kaggle	Paper Dataset	This is dataset is from the paper: Akcora, C.G., Li, Y., Gel, Y.R. and Kantarcioglu, M., 2019. BitcoinHeist: Topological Data Analysis for Ransomware Detection on the Bitcoin Blockchain. IJCAI-PRICAI 2020.
8	Behas	Researcher	Bernhard Haslhofer (Behas) is faculty member at the Complexity Science Hub Vienna, where I lead the Cryptofinance research group. Additionally, I am also one of the co-founders of iknaio Cryptoasset Analytics GmbH and a court-certified expert in IT and cryptoasset forensics.
9	Etherscamdb	Crowd Source Dataset	An open-source database to keep track of all the current ethereum scams
10	Sophoslab	Open Threat Intelligence Platform	Powered by threat intelligence, AI and machine learning from SophosLabs and SophosAI, Sophos delivers a broad portfolio of advanced products and services to secure users, networks and endpoints against ransomware, malware, exploits, phishing and the wide range of other cyberattacks. Sophos provides a single integrated cloud-based management console, Sophos Central – the centerpiece of an adaptive cybersecurity ecosystem that features a centralized data lake that leverages a rich set of open APIs available to customers, partners, developers, and other cybersecurity vendors. Sophos sells its products and services through reseller partners and managed service providers (MSPs) worldwide.
11	Theresa Github	Paper Dataset	This is a dataset from a the paper: Bitcoin as a payment method for ransomware attacks has gained popularity in the past years.
12	Killing The Bear	Open Threat Intelligence Platform	BloodCoin is repo that allows users to collect crypto wallet addresses (mainly BTC) from ransom notes. These addresses are from real and recent attacks on services like ELK, MongoDB, and other unsecured resources, allowing researchers to be aware of the latest movements and protect themselves from future attacks. BloodCoin also collects email addresses associated with ransom notes, as well as metadata such as countries, continents, services, versions and ports. Events are grouped so track data along days its possible.
13	ransomlook	Open Threat Intelligence Platform	RansomLook is an open-source project aimed at assisting users in tracking ransomware-related posts and activities across various sites, forums, and Telegram channels.
14	OFAC sanctions lists	U.S. Sanction Dataset	Ofac is the Office of Foreign Assets Control administers and enforces economic sanctions programs primarily against countries and groups of individuals, such as terrorists and narcotics traffickers. The sanctions can be either comprehensive or selective, using the blocking of assets and trade restrictions to accomplish foreign policy and national security goals.
15	Cryptocurrency Exchange Scams	Paper Dataset	This dataset is from the paper Characterizing Cryptocurrency Exchange Scams
16	Irvine	Paper Dataset	BitcoinHeistRansomwareAddressDataset. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5BG8V.

Figure 3.14: Sources with a short description

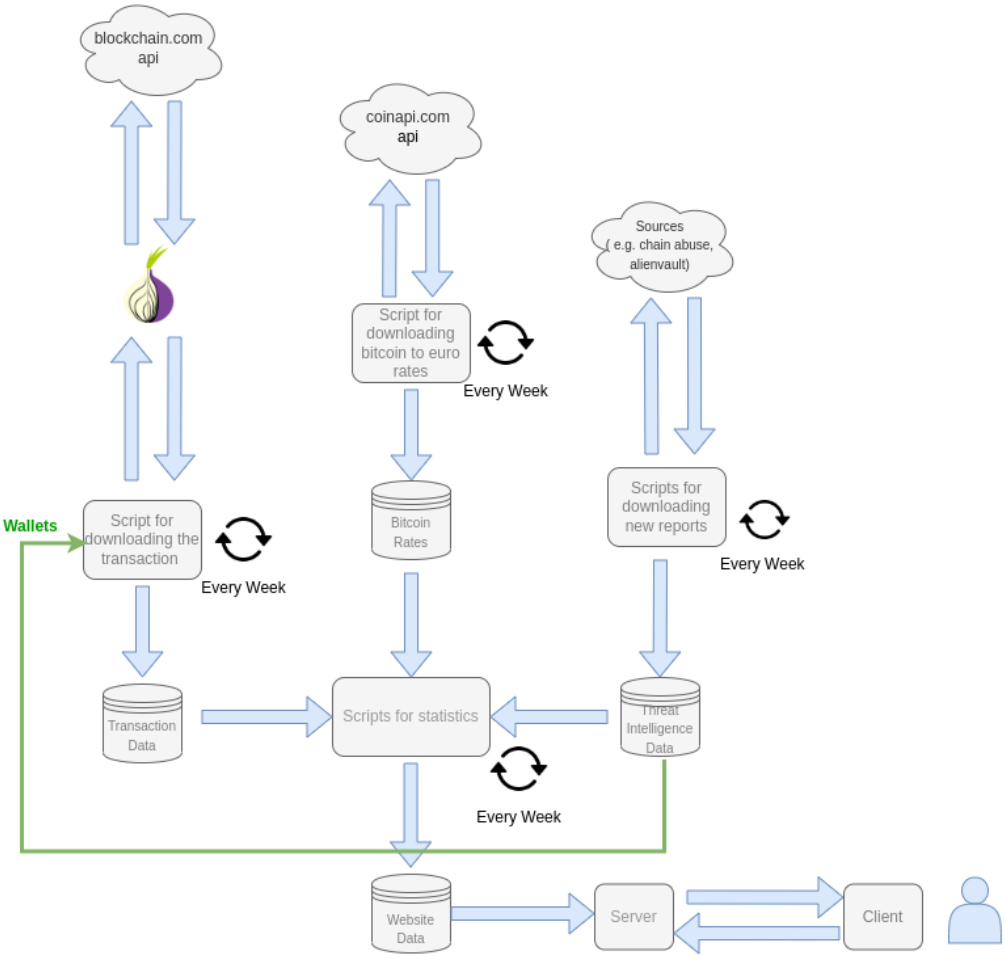

Figure 3.15: System, Overview

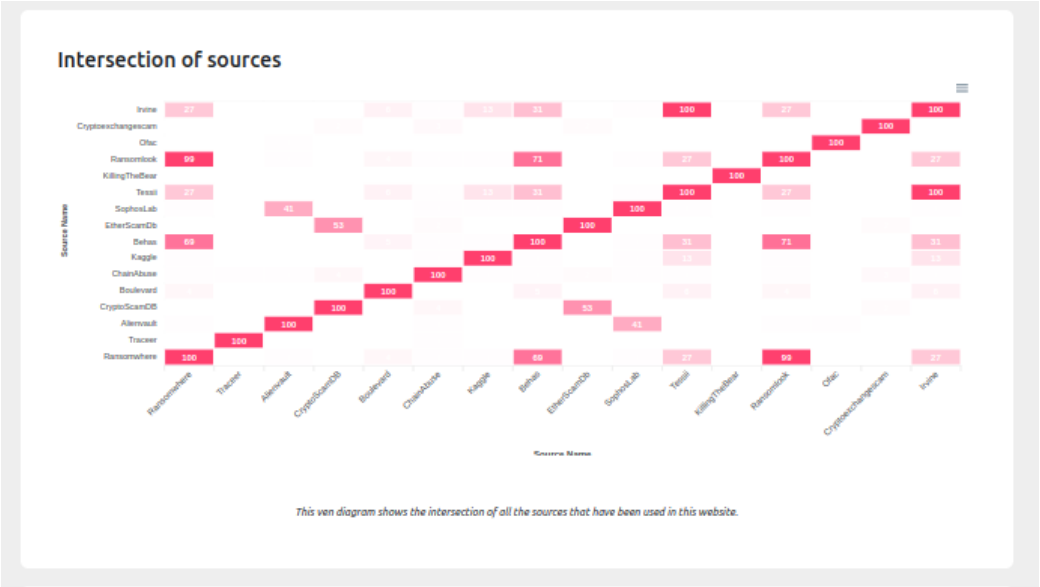

Figure 3.16: Jaccard similarity of sources

Chapter 4

Cybercrime Analysis

4.1 Overview

This chapter presents the outcomes of our research efforts. Within this chapter, we will expose the data collected, the analysis that has been conducted, and the conclusions drawn from our research. We decided to focus our research from 2017 and onward. During this period cryptocurrencies became more prevalent in illicit transactions due to the increase of the price and the maturity of the market. We opted to present the data in euros instead of Bitcoin because malicious actors predominantly demanded payments in established currencies such as euros and dollars. Displaying the amounts in bitcoin would likely result in a decreasing trend, given the rise in bitcoin prices. For instance, if a ransom were set at 500 euros, the equivalent in bitcoin would be approximately twice as much in 2020 compared to 2021 due to the increase in bitcoin's value.

4.2 Findings

4.2.1 Bitcoin Crime Overall

Before diving into a more detailed analysis, it is crucial to provide an overview. Throughout the project, we gathered data on over 35,000 malicious wallets, comprising a total of more than 198,000 transactions. This amounted to over 258,000 bitcoins or more than 5.17 billion of euros. It's important to note that these figures are conservative, as we only retained wallets from verifiable sources. To provide context, we initially collected data on four times the number of wallets and over 40 billion euros. However, these figures couldn't be substantiated due to unreliable sources. The actual numbers likely fall somewhere in between. Detailed and verified research figures are presented in the Table 4.1.

Tracked Addresses	35.549
Tracked Transactions	198.388
Stolen Bitcoin	258.494
Stolen Bitcoin In Euros	5.157.963.234

Table 4.1: Summary of our findings.

4.2.2 Bitcoin Crime over the Years

In this chapter, we use all available data sources for each crime classification to compute the number of euros that have been stolen over the years. Additionally, owing to the traceable nature of Bitcoin transactions, we can ascertain details for each transaction, enabling us to calculate the overall monetary losses at specific time frames, such as annually or monthly. This analysis aims to highlight the magnitude of criminal activities, over a timeline showing all the uphills and downhills of this crime activity.

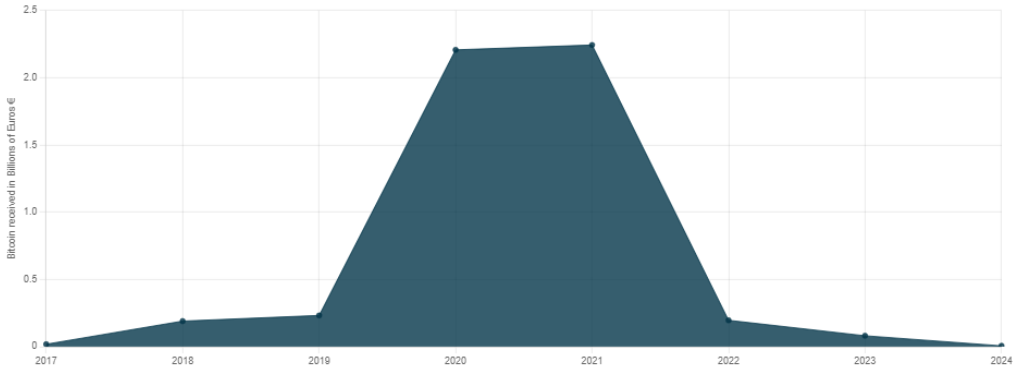

Figure 4.1: Annual revenue of cybercriminals through bitcoin.

In Figure 4.1, we can observe a surge in stolen euros orchestrated by malicious actors, reaching an all-time high in 2020 and 2021, as anticipated. This escalation can be attributed to various factors. Firstly, the unprecedented surge in Bitcoin prices incentivized criminals to resort to extortion through cryptocurrency. Secondly, the widespread adoption of remote working, prompted by the COVID-19 pandemic, provided cyber criminals with more opportunities to infiltrate companies and execute extortion schemes. Thirdly, following the war in Ukraine, numerous ransom groups led by both Russian and Ukrainian actors disbanded, e.g. the Conti Ransomware group [11] [13]. The dissolution of these groups resulted in the exposure of private communications among criminals, unveiling Bitcoin wallets utilized by them. Lastly, amidst the uncertainties brought about by the pandemic, and the perception of economic instability, individuals began turning to blockchain

investments. Unfortunately, many fell victim to scams in the process.

4.2.3 Bitcoin Crime per Abuse Classification

In this chapter, we sorted each wallet into specific groups without repeating any calculations. So, each wallet appears just once in each category total and once in the overall total. However, one wallet can end up in different categories from different sources. For example, OFAC might label a wallet as sanctioned, while RansomWhere sees it as a ransomware wallet.

The categories we used are Sanctions, Ransomware, Blackmail or Scam, Dark net Market, Bitcoin Tumbler, and Others.

Cime Classification	Stolen Money
Sanctions	2.66 Billion
Ransomware	1.92 Billion
Blackmail or Scam	634.83 Million
DarknetMarket	3.56 Million
BitcoinTumbler	745.49 Thousalnds
Other	29.01 Millions

Table 4.2: Stolen Bitcoin In Euros in each crime classification

Looking at Table 4.2, we can see the total amounts for each classification. The highest sums are concentrated in sanctions and ransomware. This aligns with our expectations, especially considering the surge in ransomware-as-a-service during 2020 and 2021. Following closely, we observe significant amounts in blackmail and scams, which is in line with the increase in crypto scams.

On the other hand, the funds associated with darknet markets and tumblers are notably lower. This was anticipated due to the inherent anonymity associated with these types of crimes.

4.2.4 Earnings in different Classifications over the years

In this section, we take a closer look at each category's trends over the years. Despite the overall peak in 2020 and 2021, the peak varies for each classification. Certain categories, such as Sanctions, Ransomware, and Scams, warrant a more in-depth exploration.

One particularly intriguing aspect is the comparison between the peaks of ransomware and sanctions. As mentioned earlier, the conditions in 2020 were conducive to a surge in ransomware attacks. In 2021, the emergence of ransomware as a service led to organizations like OFAC listing it in their sanctions. This intersection of ransomware and sanctions provides an interesting perspective on the evolving landscape of illicit financial activities. Additionally, it's noteworthy

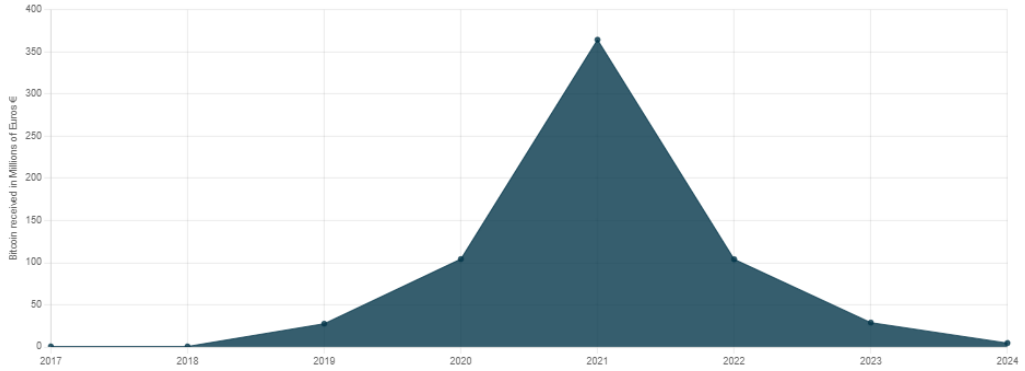

Figure 4.2: Money that has been stolen by scammers or collected by extortions over the years

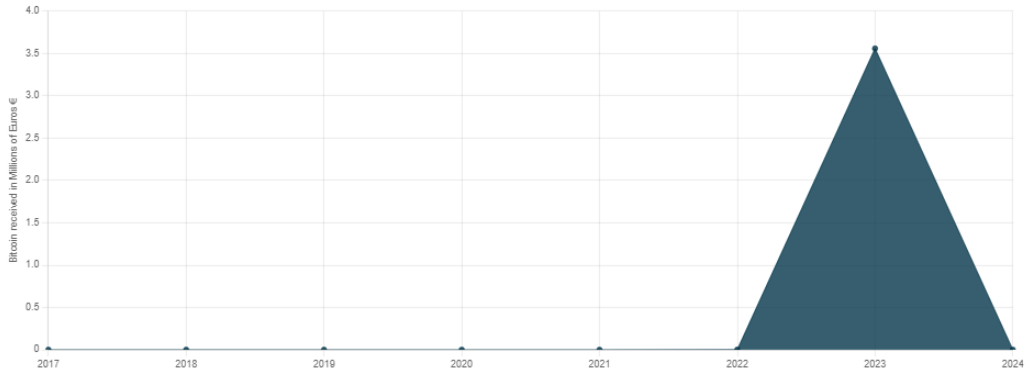

Figure 4.3: Money received by dark-net markets over the years

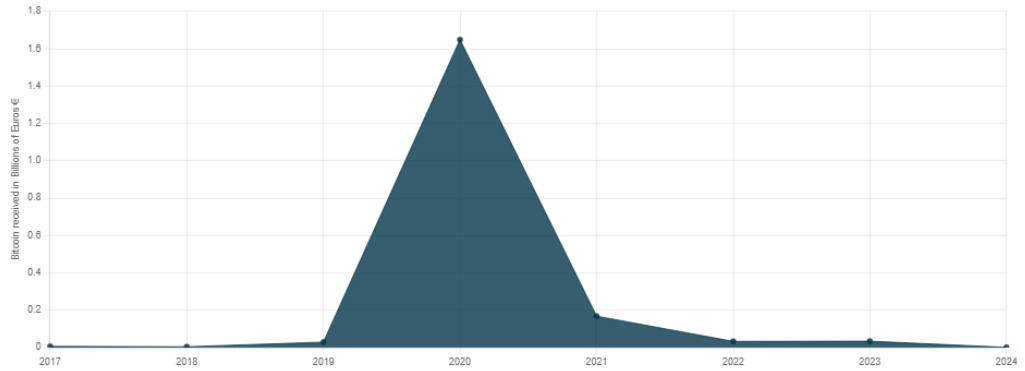

Figure 4.4: Stolen money collected by ransomware over the years

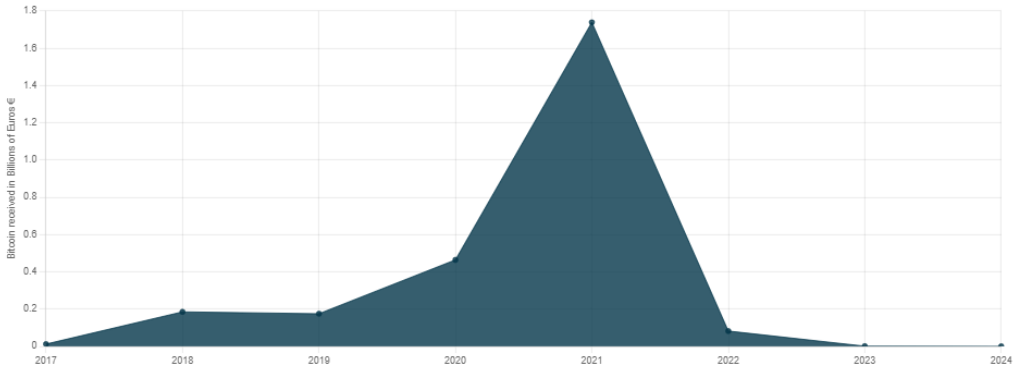

Figure 4.5: Money that has been collected over the years by wallets that have been sanctioned

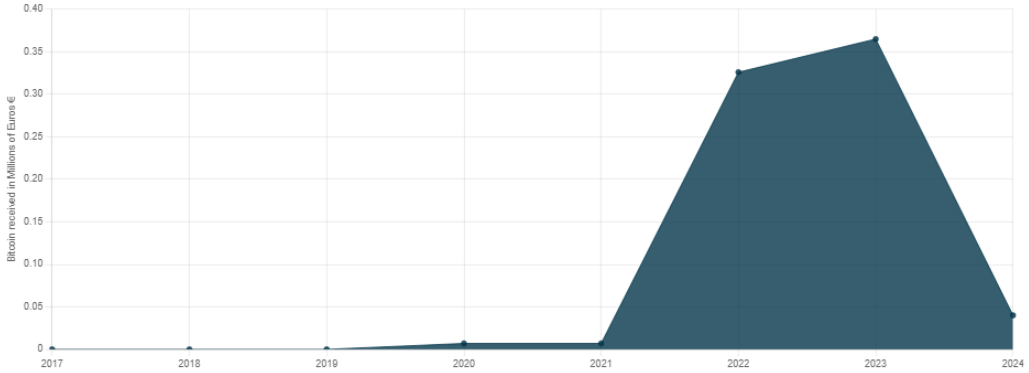

Figure 4.6: Money from wallets related to malicious bitcoin tumbling.

to observe the peak in blackmail and scams coinciding with the rise of NFT and other crypto scams.

4.2.5 Contribution of data sources in each crime classification

In this section, we extend our research to examine the contribution of each data source to the overall damage caused by cybercriminals. Building on the insights gained from earlier chapters regarding the overall damage and the impact of different classifications, we now focus on understanding the specific role played by each data source in the larger context of cybercrime.

By delving into the contribution of each source, we aim to provide an understanding of how various entities and platforms contribute to the overall financial implications of cybercriminal activities. This exploration will shed light on the dynamics between different sources and their influence on the financial landscape

Source	Contribute Wallets
Ofac	2.66 Billions
Alienvault	1.04 Billion
Ransomwhere	938.85 Millions
ChainAbuse	621.40 Millions
Ransomlook	134.70 Millions
CryptoScamDB	12.77 Millions
SophosLab	6.64 Millions

Table 4.3: Contribution (in euros) per data source.

impacted by cybercrime.

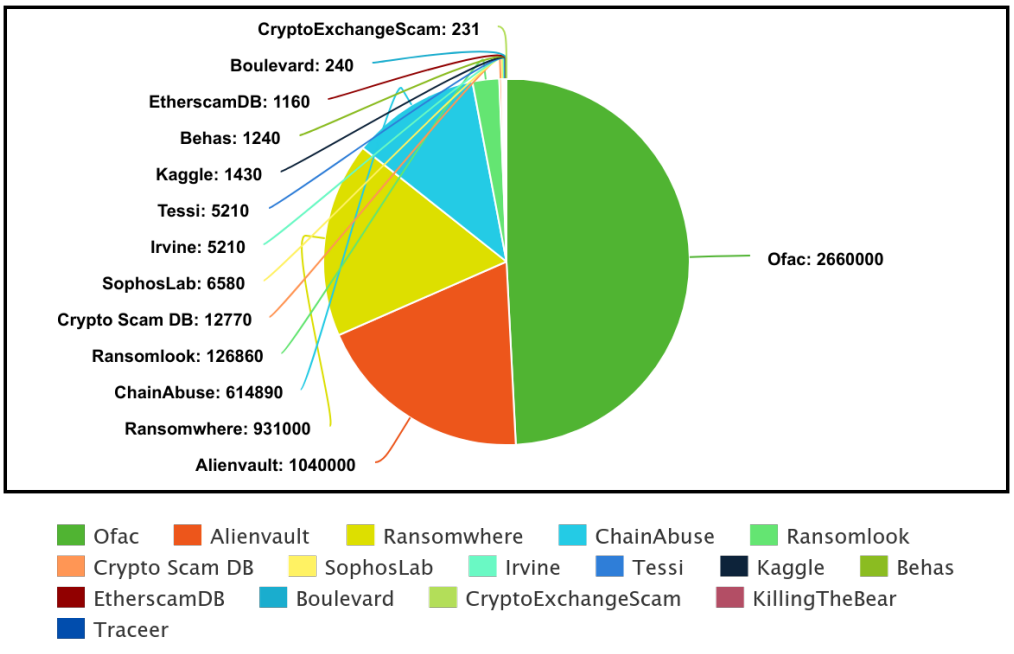

Figure 4.7: The Contribution of each source in millions of euro

Observing Table 4.3, several noteworthy patterns emerge. The most significant contributor is OFAC, which aligns with expectations given its role in sanctioning organized crime. Despite having fewer wallets, OFAC commands a substantial portion of the overall contribution. Following closely are major threat intelligence platforms, including ChainAbuse, AlienVault, and Ransomwhere, which collectively contribute significantly to the landscape. In contrast, all other sources make comparatively smaller contributions, emphasizing the concentrated impact of key

entities in shaping the overall understanding of cybercrime and its financial implications.

4.2.6 Contribution of data sources in each crime classification

We now examine the contribution of each source within each category is a crucial step in gaining a more granular understanding of the data. This analysis will unveil whether a particular source has a deep focus on specific categories or broader coverage across multiple categories.

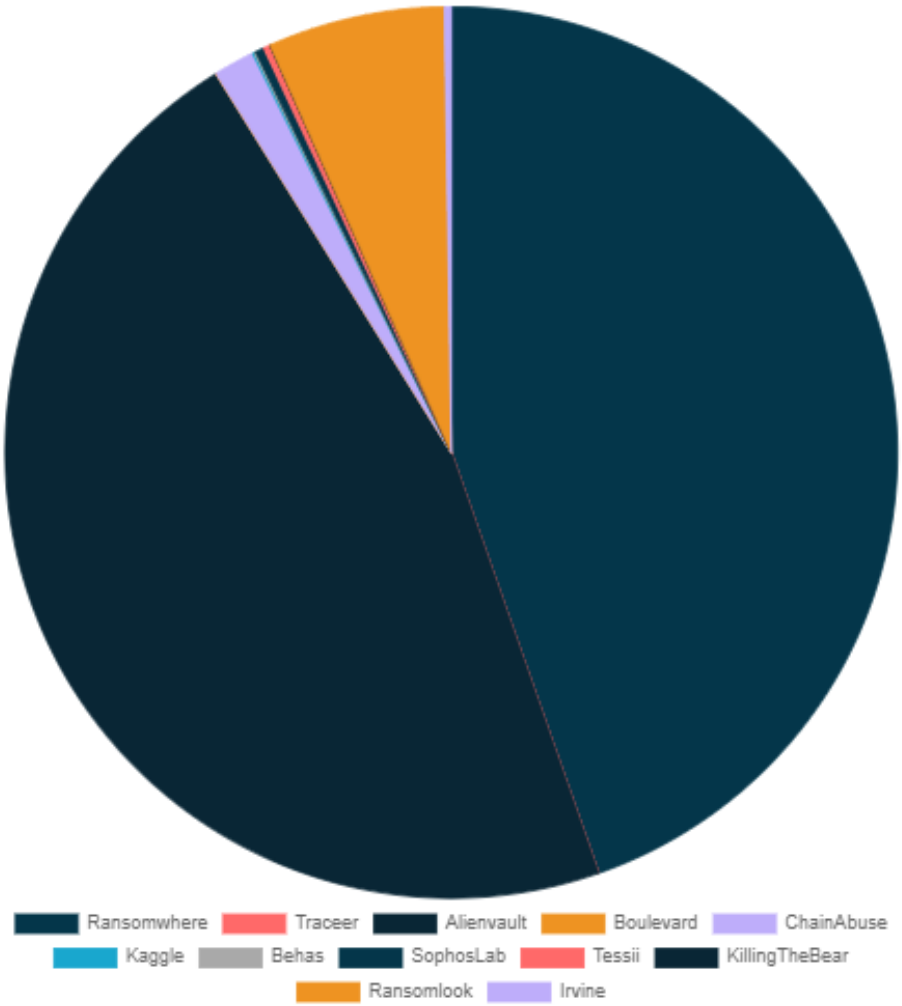

Figure 4.8: The Contribution of each source to the ransomware dataset based on the euro value.

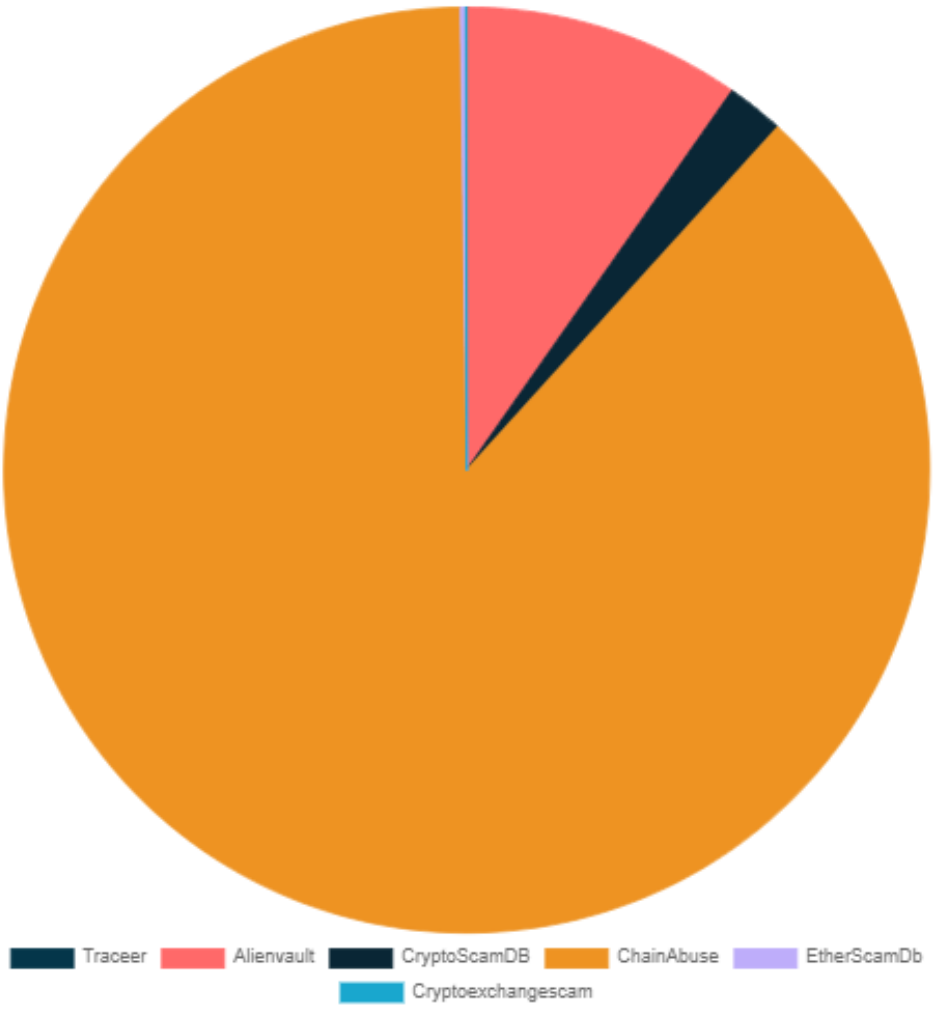

Figure 4.9: The Contribution of each source to the Blackmail and scam dataset based on the euro value

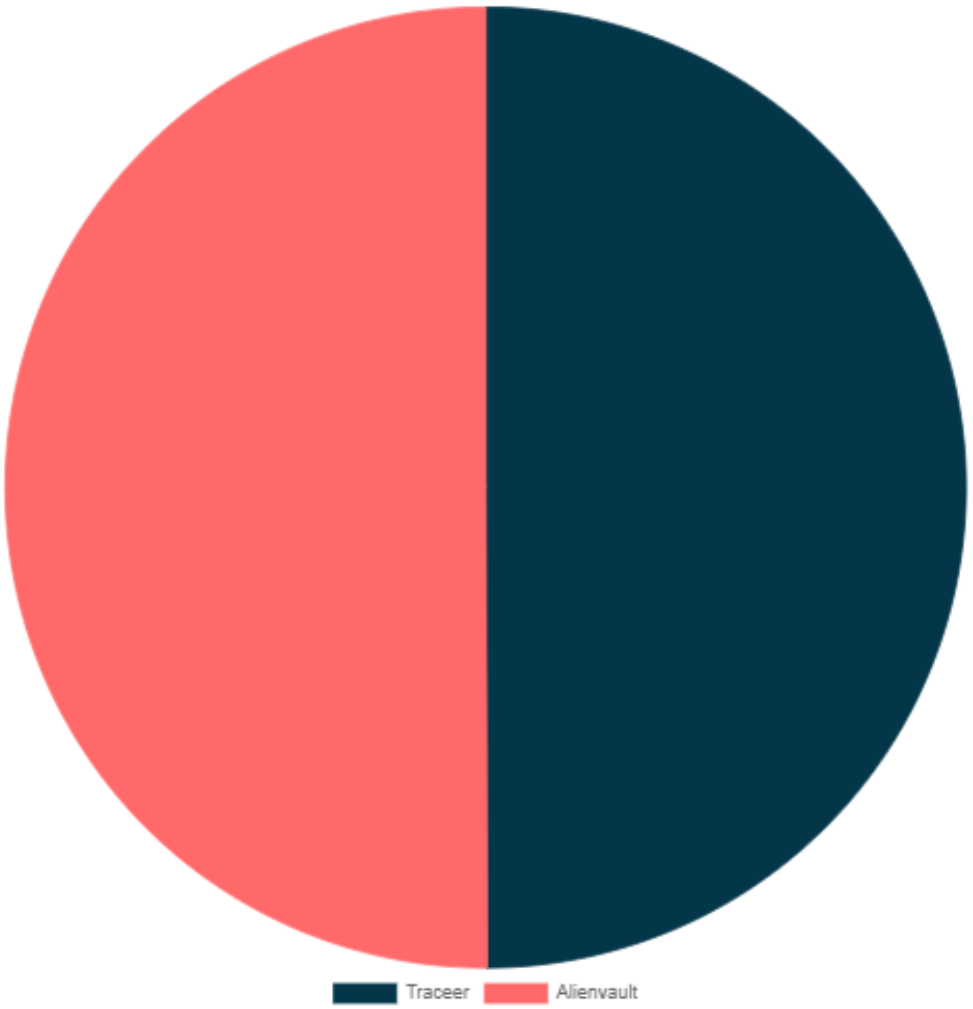

Figure 4.10: The Contribution of each source in millions of euro

Analyzing the data several intriguing patterns come to light. Firstly, OFAC stands out as the primary contributor to the overall amount in the sanctions category, reinforcing its focus on organized crime and sanctions.

Secondly, from the Figures 4.8- 4.10 ChainAbuse emerges as a noteworthy source, offering data in multiple categories in substantial portions. This can be attributed to the crowdsourcing nature of ChainAbuse, allowing it to cover a broad spectrum of cybercrime activities.

Thirdly, an interesting observation is the equal contribution of AlienVault and Ransomwhere in the ransomware category Figure 4.8. This parity is logical considering that both sources heavily rely on research-based resources, making them more inclined to contribute significantly to a subject like ransomware as opposed to areas like Nigerian scams or fake extortion.

Lastly, only crowd-sourced platforms have wallets unclassified (classified as others).

4.2.7 Source Coloration

Given the observation that the total amount is less than the sum of all sources combined, it suggests an overlap or sharing of wallets among the sources. So in this chapter we will review the intersection of the sources.It is crucial in unveiling an understanding of the reliability of certain wallets. Additionally, it provides insights into the collaborative nature or distinct perspectives of the sources. This exploration of source intersections enhances the credibility of identified wallets and enriches the overall reliability and context of the gathered data. Due to the big number of sources we choose to create a heatmap. Each axis has all the sources as value. Each value representing the jaccard similarity. Meaning more similar sources have bigger value ranging from zero to hundred.

Excluding the diagonal line, which represents self-comparisons for each source we observe some great insights. Such as

The sources irivine and Tessi are copying each other. Ransomlook and Ransomwhere has almost the same data. This is happening because Ransomlook use the data of ransomwhere.

Unexpectedly we learn that behas use the same data as ransomwhere.

Lastly, we observe that crowd-source platforms offer unique data due to their nature of them. While still maintaining a small coloration proving the validity of their data.

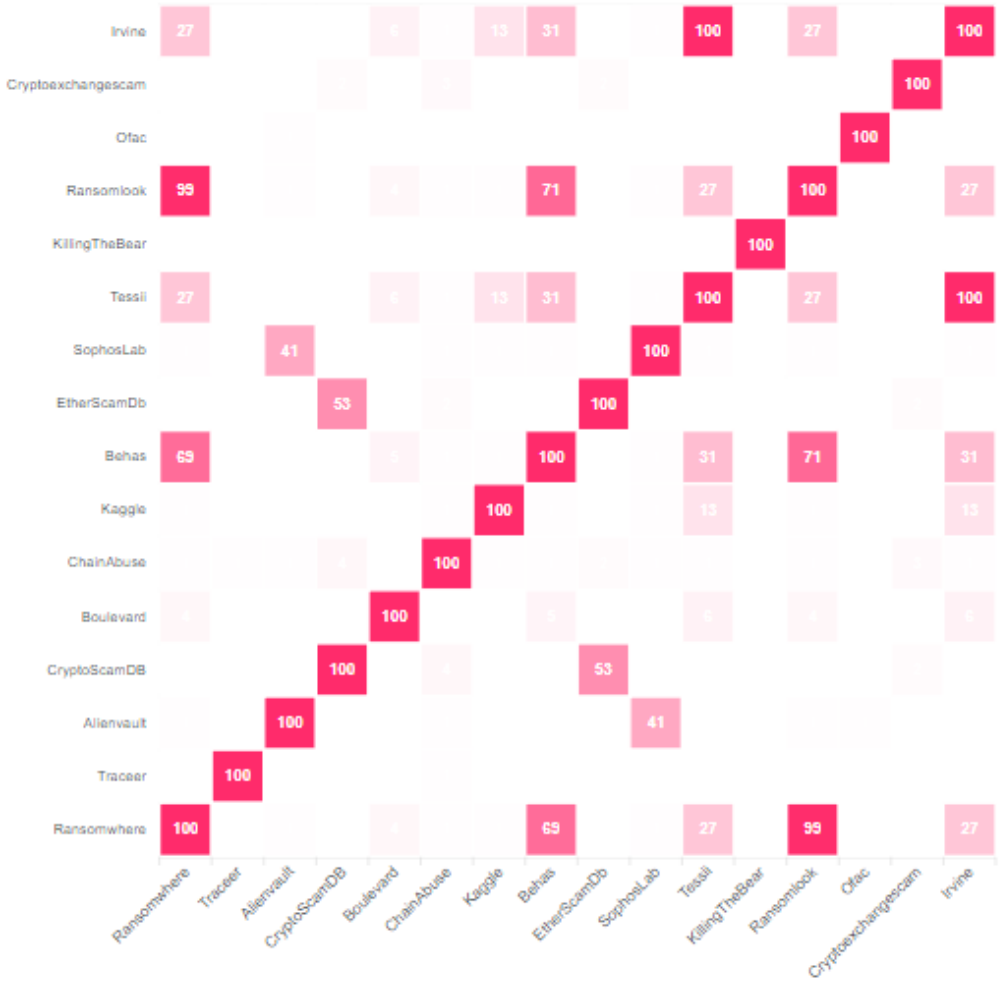

Figure 4.11: Intersection of each source with others

Chapter 5

Related Work

In order to conduct our analysis and create this system we use and extend methodologies from cryptocurrency tracking, leaked cybercrime data, ransomware analysis, and open threat intelligence platforms. The related work can be divided into two main categories, Academic Contributions and Open Threat Intelligence Platforms.

5.1 Academic Contributions

In this section, we compare our contribution with existing state-of-the-art research. Previous research studied the Ransomware Dark Web payments to quantify revenue from cryptocurrencies.

Chainalysis: [7] publishes annual reports with estimations about the total revenue of illicit activity on the Dark Web and per category. Although the analysis provides valuable policy-making insights, their methodology is proprietary.

Christin [8] crawled the Silk Road Marketplace and found it was primarily drug-oriented.

A Tale of Two Markets [13] compares the ransomware as a commodity and Ransomware as a Service (RaaS).

Gibran Gomez [10] performs the first systematic analysis on the estimation of cybercrime bitcoin revenue. They implement a tool that can replicate different estimation methodologies. They also compare existing methodologies and reveal underestimations and overestimations in quantifying the revenue of cybercriminals.

Money Over Morals [11] the authors leverage leaked chat messages to provide an in-depth empirical analysis of Conti, one of the largest ransomware groups.

Quantifying Dark Web Shops' Illicit Revenue [14] presents a methodology to estimate the size and nature of illicit commercial activity on the Dark Web, specifically focusing on single-vendor Dark Web Shops. The study reveals that in 2021, Dark Web Shops generated at least 113 million USD in revenue, with sexual abuse being the top illicit revenue category.

Ransomware payments in the Bitcoin ecosystem [15] investigates how

ransomware attacks use Bitcoin as payments. The authors discovered many ransomware groups don't make as much money as people might think. By studying a variety of ransomware families, the research showed that most attacks are not very advanced and ask for small amounts of money.

To our knowledge, our analysis is one of the largest out there. At the same time, all of our data is public and we study different types of cybercrime in the blockchain. Finally, we believe that we do not over or under-estimate the resulting revenue from the cybercriminals.

5.2 Threat Intelligence Platforms

In this section, we compare our application with other applications in this field. Many of the projects we examined have contributed to our study by introducing problem solutions, providing data, and raising issues that have ultimately enhanced our tool.

Chain Abuse: stands out as the primary reporting platform for malicious crypto activity worldwide. Anyone can visit the site and submit a report; subsequently, these reports can be displayed individually or grouped based on cryptocurrency, crime classifications, blockchain wallet, or specific criminal activities. However, as a crowded reporting platform, it segregates its user base according to three credibility levels: unverified, checked, and trusted.

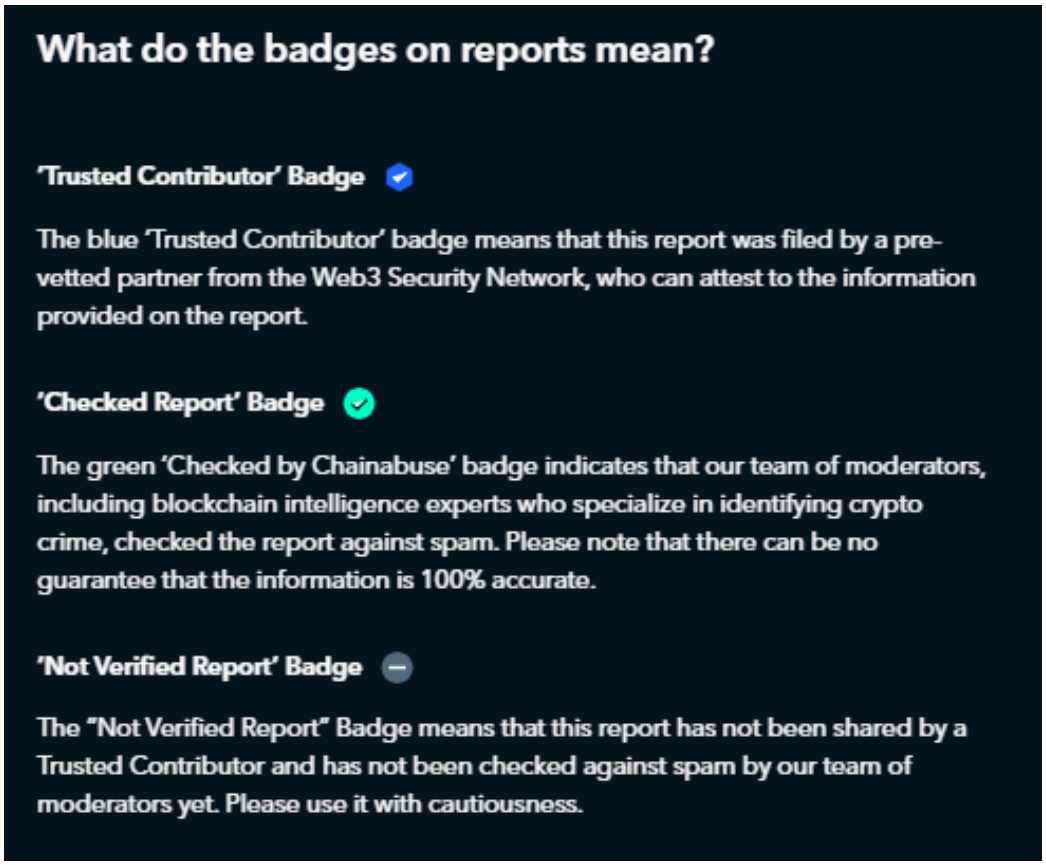

Figure 5.1: Chain Abuse badges description

Bitcoin Abuse Similar to Chainabuse, Bitcoinabuse serves as a public database of bitcoin addresses used by scammers, hackers, and criminals. Like Chainabuse, the data is sourced solely from crowd reporting, without providing any guarantees for the credibility of the sources. Bitcoinabuse also presented diagrams and statistics based on the available data. However, since 2023, Bitcoinabuse has been closed, and all data has been integrated from Chainabuse.

Chainalysis A blockchain data platform [7]. Unlike others is very user-friendly providing detailed reports every year, and smaller reports more often than that. But Chainalysis and others like it (crystal blockchain) are private companies and do not share their sources. As a result, their data can't be verified.

Ransomwhere An open, crowdsourced ransomware payment tracker. In contrast to Bitcoin Abuse and Chainabuse, Ransomwhere is primarily the work of a single researcher named Jacked Kable, who based the dataset on leaked communications of the ransomware team Conti.

AlienVault A Threat Intelligence Community platform where reports are contributed by trusted users. However, akin to other sources, the tool is primarily tailored for developers, lacking user-friendly features such as plots or statistics.

Chapter 6

Limitations

Throughout this study, we have identified four major limitations that significantly impact the usability of similar tools.

Data Formats: One significant limitation lies in the diversity of data formats chosen by each maintainer for publishing their data. While it is understandable that collecting data from various sources can be time-consuming, adapting to a specific format each time necessitates a particular structure for the fetching mechanism.

Maintenance: Despite successfully collecting a substantial amount of data from multiple sources, we have observed numerous datasets that have not been maintained over the years. This implies that, over time, our tool will require continuous updates with new sources, or it may become obsolete.

Credibility: As mentioned earlier, a notable concern is the lack of data verification in many datasets. The absence of such verification compromises the trustworthiness of any tool relying on these datasets.

Usability: Even after addressing the aforementioned limitations, the tool must be designed to be self-explanatory to ensure accessibility for every user. Many existing tools are designed for developers rather than ordinary users, thus making the public unable to comprehend the severity of the issue. It is imperative to create tools that bridge this gap and make the subject matter more comprehensible to a broader audience.

Chapter 7

Future Work

Through the development of this tool, we have successfully created a comprehensive system that effectively exposes a significant portion of cybercrime within the blockchain. Our approach involves the gathering and parsing of data from multiple open threat intelligence sources. However, it is worth noting that despite our efforts, there may exist additional sources of information that we have yet to discover or validate. Therefore, to uphold the accuracy of our tool, it is imperative to continuously update its database.

In addition to data accuracy, we recognize the importance of user experience. Thus, ongoing efforts are directed towards improving the user interface (UI) and incorporating more informative statistical plots and content. These enhancements aim to better serve the needs of users and facilitate their interaction with the tool effectively.

Initially, our focus has primarily been on Bitcoin due to the prevalence of criminal activities within its ecosystem and the pseudo-anonymity it offers, which aids in monitoring criminal wallets. However, despite the scale of crime within Bitcoin, it is essential to broaden our monitoring efforts to include other blockchains. Each blockchain presents unique insights and challenges, necessitating an extension of our tool's functionality to cover a broader spectrum of cryptocurrencies.

Moreover, leveraging the extensive open-source and reliable dataset we have compiled, we aim to develop a machine learning model [17] [18] [12] [16] [9] [5] to accurately predict various aspects related to wallet ownership and associated criminal activities. This model will be instrumental in identifying malicious wallet owners, determining their involvement in cybercrime, and classifying the types of crimes committed. Furthermore, we intend to utilize machine learning, either through the previously mentioned model or alternative approaches, to expand our research to include the graph neighbors of identified malicious wallets. This expansion will enable a more comprehensive understanding of criminal networks and aid in the development of proactive measures against cyber threats.

Chapter 8

Conclusion

In conclusion, the development and implementation of a three-layer system, comprising a parsing module, data processing, and visualizations, has proven to be effective in handling and analyzing threat intelligence datasets. This system offers valuable insights into the impact provoked by malicious actors by visualizing the extent of damage. Moreover, it provides a more comprehensive perspective, as the final damage calculation considers all transactions rather than focusing only on a report of stolen money from a specific user. The parsing module played a crucial role in extracting data from diverse sources, including API endpoints, datasets, and platforms. The data processing module efficiently transformed raw data from our sources into a uniquely structured format suitable for subsequent processing tasks. This included tasks such as transaction gathering, calculation, grouping of data, and the ability to produce comprehensive statistics. With the statistics ready a series of endpoints could produce the data upon request to the dashboard that visualizes them in the best possible way. Utilizing data visualization techniques, including charts and interactive boards, the system offered users a comprehensive overview of the analyzed data, enabling the observation of attack patterns and trends. Additionally, the system was intentionally designed to be extensible, accommodating future publicly available data and presenting it with simple yet meaningful representations.

In summary, the system not only contributes to the threat intelligence community by incorporating data from the most trusted sources but also distinguishes itself as user-friendly, reliable, and stable.

Bibliography

[1] Alievault. https://www.alienvault.com/.

[2] Bitcoin heist ransomware address dataset. https://archive.ics.uci.edu/dataset/526/bitcoinheistransomwareaddressdataset.

[3] Blockchain api. https://www.blockchain.com/explorer/api/blockchain_api.

[4] Office of foreign assets control. https://ofac.treasury.gov/.

[5] Cuneyt Gurcan Akcora, Yitao Li, Yulia R. Gel, and Murat Kantarcioglu. Bitcoinheist: Topological data analysis for ransomware detection on the bitcoin blockchain, 2019.

[6] Jack Cable. Ransomwhere: A Crowdsourced Ransomware Payment Dataset, May 2022.

[7] Chainalysis. The 2022 cryptocrime report. 2022.

[8] Nicolas Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace, 2012.

[9] Siddhartha Dalal, Zihe Wang, and Siddhanth Sabharwal. Identifying ransomware actors in the bitcoin network, 2021.

[10] Gibran Gomez, Kevin van Liebergen, and Juan Caballero. Cybercrime bitcoin revenue estimations: Quantifying the impact of methodology and coverage. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23. ACM, November 2023.

[11] Ian W. Gray, Jack Cable, Benjamin Brown, Vlad Cuiujuclu, and Damon McCoy. Money over morals: A business analysis of conti ransomware, 2023.

[12] Noor Nayyer, Nadeem Javaid, Mariam Akbar, Abdulaziz Aldegheishem, Nabil Alrajeh, and Mohsin Jamil. A new framework for fraud detection in bitcoin transactions through ensemble stacking model in smart cities. *IEEE Access*, 11:90916–90938, 2023.

[13] Kris Oosthoek, Jack Cable, and Georgios Smaragdakis. A tale of two markets: Investigating the ransomware payments economy, 2022.

[14] Kris Oosthoek, Mark Van Staalduinen, and Georgios Smaragdakis. Quantifying dark web shops' illicit revenue. *IEEE Access*, 11:4794–4808, 2023.

[15] Masarah Paquet-Clouston, Bernhard Haslhofer, and Benoît Dupont. Ransomware payments in the Bitcoin ecosystem. *Journal of Cybersecurity*, 5(1):tyz003, 05 2019.

[16] Mohamed Rahouti, Kaiqi Xiong, and Nasir Ghani. Bitcoin concepts, threats, and machine-learning security solutions. *IEEE Access*, 6:67189–67205, 2018.

[17] Mohammad Javad Shayegan and Hamid Reza Sabor. A collective anomaly detection method over bitcoin network, 2021.

[18] Kai Wang, Jun Pang, Dingjie Chen, Yu Zhao, Dapeng Huang, Chen Chen, and Weili Han. A large-scale empirical analysis of ransomware activities in bitcoin. *ACM Trans. Web*, 16(2), dec 2021.