A Web Scraping Approach Towards Cryptocurrency Investigations

Bongani Mawhayi[1], Johnny Botha[2] and Louise Leenen[1,3]
[1]University of the Western Cape, Cape Town, South Africa
[2]Council for Scientific and Industrial Research, Pretoria, South Africa
[3]The Centre for Artificial Intelligence Research, Cape Town, South Africa

4013109@myuwc.ac.za
jbotha1@csir.co.za
lleenen@uwc.ac.za

Abstract : The investigation of cryptocurrency crimes is still in its infancy with no standardised process or methodology to follow. This paper describes research that forms part of a broader project led by the second author (Botha, et al., 2025). The broader project's aim is to develop a methodology to follow when conducting cryptocurrency crime investigations. One of the steps in the proposed methodology is web scraping. The authors of this paper present a detailed exploration of web scraping techniques within the broader context of the proposed investigation methodology. In this paper, the focus is on developing a well-structured methodology for scraping social media platforms and online forums to gather data related to fraudulent activities; the goal is to find posts that include references to the wallet address of interest. This exploration uses an iterative approach; for every new cryptocurrency wallet address discovered or revealed through on-chain analysis, a parallel path is followed by scraping the Internet. If a mention of the cryptocurrency address should be discovered it is considered to be a key finding, creating a pivot point in the investigation. From a pivot point, further open-source intelligence (OSINT) techniques will be applied, though this aspect falls beyond the scope of this paper. If no relevant information or link is found, the scraping path will not be pursued, and the investigation proceeds with on-chain analysis to identify additional wallet addresses. Additionally, challenges encountered in web scraping, such as handling platform restrictions, ensuring data accuracy, and managing large volumes of data, are addressed. The goal of the proposed methodology is to enhance data extraction and analysis efficiency contributing to the proposed methodology for investigating cryptocurrency scams.

Keywords: Blockchain, Crypto-crime, Cryptocurrency, Crypto-scam, Web scraping

1. Introduction

In the rapidly expanding and evolving domain of digital finance, there has been a simultaneous rise in cryptocurrency-related scams. This rise in cryptocurrency scams has exposed a significant shortcoming in the absence of a standardised methodology for investigating such fraudulent activities. The broader project that this paper forms part of (Botha, et al., 2025), addresses this critical gap by presenting a systematic approach for investigating cryptocurrency scams, and highlights the urgent need for a structured investigative methodology in this field. This paper presents research on web scraping and provides an in-depth analysis of web scraping techniques, focusing on their application for collecting and analysing data from social media platforms and online forums to trace connections to wallet addresses implicated in scams.

Cryptocurrency scams have been prevalent since Bitcoin's early days, exploiting the lack of regulation and the novelty of the industry. There are various types of crypto scams, including giveaway scams, where scammers lure victims with promises of receiving cryptocurrencies or assets in exchange for small contributions. These scams typically start on social media platforms like YouTube, Twitter/X, and Instagram, often using fake promotions or live streams featuring celebrities to deceive users (Botha, et al., 2023). Impersonation scams follow a similar structure, with scammers pretending to be celebrities, legitimate businesses, or government agents on social media platforms. Victims are contacted via direct messages offering fake crypto investment advice and are directed to send funds, often through WhatsApp, before being asked for additional payments like withdrawal fees (Botha, et al., 2023) (Department of Financial Protection and Innovation, 2025). Phishing scams involve tricking victims into revealing sensitive information such as wallet seed phrases or login credentials by impersonating legitimate platforms. Scammers may send fake password reset emails or create convincing replicas of cryptocurrency exchange websites to gain access to victim funds (Botha, et al., 2023). Another common scam is the pump-and-dump scheme, where the price of a cryptocurrency is artificially inflated to create hype, with scammers selling off their tokens, leaving late investors at a loss. These schemes often take place on platforms like Telegram and typically involve newly formed, unregulated cryptocurrencies (Botha, et al., 2023). An example of this scam is the SaveTheKids cryptocurrency, promoted by influencers who made significant profits while leaving investors with worthless tokens. The scam falsely associated the project with the legitimate Save the Children organization by using a similar logo and vague details about charitable donations (Levy, 2023).

While the tactics that scammers deploy vary, the end goal of these scams is always the same – to exploit unsuspecting victims for financial gain. Staying informed about tactics, practising vigilance, and implementing security measures are critical steps for users and investors to safeguard their assets. For a more comprehensive analysis of cryptocurrency scams refer to Bartoletti et al (2021). This study provides an extensive review of the types of scams prevalent in the cryptocurrency domain.

This paper examines the complexities and methodologies involved in investigating cryptocurrency scams, with a particular focus on the role of web scraping in gathering evidence and tracking fraudulent activities. Section 2 explores the challenges faced by investigators, including the pseudonymous nature of cryptocurrencies, jurisdictional issues, and the use of mixers to obscure transaction flows. Section 3 outlines the methodology employed, detailing the process of on-chain analysis and web scraping to identify and track scam-related wallet addresses. Section 4 introduces a proposed approach for applying web scraping techniques to gather relevant data from various platforms, creating pivot points for further investigation. Section 5 highlights the challenges and limitations of this approach, particularly regarding data accuracy, legal concerns, and the maintenance of scraping tools. The paper concludes in Section 6, summarising the findings and offering insights into the future of cryptocurrency scam investigations.

2. Challenges in Investigating Crypto Scams

The decentralised and pseudonymous nature of cryptocurrencies can obscure the identities of perpetrators, making it difficult to attribute fraudulent activity to specific individuals or entities. While cryptocurrency transactions are not entirely anonymous, blockchain analysis tools can help trace hidden assets. There have been investigations that demonstrate the feasibility of tracing hidden cryptocurrency assets and that highlight the importance of legal action (Botha & Leenen, 2024). Moreover, jurisdictional issues and cross-border complexities further complicate the investigative process, necessitating international cooperation and coordination among law enforcement agencies. However, criminals' use of tumblers and mixers to obscure the flow of funds adds another layer of complexity. Cryptocurrency mixing or tumbling involves blending tokens with others through numerous transactions, making it difficult to trace the source and destination (Botha, et al., 2023). While investigators can identify that funds were transferred to a mixer and later received by another party, establishing a clear connection between them remains challenging (Gertenbach, et al., 2024).

Social media has become essential in cryptocurrency investigations providing real-time data and crowdsourced intelligence. Platforms like Facebook, X, and Instagram allow investigators to monitor public sentiment and track criminal activity faster than traditional media (Social Links, 2023) (TechMindXperts, 2023). This enables early detection of scams as seen when social media reports fraudulent schemes before traditional outlets (Social Links, 2023). One of the most crucial challenges, however, lies in identifying specific mentions of wallet addresses across vast amounts of data. Criminals often reference or promote illicit wallet addresses in posts, which investigators must track to link them to fraudulent activities. With platforms generating enormous amounts of data – such as Facebook's 1.7 million posts per minute – finding relevant posts that reference these addresses becomes a daunting task. Advanced scraping tools are essential for filtering through this content to identify these key references quickly. Despite these challenges, social media remains valuable for identifying emerging scams and spotting deceptive patterns (Social Links, 2023). However, there are challenges presented by certain platforms such as X, that recently made changes to its scraping policy. X imposed a paywall for application programming interface (API) access, which restricted researchers and law enforcement from scraping real-time data from the platform. The cost for enterprise-level API access skyrocketed to $42,000 per month, making it financially unfeasible for many research projects to continue (Calma, 2023). This change has had a detrimental effect on academic research, including studies related to cryptocurrency scams, misinformation, and online behaviour. The shift also creates legal uncertainties for researchers, who are concerned about potential litigation surrounding data scraping practices (Calma, 2023) (Gotfredsen, 2023).

The intersection of web scraping and cryptocurrency investigations signifies a simplification of extracting essential information from the vast digital ocean of information. Web scraping is an automated process for extracting data from websites, effectively replacing the manual process of searching and copying information. It begins with sending Hypertext Transfer Protocol (HTTP) requests to retrieve the Hypertext Markup Language (HTML) content of web pages, making use of programming libraries such as Requests in Python (Singrodia, et al., 2019). Once retrieved, the HTML content is parsed using tools like BeautifulSoup, which enables efficient extraction of meaningful data from the HTML tags of websites and storing it in a central local database, spreadsheet, or JavaScript Object Notation (JSON) file. These specialised programs or scripts can be tailored for specific websites or generalised to work with any website. These tools enable scraping tasks to be performed by

machines, reducing the reliance on humans for laborious and error-prone manual data collection processes. The data extracted through web scraping can be converted into accessible formats like JSON, XML, CSV, XLS, or RSS, suitable for advanced processing and integration (Singrodia, et al., 2019). The main goal of these specialised program scripts is to convert unstructured data from websites/platforms into organised data.

The advantages of web scraping are substantial as it provides error-free data, saves time by delivering rapid results, and consolidates all data in one accessible location. Users can choose the format in which data is stored, facilitating easy access and analysis. This automation frees users from the tedious and error-prone tasks of manual data retrieval, enabling more efficient use of resources for complex analysis and decision-making. In the context of investigating crypto scams, web scraping can be particularly valuable for extracting relevant data points such as social media post metadata – user information, location, post ID, post description – that contain the target wallet address. This automated approach significantly streamlines the process of gathering crucial evidence and identifying pivot points, which would otherwise be time-consuming and labour-intensive (Gong, 2024)

Furthermore, while web scraping offers numerous benefits, it also presents challenges, including potential legal issues and the need for continuous maintenance due to changes in website structures. The legality of web scraping is a contentious issue, and practitioners must navigate legal and ethical considerations to ensure compliance with regulations and respect for the rights of website owners (Krotov, et al., 2020). Web scraping's impact on stakeholders, including website owners and users, raises social issues that warrant careful consideration (Krotov, et al., 2020). Web scraping is an indispensable tool for automating data extraction from the web, transforming unstructured web data into structured formats that can be utilised for diverse applications.

3. Methodology

The research methodology involves the selection of a relevant cryptocurrency scam use case, which serves as a basis for evaluating the effectiveness of web scraping. After identifying wallet addresses linked to the scam, web scraping is employed to collect information from various online sources. The collected data aims to support further investigation by providing leads and contextual information for OSINT analyses. However, the scope of this paper is limited to the application of web scraping techniques, and the subsequent OSINT analysis, while integral to the overall investigation process, will be addressed in future research.

3.1 Use Case Selection

The specific choice of a case is not critical to the applicability of the proposed process, as it is designed to be versatile and applicable to a range of scams. For this investigation, a Bitcoin-related impersonation scam was selected due to its prevalence and distinctive nature within the cryptocurrency landscape (Popov, 2023). The type of scam itself was chosen at random, with the primary focus being on showcasing the proposed investigative approach. The case was chosen from ChainAbuse. In the selected case, the victim received a call from an individual impersonating a Social Security Agent. The caller falsely claimed that the victim had made a large unauthorised purchase and that their identity had been stolen. The victim was falsely accused of serious crimes and was pressured to pay a significant amount to resolve the alleged issues. The victim made a payment to a wallet address, identified as the redacted version bc1qhl...avd. The aim of this analaysis is to follow the funds associated with this wallet address and reveal additional addresses, with the a particular focus on outgoing transactions. This case exemplifies the tactics employed in impersonation scams and underscores the importance of the chosen investigative approach

3.2 On-Chain Analysis

Analysing blockchain data presents significant challenges due to its inherent complexity and the pseudonymous nature of transactions. On-chain analysis is a crucial technique employed by investigators, that involves examining the data recorded directly on a blockchain, including every transaction, smart contract execution, and network activity (Tradivest, 2024). This process allows investigators to uncover patterns and anomalies, enabling them to trace the flow of funds on the blockchain and identify suspicious transactions associated with illicit activities (Social Links, 2023). However, the data is often unstructured, complicating the extraction of meaningful insights from the transaction records. The vast amounts of information stored on blockchains can make analysis computationally intensive. Additionally, the pseudonymous nature of blockchain transactions makes it difficult to link transactions to specific users, while encryption adds another layer of complexity, requiring specialised techniques to access the data. Moreover, on-chain data frequently lacks contextual information, which hinders the understanding of the motives or purposes behind transactions (UEEx, 2024). These challenges create a

complex landscape for analysts, highlighting the necessity for effective tools like Breadcrumbs, Chainalysis, TRM Labs or Maltego to facilitate meaningful analysis for investigators (Botha & Leenen, 2024).

In this investigation, in particular, Breadcrumbs was employed for its advanced visualisation capabilities. At the same time, other tools offer similar functionalities, Breadcrumbs was selected due to its affordability and the availability of a free version, which was sufficient for the requirements of this investigation. This tool provides a visual representation of the connections between wallet addresses in the form of a graph, with each node representing an individual wallet address. It traces the origin and destination of cryptocurrency funds, facilitating the analysis of transaction relationships, fund movements, and entity affiliations. An additional advantage of Breadcrumbs is its ability to provide enriched data, and indication of whether an address is linked to an exchange, sanctioned, or reported as fraudulent or suspicious. By utilising features such as node size, line thickness, and arrow direction, Breadcrumbs helps convey the flow and significance of transactions, thereby mitigating the difficulties posed by the unstructured nature of blockchain data.

During the on-chain analysis, the process was halted after identifying 75 wallet addresses connected to the original wallet address. The authors deemed this number sufficient to initiate the investigation, balancing the need for a robust dataset with practical constraints. This sample size was considered reasonable, striking a balance between capturing meaningful relationships and avoiding computational limitations, while being sufficient for proof of concept. Extending the analysis further could introduce diminishing returns, as larger datasets would require increased computational resources, more complex visualisations, and additional processing time. The more nodes in the graph, the greater the complexity and rendering load, as the relationships between wallet addresses expand. During the on-chain analysis, certain nodes were hidden to reduce noise and declutter the graph, enabling a clearer focus on important relationships and data. This approach was essential for data organization, as it removes unnecessary information and emphasises critical connections, as shown in the filtered graph in ***Figure 1***. The following criteria were specifically used to determine which nodes to hide:

- The volume of transactions, both inbound and outbound, were examined to assess the activity of each wallet.
- The timeline of token usage with each wallet was analysed to determine if a wallet was dormant. Wallets that had not been active in the last three months were hidden, although their addresses were still documented for future reference.
- The balance of each wallet was also considered. Wallets containing substantial amounts of Bitcoin were prioritised, even if they appeared to be abandoned, dormant or a tumbler

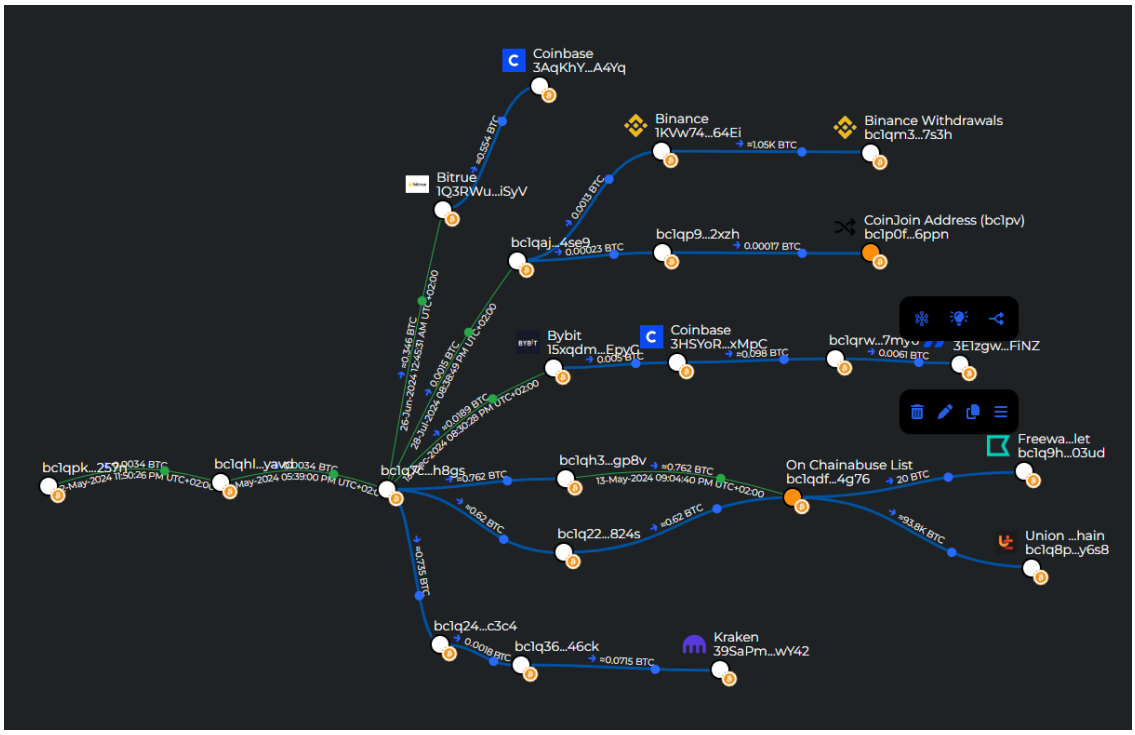


Figure 1: On-Chain Analysis

To ensure comprehensive tracking and analysis of wallet addresses during the investigation, a structured approach was implemented. Each newly discovered wallet address, whether active or dormant, was recorded in a separate file. This systematic approach ensured the preservation and organisation of all wallet addresses, facilitating further investigation through web scraping on selected social media platforms. The focus on effective data management enabled efficient tracking and analysis of wallet addresses, enhancing the overall integrity and effectiveness of the research.

Figure 1 *(only 22 nodes displayed to present the concept for simplicity and readability)* illustrates a Breadcrumbs graph, starting with the scammer's wallet address bc1qhl...yavd, which shows a single incoming transaction from the victim's wallet address bc1qpk...257n to the left of the scammer's address. This transaction serves as confirmation of the victim's involvement, as there is only a single incoming payment to the scammer's wallet and a single outgoing transaction to bc1qhl…yavd. From there, multiple outgoing transactions lead to various wallet addresses, with some of these transactions linking to well-known cryptocurrency exchanges such as Bitrue, Bybit, Kraken, Binance, Coinbase, and Bitvavo. Notably, one transaction is connected to an address listed on the Crypto Defenders Alliance (CDA) blacklist, which tracks and flags wallet addresses involved in fraudulent or suspicious activity (Crypto Defenders Alliance, 2024). Additionally, a Binance withdrawal is identified, possibly indicating a transfer to a bank, which could provide valuable clues for uncovering bank account information. Another address is linked to CoinJoin, an anonymization strategy that obscures transaction addresses and amounts, making it difficult to trace the funds (Hayes, 2024). This address is flagged (marked orange) as a high-risk address due to its use of CoinJoin. Furthermore, a separate wallet address has been linked to Chainabuse, suggesting that apart from the initial wallet address reported, other wallet addresses are actively being used in a network of scams.

All the additional information provided by Breadcrumbs is crucial for investigators, as it helps identify links to exchanges. Law enforcement can then use this data—along with the transaction ID and address information—to request subpoenas from exchanges, potentially revealing the identities of those connected to the scam. To explore additional connections, scraping techniques were then employed for all identified addresses, including those associated with the cryptocurrency exchanges.

3.3 Web Scraping

Web scraping techniques were utilised to investigate scams and gather relevant data from social media platforms and websites associated with cryptocurrency wallet addresses. Web scraping tools, such as BeautifulSoup and Selenium were employed to efficiently collect information from various online sources. These tools facilitated the extraction of data related to scam addresses, communication channels, fraudulent activities, social media handles/account names, Uniform Resource Locator (URL) and posts. There are various options for writing scraping scripts, including different libraries, modules, and programming languages. For example, Python libraries such as BeautifulSoup and Selenium are commonly used for web scraping due to their powerful capabilities and flexibility (Udofia, 2024). Choosing the appropriate tools and methods depends on the specific requirements of the data being collected and the platforms being scraped.

Both manual and automated scraping methods were employed to collect relevant data from various platforms. Manual scraping involves traditional searching and saving of information manually, which leaves it prone to human error and inconsistencies in data formatting. Although this method may be time consuming, it can be effective when automated solutions are not feasible. However, writing automated scraping scripts, despite an initial learning curve, offers a more efficient and consistent approach. Once this curve is overcome, it significantly reduces the time and effort required for data collection.

Different automated scraping techniques were employed, depending on the platform being investigated. HTML parsing, accomplished with the use of BeautifulSoup, this approach works well for platforms with static content, where HTML structure remains consistent. On the other hand, Selenium was used to handle dynamic content that only loads after interacting with the website (Udofia, 2024). These different approaches show that scraping is highly adaptable, allowing investigators to choose the most efficient approach depending on the platform's structure and the type of data required.

All the scraped data was stored using a JSON format due to its semi-structured and nested design, which facilitates the organisation and retrieval of complex, hierarchical information. Additionally, JSON is human-readable, providing a significant advantage in cryptocurrency scam investigations. However, there are instances where the JSON structure becomes more complex, such as with double-nested JSON, where JSON objects contain other objects. This added complexity can make the data harder to read and interpret manually.

However, this challenge can be mitigated by using tools like JSON readers or writing custom Python scripts to parse and organise the data effectively (Novriansyah, 2024).

The choice of which platforms to scrape is critical and based on the likelihood of obtaining useful data pertinent to the investigation. The following platforms were selected for their unique attributes and relevance to crypto scam investigations:

- **BitcoinWhosWho** is a blockchain analytics platform that offers tools to track and analyse Bitcoin transactions, addresses, and wallets. This platform enables users to look up Bitcoin addresses, verify reported scams, and monitor wallet balances. Its features, such as transaction alerts and the ability to search for associated websites or owner profiles, make it a valuable resource for promoting transparency and accountability with the Bitcoin ecosystem. This platform was chosen for its capacity to help users make informed decisions and avoid potential scams. The JSON file produced from BitcoinWhosWho offers a structured format (***Figure 2***).
- **Chainabuse** provides detailed scam reports that include wallet addresses and descriptions of fraudulent activities. The platform is utilised for its comprehensive sections such as Case Details, Scammer Information, and Description which are essential for thorough investigations. In addition, Chainabuse's community features allow for collaboration and sharing of insights among researchers, making it an invaluable tool for this study. The JSON file produced from Chainabuse has differently structured format (**Figure 3**) than shown in ***Figure 2***, and this is due to the platforms' differences.
- **X**, formerly known as Twitter, serves as a vast repository of raw data, including user profiles and user-generated content such as tweets, replies, and messages. This platform is frequently used by victims to compile informal scam reports and by scammers to post misleading content or false promises. Such instances can be crucial for investigations, as they provide pivot points for analysing behavioural patterns, detecting changes in scammers' activities, and uncovering connections to other suspicious accounts. Due to the policy changes enforced by Elon Musk (the owner), there are various consequences that may be enforced if a user scrapes data from X (Calma, 2023) (Gotfredsen, 2023). A custom script was developed to collect data from X – utilising BeautifulSoup and Selenium. This approach faced significant challenges due to X's robust bot detection mechanisms, which include CAPTCHAs and strict rate limiting. These protections make automated scraping difficult and can lead to Internet Protocol (IP) bans if not managed appropriately. A workaround to this problem is making use of X's Application Programming Interface (API). However, the free version of the API does not provide the necessary endpoints for tweet extraction, making it essential to use the paid API. The paid API allows for more efficient data collection and may help avoid the pitfalls associated with attempting to scrape X, such as IP bans and account suspensions.

4. Proposed Approach and Results

The proposed process focuses on the practical application of the research methodology and the steps to take once sufficient information has been gathered. This process is designed to ensure the effective use of collected data and to facilitate potential legal actions against perpetrators.

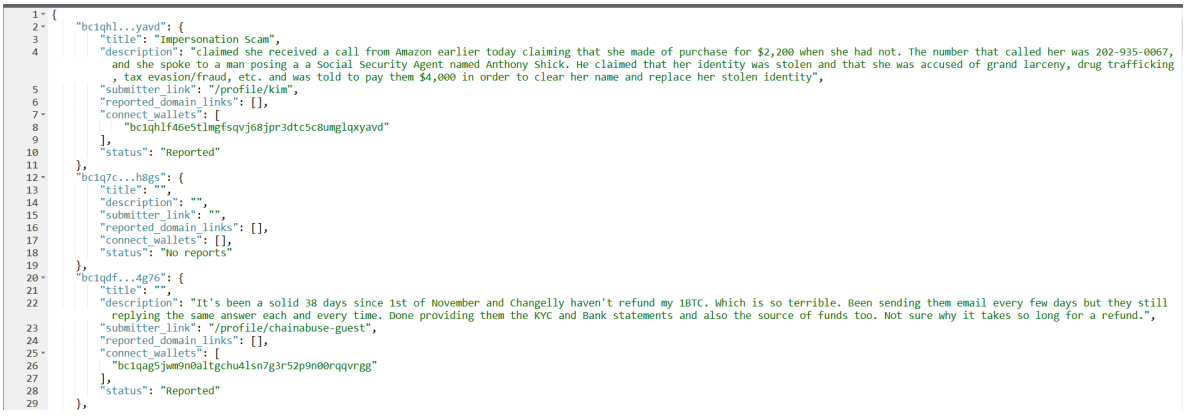
- Scraping Bitcoin Addresses: Conduct scraping to gather data on each Bitcoin address associated with reported scams. This step aims to uncover links and connections to other entities or activities related to the scams.
- Creating Pivot Points: Use the collected data to establish pivot points for the investigation. These pivot points act as crucial starting points for further analysis, guiding subsequent exploration and investigation.

```
{
    "bc1qhl...yavd": {
        "scam": "No scam found"
    },
    "bc1q7c...h8gs": {
        "scam": "No scam found"
    },
    "bc1qdf...c4g76": {
        "scam": "No scam found"
    },
    "bc1qh3...gp8v": {
        "scam": "No scam found"
    },
    "bc1q24...c3c4": {
        "scam": "No scam found"
    },
    "bc1q22...824s": {
        "scam": "No scam found"
    },
    "bc1qqp...zqyy": {
        "scam": "No scam found"
    },
    "171AvU...wnkY": {
        "scam": "No scam found"
    },
    "bc1qa9...xft5": {
        "scam": "No scam found"
    },
    "bc1qm3...7s3h": {
        "scam": [
            {
                "scam_name": "youtube scam",
                "url": "https://www.youtube.com/watch?v=DoUpLY5idis",
                "date": "Nov 18th, 21"
            },
            {
                "scam_name": "youtube scam",
                "url": "https://www.youtube.com/watch?v=DoUpLY5idis",
                "date": "Nov 18th, 21"
            },
```


Figure 2: BitcoinWhosWho JSON data: reported scams and connected websites

Figure 2 presents the JSON data from BitcoinWhosWho. The data structure includes the keys- Unique addresses (e.g., "bc1qhl...yavd") likely linked to cryptocurrency transactions, with additional attributes listed below:

- **"scam"**: Indicates if a scam is detected for the address.
- String value: "No scam found" means no scam detected.
- Array of objects: Scam detected; each object contains:
    - **"scam_name"**: Scam type (e.g., "youtube scam")
    - **"url"**: URL with scam details
    - **"date"**: Scam report date



```
{
    "bc1qhl...yavd": {
        "title": "Impersonation Scam",
        "description": "claimed she received a call from Amazon earlier today claiming that she made of purchase for $2,200 when she had not. The number that called her was 202-935-0067, and she spoke to a man posing a a Social Security Agent named Anthony Shick. He claimed that her identity was stolen and that she was accused of grand larceny, drug trafficking , tax evasion/fraud, etc. and was told to pay them $4,000 in order to clear her name and replace her stolen identity",
        "submitter_link": "/profile/kim",
        "reported_domain_links": [],
        "connect_wallets": [
            "bc1qhlf46e5tlmgfsqvj68jpr3dtc5c8umglqxyavd"
        ],
        "status": "Reported"
    },
    "bc1q7c...h8gs": {
        "title": "",
        "description": "",
        "submitter_link": "",
        "reported_domain_links": [],
        "connect_wallets": [],
        "status": "No reports"
    },
    "bc1qdf...4g76": {
        "title": "",
        "description": "It's been a solid 38 days since 1st of November and Changelly haven't refund my 1BTC. Which is so terrible. Been sending them email every few days but they still replying the same answer each and every time. Done providing them the KYC and Bank statements and also the source of funds too. Not sure why it takes so long for a refund.",
        "submitter_link": "/profile/chainabuse-guest",
        "reported_domain_links": [],
        "connect_wallets": [
            "bc1qag5jwm9n0altgchu4lsn7g3r52p9n00rqqvrgg"
        ],
        "status": "Reported"
    },
```


Figure 3: ChainAbuse JSON data: reported scams and submitted profiles

Figure 3 presents the JSON dataset from Chainabuse. The data structure includes the keys - Unique addresses (e.g., "bc1qhl...yavd") likely linked to cryptocurrency transactions, with additional attributes listed below:

- **"title"**: Name/type of scam, or empty if no scam reported.
- **"description"**: Details of the scam, or empty if no scam reported.

- **"submitter_link"**: Link to the submitter's profile (e.g., "/profile/kim").
- **"reported_domain_links"**: Array of domain links related to the scam, empty if none.
- **"connected_wallets"**: Array of connected wallet addresses, empty if none.
- **"status"**: Scam report status:
    - **"Reported"**: Scam reported.
    - **"No Reports"**: No reports submitted.

The JSON files from BitcoinWhosWho (***Figure 2***) and ChainAbuse (***Figure 3***) demonstrate how data from these platforms can reveal connections between wallet addresses, reported scams, and other entities. The inclusion of submitter profiles and associated websites further aids investigators in identifying patterns and links between fraudulent activities. Notice ***Figure 2***, which contains two YouTube links found during the scraping process. While this falls outside the scope of the paper, further off-chain analysis could advance the investigation, potentially revealing the identity behind the account that posted the video. A comparison of the data in ***Figure 2*** and ***Figure 3*** emphasises the importance of using multiple data sources – one source may lack posts mentioning the wallet address, while another may include them, as seen with the address bc1qhl…yavd.

This research endeavour centred on the utilisation of web scraping and represents significant additions to a broader initiative aimed at establishing a comprehensive methodology for investigating cryptocurrency scams.

5. Challenges and Limitations

One of the primary challenges that may appear in cryptocurrency scam investigations is that scraped data can quickly become outdated. This issue is not merely about the obsolescence of data captured at a specific moment, it also encompasses the risk of missing new scams and evolving tactics that may arise if the investigation is not continuously updated. Additionally, platforms frequently change their formats, which can render existing scraping scripts ineffective, necessitating constant updates or rewrites to maintain functionality (Jayan, 2024). This results in the ongoing task of script maintenance, adding another layer of complexity to the process.

Handling the large volumes of data generated by scraping presents another significant challenge. While scraping can streamline data collection, it does not simplify the subsequent steps of data processing. After collection, efforts are required to clean the data to correct inaccuracies and analyse it to extract meaningful insights. Moreover, effective data visualisation is necessary to present the findings clearly.

These tasks are both manual and resource-intensive but are crucial for identifying potential pivot points and drawing actionable conclusions from the collected data. In addition, many platforms such as X and Facebook, actively discourage scraping to protect user privacy and maintain platform integrity (Jayan, 2024) (Ok, 2024). These platforms employ countermeasures such as rate limits, and IP bans which can significantly hinder automated data collection efforts. To address these challenges, the use of APIs should be considered when available. APIs offer a legal and structured way to access data, often with greater reliability and adherence to platform terms of service. Utilising APIs can mitigate some of the issues associated with scraping and help ensure a more stable and compliant data collection process.

Further research needs to be conducted, including additional data sources such as Telegram, which has become popular among cryptocurrency enthusiasts – scammers have taken advantage of this, creating crypto-related scams (D'Andrea, 2024). Expanding data sources could uncover more pivot points and potentially lead to deeper insights.

6. Conclusion

This research highlighted several crucial aspects of investigating cryptocurrency scams and the application of web scraping within this context. The study highlights that digital trust is a pivotal factor in the adoption and credibility of cryptocurrencies, yet it is often undermined by fraudulent activities that exploit the platform's complexities. Key outcomes of this research include the effectiveness of web scraping, which has been demonstrated as a valuable tool for gathering relevant data from social media and online platforms. By employing web scraping techniques, investigators can uncover critical leads and context related to cryptocurrency scams. Furthermore, the integration of scraped data with traditional OSINT methods may enhance the depth and accuracy of investigations. While this paper primarily focuses on the web scraping component, it also highlights the added value of combining these techniques with broader OSINT approaches. This combination offers a more comprehensive investigation framework that may improve overall investigative outcomes. The practical implications of these findings suggest that adopting a structured methodology that combines web scraping with blockchain analysis can significantly improve the detection and analysis of

cryptocurrency scams. By using these advanced tools and techniques, investigators can take a more proactive approach to identifying fraudulent activities and, ultimately, help restore digital trust in the cryptocurrency ecosystem.

Moving forward, the research advocates for the continued development and integration of advanced investigative methodologies to better combat crypto scams. By refining web scraping techniques and leveraging them alongside blockchain analysis, stakeholders can enhance their ability to uncover fraudulent schemes and foster a safer cryptocurrency environment. The insights gained from this study provide a foundation for future research and practical applications aimed at strengthening the integrity and trustworthiness of the cryptocurrency ecosystem. With the wealth of scraped data that often-included various links and URLs, future research will look to advance the investigation further, potentially uncovering more information that could lead to revealing the identities of those behind these scams.

References

Bartoletti, M. et al., 2021. Cryptocurrency Scams: Analysis and Perspectives. *IEEE,* Volume 9, pp. 148353-148373.

Botha, J. G., Botha-Badenhorst, D. P. & Leenen, L., 2023. An analysis of crypto scams during the Covid-19 pandemic: 2020-2022. *European Conference on Cyber Warfare and Security,* Volume 18, pp. 74-85.

Botha, J. & Leenen, L., 2024. Cryptocurrency-Crime Investigation: Fraudulent use of Bitcoin in a Divorce Case. *International Conference on Cyber Warfare and Security,* Volume 19, pp. 34-42.

Botha, J., Pederson, T. & Leenen, L., 2023. An Analysis of the MTI Crypto Investment Scam: User Case. *European Conference on Cyber Warfare and Security,* pp. 89-99.

Botha, J., Singh, K. & Leenen, L., 2025. A Proposed Bitcoin Blockchain Investigation Methodology: Based on a Case Study Approach. *Journal of Information Warfare,* 24(1), pp. 1-18.

Calma, J., 2023. *Twitter just closed the book on academic research.* [Online] Available at: https://theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research [Accessed 19 November 2024].

Crypto Defenders Alliance, 2024. *The Blacklist: Track and Recover.* [Online] Available at: https://cryptodefendersalliance.com/blacklist [Accessed 12 February 2025].

D'Andrea, A., 2024. *Common Telegram Scams To Be Aware Of.* [Online] Available at: https://www.keepersecurity.com/blog/2024/09/20/common-telegram-scams-to-be-aware-of/#h-7-telegram-cryptocurrency-scams [Accessed 24 February 2025].

Department of Financial Protection and Innovation, 2025. *Crypto Scam Tracker.* [Online] Available at: https://dfpi.ca.gov/consumers/crypto/crypto-scam-tracker/#:~:text=Imposter%20Scams%20%E2%80%93%20Scammer%20impersonates%20a,to%20steal%20the%20user's%20assets).

Gertenbach, W., Botha, J. & Leenen, L., 2024. A Proposed High-Level Methodology on how OSINT is Applied in Blockchain Investigations. *Cyber Warfare and Security,* Volume 19, pp. 75-83.

Gong, J., 2024. *Speed Up Your Python Web Scraping: Techniques & Tools.* [Online] Available at: https://www.bardeen.ai/answers/how-to-web-scrape-faster

Gotfredsen, S. G., 2023. *Q&A: What happened to academic research on Twitter?.* [Online] Available at: https://www.cjr.org/tow_center/qa-what-happened-to-academic-research-on-twitter.php [Accessed 19 November 2024].

Hayes, A., 2024. *CoinJoin: What It Is, How It Works, and Privacy Considerations.* [Online] Available at: https://www.investopedia.com/terms/c/coinjoin.asp [Accessed 12 February 2025].

Jayan, J., 2024. *10 Web Scraping Challenges and Best Practices.* [Online] Available at: https://promptcloud.com/blog/web-scraping-challenges/

Krotov, V., Johnson, L. & Silva, L., 2020. *Tutorial: Legality and Ethics of Web Scraping. ,* Murray: Communications of the Association for Information Systems.

Levy, A., 2023. *How to Spot a Pump-and-Dump Cryptocurrency Scam.* [Online] Available at: https://www.fool.com/investing/stock-market/market-sectors/financials/cryptocurrency-stocks/how-to-spot-crypto-scam/#:~:text=Pump%2Dand%2Ddump%20scams%20have,positive%20news%20about%20the%20asset. [Accessed 20 March 2024].

Novriansyah, N., 2024. *Parsing and Handling Double-Nested JSON Data.* [Online] Available at: https://supersimplearn.medium.com/parsing-and-handling-double-nested-json-data-823ff6808624

Ok, J., 2024. *What is Facebook Scraping? What You Need to Know.* [Online] Available at: https://multilogin.com/blog/what-is-facebook-scraping/

Popov, C., 2023. *UK Finance's Survey: Over 70% of young adults targeted by impersonation scams.* [Online] Available at: http://bitdefender.com/en-us/blog/hotforsecurity/uk-finances-survey-over-70-of-young-adults-targeted-by-impersonation-scams [Accessed 23 January 2024].

Singrodia, V., Mitra, A. & Paul, S., 2019. *A Review on Web Scrapping and its Applications,* Coimbatore, INDIA: International Conference on Computer Communication and Informatics.

Social Links, 2023. *OSINT and Social Media Investigations: The Perfect Combination.* [Online] Available at: https://blog.sociallinks.io/osint-and-social-media-investigations-the-perfect-combination/[Accessed 18 November 2024].

TechMindXperts, 2023. *Social Media OSINT: A Comprehensive Guide to Gathering Intelligence from Social Media Platforms.* [Online] Available at: https://osintteam.blog/social-media-osint-a-comprehensive-guide-to-gathering-intelligence-from-social-media-platforms-b5dbb8d83f14 [Accessed 18 November 2024].

Tradivest, 2024. *A beginner's guide to on-chain analysis.* [Online] Available at: https://medium.com/coinmonks/a-beginners-guide-to-on-chain-analysis-1f2689efd9aa [Accessed 25 August 2024].

Udofia, E., 2024. *WebScrapping: BeautifulSoup or Selenium?.* [Online] Available at: https://medium.com/@udofiaetietop/webscrapping-beautifulsoup-or-selenium-3467edb3c0d9

UEEx, 2024. *How to Do On-Chain Analysis for Smart Crypto Investors.* [Online] Available at: https://blog.ueex.com/on-chain-analysis/