TRACKING MALICIOUS TRANSACTIONS IN CRYPTOCURRENCIES

Bhavish Dhanda
Supervisor: Dr Hassan Asghar
Co-Supervisor: Dr Benjamin Zhao

April 24, 2023

MACQUARIE
University
SYDNEY·AUSTRALIA

Department of Computing and Engineering
Macquarie University
Australia

Contents

List of Figures

Listings

ACKNOWLEDGEMENTS

I would like to acknowledge and give my warmest thanks to my supervisor Dr Hassan Asghar and co-supervisor Dr Benjamin Zhao who made this work possible. Their guidance and advice carried me through all the stages of writing my project.

STATEMENT OF CANDIDATE

I hereby declare that the work, which is being presented in the Thesis, entitled "Tracking malicious transaction in Cryptocurrencies", in partial fulfillment of the requirement for the award of the degree of Bachelor of Software Engineering (Honours) in the department of Engineering, Macquarie University. This thesis is an original piece of research work under the guidance of Dr Hassan Asghar and Dr Benjamin Zhao. The matter embodied in this thesis has not been submitted for the award of any other degree of any other academic institution.

Abstract

Cryptocurrencies, the new digital currency of the twenty first century have been becoming increasingly popular in the recent years. More and more people have started getting their hands on these currencies plus with the introduction of NFTs(Non-Fungible Tokens)[1] people are now slowly and gradually switching to digital assets like digital tokens, arts, property etc. As this becomes more and more popular its becoming hard to keep track of who is using which currency, and also with the concept of decentralization they introduce, it makes it even harder to know if someone ever dealt in crypto or not. With all of these features, arises a big issue of illegal use of crypto currencies because anyone can access it without providing any personally identifiable information and send or receive money from anywhere. Here, in this paper we are trying to find and link these so-called digital transactions to some sort of identification factor in the real world with the end goal to crack down on illegal activities in crypto.

1 Introduction

Cryptocurrencies are the new popular form of currencies which are becoming popular day by day. A few countries including but not limited to El Salvador, Central African Republic have already made Bitcoin as a legal form of tender[2]. This means that people who have bought bitcoin previously or are planning to can use it for day to day purchases. The private nature of these form of digital currencies masks a persons real identity giving them pseudo-identities to be used on the digital currency's network.

All the modern cryptocurrencies manage a public ledger which contains all the information a typical ledger would include with the striking difference of keeping user's pseudo-identities instead of real identities. This means that anyone can open as many accounts or hold as many pseudo-identities or addresses as they wish and make transactions using them. All of these features and limitations although helps users to get set up on these decentralized services quick and easily without any legal lengthy registration processes, it does leave the ground open for people to do anything they want since there is not proper regulation/monitoring in the system. Due to this, cryptocurrencies have become a very big platform for money laundering activities which needs to be regulated. **Our goal in this paper to classify and trace malicious transactions in the world of cryptocurrencies and link the illegal transactions to some personally identifiable real world information**.

Following up on that a lot of people have looked into this earlier such as Fanusie and Robinson[3], they have done a very good job at highlighting how serious of a problem is this illegal activity specifically on bitcoin. They studied 3 years worth of data from the blockchain during the periods 2013 to 2016, and discovered some very concerning patterns. According to their research services such as Silk Road and Alpha-Bay were the source of almost all of the illegal bitcoins laundered. Although most of these services were shut down sooner or later but the concerning fact was that there is almost every time a successor that arises which makes it hard to control these sorts of activities on the blockchain.

They also observed some geographical patterns to see how different regions are the source of these illegal funds and discovered the the majority of these activities were done on hidden networks like TOR which is expected of people who would want to hide their identities or make it next to impossible to identify them, but after these networks, the next biggest source was the conversion services originated from Europe followed by North America and then Asia. Another important figure that was highlighted in this paper was how dominant Bitcoin Exchanges and Mixing services are when it comes to money laundering activities. They seems to be carrying the majority of illegal funds yearly since 2013. This percentage seems to have dropped down yearly on all the individual services but the primary reason for that was the rising popularity of Bitcoin within the general public. This same observation was noted in terms of geographical distributions as popularity of Bitcoin increased over time.

Similarly a lot of other people have looked into this problem of malicious transactions in cryptocurrencies to raise awareness amongst the general public.

This had made some progress as a lot of countries are already looking at ways to prevent this from happening and setting up teams to stop this. The United States for example already have made some progress to crack down on these users such as the arrest of Roman Sterlingov[4] in 2021 who operated an illegal mixing service. Along with countries trying to regulate crypto, consumer law advocates are raising their voice demanding crypto assets to be regulated within financial laws[5]. So, all of this with the raised consumer awareness has also created an interest of people in the researching community to find ways to regulate these digital currencies which we will explore in the later sections.

In the following sections we aim to address this issue in further detail, starting with an introduction of basic concepts necessary for our research in section 2, where we try to explain what are cryptocurrencies, how do they work, followed by more advanced concepts such as mixers, verification protocols and privacy coins with detailed explanation on specific currencies. After that we explore in the following section the work people have already done to identify their learning's and potentially use them for our work, and finally we discuss what we plan to do in this research to try and solve this issue of unregulated malicious activity in cryptocurrencies.

2 Background

2.1 Crypto-Currencies

Cryptocurrencies are a digital form of currency which are controlled by a public ledger. On these public block-chains since the ledger is public everything is visible to anyone on the network. The main striking point about theses ledgers is that they collect everything except the user's real identity instead they assign a user a pseudo-identity which is basically a wallet address that exists on a particular ledger. Since this entire concept works on open source what it means is that anyone can spin up a node on their system and based on the blockchain being used, the users will get different benefits without revealing their identity. Below we will discuss some basic concepts related to cryptocurrencies and how they play an important role in developing the modern digital currencies as we have them.

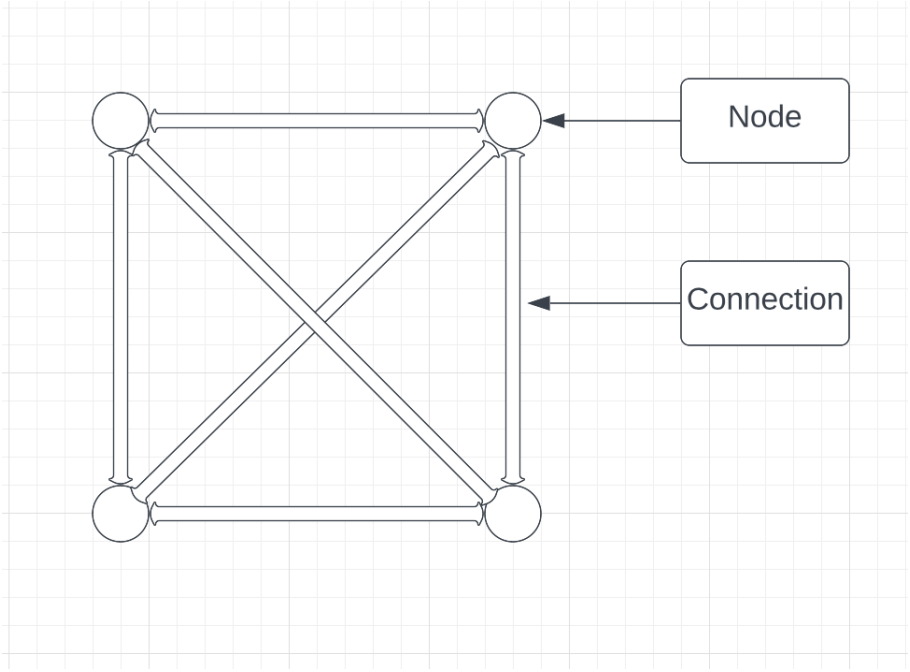

Figure 1: A very basic P2P Network

2.1.1 P2P networks

Cryptocurrencies are based on P2P networks to enable decentralization which is explained in the next section. But the point of a p2p network is to enable all the nodes to interact freely with each other without the supervision or involvement of any other party.

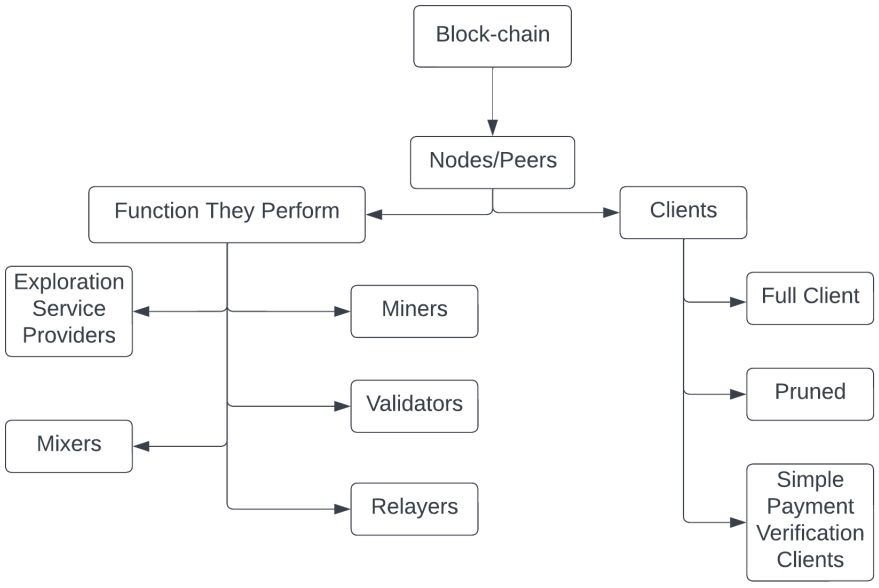

Figure 2: Different Types of nodes in typical public Blockchain

Delgado-Segura[6] and a few other authors have very nicely describe the how a p2p network builds up the entire Bitcoin blockchain. Starting with peers there are different types of peers/nodes.

1. Full which store the entire blockchain thus carrying around 406.82GB[7] of data as of 21 May 2022.
2. Pruned which only store the blockchain data for the last 2 days being around a couple of gigabytes.
3. Finally there are Simplified payment verification clients containing just the blockchain headers which is usually a few megabytes.

Then the next classification of peers can be the function they perform which can be either miners which create blocks which is the computationally most expensive task, some perform validation and relaying of information such as new blocks, addresses and transactions as they receive, some provide exploration service [8] and finally some perform mixing services, the legality of which is still under question as we discuss below. This structure of peers[1][2] varies from currency to currency depending on their revenue model usually but sometime other factors such as performance as well.

2.1.2 Decentralized Finance

Cryptocurrencies forms the so called modern day decentralized financial networks also know as DeFi. DeFi defines what modern day banking and currencies would look like and how would they perform and span across the globe. The basic ideology behind DeFi is that all the power and regulations must not be with one person or entity instead it should be publicly controlled and visible. So to achieve this what cryptocurrencies do is maintain a pubic ledger unlike a secure, secret private ledger handled by banks and various other financial institutions. This public ledger gives the mass population access to everything that is going on with the particular cryptocurrency like all the transactions, addresses, funds held by the addresses etc. The exhaustive list of these feature maintained by these currencies and how they maintained varies amongst all cryptocurrencies with the important factor being all public.

Since the entire network in DeFi works on a public blockchain, everything is controlled by the various nodes on the network, all the transactions, mining rewards, fees is also maintained by the ledger. This means that when a transaction occurs on the blockchain instead of a person approving it, all the nodes on the ledger work to sign it off by verifying multiple factors such the inputs and outputs, preventing double spending etc.

2.2 Verification Protocols

Different types of cryptocurrencies use different verification methods to verify the transactions they receive before it is accepted into the blockchain. One of the earliest one out there is Zero Knowledge proof and as cryptocurrencies evolved there came newer methods of verification ZK-snarks which is an extension of ZNP used by many of the newer privacy coins. In the following section we go through breifly about ZNP and ZK-Snarks as they might be helpful in our research moving forward to understand how different crypto-currencies work.

2.2.1 Zero Knowledge Proofs(ZNP)

Zero knowledge proofs are an identity management proof of some statement used to prove an identity or ownership in case of cryptocurrency to the public or the verifier. The fascinating part about ZNP is that one party can prove their identity/ownership to the second one without revealing anything more than their identity.

2.2.2 Zero Knowledge SNARKS(ZK-Snarks)

ZK-SNARKS is short for Zero Knowledge succinct non-interactive arguments of knowledge. Reitwiessner, Christian in their paper[9], explain ZK-Snarks by breaking it down into different elements in the case where the first party the so called prover has to convince the second one the verifier about a particular statement just by exchanging messages.

1. Succinct: This means that the size of messages is very small as compared to the actual computation needed to break down the encryption

2. Non-interactive: This means the interaction between the 2 parties is very limited, Zk-Snarks usually just have a setup phase and a message transfer for this verification process.

3. Arguments: This refers to the fact that there isn't computation power with the prover to actually break down the encryption. One can break any encryption using brute force but the harder it is that more time it takes. What the authors are trying to convey here is that the encryption is so hard to break that it will take an unimaginable amount of time to crack it.

   Out of all of these zero knowledge refers to the fact that in this entire verification process, all the verifier knows is that the statement is valid.

2.3 Bitcoin

2.3.1 Basic Overview

Bitcoin as of 7 May 2022 is the most valuable cryptocurrency[10] and one of the oldest as well. Bitcoin was originally built as a step towards decentralized currency and as we can see over the years it was highly supported by people as well. Bitcoin keeps everything decentralized by maintaining a public ledger and making the nodes make decisions such as approving transactions, giving out rewards rather then giving the entire control to a singe entity or group of people. This gives everyone visibility of what is happening with the value of the currency as well as how can they benefit from it. Bitcoin provides users with a sense of pseudo-anonymity by issuing them with wallet address without taking in any personally identifiable information. We call this identity pseudo-identity[11] is because although it seems like the user has not given his personal information but just in case something gets leaked to which the someone can link that information to the person's wallet address, he does not remain anonymous anymore.

Transactions in Bitcoin are stored in blocks to keep a record of all the transaction pubic so a user can just query a particular block to get its details which includes information such as when it was created, what all transactions are in the block, the hash of the block etc. The most important purpose of this is that this prevents double spending so every transaction is crosschecked with the ledger to ensure a user is not trying to use his tokens more than once and also to maintain the legitimacy of Blockchain.

2.3.2 Mining Bitcoin

Bitcoin mining uses a proof-of-work system to verify blocks before adding them to the blockchain. This is a distributed process where numerous nodes throughout the world try to compute and find a hash value which is less than a target specified on the blockchain. This way whosoever gets to the smallest hash earliest wins and the node is rewarded with the mining reward. After this the new block is appended to the blockchain and everyone else discards their work to try and get to the next smallest value quickest. This mining reward decreases as the number of blocks mined increased and so does the difficulty of mining gets harder i.e. the specified hash values keeps on becoming smaller. This way it is decided on the Bitcoin blockchain that the last block would be mined in the year 2140 and that is when all the Bitcoin would have been possibly mined and only the existing Bitcoin would remain in circulation.

2.4 Mixers

Mixers are an address shielding services built upon blockchain which work to protect a users identity by transferring their crypto to a different known address for a small fee. These were developed to build upon the idea to make cryptocurrencies more private than they ever could be specially considering the earlier ones like Bitcoin.

When privacy became a common concern for a lot of people using crypto, Bitcoin and other similar currencies started advising people to use a new address every time for a new transaction but not everyone was doing that. Doing this would mean, a user would have to issue a new address for every transaction he does, transfer funds into it which would always be a multiple step process so because of this very few people were practicing this. So, to tackle this problem mixers were developed.

The contradicting fact about mixers is that they are known to be a big carrier of illegal funds since their inception, still they are deemed legal. The important point to be noted here is even based on the most recent discoveries and investigations by authorities, after they crack down on a mixer being used for movement of illegal funds they are almost all the times another successor to the service similar to what happens in dark web marketplaces.

This concept of mixers also inspired the development of newer currencies especially the privacy coins as they are built to enforce privacy where some of them help users to have a different wallet address and a separate transaction address and some have mixing services inbuilt.

Researchers[12] from the University of New South Wales evaluate the impact of cryptocurrency tumblers, as their role of mixing tainted/identifiable funds with an untainted pool so as to eliminate traceability. As much as they solve the purpose cryptocurrencies were built for, which is to protect the privacy, freedom and secrecy of individuals, their misuse can help people to perform anti-social or harmful actions against others. They highlight the reason why a lot of people/authorities are pushing towards making tumbler services illegal because of their nature to mix potentially legal funds with illegal ones thereby supporting money laundering activities. This nature of tumblers has made it easier for people to perform illegal activities and made the jobs of law enforcement and regulation authorities much harder. But even if these services become illegal it is very difficult to trace them behind dark-nets such as Tor plus no one can until now stop the development of new privacy coins which as highlighted below are made on top of basic cryptocurrencies like Bitcoin to include mixing functions or services by default.

2.5 Privacy Coins

Privacy coins are the next big and emerging concept for privacy in cryptocurrency. The main reason behind these privacy coins was to enforce the privacy mechanisms which were optional in older currencies. For example to randomize an address on the bitcoin network the user would have to go through a mixing service or something of that sort and also pay a fee to the service. Privacy coins eliminate this manual step and most of them have these feature without any manual action from the users. Most of them seem to work on the basis that if you have a public transaction address and a private wallet address, it is maintaining your transaction identity on the blockchain and keep it away from de-anonymization by adding another layer. This way on the public transaction address is stored on the public ledger and the private address is only known to the user.

Now that we have a good idea of what privacy coins are and what purpose do they solve, it would be best if we discuss a couple of them.

2.5.1 Zero-coin

Narayan[11] and the rest of the authors explain how zerocoin is a new ZNK based cryptocoin made primarily to enhance the privacy of the users using it. What zerocoin seems to do is working 2 different types of token a base-coin and a zerocoin, where a base-coin is something that you would actually spend whereas the zerocoin is something that you would use to prove that you own a base-coin and these can be exchanged for each other. As mentioned earlier zerocoin uses ZNPs to prove and verify its transaction. What this means in the context of zerocoin is that whenever someone is exchanging Zero-Coin for a base-coin or vice-versa they would essentially not describe what particular target they own rather it would be more of saying that there is a big pool of target coins and the person owns one of them not knowing which one thereby breaking the link-ability between the zerocoin which is associated to a person and the base-coin which is actually used to transact.

2.5.2 Zero-cash

As highlighted by Narayan[11] and the authors Zero-cash takes privacy in cryptocurrencies to another level. Zero-cash based off zerocoin, eliminates the distinction between a Zero-cash coin and base coin. They use ZK-Snarks which is an enhancement of ZKPs. Another way Zero-cash aims to reduce traceability is by controlling what it stores on the public blockchain as the main distinction being transaction amounts. In Zero-cash only the sender, the receiver and possibly the miner if there is a miners fee know about the amounts in the transaction they were a part of, in the ledger it would appear just as if a transaction happened between two parties with no public records of the amounts associated.The biggest catch with Zero-cash as opposed to other currencies is the public parameters used to verify the token and the transaction as in this case this is usually a

gigabyte long number making it computationally extremely hard to set up the ZNP verification system.

2.6 Dark Web Markets

Dark web markets are basically online shopping platforms like Amazon etc but used for illegal activities. These are a very old form of illegal markets which have managed to stay in circulation even after various blow backs by the authorities by having some form of successor to them. The main characteristic of these markets is their privacy features to shadow their IP address and location from the rest of the network thereby protecting the anonymity of users from network layer attacks.

As much as as dark web markets are known for their use of illegal activities the launch of cryptocurrencies has only accelerated this cause. In the last few years Dark web markets have become a big source of illegal funds in cryptocurrencies and similar activites as people are using these service to protect their privacy as much as they can as we can see in [13] and [14]. Because of this nature of dark web markets it is very important for us to study about them as their use for illegal reasons is only going to increase over time as their popularity and accessibilty increases.

As highlighted by Naoki and Yaichi in their paper [15] Silk Road and AlphaBay even though they were only active for a bit over 2 years and less than 3 years respectively had the highest cumulative revenue over the years both being around 200 million and 300 million US dollars in their short active periods. On top of this it was found that in the periods of January 2017 and March 2018, Bitcoin accounted for 99.8 percent of all transactions in the dark web markets and 80 percent of those were malicious transaction i.e. were for money laundering, scams, ponzi schemes etc. Below we will go through both of them to understand how they operated and influenced the modern dark web marketplaces.

2.6.1 Silk Road

Silk Road was the first known dark-net marketplace founded by a group of people in 2011 only to a select number of people. Silk Road became increasingly popular as it was established in terms of the number of users and so did its revenue. This increased the interest of authorities towards this, and eventually in October 2013 this was shut down but this did not stop the operation like this instead it inspired more people to get involved in this. Because this growing interest shortly after its shutdown next month only in November 2013, there came up Silk Road 2.0.

2.6.2 AlphaBay

AlphaBay was founded by Christin the same person who founded Silk Road in 2011 was the second biggest dark net pool in its active period based on total revenue[15]. Within this total revenue it was estimated that around 450 million dollars accounted for 4 million transactions just using Bitcoin addresses, and after the introduction of a mixer service 340,000 wallet addresses were estimated to be for Bitcoin.

3 Related Work

In this section we will go through some related work that has already been done in this area to track down on malicious transactions in Cryptocurrencies. As we will read in the following sections a lot of work has been done to identify the illegal nodes inside the various cryptocurrency networks but not much has been done to actually map the malicious transactions into real world identities/entities.

Our main area of focus to study related works is to realise what has already been done in this area to crack down on malicious usage of cryptocurrencies, we first start by looking at various de-anonymization attempts made by people on various crypto-currencies, then moving on to ways in which people have been working to introduce anti money laundering techniques in crypto-currencies and finally we see how can we break through mixers as a large share of illegal funds have been known to pass through them[3]. The main point we need to note here is the ways people have been successful in these attempts so that we can base our research on their learnings and focus on the missing pieces.

3.1 De-anonymization Methods

Koshi, Phillip[16] along with the other authors exploit the various relay patterns that Bitcoin exhibits to identify unusual transactions.They developed a custom client called Coin-Seer specifically for data collection, and collected around five months worth of data by maintaining connections to on an average of 2678 peers simultaneously in an hour.

These guys identified 3 different relay patterns to identify what nodes might be undertaking illegal activities. The main was a single re-layer transactions: This is where they exploit the gossip protocol and the identification factor is that the node which first originated the transaction is most likely the source or the receiver of the funds.

They used a 5 step process to identify malicious IP address which started from filtering their data, hypothesizing the owner, creating Bitcoin address to IP address pairing and identifying the ownership pairings.

Looking at their approach they seem to have only identified the most incautious users. They also mention it in the end and also after going through the approach one can determine that if a user uses a third-party service such as a mixer or some sort of eWallet, their approach would fail. Their approach heavily relies on a node being able to access the IP address of a re-layer but if a re-layer is behind a dark-net like TOR or if he is using dynamic IP address or any sort of IP shadowing technique, this would fail. Considering the devices in the current times almost all Apple devices come with iCloud Private Relay [17] which would invalidate their approach from the start, and this is without the user using any special techniques to hide his IP.

Narayan[11] and other authors of the book point out in chapter 6 about various methods of de-anonymization of Bitcoin. Their first approach was to actually transact with the individual identities and they do point out that a

lot of early research was based on this. This could be a good methodology in the early days of Bitcoin where first it wasn't used as much for illegal activities, second, people weren't as much familiar with the privacy associated with cryptocurrencies. They show after making a substantial number of transactions they were able to point it what address belongs to what type of service in the blockchain. But now as the price of cryptocurrency has sky rocketed, on top of the new privacy enhancing techniques available now like mixers, wallets, Bitcoin ATMs when used cleverly make this option less and less feasible.

This takes us to the concept of network layer de-anonymization where the blockchain is our application layer and the p2p network is the actual physical network. As we go though with this research we observe network layer de-anonymization has been by far the most popular technique used by researchers all over the world to link real identities to Bitcoin addresses. This technique however starts to fail when people start using these currencies behind dark-nets. Tor for example has been a very big contributor towards money laundering in Bitcoin. The issue over here for us is that if someone is behind that kind of network, it becomes impossible for us to get to their IP address.

Xueshuo, Xie[18] and a team of researchers develop an address classifier to classify different addresses on the Bitcoin network into various classes based on their behaviour in terms of transaction activities and potentially try and de-anonymous the users to track down on illicit activities on the blockchain. They have used public APIs available from Blockchain.info to query the public ledger directly get real time transnational, node data in JSON format and convert into vectors based on their number of transactions. They were able to successfully group a collection of addresses in multiple categories like wallets, miners, illegal users and normal consumers using their address clustering model. Although the paper aimed to design an address classifier and de-anonymization users but seems like at the end they just concluded after address clustering but their model seems to be doing a very good job to cluster the address.

Biryukov, Alex and Tikhomirov, Sergei[19] attempt to de-anonymize users based on their network activity on various cryptocurrencies which included privacy coins as well. They also in the prove that anyone can run a full fledged privacy attack on cryptocurrencies without needing extraordinary resources. Alex and Sergei, the authors of this papers pose as adversaries trying to hack into the Bitcoin, Z-Cash and Monero networks. They try to exploit the basic principles of address and transaction propagation on which these cryptocurrencies work. The basic concept of their attack is then when a new node is introduced into the network or there is a new transaction the first one to propagate it throughout is usually the originator or the receiver of the network. They developed a custom client called bc-client primarily for the purpose of data collection. The more popular clients like BT-core are not made for data collection but they are for users willing to run a full node on their machines whereas bc-client was build for data collection to log all the incoming messages, IP addresses, the transaction hashes and also the timestamps of various events.

3.2 Detecting Money Laundering Techniques in Crypto

Following concerning use of crypto for money laundering activities, a lot of researchers started to look around and study for ways to identify the people and addresses involved. Below we go through some of the previous work people have done in this area and would provide valuable lessons for our research moving forward.

Weber, Mark and Domeniconi[20] along with a few more researchers develop Anti Money Laundering techniques for cryptocurrencies. The most common financial institutions like banks use something called KYC to prevent money laundering in their systems however this concept is non-existent when it comes to DeFi but they do track the activities of the user on their network. The authors of this paper try to exploit this feature of decentralized financial services in order to cut down on illegal activities on the networks.

KYC also known also as know your customer is a widely accepted technique used almost all the financial institutions throughout the world. What this means is that whenever a user is looking to use the services of one of these institutions they have to provide their basic personally identifiable information such as name, phone numbers, addresses to keep track of their activity in their institution. This concept of getting users details does not exist for decentralized financial services like cryptocurrencies, instead they work on the concept of maintaining a public ledger. A public ledger is a record of all the transactions that happened throughout the network but what these services do is instead of maintaining actual user's personal data they assign them pseudo-anonymous identities and maintain those instead. So this way even though all the information is available publicly but cannot be tracked down to an individual person easily unless the link between the real and pseudo identity gets leaked somewhere.

The authors use the elliptic data set which according to them is the largest data set on Bitcoin publicly available. This biggest advantage of this data set that it contains labelled transaction data mapped to real identities like exchanges, wallets, miners categorised as legal sources versus illegal ones like ponzi schemes, malware, scams etc. This data set contains distinct features which include the basic data from the ledger like transaction times, inputs, outputs, etc plus aggregated features which includes similar data but going one hop forward and behind.After the data they use the available feature to in to perform supervised learning for binary classification which includes techniques such as Logistic regression, multi-layer perception and random forests. Finally for their experimentation they do a 70 to 30 split for their training and test data with three different approaches Logistic Regression, Random Forest and Multi-layer perceptron to find the most efficient one eventually marking Random Forest as the most efficient one due to the nature of the data.In conclusion we can see that their technique of data classification is very efficient when it comes to identification of illegal nodes plus that combined with the visualization prototype they created called Chronograph can be a very important tool for us when it comes to at least identifying illegal nodes in a network. Also looking the various classification models they used, they did conclude Random Forest as the most

efficient one so we could learn from this apply random forest or it variations on our data directly to achieve best results. One particular thing we need to be cautious about is as they have highlighted sudden event like they are mentioning about a dark net shutdown but we need to keep an eye on events like for our research specially while refining the data as these event cause big changes in the topology of the network and are very likely to drastically decrease performance due to the sudden change in features.

A group of researchers from CSIRO , and a number of renowned universities [21] grouped together to try and work towards finding a way to detect money laundering activities on the Bitcoin network.They ran a Bitcoin core client and extracted data for a couple of years and parsed it into JSON format and extracted specific features such as timestamps, inputs, outputs to create a transaction graph. They then identified some ground truth data they had and classified it into regular users such as consumers like us and illegal. After this was done they did some feature extraction to extract extended features for a transaction such as linked input and output UXTOs and few others to make up-to 14 features each for a transaction. Post extracting the required features they used the data to train the machine learning models and used those for prediction of newer money laundering instances. The way this was achieved was they took 3 money laundering instances and ran the model by leaving out one out of 3 instances to see whether it was able to find the left out one. In the end they were able to get some goods with over 90 percent correct result in some case. The main point we can learn from here is that usually these type of illegal instances/clusters have distinguishing features which are normally missed by the human mind if they going over for example a transaction graph but machine learning algorithms can be a very effective measure for this. The distinct features of these illegal instances such as higher in-degree/out-degree ratio, smaller number of weakly connected components is something worthwhile for us to keep in find to label out these clusters.

3.3 Breaking Through Mixers

In this section we analyze some previous work done to identify the users of mixing services. Mixers are known to carry a significant proportion of illegal funds so it is important that our approach doesn't fail whenever someone uses them.

Wang, Zhipeng[22] and the other authors analyze mixers and how their presence affects and support illegal activities on the various block-chains. The main area of focus in this paper was to observe various patterns from different types of users using various mixing services and how it affects their privacy. Although based on the most recent statistics, mixers are widely popular for illegal activities given the nature of what they do and their operation, however we need to be mindful that not all transactions involving mixers are illegal. Since mixers just mix a user's address with some other random address to increase anonymity they found a way to track these mixed address by looping through their interaction with the mixer. They started with an input address and looped extensively through its transactions to see where it links to the either the same address or the input value matches the received output, in this way they checked if a used a mixing service and split his token into multiple addresses, fed into a mixer and then took them out into another address by checking the output value received they were able to confirm that the particular address was the receiver's address. This way they established a few other patterns where some people were seen to use the same address as the input and output even though the money was mixed in the middle, it still was received into the same address, this invalidated the use of the mixer.

Following this we can conclude that it is in fact possible to crack through the people using mixers to hide their identity. The only thing would be this would require an even exhaustive search to go through all the interactions within and around a mixer to get to the user's address.

4 Methodology

Now that we have identified what has already been done in this field we need to find a way moving forward to tackle this problem of the anonymous action in cryptocurrencies. It is very important that we realise what went wrong in the previous attempts and from there possibly utilise the successful approaches to get to our desired result. We need to consider in our approach that people have been successfully able to cluster the addresses in previous researches which a high success rate so instead of reinventing the wheel we should be learning from them and if possible utilizing their approach and focus only on what is missing. At this stage even though people have tried to link the wallet addresses to the real world entities but they haven't successful, so this is what we will be focusing on the most in our research.

In the following sections we will go through step by step on a proposed idea on how to solve this problem on how to get to a real world entity behind a malicious transaction. The entire process has been divided into 3 steps 3 starting with data collection followed by address filtering and finally mapping the wallet address to a real world entity. The data collection will be primarily done through publicly available APIs wherein we will directly query the blockchain, for address filtering to isolate the malicious addresses we will be taking learnings from previously done research and finally the final step will be done from scratch including building the web scraping service, finding the sources to scrape and building the algorithm to transition and crawl the neighbouring addresses in case nothing is found from the first one.

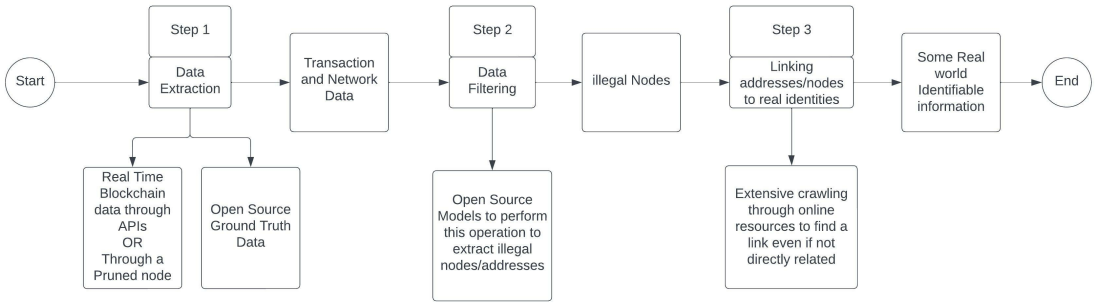

Figure 3: Planned De-anonymization Process

4.1 Step 1

4.1.1 Data Extraction

For the data extraction we will be utilizing one of the 2 approaches listed below:

1. Set up a pruned Bitcoin node on my local machine : Setting up a full node is infeasible and takes weeks to set up because of the size of blockchain

which is roughly around 400GB[7] and is increasing by the every transaction that is put into the block. So, a way around this for us is to set up a pruned node on my local machine. Pruned nodes are known for being lightweight and quickest to install, reason being they only store a couple of gigabytes of data, which includes the headers, and only the latest blocks and transactions for the last 2 days. The only issue with this approach for us is that we would need to heavily customize the client for data collection as the official ones are not developed for this purpose.

2. Public APIs :

   - Blockchain.info[23] is a online service which serves multiple needs as per the requirements of its visitors. It has wallet and exchange functionalities for people who are looking to invest in crypto using their platform. Most importantly for us, they have an exploration service for the most popular cryptocurrencies right now.

     We will be making use of their public APIs to get almost real time for the Bitcoin. Since they are running a full node they have the entire public ledger exposed using their APIs. This way we don't have to spend days and days to download hundreds of gigabytes of data to get the data. The only drawback of this approach over running a full node is that we cannot live monitor the nodes and their connected peers so we won't get their IP addresses. For our current approach this would be okay but if in future we do require the IP addresses of the connected peers we will need to look into running a full or pruned node on our machine.

   - Bitnodes.io [24] is another service which has APIs similar to blockchain.info but it has a few extra ones which would be specifically of interest to us. The striking feature of this source is the node specific data it has like the rankings of the nodes in the network, plus the details about a specific node such as it IP address, status, etc more details mentioned below in the APIs used section.

   (a) Latest Block: This API returns the information of the latest block that was mined into the blockchain such as the transactions that form the block, time the block was mined, hash of the block plus some other information.

   URL: `https://blockchain.info/latestblock`
   SAMPLE RESPONSE:

Listing 1: A Single Block

```
{
    "height": 737507,
    "time": 1653272462,
    "block_index": 1653272462,
    "txIndexes": [
```

```
        5789376937880439,
        8742777465387714,
        1623287914084078,
        3709655268110891,
        457264215309024,
        1782653476355686,
        2548581357155437,
        1696228368104285,
        8116694015932988,
        5520020294808090,
        1540679486633257,
        2116332982076570,
        4113582783598829
        // Rest of the transaction indexes
    ],
    "hash": "000000000000000000631ce095..."
}
```

(b) Transaction Information: This API takes in transaction hash as a path variable and returns all the related information about the transaction such as the timestamp, the inputs and the outputs including its values, addresses and the witness of the transaction as well.
This API will be useful to us for looping and tracking through the transactions and also find the flow of money from address to address.
URL: `https://blockchain.info/rawtx/{transactionID}`
SAMPLE RESPONSE:

Listing 2: A Single Transaction

```
{
    "hash": "7377ec70440e60d4eff3519b936...",
    "ver": 1,
    "vin_sz": 1,
    "vout_sz": 2,
    "size": 266,
    "weight": 956,
    "fee": 0,
    "relayed_by": "0.0.0.0",
    "lock_time": 0,
    "tx_index": 3557373238859613,
    "double_spend": false,
    "time": 1652535475,
    "block_index": 736356,
    "block_height": 736356,
    "inputs": [
        {
            "sequence": 4294967295,
```

```
            "witness":
                "01200000000000000000000...",
            "script": "03643c0b1b2f566961425443
                ...",
            "index": 0,
            "prev_out": {
                "tx_index": 0,
                "value": 0,
                "n": 4294967295,
                "type": 0,
                "spent": true,
                "script": "",
                "spending_outpoints": [
                    {
                        "tx_index":
                            3557373238859613,
                        "n": 0
                    }]}}],
    "out": [
        {
            "type": 0,
            "spent": true,
            "value": 636401465,
            "spending_outpoints": [
                {
                    "tx_index": 2001657083435680,
                    "n": 1
                }
            ],
            "n": 0,
            "tx_index": 3557373238859613,
            "script": "76a914536ffa992491508dc
                ...",
            "addr": "18
                cBEMRxXHqzWWCxZNtU91F5sbUNKhL5PX"
        },
        {
            "type": 0,
            "spent": false,
            "value": 0,
            "spending_outpoints": [],
            "n": 1,
            "tx_index": 3557373238859613,
            "script": "6
                a24aa21a9ed5abdb5a888efa5c4748..."
        }]}
```

(c) Reachable Nodes: This API returns the list of all the active nodes in the bitcoin network and their information such as location, IP address, client they are using, and also the latitude longitude plus some other information.
URL: `https://bitnodes.io/api/v1/snapshots/latest`
SAMPLE RESPONSE:

Listing 3: All Reachable Nodes

```
{
    "timestamp": 1653275683,
    "total_nodes": 15404,
    "latest_height": 737513,
    "nodes": {
        "54.251.65.228:8333": [
            70016,
            "/Satoshi:0.21.1/",
            1653275640,
            1037,
            737515,
            "ec2-54-251-65-228.ap-sout...",
            "Singapore",
            "SG",
            1.3036,
            103.8554,
            "Asia/Singapore",
            "AS16509",
            "AMAZON-02"
        ],
        "37.59.47.27:8333": [
            70015,
            "/Satoshi:0.17.99/",
            1652828604,
            1037,
            737515,
            "ns335655.ip-37-59-47.eu",
            null,
            "FR",
            48.8582,
            2.3387,
            "Europe/Paris",
            "AS16276",
            "OVH SAS"
        ]
        // Rest of the rechable nodes
    }
}
```

4.2 Step 2

Now that we have all the data from the block chain the task at our hand is to find the link between wallet address and some real identity. For us to be able to achieve that we need to follow a sequence of steps Once we have extracted the data from the blockchain we need to filter it out to make sure we don't focus on the nodes which are not going to get us anywhere. Specifically what we need to do is from the entire data find a group or cluster of nodes which might be involved in illegal activities and then perform analysis on them to investigate what all we can find.

4.2.1 Address Clustering

For address clustering what we will be doing is taking inspiration from one of the papers in the related works section and running a machine learning analysis to identify the nodes of interest. The first task of action for us is to identify a machine learning algorithm and a data set to train it to identify illegal nodes/addresses in a network data set. Once we can verify that we have working model with a high accuracy, we will then run the model with our data to make it predict the illegal nodes. This will isolate for us the nodes/addresses involved in illegal transaction by which can follow the trace of how the money moved across the blockchain to get a better idea of how people perform these tasks over these digital networks, and this would possibly be transferable to other cryptocurrencies as well.

4.2.2 Data Visualization

Data visualization is an optional task for us, reason being this would only be helpful to us to explain to someone the severity of the problem and since there are already a large number of tools available online for this e.g. Gephi [25] we won't have to spend too much time on this.

Once we have performed the address clustering we can organize the transaction data into input and outputs to start and visualize into a transaction graph where the nodes are the cryptocurrency address and the edges are the transaction itself. This would help us visualize how different types of addresses interact with other addresses on the network for example in real life money laundering it is very common for people to circulate money address different bank accounts and sometimes even across borders. In cryptocurrencies there is essentially no concept of borders, the entire system behaves like one isolated bank in a country. So, the overall purpose of this activity is more of just to provide more clarity on the problem.

4.3 Step 3

4.3.1 Linking Digital to Real Identities

Now that we would have identified the illegal nodes/addresses the final task at hand which is our ultimate goal for this research is to find some personally identifiable information which can help us link the digital address/identity to some real world entity or eventually to the person who was involved in the activity.

This process will involve building a web crawling service that will go through a number of online sources with the bitcoin address in question to find something linked to it. As of now we have identified a number of services listed below to crawl :

1. BitcoinWhoIsWho[26]
2. btcsniffer.com[27]
3. blockexplorer.com[28]
4. A few more as many as I can find in the following months

The way this would work technically is once we have identified that a particular address is illegal, our first hit would be at the address itself and if we don't find anything we will go one hop by hop away from the address to find something until we have something that we can use to link the address to a real entity. As much as exhaustive as this is, this way we would even be able to overcome mixers as we can see in "On How Zero-Knowledge Proof Blockchain Mixers Improve, and Worsen User Privacy"[22], where they looped through the transactions that went inside a mixer to see how one can eventually trace money that even went inside a mixer. We can follow the patterns they observed from different types of users using mixers to crack those transaction because looking at the statistics[3] mixers carry a huge portion of illegal funds throughout crypto.

Following these steps and conductive exhaustive searching and crawling there is high chance we will be able to track these illegal funds flowing though cryptocurrencies, starting with Bitcoin and eventually these can be put to use with other currencies as well since a lot of the newer currencies and coins were forked off Bitcoin only to start with.

5 Planned Timeline

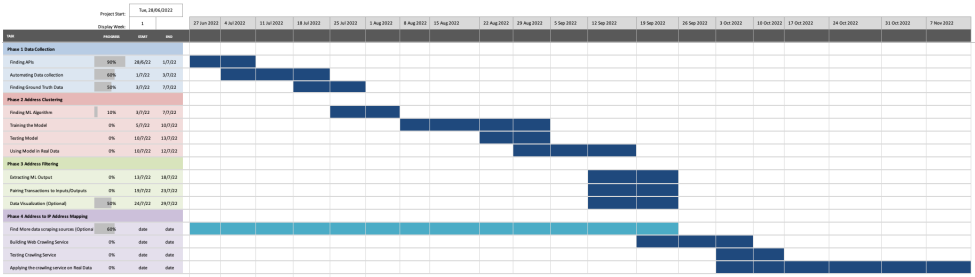

Figure 4: Gannt Chart

So, now that we have a good idea of what needs to be done over the next few months, we need to plan everything and set goals possibly for every week to keep track on things on a weekly basis and resolve any blockers as soon as possible if they arrive.

Initially for the first few weeks my focus will be on to automating the data collection, we have all the APIs ready and working so that should not take long. The only thing we will need to look at is the ground truth data. Once we have cleared out the first step of data collection, then we will need to study more about the previous research to find or develop a machine learning model to isolate the illegal nodes out of all the data, and train that using the ground truth data. After establishing that we will be left with the final part of building the web crawling service to create the mappings between the wallet addresses and real world entities.

6 Conclusions

Following this we can conclude from the previous research that a good amount of work has been done to classify the addresses which will help us in filtering out the data and our work relies on our crawling services. We will start our work with Bitcoin considering it has the maximum market cap and circulation right now so there is a high chance of finding even more reliable source to crawl over the next few months. Once we are able to get good results on Bitcoin we will be able to expand it on other currencies and coins as well as a lot of the newer privacy coins as a lot of them are a fork of bitcoin, meaning they share the same core functionality as Bitcoin with differentiation on how the higher layers work.

So, until of now we have made good progress as we have a good idea of what has already been done and also we are clear about what we are planning to do over the next months to potentially tackle this problem. We have already started the the first task of data collection and made good progress on it and the following tasks will be done once this is completed as highlighted by the timeline.

References

[1] Q. Wang, R. Li, Q. Wang, and S. Chen, "Non-fungible token (nft): Overview, evaluation, opportunities and challenges," *arXiv preprint arXiv:2105.07447*, 2021.

[2] Livemint, "This country adopts bitcoin as legal currency. details here," Apr 2022.

[3] Y. Fanusie and T. Robinson, "Bitcoin laundering: an analysis of illicit flows into digital currency services," *Center on Sanctions and Illicit Finance memorandum, January*, 2018.

[4] D. of Justice, "Individual arrested and charged with operating notorious darknet cryptocurrency mixer," April 2021. [Online; posted 28-April-2021].

[5] J. Taylor, "'Complex and volatile': cryptocurrencies should be regulated by financial watchdogs, say consumer advocates," *The Guardian*, May 2022.

[6] S. Delgado-Segura, C. Pérez-Solà, J. Herrera-Joancomartí, G. Navarro-Arribas, and J. Borrell, "Cryptocurrency networks: A new p2p paradigm," *Mobile Information Systems*, vol. 2018, 2018.

[7] B. Explorer, "Blockchain explorer," May 2022. [Online;].

[8] Bitnodes, "Live bitcoin nodes," May 2022. [Online;].

[9] C. Reitwiessner, "zksnarks in a nutshell," *Ethereum blog*, vol. 6, pp. 1–15, 2016.

[10] C. Info, "Coin market cap."

[11] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder, *Bitcoin and cryptocurrency technologies: a comprehensive introduction*. Princeton University Press, 2016.

[12] U. W. Chohan, "The cryptocurrency tumblers: Risks, legality and oversight," 2017.

[13] C. Chan, "Nanaimo men charged with operating crypto-funded drug ring on dark web — vancouver sun," 2022.

[14] J. Wilser, "Drugs, drugs and more drugs: Crypto on the dark web," Apr 2022.

[15] N. Hiramoto and Y. Tsuchiya, "Measuring dark web marketplaces via bitcoin transactions: From birth to independence," *Forensic Science International: Digital Investigation*, vol. 35, p. 301086, 2020.

[16] P. Koshy, D. Koshy, and P. McDaniel, "An analysis of anonymity in bitcoin using p2p network traffic," in *International Conference on Financial Cryptography and Data Security*, pp. 469–485, Springer, 2014.

[17] A. Support, "About icloud private relay," Apr 2022.

[18] X. Xueshuo, W. Jiming, Y. Junyi, F. Yaozheng, L. Ye, L. Tao, and W. Guiling, "Awap: Adaptive weighted attribute propagation enhanced community detection model for bitcoin de-anonymization," *Applied Soft Computing*, vol. 109, p. 107507, 2021.

[19] A. Biryukov and S. Tikhomirov, "Deanonymization and linkability of cryptocurrency transactions based on network analysis," in *2019 IEEE European symposium on security and privacy (EuroS&P)*, pp. 172–184, IEEE, 2019.

[20] M. Weber, G. Domeniconi, J. Chen, D. K. I. Weidele, C. Bellei, T. Robinson, and C. E. Leiserson, "Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics," *arXiv preprint arXiv:1908.02591*, 2019.

[21] Y. Hu, S. Seneviratne, K. Thilakarathna, K. Fukuda, and A. Seneviratne, "Characterizing and detecting money laundering activities on the bitcoin network," *arXiv preprint arXiv:1912.12060*, 2019.

[22] Z. Wang, S. Chaliasos, K. Qin, L. Zhou, L. Gao, P. Berrang, B. Livshits, and A. Gervais, "On how zero-knowledge proof blockchain mixers improve, and worsen user privacy," *arXiv preprint arXiv:2201.09035*, 2022.

[23] B. Info, "The most trusted crypto company."

[24] Bitnodes, "Bitnodes api," May 2022. [Online;].

[25] G. Info, "Gephi: The open graph viz platform."

[26] bitcoinWhoIsWho, "bitcoinwhoiswho," May 2022. [Online;].

[27] bitsniffer, "bitsniffer," May 2022. [Online;].

[28] blockexplorer, "blockexplorer," May 2022. [Online;].