Darknet and Bitcoin De-anonymization: Emerging Development

Wenlin Han, Viet Duong, Long Nguyen, and Caesar Mier

Department of Computer Science
California State University, Fullerton
Fullerton, USA
whan@fullerton.edu
vietanhduong86@csu.fullerton.edu
longnguyen0024@csu.fullerton.edu
caesarmier@csu.fullerton.edu

Abstract— **Darknet websites, the warm beds for money laundry, child pornography, and illicit drug trafficking, are built on hidden services and anonymous communication protocols. Cryptocurrencies, such as Bitcoin, is the major payment method used on Darknet. In this paper, we summarize and introduce the latest development on de-anonymization techniques used to reveal the hidden information that are helpful for crime investigation, which is a key step for the future research work.**

Keywords— Darknet, Bitcoin, Cryptocurrency, De-anonymization, Blockchain, Tor

I. Introduction

The popularity in the research of blockchain [19] and technology has increased enormously since the time Bitcoin was introduced. Cryptocurrencies are digital or virtual currencies that are designed to act as a medium of exchange. Cryptocurrencies use cryptography to secure and verify all the transactions and control the creation of new units of currencies.

Darknet websites are the websites that allow anonymous transactions between anonymous users. These websites use cryptocurrencies, such as Bitcoin, as payment methods, and they are built upon hidden services, such as Tor. Thus, they are often the warm beds for a variety of crimes, such as money laundry, child pornography, illicit drug trafficking, etc.

De-anonymization is a strategy in which anonymous data is cross-referenced with other available data sources to re-identify the anonymous data source or generalized information. Moreover, de-anonymization identifies encrypted information. You can predict the usage pattern heuristic an attacker can connect to the public that shows its addresses over 85.5 % of the transactions that use a zero-knowledge proof [1]. Hence, De-anonymization is possible to detect using some techniques that are available nowadays will be listed below that are available in the market that helps us accomplish the need to discover the hidden transaction in cryptocurrency as popular as Zcash, Ethereum, Bitcoin, etc.

In this paper, we summarize the state-of-the-art de-anonymization techniques used on Darknet websites and Bitcoin ledger including trawling, graph analysis, clustering, mixer, heuristic approach, to name a few. This paper lays a foundation for further research work. The rest of the paper is organized as follows: several de-anonymization techniques used on darknet websites are introduced in Section II. In Section III, we introduce the latest methods used to de-anonymize Bitcoin. The paper is concluded in Section IV.

II. Darknet de-anonymization

In this section, we will discuss the techniques used to detect hidden information on darknet websites.

A. Detecting Leaks

Based on the Tor services, it has three steps [2] such as exploration, candidate selection, and validation. In the beginning, the onion URLs will enter Tor exploration. Onion URLs will expand and fetch onion pages in tor exploration and open the database. Next phrase, it will go into the candidate selection, which will contain certificates, identifiers, page titles, endpoints (it will also involve the internet in this phrase) list of candidate pairs will go into the validation stage. Validation will check to see if the input candidate pair and verifies if a candidate Internet endpoint will host the hidden services. It then goes back to visit the internet endpoint to fetch contents and certificates.

They can find the location leaks by looking into this phrase 3, where all the user's activities take place there. Furthermore, using deanonymizing Tor hidden services is not a good idea when a user connects directly to Bitcoin using benign transactions and with those sensitive transactions, we can forward traffic through a chain of Tor relays or VPNs [3]. The attackers will have a hand on the sensitive data if they implement the attack by getting in the middle. The attackers will process all the fingerprints of and record of the user's transactions (going through attacker's-controlled nodes).

978-1-7281-8259-9/20/$31.00 ©2020 IEEE

B. Cell Manipulation

Services hidden in Tor are often given to users via TCP protocol in order not to expose the hidden server's IP address, also to achieve censorship [4]. Hidden services are usually abused to illegal purposes like human trafficking, pornography etc. Many studies and researches have been conducted to analyze the advantages and disadvantages of attacking schemes of hidden services in Tor. Many hidden services in Tor confuse users and while providing them services, they can hold illegal websites. This act can lead to many terrible consequences [5]. Therefore, one of the best ways to prevent this is to attack these hidden services to deanonymize them.

There are three attacking schemes that attackers often use: manipulating Tor cells, cell count based method, padding cell method [4]. Nothing is always good, and nothing is always bad. Using hidden service in Tor can keep the anonymity of the service provider; however, it also keeps the anonymity of illegal hosts.

C. JoinMarket

In the paper [6], Malte et al. perform the first measurement study of JoinMarket, a growing platform for anonymous cryptocurrency transfers. Over the course of thirteen-month research, the platform was responsible for over 29 million dollars in transactions taking place. They discovered that a well-funded attacker with a 90% success rate would require $14,000 to $54,000. JoinMarket is one of the cheapest ways to launder money in Bitcoin and is only growing in popularity.

D. Silk Road

Silk Road funds were an activity that everyone paid attention to, which is best known as the address of the FBI via 445 transactions of exactly 324 Bitcoins [7]. The algorithm chose a specific FBI address as a user of a high page-ranked node. It also determined the nodes related to interests and allowed for looking more thoroughly into high-page ranked nodes. By collecting the information and web-scraping from the user's data, the authors were able to discover direct transactions from various Bitcoin users. Moreover, they also can backtrack from any of this transaction to find out bitcoin forum users with only one hop from the Silk Road.

E. BitScope

As we briefly talked about de-anonymizing hidden services in Tor above, anonymity in Bitcoin works almost the same way. Anonymous attackers take advantage of this feature to conduct attacks. However, it can be prevented by de-anonymizing the real identities behind anonymous addresses.

However, in the real world, it is feasible but quite difficult to have ideal results. One of the most challenging issues is scalability. It is very hard to keep up with enormous blockchain data and produce results in a timely manner. That is the reason computer scientists at University of Washington created BITSCOPE. They use BITSCOPE to deal with nearly 1 billion nodes and analyze its de-anonymization [8].

F. Trawling

In the paper [9], they look at Tor's vulnerabilities as well as deanonymization. Because an attacker is connected to the Tor, they can determine the IP address of the Tor hidden service and if their nodes are being used by Tor, whether encrypted or not.

When an attack is taking place, they can discover guard nodes to lead to the next step in our given attack. Because Tor services are always online, this is a significant vulnerability that helps to deanonymize Tor. The attack during their research correctly identified two guard nodes. Their attack only took about 2 hours, where they worked with rendezvous points trying to accurately identify the guard nodes.

III. Bitcoin De-anonymization

In this section, we will introduce the latest techniques used to de-anonymize Bitcoin transactions.

A. Graph Analysis

Graph Analysis is a framework which developed to de-anonymize the identities of users that gave publicity available information.

The graph has described the outline at the start from Bitcoin, it then moves to blockchain parse Tx-Graphentity Graphuser activity revealed, and on the other side of the transaction, web-Scraping (forum) will also connect the users' ID. And any of the transactions of the activity of bitcoin will be shown through the Tx Graph, which contains transaction and user graphs.

In the paper [10], they focus on the benefits and challenges of de-anonymizing bitcoin using Graph Convolutional Networking (GCN) analysis, putting matrix factorization, and random walk on the backburner. GCN is a two-hop graph embedded computation that adds its neighbors to a node with much better, juicier, information that gives better accuracy of the results.

It gathers information on the first hop, then GCN unveils the known vertices on the second hop and updates the weight of the neural network. Some issues they came across when working with Bitcoin was that the blockchain was filled with millions of vertices and billions of edges, making it use up more memory. Other kinds of user activity can hurt the process of retrieving accurate information like distributed storage.

GCN does have a few advantages when it comes to graph analysis. Luckily, with so many empty accounts or inactive accounts, it ignores a lot of addresses. In doing so, it makes the computational process a whole lot faster. Another advantage is that a lot of real-world identities are published publicly on third party websites so we can configure the GCN to use that information to make our process a whole lot easier.

B. Stealth Addresses

This technique is the 2nd generation anonymization techniques. Stealth addresses provide the security for whoever posts a bitcoin address to receive money publicity. Although the address is not technically seen in public, multiple transactions can be linked to the same address. Stealth addresses allow a user to publish a single static identifier. By using this identifier, which is known as public key Q, the sender obtains a unique address of bitcoin and uses that to hide it from publicity. The protocol of stealth addresses is based on Diffie–Hellman key exchange on elliptic curves [11]. It tends to prevent attackers from typing information to an address that allows them to identify a bitcoin user. However, it is still possible to identify the usage of this technique so that the attackers can reconstruct the original stealth addresses Q.

C. Heuristic Approach

Bitcoin was first created in 2008 by Satoshi Nakamoto as a P2P system of electronic currency. To analyze the Bitcoin Blockchain, many different areas have been intensively researched and studied. Since the whole purpose of bitcoin and cryptocurrency changed, Cui and his fellow scientists came up with a new approach that can follow the new trend. Specifically, they use a heuristic approach to analyze the patterns of financial High-Frequency Transaction (HFT) [12]. Based on this approach, they created the de-anonymous method to check if activity information of IP matches with transaction records in blockchain. They proved the superiority of IP matching methods.

According to the paper, analyzing methods are divided into two main categories: Direct information recorded in the blockchain and indirect information of the network graph [12]. First, based on Power Law Distribution of the Bitcoin Network, the scientists discovered that direct analysis is more efficient than clustering methods because heuristic approach has better performance and can save calculation capacity. These experiments produce some amazing results.

D. CoinJoin Mixer

We mainly look at the detection of transactions involved in money laundering schemes and the possibility of modification of payment systems. Apart from the advantages of Bitcoin, it also has many disadvantages. One of them is that anyone who is a member of the network can have access to all the transactions [13]. This is quite dangerous because this data can be used for malicious purposes. Although addresses in the bitcoin network do not represent users, with basic heuristic search, one can identify a specific user they target at.

Therefore, many developers came up with services named “bitcoin mixers.” These services can anonymize transactions by obfuscating paths or mixing up addresses of senders and recipients. However, there have been many reports of fraud related to the use of these services. Gregory et al. developed an algorithm called CoinJoin [13]. But like any other algorithms, CoinJoin has pros and cons. The algorithm keeps bitcoins from getting stolen from the malicious mixer service. However, since this is a public service, anyone can make use of it, from honest users to criminals.

To solve this problem, the paper suggests a new analysis approach of the transaction: build two graphs (transactions and users’ graph). Transaction graph is like a flow of Bitcoin transactions. Vertex is transaction. Edge is bi-directional and is a bridge connecting output and input. User graph works with users. After testing, all transactions are clearly marked and tracked. This data can be used to verify the legitimacy of tracking. Therefore, CoinJoin transactions are practical and can be improved.

E. Data-Driven Approach

In the paper [14]. The researcher analyses the performance in several clustering algorithms/heuristics: Multi-input, Shadow, Consumer, Optimal. Bloom Filters are data structures that help check whether an element is a member of a set. The advantage is that it uses less space and only uses two operations: insert and query. Insert puts the data through the filter, and the query returns whether the data has been inserted before. These Bloom Filters are commonly remembered by their false-positive rates.

However, one advantage is that you never get false negatives using this algorithm. The author exploits a vulnerability in the implementation of Connection Bloom Filtering to gather real data on more than 30,000 wallets. The attack targeted the most used SPV (simplified payment verification) wallet library that uses Bloom filters called BitcoinJ. The main disadvantage with BitcoinJ is that they put both their pubkey and pubkey-hash into the Bloom filter. So, if the wallet really owns the pubkey then this means both the pubkey and pubkey-hash must be in the filter meaning it is easier for an attacker to guess true positives with a high probability.

The attack was performed on older and modern wallets, but modern wallets were the area of interest. Now that they had the wallets, they applied their heuristic to them and used the wallet data to measure the performance on ground truth. It turns out that even modern wallets have a vulnerability to being identified as an owner to many transactions. Multi-input had a mean recall of 68.59% accuracy. Shadow had 69.16%, Consumer had 69.26%, and Optimal had 69.34%. Privacy is hard to come by on Bitcoin, unlike many claims it to be.

F. Clustering

In the paper [15], they focus on the efficient automatic clustering algorithm based on blockchain and off-chain information. In the blockchain, they studied the unspent transaction outputs, or UTXO, focusing on common spending and one-time change to help cluster the addresses but may have errors.

Common spending says if two or more inputs were used in a transaction and there is only one output, then those inputs must belong to the same owner [15]. They tried to ignore transactions with multiple outputs because those can be shared with other users, so there is not enough accurate information to use.

One-time change looks at the change from a transaction and sees where it goes (a new bitcoin address). If the situation matches their algorithm, then the one-time change output and all its inputs belong to the same user. This algorithm also faces its challenges of having errors in the data but still beneficial information when trying to de-anonymize Bitcoin.

Next, they focused on off-chain information, data anywhere else other than the blockchain, to see if they could match addresses with users. They collect tags passively and actively. The passive approach they are searching through the web for any information that can help identify the users. The active approach is where they analyze Bitcoin companies and data actualization procedures. Most companies use prefixes with their addresses, so to guess the correct address, one must use many different private keys to reach the desired solution.

Their new Bitcoin address clustering algorithm is very detailed, so much so that it is better than most other algorithms. It uses not only blockchain information but also off-chain information by searching throughout the web. They also use specific off-chain data to help avoid encountering bogus information from cluster merging. They used this off-chain information to get a better understanding of the clustering information. The largest cluster they discovered had a size of 2,475,769 addresses. So, it is possible to cluster addresses more accurately using the techniques mentioned above.

The next paper we researched was *De-Anonymizing the Bitcoin Blockchain* [16]. It dives deep into Bitcoin, investigating whether utilizing different addresses for every transaction helps in keeping transaction anonymity. They do this by attempting to cluster together addresses owned by the same entity. Similar approaches have been used in the past, and they build on that using either structural properties of the bitcoin transaction graph or network-level data on user behavior. Bitcoin is popular because many believe you can hide behind a screen without anybody knowing who you are; however, all transactions are public.

Clustering of addresses based on public transaction information from the blockchain, network, cluster found by Meiklejohn et al., and clusters created by Chainalysis.com. The authors then used a graph clustering algorithm described by Zhou et al. It demanded to change all attributes to vertices on the graph. The authors then drew edges from the vertices (transactions) to each attribute associated with it. They use heuristic 1, which says if multiple inputs are used then those inputs belong to the same wallet. Heuristic 2 has to do with the change property from a transaction and that it belongs to the input user. In heuristic 3, they join single-link transactions; however, this heuristic is not very accurate.

They then used the clustering algorithm HDBSCAN because it satisfies all the requirements he initially needed. There were noisy data, varying-numbers of clusters, varying-density clusters, floating data, and large data size. The authors then compared his clusters with those generated from Chainalysis.com heuristics but to a much smaller data set. The methodology was successful in clustering the test cases though they recommend future work be done to obtain a more robust dataset.

G. Machine Learning

Another technique for classification is to apply supervised learning classification algorithms from Burks et al. [17]. Similar to neural network technique, they collect and cluster data, then encode features for bitcoin addresses about its behaviors. The behaviors would include number of transactions, the maximum balance, lifetime. The algorithms they used for classifying and predict is Support Vector Machine and Random Forest. Support Vector Machine uses a kernel to separate addresses into classes such as gambling address, mixing services address, exchanges address, and Tor market address. Random Forest uses a decision tree to classify base on the features.

In the paper [18], they used a neural network to do the work. First, they cluster addresses using k-mean clustering. Then they encode some features that they believe that would differentiate users. After that, they used a neural network to study the behaviors and classify. Then the neural network will have the ability to predict whom a strange new address belongs to.

IV. Conclusion

In this paper, we summarized and introduced the state-of-the-art research works on darknet and Bitcoin de-anonymization, which include trawling, graph analysis, clustering, mixer, heuristic approach, etc. The summary lays a foundation for future research work.

References

[1] A. Biryukov and D. Feher, “Deanonymization of hidden transactions in zcash,” 2018, https://cryptolux.org/images/d/d9/Zcash.pdf.

[2] S. Matic, P. Kotzias, and J. Caballero. Caronte: Detecting location leaks for deanonymizing tor hidden services. In CCS, 2015.

[3] A. Biryukov and I. Pustogarov. Bitcoin over Tor isn’t a good idea. In IEEE Symposium on Security and Privacy, 2015.

[4] S. Nepal, S. Dahal, and S. Shin, “Deanonymizing schemes of hidden services in tor network: A survey,” in International Conference on Information Networking, 2015.

[5] Al Jawaheri, H., Al Sabah, M., Boshmaf, F., Erbad, A. (2017): Deanonymizing Tor hidden service users through Bitcoin transaction analysis. Retrieved from: https://arxiv.org/pdf/1801.07501.pdf.

[6] M. Moser and R. B ̈ ohme, “The price of anonymity: empirical evidence ̈ from a market for Bitcoin anonymization,” J. Cybersecurity, 2017.

[7] Michael Fleder, Michael Kester, and Sudeep Pillai. 2013. Bitcoin Transaction Graph Analysis. Technical Report. Massachusetts Institute of Technology (MIT), Computer Systems Security (6.858). http://css.csail.mit.edu/ 6.858/2013/projects/mfleder-mkester-spillai.pdf.

[8] Z.Zhang, T. Zhou, Z. Xie, “BITSCOPE: Scaling Bitcoin Address De-anonymization using Multi-Resolution Clustering.” https://blog.zhen-zhang.com/bitscope-public/paper.pdf.

[9] A. Biryukov, I. Pustogarov, and R.-P. Weinmann, “Trawling for tor hidden services: Detection, measurement, deanonymization,” in Proceedings of IEEE Symposium on Security and Privacy (SP’13). IEEE Computer Society, 2013.

[10] Anil Gaihre, Hang Liu, Santosh Pandey(2019). “Deanonymizing Cryptocurrency With Graph Learning: The Promises and Challenges”. 2019 IEEE Conference on Communications and Network Security (CNS), June 2019.

[11] M. Möser and R. Böhme, “Anonymous alone? Measuring bitcoin’s second-generation anonymization techniques,” in Proc. IEEE Eur. Symp. Security Privacy Workshops (EuroS&PW), 2017, pp. 32–41.

[12] J. Cui, H. We, et al., “De-anonymizing Bitcoin Networks: An IP Matching Method via heuristic Approach,” 2019.

[13] Maksutov, A.A.; Alexeev, M.S.; Fedorova, N.O.; Andreev, D.A. Detection of Blockchain Transactions Used in Blockchain Mixer of Coin Join Type. In Proceedings of the 2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), Saint Petersburg and Moscow, Russia, 28–31 January 2019; pp. 274–277.

[14] Jonas Nick. Data-Driven De-Anonymization in Bitcoin. Master’s thesis, ETH Z ̈urich, 8 2015.

[15] Ermilov, D., Panov, M., Yanovich, Y.: Automatic bitcoin address clustering. In: 16th IEEE International Conference on Machine Learning

and Applications. pp. 461–466 IEEE (2018). https://doi.org/10.1109/icmla.2017.0-118.

[16] Bharath Srivatsan, Christina Huang, and Arvind Narayanan. “De-Anonymizing the Bitcoin Blockchain”, 2016. https://pdfs.semanticscholar.org/88a5/a57b30438e8349ea2c388085846e9187cca6.pdf.

[17] Lynne Burks, Andrew Cox, Kiran Lakkaraju, Mark Boyd, and Ethan Chan. “Bitcoin Address Classification,” Machine Learning and Deep Learning Conference 2017.

[18] Zola, F.; Eguimendia, M.; Bruse, J.L.; Orduna Urrutia, R. Cascading Machine Learning to Attack Bitcoin Anonymity. In Proceedings of the 2nd IEEE International Conference on Blockchain, Atlanta, GA, USA, 14–17 July 2019; pp. 1–8.

[19] Austin Draper, Aryan Familrouhani, Devin Cao, Tevisophea Heng, and Wenlin Han, “Security Applications and Challenges in Blockchain,” IEEE International Conference on Consumer Electronics (ICCE’19), pp. 1-4, Las Vegas, USA, Jan. 11-13, 2019.