Longitudinal Analysis of Misuse of Bitcoin*

Karim Eldefrawy[1], Ashish Gehani[1], and Alexandre Matton[2]**

[1] SRI International, USA
[2] Stanford University, USA

Abstract. We conducted a longitudinal study to analyze the misuse of Bitcoin. We first investigated usage characteristics of Bitcoin by analyzing how many addresses each address transacts with (from January 2009 to May 2018). To obtain a quantitative estimate of the malicious activity that Bitcoin is associated with, we collected over 2.3 million candidate Bitcoin addresses, harvested from the dark web between June 2016 and December 2017. The Bitcoin addresses found on the dark web were labeled with tags that classified the activities associated with the onions that these addresses were collected from. The tags covered a wide range of activities, from suspicious to outright malicious or illegal. Of these addresses, only 47,697 have tags we consider indicative of suspicious or malicious activities.

We saw a clear decline in the monthly number of Bitcoin addresses seen on the dark web in the periods coinciding with takedowns of known dark web markets. We also found interesting behavior that distinguishes the Bitcoin addresses collected from the dark web when compared to activity of a random address on the Bitcoin blockchain. For example, we found that Bitcoin addresses used on the dark web are more likely to be involved in mixing transactions. To identify mixing transactions, we developed a new heuristic that extends previously known ones. We found that Bitcoin addresses found on the dark web are significantly more active, they engage in transactions with 20 times the neighbors and 4 times the Bitcoin amounts when compared to random addresses. We also found that just 2,828 Bitcoin addresses are responsible for 99% of the Bitcoin value used on the dark web.

1 Introduction

Understanding how cryptocurrencies may affect society depends on being able to analyze their use and misuse. We present a first step in this direction. Our study shows a decline in the level of malicious Bitcoin activity over the years, when measured in terms of the number of addresses involved. The decline of Bitcoin's

* This material is based upon work supported by the National Science Foundation (NSF) under Grant ACI-1547467. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

** Research performed while visiting SRI.

usage in suspicious and malicious activities is not surprising for those who follow the space closely. There is now an increased awareness about the lack of strong anonymity in Bitcoin, in comparison with other privacy-preserving coins, such as Monero [18] and Zcash [24]. Even though Bitcoin usage in suspicious activities is declining, our study is still useful since it provides a quantitative understanding of the trend. We believe that our findings can benefit other researchers as well as help educate administrators and law enforcement as they create and implement new regulations.

1.1 Cryptocurrency Studies

Analysis of Bitcoin: One of the first attempts to analyze the Bitcoin blockchain was performed in 2012 by Ron and Shamir [21]. They studied Bitcoin's transactions graph and identified interesting patterns in it. The scale and complexity of the graph has exploded in the seven years since that study was performed. Subsequent analyses [16, 1, 23, 4, 19, 20] have used heuristics to cluster Bitcoin wallets, based on evidence of shared authority, and then perform active re-identification attacks – for example, by purchasing goods and services to classify the operators in clusters [16], and searching for transaction patterns on exchanges [20].

BitRank [3] is a proprietary wallet scoring system developed by the startup Blockchain Intelligence Group (BIG). BIG's website states that the current beta version of BitRank performs real-time risk assessment to determine the relative safety of pending Bitcoin transactions. As of January 2019, the site provides little public information about the technical details of the system. There are several other services that analyze Bitcoin (and other systems such as Ethereum and Litecoin) to aid businesses and law enforcement. These include Chainanalysis [9], CipherTrace [10], and Elliptic [11]. To the best of our knowledge, they have not published analysis that covers the material in our study for the duration we consider.

In parallel to our effort, Lee *et al.* [15] collected 27 million dark web pages and extracted a mix of 10 million unique Bitcoin, Ethereum, and Monero addresses. They classified the usage of the addresses, identified their use in the trade of illicit goods, and traced cryptocurrency flows, to reveal black money activity on the dark web. Their analysis shows that more than 80% of Bitcoin addresses found on the dark web were involved in malicious activities. The monetary value of the associated cryptocurrency activity was estimated to be $180 million.

Other Cryptocurrencies: In recent work [17], some transactions of the privacy-focused cryptocurrency Monero [18] were found to be highly linkable. We do not claim that any of our analysis or results apply to privacy-preserving cryptocurrencies. This paper only considers Bitcoin, with similar analysis of Monero left as challenging future work.

1.2 Contributions

We provide the following:

- A quantitative study on the misuse of Bitcoin in malicious contexts. Such activities are identified by collecting Bitcoin addresses that are advertised as a means of payment on dark web onions associated with a wide range of undertakings, such as selling illegal substances, human trafficking, and ransomware.
- New heuristics to identify *CoinJoin* mixing transactions. We believe that our heuristics are of independent interest.

We emphasize that our study does not claim that Bitcoin has been (or is) used only for malicious or illegal activities. Our aim is to provide a quantitative assessment of the extent of such activities. This is critical for researchers, regulators, law enforcement, and the wider community to understand the magnitude and scope of the problem. We believe that this understanding is necessary for the cryptocurrency ecosystem to mature.

1.3 Summary of Findings

We highlight some of our results below.

1. **Bitcoin Ownership and Use (in Section 3.1):** Less than 0.06% of all Bitcoin addresses own over 99% of all bitcoins. In particular, that 0.06% consists of 2,266,265 out of 397,301,155 unique addresses observed. Between January 2009 and May 2018 each address participated in at least one of the 316,386,663 transactions that we analyzed. Most addresses were used at most a few times, which is what we expect based on how wallet software is designed and used.

2. **Bitcoin on the Dark Web (in Section 3.3):** Of the 2,093,568 Bitcoin addresses found on the dark web, 276,549 were from mirrors of the *Blockchain.info* explorer. 82% of the remaining addresses were active – that is, participated in at least one transaction. In particular, there were 1,491,709 active addresses. Of these, only 47,697 had tags that we considered indicative of suspicious or malicious activities. Just 2,828 addresses owned 99% of the bitcoins that were involved in the dark web. There was a clear decline the number of Bitcoin addresses appearing on the dark web in the months in which dark web markets were taken down.

3. **Mixing Transactions (in Section 4):** The fraction of all Bitcoin addresses that participate in at least one CoinJoin transaction is only 0.4%. However, our analysis found that on the dark web, this fraction was 5 times higher – that is, 2.3% of Bitcoin addresses found here were part of CoinJoin operations.

4. **Transaction Characteristics (in Sections 5.1 and 5.2):** When considering all Bitcoin addresses, 340,138,543 (85.7%) of them have transacted with less than 10 other addresses, while only 25,7925 (0.06%) have transacted

with more than a 1,000 addresses, and only 6,178 (0.002%) transacted with more than 10,000 addresses. In contrast, 597,744 (40.1%) of Bitcoin addresses found on the dark web have transacted fewer than 10 other addresses, while 61,330 (4.1%) have transacted with more than 1,000 addresses, and 3,244 (0.2%) have transacted with more than 10,000 addresses. The higher participation in mixers is one reason that the Bitcoin addresses found on the dark web have transacted with more addresses.

1.4 Study Limitations

Given the significant scope of the effort, it had its limitations. We note three in particular:

1. *Coverage of dark web:* The data spans June 2016 to December 2017. No claim is made regarding its completeness. Section 3.2 describes our collection methodology and the resulting data.

2. *Dark web data labeling:* We relied on previous research on (thematic) labeling of dark web onions to describe the activities that they are involved in. An address that is collected from an onion inherits its labels (which we call tags). Note that only a subset of tags are indicative of suspicious or malicious activities. Section 3.2 describes how the labeling was performed in prior work.

3. *Analysis accuracy:* Since we did not have the ground truth for much of the analysis that we performed, we could not cross-check the accuracy of our inferences. The transaction graph is based on publicly available information, ensuring its reliability. There is also a basis for confidence in the labeling of the dark web data since some of it was manually verified. Since our work on detecting mixing transactions depends on heuristics, the results may have both false positives and false negatives. However, we did verify as many mixing transactions as we could.

1.5 Outline

Section 2 covers background on Bitcoin and mixers (especially CoinJoin). Section 3 provides an overview of the data sets used in our study, a characterization of behavior observed in the individual data sets, a description of our modified heuristic for detecting mixing transactions, and the properties of such transactions. Section 5 presents more details of our Bitcoin analysis and our findings. Section 6 concludes with a discussion of future work.

2 Bitcoin Preliminaries

2.1 Identifying Bitcoin Addresses

Bitcoin uses the Elliptic Curve Digital Signature Algorithm (ECDSA). Each user has at least one ECDSA key pair. A user can digitally sign a transaction with their private key. The user's public key can be used to verify that the signature is valid. The user's Bitcoin address is an encoding of the 160-bit hash of the public key [6]. A Bitcoin address contains a built-in checksum. This allows detection of malformed addresses, as may occur if the address is mistyped.

A Bitcoin address can be generated offline using wallet software. Even if it is listed on web pages, it may never be used. We only consider an address active if it has appeared on the public Bitcoin blockchain. When bitcoins are sent to an address that is well-formed but not owned by any user (or if the user has lost the corresponding private key), the bitcoins will be lost. In the latter case, the private key may be recovered by alternate means [5].

To construct a Bitcoin address, the hash of the user's ECDSA public key and checksum are converted to an alphanumeric representation. This is done using the Base58Check custom encoding scheme. The resulting address can contain all alphanumeric characters except 0, O, I, and l. Normal addresses start with 1, while addresses from script hashes begin with 3. An address that is used on the main Bitcoin network is 25-34 characters long. Most are 33 or 34 characters in length.

Initially, a regular expression was used to extract candidate Bitcoin addresses from the dark web pages. (See Section 3.2 for more information.) Of the 2.3 million found, 0.2 million failed to pass the checksum test [7]. It is expected that some addresses that were classified as inactive in the study will subsequently be used.

2.2 Mixing Transactions

A CoinJoin is a specific type of Bitcoin transaction. It enables a participant to increase their anonymity by "mixing" their payment with that of other users. Each participant creates a new Bitcoin address. Next, they construct a fixed size payment to it. This is then sent to an aggregator that collects all the participants' payments. The aggregator constructs a single transaction that includes all of the payments. This transaction is sent to all the participants for them to sign. The security of the CoinJoin depends on the fact that the transaction is not valid till every participant provides a signature. Once all the signatures are received, the transaction can be posted for inclusion in a block by a miner. Since the payments are all the same amount, there is no direct way to connect an output to a specific input. All of these steps are handled by wallet software. See the CoinShuffle paper [22] for more information.

The aggregator can be a centralized service or a peer-to-peer protocol. Some services are available on the public web, such a JoinMarket and CoinShuffle. Other services are only present on the dark web. Since the payment from a participant in a CoinJoin should only be connected to a single output, this class of transactions introduces noise in our analysis. Of equal concern is that all the other participants appear as payees. This motivated us to develop a heuristic to detect CoinJoins so that they can be excluded from selected portions of our analysis. See Section 4 for more detail.

3 Bitcoin and Dark Web Data Sets

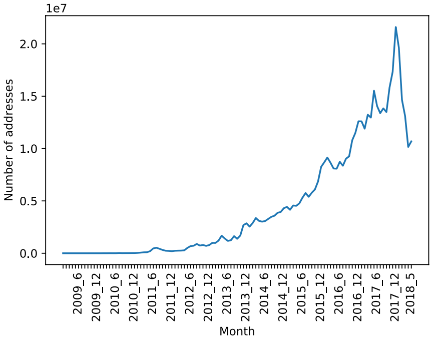

Fig. 1: Monthly used Bitcoin addresses

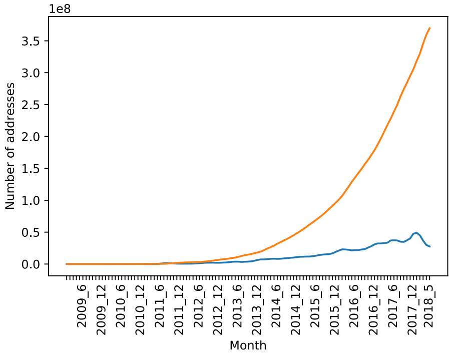

Fig. 2: Active (in last 3 months, blue) and inactive Bitcoin addresses (orange)

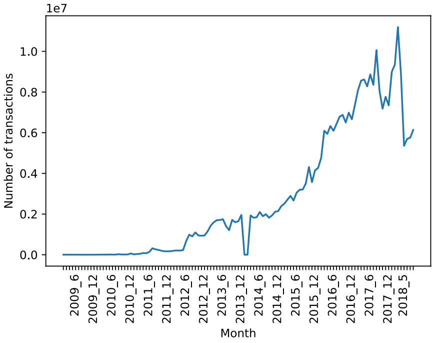

Fig. 3: Bitcoin transactions per month

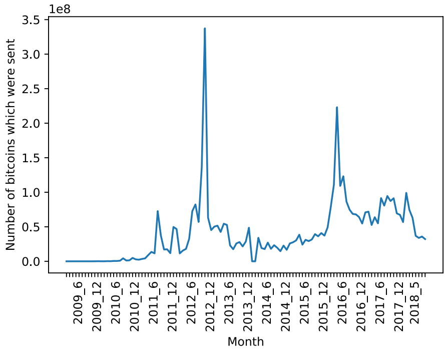

Fig. 4: Amount of bitcoins sent

3.1 Bitcoin's Blockchain

From the genesis of Bitcoin in 2009 to the end of May 2018, the blockchain contains 397,301,155 unique addresses that have participated in at least one of the 316,386,663 transactions that occurred in that timeframe. The number of addresses used in a given month has increased rapidly, as seen in Figure 1. Since we collected this information using a Bitcoin client, we also cross-checked the numbers with data from the *Blockchain.info* explorer [2]. The number of transactions per month is also increasing, as seen in Figure 3. However, the quantity of bitcoins transferred each month is significantly more volatile than the number of addresses used or transactions. This can be seen in Figure 4.

Bitcoin addresses behave very differently from each other. Most addresses are only used a few times. This is because the reuse of a single address makes a user more susceptible to deanonymization. A payment from an address must reference and not exceed the sum of past Unspent Transaction Outputs (UTXOs) to that address. To avoid overpayment, a change address is used. In principle, this can be the payer's original address. In practice, it is a different address for the aforementioned reason. The use of wallet software automates the process of using a new address for each transaction and a different change address.

Many addresses participate in several transactions. Some participate in a large number of transactions. For multiple measures, such as the number of transactions per address, or the number of bitcoins received by each address, the resulting graphs can be approximated by a Pareto power-law distribution. For instance, more than 99% of bitcoins used in transactions belong to just 0.06% of the number of addresses that have been used. (The 0.06% set consisted of 2,266,265 addresses.)

3.2 Dark Web Data

Prior work at SRI focused on collecting, labeling, and categorizing information from the dark web [12]. The effort had to first receive approval from SRI's Institutional Review Board (IRB) due to the complex legal and ethical considerations involved. We have not focused on these aspects in our research. Instead, we use data from that study, consisting of labels associated with Bitcoin addresses found on the dark web. Section 1.4 on the limitations of our work identifies this labeling of dark web data as a possible source of error.

For completeness, we briefly discuss the methodology used to collect the data from the dark web. See the description by Ghosh *et al.* [12] for further detail. An acquisition infrastructure was constructed to discover new onion websites, crawl their content, and integrate them into an indexed repository. This leveraged OnionCrawler, a fully automated crawling tool to identify new Tor onion domains. The dark web crawling system was run continuously, twice per day, to address diurnal patterns in onion site availability. If a string that matched the Bitcoin address format was found on a page, the address was associated with the onion (and its labels).

Seed data was used from previously published onion data sets, references to onions in a large collection of DNS resolver logs, and an open repository of (non-onion) web crawl data, called the Common Crawl. The automated categorization was used to label each onion with tags describing the activity found on its pages. We believe the tag provides a clear indication of the activity it refers to. We focused on the following tags in most of our analysis: `PONZI`, `MARKET`, `HUMAN_TRAFFIC`, `HACKER`, `DRUGS`, `CHILD_P`, `COUNTERFEIT`, `RANSOM`, `CASINO`, `NATIONALSEC`, `HOSTING_PROVIDER`, `PIRATEBAY`, `HITMAN`, `WEAPONS`, `JIHAD`, `EXPLOSIVES`, `CREDIT_CARD_FRAUD`, `DISCLOSURES`, `ANONYMOUS`, `PIRATE_BAY`, `MURDER`, `DOXBIN`, `ALPHA_MARKET`, `ESCORT`, `WIKILEAKS`, `DECRYPT_RANSOM`.

About 2.3 million candidate Bitcoin addresses were found in the dark web pages. As explained in Section 2.1, 0.2 million of these were strings that matched a regular expression for detecting the presence of an address on a web page but subsequently failed the Bitcoin checksum test [7]. After eliminating these false positives, we were left with 2,093,568 Bitcoin addresses. Table 1 shows how many of these addresses were associated with each of the 20 most frequent suspicious tags. More detail is provided next in Section 3.3.

	Number of Addresses	**Number of Neighbors**	**Owned Bitcoins**

CHILD_P	1,696	mean = 2,505
		median = 7
			mean = 94.41
			99% = 67.05

HUMAN TRAFFIC	1,876	mean = 2,350
		median = 7
			mean = 85.52
			99% = 60.16

MARKET	2,604	mean = 2,023
		median = 6
			mean = 63.2
			99% = 85.52

DRUGS	1,704	mean = 2,585
		median = 8
			mean = 94.11
			99% = 76.78

HACKER	1,817	mean = 2,433
		median = 8
			mean = 88.3
			99% = 65.17

PONZI	4,011	mean = 2,622
		median = 7
			mean = 5.63
			99% = 66.49

RANSOM	1,546	mean = 210
		median = 6
			mean = 0.12
			99% = 1.41

COUNTERFEIT	1,561	mean = 2,385
		median = 9
			mean = 68.80
			99% = 80.56

CASINO	1,421	mean = 2,891
		median = 7
			mean = 112.64
			99% = 89.15

NATIONALSEC	1,415	mean = 2,951
		median = 8
			mean = 112.56
			99% = 84.55

PIRATE_BAY	1,152	mean = 2,956
		median = 8
			mean = 7.27
			99% = 87.84

HOSTING_PROVIDER	1,276	mean = 2,741
		median = 8
			mean = 76.59
			99% = 89.32

CURRENCY	41,883	mean = 786
		median = 3
			mean = 10.47
			99% = 7.88

BITCOIN WALLET	2,985	mean = 5,630
		median = 4
			mean = 9.74
			99% = 154.93

FORUM SOFTWARE	1,473	mean = 1,095
		median = 3
			mean = 5.26
			90% = 52.32

REGISTRATION	1,332	mean = 3,710
		median = 11
			mean = 80.27
			99% = 88.68

HOSTING PROVIDER	1,317	mean = 4,041
		median = 21
			mean = 8.75
			99% = 0.0

ELECTRONICS	1,298	mean = 2,651
		median = 8
			mean = 75.23
			99% = 0.0

BLOG	1,220	mean = 2,798
		median = 8
			mean = 6.88
			99% = 84.11

NO_TAG	1,440,12	mean = 234
		median = 25
			mean = 0.8
			99% = 0.19

Table 1: Number of neighbors and bitcoins owned for active dark web addresses (limited to top 20 dark web tags considered suspicious or malicious)

Number of Tags (N)	**Number of Potential Addresses with N Tags**	**Number of Active Addresses with N Tags**	**Number of Tags (N)**	**Number of Potential Addresses with N Tags**	**Number of Active Addresses with N Tags**
1	143,130	41,783	14	696	69
2	45,152	2,319	15	2,210	78
3	24,331	1,519	16	175	30
4	68,236	290	17	76	17
5	4,761	67	18	127	36
6	1,382	32	19	16	3
7	176	57	20	258	53
8	598	151	21	417	344
9	482	60	22	187	13
10	284	44	23	551	41
11	290	66	24	147	33
12	485	31	25	203	65
13	841	75			

Table 2: Number of potential and active Bitcoin addresses on the dark web with N tags. Total number of addresses with tags is 296,069. Among them 47,697 are active and have tags we consider suspicious or malicious.

3.3 Bitcoin on the Dark Web

About 32% of the Bitcoin addresses that were found on the dark web – that is, 649,556 addresses – were labeled with tags. 1,444,012 addresses did not have any tags. A subset (of size 0.3 million) of the addresses were determined to be from mirrors of *Blockchain.info* explorer pages. These were eliminated from further analysis. The remaining addresses had a total of 49 unique tags associated with them.

We studied the addresses associated withe 20 most frequent suspicious tags. Table 1 reports the number of addresses, neighbors, and bitcoins associated with each of these tags. Though most addresses have few tags, some are labeled with many as seen in Table 2. Since some of the addresses may only be present on a dark web page without ever having been used, we performed the same analysis with active addresses. The results are reported in the same table to facilitate comparison. The histograms in Figures 5 and 6 depict the data from Table 2.

The 20 most frequently associated tags differ significantly when all Bitcoin addresses are considered versus when only active ones are analyzed, as can be seen from Figures 5 and 6. The inactive addresses that we eliminated appear to serve as decoys – that is, they are correctly constructed but unused. In the case of active addresses, the most frequently associated tag is "`CURRENCY`", indicating the prevalence of Bitcoin use in onions. We note that the histograms alone cannot be used to judge the significance of a topic on the dark web.

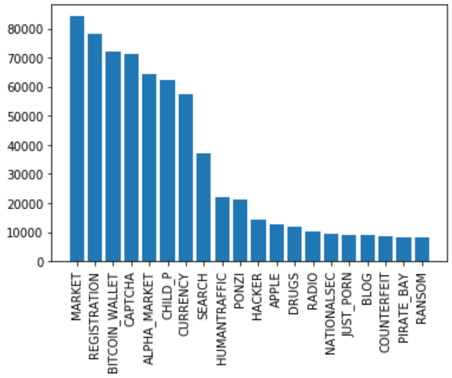

Fig. 5: Number of times a (top 20) tag appears with potential Bitcoin addresses on the dark web

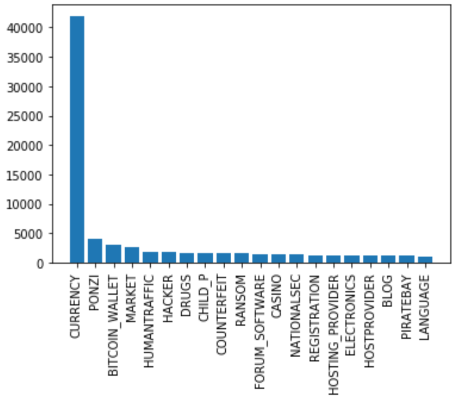

Fig. 6: Number of times a (top 20) tag appears with active Bitcoin addresses on the dark web

Of the 1,491,709 Bitcoin addresses found on dark web pages, only 47,697 had tags that we considered suspicious or malicious. The tags are shown in Table 1. To gain insight into suspicious activities that involve Bitcoin, our dark web analysis focused on addresses with these tags.

The number of addresses collected from the dark web each month grew initially, but then fell significantly. Figure 7 shows this for all Bitcoin addresses found on the dark web. To better understand usage, Figure 8 how many addresses appeared in a transaction on the blockchain for the first time in each month. The mid-2017 drops in the graphs may be explained by the seizure and shutdown of the Alphabay and Hansa dark web markets [14]. The final drop in early 2018 is due to our dark web data only extending to the end of 2017.

We note that this data must be interpreted with caution. In particular, there may be suspicious and malicious activity on the dark web that is not captured by the tags we use, creating false negatives. Further, dark web sites may reference benign addresses other than the mirrored *Blockchain.info* explorer pages that we were able to identify and exclude. This would have created false positives.

4 Detecting CoinJoins

CoinJoins are not first class primitives in Bitcoin. Hence, they cannot be definitely identified from inspecting the blockchain. In a minority of cases, a CoinJoin is listed explicitly on a web site, such as a discussion forum. In general, CoinJoin transactions must be detected using a heuristic based on their characteristics.

We build on an algorithm from Goldfeder *et al.* [13] that was designed to identify JoinMarket transactions. First, we identify the most common value (MCV) among the bitcoin amounts in the outputs of a transaction. The number of out-

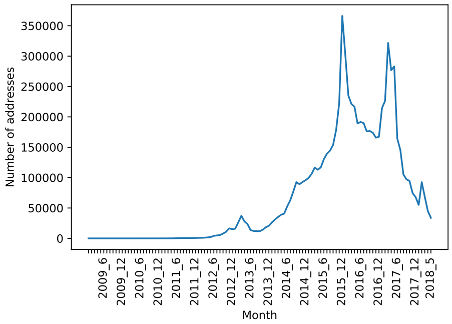

Fig. 7: Number of times a Bitcoin address found on the dark web appears in a transaction on the blockchain

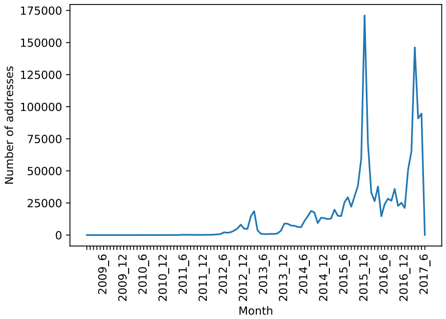

Fig. 8: Number of times a Bitcoin address found on the dark web *first* appears in a transaction on the blockchain

puts that have this value is considered to be the number of participants in such a transaction. In addition, the following three conditions must be satisfied:

1. The number of participants should be more than half the number of outputs. This is because up to half the outputs could be to change addresses.
2. The number of participants should be less than or equal to the number of inputs. This is because each participant must use at least one address as the source of their payment.
3. There should be at least one possible match between the inputs and the outputs, after considering the Bitcoin transaction fees and a liquidity payment (that is explained below).

Some services, such as JoinMarket, have users that continuously provide their bitcoins for use in CoinJoin transactions. These users serve as liquidity providers so that others who want to perform a CoinJoin can easily find peers with whom they can engage in such transactions. In exchange, such liquidity providers receive a percentage P of the MCV.

Our objective is then to find a set of disjoint input sets (S) so that each one can be matched with an output address (with a `change` address, denoted as *chng*). Each match translates to the following equation, with P being the max percentage of what a CoinJoin user pays for the liquidity provider, n being the number of participants, and MCV is the most common value in the transaction:

$$\forall \;\; inpt \in \;\; S : inpt \in [MCV \cdot (1-P) + chng, \quad MCV \cdot (1+(n-1)P) + chng + fees] \tag{1}$$

Since a liquidity provider may receive fees from $n-1$ other members, the upper limit of the interval contains a factor of $(1+(n-1)P)$. In our analysis, the payment P to the liquidity provider is allowed to be up to 2%.

4.1 Algorithm Details

The heuristic used to identify CoinJoin transactions is described detail in Algorithm 1. The general problem of finding a set S that satisfies equation 1 is NP-complete. It is harder than the problem of variable-sized bin-covering in the unit supply model. Approximation algorithms for generalized and variable-sized bin-covering do exist. However, the intervals in our setting are small enough that most instances can be easily eliminated. Indeed, in most cases the fees and the percentage given to the liquidity providers are usually very low when compared to the inputs. When this fact is taken into account, the problem becomes tractable.

Our heuristic solves the problem using the following steps. First, the outputs are computed by adding each change address to an MCV. Next, we perform a depth-first search of a tree. The nodes of the tree correspond to the output that is being taken into account, and a list of the remaining inputs. At a specific node, we look for all possible sets of remaining inputs that can satisfy equation 1 with respect to the output at the node. A new child node with the next output is created for each feasible set.

This approach avoids the exponential explosion that would result from exhaustively generating all possible combinations of sets of inputs. In practice, we found that when a solution exists, the depth-first search usually found it quickly. When there is no solution, the analysis must still traverse the entire tree.

In Algorithm 1, the function *subsets_between_two_values* recursively computes all subsets of the list provided as the first argument, subject to the constraint that their sum must be between the second and third arguments. This function's algorithm is also NP-hard. For instance, if the second argument is 0 and the third is $+\infty$, then it must return all possible combinations of elements of the input list (provided as the first argument). This will be a set of length 2^n.

Finally, we use simple rules to filter cases that are unlikely to be CoinJoins. One example is checking whether a transaction involves known addresses, such as those of SatoshiDice or other similar services. We also check that the fees are below a threshold fraction of the MCV. These rules allow us to reject many transactions early. These optimizations are not described in Algorithm 1.

Algorithm 1: CoinJoin Identification Heuristic

```
input  : A transactions $T$ with a list of inputs and a list of outputs (addresses
         + Bitcoin amounts)
output: A boolean indicating (if assigned True) that the function is classified
         as a CoinJoin

Find the most common value $MCV$ among outputs, and its number of
appearance $n\_participants$;
if $n\_participants < \left\lfloor \frac{length(outputs)+1}{2} \right\rfloor$ then
    return False;
end
if $n\_participants > length(inputs)$ then
    return False;
end
if $length(inputs) > 17$ then
    return True;
end
$new\_outputs \leftarrow$ array of length n_participants with value MCV in all cases;
$i \leftarrow 0$;
for $value$ in $outputs$ do
    if $value \neq MCV$ then
        $new\_outputs[i] \leftarrow new\_outputs[i] + value$;
        $i \leftarrow i + 1$;
    end
end
Sort inputs and new_outputs in decreasing order; //This does not really change
anything, it is just performed for convenience
$remaining\_inputs\_list \leftarrow [(new\_inputs, 0)]$; //This list contains sublists. Each
one of them is a node in the tree, representing a list of remaining inputs, and
the index of the output which has to be considered. We use this as a LIFO list
to make the tree search depth-first oriented.
$fees\_to\_provider \leftarrow max(2 * MCV/100, 0.0001BTC)$;
while $remaining\_inputs\_list$ is not empty do
    $remaining\_inputs, output\_index \leftarrow remaining\_inputs\_list.\mathrm{pop}()$;
    $current\_output \leftarrow outputs[output\_index]$;
    $lower\_limit \leftarrow current\_output - fees\_to\_provider$;
    $upper\_limit \leftarrow$
    $current\_output + fees + fees\_to\_provider * (n\_participants - 1)$;
    $new\_set\_of\_feasible\_inputs \leftarrow$
    $subsets\_between\_two\_values(remaining\_inputs,$
    $lower\_limit, upper\_limit)$;
    if $output\_index = length(new\_outputs) - 1$ and $new\_set\_of\_feasible\_inputs$
    is not empty then
        return True;
    end
    for $remaining\_inputs$ in $new\_set\_of\_feasible\_inputs$ do
        $remaining\_inputs\_list.append($
        $(remaining\_inputs, output\_index + 1))$;
    end
end
return False;
```

4.2 Analysis Results

The heuristic outlined above performs well in practice for CoinJoins with less than 18 inputs. Out of about 400,000 transactions that satisfy the two first conditions, the algorithm requires 180 seconds of computation time on a 2016 Macbook Pro. The heuristic identified 157 transactions that required deeper analysis.

More than 90% of all CoinJoin transactions have less than 18 inputs [13]. We automatically consider transactions with more than 17 inputs that pass the two first conditions to be CoinJoins. We found that 18% of all transactions considered to be CoinJoins by our heuristic have more than 17 inputs. Among transactions with fewer than 18 inputs, between 25% and 50% of those that satisfy the first two conditions also meet the third criterion. We concluded that between 4.5% and 9% of the CoinJoin transactions have more than 17 inputs. This is close to the result reported by Goldfeder *et al.* [13].

According to the heuristic, 114,925 transactions were CoinJoins. This represents 0.036% of the transactions that were analyzed. A total of 2,035,978 addresses were part of these CoinJoin transactions. This set of addresses was intersected with those found on the dark web, allowing us to conclude that over 2.3% of the addresses on the dark web have been CoinJoin participants. In contrast, this is only true for 0.4% of all Bitcoin addresses. We could not identify a specific dark web category that used more CoinJoins than others.

Dark web addresses appear to be 5 times more likely to participate in CoinJoin transactions. We noted with interest that only 2.3% of addresses appearing on the dark web have participated in CoinJoins, since that is a small fraction. It is conceivable that this is due to the use of alternative mixing approaches.

5 Bitcoin Neighborhood Analysis

We first report our findings from analyzing the activity of all addresses in the Bitcoin blockchain. After this, we focus on the subset of addresses that have participated in transactions as well as appeared on the dark web.

5.1 Across the Blockchain

A wide range of behaviors were exhibited by the 397,016,130 Bitcoin addresses that we analyzed. To characterize them, we studied how many other addresses an address has transacted with, how many transactions it has been involved in, the amount of bitcoin that has flowed into it, from it, and is owned by it. Table 3 reports our findings.

To identify the neighbors of addresses we constructed the transaction graph, with one vertex per address, and undirected edges between two addresses if they are both listed in (at least) one transaction, with one as a sender and the other as a receiver. As noted earlier, we excluded CoinJoins before inferring the neighbor relationship. We also do not consider two senders (or receivers) in the same transaction to be neighbors.

For all (397,301,155) addresses with at least one transaction	**Number of Neighbors**	**BTC In**	**BTC Out**	**BTC Owned**	**Number of Tx's**
Mean	11.92	10.0687	10.03	0.04	3.62
Std	372.26	989.07	987.60	22.55	316.79
Median	3	0.05	0.048	0.00	2
Max	4,586,602	9,351,251	9,356,600	175,236	3,195,815
Min	1	0	0	0	1
Percentile 90%	19	4.37	4.31	0	2
Percentile 99%	137	126.61	126.11	0.03	24
Percentile 99.9%	758	965.99	963.64	1.99	197
Percentile 99.99%	2,846	8,329.38	8,314.63	41.10	1,059
Number of addresses that hold more than 99% of bitcoin: 2,266,265					

Table 3: Characterization of all addresses in terms of neighbors, transactions, and amount of bitcoin in/out/owned. (BTC = Bitcoin, Tx = transaction)

Number of Addresses with ...		
Less than 10 neighbors	34,013,8543	85.67%
More than 1000 neighbors	257,925	0.06%
More than 10000 neighbors	6,178	0.00156%

Table 4: Breakdown of neighbor count of all (397,301,155) addresses with at least one transaction

The standard deviation of the number of neighbors per address is large, significantly exceeding the 99% percentile. This indicates that the extreme values are located far from the average. This is also confirmed by the fact that the mean is much larger than the median. While 50% of the addresses have transacted with less than 3 other addresses, approximately 6,000 addresses have more than 10,000 neighbors. A few addresses have more than a million neighbors. The latter addresses are probably not manually controlled by humans. Most outliers are addresses that come from exchange services, which are involved in many transactions.

The results for the number of transactions exhibit similar characteristics, with a large standard deviation. The mean and the median are closer. Most addresses are involved in few transactions. Specifically, the number of transactions is smaller than the number of neighbors for most addresses.

The quantity of bitcoin owned by each address also varies widely. The average is 0.04 bitcoin, while the standard deviation is 500 times larger. Most addresses have no bitcoins left. This is explained by the fact that in a transaction the sender needs to use all the bitcoins from each past input referenced. If there is an excess it must either be sent to a change address or it will become part of the fee to the miner.

We found that the addresses that owned the largest quantities of bitcoin corresponded to the ones listed on websites that track wallet addresses with large holdings [8]. Most such addresses belong to exchanges. An exception is “1KAt6STtisWMMVo5XGdos9P7DBNNsFfjx7”, which was ranked sixth at the time of writing. Each of the top six addresses own more than 0.5% of the total number of bitcoins.

5.2 Addresses on the Dark Web

The statistics for the addresses used on the dark web differ significantly from those of addresses across the entire blockchain. On the dark web, 90% of the Bitcoin addresses have transacted with up to 400 other addresses, participated in over 200 transactions, and been involved with 12 bitcoins. Across the entire blockchain, 90% of the addresses have transacted with fewer than 20 other addresses and only dealt with amounts totaling 4 bitcoins. The differences can be seen in Tables 3 and 6.

For (1,491,709) dark-web addresses with at least one transaction	**Number of Neighbors**	**BTC In**	**BTC Out**	**BTC Owned**	**Number of Tx's**
Mean	255.09	153.97	152.91	1.12	143.68
Std	3,723.62	14,554.28	1,4536.67	239.37	5,102.65
Median	23	0.10	0.097	0.00	4
Max	2,277,764	9,351,251	9,350,599	175,236	3,195,815
Min	1	0	0	0	1
Percentile 90%	426	12.85	12.61	0.00089	220
Percentile 99%	27,45	375.15	366.29	0.21	1459
Percentile 99.9%	20,891	10577.64	10,287.04	40.00	7,674
Percentile 99.99%	114,947	226,413.71	226,405.41	800.00	8,3299
Number of addresses that hold more than 99% of the bitcoin (limited to addresses found on the dark web): 2,828					

Table 5: Characterization of Bitcoin addresses found on the dark web, in terms of neighbors, transactions, and amount of bitcoin in/out/owned. (BTC = Bitcoin, Tx = transaction)

The same analysis for active addresses found on the dark web indicates that they transact more than addresses on the Bitcoin blockchain. This can be seen by comparing Table 3 with Tables 5, 7, and 8. The average amount of bitcoin owned is also larger for addresses found on the dark web. In addition, 99% of the coins touched by dark web addresses are owned by just 2,828 dark web addresses.

These results need to be interpreted with caution. The addresses found on the dark web were publicly accessible. This may have skewed the analysis in favor of addresses that are more popular and frequently used. This could explain the significant difference in the characteristics of addresses found on the dark web

Number of Addresses with	**Absolute Number**	**Percentage**
Less than 10 neighbors	597,744	40.09%
More than 1000 neighbors	61,330	4.11%
More than 10000 neighbors	3,244	0.22%

Table 6: Breakdown of transaction neighbor counts for active addresses found on the dark web

in comparison to those across the entire blockchain. This may also account for the fact that 10% of the Bitcoin addresses found on the dark web participate in more than 220 transactions each.

The difference in the number of neighbors per address is even larger, this can be explained by the observation that dark web addresses are more likely to use mixing methods (as our CoinJoin analysis indicated), and those methods will increases the neighbors in our analysis. Also, the sum of the bitcoins owned by these addresses represent less than 10% of all bitcoins. This number is far from exact, and is in fact much smaller, as several of the richest addresses have been cited in forms and discussion on the dark web, so can be found in this set.

For (35,492) dark web addresses with at least one CoinJoin transaction	**Number of Neighbors**	**BTC In**	**BTC Out**	**BTC Owned**	**Number of Tx's**
Mean	1,745	2,618	2,612	7.07	1,429
Std	12,726	71,325	71,335	616	28,479
Median	159	1.99	1.97	0	48
Percentile 90%	2341	100	100	0	732

Table 7: Characterization of Bitcoin addresses found on the dark web, with at least one CoinJoin transaction, in terms of neighbors, transactions, and amount of bitcoin in/out/owned. (BTC = Bitcoin, Tx = transaction)

Participation in a CoinJoin is unusual (as can be seen in the statistics reported in Section 4). This motivated us to study Bitcoin addresses found on the dark web that have participated in at least one CoinJoin transaction. Table 7 reports the results. In particular, the mean and standard deviation of both the number of neighbors and exchanged Bitcoins are higher than for addresses that do not participate in a CoinJoin.

Assume that the more an address participates in transactions, the higher the chance that it will be part of a CoinJoin. This would explain why the Bitcoin addresses that appear most often on the dark web are likely to be part of CoinJoin transactions. However, we found that even Bitcoin addresses on the dark web that appear at the median frequency are more likely to have participated in

For (7,713) dark web addresses with at least one malicious tag	**Number of Neighbors**	**BTC In**	**BTC Out**	**BTC Owned**	**Number of Tx's**
Mean	1,465	8,234	8,213	22.8	1,527
Std	1,8351	152,755	15,2754	1,269	22,023
Median	6	2.55	2.48	0	3
Percentile 90%	347	295	295	0.004	217

Table 8: Characterization of Bitcoin addresses found on the dark web, with at least one malicious tag, in terms of neighbors, transactions, and amount of bitcoin in/out/owned. (BTC = Bitcoin, Tx = transaction)

CoinJoins. An explanation supported by the data is that transactions associated with Bitcoin addresses found on the dark web involve larger amounts, motivating increased caution.

We stress that the dark web is also used for several legitimate activities. As an additional filter for teasing out real suspicious or malicious activities, we focus on the set of dark web addresses that contained tags associated with what we judged as the most suspicious (and in cases very obvious malicious) activities, i.e., addresses containing at least one tag from the following list: `PONZI, MARKET, HUMANTRAFFIC, HACKER, DRUGS, CHILD_P, COUNTERFEIT, RANSOM, CASINO, NATIONALSEC, HOSTING_PROVIDER, HITMAN, WEAPONS, JIHAD, EXPLOSIVES, CREDIT_CARD_FRAUD, DISCLOSURES, ANONYMOUS, PIRATE_BAY, WIKILEAKS, MURDER, MARKET, ESCORT, DECRYPT_RANSOM`. We note that some addresses associated with some of these tags are not active on the blockchain so not all these tags show up in all our analysis.

We notice that these addresses (see results in Table 8) do not do many more transactions that the whole dark web address set, but these figures remain much bigger than the ones obtained from regular addresses. Moreover, the BTC amounts these dark web addresses handle are even larger. Even the median has a higher value. As the size of the set is small, these addresses are probably among the most well-known addresses used for malicious activities, and a lot of them are used extensively.

6 Future Work

This study provides a quantitative characterization of suspicious and malicious activities involving Bitcoin. In addition to addressing the limitations discussed in Section 1.4, we envision the following avenues of research in future work.

1. Similar analyses could be performed for other popular cryptocurrencies, such as Bitcoin forks, Ethereum, and Litecoin. In particular, comparing results from other cryptocurrencies to those from Bitcoin may yield new insights.
2. Augmenting the data sets used in this study with ones that may help attribute malicious activities to geographic location. This could include data

mapping addresses to well-known wallets or entities, as well as to IP addresses (for which geolocation data is typically available).

3. Studying cross-cryptocurrency transaction activity could enable detection of synchronized addresses. This may provide a new means for detecting when seemingly unrelated addresses are controlled by the same user or pertain to coordinated activity. Detecting synchronized activity may also offer insight into significant events in the history of cryptocurrencies.

References

1. Elli Androulaki, Ghassan Karame, Marc Roeschlin, Tobias Scherer, and Srdjan Capkun, **Evaluating user privacy in Bitcoin**, *Financial Cryptography and Data Security, Lecture Notes in Computer Science, Vol. 7859, Springer*, 2013.
2. Blockchain.info Bitcoin explorer, `https://www.blockchain.com/explorer`
3. Blockchain Intelligence Group, `https://blockchaingroup.io/`
4. Jeremiah Bohr and Masooda Bashir, **Who uses Bitcoin? An exploration of the Bitcoin community,** *12th International Conference on Privacy, Security, and Trust*, 2014.
5. Joachim Breitner and Nadia Heninger, **Biased nonce sense: Lattice attacks against weak ECDSA signatures in cryptocurrencies**, *23rd International Conference on Financial Cryptography and Data Security*, 2019.
6. Version 1 Bitcoin Addresses, `https://en.bitcoin.it/wiki/Technical_background_of_version_1_Bitcoin_addresses`
7. Bitcoin forum: Validating Bitcoin addresses, `https://bitcointalk.org/index.php?topic=1026.0`
8. Largest bitcoin holdings, `https://bitinfocharts.com/top-100-richest-bitcoin-addresses.html`
9. Chainanalysis Platform, `https://www.chainalysis.com/`
10. CipherTrace Platform, `https://ciphertrace.com/`
11. Elliptic Platform , `https://www.elliptic.co/`
12. Shalini Ghosh, Ariyam Das, Phil Porras, Vinod Yegneswaran, Ashish Gehani, **Automated categorization of Onion sites for analyzing the Darkweb ecosystem**, *23rd ACM International Conference on Knowledge Discovery and Data Mining*, 2017.
13. Steven Goldfeder, Harry Kalodner, Dillon Reisman, and Arvind Narayanan, **When the cookie meets the blockchain: Privacy risks of web payments via cryptocurrencies**, *18th Privacy Enhancing Technologies Symposium*, 2018.
14. Andy Greenberg, **Global police spring a trap on thousands of dark web users**, *Wired*, `https://www.wired.com/story/alphabay-hansa-takedown-dark-web-trap/`, 20th July, 2019.
15. Seunghyeon Lee, Changhoon Yoon, Heedo Kang, Yeonkeun Kim, Yongdae Kim, Dongsu Han, Sooel Son, and Seungwon Shin, **Cybercriminal Minds: An investigative study of cryptocurrency abuses in the Dark Web**, *26th Annual Network and Distributed System Security Symposium (NDSS)*, 2019.
16. Sarah Meiklejohn, Marjori Pomarole, Grant Jordan, Kirill Levchenko, Damon McCoy, Geoffrey Voelker, and Stefan Savage, **A fistful of bitcoins: Characterizing payments among men with no names**, *13th ACM Internet Measurement Conference (IMC)*, 2013.

17. Andrew Miller, Malte Moser, Kevin Lee, and Arvind Narayanan, **An empirical analysis of linkability in the Monero blockchain**, *arXiv:1704.04299*, 2017.
18. Monero, `https://getmonero.org/`
19. Till Neudecker and Hannes Hartenstein, **Could network information facilitate address clustering in Bitcoin?**, *4th Workshop on Bitcoin and Blockchain Research*, 2017.
20. Stephen Ranshous, Cliff Joslyn, Sean Kreyling, Kathleen Nowak, Nagiza Samatova, Curtis West, and Samuel Winters, **Exchange pattern mining in the Bitcoin transaction directed hypergraph**, *4th Workshop on Bitcoin and Blockchain Research*, 2017.
21. Dorit Ron and Adi Shamir, **Quantitative analysis of the full Bitcoin transaction graph**, *17th International Conference on Financial Cryptography and Data Security, Lecture Notes in Computer Science, Vol. 7859, Springer*, 2013.
22. Tim Ruffing, Pedro Moreno-Sanchez, Aniket Kate, **CoinShuffle: Practical decentralized coin mixing for Bitcoin**, *19th European Symposium on Research in Computer Security*, 2014.
23. Michele Spagnuolo, Federico Maggi, and Stefano Zanero, **BitIodine: Extracting intelligence from the Bitcoin network**, *Financial Cryptography and Data Security, Lecture Notes in Computer Science, Vol. 8437, Springer*, 2014.
24. Zcash: Privacy-protecting Digital Currency, `https://z.cash/`