User Privacy in the Public Bitcoin Blockchain

Jaume Barcelo

***Abstract*—Bitcoin is a peer-to-peer electronic cash system that maintains a public ledger with all transactions. The public availability of this information has implications for the privacy of the users. The public ledger consists of transactions that transfer funds from a set of inputs to a set of outputs. Both inputs and outputs are linked to Bitcoin addresses. In principle, the addresses are pseudonymous. In practice, it is sometimes possible to link Bitcoin addresses to real identities with the consequent privacy leaks. The possibilities of linking addresses to owners are multiplied when addresses are reused to receive funds multiple times. The reuse of addresses also multiplies the amount of private information that is leaked when an address is linked to a real identity. In this work we describe privacy-leaking effects of address reuse and gather statistics of address reuse in the Bitcoin network. We also describe collaborative (CoinJoin) transactions that prevent the privacy attacks that have been published in the literature. Then we analyze the Blockchain to find transactions that could potentially be CoinJoin transactions.**

***Index Terms*—Bitcoin, cryptocurrency, privacy, address reuse, CoinJoin**

I. Introduction

BITCOIN is a peer-to-peer electronic cash system [1] that maintains a public ledger with all transactions. All the transactions need to be available to the peer-to-peer network that guarantees the security of the system. Any full node of the network stores in a database all the transactions in the history of Bitcoin. This database is typically referred to as the *Blockchain*.

With traditional cash, there is no public record of all the transactions. And the traditional banking system keeps the transactions of their customers private. The public availability of the *Blockchain* represents a novelty and its privacy implications are worth studying. Ideally, any new payment system should offer privacy guarantees at least as good as traditional systems.

To receive a Bitcoin payment, the payee must provide the payer a Bitcon address to which the payment will be sent. Both the Bitcoin community and previous research [1], [2] agree that address reuse is in general a bad practice. It is recommended to generate a new address for each payment to be received. These addresses that are used only once are called disposable addresses.

In this paper we first review the basic elements of the Bitcoin system necessary for the subsequent discussion. Then we discuss the necessity of avoiding address reuse to prevent unnecessary privacy leakage. After that, we analyze the Blockchain to determine to which extent address reuse occurs in the Bitcoin community. As disposable addresses do not offer total protection against privacy leakage, we detail the remaining risks as well as possible solutions.

The author is with Universitat Pompeu Fabra. Roc Boronat 138, 08018 Barcelona, Catalunya, Spain. E-mail: jaume.barcelo@upf.edu

II. Bitcoin and Address Reuse

Bitcoin is a protocol, a network, and an Internet currency unit. We capitalize the word when we refer to either the protocol or the network.

Transactions are a fundamental element of Bicoin. Payments require transactions, and these transactions are shared with the network and securely stored in the Blockchain. Each transaction consumes some inputs and creates some outputs. The inputs and outputs have a value in bitcoins. For a regular transaction to be valid, it is required that the total value of the outputs does not exceed the total value of the inputs.

The outputs of one transaction can be used as inputs of later transactions. Critically, each output can be used only once. In the Bitcoin jargon, available outputs are also referred to as available coins. The network does not accept transactions that try to spend coins that have been spent before.

An example transaction is presented in Listing 1. Long hexadecimal strings have been trimmed to save space. A transaction is identified by a hash, which is the first field. This particular transaction has a single input (“in”) which is the first output (“n”:0) of a previous transaction with a hash starting with a777.

The example transaction has two outputs worth 22 bitcoins and approximately 0.87 bitcoins. The “scriptSig” field in the input contains the signature required to spend the coins in the input. The “scriptPubKey” field in the outputs describes the signature required to spend those outputs.

Listing 1. Example Bitcoin Transaction

```
{
"hash":"1093[...]",
  "ver":1,
  "vin_sz":1,
  "vout_sz":2,
  "lock_time":0,
  "size":258,
  "in":[
    {
      "prev_out":{
        "hash":"a777[...]",
        "n":0
      },
      "scriptSig":"3045[...]"
    }
  ],
  "out":[
    {
      "value":"22.00000000",
      "scriptPubKey":"OP_DUP OP_HASH160 17ed[
          ...] OP_EQUALVERIFY OP_CHECKSIG"
    },
    {
      "value":"0.87213300",
      "scriptPubKey":"OP_DUP OP_HASH160 9319[
          ...] OP_EQUALVERIFY OP_CHECKSIG"
```

```
      }
    ]
}
```

Other important elements of Bitcoin are public/private asymmetric cryptographical keys. The public key is hashed and coded with some redundancy into base58 addresses. These addresses are alphanumeric chains that can be used to receive funds. The outputs of a transaction can be sent to an address, which is simply a convenient representation of a public key.

In order to spend an output, it is necessary to offer proof of ownership of the address and, consequently, of the output. This proof is the evidence of knowledge of the private key corresponding to the address. A user willing to spend an output sent to a given address must provide the public key that hashes to that address and a valid signature. The signature can be generated only by the owner of the private key.

Although address reuse is discouraged, it is technically possible and easy to use an address multiple times to receive transaction outputs. The owner of the address' private key will be able to spend each of those outputs once. Another particularity of bitcoin is that there is no practical limit on the number of addresses that can be generated. Each address is 160 bits long and therefore, even though the addresses are randomly generated, collisions do not occur in practice. It is possible for users to generate new (disposable) addresses for each incoming payment. This is the recommended practice.

There is particular form of output, the *change* output, that is relevant for the later discussion. A user willing to send a Bitcoin payment needs to combine in a transaction a number of inputs of value equal or larger than the desired payment. If the value of the inputs is larger than the output, it is likely that the payer does not want to loose the difference. The difference between the total value of the inputs and amount to be payed is called the change. In order to keep the change, the payer creates a transaction with two outputs: the actual payment and the change. The payment is sent to the payee and the change is sent to an address controlled by the payer. The second output of the Listing 1 is probably an example of a change output.

A. Address Reuse

It may be tempting to use Bitcoin addresses like regular bank account numbers. That is, using a single address to receive many, or even all, payments.

The difference is that bank account transactions and balances are kept private by the bank. In contrast, all the transactions and balances in Bitcoin are publicly available to anyone with Internet access. In principle, Bitcoin address are pseudonymous and are not necessarily linked to real identities. In practice, the pseudonymity can be a very thin line of protection and there are many opportunities to link Bitcoin addresses to owners.

When a particular address is linked to its owner, all the transactions related to that address are also linked to the owner. Address reuse increases the chances that an address is linked to its owner. Furthermore, address reuse also increases the amount of information that is revealed when the address is linked to its owner. The original Bitcoin paper [1] discourages address reuse.

B. Address Reuse in the Blockchain

Despite the fact that address reuse was discouraged since the inception of Bitcoin, we can find evidence of address reuse in the Blockchain. We use Obelisk and Libbitcoin to download and query the Blockchain for evidence of address reuse. Both Obelisk and Libbitcoin are open source projects under heavy development. Our source code to generate the data and the plots presented in the paper is also available in github[1]. We sweep all the transactions in the Bitcoin history until January 2014 and, for each address, we count how many times it appears as an output of a transaction.

Some statistics of the distribution are presented in Table I. A histogram of the data is presented in Fig. 1. For readability reasons, we only show the first 100 positions of the histogram. The distribution has a long tail and some addresses are used over one million times.

C. Wikileaks Privacy Leakage

Wikileaks funding campaign is an example of address reuse. At the time of this writing, Wikileaks donation webpage offers by default a reused address. It also offers the donors the possibility of generating a new (disposable) address by simply clicking a button. If the donor uses the default reused address, it is likely that some private information is leaked.

The Blockchain contains all the details of the transactions involving the reused public address. A Blockchain explorer website (such as `blockchain.info`) can be used to browse all those details. At the time of this writing, Wikileak's public address has received over 3,858 bitcoins in 2477 transactions. The source addresses of each incoming transaction are also public.

These source addresses are pseudonymous and are not, in principle, linked to their owners. Nevertheless, there are several situations in which the association between addresses and owners is possible. The most obvious example is when the owner publicizes the address in a personal webpage, blog, forum or any other Internet site. A second example is when the owner of the address uses it to make or receive non anonymous payments. The other party involved in the payment learns the identity of the owner of the address.

Combining the facts that some of the addresses donating to Wikileaks are public and some of the owners of those addresses are known, the result is that the identity of some Wikileaks donors is leaked.

There are advanced techniques that we review later in this work that can result in increased privacy leakage. These techniques exploit the usual behaviour of Bitcoin *wallets* and *change addresses*.

D. Bitcoin Wallets

From the previous section it should be clear that address reuse has negative consequences to the user's privacy. A partial solution to Bitcoin's privacy weaknesses is the use of disposable addresses. Disposable addresses are used only

[1] https://github.com/jbarcelo/txfillstat

TABLE I
ADDRESS REUSE STATISTICS

Mean	3.18
Min	1
25th perc.	1
50th perc.	1
75th perc.	1
Max	1,238,931
Number of addresses	12,963,199
Number of uses	41,244,997
Addresses used once	10,476,899
Addressed used twice	1,397,373
Used over 100 times	25,004

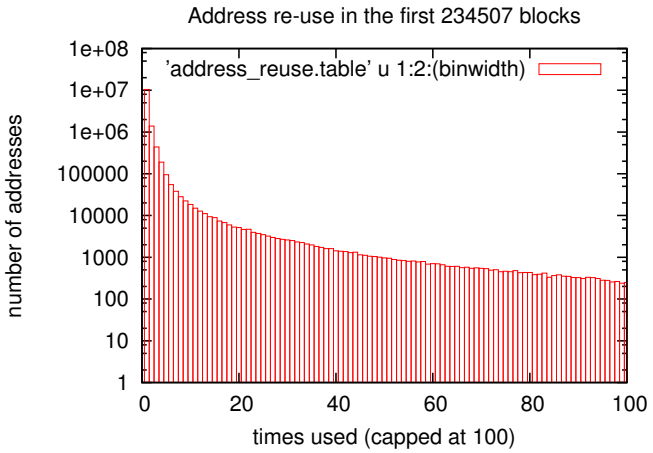

Fig. 1. Number of addresses for reuse factor from 1 to 100. Note the log scale.

once and therefore make it more difficult to link different transactions to the same user.

The reference Bitcoin peer implementation, *bitcoind* and many other software packages make it possible for the Bitcoin user to comfortably handle a large number of addresses. The software that takes care of Bitcoin addresses is called a Bitcoin wallet. This software generates as many addresses as needed. Some of these addresses are used for receiving payments while others are used to receive change outputs. The software does not reuse change addresses and therefore, if the user is cautious enough to avoid address reuse for payment reception, addresses are used a single time to receive funds.

The software wallet also computes the sum of the funds available in all the addresses to present a total balance to the user. The wallets also assists the users in the creation of transactions. Among all the outputs available to the user, the wallet picks those that are used as inputs of the new transaction. Change addresses are also handled transparently for the user.

The software keeps all the private keys of the user. These keys are typically in an encrypted file with a single password that the user needs to remember.

E. Wallet Privacy Leakage

Common operation of Bitcoin wallets may leak some information about the users. This leakage has been exploited in the past [2]. The first form of privacy leakage is when a wallet combines different outputs as inputs for a single transaction. This combination is necessary when the software creates a transaction for a payment value that exceeds the value of the individual available coins available in the wallet. When the transaction is broadcast, an attacker can infer that all the inputs of the transaction and their associated addresses belong to the same user.

It is important to highlight at this point that the Bitcoin protocol does not require that all the inputs belong to the same wallet. This is simply a common practice in widespread software wallets. Therefore it is not a protocol vulnerability, but an implementation vulnerability.

A second possible attack involves change addresses. There are techniques to try to infer which of the outputs of a particular transaction is a change output. For example, in a transaction which presents two outputs and one of them is smaller than any of the inputs, it can be assumed that the small output is a change output. Again, this is simply an assumption that relies on the way that popular wallets operate today and by no means is an inherent restriction of the protocol. If a change output of a transaction is identified and later used as an input in another transaction, an attacker may infer that all the inputs of the first and second transaction belong to the same wallet.

As most of the transactions have a change output and this output can be an input of another transaction, a number of transactions can be chained together in a privacy leaking chain.

Privacy invading techniques in the literature rely on heuristics that take advantage of *idioms of use*. It is noted in [2] that a change of current Bitcoin practices may invalidate current privacy attack methods.

To the best of the author's knowledge, techniques currently available in the literature cluster addresses belonging to the same wallet but do not attempt to link different wallets of the same user. Therefore, the use of different wallets may offer partial protection against published privacy attacks.

F. Enhanced Privacy Measures

The privacy attacks described in the previous subsection take advantage of the current behaviour of software wallets that create transactions in which all of the inputs belong to the same user. The attacks can be disrupted by introducing transactions in the Blockchain that intendedly violate the assumptions on which the privacy attacks rely. If different users collaborate in creating a transaction and the coins that are used as inputs belong to the different users, the attacks currently available in the literature will no longer be reliable. These transactions create an additional layer of confusion which results in additional protection for the privacy of the users.

The combination of inputs of different users has been termed *CoinJoin* by the Bitcoin community. In the simplest case, two users combine two inputs to make two payments in

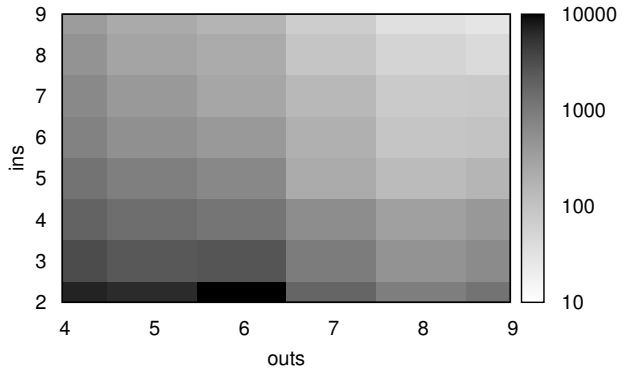

Fig. 2. Number of potential CoinJoin transactions for a given number of inputs and outputs.

a single transaction. A simple CoinJoin transactions has two or more inputs from two different users, two outputs which are the two payments, and two change addresses for the two payers. If the payments are of similar amounts, a payee cannot determine which inputs belong to which payer. Similarly, a Blockchain observer cannot link different inputs as belonging to the same user.

There is no limit in the number of users that can combine their payments in a single Bitcoin transaction. In fact, a larger number of collaborators makes it even more difficult for the attacker to extract information from the Blockchain. The only limitation is that to participate in a CoinJoin transaction, it is necessary to find other users that want to make a payment at the same time. At the time of this writing, mainstream wallets do not offer automated tools for finding collaborators for a CoinJoin transactions and therefore CoinJoin is seldomly used.

We scan the Blockchain for transactions with two or more inputs and four or more outputs as they are possible CoinJoin transactions. In Fig. 2 we present results for a number of inputs and outputs lower than 9.

III. Related Work

The dangers of address reuse are mentioned in the original Bitcoin paper [1]. An insightful analysis of the anonymity of the Bitcoin network is presented in [3]. Among other aspects, this paper covers the mechanisms to cluster public keys belonging to the same user. The paper also shows how external information can be used to identify the users. In some cases, the information publicly available on the Internet can be enough to identify users. Centralized servers such as exchanges are in an even better position to de-anonymize the network.

A more recent work is [2], which describes in more detail the privacy attacks and the encountered difficulties in the process of de-anonymizing the Blockchain. The authors of this paper actively participate in the Bitcoin ecosystem to be able to tag the most important services in the graphs derived from the Blockchain. The Wikileaks case that is mentioned in the present work is also considered in [2].

Both [3] and [2] pay attention to bitcoin thefts. In some cases, the authors can track the money until a fraction of the theft reaches an exchange.

A simulation study on Bitcoin privacy [4] recommends the use of disposable addresses by the merchants for increased privacy protection. The paper also mentions the collaborative construction of transactions and considers that its widespread use is unlikely.

IV. Conclusion

The public availability of all Bitcoin transactions poses privacy-related challenges on the network. In this work we discuss address reuse and CoinJoin transactions as two key privacy elements in Bitcoin. We present data regarding address reuse in the blockchain as well as possible CoinJoin transactions. The Bitcoin community is advancing in the understanding of privacy threats and developing mitigating techniques.

Acknowledgment

The author would like to thank Amir Taaki and the other developers of libbitcoin and obelisk for providing the code used to gather the statistics.

References

[1] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” 2008.
[2] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, and S. Savage, “A fistful of bitcoins: characterizing payments among men with no names,” in *Proceedings of the 2013 Internet Measurement Conference*. ACM, 2013, pp. 127–140.
[3] F. Reid and M. Harrigan, “An analysis of anonymity in the bitcoin system,” in *Security and Privacy in Social Networks*. Springer, 2013, pp. 197–223.
[4] E. Androulaki, G. Karame, M. Roeschlin, T. Scherer, and S. Capkun, “Evaluating user privacy in bitcoin.” *IACR Cryptology ePrint Archive*, vol. 2012, p. 596, 2012.