Bitcoin Transaction Graph Analysis

Michael Fleder
mfleder@mit.edu

Michael S. Kester
kester@eecs.harvard.edu

Sudeep Pillai
spillai@csail.mit.edu

January 3, 2014

1 Introduction

Bitcoins have recently become an increasingly popular cryptocurrency through which users trade electronically and more anonymously than via traditional electronic transfers. Bitcoin's design keeps all transactions in a public ledger. The sender and receiver for each transaction are identified only by cryptographic public-key ids. This leads to a common misconception that it inherently provides anonymous use. While Bitcoin's presumed anonymity offers new avenues for commerce, several recent studies raise user-privacy concerns. We explore the level of anonymity in the Bitcoin system. Our approach is two-fold: (i) We annotate the public transaction graph by linking bitcoin public keys to real people - either definitively or statistically. (ii) We run the annotated graph through our graph-analysis framework to find and summarize activity of both known and unknown users.

2 Contributions

We present a bitcoin transaction-graph-annotation system in two parts. First, we developed a system for scraping bitcoin addresses from public forums. Second, we include a mechanism for matching users to transactions using incomplete transaction information. For example, suppose we hear Bob say to Alice: "I sent you $100 in bitcoins yesterday at noon"; though we don't know the exact time of the transaction (since "at noon" could easily mean 11:50 or 12:10) or the exact amount in bitcoins (exchange rates fluctuate significantly), we can generate candidate transaction matches and associated matching probabilities.

We also present a graph-analysis framework capable of tracing and clustering user activity. For example, our framework suggested the FBI seizure of Silk Road assets as "interesting" activity on 10/25/2013 without prior knowledge of the FBI or Silk Road public keys. Furthermore, our system found close links between Silk Road and real users identified with our annotation system.

3 Background

Recently, several research studies [3, 2, 4] have suggested the potential privacy limitations with bitcoin transactions. [3] investigates an alleged theft by leveraging external sources of information and combining them with techniques such as context discovery and flow analysis. [4], on the other hand analyzes statistical properties of the transaction graph to answer questions about typical user behavior, spending/acquiring habits, and flow of bitcoins between multiple accounts belonging to the same user. Realizing the need for stricter privacy in the bitcoin graph, the authors in [2] suggest an extension to bitcoin that augments the protocol to allow for fully anonymous currency transactions.

4 Threat Model

4.1 Attacker Goal: Tie "real" names to transactions

The "real" name here may be a person's true name or username from an online public forum (or any other public data source). The goal is to associate numerous unrelated cryptographic IDs with an actual user.

4.2 Attacker Capabilities

First, an attacker has access to all public information including forums, donation sites, and public social networks from which one can scrape bitcoin addresses that have been intentionally or unintentionally1 divulged. That is, an attacker may scrape ("real" name, public key) pairs from web sites.

Second, an attacker may also "overhear" imprecise transaction information from known users. For example, an attacker may have overheard "Alice, its Bob. I sent you $100 bitcoins yesterday at noon." That is, an attacker may hear ("real" name, some rough transaction info) pairs.

Intentional Leak **Unintentional Leak**

Figure 1: On the **left**: Annotated Transaction Graph. On the **right**: Silk Road owner Dread Pirate Roberts unintentionally reveals his public key in an online forum bitcointalk.org

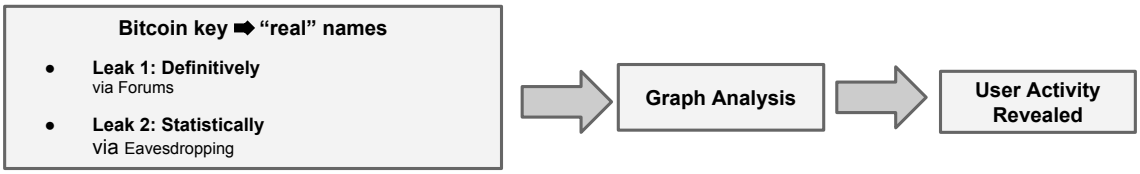

Figure 2: Attack Model

5 Implementation

In this section we shall describe the several steps involved in revealing bitcoin user activity information by leveraging publicly available transaction information. As described in the earlier section, we investigate both, statistical and definitive approaches.

5.1 Pre-processing

As a precursor to both the above mentioned approaches, the raw transactions have to be extracted from the full blockchain. As of Dec 13, 2013, approximately 275,000 blocks have been mined in the bitcoin block chain. Each block contains on the order of hundreds of transactions. We describe the blockchain parsing in the following section.

5.1.1 Blockchain Parsing

While the standard bitcoin client `bitcoin-0.8.5`[1] automatically downloads the whole blockchain in a P2P fashion, we noticed a significantly reduced network download rate which prompted us to download a torrent [2] from [3] quickly. The remaining blocks were updated automatically by the bitcoin client after which it was

[1] bitcoin-0.8.5 `http://bitcoin.org/en/download`
[2] `http://sourceforge.net/projects/bitcoin/files/Bitcoin/blockchain/bootstrap.dat.torrent/download`
[3] `http://sourceforge.net/projects/bitcoin/files/Bitcoin/blockchain/`

indexed. While previous works [2, 4] employed a forked version of `bitcointools` [4], the newer bitcoin clients indexed the full blockchain using LevelDB instead making the publicly available `bitcointools` obsolete. Instead, we used Armory [5] to parse through the blockchain, and wrote wrapper classes that extracted the relevant information required to construct the transaction graph.

5.1.2 Web Scraping

Many users, in particular early adopters, are interested in driving bitcoin use into more mainstream public use. One way they do this is to try to encourage transactions. A common practice is to attach a bitcoin address as a signature to emails or forum posts. In forum posts especially, users contribute to the community, for example with new mining software or a tutorial on how to get set up to use bitcoins, and leave their address in the signature block. They expect to receive tips from forum readers that find their post helpful. This practice created a natural attack vector to the anonymity of the block chain. We can easily tie user information to transactions in the block chain.

We used a python package called Scrapy[6] to fetch and parse the forum pages(fig. 3). We wrote a spider that crawls bitcointalk.org in a breadth-first manner looking for post signatures that might contain bitcoin addresses (i.e. it matched the regular expression r'1.{26,33}'). We then took this string and verified that it was a legitimate bitcoin public key (bitcoin addresses include a built-in checksum) to avoid attempting to annotate a large number nodes that can't possibly appear in the blockchain.

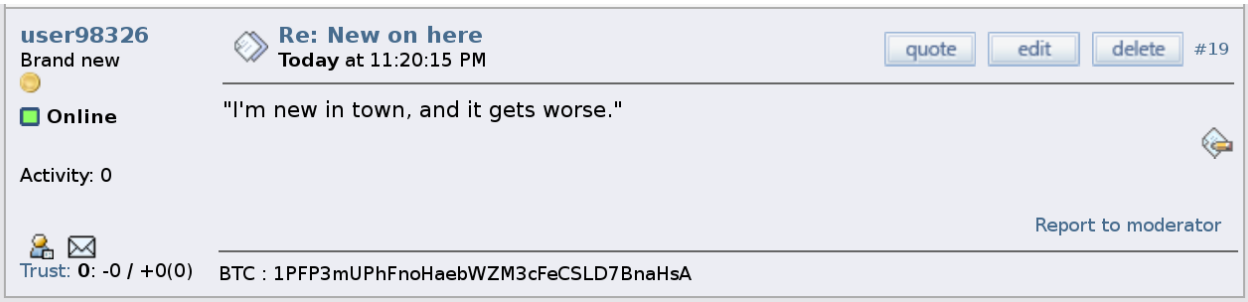

Figure 3: A typical user signature line that includes a bitcoin address for 'tipping'.

We were able to find a large number of forum users that can be directly linked to their keys in the transaction graph. We ran the scraping code for just under 30 hours. During this time we followed links up to four deep from the home page. This covered a total of 44,086 pages and 89,088 posts that included a valid bitcoin address. Of this, we found 2,322 unique users with 2,404 addresses that passed our validation.

5.2 Transaction Fingerprinting

Here we analyze the difficulty in taking rough information regarding transaction time and value, and matching it to an exact transaction in the blockchain. For example, if we overhear Bob telling Alice, I sent you $100 USD yesterday at noon, we examine the difficulty of finding a matching transaction in the blockchain. Suppose we assume the value of bitcoins fluctuated $1 USD yesterday, and that the noon timestamp is accurate to within 5 minutes. Then we have a number of candidate transactions to examine. Continuing with this example, we convert from USD to bitcoins using daily market prices from BlockChain[7]. Next, we examine all transactions occurring in both ranges of [$99, $101] and [11:55 AM, 12:05 PM].

To generalize this example, we examine every transaction in the block chain, and then create time and USD value windows by varying amounts to see how many other transactions will match this weaker, window criteria. The figure below shows, for given USD and time windows, the average number of transactions that will match any particular transaction. Bitcoins have over time become more popular and frequently traded over time, so more candidate transactions will match the given dollar and time criteria in recent months.

[4] `https://github.com/harrigan/bitcointools`
[5] `https://github.com/etotheipi/BitcoinArmory`
[6] http://scrapy.org/
[7] https://blockchain.info/charts/market-price, December 12, 2013

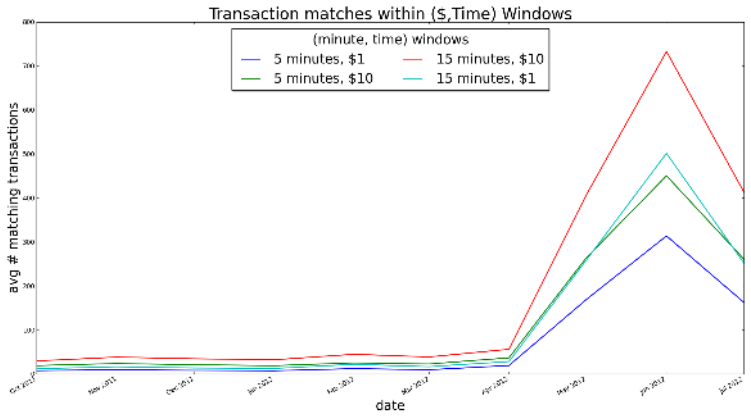

Figure 4: Transaction ambiguity resulting from inexact time and approximate USD worth

Continuing with the example from the start of this section, we may identify Bobs public key with probability 1/10 assuming the conversation took place in March 2012.

Whether by using this fingerprinting tool or by scraping data from online sources, we are able to annotate the blockchain with additional, user-identifying information. In the former case, the annotations may have associated probabilities.

5.3 Graph Analysis

We developed a graph analysis framework in order to de-anonymize users' identities given publicly available information such as scraped bitcoin forum users, and bitcoin transaction information. Figure 5 outlines the different components of our framework.

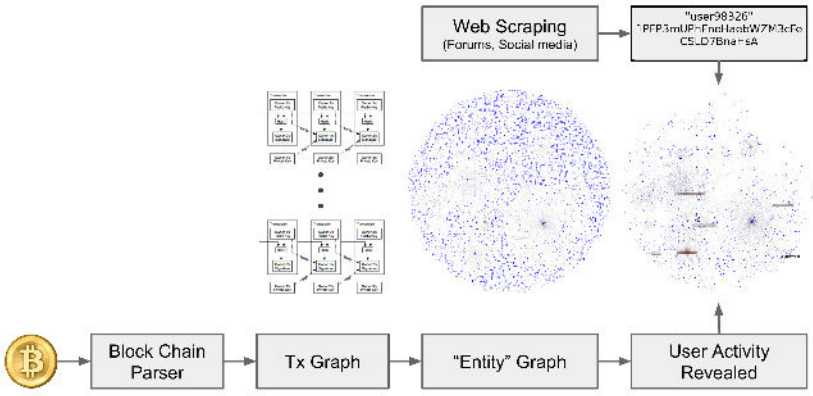

Figure 5: The graph analysis pipeline that was used to reveal user identity constructs a user network graph as shown, and annotates the users in the graph with web scraped results.

5.3.1 The Transaction Graph

Once the transaction records are extracted from the blockchain, we construct a transaction graph that gives an intuition towards the flow of bitcoins between public key addresses over time. More specifically, the transaction graph is a directed graph where the nodes denote public addresses of anonymous individuals or "entities" and the directed edge represents a particular transaction from a source address to a target address. Since both the source and the target "entities" can arbitrarily generate new public-private key pairs for each subsequent transaction, many public key addresses may only appear once or a few times in the transaction graph. Additionally, typical transactions in today's blockchain are multi-input/multi-output transactions. For a more detailed reference, please refer to [1]. We use similar techniques as described in [2], to prune transactions in our graph. For our experiments, we construct the transaction graph for a 24-hour period on October, 25th, 2013. The transaction graph consists of 89,806 transactions, with 80,030 unique vertices (or public key addresses). We also constructed a transaction graph for a 7 month period consisting of 1,669,728 transactions, spanning the months of Mar 2013 through Oct 2013, in an attempt to reveal any links between the bitcoin forum users and the Silk Road nodes, before it was shut down.

5.3.2 The User Graph

In this section, we focus on the 1-day transaction graph constructed and describe our findings on particular activities that are immediately visible after our graph analysis algorithm is applied. Using the transaction graph, we construct a proxy directed graph called the user graph U similar to that described in [2], where the user or entity consists of a collection of public key addresses that were used during separate transactions. As noted in [1], we link together transactions with multi-inputs as originating from the same user. This allows the creation of a user graph by performing a transitive closure on the set of public key address involved in all transactions, after the multi-input public keys are linked together. We use existing tools [8] to construct the entity/user graph, where the vertices now represent physical users/entities, and edges rerepsent a transaction between a source user and a target user. As we construct the user graph from a transaction graph spanning a 24-hour period, our user netowrk is not quite indicative of the true user network as several public key addresses that may have appeared before or after the 24-hour transaction period are not used to link addresses. The resulting user network for the Oct 25, 2013 consists of 54,941 users with 89,806 edges.

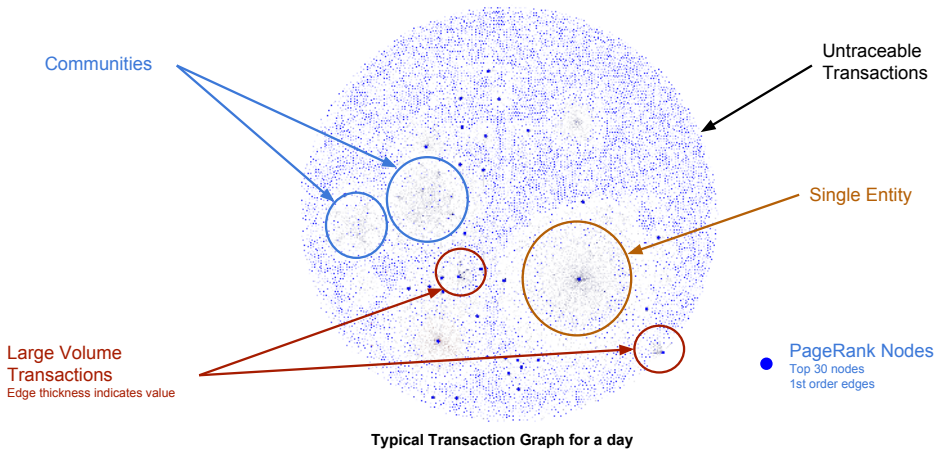

Figure 6: An "entity" graph for a typical day (Oct 25, 2013) labeled with the top 30 page ranked nodes. While a significant set of users are untraceable, several different activities and network layouts are noticeable such as communities, single entities, and large volume transactions.

[8] http://compbio.cs.uic.edu/data/bitcoin/

5.4 Page Rank

Due to the nature of bitcoin transactions in the directed user graph, we see a direct resemblance of this graph to those constructed by search-engines. Most search-engines, in particular Google, use PageRank as a metric to rank websites based on their importance. Intuitively, the algorithm prefers nodes in a directed graph that are most easily reached, or in our case, nodes that receive a large enough traffic to be labeled as important. We use PageRank as a guide to determine the most intersting nodes, or users in our user graph to further investigate their linkage with known forum users. Figure 6 shows the user graph for Oct 25, 2013, and labels the top page ranked nodes with a larger node size. As expected, most of the users in the graph are not connected implying that these nodes are not of much importance as they are not traceable to other users. One also notices several other types of activities or transactions involving communities, single entities, and even large volume transactions as indicated by the thickness of the edges. Given some of the most popular public key addresses from BlockChain.info[9], we were able to determine that one of the single entity nodes with a high in and out-degree was in fact a Bitcoin gambling website called SatoshiDICE[10].

5.5 User De-anonymization

An activity of particular interest was the seizure of Silk Road funds to a publicly known address of the FBI[11] via 445 transactions of exactly 324 Bitcoins. Our graph analysis algorithm picked out this specific FBI address as a user of high importance (high page-ranked node). This validates our algorithm to appropriately pick out nodes of particular interest, and allows for further investigation of these high page-ranked nodes. Given this information and web-scraped bitcoin forum user's information, we were also able to backtrack from these transactions to uncover bitcoin forum users that were only a single hop away from the Silk Road nodes. This implies that the bitcoin forum users transacted with some user that had directly transacted with the Silk Road. Due to DPR's[3] arrest earlier that month, we analysed transactions up to 7 months before the seizure (Mar 25, 2013 through Oct 25, 2013). We were also able to uncover direct transactions from multiple bitcoin forum users to SatoshiDICE[12], implying that they may have gambled at some point during that 7 month period. Interestingly enough we were also able to find direct transactions to Wikileaks[13] from a few of the bitcoin forum users.

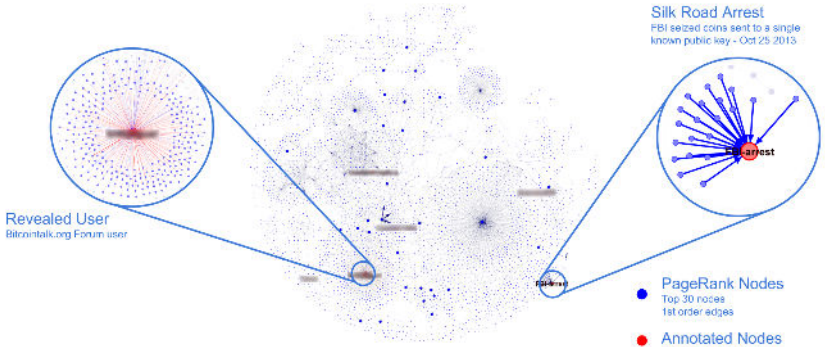

Figure 7: The transaction graph for Oct 25, 2013 showing the top page ranked nodes and their first order edges with annotations from web-scraped results. Several noticeable activities, including the seizure of bitcoins from Silk Road entities to a single known FBI address, tend to be involved with the top page ranked nodes.

[9] `https://blockchain.info/`
[10] `http://legacy.satoshidice.com/`
[11] FBI: 1FfmbHfnpaZjKFvyi1okTjJJusN455paPH
[12] SatoshiDICE 48%: 1dice8EMZmqKvrGE4Qc9bUFf9PX3xaYDp
[13] `http://wikileaks.org/` : 1HB5XMLmzFVj8ALj6mfBsbifRoD4miY36v

6 Conclusion

In conclusion, we showed that by leveraging several sources of publicly available information via web-scraped forums and Bitcoin's transaction ledger, the bitcoin transaction network is shown to be not entirely anonymous. Furthermore, we were able to tie bitcoin forum users with the original Silk Road nodes with only a single intermediary. We were also able to successfully find transactions that directly linked the scraped bitcoin forum users with known entities like SatoshiDICE, and Wikileaks implying that they may have dealt with, supported, or interacting with such entities.

References

[1] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system.

[2] Fergal Reid and Martin Harrigan. An Analysis of Anonymity in the Bitcoin System. *arXiv*, 2011.

[3] Dorit Ron and Adi Shamir. How did dread pirate roberts acquire and protect his bitcoin wealth?

[4] Dorit Ron and Adi Shamir. Quantitative analysis of the full bitcoin transaction graph. Cryptology ePrint Archive, Report 2012/584, 2012. `http://eprint.iacr.org/`.