BlockTag: Design and Applications of a Tagging System for Blockchain Analysis

Yazan Boshmaf, Husam Al Jawaheri, Mashael Al Sabah

BlockTag: Design and Applications of a Tagging System for Blockchain Analysis

Yazan Boshmaf[1], Husam Al Jawaheri[2], and Mashael Al Sabah[1]

[1] Qatar Computing Research Institute, Doha, Qatar
[2] Qatar University, Doha, Qatar

Abstract. Annotating blockchains with auxiliary data is useful for many applications. For example, criminal investigation of darknet marketplaces, such as Silk Road and Agora, typically involves linking Bitcoin addresses, from which money is sent or received, to user accounts and web activities. We present BlockTag, an open-source tagging system for blockchains that facilitates such tasks. We describe BlockTag's design and demonstrate its capabilities through a real-world deployment of three applications in the context of privacy research and law enforcement.

Keywords: Blockchain · Tagging · Bitcoin · Privacy · Law Enforcement

1 Introduction

Public blockchains contain data that describe various financial transactions. As of December 2018, Bitcoin's blockchain amounted to 18.5GB of raw data and is growing rapidly. Such data is crucial for understanding different aspects of cryptocurrencies, including their privacy properties and market dynamics. Blockchain analysis systems, such as BlockSci [16], have enabled "blockchain science" by addressing three pain points: Poor performance, limited capabilities, and cumbersome programming interfaces. These systems, however, are focused on analyzing on-chain data and are not designed to incorporate off-chain data into their analysis pipeline. This limitation makes it difficult to use existing blockchain analysis systems for tasks that require linking off/on-chain data and searching for vulnerabilities or clues, which are common in privacy research and law enforcement.

We present BlockTag: An open-source tagging system for blockchain analysis. BlockTag uses vertical crawlers to annotate on-chain data with customizable, off-chain tags. In BlockTag, a tag is a mapping between a block, a transaction, or an address identifier and external auxiliary data. For example, the system can tag a Bitcoin address with the Twitter user account of its likely owner. BlockTag also provides a novel query interface for linking and searching. For example, BlockTag provides best-effort responses to high-level queries used in e-crime investigations, such as "which Twitter user accounts paid $\geq$ ฿10.0 to Silk Road in 2014."

We designed BlockTag based on the observation that blockchain analysis systems transform raw blockchain data into a stripped-down, simple data structure which can fit in or map to OS memory. Therefore, on-chain data that is not part of basic transaction information, such as hashes, scripts, and off-chain auxiliary

data, cannot be part of this data structure and must have their own mappings. This naturally leads to a layered system architecture where a tagging layer sits on top of an analysis layer with a well-defined and extendable cross-layer interface.

In our implementation, we used BlockSci for analysis as it is much faster than its contenders. For BlockTag, we developed four vertical crawlers that annotate Bitcoin addresses with three types of tags: User tags representing BitcoinTalk and Twitter user accounts, service tags representing Tor hidden service providers, and text tags representing user-generated Blockchain.info labels. We extended BlockSci's analysis library and implemented a novel query engine to link, search, and aggregate off/on-chain Bitcoin data in a SQL-like syntax.

We deployed BlockTag in January 2018 for three months on a single, locally-hosted, machine. As of March 2018, the crawlers have ingested about 5B tweets, 2.2M BitcoinTalk user profiles, 1.5K Tor onion pages, and 30K Blockchain.info labels. This has resulted in 45K user, 88 service, and 29K text tags. In addition to BlockTag, our *contributions* include the following findings from three real-world applications that demonstrate BlockTag's capabilities:

1) Linking: We showed how to deanonymize Tor hidden service user by linking their Bitcoin payments to their social network accounts. In total, we were able to link 125 user accounts to 20 service providers, which included illegal and controversial ones, such as Silk Road and The Pirate Bay. Such deanonymization is possible because of Bitcoin's pseudo-anonymous privacy model and the lack of retroactive operational security, as originally highlighted by Satoshi [22]. From a law enforcement perspective, BlockTag offers a valuable capability that is useful in e-crime investigations. In particular, showing a link between a user account on a website and illegal activities on darknet marketplaces could be used to secure a subpoena and collect more information about the user from the website [27].

2) Market economics: We analyzed the market of Tor hidden services by calculating their "balance sheets." We found that WikiLeaks is the highest receiver of payments in terms of volume, with about 26.4K transactions. In terms of the total value of incoming payments, however, Silk Road tops the list with more than ₿29.6K received on its address. We also found that total value of incoming and outgoing payments of service addresses are nearly the same, meaning they have nearly-zero balances. This suggests that service providers do not keep their bitcoins on the addresses on which they receive payments, but distribute them to other addresses. From transaction dates of service addresses, we found that all but three of the top-10 revenue making service providers are active in 2018.

3) Forensics: We performed an exploratory investigation of MMM: One of the world's largest Ponzi schemes. In total, we were able to link 24.2K users and 202 labels to MMM and its affiliates using BlockTag's full-text search capabilities. We found that all of the linked users are BitcoinTalk users who are mostly male, 20–40 years old, and are located worldwide in more than 80 countries. Moreover, we found that only 313 of these users have logged in to the forum at least once a day and made one or more activities, such as writing posts. After further analysis, we found that all of the linked user accounts were created as part of the "MMM Extra" scheme, which promises "up to 100% return per month for performing simple daily tasks that take 5–15 min." We also used BlockTag to retrieve and

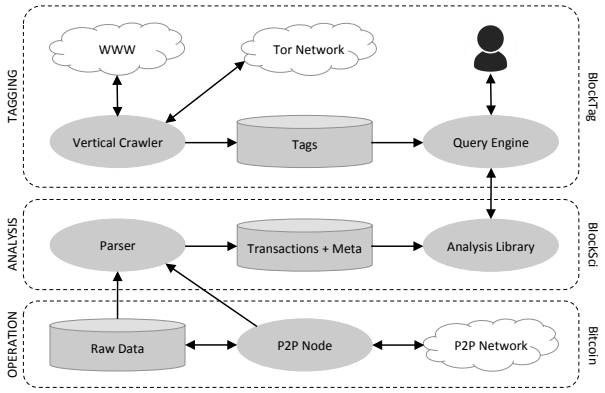


Fig. 1: Layered blockchain system architecture.

model MMM transactions as a directed graph, consisting of 14.3K addresses and 32.K transactions. We found that two of the top-10 ranked addresses, in terms of their PageRank, have been flagged on BitcoinTalk as known scammer addresses. As of December 1, these addresses have received more than ฿2M combined.

2 Design and architecture

BlockTag's design follows a layered system architecture. As depicted in Figure 1, each layer in the stack is responsible for a separate set of tasks and can interact with other layers through programmable interfaces. We present a high-level view of BlockTag's design, and leave the details in the technical report [7].

Tags. In BlockTag, a tag is a mapping between a block, a transaction, or an address identifier and a list of JSON-serializable objects. Each object specifies the type, the source, and other information representing auxiliary data describing the tagged identifier. As raw blockchain data is stored in a format that is efficient for validating transactions and ensuring immutability, the data must be parsed and transformed into a simple data structure that is efficient for analysis. For example, BlockSci uses a memory-mapped data structure to represent core transaction data as a graph. All other transaction data, such as hashes and scripts, are stored separately as mappings that are loaded when needed. BlockTag follows this design choice, and uses a persistent key-value database, such as RocksDB [12], with an in-memory cache in order to store and manage blockchain tags, as they can grow arbitrarily large in size.

BlockTag defines four types of tags, namely user, service, text, and custom tags. A user tag represents a user account on an online social network, such as BitcoinTalk and Twitter. A service tag represents an online service provider, such as Tor hidden services like Silk Road and The Pirate Bay. A text tag represents a user-generated textual label, such as address labels submitted to Blockchain.info. A custom tag can hold arbitrary data, including other tags, and is usually used by analysts to create tags manually.

Vertical crawlers. In BlockTag, a vertical crawler is used to scrape a data source, typically an HTML website or a RESTful API, in order to automatically create block, transaction, or address tags of a particular type using a website-specific parser. A crawler can be configured to run according to a crontab-like schedule, and to bootstrap on the first run with previously crawled raw HTML/JSON data, which can also be used to initialize blockchain tags.

For example, BitcoinTalk, the most popular Bitcoin forum with more than 2.2M users and 42.2M posts, is a good data source to collect public Bitcoin addresses and their associated user accounts. Behind the scene, a BitcoinTalk user crawler downloads user account pages through a URL that is unique for each user account. In addition to a BitcoinTalk user crawler, BlockTag implements a Twitter user crawler that consumes Twitter's API, a Tor hidden service crawler that scrapes onion landing pages of Ahmia-indexed service providers, and a Blockchain.info text crawler that scrapes textual labels that are self-signed by address owners or submitted by arbitrary users. By default, the vertical crawlers create Bitcoin address tags, but can be configured to scrape auxiliary data of other cryptocurrencies, including Litecoin, Namecoin, and Zcash.

Query engine. BlockTag query engine is inspired from NoSQL document databases, such as MongoDB [9], where queries are specified using a JSON-like structure. Selecting, grouping, and aggregating transactions is provided through a simple query interface. To write a query, the analyst starts with specifying block, transaction, or address properties to which the results should match. BlockTag treats each property as having an implicit boolean AND, but also supports boolean OR queries using a special operator. In addition to exact matches, BlockTag has operators for string matching, numerical comparisons, etc. The analyst can also specify the properties by which the results are grouped. Finally, the analyst can specify which properties to return per result. While this query interface is suitable for many tasks, BlockTag's Python package also exposes lower-level functionality to analysts who have tasks with more sophisticated requirements.

One important capability of BlockTag's query engine is address clustering [18], which can be configured to operate on a particular source, namely inputs, outputs, or both, using one of the supported clustering methods. Address clustering expands the set of Bitcoin addresses that are mapped to a unique user, service, or text tag through a technique called closure analysis. As a result, this allows the analyst to identify more links between different tags by considering a larger number of transactions in the blockchain.

BlockTag supports multiple address clustering methods. The first method is based on the heuristic proposed by Meiklejohn et al. [18], which works as follows: If a transaction has addresses A and B as inputs, then A and B belong to the same cluster. The rationale behind this heuristic is that such addresses are highly likely to be controlled by the same entity. While efficient, this method can result in large clusters that include addresses which belong to different entities, due to mixing services, exchanges, mining pools and CoinJoin transactions. In order to tackle this issue, BlockTag implements a novel minimal clustering method that prematurely terminates the original clustering method before the clusters

		# addresses	
Source	Type	Original	Clustering
BitcoinTalk	User	40,970	19,213,141
Twitter	User	4,183	623,189
Tor Network	Service	88	–
Blockchain.info	Text	29,643	–

Table 1: Summary of created tags.

grow to their maximum size. Minimal clustering includes a final trimming phase to find clusters that share at least one address and consequently merges them, after which they are conditionally removed if their size exceeds a defined limit, which defaults to cluster size > 1 (i.e., unconditional removal of merged clusters). Doing so ensures that the clusters are mutually-exclusive and likely to belong to separate entities, but also means the clusters are smaller than usual, reducing the chance of linking different tags as a result.

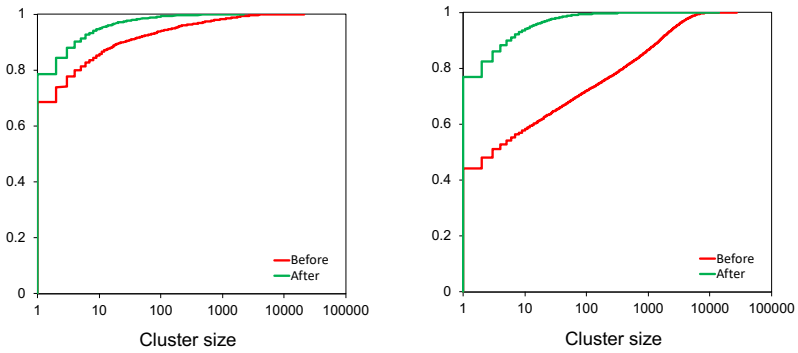
3 Real-world deployment

We deployed BlockTag on a single machine from January 1–March 21, 2018.[3] The machine was running Ubuntu v16.04.4 LTS, Bitcoin Core v0.16.0, and BlockSci v0.5.0 on two 2GHz quad-core CPUs, 128GB of system memory, and 2TB of NAS storage. We used BlockTag to tag Bitcoin's blockchain at the address level. As of March 2018, the crawlers have ingested nearly 5B tweets, 2.2M BitcoinTalk profiles, 1.5K Tor onion pages, and 30K Blockchian.info labels, resulting in 45K user, 88 service, and 29K text tags. We used a previously collected dataset consisting of 4.8B tweets, which were posted in 2014, to bootstrap Twitter user tags. Moreover, for the first application where we link users to services, we configured address clustering for inputs from user tags using the minimal clustering method. We summarize the created tags in Table 1.

Figure 2 shows the CDFs of the cluster size for BitcoinTalk and Twitter user tags, before and after the trimming phase of minimal clustering. As illustrated in the figure, there is a significant drop in the size of clusters after trimming; the average size of a cluster decreased from 75 addresses to 7 for Twitter users, and from 452 addresses to 6 for BitcoinTalk users. The standard deviation also decreased from 606 to 67 and from 1194 to 114, respectively. This suggests that cluster sizes are getting closer to the mean. In fact, more than 90% of the users in both sources have 10 addresses or less in their clusters after trimming. The figure also suggests that more BitcoinTalk users have larger cluster size than Twitter users, as shown by the difference in their before/after distributions.

To cross-validate minimal clustering, we used WalletExplorer: An online service that uses a similar approach to find and tag clusters based on aggregated information from the web. We crawled cluster information from WalletExplorer for both user tag sources. Overall, we found that our closure analysis coincides

[3] For research ethics considerations, please refer to the technical report [7].



(a) Twitter (b) BitcoinTalk

Fig. 2: CDFs of minimal clustering's cluster size before and after trimming

with that obtained from WalletExplorer. All clusters that had less than 700 addresses were untagged on WalletExplorer, which means it is likely that these are user clusters, not services. When we used this number as a limit for the trimming, the percentage of clusters with size 700 or less changed from 83% to 99.95% for BitcoinTalk users and from 97.63% to 99.75% for Twitter users.

4 Applications

4.1 Linking users to services

In e-crime investigations of Tor hidden services, analysts often try to link cryptocurrency transactions to user accounts and activities. This can start with a known transaction that is part of a crime, such as a Bitcoin payment to buy drugs on Silk Road. Alternatively, a wider search criteria can be used to understand the landscape of activities of illegal services, such as finding service providers that receive the most payments. Either way, the analysts need to link users to services, which is a core feature of effective blockchain analysis.

BlockTag was able to link 28 Twitter user accounts to 14 service providers via 167 transactions and 97 BitcoinTalk user accounts to 20 service providers via 115 transactions. Some of these users were linked to multiple service providers. In total, 125 users were linked to 20 services. The results suggest that although Twitter users are smaller in number compared to BitcoinTalk users, they are more active and have a larger number of transactions with services. In fact, some of these users are "returning customers," as they have performed multiple transactions with the same service provider.

From services perspective, Table 2 lists the top-10 service providers ranked by how many users were linked to them. The list is topped by WikiLeaks, which is a service that publishes secret information provided by anonymous sources, with 46 linked users. This is followed by Silk Road, the famous darknet marketplace, with

	# linked users		
Name	Twitter	BitcoinTalk	Total
WikiLeaks	11	35	46
Silk Road	4	18	22
Internet Archives	3	13	16
Snowden Defense Fund	3	8	11
The Pirate Bay	3	7	10
DarkWallet	9	1	10
ProtonMail	1	7	8
Darknet Mixer	1	2	3
Liberty Hackers	0	2	2
CryptoLocker Ransomware	1	0	1

Table 2: Top-10 linked service providers.

22 linked users whose spent coins have been seized by the FBI. Although the Silk Road address was seized, it still appears in transactions until recently. However, based on further analysis, we found that a number of transactions were performed prior to the seizure. Ranked fifth, The Pirate Bay, which is known for infringing IP and copyright laws by facilitating the distribution of protected digital content, was linked to 10 users. As the linked users have accounts with various personally identifiable information (PII), their identities could be deanonymized. We next focus on two case studies that illustrate this threat.

Actionable links. Purchasing products and services from darknet marketplaces is generally considered illegal and calls for legal action. Some of the 22 users who are linked to Silk Road through transactions with seized coins shared enough PII to completely deanonymoize their identity. For example, one user is a teenager from from the U.S. The user has been a registered BitcoinTalk member since 2013, and has a transaction with Silk Road in 2013, the takedown year. The corresponding user account points to his personal website, which contains links to his user profiles on Facebook, Twitter, and Youtube. Even if users do not share PII or use fake identities on their accounts, simply having an account on social networks is enough to track them online, or even secure a subpoena to collect identifiable information, such as login IP addresses. For example, three out of the 18 BitcoinTalk users recently logged in to the forum.

A matter of jurisdiction. One of the users who are linked to The Pirate Bay is a middle-aged man from Sweden. The Pirate Bay was founded by a Swedish organization called Piratbyrån. Furthermore, the original founders of the website were found guilty in the Swedish court for copyright infringement activities. Since then, the website has been changing its domain constantly, and eventually operated as a Tor hidden service. Consequently, having such a link to The Pirate Bay through recent transactions in Sweden can lead to legal investigation, at least, and potentially be incriminating.

4.2 Market economics

Keeping track of market statistics of Tor hidden services is useful for identifying thriving services, measuring the impact of law enforcement, and prioritizing e-

	Volume	Flow of money (฿)		Lifetime (dd/mm/yyyy)		
Name	(# txs)	Incoming	Outgoing	First tx	Last tx	# days
Silk Road	1,242	29,676.99	29,658.80	02/10/2013	19/03/2018	1,628
WikiLeaks	26,399	4,043.00	4,040.74	15/06/2011	21/03/2018	2,470
VEscudero Escrow Service	192	842.42	842.42	27/05/2012	20/08/2017	1,910
Internet Archives	2,957	775.86	746.89	06/09/2013	21/03/2018	1,656
Freenet Project	280	691.87	687.62	23/02/2011	16/03/2018	2,577
Snowden Defense Fund	1,722	218.95	218.95	11/08/2013	18/03/2018	1,680
ProtonMail	3,096	208.40	208.36	17/06/2014	18/03/2018	1,369
Ahmia Search Engine	1,423	176.51	176.50	27/03/2013	06/03/2018	1,652
DarkWallet	983	114.62	97.40	16/04/2014	02/11/2016	931
The Pirate Bay	1,214	76.80	76.80	29/05/2013	21/08/2017	1,544

Table 3: Balance sheet of top-10 service providers ranked by incoming coins.

crime investigations. As such, an analyst may start with calculating a financial "balance sheet" for service providers, which includes the number of transactions with which a service is involved (i.e., volume), the amount of coins a service has received or sent (i.e., money flow), and the difference between the timestamps of the last and first transactions (i.e., operation lifetime). Table 3 shows the balance sheet of the top-10 service providers ranked by incoming coins.

Volume. While the number of created service tags is small, the corresponding service providers have been involved in a relatively large number of transactions. For example, WikiLeaks tops the list with 26.4K transactions. The Darknet Mixer, which did not make it to the top-10 list in Table 3, has a volume of 22.1K transactions that is greater than the remaining services combined. One explanation for this popularity is that users are actually aware of the possibility of linking, and try to use mixing services in order to make traceability more difficult and improve their anonymity.

Money flow. One interesting observation is that service providers have a nearly zero balance, which means almost the same amount of coins comes in and goes out of their addresses. This indicates that the coins is likely distributed to other addresses and is not kept on payment-receiving addresses. One explanation for this behavior is that by distributing coins among multiple addresses, a service provider can reduce coin traceability. Moreover, service providers still need to distribute their revenues among owners and sellers. Among all service providers listed in Table 3, Silk Road stands out with an income of ฿29.6K.

Lifetime. The services vary in their lifetime, ranging from two to seven years of operation. The first transaction date indicates the date on which the service provider started receiving payments through the tagged addresses. Looking at last transaction dates, all but three services are still active in 2018. For example, Silk Road has been receiving money since October 2013, even after the address has been seized by the FBI and its coins auctioned for sale in June, 2014. However, a large number of post-seizure transactions appear to be novelty tips.

4.3 Forensics

Organizations responsible for consumer protection, such as trade commission agencies and financial regulatory authorities, have a mandate to research and identify fraud cases involving cryptocurrencies, including unlawful initial coin offerings and Ponzi schemes. Given the popularity of Ponzi schemes in Bitcoin [28,29], we focus on this type of fraud and show how BlockTag can help analysts flag users who are likely victims or operators of such schemes.

A Ponzi scheme, also known as a high yield investment program, is a fraudulent financial activity promising unusually high returns on investment, and is named after a famous fraudster, Charles Ponzi, from the 1920s. The scheme is designed in such a way that only early investors will get benefits and once the sustainability of the scheme is at risk the majority of shareholders will lose the money they invested [1]. Among various Ponzi schemes in Bitcoin, MMM is considered one of the largest schemes that is hard to detect solely based on blockchain transaction analysis [2], highlighting the need for a systematic integration of auxiliary data into blockchain analysis. As such, an analyst can start the investigation with BlockTag using a full-text search query of keywords associated with MMM scheme, such as its name, without requiring prior knowledge of who is involved in the scheme or how it works.

BlockTag's search returned 24.2K user accounts, all of which are BitcoinTalk users, and 202 Blockchain.info text labels. For BitcoinTalk user accounts, the full-text search matched the website property of an account, which contained a URL pointing to the user's profile on MMM website. As for Blockchain.info text tags, the search matched the self-signed label property, which contained "mmm" substring, as summarized in Table 4. We next analyze the user accounts looking for clues related to MMM operation.

User demographics. Out of 24.2K users, 52.86%, 18.31%, and 12.48% shared their gender, age, and geo-location information, respectively. Based on this data, we found that the users are mostly male (75.44%), between 20–40 years old (average=32), and are located worldwide in more than 80 different countries. However, 70.69% of the users were located in only five countries, namely Indonesia, China, India, South Africa, and Thailand. Interestingly, most of these countries have a corresponding MMM label, as listed partially in Table 4.

Forum activity. Using activity-related properties of user accounts, we found that 99.44% of the users registered on the forum between August 2015–March 2016. Moreover, 98.21% of the users made their last activity on the forum during the same period. This suggest that users have short-lived accounts. In fact, we found that 94.25% of the users were active for 30 days or less, and that 78.45% of users were dormant, meaning they were active for less than a day after registration. This also suggests that most of the users are not engaged with the forum. Indeed, only 313 users made at least one activity, and even for these users, they never engaged with the forum for more than once a day, on average. After manually inspecting the accounts on the website, we found that most of them were created as part of its "MMM Extra" scheme, which promises "up to 100% return per

Label	Frequency
mmm universe.help	46
mmm global	13
bonus from mmm universe.help	9
mmm indonesia	6
mmm nusantara	4
mmm china	2
mmm india	2
mmm indonesia	2
mmm philippines	2
mmm russia	2

Table 4: Top-10 frequent MMM labels.

month for performing simple daily tasks that take 5–15 min," such as promoting MMM on social networks. This was evident from the accounts' signatures, which the crawler did not parse, that included messages such as "MMM Extra is the right step towards the goal" and "MMM participants get up to 100% per month."

Financial operation. We can invistigate how MMM scheme operates financially through transaction graph analysis [25]. In this analysis, Bitcoin transactions are modeled as a weighted, directed graph where nodes represent addresses, edges represent transactions, and weights represent information about transactions, such as input/output values and dates. Analyzing the topological properties of this graph can provide insights into which addresses are important and how the money flows. For example, having a few "influential" nodes and a small clustering coefficient suggest that most of the money funnels through these nodes and does not flow back to others, which are indicative of a Ponzi operation [28,29,2]. In BlockTag, an analyst can easily model case-specific transaction graphs by linking tags based on some search criteria.

We modeled and analyzed five transaction graphs, one for every combination of tag types, as summarized in Table 5. The MMM transaction graph includes addresses of any type, and consisted of 14.3K addresses (i.e., order) and 32.5K transactions (i.e., size). This graph is also sparsely connected, as suggested by the small-sized largest strongly connected component (LSCC), low clustering, and long distance measures. Moreover, it consists of two subgraphs, the user→user subgraph, which is also sparsely connected, and the label→label subgraph, which is dense and small. Even though the two subgraphs are loosely connected through only 170 edges, an order of magnitude more money has flown from users to labels than the reverse direction.

To find influential nodes in the graph, we computed their PageRank, where weights represented input address values of transactions. All of the top-10 ranked nodes were located in the user→user subgraph, which mapped to unique BitcoinTalk users. After manually inspecting the corresponding accounts, we found that the first and the third users have been reported as scammers on BitcoinTalk for operating fraudulent services, namely Dr.BTC and OreMine.org. While the first user has received a total of Ƀ426.7K on her address, the third has received a total of Ƀ1.8M on his address that is associated with Huobi, an exchange service, suggesting that the user has exchanged the received coins.

Type				LSCC			Triangles		Distance	
Input	Output	n	m	n	m	$\bar{C}$	#	%closed	d	r
User	User	14,227	31,819	5,850	17,498	0.11	6,566	0.08	17	7
User	Label	129	125	1	0	0.00	0	0.00	0	0
Label	User	64	45	1	0	0.00	0	0.00	0	0
Label	Label	61	508	20	246	0.64	943	61.04	3	2
Any	Any	14,319	32,497	5,934	18,128	0.11	7,576	0.09	17	7

Table 5: Properties of MMM transaction graphs where n is the order, m is the size, $\bar{C}$ is the average clustering coefficient, d is the diameter, and r is the radius.

5 Discussion

Limitations. BlockTag's main limitation is the validity of its tags, since they are created automatically by crawlers from open, public data sources. This limitation is part of a larger problem that is common with Internet content providers, such as Google and Facebook, especially when content is generated mostly by users [30,17]. In general, the validity issue is especially important for user identities, as fraudsters can always create fake accounts in order to hide their real identity [13]. While doing so improves their anonymity, law enforcement agencies can use the links found through BlockTag to secure a subpoena in order to collect more information about suspects from website operators [27].

Work in-progress. We are designing BlockSearch, an open-source Google-like searching layer that sits on top of BlockTag. BlockSearch allows analysts to search blockchains for useful information in plain English and in real-time, without having to go through the hassle of performing low-level queries using BlockTag. The system also provides in a dashboard for analysts that displays real-time results of important queries, such as the ones we used in the paper. Based on feedback from trade commission agencies and financial regulatory authorities, such capabilities are extremely helpful to protect customers, comply with know you customer (KYC) and anti-money laundering (AML) laws, and draft new, investor-friendly cryptocurrency regulations.

Future work. In order to address the main limitation of BlockTag, we plan to define confidence scores for tag sources. The scores can be computed using various "truth discovery" algorithms [10], which are generally based on the intuition that the more sources confirm a tag the more confidence is assigned to it.

BlockTag is modular by design. This means we can easily enhance or add new capabilities. As such, we plan to implement more vertical crawlers for services such as WalletExplorer, ChainAlysis, BitcoinWhosWho, and Reddit. We also plan to support more clustering methods and develop a systematic way to automatically tag clusters, in addition to blocks, transactions, and addresses, based on label propagation algorithms [15].

6 Related work

Analysis systems. Blockchain analysis systems parse and analyze raw transaction data for many applications. Recently, Kalodner et al. proposed BlockSci [16], an open-source, scalable blockchain analysis system that supports various blockchains and analysis tasks. BlockSci incorporates an in-memory, analytical database, which makes it several hundred times faster than its contenders. While there is a minimal support for tagging in its programming interface, BlockSci is designed for analysis of core blockchain data. At the cost of performance, annotation and tagging can be integrated into the analysis pipeline through a centralized, transactional database. For example, Spagnuolo et al. proposed BitIodine [26], an open-source blockchain analysis system that supports tagging through address labels. However, BitIodine, relies on Neo4j [20], a general-purpose graph database that is not designed for blockchain data and its append-only nature, which makes it inefficient for common blockchain analysis tasks, such as address linking. In contrast, BlockTag is the first open-source tagging system that fills this role.

Linking. The impact of Bitcoin address linking on user anonymity and privacy has been known for a while now [24,19,11,14]. Fergal and Martin [24] showed that passive analysis of public Bitcoin information can lead to a serious information leakage. They constructed two graphs representing transactions and users from Bitcoin's blockchain data and annotated the graphs with auxiliary data, such as user accounts from BitcoinTalk and Twitter. The authors used visual content discovery and flow analysis techniques to investigate Bitcoin theft. Alternatively, Fleder et al. [14] explored the level of anonymity in the Bitcoin network. The authors annotated addresses in the transaction graph with user accounts collected from BitcoinTalk in order to show that users can be linked to transactions through their public Bitcoin addresses. These studies show the value of using public data sources for Bitcoin privacy research and law enforcement, which is our goal behind designing BlockTag.

Tor and darknet markets. Tor hidden services have become a breeding ground for darknet marketplaces, such as Silk Road and Agora, which offer illicit merchandise and services [5,21]. Moore and Rid [21] studied how hidden services are used in practice, and noted that Bitcoin was the dominant choice for accepting payments. Although multiple studies [14,18] showed that Bitcoin transactions could be linked to identities, Bitcoin remains the most popular digital currency on the Dark Web [8], and many users choose to use it despite its false sense of anonymity. Recent research explored the intersection between Bitcoin and Tor privacy [3,4], and found that legitimate hidden service users and providers are one class of Bitcoin users whose anonymity is particularly important. Moreover, Biryukov et al. [5] found that hidden services devoted to anonymity, security, human rights, and freedom of speech are as popular as illegal services. While BlockTag makes it possible to link users to such services, we designed it to help analysts understand the privacy threats and identify malicious actors.

Forensics. Previous research showed that cryptocurrencies, Bitcoin in particular, have a thriving market for fraudulent services, such as fake wallets, fake mining pools, and Ponzi schemes [28,6]. Recently, Bartoletti et al. [2] proposed a data mining approach to detect Bitcoin addresses that are involved in Ponzi schemes. The authors manually collected and labeled Bitcoin addresses from public data sources, defined a set of features, and trained multiple classifiers using supervised machine learning. The best classifier correctly labelling 31 addresses out of 32 with 1% false positives. Interestingly, MMM was excluded because it had a complex scheme. In concept, BlockTag complements such techniques by providing an efficient and easy way to collect and explore data that is relevant to the investigation. This data can be then analyzed using different machine learning and graph algorithmic techniques with the help of existing tools [29].

7 Conclusion

State-of-the-art blockchain analysis systems, such as BlockSci, while efficient, are not designed to annotate and analyze auxiliary blockchain data systematically. We presented BlockTag, an open-source tagging system for blockchains. We used BlockTag to uncover privacy issues with using Bitcoin in Tor hidden services, and to flag Bitcoin addresses that are likely to be part of a large Ponzi scheme.

References

1. Artzrouni, M.: The mathematics of ponzi schemes. Mathematical Social Sciences **58**(2), 190–201 (2009)
2. Bartoletti, M., Pes, B., Serusi, S.: Data mining for detecting bitcoin ponzi schemes. arXiv preprint arXiv:1803.00646 (2018)
3. Biryukov, A., Khovratovich, D., Pustogarov, I.: Deanonymisation of clients in bitcoin p2p network. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. pp. 15–29. ACM (2014)
4. Biryukov, A., Pustogarov, I.: Bitcoin over tor isn't a good idea. In: Security and Privacy (SP), 2015 IEEE Symposium on. pp. 122–134. IEEE (2015)
5. Biryukov, A., Pustogarov, I., Thill, F., Weinmann, R.P.: Content and popularity analysis of tor hidden services. In: Distributed Computing Systems Workshops (ICDCSW), 2014 IEEE 34th International Conference on. pp. 188–193. IEEE (2014)
6. Bohr, J., Bashir, M.: Who uses bitcoin? an exploration of the bitcoin community. In: 2014 Twelfth Annual Conference on Privacy, Security and Trust (PST). pp. 94–101. IEEE (2014)
7. Boshmaf, Y., Jawaheri, H.A., Sabah, M.A.: Blocktag: Design and applications of a tagging system for blockchain analysis. arXiv preprint arXiv:1809.06044 (2018)
8. Castillo, M.d.: Bitcoin remains most popular digital currency on dark web. `https://bit.ly/2UOZNS6` (2016), [online; accessed 01-July-2018]
9. Chodorow, K.: MongoDB: The Definitive Guide: Powerful and Scalable Data Storage. " O'Reilly Media, Inc." (2013)
10. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: the role of source dependence. Proceedings of the VLDB Endowment **2**(1), 550–561 (2009)

11. DuPont, J., Squicciarini, A.C.: Toward de-anonymizing bitcoin by mapping users location. In: Proceedings of the 5th ACM Conference on Data and Application Security and Privacy. pp. 139–141. ACM (2015)
12. Facebook: RocksDB: An embeddable persistent key-value store for fast storage. `https://rocksdb.org` (2014), [online; accessed 01-July-2018]
13. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Communications of the ACM **59**(7), 96–104 (2016)
14. Fleder, M., Kester, M.S., Pillai, S.: Bitcoin transaction graph analysis. arXiv preprint arXiv:1502.01657 (2015)
15. Gregory, S.: Finding overlapping communities in networks by label propagation. New Journal of Physics **12**(10), 103018 (2010)
16. Kalodner, H., Goldfeder, S., Chator, A., Möser, M., Narayanan, A.: Blocksci: Design and applications of a blockchain analysis platform. arXiv preprint arXiv:1709.02489 (2017)
17. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J.: A survey on truth discovery. ACM Sigkdd Explorations Newsletter **17**(2), 1–16 (2016)
18. Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G.M., Savage, S.: A fistful of bitcoins: characterizing payments among men with no names. In: Proceedings of the 2013 conference on Internet measurement conference. pp. 127–140. ACM (2013)
19. Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G.M., Savage, S.: A fistful of bitcoins: characterizing payments among men with no names. In: Proceedings of the 2013 conference on Internet measurement conference. pp. 127–140. ACM (2013)
20. Miller, J.J.: Graph database applications and concepts with neo4j. In: Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA. vol. 2324, p. 36 (2013)
21. Moore, D., Rid, T.: Cryptopolitik and the darknet. Survival **58**(1), 7–38 (2016)
22. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system. Bitcoin.org (2008)
23. Reid, F., Harrigan, M.: An analysis of anonymity in the bitcoin system. In: Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. pp. 1318–1326. IEEE (2011)
24. Reid, F., Harrigan, M.: An analysis of anonymity in the bitcoin system. In: Security and privacy in social networks, pp. 197–223. Springer (2013)
25. Ron, D., Shamir, A.: Quantitative analysis of the full bitcoin transaction graph. In: International Conference on Financial Cryptography and Data Security. pp. 6–24. Springer (2013)
26. Spagnuolo, M., Maggi, F., Zanero, S.: Bitiodine: Extracting intelligence from the bitcoin network. In: International Conference on Financial Cryptography and Data Security. pp. 457–468. Springer (2014)
27. Theymos: DPR subpoena. `https://bitcointalk.org/index.php?topic=881488.0` (2014), [online; accessed 01-July-2018]
28. Vasek, M., Moore, T.: There's no free lunch, even using bitcoin: Tracking the popularity and profits of virtual currency scams. In: International conference on financial cryptography and data security. pp. 44–61. Springer (2015)
29. Vasek, M., Moore, T.: Analyzing the bitcoin ponzi scheme ecosystem. In: Financial Cryptography (2018)
30. Yin, X., Han, J., Philip, S.Y.: Truth discovery with multiple conflicting information providers on the web. IEEE Transactions on Knowledge and Data Engineering **20**(6), 796–808 (2008)