Multi-Class Bitcoin-Enabled Service Identification Based on Transaction History Summarization

Conference Paper · July 2018

DOI: 10.1109/Cybermatics_2018.2018.00208

3 authors:

Kentaroh Toyoda
Agency for Science, Technology and Research (A*STAR)
84 PUBLICATIONS **223** CITATIONS
SEE PROFILE

Tomoaki Ohtsuki
Keio University
512 PUBLICATIONS **3,626** CITATIONS
SEE PROFILE

P. Takis Mathiopoulos
National and Kapodistrian University of Athens
261 PUBLICATIONS **3,727** CITATIONS
SEE PROFILE

2018 IEEE Confs on Internet of Things, Green Computing and Communications, Cyber, Physical and Social Computing, Smart Data, Blockchain, Computer and Information Technology, Congress on Cybermatics

Multi-class Bitcoin-enabled Service Identification Based on Transaction History Summarization

Kentaroh Toyoda*, Tomoaki Ohtsuki*, and P. Takis Mathiopoulos†
* Dept. of Information and Computer Science, Keio University, Japan
† National and Kapodistrian University of Athens, Greece
Email: toyoda@ohtsuki.ics.keio.ac.jp

***Abstract*—In recent years, Bitcoin has been used for many services and purposes, e.g. gambling, marketplace, but also even as an investment scam. In order to clarify how Bitcoin is used, it is in great importance to identify what kind of services are operated by Bitcoin addresses. In this paper, we propose a multi-class service identification scheme in Bitcoin based on novel transaction history summarization. Our novelty is to propose how transaction history is retrieved and how the retrieved transactions are processed for better identification. When a Bitcoin address is given, the characteristics of its transaction history is calculated as features. Then, the set of calculated features is fed into a supervised classifier and the services operated by the given Bitcoin addresses are identified among seven major services: (i) exchange, (ii) faucet, (iii) gambling, (iv) investment scam, (v) marketplace, (vi) mining pool, and (vii) mixer. To our knowledge, we are the first to propose a multi-class identification. We show that our scheme achieves 72% of accuracy through performance evaluation with more than 26,000 Bitcoin addresses that have been used for seven services/purposes from Jan. 2009 to Feb. 2017.**

I. Introduction

It has been nine years since Bitcoin was released [1]. Fig. 1 shows the market price chart of BTC/USD [2]. As can be seen from this chart, although the value of Bitcoin is almost zero until 2013, then it surges exponentially. Bitcoin's surge not only makes more and more people exchange Bitcoin for speculation [3], but also yields many Bitcoin-enabled services. These services include not only traditional ones, e.g., gambling, faucet, marketplace, but also Bitcoin (cryptocurrency) specific services, e.g., mining pool and mixing. Furthermore, it is also abused for illegal purposes such as investment scams [4]. In this situation, researchers have started to analyze the statistics of Bitcoin-enabled services (e.g. [5], [6]). For example, transaction volume distribution by services such as exchanges and gambling is shown in [6].

Although there exist many papers that analyze the statistics of Bitcoin-enabled services, in this paper, we are trying to give answers to the following question: *When a Bitcoin address is given, can we identify its purpose or service by leveraging its transaction history?* Clearly, as any approved Bitcoin transactions are visible in the blockchain, we can extract the characteristics of transactions, e.g., frequency of transactions, which can be useful for service identification. Such information is very useful not only for economical purposes but also to identify fraudulent activities. Until now, there exist two works that are closely related to Bitcoin-enabled service identification: (i) an identification scheme of Bitcoin exchanges [7] and HYIP (High Yielding Investment Program), which is an investment scam, [8], respectively. However, these works propose a binary classification approach, i.e. whether or not given Bitcoin addresses are used for exchange (or HYIP). Realizing a multi-class service identification scheme is beneficial for many things: (i) fraud detection and (ii) enriching the dataset of labelled Bitcoin addresses, which can be used for statistical analysis.

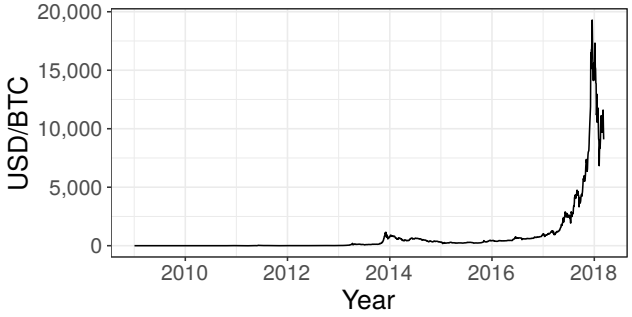


Fig. 1. Market price chart of USD/BTC. Source: blockchain.info

In this paper, we propose a multi-class Bitcoin-enabled service identification based on novel transaction history summarization. There exist two important points for better identification. The first point is the way to retrieve transaction history. For this, we propose two schemes of transaction history retrieval: (i) address-based and (ii) owner-based schemes. In the address-based scheme, when a Bitcoin address is given, any transactions where it is involved either in the inputs or outputs are retrieved and then features are extracted. In contrast, in the owner-based scheme, other addresses controlled by its owner are also extracted with the help of address clustering [5]. The second important point is the way to extract features from the retrieved transactions. For this, our key idea is an elaborate way of pre-processing of transaction history and feature extraction. For example, change in transactions, which is not necessary to feature extraction, is removed. In addition, the amount of BTC is converted to USD to take into account Bitcoin's high volatility. A set of features together with service labels are trained with a machine learning classifier, e.g. Random Forests [9]. Once a machine learning classifier has been trained, the service/purpose of a Bitcoin address

978-1-5386-7975-3/18/$31.00 ©2018 IEEE
DOI 10.1109/Cybermatics_2018.2018.00208

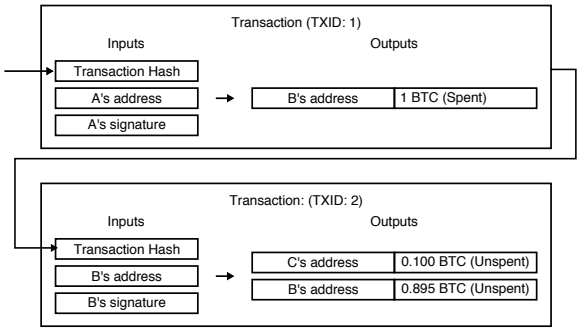


Fig. 2. Two simplified Bitcoin transactions.

can be inferred. This may be used for the service/purpose identification of unlabeled Bitcoin addresses.

The service classification accuracy of the proposed scheme was evaluated by means of computer simulation with more than 26,000 labelled Bitcoin addresses, which are obtained from `Blockchain.info/tags`, `WalletExplorer`, and `BitcoinTalk`. It is shown that our owner-based scheme and the address-based scheme achieve 72% and 70% of classification accuracy, respectively. Furthermore, more detailed classification accuracy is shown as a confusion matrix. The contributed features and their distribution are also shown visually to understand the difference of services.

The remainder of this paper is organized as follows: Section II deals with the preliminaries. Section III deals with the proposed scheme. Then, the performance evaluation is shown in Section IV. Finally, the conclusions of this paper are described in Section V.

II. Preliminaries

A. Fundamental of Bitcoin

Bitcoin is a decentralized cryptocurrency where no trusted authority, i.e. bank, exists [1]. Bitcoin is operated over P2P (Peer-to-Peer) network and a message called *transaction* is used to transfer Bitcoin. BTC is used as the unit. Fig. 2 shows an example of two simplified Bitcoin transactions. In this figure, user A sends 1 BTC to user B in the transaction of which TXID is 1. Then, in the second transaction (TXID is 2), user B spends 0.01 BTC of 1 BTC received from user A to user C. As can be seen from these transactions, addresses are specified in inputs and outputs. An address is calculated from a hash value of a public key. In general, a signature should be attached to a transaction to spend his/her Bitcoin. The signature must be signed by the private key of the pairwise public key, so that the participants of Bitcoin can verify the signature. Any verified transactions are broadcast by the participants to the Bitcoin P2P network. Shortly, the verified transactions are stored in a series of blocks *blockchain* by so-called *miners*. Miners compete each other to get incentive including newly minted Bitcoin and transaction fee for each block. A transaction for the incentive is referred to as *coinbase* transaction.

TABLE I
The list of major services operated with Bitcoin.

Service	Description
Exchange/Wallet	Exchanging among fiat currencies and Bitcoin and manages users' Bitcoin
Faucet	Offering free but small amount of Bitcoin in return for solving CAPTCHA, or clicking advertisements.
Gambling	Gambling games, e.g. dice and roulette.
HYIP	Investment program that promises high interest return, e.g. 1% per day.
Marketplace	Payment service, e.g. escrow, is offered in an online marketplace.
Mining pool	Cooperative mining team that shares computational power to find blocks. If one of the pool members finds a block, its minted Bitcoin is shared by them.
Mixer	Laundering a several Bitcoin transactions to avoid Bitcoin flow tracking.

Bitcoin is said to possess two key features: *Transparency* and *Pseudo-anonymity*. On the one hand, the transparency of Bitcoin comes from the fact that any transactions are stored in the blockchain and publicly available. On the other hand, the pseudo-anonymity comes from the fact that a Bitcoin address is computed from a public key which is not related with individuals. This means that anyone can hold a number of Bitcoin addresses to avoid tracking of the Bitcoin.

B. Services Operated with Bitcoin

Recently, many Bitcoin-enabled services are yielded. TABLE I lists the major services operated with Bitcoin. Exchange and wallet services offer exchange in between cryptocurrencies and fiat currency, and most of them also manage users' Bitcoin. Faucet is a service that very small amount of Bitcoin, e.g. $10^{-6} - 10$ Bitcoin, is given in return for solving CAPTCHA, or clicking advertisements on the website. Gambling is one of the most actively operated services [6]. For example, `SatoshiDICE` offers a provably fair betting game. HYIP is an investment fraud that scammers lure investors to promise high interest, e.g., more than 1% per day [4]. Mining pool is a team for miners to collaboratively share their computational power to find blocks. Since it is infeasible for a solo miner to find blocks, this way enables miners to earn Bitcoin more steadily. Marketplace is an online place where products and services can be purchased. Such a marketplace typically offers both legal goods, e.g. books and electronics, and illegal goods, e.g. drug. Mixer offers a laundry service when a user sends Bitcoin to an address. Mixer promises to send Bitcoin to the address so that it has no association with the ones sent to them [5].

There exist many works that analyze Bitcoin-enabled services (e.g. [4], [10], [11]). For example, the flow of Bitcoin related with `Mt.Gox`, which was the biggest Bitcoin Exchange in 2012, was analyzed [10]. In [11], the reverse-engineering methods were taken to understand the mode of operation of mixer services. The statistics of marketplaces are shown in [12], [13].

However, there are few works that identify a service operated by a given Bitcoin address by leveraging the charac-

teristics of transaction history. The characteristics of Bitcoin transaction history differs by services. For example, it is typical that very small amount of Bitcoin is frequently spent from the addresses of faucet. Such a characteristics might be helpful to service identification. To our knowledge, two closely related ideas have been studied so far [7], [8]. In the next section, these two schemes are summarized.

C. Service Identification

In [7], an identification scheme of Bitcoin exchanges was proposed. Totally 19 features are used to identify the Bitcoin addresses of exchange. For example, the features include *total_bitcoin_received* (how much Bitcoin the address received from transaction outputs over the full time window) and *total_bitcoin_spent* (how much Bitcoin the address spent as transaction inputs over the full time window). In [8], an identification scheme of HYIP operators' Bitcoin addresses was proposed. Although it is generally difficult to collect HYIP operators' Bitcoin addresses, these are collected by scraping the topics in the investors-based games section of `BitcoinTalk`, which is one of the biggest online Bitcoin forum.

However, these schemes focus on two-class identification, i.e. whether or not a Bitcoin address is used for exchange (or HYIP). Obviously, it is in great demand to extend two-class identification to multi-class identification from many aspects, such as economics. As described in the above, a few researchers have studied Bitcoin services and its economy [5], [6], However, it is typical that labelled Bitcoin addresses are obtained from `Blockchain.info/tags`, `WalletExplorer`, and `BitcoinTalk` and thus very limited. Multi-class identification will enable to label a number of Bitcoin addresses of which its service/purpose is unknown.

III. Proposed Scheme

We propose a multi-class service identification of Bitcoin addresses based on novel transaction history summarization. When a Bitcoin address is given to identify its service/purpose, there are two important points for better identification. The first one is the way to retrieve transaction history. Regarding the transaction retrieval, two schemes are proposed: (i) address-based and (ii) owner-based schemes. In the address-based scheme, when a Bitcoin address is given, any transactions where it is involved either in the inputs or outputs are retrieved and then features are extracted. In contrast, in the owner-based scheme, other addresses that an owner may control are also extracted with the help of address clustering. Then, all the transaction history of the addresses is retrieved and features are extracted. Hence, the owner-based scheme identifies the service by leveraging how the owner of a given Bitcoin address spends/receives Bitcoin. The second one is the way to extract features from the retrieved transactions is very important for the better identification. The second point is how to process the retrieved transactions. For this, our key idea is an elaborate way of pre-processing of transaction history and feature extraction, which are described in Section III-B and III-C, respectively. For example, unnecessary entries in each transaction, e.g. change, are removed. In addition, the amount of BTC is converted to USD to take into account Bitcoin's high volatility.

In the following, we explain the address-based scheme and then the owner-based one.

Fig. 3 illustrates the flow of the address-based scheme. It mainly consists of the following phases: (i) Collecting labelled Bitcoin addresses, (ii) retrieving and pre-processing transactions, (iii) feature extraction, and (iv) training a machine learning classifier.

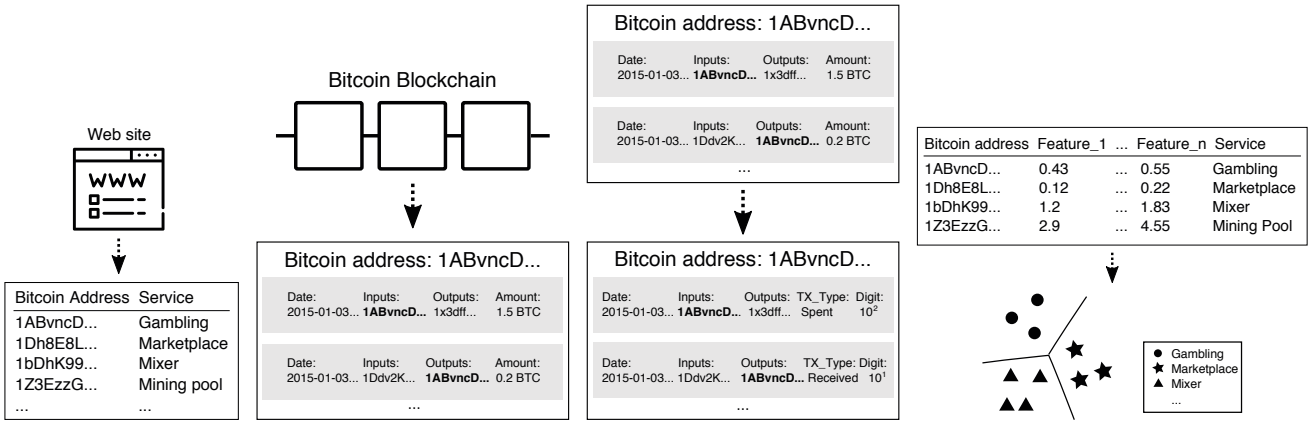
A. Collecting labelled Bitcoin addresses

At first, a set of labelled Bitcoin addresses should be obtained. Here, "labelled" Bitcoin addresses are Bitcoin addresses where their services/purposes are known. For this, we explored two websites `WalletExplorer` and `Blockchain.info/tags` that list Bitcoin addresses with their service names. Fig. 4 shows the examples of the labelled Bitcoin addresses in the Internet. As can be seen from theses figures, `WalletExplorer` and `Blockchain.info/tags` offer the service name and its Bitcoin addresses. By searching these service names on web search engines, their service/purpose can be obtained.

B. Retrieving and Processing Transactions

In the address-based scheme, for each Bitcoin address, transactions where a given Bitcoin address is included in inputs and outputs are retrieved from the blockchain. In contrast, in the owner-based scheme, the transaction history of Bitcoin addresses that the owner of the given Bitcoin address might control is retrieved. This is because it is not uncommon to operate services with multiple Bitcoin addresses. To infer other controlled Bitcoin addresses, address clustering is leveraged [14]. We use the heuristic that any input addresses in each transaction are assumed to be controlled by the same owner. However, when applying this heuristic, several owners are found to have a large number of Bitcoin addresses. Hence, to reduce the calculation time, we randomly sample N_{addr} addresses from them. In this paper, N_{addr} is fixed to 100.

However, some Bitcoin addresses have a large number of transactions, which may take long time for pre-processing and feature extraction. Hence, we retrieve at most N_{TX} subsequent transactions. In this paper, N_{TX} is fixed to 1,000. The retrieved transactions are then pre-processed for feature extraction. The retrieved transactions are classified into three types: (i) spent transactions, (ii) received transactions, and (iii) coinbase transactions. Fig. 5 illustrates an example of three types of transactions when processing Bitcoin address `1a2c...`. Spent transactions are the transactions where the given Bitcoin address spends Bitcoin, i.e. the address appears in the inputs of the transactions. Received transactions are the transactions where the given Bitcoin address receives Bitcoin, i.e. the address only appears in the outputs of the transactions. Lastly, coinbase transactions are the transactions of which inputs are null while the address appears in the outputs.



(a) Collecting class labels. (b) Retrieving transaction history. (c) Pre-processing transaction history. (d) Training a classifier.

Fig. 3. Overview of the address-based Bitcoin-enabled service identification scheme.

Wallet Xapo.com (show transactions)

Page 1 / 10123 Next... Last (total addresses: 1,012,203)

address	balance
3HPF1x3g34oeiakfmrUb8Ej6mbWSj8CpZc	1582.12161469
3FxNPjJmcQsw5czeheoLMub7sveWW1BNFK	419.3315891
3NHB7u25szhW9vz3ikeBbDN8RGE5LC3VSR	389.98092455
3Pei4hHgGpXQAJFremyy3yNREcLzvBvh4E	370.98485997
3BuQmbmdce3e31GEovq5SgowLdfMgJzLDE	355.59097868

(a) `WalletExplorer`. In this figure, a part of Bitcoin addresses controlled by an online wallet service, Xapo.com, are displayed.

Address	Tag
1BteW1uy7zNXMNqhFdvFafbJLd1HcHVPLx	Ripplepay
1MyZjxnLgun6APrDkkh7ffrQJyy6xbuDho	FreedomBox Foundation
1Gpa3NKn8nR9ipXPZbwkjYxqZX3cmz7q97	Ancientbeast.com
1Pug3dAjqXYUkYkjppHjQyZia2xgM79YZV	Blockbox Linux
1wdociqV3xXf8AnEeoPR2jzvVpk1ptH9N	Osiris-sps.org
18S8ugWEuWLbMP9DBpBdDk9SN6CiRxZB8S	The Free Network Foundation
17RTTUAiiPqUTKtEggJPec8RxLMi2n9EZ9	Bitcointalk Forums

(b) `Blockchain.info/tags`.

Fig. 4. Examples of the labelled Bitcoin addresses in the Internet.

After other controlled Bitcoin addresses are retrieved, their transaction history is merged in order, pre-processed, and features are extracted as same as the address-based scheme.

Each transaction is then pre-processed for feature extraction. Fig. 6 shows an example of Bitcoin address `1a2c...`'s pre-process. The pre-processing phase consists of three phases: (i) change removal, (ii) currency conversion, and (iii) digit conversion.

Change removal: The change part of every spent transaction is removed. Fig. 6(a) illustrates an example of change removal of Bitcoin address `1a2c...`. As can be seen from this figure,

Inputs	Outputs	Amount
`1a2c...`	`1x4c...`	1 BTC

(a) Spent transactions.

Inputs	Outputs	Amount
`1x4c...`	**`1a2c...`**	3 BTC

(b) Received transactions.

Inputs	Outputs	Amount
	`1a2c...`	12.67 BTC

(c) Coinbase transactions.

Fig. 5. An example of `1a2c...`'s spent, received, and coinbase transactions.

when `1a2c...` appears not only inputs but also outputs of a transaction, such output(s) are assumed to be the change that the owner of `1a2c...` takes and thus removed.

Currency conversion: Since the price of BTC is highly volatile as shown in Fig. 1, the amount of transactions is converted from BTC to USD by referring a daily currency conversion rate. This way eliminates the Bitcoin's volatility effect in feature extraction. In practice, we use a daily currency conversion rate offered by Blockchain.info [2].

Digit conversion: When extracting the characteristics of how much money is transferred, we are not interested in the exact transferred amount, but how large amount is transferred, i.e. the digit of amount. Hence, each amount x in USD is converted to $10^{\lfloor \log_{10}(x) \rfloor}$. Fig. 6(c) illustrates an example of digit conversion. In this case, 450 USD is converted to 10^2.

C. Feature Extraction

A set of features is extracted from the pre-processed transactions. Feature selection is important for training the characteristics of each service, when training a machine learning classifier.

TABLE II lists the calculated features. f_{TX} is defined as the frequency of transactions per day. Some services very frequently issue transactions, but some may not. r_{received} and r_{coinbase} are the features that represents the ratio among

Inputs	Outputs	Amount
1a2c...	1x4c...	1 BTC
	1a2c...	0.9 BTC

→

Inputs	Outputs	Amount
1a2c...	1x4c...	1 BTC

(a) Change removal (only for spent transactions).

Inputs	Outputs	Amount
1a2c...	1x4c...	0.02 BTC

→

Inputs	Outputs	Amount
1a2c...	1x4c...	2.3 **USD**

(b) Currency conversion from BTC to USD.

Inputs	Outputs	Amount
1x4c...	1a2c...	450 USD

→

Inputs	Outputs	**Digit**
1x4c...	1a2c...	$\mathbf{10^2}$

(c) Extracting the digit of USD amount.

Datetime	Inputs	Outputs	TX type	Digit
2015-01-23 4:50:00	1a2c...	1x4c...	Spent	10^0
2015-01-25 12:10:30	1x4c...	1a2c...	Received	10^{-2}
2015-01-25 18:20:00		1a2c...	Coinbase	10^2
...	...	...	...	...

(d) Pre-processed transactions.

Fig. 6. An example of pre-processing on Bitcoin address 1a2c...'s transactions.

TABLE II
THE LIST OF CALCULATED FEATURES.

FEATURE	DEFINITION
f_{TX}	The frequency of transactions which is defined as the number of all transactions per day.
r_{received}	The ratio of received transactions to all transactions.
r_{coinbase}	The ratio of coinbase transactions to all transactions.
$f_{\mathrm{spent}}(10^i)$	The frequency of digit i in USD appeared in spent transactions, where $i \in (10^{-3}, 10^{-2}, \cdots, 10^6)$
$f_{\mathrm{received}}(10^i)$	The frequency of digit i in USD appeared in received transactions, where $i \in (10^{-3}, 10^{-2}, \cdots, 10^6)$
r_{payback}	Payback ratio defined as the ratio of Bitcoin addresses that appear in both inputs and outputs.
$\bar{N}_{\mathrm{inputs}}$	The mean value of the number of inputs in the spent transactions
$\bar{N}_{\mathrm{outputs}}$	The mean value of the number of outputs in the spent transactions

spent, received, and coinbase transactions. For example, mining pool's r_{coinbase} may be much higher than the others'. $f_{\mathrm{spent}}(10^i)$ and $f_{\mathrm{received}}(10^i)$ are features that characterize how much amount of money are frequently transferred. For example, the transferred value of faucet may be typically small, e.g. less than 1 USD. In contrast, that of marketplace may be much higher than 1 USD. r_{payback} is a service pays back to the Bitcoin addresses that spent some before . Hence, r_{payback} of HYIP and gambling is much higher than the others, since they often pay back some money to investors or players. In contrast, pay back seldom occur for faucet and mining pool.

D. Training a Machine Learning Classifier

After a series of features is calculated, a machine learning classifier is trained by using labelled services together with a set of features. Any supervised machine learning classifier, e.g. RF (Random Forests) [9], XGBoost (eXtreme Gradient Boosting) [15], SVM [16], and neural network [17], does the job.

Once a machine learning classifier has been trained, the service/purpose of a given Bitcoin address can be inferred. In particular, when a Bitcoin address whose label is unknown is given, the set of features is calculated. Then, the features are input into the machine learning classifier and the service where a given Bitcoin address might be operated is identified.

TABLE III
THE NUMBER OF OWNERS AND ADDRESSES BY SERVICES

SERVICE	# OWNERS	# ADDRESSES
Exchange/Wallet	157	10,469
Faucet	61	340
Gambling	89	6,734
HYIP	956	2,026
Mining pool	38	1,645
Marketplace	17	1,900
Mixer	32	3,199
Total	1,360	26,313

IV. PERFORMANCE EVALUATION AND DISCUSSION

To show the effectiveness of the proposed scheme, (i) accuracy, (ii) confusion matrix, (iii) contributed features, and (iv) distribution of features by services are evaluated. Accuracy is defined as the ratio of correctly identified addresses to the entire set of addresses. Accuracy is calculated for (i) the owner-based scheme and (ii) the address-based scheme. For the owner-based scheme, the first heuristic of address clustering is applied, meaning that any addresses in the inputs of a transaction are assumed to be controlled by the same owner. In contrast, in the address-based scheme, only transactions where a given Bitcoin address is involved are used for feature extraction.

RF is used for the machine learning classifier [9]. In general, two parameters must be specified for RF. The first parameter, which is the number of decision trees, is set to 100. The second one, which is the number of used features in each split of a decision tree, is set to $\sqrt{|F|}$ where $|F|$ is the total number of features.

TABLE III lists the dataset used for evaluation. We obtained the pairs of service name and its Bitcoin address from WalletExplorer and Blockchain.info/tags except for HYIP. By searching these service names on web search engines one by one, their service/purpose can be obtained. HYIP operators' name and address are scraped from BitcoinTalk, and thus the number of HYIP is much larger than that of the others. As a result, our dataset contains totally 1,360 owners and 26,313 Bitcoin addresses. Our dataset is available in our cloud storage[1]. Transactions in blocks from 1 to 452,242, which ranges from 9 Jan. 2009 to 9 Feb. 2017, are used.

[1] https://goo.gl/sQJKdx

TABLE IV
CONFUSION MATRIX OF THE OWNER-BASED SCHEME.

	Exchange	Faucet	Gambling	HYIP	Marketplace	Mixer	Mining pool
Exchange	**0.41**	0.04	0.14	0.12	0.17	0.05	0.06
Faucet	0.04	**0.80**	0.04	0.09	0.00	0.00	0.03
Gambling	0.10	0.06	**0.54**	0.15	0.11	0.00	0.05
HYIP	0.04	0.12	0.09	**0.72**	0.01	0.00	0.02
Marketplace	0.05	0.00	0.08	0.00	**0.85**	0.02	0.00
Mixer	0.02	0.00	0.00	0.00	0.04	**0.94**	0.01
Mining pool	0.12	0.09	0.06	0.03	0.00	0.01	**0.70**

Since the number of owners (and addresses) differ by classes, the sampling method is applied. Specifically, for each evaluation, the following procedure is repeated by 100 times and the accuracy is calculated.

1) The entire dataset is randomly sampled so that the number of addresses/owner of each service be the same. (17 owners are randomly sampled from each class in the owner-based scheme, while 100 addresses are randomly sampled from each class in the address-based scheme, respectively.)
2) The set of proposed features are calculated with the sampled dataset.
3) Calculate accuracy by evaluating our scheme with 10-fold CV (Cross Validation).

The contributed features are evaluated based on information gain.

$$IG(\mathcal{C}, f) = H(\mathcal{C}) - H(\mathcal{C}|f), \tag{1}$$

where $IG(\mathcal{C}, f)$ is the information gain when a feature f is chosen and $\mathcal{C}$ is the class, $H(\cdot)$ is the entropy, respectively. Intuitively, when the contribution of f_i is higher than f_j, $IG(\mathcal{C}, f_i) > IG(\mathcal{C}, f_j)$.

A. Accuracy and Confusion Matrix

The overall accuracy by 0.72 by the owner-based scheme, while 0.70 by the address-based scheme. In the following, only the results of the owner-based scheme are shown[2]. TABLE IV shows the confusion matrix of each service by the owner-based scheme. In this table, each row indicates an actual service, while each column indicates a predicted service, respectively. From the confusion matrix, not only classification accuracy, but also the ratio of misclassified services can be clearly seen. For example, 0.80 (= 80%) of faucet services are correctly identified, while 9% of them are misclassified as HYIP. As can be seen from TABLE IV, the classification accuracy of mixer, marketplace, faucet, and HYIP is more than 0.7, while that of exchange and gambling is less than 0.54. From the misclassified services, services that resemble each other are clarified. For example, faucet and HYIP somewhat resemble, since 12% of HYIP is misclassified to faucet and 9% of faucet is misclassified to HYIP, respectively. Similarly, gambling and HYIP also resemble, since 9% of HYIP is misclassified to gambling and 15% of HYIP is misclassified to gambling, respectively.

[2]The more detailed comparison between the address-based and the owner-based schemes will be clarified in the full paper.

TABLE V
TOP 10 CONTRIBUTED FEATURES.

NAME	INFORMATION GAIN
f_{TX}	0.52
$f_{\mathrm{received}}(10^{-3})$	0.44
r_{received}	0.35
r_{payback}	0.31
N_{inputs}	0.27
N_{outputs}	0.21
$f_{\mathrm{received}}(10^{0})$	0.19
$f_{\mathrm{received}}(10^{-1})$	0.18
$f_{\mathrm{received}}(10^{2})$	0.18
$f_{\mathrm{received}}(10^{1})$	0.16

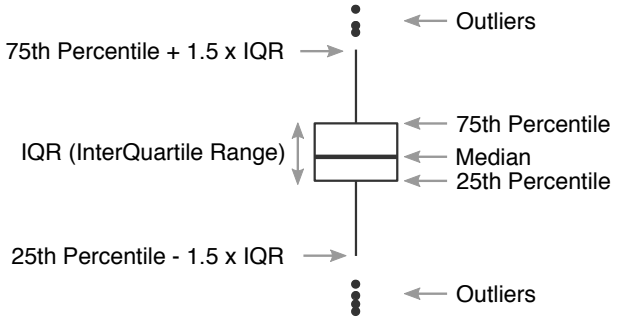
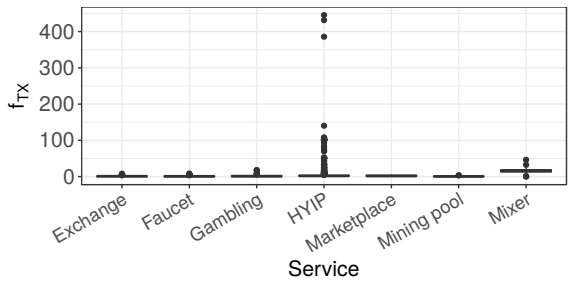


Fig. 7. Explanation of boxplot.

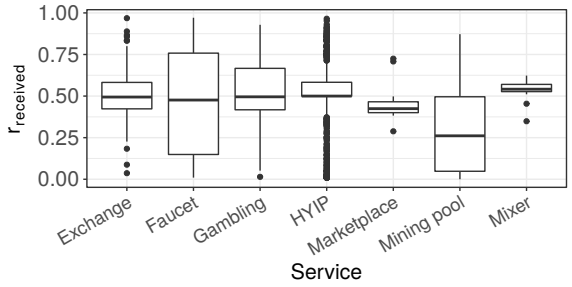
B. Distribution of Contributed Features by Services

TABLE V lists top 10 contributed features. From this table, it can be seen that f_{TX} is the most contributed feature followed by $f_{\mathrm{received}}(10^{-3})$, r_{received}, and r_{payback}. To visually see the difference of these features by services, the distribution of these features is plotted by services with the expression of boxplots. Fig. 7 illustrates the examination of boxplot. Boxplot is a good way to see the distribution of a feature value by services. The top and bottom of a box and the horizontal line in the box indicate the range of quartiles, where each box is structured with 75th percentile, 25th percentile, and median (50th percentile), respectively. From the top and bottom of a box, two lines are vertically drawn. The edges of these lines indicate the boundaries of outliers. Hence, if the box is "pressed" or the length of the box is short, it means that such a feature value is concentrated on a specific value within a service. In contrast, the size of box is large, it means that such a feature value is widely distributed within a service.

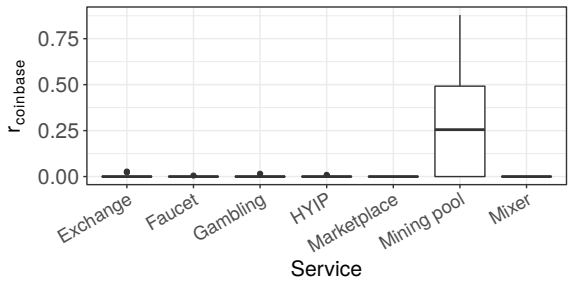
Fig. 8(a) shows the boxplots of f_{TX} by services. From this figure, it can be seen that f_{TX} of mixer and HYIP is much higher than that of the others, meaning that f_{TX} is effective to identify mixer and HYIP. Fig. 8(b) shows the boxplots of r_{received} by services. From Fig. 8(b), the median of mining pool's r_{received} is 0.2 and much lower than the others, while others' r_{received} are around 0.5. This is because a certain amount of transactions of mining pool are coinbase. To prove this, Fig. 8(c) shows the distribution of r_{coinbase}. It is clear that the transactions of mining pool are coinbase. Fig. 8(d) shows the boxplots of r_{payback} by services. As can be seen



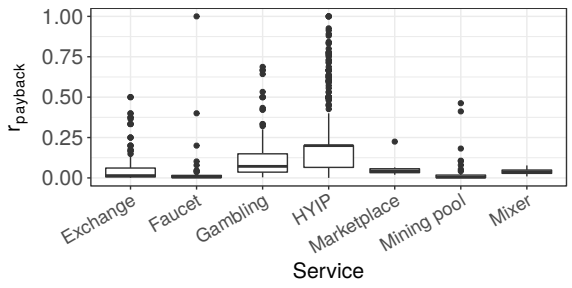
(a) f_{TX}



(b) r_{received}



(c) r_{coinbase}



(d) r_{payback}

Fig. 8. Boxplots of contributed features by services.

from Fig. 8(d), the median of r_{payback} of gambling and HYIP is higher than the others. This is because gambling and HYIP often pay back to wagers and investors. Hence, r_{payback} is effective to identify HYIP and gambling.

Fig. 9 shows the distribution of $f_{\mathrm{received}}(x)$ by services. In Fig. 9, each figure can be seen as a histogram of digits in USD, but each bin is replaced by a boxplot. As can be seen from this figure, the characteristics of each service is clearly seen. For example, the distribution of digits of faucet and HYIP are tailored to smaller amount, i.e. ranging from 10^{-2} to 10^{1}. Since faucet typically offers small amount of Bitcoins and investors of HYIPs also invest a few USD, this result is comprehensive. In contrast, that of exchange and marketplace is mostly ranging from 10^{0} to 10^{2}, which is also intuitive. Furthermore, it can be also seen that all of the digit distribution is bell-shaped, however, their kurtosis and peak are different by services. For example, the peak of exchange is 10^{1}, while that of gambling is 10^{0} and that of mixer is 10^{2}, respectively.

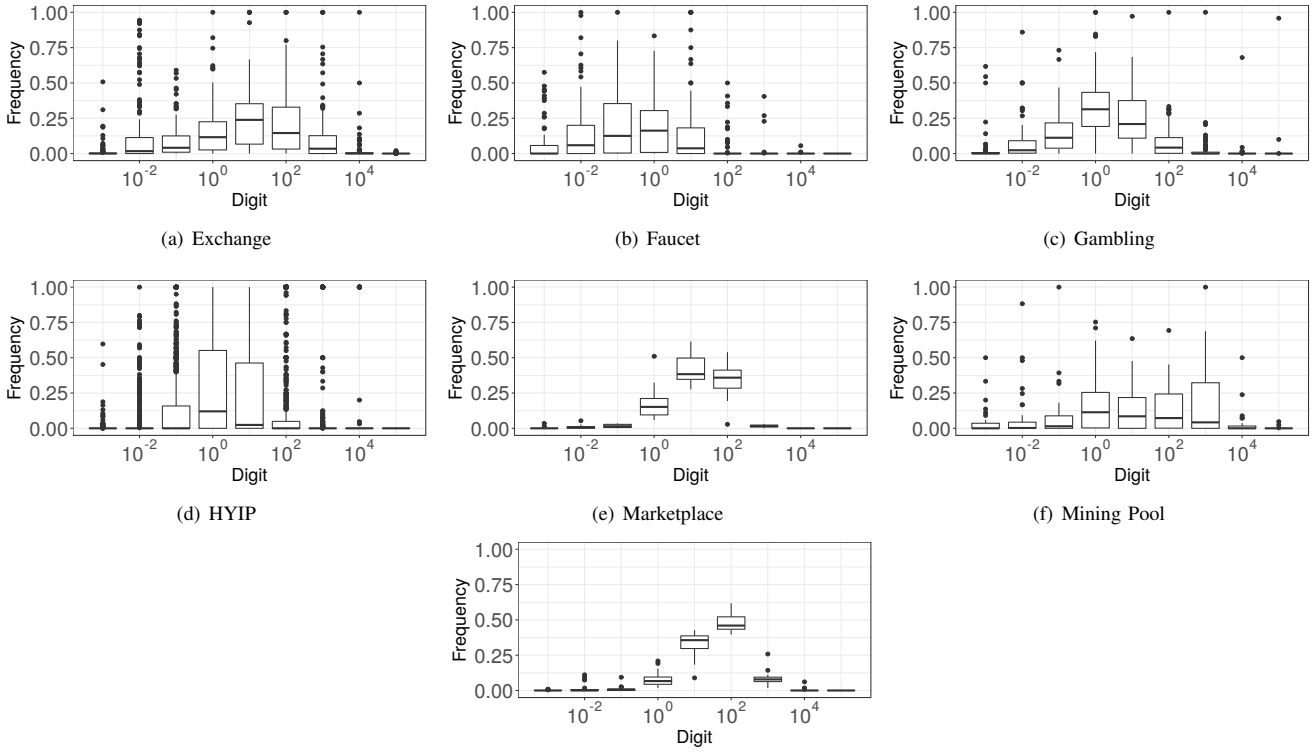
It is worth noting that several characteristics of each service can be seen from the transaction history, even though Bitcoin has a nature of pseudo-anonymity.

V. Conclusions

In this paper, we have proposed a multi-class service identification of Bitcoin addresses based on novel transaction history summarization. Realizing a multi-class service identification scheme is beneficial for many things: (i) fraud detection and (ii) enriching the dataset of labelled Bitcoin addresses, which can be used for statistical analysis. The novelty of our scheme is two-fold: The first one is the way to retrieve transaction history. We propose two schemes of transaction history retrieval: (i) address-based and (ii) owner-based schemes. In the address-based scheme, when a Bitcoin address is given, any transactions where it is involved either in the inputs or outputs are retrieved and then features are extracted. In contrast, in the owner-based scheme, other addresses controlled by its owner are also extracted with the help of address clustering. The second important novelty is the way to extract features from the retrieved transactions. For this, our key idea is an elaborate way of pre-processing of transaction history and feature extraction. For example, change in transactions is removed currency and digit conversion. A set of features together with service labels are trained with a machine learning classifier. Once a machine learning classifier has been trained, the service/purpose of a Bitcoin address can be inferred. This may be used for the service/purpose identification of unlabeled Bitcoin addresses.

The classification performance of the proposed scheme was evaluated with more than 26,000 labelled Bitcoin addresses. It is shown that the owner-based scheme achieves 72% of classification accuracy, while the address-based scheme achieves 70%, respectively. Furthermore, classification accuracy by services is shown as confusion matrix. As a result, several services, e.g. faucet, mixer, and marketplace are well identified, while exchange and gambling are not. The distribution of the contributed features is also shown by services. From the result, it has been shown that the characteristics of each service is visually clarified. For example, the payback ratio of gambling and HYIP is much higher than that of others, because they pay back to wagers and investors in nature.

However, the research of service identification in cryptocurrency has just started. In the future work, we will not only



(a) Exchange (b) Faucet (c) Gambling

(d) HYIP (e) Marketplace (f) Mining Pool

(g) Mixer

Fig. 9. Distribution of $f_{\mathrm{received}}(x)$

try to improve the identification accuracy, but also extend our scheme to other cryptocurrencies, e.g. Ethereum.

ACKNOWLEDGMENT

This work is partly supported by the Grant KAKENHI (No.18K18162) from Ministry of Education, Sport, Science and Technology, Japan.

REFERENCES

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Self published*, pp. 1–9, 2008.

[2] "BTC to USD: Bitcoin to US Dollar Market Price - Blockchain," https://blockchain.info/charts/market-price?timespan=all, (Accessed on 03/25/2018).

[3] C. Sas and I. E. Khairuddin, "Design for Trust: An Exploration of the Challenges and Opportunities of Bitcoin Users," in *Proc. of CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 6499–6510.

[4] M. Vasek and T. Moore, "There's No Free Lunch, Even Using Bitcoin: Tracking the Popularity and Profits of Virtual Currency Scams," in *Proc. of Financial Cryptography and Data Security (FC)*, 2015, pp. 44–61.

[5] 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 *Proc. of Internet Measurement Conference (IMC)*. ACM, 2013, pp. 127–140.

[6] M. Lischke and B. Fabian, "Analyzing the Bitcoin Network: The First Four Years," *Future Internet*, vol. 8, no. 1, pp. 1–40, 2016.

[7] S. Ranshous, C. A. Joslyn, S. Kreyling, K. Nowak, N. F. Samatova, C. L. West, and S. Winters, "Exchange Pattern Mining in the Bitcoin Transaction Directed Hypergraph," in *Proc. of Workshop on Bitcoin and Blockchain Research (BITCOIN)*. Springer, 2017.

[8] K. Toyoda, T. Ohtsuki, and P. T. Mathiopoulos, "Identification of High Yielding Investment Programs in Bitcoin via Transactions Pattern Analysis," in *Proc. of Global Communications Conference (GLOBECOM)*. IEEE, 2017.

[9] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] D. Ron and A. Shamir, "Quantitative analysis of the full Bitcoin transaction graph," in *Proc. of Financial Cryptography and Data Security (FC)*, 2013, pp. 6–24.

[11] M. Möser, R. Böhme, and D. Breuker, "An inquiry into money laundering tools in the Bitcoin ecosystem," in *Proc. of eCrime Researchers Summit (eCRS)*. IEEE, 2013, pp. 1–14.

[12] N. Christin, "Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace," in *Proc. of International Conference on World Wide Web (WWW)*. ACM, 2013, pp. 213–224.

[13] K. Soska and N. Christin, "Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem," in *Proc. of USENIX Security Symposium*, 2015, pp. 33–48.

[14] E. Androulaki, G. O. Karame, M. Roeschlin, T. Scherer, and S. Capkun, "Evaluating user privacy in Bitcoin," in *Proc. of Financial Cryptography and Data Security (FC)*. Springer, 2013, pp. 34–51.

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 785–794.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[17] S. Haykin and N. Network, "A comprehensive foundation," *Neural Networks*, vol. 2, no. 2004, 2004.