IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 26, 2019, accepted May 31, 2019, date of publication June 5, 2019, date of current version June 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2921087

A Novel Methodology for HYIP Operators' Bitcoin Addresses Identification

KENTAROH TOYODA[1,2], (Member, IEEE), P. TAKIS MATHIOPOULOS[4], (Senior Member, IEEE), AND TOMOAKI OHTSUKI[3], (Senior Member, IEEE)

[1]Agency for Science, Technology, and Research (A*STAR), Singapore Institute of Manufacturing Technology, Singapore 138634
[2]Faculty of Science and Technology, Keio University, Kanagawa 223-8522, Japan
[3]Department of Information and Computer Science, Keio University, Kanagawa 223-8522, Japan
[4]Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, 161 22 Athens, Greece

Corresponding author: Kentaroh Toyoda (kentaroh_toyoda@simtech.a-star.edu.sg)

This work was supported by JSPS (Japan Society for the Promotion of Science) KAKENHI Grant Number JP18K18162.

ABSTRACT Bitcoin is one of the most popular decentralized cryptocurrencies to date. However, it has been widely reported that it can be used for investment scams, which are referred to as high yield investment programs (HYIP). Although from the security forensic point of view it is very important to identify the HYIP operators' Bitcoin addresses, so far in the open technical literature no systematic method which reliably collects and identifies such Bitcoin addresses has been proposed. In this paper, a novel methodology is introduced, which efficiently collects a large number of the HYIP operators' Bitcoin addresses and identifies them based upon a novel analysis of their transactions history. In particular, a scraping-based method is first proposed which is able to collect more than 2,000 HYIP operators' Bitcoin addresses from the Internet thus providing a large number of the HYIPs' samples. Second, a supervised machine learning technique, which classifies, whether or not, specific Bitcoin addresses belong to the HYIP operators, is introduced and its performance is evaluated. The proposed classification method is based upon two novel approaches, namely the rate conversion technique that mitigates the effect of Bitcoin price volatility and the sampling technique that reduces the computational amount without sacrificing the classification performance. By employing close to 30,000 real Bitcoin addresses, extensive performance evaluation results obtained by means of computer simulation experiments have shown that the proposed methodology achieves excellent performance, i.e., 95% of the HYIP addresses can be correctly classified, while maintaining a false positive rate less than 4.9%. In order to further validate the proposed classifier's ability to detect the HYIP operators' Bitcoin addresses, our designed classifier has been tested against a recently published list of the HYIP addresses maintaining its excellent detection accuracy by achieving a 93.75% success rate.

INDEX TERMS Bitcoin, blockchain analysis, forensics, data mining, HYIP (high yield investment programs).

I. INTRODUCTION

In recent years, cryptocurrency has been widely accepted as a new digital currency in the world, with Bitcoin being possibly the most popular digital currency [1]. Bitcoin is a decentralized cryptocurrency for which no central authority is required to control it. Unlike other conventional non-cryptocurrencies, Bitcoin possesses two key features: (i) *Transparency* and (ii) *Pseudo-anonymity*. The former is typically assured because all transactions are kept at a decentralized ledger called blockchain and can be publicly observed. The latter is achieved by introducing Bitcoin address computed from a user's public key, and thus no real identifier, e.g. the user name, is embedded on the transactions.

Recent studies have revealed that the property of pseudo-anonymity is abused by investment scams, which are referred to as HYIP (High Yield Investment Programs) [2], [3]. HYIP is a rather popular scam that operators use to lure investors to promise high interest payment, e.g., 1-2% interest per day. Although HYIP related activities emerged in the 1920's [4], the dramatic increase of Bitcoin-enabled HYIPs has been witnessed in recent years. For example, one of the

The associate editor coordinating the review of this manuscript and approving it for publication was Donato Impedovo.

Bitcoin-enabled HYIP operators, who operated the Bitcoin Saving and Trust (BTCST) that offered 7% daily interest to investors and raised 700,000 BTC (worth $4.5 million based on the average trading value) [5], was charged by the Security and Exchange Commission (SEC) [6]. It is noted that a recent study has shown that there exist many statistical aspects of HYIP, e.g. its lifetime and advertisements [2], [3]. The threatening fact is that HYIP operators can start new HYIPs again and again in almost "no-time", as Bitcoin addresses can be generated unlimitedly. Therefore, it is very important to identify HYIP operators' Bitcoin addresses and associated transactions related to fraud by extracting features from the transaction history.

Capitalizing on this need, we have recently proposed a HYIP operators' Bitcoin addresses identification methodology [7]. The main idea behind this work is that when a Bitcoin address is tested, its transactions characteristics, which are also known as features, are calculated from its transaction history, and then designing a machine learning classifier by training the characteristics of HYIP and non-HYIP. As the work reported in [7] is rather preliminary, there are a number of important issues which need to be further investigated. The first one is that the dataset used for evaluation in [7] is rather limited, as only 43 HYIPs and 1,523 non-HYIPs have been considered. Hence, an efficient HYIP operators' address collection method is required in order to obtain reliable results. The second one is that, since [7] does not consider the volatility of Bitcoin, the calculated features related with the amount of transferred Bitcoin have not been accurately calculated. The third one is that with the approach considered in [7] it is not feasible to process a large number of transactions of "giant" Bitcoin addresses and owners.

Motivated by the above, in this paper we extend our previous work [7] toward more solid and accurate HYIP owners' Bitcoin addresses identification solutions by individually addressing the above mentioned issues. Firstly, in order to increase the number of dataset, we leverage the fact that many HYIPs are introduced on *Investor-based games* section on bitcointalk.org,[1] which is a major Bitcoin online forum. We then identify more than 2,000 HYIP operator's Bitcoin addresses from the collected TXID (Transaction Identifier) by making decisions based upon the context of posts. Secondly, when transactions are processed to calculate features, the unit is converted from BTC to USD through the use of chart rate offered by a Bitcoin exchange. As it will be shown later, this simple step significantly improves classification performance. Lastly, a sampling approach is introduced to reduce the computation complexity when a large number of transactions or addresses are retrieved. It will be also shown that, even if sampling is executed, the actual classification performance does not degrade much.

We have evaluated the classification performance of the proposed methodology by means of computer simulation employing 2,026 HYIP operators' Bitcoin addresses and 26,967 non-HYIP ones obtained from bitcointalk.org and Blockchain.info. These experiments have shown that 95% of HYIPs are correctly classified, while maintaining false positive rate less than 4.9%. Furthermore, the computational time, the contributing features, and their distribution by HYIP and non-HYIPs are shown. Finally, to test the generality of the proposed classifier to detect HYIP operators' Bitcoin addresses, an additional experiment without considering our previous dataset collection has been run. In particular, our designed classifier is tested against the HYIP address list offered by Bartoletti *et al.* [8] which consists of 32 HYIP operators' Bitcoin addresses. The obtained results have shown that, also for this experiment, the detection accuracy is 0.9375, again verifying that the proposed methodology can be effectively used for the forensics of Bitcoin-related fraud.

[1] https://bitcointalk.org/index.php?board=207.0

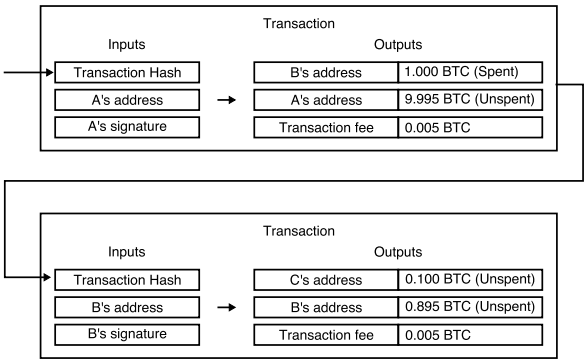

FIGURE 1. An example of two Bitcoin transactions.

The remainder of this paper is structured as follows. After this introduction, Section II presents the background information on Bitcoin which is related to this paper. In Section III, the detailed description of the proposed methodology can be found. In Section IV, various performance evaluation results are presented and discussed. Finally, the conclusions of this paper are given in Section V.

II. PRELIMINARIES

In this section, the most important operational procedures of Bitcoin are presented and their relationship to the research area and topic of the paper is explained. Subsequently, fraudulent activities that coexist with Bitcoin and its conventional studies are discussed. Finally, the HYIP scam activities and their operation are briefly reviewed.

A. TRANSACTIONS FUNDAMENTALS

Bitcoin is a decentralized cryptocurrency that works in a P2P (Peer-to-Peer) network [1]. FIGURE 1 illustrates an example of two Bitcoin transactions. As illustrated in FIGURE 1, Bitcoin is transferred among Bitcoin addresses via a message format called *transaction*. For each of the two Bitcoin transactions illustrated in this figure, the senders and recipients of Bitcoin are identified as inputs and outputs, respectively. A Bitcoin address is created from a pair of ECDSA (Elliptic

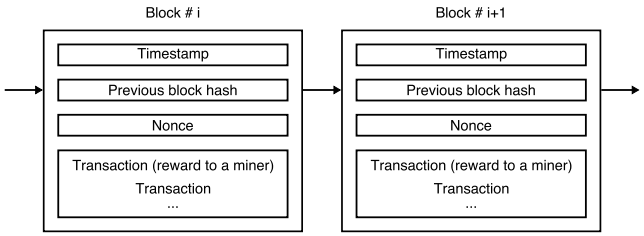

FIGURE 2. A sequence of blocks used to construct a blockchain.

Curve Digital Signature Algorithm) private and public keys. When a user creates a transaction to send a certain amount of Bitcoin to a specific Bitcoin address, a message signature that can be calculated from its pairwise private key is required. A transaction is sent to the P2P network and its validity is checked, e.g. whether or not the inputs of a transaction have not been previously spent, and the attached signature is validated by participating nodes. If identified to be valid, this transaction must be agreed by every participant of Bitcoin and stored permanently in the Bitcoin blockchain. In Bitcoin, a set of approved transactions are stored in a *block* and newly created blocks are periodically distributed among all nodes of the P2P network. However, since no single trusted authority exists in Bitcoin, there is a possibility that such blocks are abandoned and the already spent Bitcoin might be used again. Such transactions might be double-spending and thus must be avoided. In fact, Bitcoin avoids double-spending by rewarding Bitcoins to rational nodes as incentive. In Bitcoin, a block is created by solving a computational puzzle that is difficult to solve but easy to check. More specifically nodes, which are often referred to as *miners*, are required to find a nonce (i.e. a number used once) under the condition that the result of a (SHA-256) hash value together with a reference to the previous block and a set of the unapproved transactions lowers the specified target value. The first miner that identified such a nonce can acquire the newly minted Bitcoin through a so-called *coinbase* transaction, and all transaction fees included in the block. FIGURE 2 illustrates a sequence of blocks used to construct a blockchain. Since the previous block is required to create the next block, this results in an ever-growing chains of blocks, i.e., a blockchain. It is noted that as a "by-product" of this block creation process, valid transactions agreed by the nodes are permanently stored.

B. BITCOIN ADDRESSES CLUSTERING

As the complete Bitcoin transactions are available via the blockchain, it is interesting to analyze them for the better understanding of how Bitcoin is used, e.g., the anonymity it offers to its users. For this, let's recall that anyone can manage a number of Bitcoin addresses. To infer Bitcoin addresses controlled by the same owner, there exist several techniques proposed in the past which are usually referred to as AC (Address Clustering), e.g. [9]–[12]. For example in [9], Androulaki *et al.* showed that two heuristics are effective to tie a set of Bitcoin addresses to its owner. The first

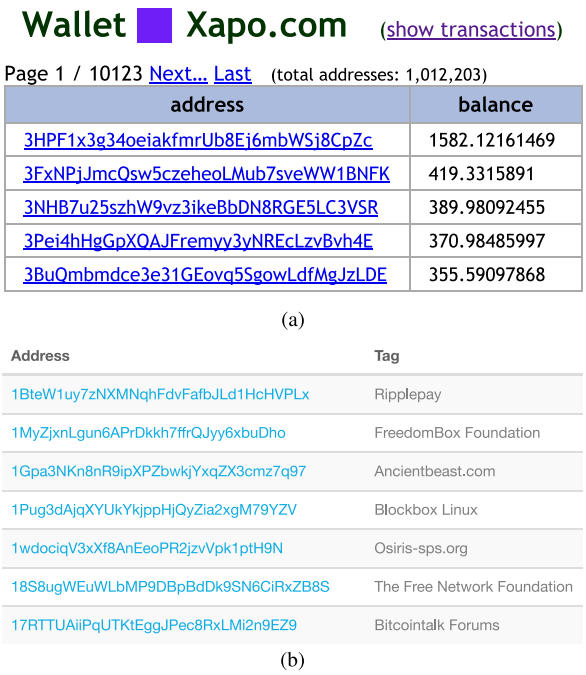

FIGURE 3. Collecting labelled Bitcoin addresses from the Internet. (a) WalletExplorer.com. In this figure, a part of Bitcoin addresses controlled by an online wallet service, Xapo.com, are displayed. (b) Blockchain.info/tags.

such heuristic is that any input addresses in a transaction are assumed to be owned by the same entity. The second heuristic is that, if and only if, the number of addresses at the output is two and one of them has never appeared in the blockchain, such address is for the sending entity to accept changes. Nick in [13] proposed two other heuristics that are applicable only to general Bitcoin wallet applications, e.g. Bitcoin Core and Electrum. The first one is to leverage the fact that such wallet applications can only send, by default, Bitcoin to one Bitcoin address. This means that at most two Bitcoin addresses appear in the outputs of transactions, i.e. one is recipient's address and the other is for the change. The second heuristic is that when the output value of a Bitcoin transaction is lower than any of its inputs, such output Bitcoin address is likely to be a change address. Recently in [14], Neudecker and Hartenstein proposed a sophisticated AC technique that combines the previously proposed heuristics and as well as IP address information extracted from observing the message flood process of the Bitcoin network. In the same reference, two monitoring peers were deployed on the Bitcoin network and the location where a transaction is issued is inferred by leveraging the information measured at the two peers with a GeoIP service, which provides the rough location of client by querying its IP address.

The main limitation with the identification of Bitcoin addresses controlled by the same user is that these Bitcoin addresses do not involve any information that links its owners' identifiers. Hence, there have been many attempts to retrieve the information from the other sources, e.g. websites and posts on SNS (Social Networking Services), and

combine it with the data from the blockchain. For example, Meiklejohn *et al.* have studied how Bitcoin transactions are used via transaction graph analysis based on AC [11]. In particular, they have shown that Bitcoin is used for various services, e.g., (i) Mining pools; (ii) Wallets; (iii) Exchanges; (iv) Vendors, (v) Gambling, and (vi) Money laundry. They have further analyzed the transaction volume and graph network among Bitcoin addresses used for these services. More recently, Lischke and Fabian have analyzed the development and advances of Bitcoin during the first four years of their operation [15]. McGinn *et al.* have developed a tool to visualize Bitcoin transaction patterns [16]. Rahouti *et al.* have surveyed the recent threats and security solutions to Bitcoin [17]. Fleder *et al.* [12] developed a system which ties real names or entities to Bitcoin addresses by scraping the topics of bitcointalk.org, which is the biggest Bitcoin forum. Spagnuolo *et al.* [18] also developed a powerful tool called BitIodine to cluster Bitcoin addresses and tie them with identifiers used in bitcointalk.org and bitcoin-otc, which is an OTC (Over The Counter) market for Bitcoin. They have demonstrated its effectiveness by using their own tool to identify Bitcoin addresses owned by Ross William Ulbricht, who was the creator and operator of `Silk Road` and ransom payment caused by `CryptoLocker` [19].

C. HYIP

Bitcoin has been also misused in various investment scams, e.g. [20]–[22]. For example, HYIP, which is a classical investment fraud scheme that offers high interest payments, e.g., more than 1% per day, while the earned interest is collected from new investors [20]. HYIP itself is not new as it has been known since the 1920s. Charles Ponzi was one of the most famous HYIP operators who promised investors a 50% profit within 45 days, or 100% profit within 90 days by claiming that he runs arbitrage [21]. Because of this scam, HYIP is also known as Ponzi scheme. More recently, in 2009, Bernie Madoff pled guilty to swindling investors by 64.8 billion USD, which is the largest Ponzi scheme in history [22]. Next we will briefly present the fundamentals of the HYIP scheme.

Let us consider a HYIP which offers daily interest of 2% and assume that an investor, by investing 0.01 `BTC` into this HYIP, he/she will receive 0.0002 `BTC` per day. Thus, in 50 days, the earnings will be the same as the investment, i.e., 0.01 `BTC`. If this HYIP continues to pay for more than 50 days, this investor can yield a profit. It is noted that, in general, HYIP investments do not allow early withdrawal, unless high penalties, e.g. 10% (or even more) of the total investment, are paid. However, it is obvious that this HYIP will, sooner or later, disappear so that many investors will lose their investment. Vasek and Moore in [2] analyzed the statistical aspect of Bitcoin-enabled HYIPs. They showed that the median lifetime of HYIP is 37 days looking at 23 HYIPs that existed from January 2, 2013 through September 9, 2014. Furthermore, they showed that these HYIPs totally earned 1,562 BTC (843,000 USD). They have also identified the differences between successful and non successful HYIPs by investigating HYIP-related threads in bitcointalk.org [3]. They investigated the relationships between the lifetime of the scam, the profiles of the scammers and their victims, and how much interactive the threads are on scams. It is interesting to note that, as reported in [5], a Bitcoin-enabled HYIP operator was charged by the Security and Exchange Commission (SEC) [6]. This HYIP, Bitcoin Saving and Trust (BST), offered 7% daily interest to investors and raised 700,000 `BTC`.

In these days, it is relatively easy to collect investors and operate HYIP with Bitcoin. The operators can advertise on the Internet and swindle investments from all over the world [23]. Even if a HYIP collapses, operators can restart quite easily another HYIP, because any number of Bitcoin addresses can be generated at almost no cost. Furthermore, there exists a "ready-to-use" script code of HYIPs [24], e.g., the well-known HYIP manager script `Gold Coders`. All-in-all, since the operation of HYIP together with Bitcoin can be easily implemented, it is expected that the economical losses by Bitcoin-enabled HYIPs will become even higher in the near future.

D. MOTIVATION AND CONTRIBUTIONS

Motivated by the urgent issue mentioned above, we previously proposed a basic methodology to identify Bitcoin addresses which are used to operate HYIPs [7]. However, it is noted that there exists at least three shortcomings in our conference paper [7]. The first one is related to the dataset used for evaluation is small: Only 43 HYIPs and 1,523 non-HYIPs were included. To increase the number of non-HYIP Bitcoin addresses, not only Blockchain.info/tags but also WalletExplorer.com will be used. In contrast, since there is no available website that explicitly offers HYIP operators' Bitcoin addresses, an efficient HYIP operators' address collection scheme is necessary. The second shortcoming of [7] is that the features related with the amount of transferred Bitcoin are not well calculated, since we did not consider the volatility of Bitcoin. For this reason, BTC should be converted to real currency value before feature extraction. The last shortcoming is that it is infeasible to process a large number of transactions of "giant" Bitcoin addresses and owners. For example, the number of transactions related with Bitcoin address `1N52wHoVR79PMDishab2XmRHsbekCdGquK` in Apr. 2018 was more than 97,000. Furthermore, by applying the AC technique, the owner of this address is inferred to control more than 1 million Bitcoin addresses. Obviously, it is necessary to reduce the required computational time at feature extraction.

Here we will present a generic methodology which deals with these shortcomings in a more systematic and complete manner. In this paper, apart from introducing and analyzing the performance of a generic and accurate HYIP owners' Bitcoin addresses identification methodology, it further proposes the following new ideas:

1) A novel dataset collection approach, which significantly increases the number of HYIP owners' Bitcoin addresses is introduced. Specifically, as many as

2,134 HYIP owners' Bitcoin addresses have been identified, which is significant by considering the fact that our previous work could only collect 43.

2) A solid identification methodology is proposed, which consists of several key ideas such as unit conversion and sampling approach, to realize the lightweight, fast, and accurate identification. Our sampling approach significantly decreases the computational time of preprocessing while maintaining good classification performance. If the sampling method is not used, it takes more than two hours to process a Bitcoin address with 10,000 transactions. In contrast, with our sampling method, it only requires one minute for the whole process without sacrificing the best classification performance.
3) The proposed classifier is further tested against a dataset which involves 32 HYIP owners' Bitcoin addresses [8]. As it will be shown, 30 out of 32 such addresses are successfully identified, i.e. the proposed methodology achieves a high accuracy of 93.75%.

III. METHODOLOGY

This section presents the proposed methodology using a two-step approach, namely: (i) Scraping-based HYIP transactions and operators' Bitcoin addresses collection; and (ii) Identification process.

A. SCRAPING-BASED HYIP TRANSACTIONS AND OPERATORS' BITCOIN ADDRESSES COLLECTION

In general, it is difficult to collect a large number of HYIP transactions and its operators' Bitcoin addresses. To overcome this, we leverage the fact that many HYIPs are introduced on the *Investor-based games* section on bitcointalk.org, which is a major Bitcoin online forum. FIGURE 4(a) illustrates an example of topics in Investor-based games section on BitcoinTalk.org, which clearly shows that HYIP operators (and sometimes investors) create topics to lure investors together with their HYIPs' rule. In such topics, it can be seen that investors and operators sometimes post the proofs of investment or withdrawal of interest with their TXID (transaction identifier). FIGURES 4(b) and 4(c) illustrate two such examples of investment and payment proof with TXID, respectively. Clearly, both HYIP operators and investors have reasons to post such investment and payment proof: i) On the one hand, HYIP operators want more and more investments; and ii) On the other hand, investors need more investors for their interest to be paid. Consequently, HYIP transactions and HYIP operators' Bitcoin addresses can be scraped from TXIDs that appear in such posts. The following algorithms are used to retrieve HYIP operators' Bitcoin addresses.

1) **Collect HYIP-related topics:** The topic titles, the number of page view, the number of replies, and the links of topics are retrieved from the *Investor-based games* section of BitcoinTalk.org.

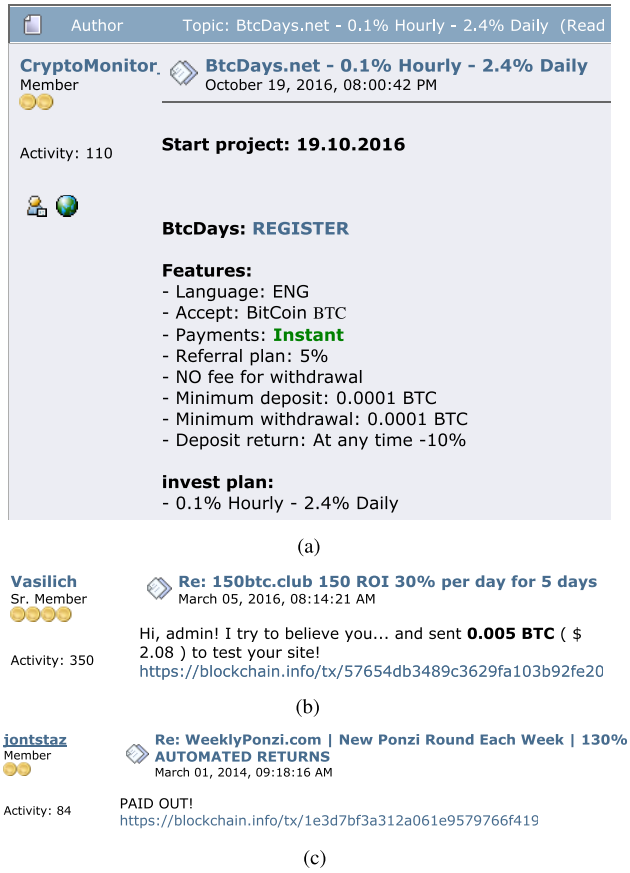

FIGURE 4. The procedure for collecting HYIP transactions and HYIP operators' Bitcoin addresses from bitcointalk.org. (a) An example of HYIP promoting on bitcointalk.org. (b) A post to prove an investment with TXID. (c) A post to prove a successful payment with TXID.

2) **Collect posts with TXIDs in the topics:** Since Bitcoins' TXIDs can be expressed as 64 characters including alphabets from 'a' to 'f' and digits '0' to '9', for each post of the collected topics, any TXID is extracted with the regular expression `[a-fA-F0-9]{64}` that can extract any TXID. Here, `[a-fA-F0-9]` matches a letter from 'a' to 'f', or from 'A' to 'F', or from '0' to '9', and `{64}` denotes that the previous pattern must be repeated 64 times. Hence, `[a-fA-F0-9]{64}` can match every possible TXID in the posts.
3) **Identify HYIP Bitcoin addresses from the posts with TXIDs:** For each post that involves a TXID, a HYIP Bitcoin address is extracted by manually judging the context of posts with TXIDs. Specifically, when a post is to show a payment proof, Bitcoin addresses that appear in the inputs of the TXID are owned by HYIP operators with high probability. In contrast, when a post is with an investment proof, a Bitcoin address that appears in the outputs might be owned by HYIP operators. Therefore, when a post with TXID is given, it is necessary to extract HYIP Bitcoin addresses according to the context of posts.

By following the above steps, topics in the Investor-based games section have been retrieved. FIGURE 5 illustrates the

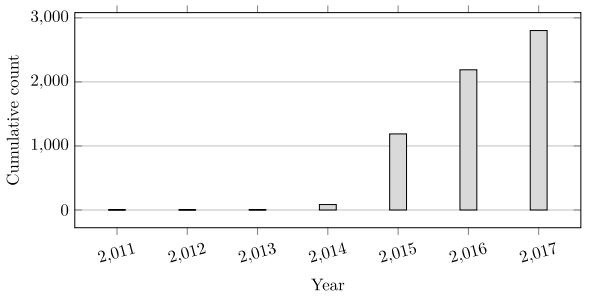

FIGURE 5. The cumulative count of created HYIP topics for the years 2011-2017.

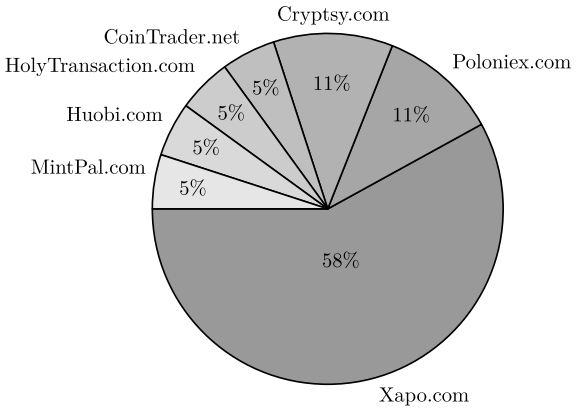

FIGURE 6. The breakdown of online wallets used by HYIPs for the years 2011-2017.

cumulative count of topics for the years 2011-2017. The first topic is posted in June 2011 but very few topics are created. However, the number of HYIP-related topics suddenly increases in 2015. From each topic, we obtained 1,187 TXID that are related with HYIP, out of which 1,187 TXID, 81 TXID are posted with investment proof while 1,106 are with payout proof. From 1,187 transactions, 2,134 HYIP operators' Bitcoin addresses have been manually extracted. By checking WalletExplorer.com, 100 of the 2,134 Bitcoin addresses are found to be controlled by online wallets. FIGURE 6 shows the breakdown of online wallets used by HYIPs. As can be seen from this figure most of them use Xapo.com, which, at the time of writing this paper, is one of the biggest online wallets.

B. IDENTIFICATION PROCESS

The idea behind the novel identification process is to extract the characteristics of transactions, which are so-called *features*, by Bitcoin addresses and train a machine learning classifier which detects whether or not a given Bitcoin address is controlled by HYIP or non-HYIP. Our proposed methodology consists of the following steps:

1) Collecting HYIP operators' Bitcoin addresses and other (non-HYIP) addresses;
2) Retrieving transactions;
3) Pre-processing transactions;
4) Extracting features; and
5) Training a machine learning classifier with the features.

For each Bitcoin address, transactions where a given Bitcoin address is included in either inputs and outputs are retrieved from the blockchain. If the entity-based scheme is applied, not only a single Bitcoin address but other addresses that may be controlled by the same user are also retrieved by applying the AC technique. Each transaction is then pre-processed for feature extraction in two steps: (a) Change removal; and (b) Digit conversion.

After pre-processing, a number of features are calculated, e.g. f_{TX}, which is the frequency of transactions per day. Once the features are calculated, a machine learning classifier, e.g. RF (Random Forests) [25], XGBoost (eXtreme Gradient Boosting) [26], is trained by using labels (HYIP or non-HYIP) together with a set of the extracted features. After the classifier is properly trained, whether a given Bitcoin address is operated for HYIP can be inferred. In particular, when a Bitcoin address whose label is unknown is given, then the set of features is calculated. These features are used as inputs into the trained classifier and it can be clarified whether or not a given Bitcoin address is used for HYIP operation. Next, the detailed algorithms will be described.

1) TRANSACTIONS RETRIEVAL FROM BLOCKCHAIN

For each Bitcoin address, transactions, for which a given Bitcoin address is included in inputs or outputs, are retrieved from the blockchain. However, since some Bitcoin addresses have a large number of transactions, to reduce the computational complexity, at most n_{TX} subsequent transactions are retrieved. In the entity-based scheme, not only the given address but also other addresses controlled by its user are also retrieved with the help of AC. As far as the operation of the AC is concern, 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 possess a large number of Bitcoin addresses. Hence, to reduce the computational time, we randomly sample n_{addr} addresses from them.

2) TRANSACTIONS PRE-PROCESSING

The retrieved transactions are then pre-processed for feature extraction by employing a four-step procedure. The first step removes change parts in spent transactions, which are transactions where the given Bitcoin address spends Bitcoin, i.e. the address appears in the inputs of the transactions. FIGURE 7(a) illustrates an example of this phase. In the following, it is assumed that a given Bitcoin address is `1abc...` and wants to send 1 `BTC` from its controlled 1.9 `BTC` to the receiver `1def...`. Since in Bitcoin, `BTC` cannot be split, the sender of Bitcoin must specify another Bitcoin address to receive a change. Hence, if the same Bitcoin addresses appear in both the inputs and outputs, it is safe to assume that the sender received his/her change of 0.9 `BTC` to `1abc...`.

The second step removes any outputs other than the given Bitcoin address in the received transactions where the given Bitcoin address receives Bitcoin. In other words, such

Inputs	Outputs	Amount
1abc…	1def…	1 BTC
	1abc…	0.9 BTC

→

Inputs	Outputs	Amount
1abc…	1def…	1 BTC

(a)

Inputs	Outputs	Amount
1c3h…	1abc…	1 BTC
	1ghi…	0.9 BTC
	1jkl…	0.8 BTC
	1cUX…	0.2 BTC

→

Inputs	Outputs	Amount
1c3h…	1abc…	1 BTC

(b)

Datetime	Inputs	Outputs	Amount
2018-01-01 01:02:03	1abc…	1def…	0.001 BTC

↓

Datetime	Inputs	Outputs	Amount
2018-01-01 01:02:03	1abc…	1def…	14.586 **USD**

(c)

Inputs	Outputs	Amount
1def…	1abc…	123 USD

→

Inputs	Outputs	**Digit**
1def…	1abc…	$\mathbf{10^2}$

(d)

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

(e)

FIGURE 7. An example of pre-processing on Bitcoin address `1abc...`'s transactions. (a) Change removal of spent transactions. (b) Removal of unnecessary outputs in received transactions. (c) Currency conversion from BTC to USD. (d) Extracting the digit of USD amount. (e) Pre-processed transactions.

TABLE 1. Recent USD/BTC daily conversion rates.

DATE	RATE (USD/BTC)
2018-09-01	7,100.95
2018-09-02	7,247.94
2018-09-03	7,260.95
2018-09-04	7,326.85
2018-09-05	7,113.07
2018-09-06	6,433.27
2018-09-07	6,444.80

an address only appears at the outputs of the transactions because any other outputs entries are not related with the given Bitcoin address. FIGURE 7(b) illustrates an example of pre-processing against the received transactions of `1abc...`. In this example, a Bitcoin address `1c3h...` sends not only 1 `BTC` to `1abc...` but also to `1ghi...`, `1jkl...`, and `1cUX...`. Since `1abc...`'s transactions are pre-processed, outputs except for `1abc...` are removed.

After removing the unnecessary parts of transactions, the third step of the pre-processing is to convert the unit from BTC to USD and extract the digits of each output entry in the transactions. A daily currency conversion rate offered by

TABLE 2. The list of calculated features.

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

Blockchain.info [27] is leveraged for this currency conversion. TABLE 1 shows typical daily conversion rates available at the time of writing this paper. As shown in FIGURE 7(c), the amounts in the transactions are converted from BTC to USD, e.g. by using TABLE 1. Then, since we are more interested in the digit of transacted amount than the exact amount, the amount x in USD is converted to $10^{\lfloor \log_{10}(x) \rfloor}$, where $\lfloor r \rfloor$ denotes the greatest integer less than or equal to the real number r. FIGURE 7(d) illustrates an example of such digit conversion, where 123 USD are converted to 10^2.

The last step of the pre-processing is to label "Spent", "Received", or "Coinbase" according to the transaction type. If a given Bitcoin address spends/receives Bitcoin, its type is labelled as "Spent"/"Received", respectively. Similarly, if a transaction is coinbase, it is labelled as "Coinbase". For instance, we assume to process the transactions where a given Bitcoin address `1abc...` is involved. If this address is included in the inputs of a transaction, this means that `1abc...` is willing to spend Bitcoin and thus this transaction is labelled as "Spent". Similarly, if `1abc...` has appeared on an output of a transaction, it means that `1abc...` receives Bitcoin and thus it is labelled as "Received". Finally, if any input is empty while `1abc...` has appeared on its output, such transaction is "Coinbase", which is a special transaction for miners to receive rewards in return for mining.

3) FEATURE EXTRACTION

After the pre-processing phase, features are extracted from the pre-processed transactions. TABLE 2 lists the calculated features and their meaning. The transaction characteristics of each Bitcoin address are represented as a set of features. Some services very frequently issue transactions, but some may not. Apart from f_{TX}, which has been previously introduced, r_{received} and r_{coinbase} are the features that represent the ratio among spent, received, and coinbase transactions. For example, the feature of mining pool's r_{coinbase} may be much higher than that of the other services. $f_{\mathrm{spent}}(i)$ and $f_{\mathrm{received}}(i)$ are features that characterize the amount of money which is frequently transferred. For example, the transferred value of faucet may be typically small, e.g. less than 1 USD which is in sharp contrast to the everyday marketplace where it is usually much higher than 1 USD. r_{payback} is a service which

pays back to the Bitcoin addresses that spent some amount of Bitcoin before. Hence, r_{payback} of HYIP is much higher than the other features, since HYIP often pays back some money to investors. In contrast, pay back may seldom occur in faucet and mining pool.

4) TRAINING A SUPERVISED MACHINE LEARNING CLASSIFIER BASED ON FEATURES

A supervised machine learning classifier is trained by combining calculated features with HYIP or non-HYIP labels. It is underlined that any classifier can be used, e.g., RF [25] and XGBoost [26]. The fundamental idea of training a supervised machine learning classifier is to find the best splitting functions with given features. In other words, the process of training a classifier is to train splitting functions which effectively classify a given data into the correct class. Once a classifier is trained, it can be used to classify unlabeled Bitcoin addresses into HYIP or non-HYIP. In particular, when a Bitcoin address whose label is unknown is given, the set of features is calculated. Then, the features are used as inputs to the classifier which decides, whether or not, a given address is employed by a HYIP operator.

IV. PERFORMANCE EVALUATION AND DISCUSSION

The proposed HYIP operators' Bitcoin addresses classification scheme is evaluated using both scraped HYIP operators' Bitcoin addresses and non-HYIP ones. Specifically, TPR (True Positive Ratio) and FPR (False Positive Ratio) are evaluated by means of the classification accuracy. On the one hand, TPR is defined as the ratio of correctly classified HYIPs addresses. On the other hand, FPR is defined as the ratio of misclassified non-HYIPs addresses. TPR and FPR are evaluated by 10-fold cross validation. The evaluation is repeated 100 times and the classification results are averaged by its results. The detection performance with different classifiers is then evaluated. Furthermore computational time is evaluated by varying n_{TX}. In addition, we qualitatively evaluate how each feature contributes to classification using the following expression:

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

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

Our entire dataset consists of 26,967 non-HYIP addresses and 2,026 HYIP operators' Bitcoin addresses. Non-HYIP addresses are collected from the websites WalletExplorer.com and Blockchain.info/tags and their classes are listed in TABLE 3. In contrast, HYIP operators' addresses are collected by the proposed methodology previously presented in Section III-A. Our dataset is disclosed in our git repository, so that others can reproduce the obtained results.[2] After applying address clustering to these Bitcoin addresses, the number of their owners are found to be and 1,813 and 955. Hence, when evaluating the address-based scheme, the dataset size is 28, 993 (= 26, 967 + 2, 026) whereas for the owner-based scheme, it is 2, 768 (= 1, 813 + 955). By down-sampling the dataset, we vary the ratio of HYIP operators' Bitcoin addresses in the entire dataset, i.e. r_{HYIP} takes values from 0.1 to 0.5. Note that, since the number of Bitcoin addresses depends on the classes, the entire dataset is down-sampled so that the number of each class's Bitcoin addresses is all of the same length. As will be shown later in Section IV-A, since the classification performance of the RF outperforms other classifiers, the RF has been used as the machine learning classifier [25], in which the following parameters were varied: (i) the number of decision trees, i.e. N_{tree}; and (ii) the number of used features in each split in a decision tree, i.e. M_{try}. We set N_{tree}= 500 and M_{try}= $\sqrt{|F|} = \lfloor\sqrt{26}\rfloor = 5$ where $|F|$ is the number of features, respectively. We further used the `mlr` package in R language for evaluation [28] and employed the blockchain with block height from 1 to 452,242, taken for the time period from Jan. 9, 2009 up to Feb. 9, 2017. The blockchain is processed in the SQLite database with the help of *blockparser* [29]. The entire procedure is computationally executed on a Linux workstation equipped with Intel(R) Xeon(R) CPU E5-2603 v4 @ 1.70GHz and 128 GiB RAM.

TABLE 3. Non-HYIP classes used for our dataset.

CLASS	DESCRIPTION
BitcoinTalk User	User on a Bitcoin online forum bitcointalk.org.
Donation	Account to accept donation.
Exchange/Wallet	Exchanging among fiat currencies, e.g. US Dollar and Japanese Yen, and Bitcoin and manages users' Bitcoin
Faucet	Service that offers small amount of Bitcoin in return for solving CAPTCHA, or clicking advertisements.
Gambling	Gambling games, e.g. dice and roulette.
Lending	Lending platform for business loans.
Marketplace	Payment service, e.g. escrow, is offered in an online marketplace.
Mining Pool	Team of miners that share their computational power to find blocks. If one of the miners of a team found a block, its incentive is shared by the miners in the team.
Mixer	Service that mixes several Bitcoin transactions to avoid Bitcoin flow tracking.
Payment	Payment service that accepts Bitcoin for the payment of online shopping.

Next, the detection performance comparison between commonly used classifiers will be presented, followed by the impact of n_{TX}, and r_{HYIP} on the classification performance. Then, the misclassified non-HYIP classes and the impact of the contributed features on the accuracy of the proposed classification process are discussed. Finally, to show that our methodology is practical, unlabeled Bitcoin addresses are tested with the trained classifier and HYIP owners' Bitcoin addresses are newly obtained.

A. PERFORMANCE COMPARISON OF DIFFERENT CLASSIFIERS

In order to choose the most appropriate classifier to test the proposed methodology, several readily available

[2] https://goo.gl/k5PCOZ

TABLE 4. Comparison of classification results by different classifiers.

Classifier	TPR	FPR
RF	**0.9433**	**0.0641**
XGBoost	0.9393	0.0657
Neural Network	0.9240	0.0868
SVM	0.9139	0.0819
k-NN	0.8789	0.1137

TABLE 5. Computational time required for processing the transactions of a Bitcoin address. (i) TX (transactions) retrieval; (ii) Pre-processing & feature extraction; and (iii) Prediction.

	Time [seconds]		
n_{TX}	TX Retrieval	Pre-processing & Feature Extraction	Prediction
100	1.9	71.7	0.19
1,000	5.1	685.4	0.20
10,000	77.4	8139.4	0.21

classifiers have been selected and their performance has been tested. In particular, the following five classifiers have been considered:

- RF
- XGBoost
- Neural network (feed-forward neural network)
- SVM (Support Vector Machine)
- k-NN (Nearest Neighbors)

Furthermore, the following hyperparameters for each classifier have been chosen: The number of trees for RF is set to 100. The number of rounds for XGBoost is set to 100. In neural network, the number of hidden layer and units are 1 and 5, respectively. RBF (Radial Basis Function) is chosen for SVM's kernel function and γ is set to 0.04. $k = 9$ is set for k-NN (Nearest Neighbors). Furthermore, the owner-based scheme with currency conversion is adopted for each classifier. The performance metrics (TPR and FPR) are evaluated under $r_{\mathrm{HYIP}} = 0.5$. TABLE 4 presents the comparison of classification performance for the five classifiers. From these results, it can be seen that the RF classifier outperforms the other four. As such it will be employed to obtain the detailed performance result which will be presented in the next section. It is also noted that, with the exception of the k-NN classifier, all the other four achieve similar performance. Consequently, it can be claimed that an additional advantage of our novel methodology is that it achieves excellent detection performance regardless of the choice of the employed classifier provided that specific hyperparameters are set for each classifier.

B. IMPACT OF n_{TX}

In the proposed methodology, each feature is calculated from n_{TX} subsequent transactions of each address or owner to shorten, as much as possible, the computational time without sacrificing the classification accuracy. Hence, (i) classification accuracy (TPR and FPR) and (ii) computational time are evaluated by varying n_{TX}. FIGURE 8 illustrates the TPR and FPR of the owner-based and address-based schemes vs. n_{TX}, when $r_{\mathrm{HYIP}} = 0.5$. These results clearly show that both TPR and FPR improve as n_{TX} increases up to $n_{\mathrm{TX}} = 100$. In other words, for $n_{\mathrm{TX}} < 100$, the transaction history is not accurately characterized. TABLE 5 shows the computational time required to process the transactions of a Bitcoin address for $n_{\mathrm{TX}} = 100, 1,000$, and $10,000$. In particular, it presents the computational time of three phases, i.e. (i) TX (transactions) retrieval; (ii) Pre-processing & feature extraction; and (iii) Prediction with a supervised machine learning classifier. Clearly, the computational time of TX retrieval and pre-processing & feature extraction increases as n_{TX} gets larger. In particular, when $n_{\mathrm{TX}} = 10,000$, more than two hours are required to perform pre-processing and feature extraction of a Bitcoin address. On the contrary, the computational time for prediction is rather insensitive to n_{TX} variations. Thus $n_{\mathrm{TX}} = 100$ has been selected for further evaluation by considering the compromise between classification accuracy and computational time to obtain the various performance results.

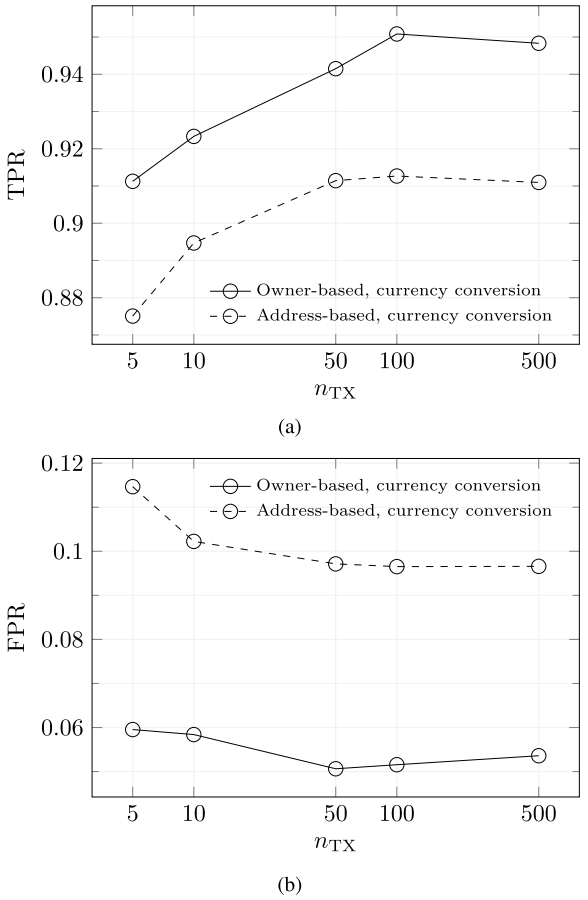

FIGURE 8. Classification performance by varying n_{TX}. (a) TPR. (b) FPR.

C. CLASSIFICATION PERFORMANCE BY r_{HYIP}

FIGURE 9 illustrates the TPR and FPR performance vs. r_{HYIP} for the owner-based and address-based schemes, with or without currency conversion. It can be observed

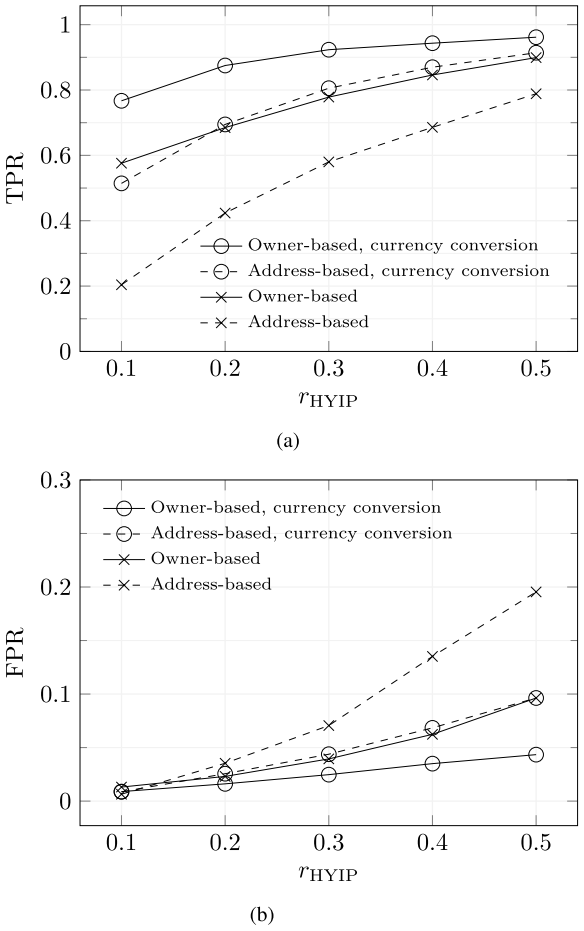

FIGURE 9. Classification performance by varying the ratio of HYIP r_{HYIP}. (a) TPR. (b) FPR.

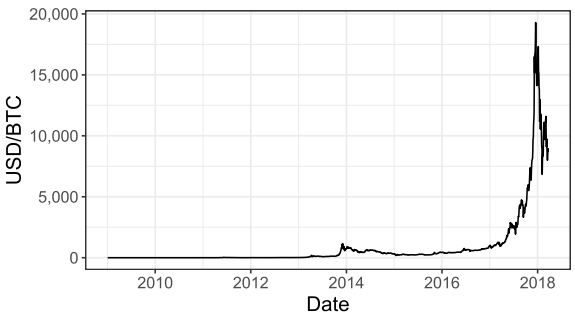

FIGURE 10. USD/BTC exchange rates. (Source: Blockchain.info).

from FIGURE 9(a) that, for all four schemes, TPR gradually improves as r_{HYIP} increases. For example, when $r_{\mathrm{HYIP}} = 0.5$, meaning that an equal number of HYIP and non-HYIP Bitcoin addresses are trained and classified, TPR $= 0.95$ when the owner-based scheme with currency conversion is applied. Even in the case of $r_{\mathrm{HYIP}} = 0.1$, the TPR performance remains high at 0.91. It is interesting to note that TPR improves when the owner-based scheme is employed. In other words, this means that in this case the characteristics of the history of transactions are better incorporated when the address clustering technique is considered. Furthermore,

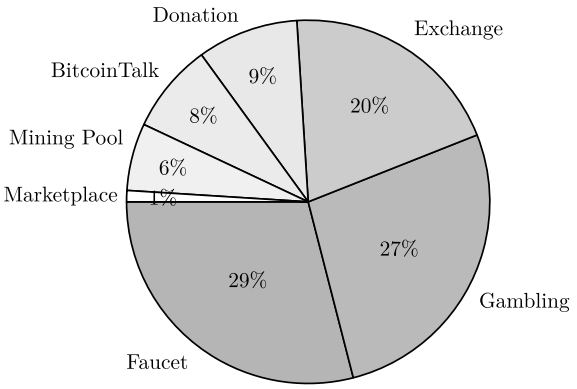

FIGURE 11. Breakdown of classes mis-classified as HYIP.

by applying currency conversion, the effects of the high fluctuation of the Bitcoin price, which has been recently witnessed (see FIGURE 10), on the performance of the proposed scheme can be significantly reduced. For example, even if a single address is given, the address-based scheme achieves TPR $= 0.90$.

On the other hand, as it can be observed from FIGURE 9(b), as r_{HYIP} increases FPR also increases, i.e. the classification performance decreases. This is actually not surprising because fewer number of non-HYIP classes can be trained when r_{HYIP} increases. Similarly with TPR, the FPR of the owner-based schemes is again better as compared to that of the address-based scheme. Furthermore, the FPR of the scheme with currency conversion is also better than that of the scheme without currency conversion. Note that when $r_{\mathrm{HYIP}} = 0.5$, FPR of the owner-based scheme with currency conversion is still less than 5%.

D. MISCLASSIFIED NON-HYIP CLASSES

It is important to elaborate on what types of non-HYIP Bitcoin addresses are mis-classified. FIGURE 11 shows the breakdown of such mis-classified non-HYIP classes. These results have been obtained through our experiments by counting the mis-classified case, divided by the total cases, when $r_{\mathrm{HYIP}} = 0.5$ and the owner-based scheme with currency conversion is applied. From this figure, the top two mis-classified classes are (i) Faucet and (ii) Gambling. We first explain why so many faucets are recognized as HYIP. Faucet is a service that offers small amounts of Bitcoin in return for solving CAPTCHA, or clicking advertisements. To clarify the relationship between HYIP and faucet, we verify how much faucet Bitcoin addresses appear in the transactions related with HYIP. As a result, 95 of 360 faucet addresses in our dataset are found to transfer Bitcoin to HYIP operators' Bitcoin addresses. From this fact it can be inferred that some HYIP and faucet are operated by same persons. Indeed our research has verified that several faucets are introduced on the investor-based games section in bitcointalk.org. As far as the results for the gambling are concerned, they also make sense as essentially the characteristics of HYIP are similar to that of gambling as both accept investment and return some.

TABLE 6. The 10 highest information gains of the contributing features calculated by the owner-based scheme. A higher value indicates higher contribution to classification.

NAME	INFORMATION GAIN
f_{TX}	0.36
$f_{\mathrm{received}}(-3)$	0.30
r_{received}	0.25
r_{payback}	0.25
N_{inputs}	0.17
$f_{\mathrm{received}}(-1)$	0.14
$f_{\mathrm{received}}(0)$	0.14
$f_{\mathrm{received}}(-2)$	0.13
$f_{\mathrm{received}}(1)$	0.12
N_{outputs}	0.10

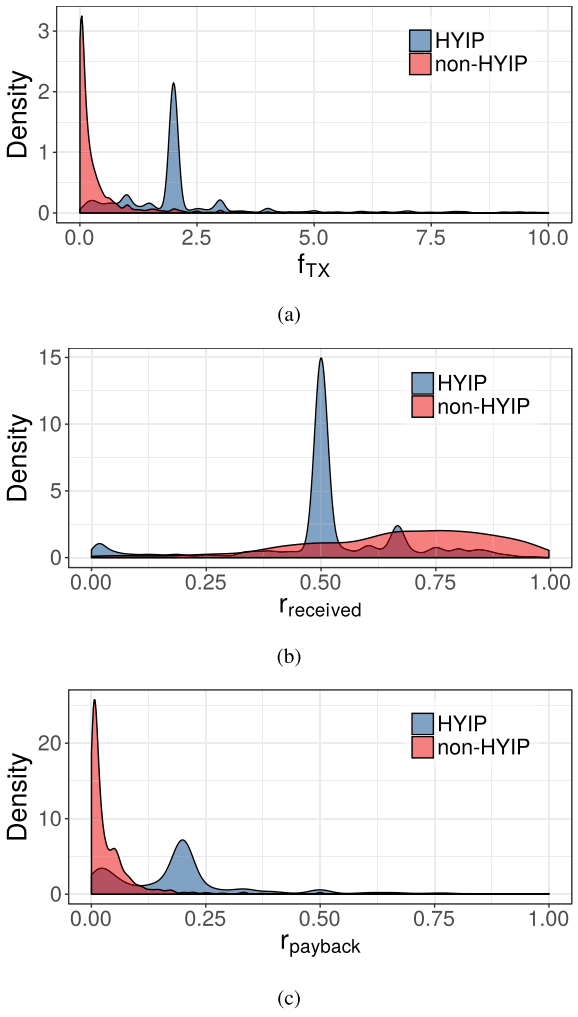

FIGURE 12. Density functions of the three highest contributing features for HYIP and non-HYIP. (a) f_{TX}. (b) r_{received}. (c) r_{payback}.

E. CONTRIBUTING FEATURES

In this section, we discuss the effects on the classification performance of the contributing features which can be used to distinguish HYIP from non-HYIP. TABLE 6 lists the information gains of the 10 highest contributing features calculated by the owner-based scheme, where f_{TX} is the highest contributing feature together with other features, including several $f_{\mathrm{received}}(\cdot)$.

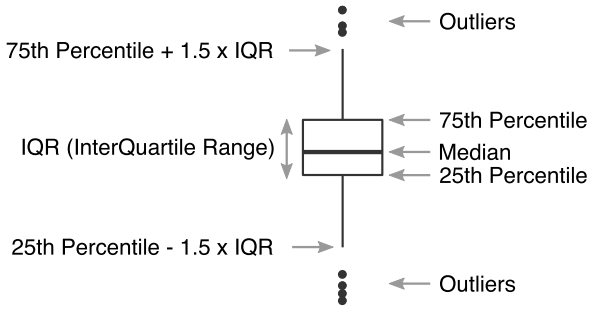

FIGURE 13. Boxplot schematic representation.

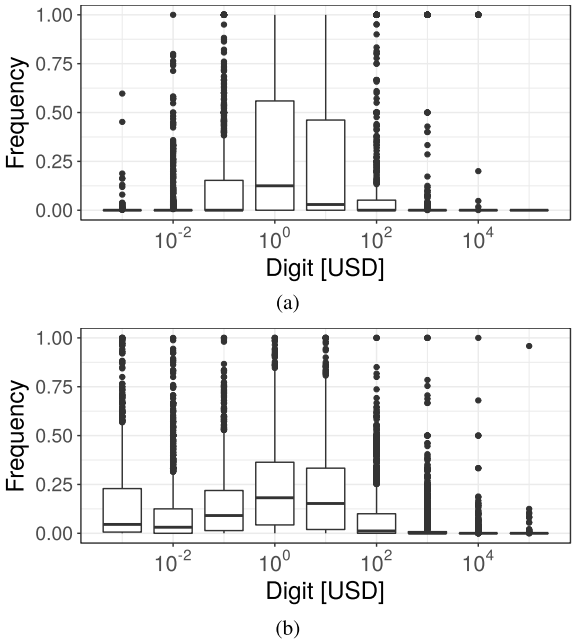

FIGURE 14. The distribution of $f_{\mathrm{received}}(\cdot)$ for HYIP and non-HYIP. (a) HYIP. (b) non-HYIP.

Next the density functions of the three highest contributing features, which are f_{TX}, r_{received}, and r_{payback}, for both HYIP and non-HYIP are presented in FIGURE 12. From these results, it can be concluded that indeed f_{TX} is the most important contributing feature, as there is a clear distinction in the density functions between HYIP and non-HYIP cases where each has a distinct peak at $f_{\mathrm{TX}} \approx 2.0$ and 0.125, respectively.

We then discuss the difference of the distribution of r_{received} by HYIP and non-HYIP. FIGURE 12(b) shows the density function of r_{received} by HYIP and non-HYIP. HYIP's r_{received} is concentrated around 0.5. In contrast, that of non-HYIP is widely distributed from 0.25 to 1.

FIGURE 12(c) shows the distribution of r_{payback} by HYIP and non-HYIP. To focus on HYIP's r_{payback}, a sharp peak can be seen around 0.175, which is much larger than that of non-HYIP. This clearly reflects the intrinsic nature of HYIP, i.e. some small amount of Bitcoin is paid back to investors.

TABLE 7. List of HYIP operators' Bitcoin addresses offered in [8] and successfully detected with our classifier.

	NAME	ADDRESS
1	Nanoindustryinv.com	1Ee9ZiZkmygAXUiyeYKRSA3tLe4vNYEAgA
2	GrandAgoFinance	1MzNQ7HV8dQ6XQ52zBGYkCZkkWv2Pd3VG6
3	Leancy	145SmDToAhtfcBQhNxfeM8hnS6CBeiRukY
4	Minimalist10	1FuypAdeC7mSmYBsQLbG9XV261bnfgWbgB
5	MiniPonziCoin	1F8ZKpjMDpnpF79mZ1pxZRoNKZgXm4Tf1d
6	120cycle	1E5MCTtXn7n2svpZ1bDHZXndY9K7qQeqZP
7	10PERCENTBTC	1BtcBoSSnqe8mFJCUEyCNmo3EcF8Yzhpnc
8	PonziIO	1ponziUjuCVdB167ZmTWH48AURW1vE64q
9	LaxoTrade	1LaxoTrQy51LnB289VmoSAgN6J6UrJbfL9
10	OpenPonzi	1BmZW65ZoeLa1kbL9MPFLfkS818mqFUSma
11	BTC-doubler.us	1AQp51H22WHDzLgK64NoUo3Bg3T183QR22
12	BTC-doubler.com	178BzARKjkszrTyx4TxBKHhzGLZijdE26e
13	investorbitcoin.com	1CpVAEg4BgVzjiHshgeZfitZLV1t1zo6Qg
14	Ponzi120	12PoNZiEtabwkCU4YFffshWNF1cRiAk5nq
15	RockwellPartners	139eeGkMGR6F9EuJQ3qYoXebfkBbNAsLtV
16	Twelverized	114Ap9G5nu78vESC648amPwSeqUorPtV5L
17	CRYPTOSX2	19YZYfMB3mfX8AixzV7aLqXuViDcntrfcK
18	1hourbtc.pw	1BsjsaHST2Qohs8ZHxNHeZ1UfWhtxoKHEN
19	bitcoindoubler.fund	1FNtgGsHhymmEUMXrMiFeMtZbuagnnS59c
20	doublebitcoin.life	1zmeu5BeWBprWyPv5ntNZKR7uThXaG9ic
21	bitcoincopy.site123.me	1EaSVdRuzcz4yjnTmibabyyrczvaQS8hAJ
22	bitcoinprofit2	1AXTqWYz1Bd3LZnq1Zf9vsgFBpqrKkHopx
23	invest4profit	1PZ8E5oT7EUVgEVz1Ggc7bjXe2byxr7wxG
24	1getpaid.me	1GetPaiDxjEuWN3KJTnY9Cbqv9QcR8zcME
25	Ponzi.io(change)	14ji9KmegNHhTchf4ftkt3J1ywmijGjd6M
26	igjam.com	1AQxcdPgMTTQghPXt1EXHU8vEjSn2kYrPQ
27	7dayponzi	195o79saDhUNHJ4DeMBYMekLmrQ848APxA
28	world-btc.online	1A88teD6QqXRHBMCyCkoxxBQHpJAztUz6e
29	bestdoubler.eu	13NZxtAnKk5mbCUHpxHqKwWTDJzFHMGHLh
30	bitcoindoubler.prv.pl	18Smkvyf3gJN4z59FhjJsCu6NhSYmZkNvG

Since several $f_{\text{received}}(\cdot)$ are listed on the contributing features in TABLE 6, we have also obtained the distribution of $f_{\text{received}}(\cdot)$ for both HYIP and non-HYIP. For these distributions, since we want to observe the distributions of multiple $f_{\text{received}}(\cdot)$, it is more convenient to present them using the boxplot representation illustrated in FIGURE 13. In this figure, the top and bottom of a box and a horizontal line in the box indicate the range of quartiles, where each box is structured with 75$^{\text{th}}$ percentile, 25$^{\text{th}}$ percentile, and median (50$^{\text{th}}$ 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'', i.e. the length of the box is rather short, it means that such feature values are sharply distributed within a class. In contrast, if the size of the box is large, it means that such a feature value is widely distributed within a class.

FIGURE 14 illustrates the distribution of $f_{\text{received}}(\cdot)$, from where it can be clearly observed that the difference of the distribution of $f_{\text{received}}(-3)$, $f_{\text{received}}(-2)$, $\cdots$, and $f_{\text{received}}(1)$ by HYIP and non-HYIP. HYIP typically receives Bitcoin ranging from 10^{-1} to 10^{2}, while non-HYIP receives more widely from 10^{-3} to 10^{2}. In particular, the median of $f_{\text{received}}(-3)$ of HYIP is much smaller than that of non-HYIP. In fact this explains why the $f_{\text{received}}(-3)$ is the most contributing feature among $f_{\text{received}}(\cdot)$.

F. EXPERIMENTS USING RANDOMLY CHOSEN BITCOIN ADDRESSES

In order to generalize the procedure for evaluating the proposed classifier's ability to detect HYIP operators' Bitcoin addresses, we have run additional experiments without considering our previous dataset collection. In particular, we tested our classifier against the HYIP address list offered in [8] which consists of 32 HYIP operators' Bitcoin addresses. For this evaluation, the RF classifier with $N_{\text{tree}} = 500$ has been chosen. From a total of 32 HYIP addresses listed in [8], the proposed methodology has successfully detected 30 (see TABLE 7) achieving a HYIP detection accuracy of 93.75%. This result is very encouraging and in fact verifies that the proposed methodology can be effectively used for the forensics of Bitcoin-related fraud.

V. CONCLUSIONS

In this paper, we have proposed a novel HYIP identification methodology which accurately classifies whether or not a specific Bitcoin address belongs to HYIP operators. Apart from introducing and analyzing the performance of a generic and accurate HYIP owners' Bitcoin addresses identification methodology, we have also proposed a novel dataset collection approach, which significantly increases the number of HYIP owners' Bitcoin addresses obtained through scraping the HYIP-related topics in the Bitcoin forum. A solid identification methodology has been proposed, which consists of several key ideas such as unit conversion and a sampling approach, to realize the lightweight, fast, and accurate identification. The idea behind the identification scheme is to extract the features of transactions and train a machine learning classifier that outputs whether a specific Bitcoin address is controlled by HYIP or non-HYIP.

Through various systematic simulation experiments it has been shown that the owner-based approach with currency conversion achieves TPR (True Positive Rate) $=$ 0.95 and FPR (False Positive Rate) $=$ 0.049. In addition, the proposed sampling approach has been shown to be effectively reducing the computation complexity while maintaining the high classification accuracy. We have also explained the reason why several features, e.g. f_{TX}, $f_{\text{received}}(10^{-3})$, and r_{received}, contribute for identification by analyzing the distribution of features by HYIP and non-HYIP. Finally, in order to verify the proposed classifier's ability to detect HYIP operators' Bitcoin addresses, our classifier has been tested against a HYIP address list offered in [8] and proven that its detection accuracy achieves 93.75%, which is a very positive result for the viability of Bitcoin-related fraud detection.

REFERENCES

[1] S. Nakamoto, ''Bitcoin: A peer-to-peer electronic cash system,'' Working Paper, 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf
[2] M. Vasek and T. Moore, ''There's no free lunch, even using Bitcoin: Tracking the popularity and profits of virtual currency scams,'' in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, Jul. 2015, pp. 44–61.
[3] M. Vasek and T. Moore, ''Analyzing the Bitcoin Ponzi scheme ecosystem,'' in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, Feb. 2019, pp. 101–112.

[4] Biography.com. (2015). *Charles Ponzi-Criminal-Biography*. Accessed: Oct. 10, 2018. [Online]. Available: https://www.biography.com/people/charles-ponzi-20650909
[5] J. Simser, "Bitcoin and modern alchemy: In code we trust," *J. Financial Crime*, vol. 22, no. 2, pp. 156–169, May 2015.
[6] SEC.gov. (Jul. 2013). *SEC Charges Texas Man With Running Bitcoin-Denominated Ponzi Scheme*. Accessed: Oct. 10, 2018. [Online]. Available: https://www.sec.gov/news/press-release/2013-132
[7] K. Toyoda, T. Ohtsuki, and P. T. Mathiopoulos, "Identification of high yielding investment programs in Bitcoin via transactions pattern analysis," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.
[8] M. Bartoletti, B. Pes, and S. Serusi, "Data mining for detecting Bitcoin Ponzi schemes," in *Proc. Crypto Valley Conf. Blockchain Technol. (CVCBT)*, Jun. 2018, pp. 75–84.
[9] E. Androulaki, G. O. Karame, M. Roeschlin, T. Scherer, and S. Capkun, "Evaluating user privacy in Bitcoin," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, Apr. 2013, pp. 34–51.
[10] D. Ron and A. Shamir, "Quantitative analysis of the full Bitcoin transaction graph," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, Apr. 2013, pp. 6–24.
[11] 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. Conf. Internet Meas. Conf.*, Oct. 2013, pp. 127–140.
[12] M. Fleder, M. S. Kester, and S. Pillai, "Bitcoin transaction graph analysis," Feb. 2015, *arXiv:1502.01657*. [Online]. Available: https://arxiv.org/abs/1502.01657
[13] J. D. Nick, "Data-driven de-anonymization in Bitcoin," Ph.D. dissertation, ETH Zurich, Zürich, Switzerland, 2015.
[14] T. Neudecker and H. Hartenstein, "Could network information facilitate address clustering in Bitcoin?" in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, Nov. 2017, pp. 155–169.
[15] M. Lischke and B. Fabian, "Analyzing the Bitcoin network: The first four years," *Future Internet*, vol. 8, no. 1, p. 7, 2016.
[16] D. McGinn, D. Birch, D. Akroyd, M. Molina-Solana, Y. Guo, and W. J. Knottenbelt, "Visualizing dynamic Bitcoin transaction patterns," *Big Data*, vol. 4, no. 2, pp. 109–119, Jun. 2016.
[17] M. Rahouti, K. Xiong, and N. Ghani, "Bitcoin concepts, threats, and machine-learning security solutions," *IEEE Access*, vol. 6, pp. 67189–67205, 2018.
[18] M. Spagnuolo, F. Maggi, and S. Zanero, "BitIodine: Extracting intelligence from the Bitcoin network," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, Nov. 2014, pp. 457–468.
[19] K. Jarvis. (2014). *Cryptolocker Ransomware*. [Online]. Available: http://www.secureworks.com/cyber-threat-intelligence/threats/cryptolocker-ransomware
[20] T. Moore, J. Han, and R. Clayton, "The postmodern Ponzi scheme: Empirical analysis of high-yield investment programs," in *Proc. Int. Conf. Financial Cryptogr. Data Secur.*, Feb. 2012, pp. 41–56.
[21] TIME. (Jan. 1931). *Business Finance: Ponzi Payment*. Accessed: Oct. 10, 2018. [Online]. Available: http://content.time.com/time/magazine/article/0,9171,930255,00.html
[22] F. Smith, "Madoff Ponzi scheme exposes the myth of the sophisticated investor," *Univ. Baltimore Law Rev.*, vol. 40, pp. 215–282, Jul. 2010.
[23] J. Drew and T. Moore, "Automatic identification of replicated criminal websites using combined clustering," in *Proc. IEEE Secur. Privacy Workshops*, May 2014, pp. 116–123.
[24] J. Neisius and R. Clayton, "Orchestrated crime: The high yield investment fraud ecosystem," in *Proc. APWG Symp. Electron. Crime Res.*, Sep. 2014, pp. 48–58.
[25] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[27] BTC to USD. *Bitcoin to US Dollar Market Price–Blockchain*. Accessed: Oct. 10, 2018. [Online]. Available: https://blockchain.info/charts/market-price?timespan=all
[28] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, Z. Jones, and G. Casalicchio, *MLR: Machine Learning in R, R package version 2.9*. [Online]. Available: https://CRAN.R-project.org/package=mlr
[29] znort987. *GitHub-Znort987/blockparser: Simple C++ Bitcoin Blockchain Parser*. Accessed: Oct. 10, 2018. [Online]. Available: https://github.com/znort987/blockparser

KENTAROH TOYODA was born in Tokyo, Japan, in 1988. He received the B.E., M.E., and Ph.D. (Engineering) degrees from the Department of Information and Computer Science, Keio University, Yokohama, Japan, in 2011, 2013, and 2016, respectively. He was an Assistant Professor with Keio University from 2016 to 2019, and is currently a Scientist with Agency for Science, Technology, and Research (A*STAR), Singapore Institute of Manufacturing Technology, Singapore, and a Visiting Assistant Professor with the Faculty of Science and Technology, Keio University. His research interests include blockchain analysis and application, security and privacy for emerging systems and services, and wireless healthcare monitoring. He is a member of IEEE and IEICE. He was a recipient of the IEICE Communication Society Encouragement Awards, in 2012 and 2015, respectively, the Telecom System Technology Encouragement Award, in 2015, and the Fujiwara Foundation Award, in 2016.

P. TAKIS MATHIOPOULOS (SM'94) received the Ph.D. degree in digital communications from the University of Ottawa, Ottawa, ON, Canada, in 1989. From 1982 to 1986, he was with Raytheon Canada Ltd., working in the areas of air navigational and satellite communications.

In 1989, he joined the Department of Electrical and Computer Engineering (ECE), University of British Columbia (UBC), Vancouver, BC, Canada, as an Assistant Professor and where he was a Faculty Member, until 2003, holding the rank of a Professor, from 2000 to 2003. From 2000 to 2014, he was the Director and then the Director of Research with the Institute for Space Applications and Remote Sensing (ISARS), National Observatory of Athens (NOA), where he established the Wireless Communications Research Group. From 2000 to 2004, as the ISARS Director, he led the Institute to a significant expansion in R&D growth and international scientific recognition. For these achievements, ISARS has been selected as the National Centre of Excellence, from 2005 to 2008. From 2008 to 2013, he was a Guest Professor with Southwest Jiaotong University, Chengdu, China. He has also been appointed by the Government of China as a Senior Foreign Expert with the School of Information Engineering, Yangzhou University, Yangzhou, China, from 2014 to 2017, and also with the Graduate School of Science and Technology, Keio University, as a Guest Professor (Global), from 2015 to 2016 and from 2017 to 2018, respectively, under the Top Global University Project of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of the Government of Japan. Since 2014, he has been a Professor of telecommunications with the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece. For the last 25 years, he has been conducting research mainly on the physical layer of digital communication systems for terrestrial and satellite applications, and in the fields of remote sensing and the Internet of Things. In these areas, he has coauthored more than 120 journal papers published mainly in various IEEE journals, one book (edited), five book chapters, and more than 140 conference papers.

He has been a member of the Technical Program Committees (TPC) of more than 70 international IEEE conferences. As a Faculty Member UBC, he received an Advanced Systems Institute (ASI) Fellowship and a Killam Research Fellowship. He was also a co-recipient of two best conference paper awards. In 2017, he has been presented by the Satellite and Space Communication Technical Committee of the IEEE Communications Society their 2017 award for outstanding contributions in the field of satellite and

space communications. He has been a TPC Vice Chair for the 2006-S IEEE VTC and 2008-F IEEE Vehicular Technology Conference (VTC), and a Co-Chair of the International Conference on Future Information Technology (FITCE) 2011 and AUTOMOTIVE17. He has regularly served as a consultant for various governmental and private organizations. He has been or currently serves on the editorial board of several archival journals, including the *IET Communications* and the IEEE TRANSACTIONS ON COMMUNICATIONS, from 1993 to 2005. Since 1993, he has been serving on a regular basis as a Scientific Advisor and a Technical Expert for the European Commission (EC). From 2001 to 2014, he has served as a Greek Representative to high-level committees in the EC and the European Space Agency. He has delivered numerous invited presentations, including plenary and keynote lectures, and has taught many short courses all over the world.

TOMOAKI OHTSUKI received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1990, 1992, and 1994, respectively, where he was a Postdoctoral Fellow and a Visiting Researcher in electrical engineering, from 1994 to 1995. From 1993 to 1995, he was a Special Researcher of Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. From 1995 to 2005, he was with the Science University of Tokyo. From 1998 to 1999, he was with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley. In 2005, he joined Keio University, where he is currently a Professor. He has published more than 140 journal papers and 340 international conference papers. His research interests include wireless communications, optical communications, signal processing, and information theory.

He is a Fellow of the IEICE. He was a recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando ,Memorial Young Engineering Award, the Ericsson Young Scientist Award 2000, the IEEE 1st Asia-Pacific Young Researcher Award 2001, the 2002 Funai Information and Science Award for Young Scientist, the 5th International Communication Foundation (ICF) Research Award, the 2011 IEEE SPCE Outstanding Service Award, the 28th TELECOM System Technology Award, ETRI Journal's 2012 Best Reviewer Award, and the 9th International Conference on Communications and Networking in China 2014 (CHINACOM'14) Best Paper Award. He served as a Chair for the IEEE Communications Society, Signal Processing for Communications and Electronics Technical Committee. He has served as a General Co-Chair and a Symposium Co-Chair of many conferences, including the IEEE GLOBECOM 2008, SPC, IEEE ICC 2011, CTS, and IEEE GLOBECOM 2012, SPC. He served as a Technical Editor for the *IEEE Wireless Communications Magazine*. He is currently serving as an Editor for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, and Elsevier *Physical Communications*. He gave tutorials and keynote speech at many international conferences, including the IEEE VTC, IEEE PIMRC, and so on. He is also serving as a President of Communications Society of the IEICE.