Article | Received 1 November 2022; Accepted 6 January 2023; Published 14 January 2023
https://doi.org/10.55092/blockchain20230004

Detecting phishing gangs via taint analysis on the Ethereum blockchain

Kangrui Huang[1], Weili Chen[2,*] and Zibin Zheng[3]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]School of Information Science and Technology Guangdong University of Foreign Studies, Guangzhou, China
[3]School of Software Engineering, Sun Yat-Sen University, Guangzhou, China;
E-mail: zhzibin@mail.sysu.edu.cn

* Correspondence author; E-mail: mathutopia@163.com.

Abstract: Blockchain technology has created a new cryptocurrency world and attracted a lot of attention. It also attracts scams, for example, phishing scam, a typical fraud, has been found making a notable amount of money in the blockchain ecosystem, which has a very negative impact. Considering the whole life cycle of a phishing scam, this paper proposes the concept of a phishing gang, that is, a set of accounts that serve for phishing activity and belong to the same entity on the blockchain. As phishers often use multiple accounts to commit phishing scams and money laundering, detecting phishing gangs in the blockchain ecosystem is a real and critical problem. To help deal with this issue, this paper proposes a method of detecting phishing gangs on the Ethereum blockchain. Specifically, we first construct a transaction network with a graph structure by mining the transaction record and the account labels of the Ethereum blockchain. Next, we propose the base and improvement methods of taint analysis, aiming to evaluate the taint score of each account by tracking the fund flow of phishing accounts. Then, with the results of taint analysis and some heuristic means, all accounts in the transaction network are divided into five categories. Based on this, we propose a heuristics algorithm for phishing gang detection. And we also summarize gang patterns and reveal money laundering in phishing activities. Experimental results indicate that the proposed framework can be used to build a uniform platform to monitor every account on the Ethereum blockchain for early warning of phishing scams and detection of the phishers' money laundering and cashing process.

Keywords: blockchain; ethereum; phishing scam; money laundering; phishing gang; taint analysis

1. Introduction

Recently, blockchain has become one of the most popular buzzwords. It is described as a disruptive technology that is going to revolutionize many traditional industries. Generally speaking, a blockchain can be described as a distributed and trusted database maintained by a peer-to-peer network through a special consensus mechanism. A blockchain usually implements a cryptocurrency (or a virtual currency) and it can be exchanged with other cryptocurrencies or fiat money through *exchanges*. Ethereum [1], a famous blockchain platform that supports smart contracts, can implement plenty of

financing methods and the corresponding cryptocurrency *Ether* becomes the second largest cryptocurrency (the largest is Bitcoin).

The boom in blockchain technology has given rise to a large number of blockchain application scenarios, while at the same time has become a high incidence of various cybercrimes. Initial coin offering (ICO) is a financing method for the blockchain industry, which is used by start-ups to raise cryptocurrency to fund their ideas and businesses by issuing their cryptographic coins. However, more than 10% of ICOs released on Ethereum have been reported to suffer from a variety of scams, including phishing, hack schemes, *etc.*[2] Furthermore, cryptocurrencies are also used for criminal activities in which the perpetrators hope to cover up their traces. Money laundering in the blockchain ecosystem has become an important means of obtaining illegal proceeds. The endless variety of malicious activities and services has seriously hindered the healthy development of blockchain. Therefore, the detection of these fraud activities has become an urgent and critical problem in the blockchain ecosystem and attracted a wealth of research attention.

The phishing scam is a new type of cybercrime that arises along with the rise of online business [3]. At present, it has been found in the blockchain ecosystem. According to the report of *Chainalysis*, there have been approximately 30,000 victims of cybercrime on Ethereum losing on average $7,500 each [4]. A widely known example is the phishing scam on Bee Token ICO [5]. Bee Token is a blockchain-based home-sharing service and plans to launch its public ICO in 2018. However, before the startup runs its token sale, the phisher sent a fake email to would-be buyers stating the opening of the ICO. In the email, the phisher promised the investors a 100% bonus for all contributions within the next 6 hours and a double value of the Bee token within the next 2 months. This phishing scam eventually gathered about $1 million from investors in only 25 hours. These examples show that detecting and preventing phishing scams is an urgent problem on the Ethereum blockchain.

Phishing detection has been extensively studied in the past decades and many methods have been proposed [6-9]. However, phishing scams in the blockchain era have many unique characteristics compared with those in traditional online business scenarios. First of all, instead of cash, cryptocurrencies become phishing targets. And phishers need to use their accounts on the blockchain. Second, the ill-gotten cryptocurrencies have to be cashed through exchanges for fiat money (i.e., to convert the ill-gotten cryptocurrencies into fiat money). And before cashing out, money laundering is often required to prevent the source of illegal funds from being detected. Third, the transaction records of public blockchain are publicly accessible, which provides a new data source for phishing detection.

Based on these new characteristics, new anti-phishing and anti-money laundering methods can be proposed in the blockchain ecosystem. In this work, we focus on the issue of phishing gang detection on the Ethereum platform and propose a comprehensive approach to prevent phishing scams and their money laundering. *Phishing gang* can be defined as the set of accounts that serve for phishing activity and belong to the same entity in the blockchain ecosystem. By mining the transaction records, we propose to build a framework for automatically detecting phishing gangs via taint analysis. Taint analysis is a tracking method for anti-money laundering. As phishing accounts transfer ill-gotten funds out of money laundering, the taint score of each account can be evaluated by tracking and analyzing the transaction flow of these funds, which refers to the correlation between other accounts and phishing accounts. Specifically, as shown in Figure 1, we first collect data including external transactions and internal transactions on the Ethereum ledger, as the input of transaction network construction, and extract some transaction features. Then, account labels are extracted from *Etherscan*(etherscan.io), a

famous block explorer and analytics platform for Ethereum. It usually labels addresses (an address can be seen as an account in the Ethereum platform) and issues warning messages. Finally, a three-step detection framework is established to detect the phishing gangs, and their patterns are classified and displayed. We also revealed and analyzed the phishing money laundering on the basis of phishing gangs.

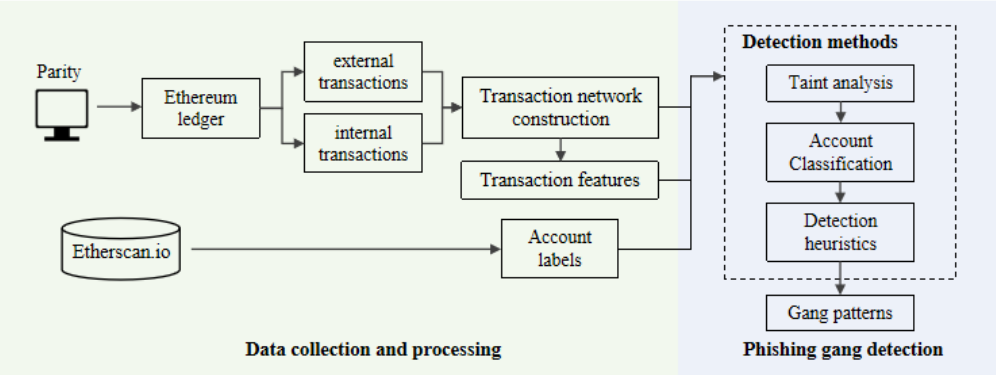


Figure 1. The framework.

The main contributions of this paper can be summarized as follows: 1) to the best of our knowledge, we are the first to define phishing gang and mine the transaction records for phishing gang detection in the Ethereum platform; 2) novel taint analysis method for determining the risk factors of accounts on Ethereum is proposed; 3) the criteria of account classification and a heuristic algorithm for phishing gang detection are presented; 4) gang patterns and the whereabouts of phishing money laundering are analyzed. The proposed framework can be used to give a warning message to investors and to prevent the phishers' cashing process if it is embedded into exchanges and special platforms like decentralized exchanges. Since cybercrime has become a problem that cannot be ignored in blockchain ecology, we hope our preliminary study might serve as a modest spur to induce extensive efforts in this field.

The remainder of the paper is organized as follows. Section 2 provides some background of the Ethereum, anti-money laundering in the blockchain ecosystem, and phishing scam detection. The problem definition and data description are discussed in Section 3. Taint analysis, account classification, and detection heuristics are proposed in Section 4. Experimental results and application are summarized in Section 5. Finally, we draw the conclusion in Section 6.

2. Background and related work

In this section, we provide some background to understand the illegal activities in the blockchain ecosystem. Specifically, we first introduce the Ethereum platform. Then, a review of the current status of anti-money laundering in the blockchain ecosystem was made. Finally, we introduce phishing scam detection.

2.1. Ethereum

After Satoshi Nakamoto used blockchain technology to create Bitcoin in 2009[10], many new blockchain ecosystems and related cryptocurrencies have been introduced. According to market capitalization, Ethereum[11] is the second most popular blockchain after Bitcoin, and it is currently the largest blockchain-based platform that provides a Turing-complete language to support smart contracts[12]. A smart contract is a digital contract residing in the blockchain and can self-execute when the preset conditions are

met. Due to the self-executing and unforgeable characteristics, smart contracts can be used to implement some useful economic functions. For example, it is very easy to create a new digital token[13] through smart contracts in the Ethereum platform. Tokens in the Ethereum ecosystem can represent any fungible tradable goods. To implement some basic features in a standard way, a token is essentially compatible with the Ethereum wallet and other tokens using the same standards, which means that it can be exchanged with Ether (the cryptocurrency of Ethereum, also named ETH) and other tokens conveniently.

2.2. Anti-money laundering in blockchain ecosystem

Anti-money laundering is an analysis of the transaction characteristics of fraud activities after obtaining a certain amount of illegal income and is an important research field in blockchain supervision. The research on the detection of money laundering activities mainly focuses on the detection of mixing services and the investigation of suspicious transaction patterns. Paper[14] used reverse-engineering methods to analyze the operation modes of several money laundering services, and invested Bitcoin to trace anonymized transactions and analyze their money laundering patterns. Paper[15] first proposed the problem of detecting mixing services and solved it as a community outlier detection problem. Once mixing services are detected, it can further analyze whether the addresses or accounts that interact with these services participate in illegal activities. And paper[16] uses feature analysis to characterize the transaction patterns of bitcoin money laundering and develops classifiers to detect money laundering through some network embedding methods, such as Deepwalk, Node2vec, *etc.*

The analysis of money laundering funds flow is to study how the illegal cryptocurrency incomes are transacted on the chain or flow between different blockchain platforms. Paper[17] proposed a method based on Breadth First Search (BFS) algorithm to track the funds flow after Bitcoin fraud thefts. Paper[18] proposed a Bitcoin transaction data analysis framework called Graph Sense, which can track the flow of Bitcoin and find the path of connections between the given address labels. Paper[19] proposed several iterative algorithms for illicit funds flow analysis on the Bitcoin network to assess the trustworthiness of Bitcoin wallet addresses. And paper[20] analyzed several different illegal fund transfer strategies of Bitcoin and uses the method of setting control groups to compare the accuracy of various taint analysis methods.

2.3. Phishing scam detection

The term "phishing" comes from the fact that cyber-attackers are fishing for data but the techniques they used are more sophisticated than simplistic fishing [7]. Phishing is defined as the art of impersonating an official website or a message from authorities aiming to acquire users' private information such as usernames, passwords, and account numbers [8]. It usually begins with sending a fake message that seems to be from an authentic organization to victims, with the purpose of asking them to update their information by clicking a disguised URL that directs them to the phishing website. In realistic phishing scams, various methods have been used to distribute fake messages and URLs, such as e-mail, Google ads [21], Slack, and Reddit [22].

Since the invention of phishing attacks, several methods based on visual similarity [23], association rules [8], and support vector machine (SVM) [9] have been proposed for phishing detection. Moreover, various factors have been proven to be related to phishing detection, such as webpage source code [24], reputation of user location [25], URL features [26], and CSS features [27]. Additionally, there are also some tools in the industry to prevent phishing scams in the blockchain ecosystem. For example,

EtherAddressLookup detects phishing domains and Blocking Policy maintains a phishing URL blacklist. However, these tools only support the detection of phishing websites, and cannot detect fraud accounts that do not make use of websites.

Like in traditional markets, phishing in the blockchain ecosystem also has three stages. Figure 2 shows the life cycle of a phishing scam in the blockchain ecosystem. It is worth noting that most traditional anti-phishing models usually function before the victim is deceived. In the blockchain ecosystem, however, we can build models to detect and prevent phishing scams before and after the victim is deceived. It is feasible because 1) all the transactions are recorded in the public blockchain, and thus we can mine the transaction history to warn suspect phishing accounts; 2) the ill-gotten gains are cryptocurrencies and have to be cashed through exchanges, and thus we can prevent the cashing process by cooperating with the exchanges. Based on the two characteristics, we can build a model to detect the accounts involved in phishing scams and prohibit the phisher to withdraw money from the exchanges.

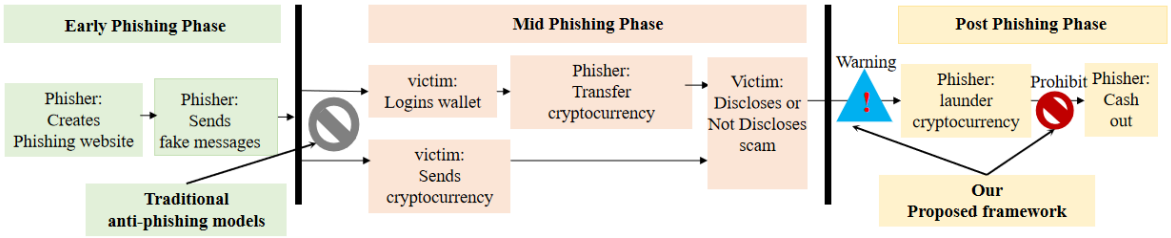


Figure 2. The phishing life cycle and the difference of anti-phishing between traditional and our proposed framework.

3. Problem definition and data description

In this section, we will first give a concrete definition of a phishing gang on the Ethereum blockchain. Then we will provide an overview of the dataset used for detecting phishing gangs, including the transaction network, which is constructed on the basis of transaction records on Ethereum, and the account labels, which are verified accounts with significant attributes.

3.1. Problem definition

As shown in Figure 2, phishing activity in the blockchain ecosystem has three phases. Most traditional anti-phishing models usually function in the early phishing phase. In the blockchain ecosystem, however, we can prevent phishing scams before and after the victim is deceived by detecting the fraud accounts involved in the phishing activity. It is worth noting that a phisher usually controls multiple accounts to put the phishing scams into effect. Specifically, a phisher usually uses several accounts to accumulate cryptocurrency from the victims and plenty of accounts to launder cryptocurrency, aiming to cash through exchanges to get fiat money. The set of these accounts which serve for phishing activity and belong to the same entity is a phishing gang. Detecting a phishing gang is vital in preventing phishing scam as 1) it helps to understand the entire process of a phishing scam and prevent it as a whole; 2) it find out some potential scam accounts which may not be labeled; 3) it reveals the laundering techniques and patterns of a phishing scam. Our framework to detect phishing gangs can function in the post phishing phase, which can provide a more comprehensive understanding of the accounts

github.com/409h/etheraddresslookup
github.com/metamask/eth-phishing-detect

related to phishing scams and give an early warning to investors. When a phishing gang is detected, we can build a "firewall" in the exchanges to prevent the scam accounts from cashing. The firewall can enable victims to identify the whereabouts of the lost funds and recover the funds by legal means. Of course, this firewall requires cooperation between the government and the exchanges. For example, Japan's national police agency announced 170 cases of suspected money laundering related to cryptocurrencies over a period of six months [28]. It is capable of doing this because the government has updated its laws on preventing the transferring of criminal proceeds and issued licenses to cryptocurrency exchanges in order to report suspected cases of money laundering. In this scenario, the proposed framework, combined with other regulatory rules, can work as an effective tool for preventing phishing scams.

3.2. Data description

3.2.1. Transaction network

This study aims to detect phishing gangs on Ethereum based on Ether transaction tracking. To obtain the Ethereum transaction network, we adopt the same method used in [29] to collect data by running an Ethereum client, Parity, which synchronizes all historical transaction records from the Ethereum blockchain. Since very few phishing addresses were reported before block height 2 million, we extract all the Ether transaction records between August 3, 2016, and October 6, 2020 (to be exact, from block height 2,000,000 to block height 11,000,000), including external transactions and internal transactions, as the input of network construction. And we have removed some irrelevant transactions such as the token transfer events, minting events, *etc.*

With data processing, we can abstract the transaction network as a Temporal Weight Multidigraph (TWMDG)[26]. Specifically, $G = (V, E)$, where V is a set of nodes and E is the set of edges. Each node represents a unique account address and each edge is represented as $e = (v_i, v_j, a_{ij}, t_{ij})$, where v_i is the source address, v_j is the target address, a_{ij} is the amount value of ETH which v_i transfers to v_j, and t_{ij} is the timestamp. It should be noted that there may be multiple edges between two nodes in TWMDG, depending on the number of transactions between the two related addresses. (In this case, account, address, and node are different expressions of the same thing, hence we use these names interchangeably in the following.)

3.2.2. Account labels

After constructing the transaction network, we should make extensive use of knowledge about the characteristics of accounts in the Ethereum ecosystem. The accounts on Ethereum can be divided into two types: Externally Owned Accounts (EOAs) and Contract Accounts (CAs). EOAs are similar to general bank accounts and controlled by people who hold the public-private key pairs, whereas CAs are smart contracts and ruled by executable code files. In addition, if an EOA was never the source address of a transaction, we define it as inactive, which may be a fresh address without any transfer transaction or used to burn crypto assets [30]. We can't track any Ether transaction followed by inactive EOAs.

Additionally, one of the most important tasks in detecting phishing gangs is to identify enough phishing accounts and other crux accounts related to money laundering. To do so, we utilize the Label Cloud provided by Etherscan, which serves as an Ethereum block explorer. By crawling the website, we extract all the necessary account labels that appear in the transaction network we construct and categorize them into five kinds:

www.parity.io/ethereum/
etherscan.io/accounts/labelcloud/

Phishing, Hack, Token, Exchange, and DEX (Decentralized Exchange). All kinds of labels with their numbers and description are listed in Table 1.

Table 1. Account Labels.

Label	Type	Count	Description
Phishing	EOA/CA	3,435	Verified phishing account, which phishers used for defraudation
Hack	EOA/CA	47	Fraud account related to hack, which may participate in phishing activity
Token	CA	537	Smart contract that allows customers to transfer tokens
Exchange	EOA/CA	278	Centralized exchange which is the platform that allows customers to buy and sell Ether for other assets
DEX	EOA/CA	2,378	Decentralized exchange that trades occur directly between users (peer to peer) through an automated process, including EtherDelta, Uniswap, Bancor, 0x and so on

4. Methods

In this section, we will introduce the methods for phishing gang detection. We first describe our taint analysis method for determining the taint scores of accounts in the transaction network. Based on the taint scores and transaction features of accounts, we then classify accounts into five categories, some of which are involved in phishing gangs and others are not. Finally, we propose a heuristic algorithm to detect phishing gangs.

4.1. Taint analysis

A reasonable and accurate tainting method is the key to phishing gang detection. Based on our TWMDG network with account labels, we propose a new taint analysis method for determining the taint scores of accounts that received Ether from phishing accounts on Ethereum. The higher the taint score of an account, the more likely the account is connected to a phishing scam. In our method, tainting starts from all the phishing nodes whose taint scores are fixed values set to 1. Contrarily, the other nodes get initial taint scores set to 0. And we calculate the taint score for each node in an iterative fashion using the equation as follows:

$$T(v_j) = \sum_{e_{ij} \in \Gamma_j} \frac{T(v_i) \times a_{ij}}{A_j} \tag{1}$$

where $T(v_j), T(v_i)$ are taint scores of node v_i and node v_j, Γ_j represents the set of all in-degree edges (symbolizing deposit transactions) of v_j, $e_{ij} = \{v_i, v_j, a_{ij}, t_{ij}\}$ is the edge from v_i to v_j with amount Ether a_{ij} and timestamp t_{ij} as defined in TWMDG, and A_j represents the sum amount Ether of all in-degree edges in Γ_j. Following the equation, each node received a taint score from every one of its predecessor nodes, based on the amount received via its in-degree edge and the taint score of its in-degree neighbor. There may be some contract nodes whose in-degree edges are all used for contract calls with 0 ETH so we set their taint scores to be zero instead of computing by the equation. It is worth mentioning that the taint score spread ends in the nodes labeled with Exchange, DEX, and Token because they are usually the end goals or exit points of taint Ether.

The basic principle behind this method is to estimate the ratio of taint funds aggregation of each node, which means the proportion of the money originally from phishing nodes in the total funds each node received. According to the tainting equation, the higher the ratio of taint funds aggregation of a node, the higher the taint score the node will get. For the reason that a common account would not receive most of the money from phishing accounts, if a node owns a high ratio of taint funds aggregation, it is likely to be controlled by the phisher and used for money laundering. With this approach, we can detect the phishing gang members served for laundering activities by their taint scores.

However, we find that a part of the accounts has different transaction activities in different periods, for example, some accounts served for money laundering had common deposit transactions several weeks before their laundering transactions, which leads to a dropping taint score as we take into consideration all in-degree edges in the whole period. To prevent this situation, when the network spans a long period, we should ameliorate our tainting method for better accuracy. We set a time range of t_{lim} for selecting in-degree edges and define $\Delta_j = \{e_1, e_2, \ldots, e_m\}$ as a maximal subset of Γ_j, where the timestamp sequence $\{t_1, t_2, \ldots, t_m\}$ of edges should satisfy $t_x \leq t_{x+1} (1 \leq x \leq m-1)$ and $t_m - t_1 \leq t_{lim}$. Then we change the tainting equation into:

$$T(v_j) = \max_{\Delta_j \subseteq \Gamma_j} \{ \sum_{e_{ij} \in \Delta_j} \frac{T(v_i) \times a_{ij}}{A'_j} \} \tag{2}$$

where A'_j represents the sum amount Ether of in-degree edges in Δ_j. By this equation, we extract every different maximal subset in the time range of t_{lim} from the set of all in-degree edges and find the maximum taint score computed in each subset by the same formula of Equation 1. In this improvement method, we can taint the node by focusing on the time frame when it is most involved in money laundering activities.

4.2. Account classification

Although the accounts of a phishing gang should belong to the same entity, they undertake different functions in the complete period of phishing activity. To propose the heuristic method for detecting phishing gangs, we should understand what roles the accounts play in phishing activities and identify them based on the taint scores and transaction behaviors of accounts. Therefore, utilizing taint analysis results and the information available in the transaction network with account labels, we classify the accounts into five categories: gangster account, exchange account, bridge Account, suspect account, and common account. For each category, we will not only interpret its role and function but also define its identity criteria as follows.

Gangster account A gangster account is a fraud account related to phishing scams. The phishers utilize this type of account to accumulate ill-gotten Ether from the fraud they commit. It is worth noting that a phisher may control multiple gangster accounts used in the same scam or different scams, and sometimes there are transfers of funds among the gangster accounts. Obviously, a phishing account is a gangster account. Moreover, we find some phishing activities are correlated to the accounts used for hack fraud. As a result, we classify all the accounts labeled with Phishing or Hack into gangster accounts. A gangster account is the main member of a phishing gang and a source of phishing funds.

Exchange account An exchange account is the end place where the phishers can exchange the obtained Ether with other assets to cash out their profit. There are two

types of exchange in Ethereum: one is centralized exchange performed as a 3rd platform, and the other one is decentralized exchange (DEX) which does not rely on a 3rd party. Additionally, we also take into account the token contract that may be used by phishers to buy and transfer tokens with Ether transactions. In this way, the token contract can serve as an exit point from the Ether transaction network to the token transaction network. To sum up, we classify all the accounts labeled with Exchange, DEX, or Token into exchange accounts. As mentioned in section 4.1, each exchange account is the end node of taint analysis and obtains a zero taint score. Therefore, we consider that the exchange account doesn't belong to the phishing gang, but it can help us to know about the destination to which the phishing funds have gone.

Bridge account A bridge account serves for money laundering and functions as a bridge between a gangster account and an exchange account or another bridge account. It usually has only a small number of deposit and transfer transaction neighbors. According to taint analysis, a node with a high taint score is a bridge account, and we should set the threshold of the taint score to identify it explicitly. Nevertheless, the node with not so high taint score but behaves like a bridge account should be considered. Because there may be some unknown phishing accounts without being labeled, and they are not used in taint analysis, which leads to low taint scores of their bridge accounts. For these reasons, we set two rules for bridge account classification as shown in Table 2. In R1, we only set a threshold of taint score as $\sigma_1(\sigma_1 \leq 1)$ to identify an account as a bridge account, for the hypothesis that no other account can reach this high taint score except a gangster account. In R2, we set another smaller threshold of taint score as $\sigma_2(\sigma_2 < \sigma_1)$, and limit the number of in-degree neighbors as well as the number of out-degree neighbors to no more than 3. If an account doesn't satisfy R1 but satisfies R2, it probably receives a small part of the money from verified phishing accounts but receives a big part of the money from unknown phishing accounts, so we also regard it as a bridge account. As it should be, bridge accounts belong to a phishing gang.

Table 2. Rules for bridge account classification.

Rule	Taint score	In-degree neighbors	Out-degree neighbors
R1	$\geq \sigma_1$	-	-
R2	$\geq \sigma_2$	≤ 3	≤ 3

Suspect account Since some phishing scams have not been reported, there may be some unknown phishing accounts that are non-labeled in the network as mentioned above. A suspect account is an account that can be a non-labeled phishing account. It behaves like a gangster account which usually accumulates a wealth of Ether from victims and then launders through a few transfer transaction neighbors. However, we can not make sure that it is a gangster account but can only identify it as a suspect object based on the information in the transaction network. In the following, we define seven criteria for suspect account classification and set the range of each criterion according to the statistics of verified phishing accounts.

- C1: Number of in-degree neighbors. A suspect account should have a big number of in-degree neighbors. In this paper, we set the in-degree neighbors' number range criterion to be between >10 and <230, for the reason that the threshold 230 is at the 99th percentile of the values from all verified phishing accounts.

See the proofs of self-defined criteria with the data and codes: https://github.com/Cohnry/Phishing-Gang-on-Ethereum/tree/master

- C2: Number of out-degree neighbors. A suspect account should have a few out-degree neighbors to transfer its profit. In this paper, we set the out-degree neighbor's number range criterion to be between >0 and <66 for the reason that the threshold 66 is at the 99th percentile of the values from all verified phishing accounts.
- C3: Ratio of the number of in-degree neighbors to the number of all neighbors. A suspect account should have many in-degree neighbors but a smaller number of out-degree neighbors. We use the ratio of the number of in-degree neighbors to the number of all neighbors to symbolize this feature. And in this paper, we set this criterion range to be more than 0.7, to which most of the verified phishing accounts conform.
- C4: Ratio of the number of in-degree edges to the number of in-degree neighbors. Since a verified phishing account usually behaves more like a "one-shot" deal account, which means few in-degree neighbors have multiple transactions with it, a suspect account should have a low ratio of the number of in-degree edges to the number of in-degree neighbors. In this paper, we set this criterion range to be less than 2.0, which is at the 95th percentile of the values from all verified phishing accounts.
- C5: Ratio of the number of out-degree edges to the number of out-degree neighbors. As a verified phishing account usually has a small number of transfer transactions with one of its out-degree neighbors, a suspect account should also have a low ratio of the number of out-degree edges to the number of out-degree neighbors. In this paper, we set this criterion range to be less than 4.0, which is at the 95th percentile of the values from all verified phishing accounts.
- C6: Sum amount Ether of all in-degree edges. We do not consider the accounts only with 0 ETH transactions and thus set the criterion range of the sum amount Ether of all in-degree edges to be more than zero.
- C7: Sum amount Ether of all out-degree edges. As above, we also set the criterion range of the sum amount Ether of all in-degree edges to be more than zero.

With these criteria, we select a set of suspect accounts and verified phishing accounts. The proportion of the verified phishing accounts in this set is on the order of 10^{-3}, greater than the proportion of all verified phishing accounts in the set of all accounts in a network, which is on the order of 10^{-5}. It is on the cards that a suspect account can be a member of a phishing gang, depending on its transaction correlation with other members, which we will define in section 4.3.

Common account The rest accounts are classified into common accounts. Similar to the bridge accounts, some common accounts may have only a few transaction neighbors. We called them low-degree common accounts. A low-degree common account is an account that belongs to an ordinary user on Ethereum. Correspondingly, there are common accounts with a big number of transaction neighbors, called high-degree common accounts. These accounts may provide cryptocurrency services with different purposes, such as DeFi, DApp, coin mixing, and non-custodial anonymous transactions which the phishers may use for money laundering. We set the high-degree common accounts classification to be the common accounts with at least 100 transaction neighbors in a network, including in-degree neighbors and out-degree neighbors. In this case, a high-degree common account is very likely to provide a certain service.

4.3. Detection heuristics

On the basis of account profiling and the transaction behaviors of phishing scams, we illustrate three heuristic rules for phishing gang detection: 1)as a phishing account would transfer its ill-gotten Ether to its partners, the gangster accounts and bridge

accounts that received Ether from it belong to the same phishing gang; 2)gangster accounts transferring their Ether to some members of the phishing gang should be included; 3)suspect accounts which have certain sum amount Ether of deposit or transfer transactions with other members should also be included in the same phishing gang. According to these, we propose a heuristic method for detecting the phishing gang by tracking a verified phishing account in a transaction network. Full detection heuristics are presented in Algorithm 1.

Algorithm 1 Detection of Phishing Gang

Input:
A verified phishing account P;
A TWMDG network $G(V, E)$;
Output:
A set of accounts $\mathcal{P}$, indicating the identified phishing gang;
1: label every account of V with whether it has been visited and initial it to be non-visited account;
2: set $\mathcal{A} = \{P\}$;$\mathcal{P} = \{\}$;$\mathcal{S} = \{\}$;
3: **while** $\mathcal{A}! = \emptyset$ **do**
4: select an account v in $\mathcal{A}$, and set it to be visited account;
5: find all gangster accounts $\mathcal{N}_g$, all bridge accounts $\mathcal{N}_b$, and all suspect accounts $\mathcal{N}_s$ in the non-visited neighbor accounts of v;
6: $\mathcal{P} = \mathcal{P} \cup \{v\}$;
7: $\mathcal{A} = \mathcal{A} \cup \mathcal{N}_g \cup \mathcal{N}_b$;
8: $\mathcal{A} = \mathcal{A} - \{v\}$;
9: $\mathcal{S} = \mathcal{S} \cup \mathcal{N}_s$;
10: **end while**
11: **for all** s such that $s \in \mathcal{S}$ **do**
12: calculate the sum amount Ether A_1 of all in-degree edges of s, and the sum amount Ether A'_1 of its in-degree edges from the accounts in $\mathcal{P}$;
13: calculate the sum amount Ether A_2 of all out-degree edges of s, and the sum amount Ether A'_2 of its out-degree edges to the accounts in $\mathcal{P}$;
14: **if** $A'_1/A_1 \geq \sigma_2$ **or** $A'_2/A_2 \geq \sigma_2$ **then**
15: $\mathcal{P} = \mathcal{P} \cup \{s\}$;
16: **end if**
17: **end for**
18: **return** $\mathcal{P}$;

The inputs of this algorithm include a verified phishing account and a TWMDG network. It returns a set of accounts that probably belong to a phishing gang. It first labels every account in the network with whether it has been visited or not. Initially, every account is a non-visited account (Line 1). The algorithm creates an interim set $\mathcal{A}$ with one element, the verified phishing account, and creates two empty sets $\mathcal{P}$ and $\mathcal{S}$ (Line 2). The set $\mathcal{P}$ is used for collecting accounts of the phishing gang and will be returned after the algorithm. And the set $\mathcal{S}$ is used for collecting suspect accounts, some of which may be added into $\mathcal{P}$ later. While the interim set $\mathcal{A}$ is not empty (Line 3), the algorithm first selects an account in the interim set as a visited account (Line 4), and find all the gangster accounts, bridge accounts and suspect accounts in its non-visited neighbors (Line 5). Then we update the set $\mathcal{P}$ by adding the visited account (Line 6). The interim set adds all the selected gangster accounts and bridge accounts (Line 7), and deletes the visited accounts (Line 8). The set $\mathcal{S}$ adds all the selected suspect accounts (Line 9). Finally, we determine which suspect accounts of $\mathcal{S}$ are added into $\mathcal{P}$. For each suspect accounts in $\mathcal{S}$ (Line 11), we not only calculate the sum amount Ether of all its in-degree edges (i.e., A_1) and the sum amount Ether of its in-degree edges from the accounts in $\mathcal{P}$ (i.e., A'_1), but also calculate the sum amount Ether of all its in-degree edges (i.e., A_2) and the sum amount Ether of its out-degree edges to the accounts in $\mathcal{P}$ (i.e., A'_2) (Line 12, 13). Then we define that if A'_1/A_1 or A'_2/A_2 bigger

than σ_2, the suspect account is considered as a member of the gang (Line 14, 15). It's easy to see the calculation of both ratios is similar to that of the taint score when we set the score of gangster and bridge members to 1. By reference to the second rule for bridge account classification, the threshold is set to be σ_2. As a result, the returned set (Line 18) contains at most three kinds of accounts: gangster, bridge, and suspect accounts.

By this algorithm, we detect a phishing gang with at least one verified phishing account. If there are multiple verified phishing accounts in a phishing gang, the algorithm can return the same result with each of these accounts' input, that is, each verified phishing account only corresponds to a unique gang. However, there may be an intersection between different phishing gangs, because some suspect accounts may appear in more than one phishing gang. These accounts should be paid more attention to since they link to different gangs and are likely to be fraud accounts.

Algorithm 1 rely on taint analysis and account classification. According to the improvement method of taint analysis, the time complexity required to calculate the taint score of each node v_j is $O(|\Gamma_j| \log(|\Gamma_j|))$ due to the need to sort the edges by timestamp. However, when performing the detection algorithm, we only focus on nodes with a taint score $\leq \sigma_2$, which means that the time complexity of the taint analysis is determined by the parameter σ_2 and will not increase with the size of the transaction network. Similarly, it is not necessary to perform account classification on all nodes in the transaction network. It is sufficient to classify each node connected to the source node P as Algorithm 1 traverses it. Therefore, the computation complexity of Algorithm 1 is only related to the size of the connected sub-network of the verified phishing account.

5. Results

In this section, we present the experimental results of our proposed phishing gang detection framework. First, we describe the experimental implementation in detail. Next, we show the statistics of detection results and analyze some patterns of phishing gangs. Finally, we reveal how the phishing gangs we detected launder money in the Ethereum transaction network.

5.1. Experiment settings

5.1.1. Data selection

As discussed in section 3.2, we finally obtain the Ether transaction network with more than 87 million nodes and 519 million edges. However, there are only 3,435 phishing nodes in the network. Given that the network spans more than 3 years, we should focus on the periods of phishing activities. By analyzing the timestamps of all edges connected to phishing nodes, we find that there are almost 80% of phishing accounts whose transaction durations are less than 3 months or one quarter. In fact, owing to the fast-growing numbers of nodes and edges in the transaction network with the development of Ethereum and the passage of time, it is impractical and unnecessary to set the whole large-scale history network as input. Accordingly, to limit the computational resources and the time required for the following analyses and detection, selecting the three-month period sub-networks with relatively high phishing activities from the whole network as salient samples is the better choice. For this purpose, we consult two criteria for periods selection: (1)the number of active phishing accounts (the phishing nodes with at least one edge in the period); (2)the amount of phishing transfer ETH which represents the phishing profit which will flow to the network and be cashed through exchanges.

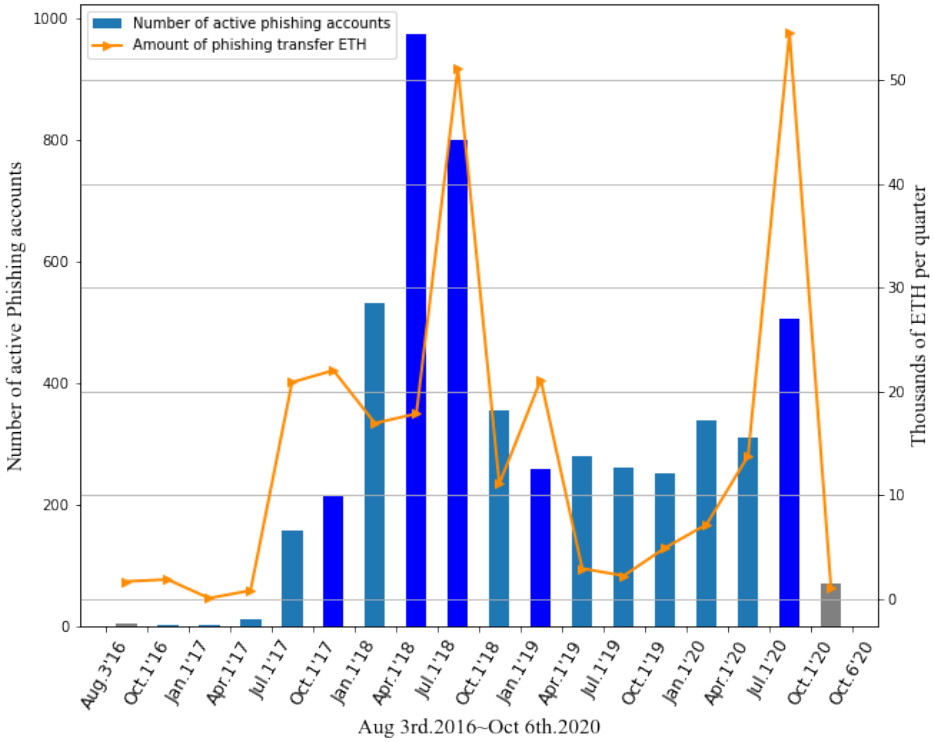


Figure 3. Number of active phishing accounts and amount of phishing transfer ETH in every three months from August 3rd, 2016 to October 6th, 2020. (The first and the last bar which are colored gray do not span three months enough so they will not be considered. The salient blue bars are the periods we consider to represent relatively high phishing activities.)

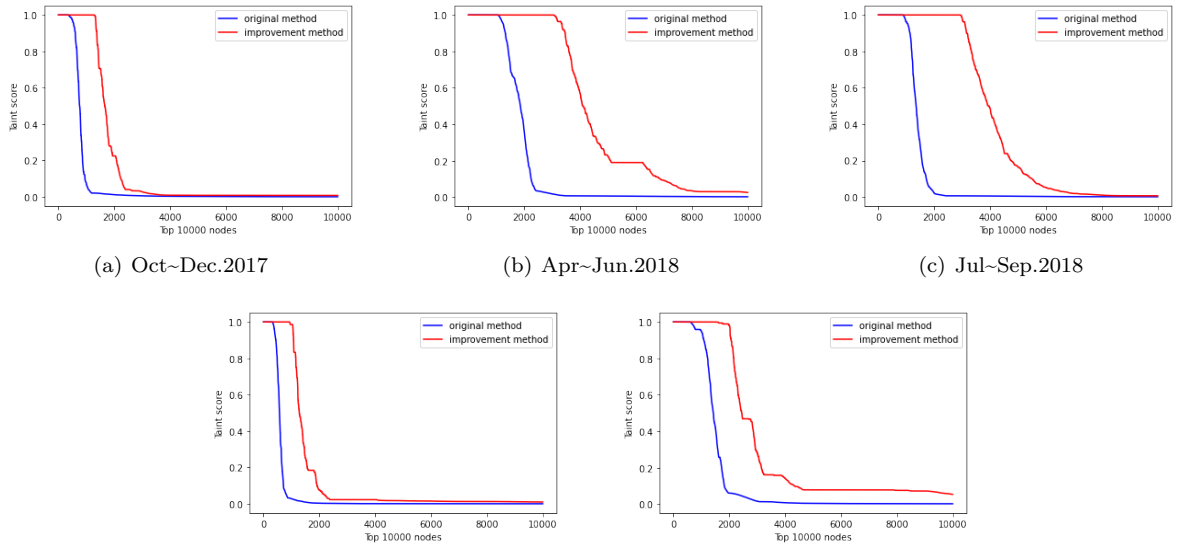
Figure 3 shows the statistics of both criteria every three months of the whole network period. It is easy to find four peaks of the amount of phishing transfer ETH and the highest value of the number of active phishing accounts, which we consider to represent relatively high phishing activities. Their corresponding periods whose bars are colored salient blue are as follows: Oct~Dec.2017, Apr~Jun.2018, Jul~Sep.2018, Jan~Mar.2019 and Jul~Sep.2020. And these five periods are the time frames in which we detect phishing gangs. Table 3 shows the information of five selected sub-networks.

Table 3. Information of five sub-networks

Time frame	Nodes	Phishing nodes	Edges
Oct~Dec.2017	9,595,165	214	39,319,247
Apr~Jun.2018	10,225,586	974	40,522,241
Jul~Sep.2018	7,532,607	800	35,780,650
Jan~Mar.2019	5,886,458	259	27,554,330
Jul~Sep.2020	12,370,866	505	64,827,148

5.1.2. Parameter estimation

Our proposed methodology for detecting phishing gangs relies on three parameters: the time range for tainting method: t_{lim} and two thresholds of taint score: σ_1 and σ_2. Based on statistics from the transaction network, few continuous transaction activities last more than half of the month, so we set $t_{lim} = 1,296,000$ seconds, computed by 15 days. With this setting, we test the original method and the improved method of taint analysis in five three-month period sub-networks. And we display the top 10,000 highest taint scores for both methods of each sub-network in Figure 4.



(a) Oct~Dec.2017 (b) Apr~Jun.2018 (c) Jul~Sep.2018

(d) Jan~Mar.2019 (e) Jul~Sep.2020

Figure 4. Top 10,000 highest taint scores for both methods tested respectively in five sub-networks. (The blue line is the result of original method and the red line is the result of improvement method.)

In each sub-network, the improvement method gets higher taint scores and has better differentiation in roughly the top 4,000 tainted nodes. Consequently, we'd better refer to the result of the improvement method and consider the top tainted nodes right until the lines start to have a linear pace. Based on this, we set the small threshold of taint score σ_2 to be 0.1, the value close to lines with linear pace in all the subgraphs. And we set the relatively large threshold of taint score σ_1 to be 0.5, for the reason that non-phishing gang members are hard to reach this high score which represents at least half of the funds it received are originally from phishing nodes in a period of 15 days. We also combined the tainted nodes of all subgraphs and obtained a cumulative distribution of taint score and number of nodes, as shown in Figure 5. The distribution curve consists of three curves with significantly different slopes. And the parameters $\sigma_1 = 0.5$ and $\sigma_2 = 0.1$ are near the two divisions (the blue dashed lines), which validates the reasonableness of our parameter selection.

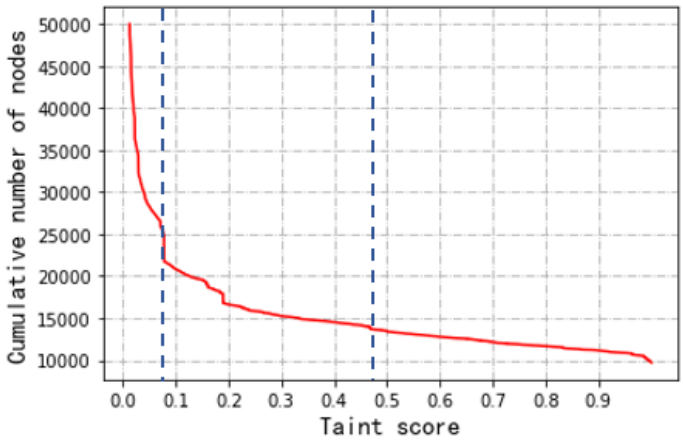


Figure 5. Cumulative distribution of taint score and number of nodes.(The distribution curve can be divided into three sections by the blue dashed line according to the difference of slope.)

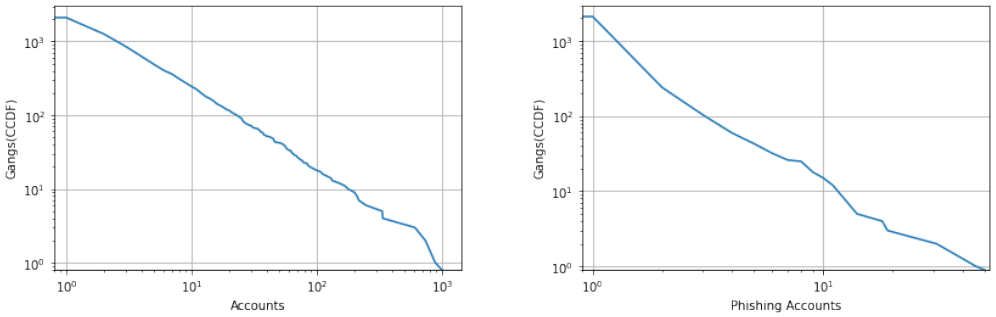
5.2. Gangs statistics and patterns

5.2.1. Statistics

Table 4. Identified phishing gangs of five sub-networks.

Time frame	Phishing gangs	Single -node	Multiple -node	Total nodes in gangs	Phishing nodes in multiple-node gangs
Oct~Dec.2017	176	51	125 (71.02%)	1,918	163 (76.17%)
Apr~Jun.2018	724	310	414 (57.18%)	4,510	664 (68.17%)
Jul~Sep.2018	598	256	342 (57.19%)	4,458	544 (68.00%)
Jan~Mar.2019	217	83	134 (61.75%)	1,385	173 (66.80%)
Jul~Sep.2020	389	146	243 (62.47%)	2,628	359 (71.09%)
Total	2,104	846	1,258 (59.79%)	14,899	1,903 (69.23%))

Based on the settings, we can implement the heuristic method to detect phishing gangs in five transaction sub-networks on Ethereum. As mentioned in section 4.3, a verified phishing account corresponds to a phishing gang, but the detection algorithm may return the same gang with different phishing accounts input. Thus, we check off duplicate gangs and the experimental results are given in Table 4. In each sub-network, hundreds of phishing gangs have been identified, and thousands of nodes are involved. Nevertheless, some of the gangs are single-node gangs each consisting of only an initial phishing node. In essence, they do not have the characteristics of a gang. There are three possible reasons for the emergence of single-node gangs. Firstly, the phishing account has not yet carried out money laundering activities. Secondly, the phishing account has cashed out directly from the exchange. Thirdly, the phishing account has used a communal money laundering service that allows funds to flow to a common account with an extremely low taint score. In summary, the phishing entity has not used other accounts to complement the phishing fraud actions.



(a) CCDF showing the number of phishing gangs each maps to a minimum number of accounts. For instance, there are about 100 gangs that consist of at least 20 accounts.

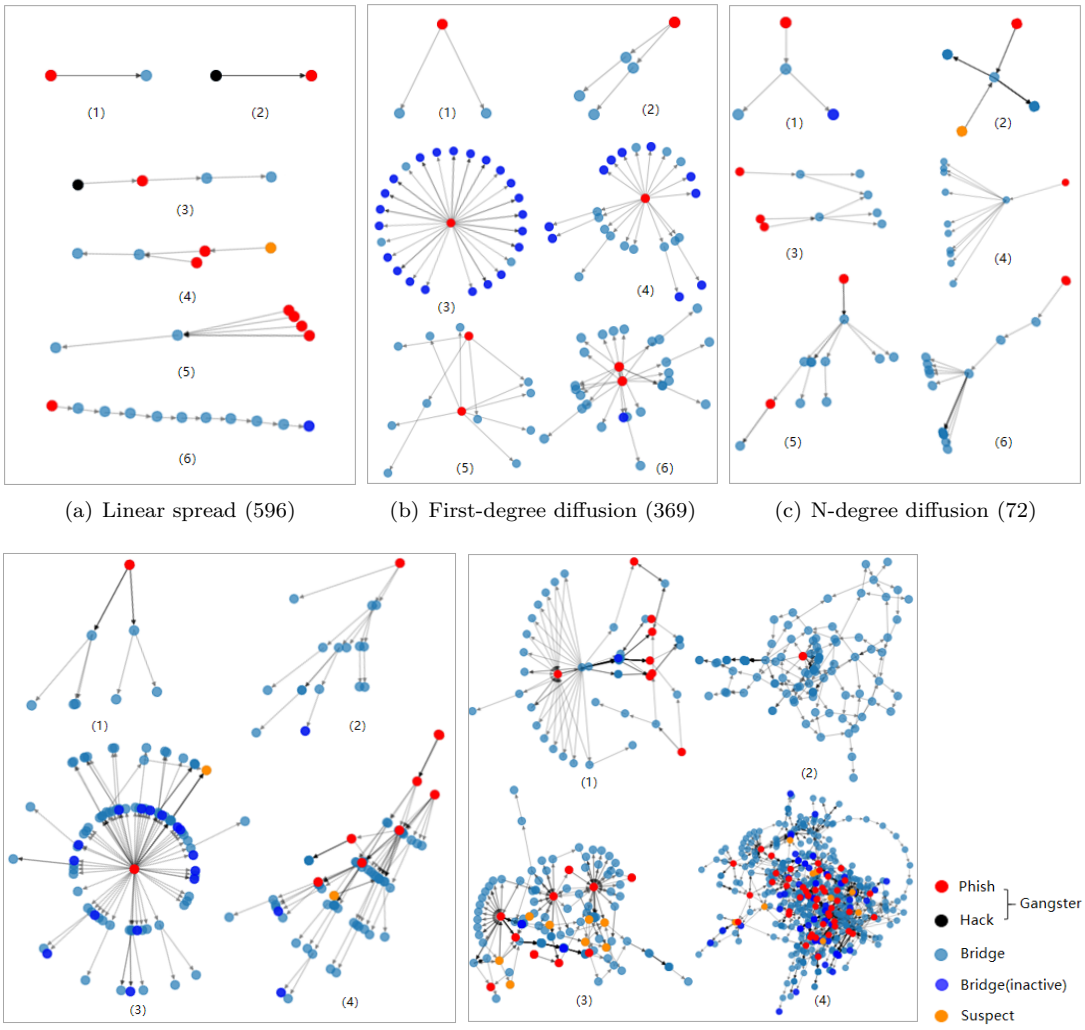
(b) CCDF showing how many phishing gangs each contains a minimum number of phishing accounts. For instance, there are about 100 gangs that contains 3 or more verified phishing accounts.

Figure 6. Complementary cumulative distribution of accounts in the phishing gangs.

Oppositely, other gangs each containing more than one node are named multiple-node gangs. Table 4 also demonstrates the number and proportion of single-node gangs and multiple-nodes gangs. It is easy to see that the multiple-nodes gangs constitute the majority in each sub-network. In total, we associate 14,899 nodes with 2,104 phishing gangs. Out of these, 846 (40.21%) are single-node gangs, and 1,258 (59.79%) are multiple-node gangs. Moreover, we can observe that there are 1,903 phishing nodes involved in multiple-node gangs, and their proportion of all verified phishing nodes is

69.23%, which means most of the phishing accounts have their partners.

As each phishing gang has its quantity scale of accounts, we can explore the full distribution with a complementary cumulative distribution function (CCDF), which is illustrated in Figure 6(a). It can be seen that the larger the number of accounts, the smaller the number of gangs, and not a phishing gang consists of more than 1,000 accounts. Meanwhile, each phishing gang owns one or more verified phishing accounts. Figure 6(b) illustrates the distribution of how many phishing accounts cooperate in a gang (again using CCDF). There, We also can see that only hundreds of gangs contain multiple phishing accounts, and not a phishing gang consists of more than 50 phishing accounts. It should be noted that the larger number of accounts doesn't mean that there are more phishing accounts in a gang. Out of 2,104 gangs, the biggest one with 885 accounts only contains an initial phishing account. And the gang with the most (that is 46) phishing accounts, totally consists of 738 accounts.



(a) Linear spread (596) (b) First-degree diffusion (369) (c) N-degree diffusion (72)

(d) Multiple diffusion (188) (e) Complex pattern (33)

Figure 7. Patterns of phishing gangs.

5.2.2. Patterns

The phishing gangs detected from the five transaction sub-networks have a variety of structures and patterns of internal transactions, many of which are similar. In order to better characterize the phishing gangs, we draw pattern legends of the gangs based

on the transactions and account types within the multiple-node gangs and classify these gangs into five patterns, which are linear spread, first-degree diffusion, n-degree diffusion, multiple diffusion, and complex pattern.

Figure 7 shows the typical pattern legends of the five gang patterns respectively. Each legend uses different colored nodes to mark the different account types. Red nodes indicate verified phishing accounts, black nodes indicate hack accounts, and they are both gangster accounts. Blue nodes indicate bridging accounts, and salient blue nodes indicate inactive bridging accounts, which are accounts without any transfer transactions as described in section 3.2.2. And orange nodes indicate suspect accounts, which are potential phishing accounts. As there will be multiple transactions between accounts, there will also be multiple edges between nodes in the legend, but they will overlap and appear as darker edges. In addition, the longer the edge, the larger the transaction amount.

The characteristics of the five gang patterns are described as follows. As shown in Figure 7(a), a linear spread pattern means that the gangster accounts in the gang will transfer funds from one account to another in a linear path of transferring funds with no bifurcation. 596 gangs belong to the linear spread pattern. The first-degree diffusion pattern is shaped as Figure 7(b), which means that the gangster accounts only disperse funds to multiple accounts in the gang at first, and these accounts will not disperse their funds at a later date. There are 369 gangs that belong to the first-degree diffusion pattern. In contrast to first-degree diffusion, the n-degree diffusion pattern is shaped as Figure 7(c), which means the gangster accounts transfer funds to only one account in a linear path at first, but at some point during the subsequent transfer of funds from that account, will disperse funds to multiple accounts (perhaps in the gangster account's second-degree neighbors, or third-degree neighbors, or fourth-degree neighbors, *etc.*), and then not disperse again. There are 72 gangs that belong to the n-degree diffusion pattern.

And multiple diffusion pattern is a combination of first-degree diffusion and n-degree diffusion, which is shaped as Figure 7(d). That is, the gangster accounts disperse funds to multiple accounts in the gang at multiple stages, and 188 gangs belong to this pattern. Lastly, the complex pattern is shown in Figure 7(e), which is an extremely complex gang pattern due to the excessive number of nodes in the gang and the existence of multiple circulating capital flows. There are 33 gangs that belong to the complex pattern. A few of the complex patterns have similar shapes, but there are few rules to follow. They may involve multiple phishing fraud activities, or use some large-scale money laundering services. All types of gang patterns essentially reflect how phishing entities use the accounts in the gang to disperse phishing funds, that is, the form of gang money laundering.

5.3. Reveal money laundering

Money laundering is an important reason for the emergence of phishing gangs. Detecting phishing gangs will help to further understand the means and whereabouts of phishing money laundering. The gang patterns reveal the decentralized transfer of funds within the phishing gangs, that is, the phishing entity uses each account under its control to launder money. However, in addition to transferring funds to internal accounts, accounts in a gang also transfer funds to accounts outside the gang, which cannot be shown in the gang pattern. To better understand the gang's money laundering activities, we further analyze the external accounts receiving funds from each account in the gang, thereby searching the gang's money laundering destination.

As described in section 4.2, there are two types of accounts outside the group, exchange accounts, and common accounts. And common accounts are divided into

low-degree accounts and high-degree accounts. If the gang accounts transfer funds to exchange accounts, it is to complete the final step of the money laundering activity - cashing out so that the real proceeds can be obtained. Otherwise, if the gang accounts transfer funds to common accounts, it needs to be discussed according to the situation. The experiment found that the number of funds transferred from gang accounts to low-degree common accounts tends to be small. We believe that it is a means used by phishing entities to enhance anonymity, which is to prevent the regulatory system from fraud identification of phishing entities by trading a small number of funds to some common accounts. The transfer of funds from gang accounts to high-degree common accounts, on the other hand, is often using some kind of service on Ethereum, which may be a cash arbitrage tool or a money laundering tool. In any case, it is worth paying attention to the high-degree accounts receiving funds from gang accounts.

To this end, we count the number of associated gangs (*i.e.* how many gangs' funds have been received) and the number of total receiving funds from gang accounts in all high-degree common accounts and rank them from large to small. Table 5 shows the top five high-degree common account addresses that have the largest number of associated gangs and the largest number of receiving funds from the gangs. It can be seen that some high-degree accounts have received funds from many phishing gangs. They act like public money laundering accounts, which transfer some funds anonymously to many other accounts. But it is worth noting that the high-degree account with the largest number of associated gangs receives very little funds from gang accounts, which may be used to collect service fees. In addition, some of the high-degree accounts have received a large number of funds from gang accounts, which may be used for cash out. From the Etherscan website, it can be seen that the accounts with the largest number of received funds belong to SushiSwap. SushiSwap is a new type of automated trading platform, which can actually be classified as a decentralized exchange (DEX). For the reason that it has not been fully started, its accounts have not been labeled as exchange accounts. Two phishing gangs have made use of this new cash-out tool to obtain large profits.

Table 5. Top five high-degree common account addresses that have the largest number of associated gangs and the largest number of receiving funds from the gangs.

Account address	**Associated gangs**	**Receiving funds(ETH)**
Top five account addresses have the largest number of associated gangs		
0064454b14cddc990ed6520ef94d3e28fe7c41b6	63	3.12
b89dd647788c54a03ba290a6320ba53566afb57d	39	209.05
af1931c20ee0c11bea17a41bfbbad299b2763bc0	32	766.01
667f17d272c205009873dd7973a2ffd185f926bb	29	25.98
b7cabd1e6598209ec21a5a7ad72cc76b4ce4fc85	24	51.18
Top five account addresses have the largest number of receiving funds		
d9e1ce17f2641f24ae83637ab66a2cca9c378b9f	2	5,523,225.40
7ed1e469fcb3ee19c0366d829e291451be638e59	18	257,871.55
9bb3fdb9cd7b6d63abfb493a362845ebac5f94c7	2	40,275.71
f65ce7833dacabd7e5c264421fe1f0521e98c0f0	4	16,020.51
7a343cce0fbe8d33e25cd8765950cb5c67c93f36	16	12,153.39

The exchange accounts also receive a large amount of money from the phishing gangs, because they are the final destination of money laundering. We also count the number of associated gangs and the number of total receiving funds from gang accounts in all exchange accounts and rank them from large to small. Table 6 shows the top five

exchange account addresses that have the largest number of associated gangs and the largest number of receiving funds from the gangs and their corresponding labels. It can be observed that ShapeShift and Binance are places where many phishing gangs choose to cash out. The number of their associated gangs is much larger than that of HitBTC, which ranks third. At the same time, the sum of funds they receive from the gangs is also in the top five. And 0x and Uniswap, as popular decentralized exchanges in Ethereum, have received huge amounts of money from the gangs, more than 200 times that of ShapeShift, which ranks third in receiving funds.

SushiSwap, as mentioned above, also receives far more funds from the gangs than ShapeShift. It suggests that although traditional exchanges (*i.e.* ShapeShift and Binance) tend to be the destination for most phishing gangs to launder money, the new decentralized exchange is the place where the phishing gangs choose to cash in huge amounts of money. This is probably because the traditional exchanges have been established for a long time, and the supervision will become more strict. The phishing gangs would not risk transferring huge amounts of money to their accounts. However, the new decentralized exchanges on Ethereum are often in a period of growing wild with little supervision. At the same time, they also have the characteristics of convenient and fast trading, through which the phishing gangs can quickly cash out a large number of funds.

Table 6. Top five exchange account addresses that have the largest number of associated gangs and the largest number of receiving funds from the gangs.

Account address	**Associated gangs**	**Receiving funds(ETH)**	**Label**
Top five account addresses have the largest number of associated gangs			
70faa28a6b8d6829a4b1e649d26ec9a2a39ba413	276	541,467.97	ShapeShift
3f5ce5fbfe3e9af3971dd833d26ba9b5c936f0be	151	250,079.15	Binance
a12431d0b9db640034b0cdfceef9cce161e62be4	50	1,740.30	HitBTC
7a250d5630b4cf539739df2c5dacb4c659f2488d	47	102,629,847.21	Uniswap
2819c144d5946404c0516b6f817a960db37d4929	46	15,830.40	Remitano
Top five account addresses have the largest number of receiving funds			
c02aaa39b223fe8d0a0e5c4f27ead9083c756cc2	8	128,363,545.87	0x
7a250d5630b4cf539739df2c5dacb4c659f2488d	47	102,629,847.21	Uniswap
70faa28a6b8d6829a4b1e649d26ec9a2a39ba413	276	541,467.97	ShapeShift
3f5ce5fbfe3e9af3971dd833d26ba9b5c936f0be	151	250,079.15	Binance
6cc5f688a315f3dc28a7781717a9a798a59fda7b	2	157,690.63	OKEx

6. Conclusion

In this article, we conducted the first systematic study of phishing gangs detection on the Ethereum blockchain via taint analysis. Specifically, a three-step framework was proposed to identify phishing gangs in the transaction network which we constructed with the historical transaction records on the blockchain. Firstly, we conducted a novel taint analysis for determining the taint scores of accounts in the transaction network. Secondly, based on the taint scores and transaction features of accounts, the accounts in the network were divided into five categories: gangster account, exchange account, bridge Account, suspect account, and common account. Thirdly, phishing gang detection heuristics were implemented. Experiments on real-world Ethereum transaction records

demonstrated most phishing accounts have partner accounts aiming to complete phishing scams and money laundering together. By statistics of the distribution of accounts in all identified phishing gangs, it was found that the number of accounts varied greatly among gangs, and the larger the number of accounts, the smaller the number of gangs. Finally, we analyzed and summarized the patterns of money transfers within phishing gangs, which were divided into five types: linear spread, first-degree diffusion, n-degree diffusion, multiple diffusion, and complex pattern. Furthermore, ShapeShift and Binance exchanges are the destinations of money laundering for most gangs, and they obtained the two largest numbers of associated gangs. Nevertheless, some gangs would choose to transfer large amounts of funds to new decentralized exchanges(i.e., 0x, Uniswap, and SushiSwap), which made their total receiving funds from the gangs far greater than other accounts.

By detecting phishing gangs, we can blacklist the accounts in the gangs and make them public in the blockchain community. At the same time, penalties for phishers could be established by negotiating with the judiciary and exchanges to build a firewall that can prevent the accounts in the blacklist from cashing out and give more restrictions in the blockchain ecosystem. Due to the anonymity of the blockchain, we cannot track the real identities of the phishers, which means that other legal tools are needed to discipline them in the real world. However, if the phishers are identified in the real world, the blacklist based on the phishing gangs can be used as evidence in the courts of law.

Detecting phishing scams and anti-money laundering in the blockchain ecosystem are real and critical problems. As a preliminary study in this field, we hope this work might serve as a modest spur to induce extensive efforts in this field. Many issues need further research, such as theoretical analysis of modeling the different phases of a phishing scam, a comprehensive system of indicators to distinguish phishing gangs from other communities, a more systematic and rigorous phishing gang detection algorithm, getting more accurate ground truth data and build more effective classification models. In addition, phishing gangs with complex patterns involve many accounts and transactions, and the operation mode of these gangs is difficult to show intuitively. Therefore, the complex pattern should be properly decomposed for analysis and research, and its money laundering strategy needs further exploration. In the future, we are going to study these issues in greater depth and build a website for phishing gangs warning and illegal financial chain display by using the proposed framework.

Acknowledgments

The work described in this paper is supported by the National Natural Science Foundation of China (62032025), Guangdong Basic and Applied Basic Research Foundation (2021A1515011939), and the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study, Grant No. SN-ZJ-SIAS-001.

References

[1] Ethereum. Welcome to Ethereum. Available: https://www.ethereum.org (accessed on 30 April 2018)

[2] Camilla H. Almost 10% of all money invested in initial coin offerings using cryptocurrency Ethereum has been stolen. Available: https://www.businessinsider.sg/ethereum-cyber-criminals-icos-threft-2017-2017-8 (accessed on 30 April 2018).

[3] Liu J, Ye Y. Introduction to e-commerce agents: Marketplace marketplace solutions, security issues, and supply and demand. *E-Commerce Agents,* Berlin/Heidelberg, Springer, 2001. 1-6.

[4] Chainalysis. The Rise of Cybercrime on Ethereum. Available: https://blog.chainalysis.com/the-rise-of-cybercrime-on-ethereum (accessed on 30 April 2018).
[5] Paul D. 100s of ETH Stolen After Bee Token ICO Email List Hacked. Available: https://theripplecryptocurrency.com/bee-token-scam (accessed on 30 April 2018)
[6] Khonji M, Iraqi Y, Jones A. Phishing detection: a literature survey. *IEEE Commun Surv Tutor* 2013, 15(4): 2091-2121.
[7] James L. *Phishing Exposed,* Amerstand: Elsevier, 2005.
[8] Abdelhamid N, Ayesh A, Thabtah F. Phishing detection based associative classification data mining. *Expert Syst Appl* 2014, 41(13): 5948-5959.
[9] Zouina M, Outtaj B. A novel lightweight URL phishing detection system using SVM and similarity index. *Hum-Cent Comput Info* 2017, 7(1): 1-13.
[10] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. *DBR*, 2008: 21260.
[11] Wood G. Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper* 2014, 151(2014): 1-32.
[12] Nick S. Smart contracts: Building blocks for digital markets. Available: http://www.fon.hum.uva.nl (accessed on 1 October 2017)
[13] Ethereum. Create your own crypto-currency with Ethereum. Available: https://www.ethereum.org/token (accessed on 30 April 2018)
[14] Möser M, Böhme R, Breuker D. An inquiry into money laundering tools in the Bitcoin ecosystem. In *2013 APWG eCrime researchers summit,* New York: IEEE, 2013.
[15] Prado-Romero MA, Doerr C, Gago-Alonso A. Discovering bitcoin mixing using anomaly detection. *Iberoamerican Congress on Pattern Recognition,* Cham: Springer, 2017.
[16] Hu Y, Seneviratne S, Thilakarathna K, Kensuke F, Aruna S. Characterizing and detecting money laundering activities on the bitcoin network. *arXiv* 2019, arXiv:1912.12060.
[17] Phetsouvanh S, Oggier F, Datta A. Egret: Extortion graph exploration techniques in the bitcoin network. *2018 IEEE International conference on data mining workshops (ICDMW),* New York: IEEE, 2018.
[18] Haslhofer B, Karl R, Filtz E. O Bitcoin Where Art Thou? Insight into Large-Scale Transaction Graphs. *SEMANTiCS (Posters, Demos, SuCCESS)* 2016.
[19] Hercog U, Povše A. Taint analysis of the Bitcoin network. *arXiv* 2019, arXiv:1907.01538.
[20] Tironsakkul T, Maarek M, Eross A, Just M. Probing the Mystery of Cryptocurrency Theft: An Investigation into Methods for Taint Analysis. *arXiv* 2019, arXiv:1906.05754.
[21] Tesa. Lost Ethers in Google ad phishing scam. Available: https://ethereum.stackexchange.com/questions/8565/lost-ethers-in-google-ad-phishing-scam (accessed on 30 April 2018)
[22] Mary-Ann R. Ethereum under siege: Scammers make $700,000 in 6 days from Slack and Reddit phishing attacks. Available: https://www.ibtimes.co.uk/ether eum-under-siege-scammers-make-700000-6-days-slack-reddit-phishing-attacks-162 9866 (accessed on 30 April 2018)
[23] Medvet E, Kirda E, Kruegel C. Visual-similarity-based phishing detection. *Proceedings of the 4th international conference on Security and privacy in communication netowrks.* 2008. pp.1-6.
[24] Alkhozae MG, Batarfi OA. Phishing websites detection based on phishing characteristics in the webpage source code. *Int J Inf Commun Technol* 2011, 1(6).
[25] McGeehan R, Popov LT, Palow CW, Read RJ, Keyani P. Preventing phishing

attacks based on reputation of user locations. U.S. Patent No. 9,576,119. 21 Feb. 2017.

[26] Jain AK, Gupta BB. PHISH-SAFE: URL features-based phishing detection system using machine learning. *Cyber Secur* 2018. 467-474.

[27] Jian M, Wenqian T, Pei L, Tao W, Zhenkai L. Phishing website detection based on effective CSS features of Web pages. *International Conference on Wireless Algorithms, Systems, and Applications,* Cham: Springer, 2017. pp.804-815.

[28] Antonovici. Japan records 170 money-laundering cases related to cryptocurrency over six months. Available: https://cryptovest.com/news/japan-records-170-money-laundering-cases-related-to-cryptocurrency-over-six-months/ (accessed on 30 April 2018)

[29] Lin D, Wu J, Yuan Q, Zheng Z. T-edge: Temporal weighted multidigraph embedding for ethereum transaction network analysis. *Front Phys* 2020, 8, 204.

[30] Victor F. Address clustering heuristics for Ethereum. *International conference on financial cryptography and data security,* Cham: Springer, 2020. pp.617-633.

[31] Lee XT, Khan A, Sen Gupta S, Ong YH, Xuan L. Measurements, analyses, and insights on the entire ethereum blockchain network. *Proceedings of The Web Conference,* 2020. pp.155-166.

[32] Chen W, Guo X, Chen Z, Zheng Z,Lu Y. Phishing Scam Detection on Ethereum: Towards Financial Security for Blockchain Ecosystem. *IJCAI* 2020.pp.4506-4512.