XBlock-ETH: Extracting and Exploring Blockchain Data From Ethereum

Preprint · October 2019

CITATIONS
0

READS
203

3 authors, including:

Zibin Zheng
Sun Yat-Sen University
254 PUBLICATIONS **7,226** CITATIONS
SEE PROFILE

Hong-Ning Dai
Macau University of Science and Technology
126 PUBLICATIONS **2,251** CITATIONS
SEE PROFILE

XBlock-ETH: Extracting and Exploring Blockchain Data From Etherem

Peilin Zheng
School of Data and Computer Science
Sun Yat-sen University
Guangzhou, China
zhengpl3@mail2.sysu.edu.cn

Zibin Zheng*
School of Data and Computer Science
Sun Yat-sen University
Guangzhou, China
zhzibin@mail.sysu.edu.cn

Hong-Ning Dai
Faculty of Information Technology
Macau University of Science and Technology
Macau, SAR
hndai@ieee.org

***Abstract*—Blockchain-based cryptocurrencies have received extensive attention recently. Massive data has been stored on permission-less blockchains. The analysis on massive blockchain data can bring huge business values. However, the lack of well-processed up-to-date blockchain datasets impedes big data analytics of blockchain data. To fill this gap, we collect and process the up-to-date on-chain data from Ethereum, which is one of the most popular permission-less blockchains. We name these well-processed Ethereum datasets as XBlock-ETH, which consists of the data of blockchain transactions, smart contracts, and cryptocurrencies (i.e., tokens). The basic statistics and exploration of these datasets are presented. We also outline the possible research opportunities. The datasets with the raw data and codes have been publicly released online.**

I. Introduction

Blockchain has attracted extensive attention from both academia and industry in the recent years. Among the diverse blockchain systems, substantial efforts have been made on the permission-less blockchain (or public blockchain) due to its decentralization [1]. The idea of permission-less blockchain was firstly proposed and implemented on Bitcoin [2]. In a blockchain system, each peer holds a ledger being considered as a public tally that is essentially temper-resistant. Ethereum [3] is another most popular permission-less blockchain system that enables Turing-complete smart contracts.

The proliferation of blockchain systems has lead to the generation of massive amount of blockchain data. Take Bitcoin as an example. There are nearly 242 GB Bitcoin data by the third quarter of 2019 as reported by Statista (https://www.statista.com/). In this paper, we focus on the data of Ethereum rather than Bitcoin, since Ethereum provides richer data. For another example, more than 16,000,000 smart contracts are deployed on Ethereum. As the Ethereum community has published two token protocol to enable easier Initial Coin Offerings (so-called ICO) for users [4], over 100,000 kinds of ERC20 token and 1,600 kinds of ERC721 token are available to be transferred on Ethereum where ERC stands for Ethereum Request for Comments.

The massive blockchain data provides researchers with both huge business values and great opportunities [5] due to openness, decentralization and temper-resistance of blockchain systems. Take business trading data as an example. In the past, it is difficult for researchers to obtain the real business trading data because of the privacy or ownership concerns of data owners. However, all the data in incumbent blockchain systems are all publicly available. Meanwhile, the blockchain data in permission-less blockchains can be accessed almost everywhere due to the decentralization of blockchain systems. Moreover, distributed consensus of blockchains also guarantees the temper-resistance of blockchain data. In addition to blockchain transactions, Ethereum (or its alternatives) also consists of both smart contracts and cryptocurrencies. Big data analytics of blockchain data can advance the developments in fraud detection of transactions, vulnerability detection of smart contracts and software development of smart contracts, etc.

However, there are a number of challenges in big data analytics of blockchain data, especially in Ethereum: **(1) Difficulty in data synchronization at Blockchain peer.** Due to the bulky size of blockchain, it takes a long period to fully synchronize entire blockchain data at a node (i.e., a peer) newly connected with the blockchain. For example, it takes more than one week and over 500 GB storage space to fully synchronize the entire Ethereum at a peer. The high expenditure of massive storage space and network bandwidth due to blockchain data synchronization impedes the analysis of blockchain data. **(2) Challenge in blockchain data extraction and procession.** Blockchain data is stored at clients in heterogeneous and complex data structures, which cannot be directly analyzed. Meanwhile, the underlying blockchain data is either binary or encrypted. Thus, it is a necessity to extract and process binary and encrypted blockchain data so as to obtain valuable information. However, it is non-trivial to process heterogeneous blockchain data since conventional data analytic methods may not work for this type of data. **(3) Absence of general data extract tools for blockchains.** Although many studies provide open source data extraction tools of blockchain data, most of them can only support to extract partial blockchain data (not all the data). Moreover, most of existing tools can only fulfil specific research tasks. **(4) Absence of basic data explorations for blockchains.** Existing studies only focus on specific data analysis of blockchain data, e.g., transaction graph [6], contract security [7]. However, the basic data explorations like statistic analysis, text analysis and data visualization are missing in most of existing tools.

To address the above challenges, we propose a blockchain

data analytics framework namely XBlock-ETH to analyze Etherium data. In particular, we extract raw data consisting of 8,100,000 blocks of Ethereum. The raw data includes three types of blockchain data: *blocks*, *traces*, and *receipts*. Since the analysis on the raw blockchain data is difficult, we process and categorize the obtained Etherium Blockchain data into six datasets: **(1) Block and Transaction**, **(2) Internal Ether Transaction**, **(3) Contract Information**, **(4) Contract Calls**, **(5) ERC20 Token Transactions**, **(6) ERC721 Token Transactions**. It is non-trivial to process the raw since it requires substantial efforts in extracting useful information from raw data and associating with six datasets. We then conduct statistic analysis on the six refined datasets. We also outlook the potential applications of XBlock-ETH, such as blockchain system analysis, smart contract analysis, and cryptocurrency analysis.

In summary, we highlight the major contributions of this paper as follows:

- The XBlock-ETH data contain the comprehensive on-chain data in contrast of previous works (only cover partial Etherium data). In particular, it includes blockchain data, smart contract data, and cryptocurrency data. Moreover, the well-processed datasets can be easily used for data exploration. Furthermore, XBlock-ETH data formally released online[1] has been periodically updated.
- The XBlock-ETH framework also offers basic statistic and exploration functions to analyze blockchain datasets.This paper also outlines the research opportunities brought by XBlock-ETH. In particular, we discuss the applications of XBlock-ETH in aspects of blockchain system analysis, smart contract analysis and crytocurrency analysis.

The rest of this paper is organized as follows. Section II first gives an overview of blockchain and smart contract technologies. Sections III, IV then present raw data acquisition from Ethereum and data exploration of six datasets. Section V discusses the applications of XBlock-ETH data. Section VI surveys related work. Finally, the paper is concluded in Section VII.

II. Background

Figure 1 presents an overview of Ethereum blockchain, which consists of a number layers from bottom to top: peers, blockchain, smart contracts, and tokens. We next review basic concepts of each layer in Ethereum.

A. Peer and Blockchain

In a nutshell, a blockchain is essentially a chain-like data structure consisting of a number of consecutively-connected blocksblocks. The chain has been maintained by all the peers in a peer-to-peer blockchain network. In a period of time, only one block can be confirmed by the entire blockchain network through a consensus protocol. The block containing the confirmed transactions at that time and the hash value

[1]http://xblock.pro/dataset

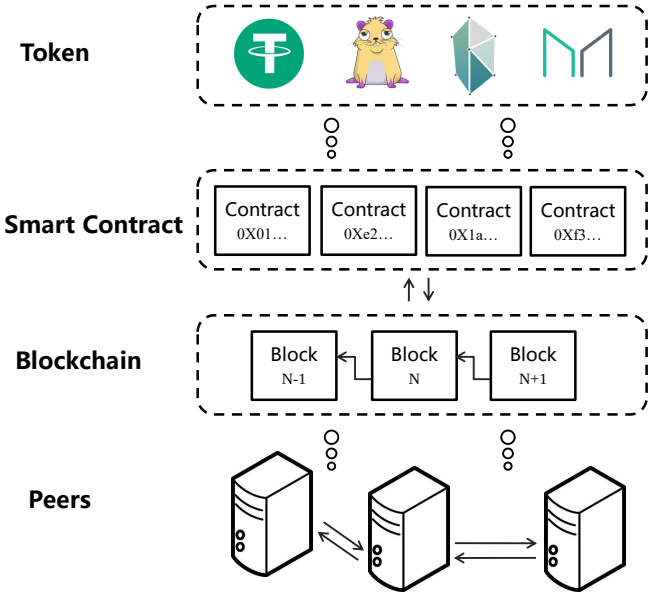


Fig. 1. Overview of Ethereum Blockchain

of the previous block has been generated by a *peer* (a.k.a. miner). After being generated, the block will be validated independently by the other peers. Once the block is validated and confirmed by most of peers in the blockchain network, the transactions in the block will be considered as *completed*. In this way, each peer can trust the whole blockchain (a.k.a. ledger) since the transactions have been validated by all the peers. In other words, blockchain enhances trustworthiness of transactional data through duplicating computation and storage at all the peers.

Thanks to the completeness of the blockchain data in each permission-less blockchain peer, researchers can obtain the entire blockchain data via connecting a blockchain peer the blockchain network. The blockchain data that consists of all the operations done by the users and miners in the blockchain contains substantial business values. For example, the transactional records are essentially operations done by different business parties. The analysis on the blockchain data can help to understand user behaviours in a real-world economic system (e.g., money transferring). Meanwhile, there is a rapid growth of blockchain data, especially in Bitcoin and Ethereum, with the proliferation of blockchain users and transactions. The analysis on blockchain data can be also beneficial to predict the economic trend.

B. Smart Contract

Smart contract that was proposed even earlier than blockchain [8] is a promising technology to reshape the modern industry. Blockchain-based smart contracts are essentially computer programs, in which the execution states are stored on top of blockchain. The blockchain transactions are the messages representing the deployment or invocations of smart contracts. Therefore, blockchain guarantees the trustworthiness of smart contracts.

The incumbent blockchain systems have enable smart contracts. For example, Bitcoin enables users to run a simple script program during the execution of transactions. This script can be regarded as a simple blockchain-based smart contract. However, the Bitcoin script is not Turing-complete so that it cannot enables complex logic expressions in the contract. In contrast, Ethereum enables Turing-complete smart contracts. In Ethereum, smart contract is executed in the environment called Ethereum Virtual Machine (EVM). EVM reads and writes the states (stored in the key-value like database) as the actions defined in a smart contract. During the contract execution, a miner uses "Gas" as a unit to evaluate the consumption of one smart contract. After running the contract , the contract user is charged by the "`GasUsed`" and "`GasPrice`". The more "`GasPrice`" that the users promise to pay for the miner, the faster the contract executes. After the transactions (i.e., operations) are done, EVM will generate a hash value of the state and record it into the blockchain. Therefore, we can learn from Figure 1 that smart contracts on Ethereum are not directly stored on blockchain. They are essentially stored in the states that have been operated by the blockchain.

C. Tokens and clients

It is worth mentioning that Ethereum has two standard token protocols (a.k.a. templates) of smart contracts [4], [9]. These token protocols define the standard variables, functions, and interfaces in the smart contract. With the protocols, users can issue tokens (or so-called cryptocurrencies) based on smart contracts on top of Ethereum. There are four typical tokens USDT[2], Cryptokitties [10], Kyber [11], MarkerDAO[3] as shown in Figure 1 (i.e., the top layer). For an example, a user can publish an ERC20 contract on Ethereum issuing tokens to others. After that, any other users (even contracts) can receive or send the token without a centralized authority (e.g., stock exchange). The standard token protocols greatly enrich the ecosystem of Ethereum so as to make Ethereum become a more flexible financial system. In Section IV-E and IV-F, we will explore the data of tokens in Ethereum.

Ethereum allows that any computer programs can join into the network if they meet the requirement of the protocol just like P2P protocols (e.g., BitTorrent). As a result, there are a number of diverse Ethereum clients that can validate the blocks and transactions. Among most of Ethereum clients, Go-Ethereum (Geth) and Parity have been the most widely used according to the statistic from Ether nodes[4]. Both of them provide JSON-RPC interfaces for users to interact with Ethereum blockchain. Through the JSON-RPC interfaces, user can obtain the blockchain data from Ethereum. Geth has been generally used in many previous studies while the interfaces designed in Geth is not suitable for data acquisition. Even though many researchers attempted to modify source codes of Geth to obtain the detailed run-time data, the whole procedure of the code modification is time consuming and complex. In addition, the obtained data is not absolutely accurate in some cases. Different from Geth, Parity better designs the interfaces so that it can obtain the index of each block corresponding to each piece of the data that we need. The details on data acquisition of blockchain data will be described in Section III.

[2]https://tether.to/
[3]https://makerdao.com/
[4]https://ethernodes.org

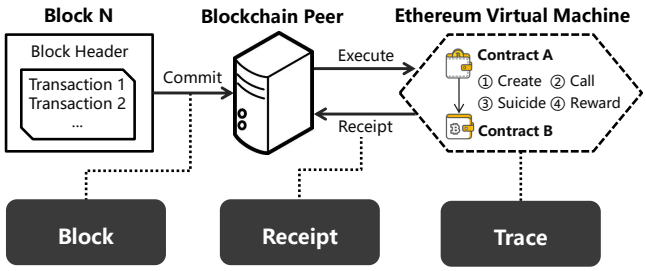


Fig. 2. Raw data collection during Ethereum transaction flow

III. Raw data extraction from Ethereum

This section describes the procedure how the raw data was obtained from Ethereum blockchain. Figure 2 illustrates the typical Ethereum transaction execution flow from Block N to EVM through Blockchain peer. During this procedure, we collect the three types of blockchain raw data: Block, Receipt and Trace. We next describe the details on the composition and acquisition of each kind of raw data.

A. Block

Block data is directly stored in Ethereum blockchain. Each block consists of two components:

- **Block Header:** Block header is the basic information of a block, including the miner's address, timestamp, gas limit, etc.
- **Block Transactions:** Block transactions constructs the body of the block. Each transaction consists of the fields: From, To, Value, Input, etc. If the transaction is used to deploy a contract, the *To* field is "`null`" in the block transaction.

Almost all the Ethereum clients including Geth and Parity offer the interfaces to query the blocks. For example, "`eth_getBlock`" is available in both Geth and Parity with the similar efficiency.

However, we can only obtain little information about the blockchain users through analyzing the block data. This is because the input of block transaction only represents operations to EVM in the contract deployment phase while the contract code will be stored only at the end of the transaction execution and it is not the same as the input of the transaction. Thus, we cannot obtain the exact contract code in the block transaction. Meanwhile, in the contract invocation phase, we cannot know whether the transaction is executed successfully or what kinds of error thrown during the transaction execution since sometimes a contract will send messages or cryptocurrencies to other contracts.

B. Trace

Trace data is essentially the detailed run-time data that was generated in EVM (e.g., internal contract calls, transferring money from the contract to a person). Trace data cannot be directly obtained or observed from the block data, but can be recorded during the contract execution. In this paper, trace data is referred to the data that cannot be obtained before or after the transaction execution, but only appears during the execution. Trace data includes the following types:

- **Create** is the trace including the creator, code, and initial balance when a smart contract is deployed. The creator of a contract can be a person or another smart contract.
- **Call** occurs when money or messages are transferred through different Ethereum addresses. Contract call or Ether transferring is shown as a "`Call`" trace.
- **Suicide** is the trace that smart contract "`suicide`" deletes its code, and refunds the value to a specific account.
- **Reward** is the trace that miners get the Ether reward when they mine a block. The reward value varies depending on the contribution of the miners.

In Geth, the interface of trace is "`debug_traceTransaction`". However, this interface returns all the operations during the transaction, resulting in large resource consumption and low efficiency. Thus, many previous studies attempt to modify the source codes of Geth to obtain the detailed run-time data, while this procedure is extremely time consuming.

In this paper, we adopt "`parity_trace`" in Parity to obtain the trace data. This interface is provided and maintained by the official developer so that the correctness is guaranteed in contrast to Geth. Meanwhile it also provides enough information that we need, such as the basic trace types and errors. Moreover, another advantage of Parity is the updating convenience as the data is indexed by blocks.

C. Receipt

After the transaction is executed, some of the Ethereum states have been changed (e.g., the balance of the account in a token contract). Then the clients need to know what have been changed. To reduce the query overhead of clients, many contracts leave a kind of outputs called "*Event*" in the execution. For example, a standard token contract will output a "`Transfer(from,to,value)`" event to let the clients know what happens during the execution. This kind of outputs is an one-way output, as it is just written in the receipt of the transaction, and can be read by external clients or persons but cannot be read by internal EVMs.

Section IV will then give the statistics of Ethereum data. In particular, there are over 100,000 kinds of cryptocurrencies using smart contracts on Ethereum. As for these token contracts, the receipt data is the important source to learn about the holders, owners, and the user behaviors. Thus, it is necessary to obtain receipt data.

Both Geth and Parity provide the interfaces to get the transaction receipts. The main difference between Geth and Parity interfaces lies in the query index of the receipts. In particular, the receipt of the interface of Geth is "`eth_getTransactionReceipt`" that is indexed by the transaction hash, while the interface of Parity is "`parity_getBlockReceipts`" that is indexed by block number. In this way, Parity is much more efficient than Geth since it can return a batch of receipts in one query.

In summary, there are three kind of raw datasets that can be obtained in Ethereum: block, trace, and receipt. Because of the massive volume and redundant information of the raw data, data procession is necessary to simplify data representation and fasten data analysis for the further study. After compression, the size of the data is about 313 GBytes.

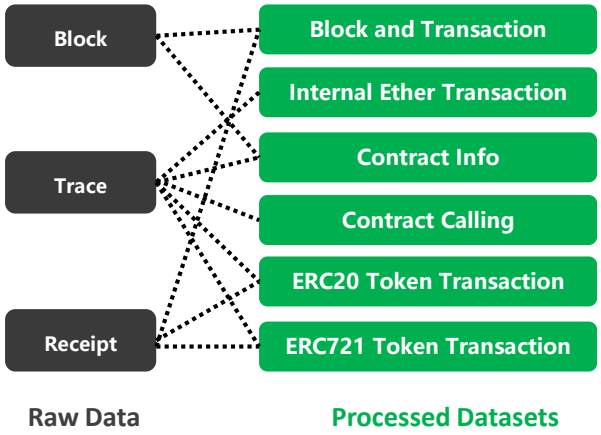


Fig. 3. Mapping from raw data to datasets

TABLE I
STATISTICS OF DATASET 1

Statistics	Values
No. of Blocks	8,100,000
No. of Transactions	491,562,222
No. of Miner Addresses	5,122
Mean of Transaction Counts per Block	60.68
Mean of Block Time	15.33 seconds
Mean of Block Size	11,457 bytes

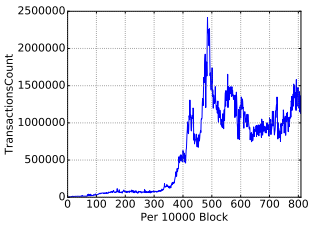
IV. DATA EXPLORATION OF ETHEREUM

In this section, we process the obtained raw data from Ethereum and divide it into six datasets: (1) Block and Transaction, (2) Internal Ether Transaction, (3) Contract Info, (4) Contract Call, (5) ERC20 Token Transaction, (6) ERC721 Token Transaction. The relationship from the raw data to the processed datasets is shown in Figure 3. We can easily observe that the trace data has been the most widely used in the data process. This section will introduce how the datasets are generated, with statistics and observations.

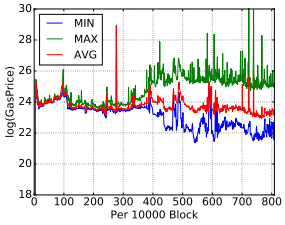
A. Dataset 1: Block and Transaction

To investigate the basic statistics of Ethereum, we extract the information about the blocks and the transactions

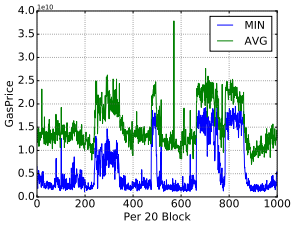
(a) Word Cloud of Miners' Text



(b) Transaction Count



(c) Macro view of GasPrice

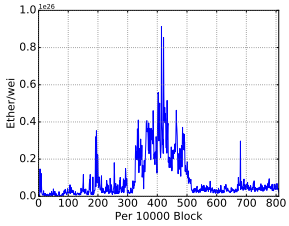


(d) Micro view of GasPrice

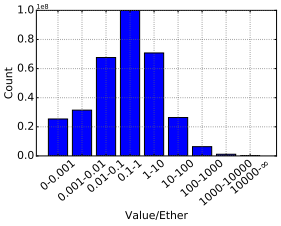
Fig. 4. Visualization of Dataset 1

TABLE II
STATISTICS OF DATASET 2

Statistics	Values
No. of Ether Transactions	329,020,692
No. of Addresses	54,720,018
Mean of Amount of Ethers	22.30
Maximum of Amount of Ether	11,901,464.24



(a) Ether Transferred Amount



(b) Ether Transaction Distribution

Fig. 5. Visualization of Dataset 2

inside the blocks. In particular, there are 8,100,000 blocks and 491,562,222 transactions generated from the block data. For each block, we also obtain the statistic values of the "`gasPrice`": minimum, average, and maximum. Meanwhile, corresponding to the hash of each transaction, the fields of "`minerReward`", "`gasUsed`" and "`error`" are extracted from the receipt and trace.

Regarding the miners of the Ethereum blockchain, there are 5,122 unique addresses of miners as shown in Table I. It implies that there are no more than 5,122 peers that serve as miners since one peer may own more than one addresses. Meanwhile, each miner has the right to write extra texts in the block. So, we also use the word cloud to analyze the texts of miners. Figure 4(a) shows the visualization of the texts of the word cloud. The results show that there are texts left by the mining pool, since most miners are in the mining pool and they have left their names in the blocks to promote their mining capability.

As shown in Table I, the mean of transaction counts per block is 60.68, and the block time is 15.33 seconds. In other words, the average throughput of Ethereum is about 4 transactions per second. Even when most of the network is active, as shown at 4,900,000 blocks in Figure 4(b), the throughput is about 16.7 transactions per second. This result implies that Ethereum still has a long way to go to support real-time Internet applications.

In Ethereum, a miner has a higher priority to package the transactions with higher "`gasPrice`" into the block. The visualization of "`gasPrice`" is shown in Figures 4(c) and 4(d). In a macro view, the "`gasPrice`" is gradually decreasing with the development of the Ethereum community, except for several peaks caused by extremely frequent transaction when the network is congested. In a micro view, we extract the time from 8,000,000 to 8,020,000 blocks and find that such fluctuations of "`gasPrice`" can be observed by the tidal law. This observation implies that the fluctuations of "`gasPrice`" can potentially be predicted.

B. Dataset 2: Internal Ether Transaction

Ether is the native cryptocurrency of Ethereum. The transactions of Ether not only happen in the transactions recorded in the block, but also occur during the smart contract execution. For example, if someone asks a smart contract to send 10 Ethers to another one, the Ether transaction from the contract will not be observed in the block. In some blockchain explorers such as Etherscan[5], this kind of transactions is also called "*Internal Transaction*". To investigate all the Ether transactions, we process the block and trace data to conduct the internal Ether transaction dataset. As shown in Table III, 329,020,672 Ether transactions which occur among 54,720,018 addresses are collected.

The values of Ether have a large variance, as the maximum is 11,901,464.24 Ethers (about 2 billions dollars now) but the mean is only 22.30 Ethers. Figure 5(a) presents statistics on the total transaction amount of every 10,000 blocks. It is shown that the most active time for Ether transaction is the time during 4,000,000 to 4,300,000 blocks, matching with the most active time of Initial Coin Offering (ICO). Regarding the Ether distribution as shown in Figure 5(b), we find that most of Ether transactions fall in the range from 0.1 Ether to 1 Ether, indicating that most of transactions only transfer small amounts of Ethers.

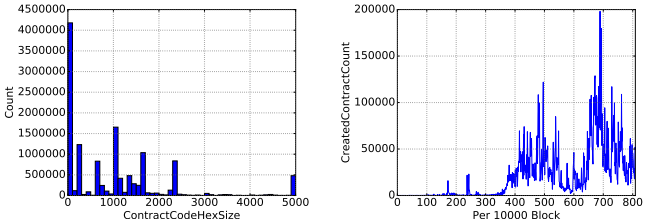
C. Dataset 3: Contract Info

Ethereum can be considered as a platform for smart contracts. To investigate all the smart contracts on Ethereum, we process the trace data to get the basic information of smart contracts, including the creator, created-time, initial value,

[5]http://etherscan.io

TABLE III
STATISTICS OF DATASET 3

Statistics	Values
No. of Created Contracts	16,609,273
No. of Creator Addresses	133,484
No. of Deleted Contracts	5,564,823
No. of Refunded Addresses	19,133,481
Mean of Contract Hex Code Size	958.20



(a) Contract size distribution (b) Count of created contracts

Fig. 6. Visualization of Dataset 3

contract code, creation code. Some smart contracts can be deleted and refund Ethers to someone if they set a "SUICIDE" operation code inside a function. Therefore, we can observe the actions of contract deletions. According to the statistics in Table III, there are 16,609,273 smart contracts created by 133,484 addresses. It implies that there should be a number of users who create multiple contracts.

An abnormal phenomenon observed from Table III is that 5,564,823 contracts are deleted while they refund the Ether balance to 19,133,481 addresses. Generally, a smart contract will not refunds Ethers to multiple addresses during deletion. The reason behind this abnormal phenomenon is that Ethereum has suffered from a Denial of Service (DoS) attacks, in which attackers use a vulnerability of the price of "SUICIDE" to create accounts in Ethereum. Before the vulnerability is fixed, a great amount of contracts are deleted to direct to empty address, leading to many Ethereum peers shutting down as indicated in previous work [12].

Regarding the contract code, we translate the bytecode into hexadecimal code. Figure 6(a) gives the statistics of contract size. Particularly, the mean of contract size is 958.20, indicating that the smart contracts take up little space of storage. The contract size distribution also implies that the sizes of most contracts have focused on some clusters. This indicates that many smart contracts may look similar. This similarity will be further investigate in Dataset 4. Figure 6(b) presents the count of created contracts. It is shown in Figure 6(b) that the number of new smart contracts is increasing, especially at the time after the concept of "ICO" [13] comes out.

D. Dataset 4: Contract Call

In EVM, a smart contract can call another one to invoke some codes or functions. To investigate the calls among the Ethereum contracts (which are represented as addresses),

TABLE IV
STATISTICS OF DATASET 4

Statistics	Values
No. of Contract Calls	1,148,572,009
No. of Calls with Inputs	639,336,722
No. of Calls with Errors	169,463,261

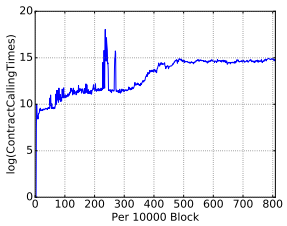
we extract Contract Calls in the execution from the trace dataset. The contract call dataset includes the caller, called address, calling function. As shown in Table IV, it consists of 1,148,572,009 Contract Calls, among which 639,336,722 contain input codes and 169,463,261 contain errors.

Figure 7 gives the visualization of Contract Calls. In particular, Figure 7(a) and Figure 7(c) show that, during the time from 2,300,000 to 2,460,000 blocks, contract calls and errors occur very frequently. This is caused by the DoS attacks mentioned in the above subsection, as the attackers invoked a large number of contracts in batches and some of them throw errors. Figure 7(b) gives the distribution of call types. In particular, Figure 7(b) shows that most of developers prefer to use "`call`" and "`delegatecall`" rather than "`staticcall`" and "`callcode`", since the logic of "`call`" and "`delegatecall`" is clearer and more practical than other two calls. Figure 7(d) shows the error types during calling contract, indicating that most of errors are caused by "Out of gas", which is mainly resulted from the wrong settings of message senders. The second most common error is "`Reverted`", which is a manually-thrown exception by the developers. Moreover, other errors such as "Bad instruction" and "Bad jump destination" are often caused by the contract codes themselves.

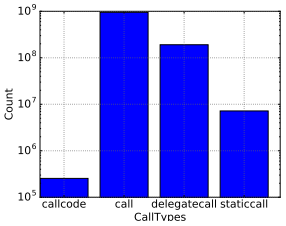
Generally, the compiler of smart contracts will use the hash value of function name and parameters as the entry of the function. In other words, in Ethereum smart contracts, the identical function in source code will have the identical entry in the complied contract code. We then count the calling contract functions to see what functions are the most common ones. The distribution of top-10 functions is shown in Figure 7(e). The results show that most of the calling functions concentrated on some types of them. For example, top-10 functions have occupied 46.32% of the contract calls. Moreover, after verifying the hash values of functions with the open-source contracts, we obtain the functions in source code. We then have the top-3 functions: "`transfer(address,uint256)`", "`balanceOf(address)`" and "`transferFrom(address,address,uint256)`". This result implies that the most common contract calls are about tokens and there might be a great similarity among the contracts due to the similar calls.

E. Dataset 5: ERC20 Token Trasnaction

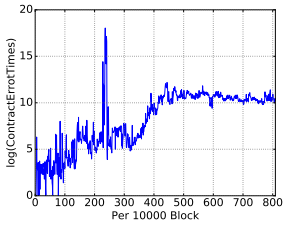
From the above analysis, we observe that the most active smart contracts on Ethereum now are the token contracts. We



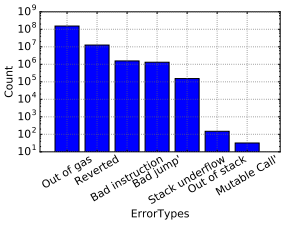
(a) Count of Contract Call



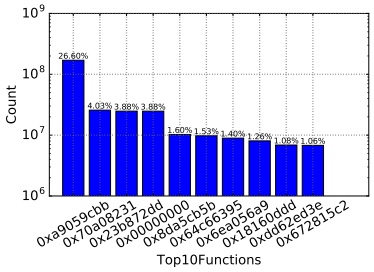
(b) Call Type Distribution



(c) Count of Contract Error



(d) Error Type Distribution



(e) Calling Count of Top 10 Contract Function

Fig. 7. Visualization of Dataset 4

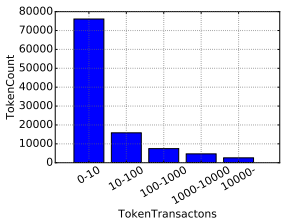
next further investigate the token contracts. In order to collect the information of tokens, we process the receipt dataset to extract the standard events, which are defined in the standard ERC20 protocol of Ethereum community [4]. Additionally, each ERC20 token contains basic information like name, symbol, total supply, etc. We then send calls to the local Ethereum peers to collect such basic information of ERC20 tokens.

As shown in Table V, 106,683 smart contracts are considered as ERC20 contracts, since they output the events that are defined as the standard ERC20 token transactions. There are 227,698,645 ERC20 transactions among 42,146,575 holder addresses. Generally, the number of holder addresses could be much more larger than that of exact human holders because a user may own several addresses. Meanwhile some token issuers will send the tokens to other users without their permissions (also called *token air-drop* [14]).

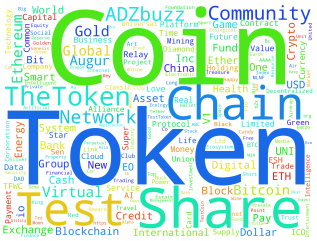
Figure 8(a) shows the transaction count distribution for each ERC20 token. We can easily observe the Matthew effect [15] from Figure 8(a) as most of token transactions happen in few token contracts. Figure 8(b) presents the word cloud of names of ERC20 tokens. It is shown in Figure 8(b) that the most common words are "Chain", "Coin", and "Share", on which the most ERC20 tokens focus. In addition, another common

TABLE V
STATISTICS OF DATASET 5

Statistics	Values
No. of ERC20 Contracts	106,683
No. of ERC20 Transactions	227,698,645
No. of Holder Addresses	42,146,575



(a) ERC20 Popularity Distribution



(b) Word Cloud of ERC20 Tokens

Fig. 8. Visualization of Dataset 5

word is "Test", implying that many ERC20 contracts deployed on Ethereum are just for the testing purpose.

F. Dataset 6: ERC721 Token Trasnaction

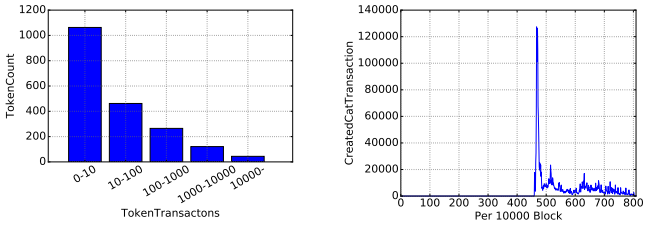
ERC721 token is another contract protocol proposed by Ethereum community [9]. Different from ERC20 token, ERC721 token is indivisible. In the contract function, the parameter is not the value of token but the token ID. For example, a virtual pet in smart contract could be a ERC721 token, which is not separable but can be transferred.

Table VI presents the statistics of ERC721 contracts. We find that 1,954 ERC721 contracts contain 7,524,827 token transactions and 414,829 holder addresses. It is worth mentioning that some of the collected contracts do not follow the standard ERC721 protocol exactly. These contracts are also included in the dataset since they output the token transferred events in the receipt. Figure 9(a) shows the popularity distribution of ERC721 tokens. Compared with ERC20 tokens, the amount of ERC721 tokens is much lower. The major reason is that ERC721 applications require much more workloads on visualization at each token, consequently improving the development difficulty.

We also investigate a popular ERC721 token contract called CryptoKitties. It is one of the most famous ERC721 token contracts, selling the virtual cats as tokens. Each cat is represented as a token in the ERC721 contract. We count the turnover times distributed by birth block of the cats, as shown in Figure 9(b). Figure 9(b) also shows that the cats that were born in 4,500,000 to 5,000,000 blocks have the higher turnover times than others. At that time, the type of CryptoKitties reaches the peak. The time to obtain the peak in Figure 9(b) is almost the same as that in Figure 4(b) and Figure 4(c), implying that the popularity of CryptoKitties leads to the congestion of Ethereum.

TABLE VI
STATISTICS OF DATASET 6

Statistics	Values
No. of ERC721 Contracts	1,954
No. of ERC721 Transactions	7,524,827
No. of Holder Addresses	414,829



(a) ERC721 Popularity Distribution (b) CryptoKitties Turnover Times

Fig. 9. Visualization of Dataset 6

V. APPLICATIONS OF XBLOCK-ETH

This section presents applications of XBlock-ETH framework. As shown in Figure 1, the architecture of Ethereum consists of peers, blockchain, smart contracts and tokens. Thus, we also categorize the applications according to top-3 layers (i.e., blockchains, smart contracts and tokens). Meanwhile, we also discuss the research opportunities in each layer.

A. Blockchain System Analysis

Since XBlock-ETH processes data from realistic blockchain systems, it can be used to support the following applications.

1) Decentralization Analysis: The decentralization is one of the key features of blockchain systems. However, there are few studies on the decentralization evaluation of the blockchain systems. In particular, the work of [16] presents the measurement of the mining pool for Bitcoin. Although Gencer et al. [17] present a measurement study on the decentralization level of Bitcoin and Ethereum, their study only consider several metrics such as network bandwith, mining power and fairness. In contrast, our XBlock-ETH data offers a more comprehensive measurement on Ethereum. Moreover, our work can be used to analyze the decentralization of users, contract owners and miners. In addition, our XBlock-ETH can also be used to make comparison with other blockchain systems, such as Bitcoin, EOS or other blockchain systems.

2) Gasprice Prediction: Since the transaction fees are equal to "`gasPrice`" times "`gasUsed`", the users can control the "`gasUsed`" in a reasonably low range to minimize the transaction fees charged by miners. Meanwhile, we can learn from Section IV-A that there is always a gap between the minimum "`gasPrice`" and the average "`gasPrice`" in a block, leading to the opportunity to save fees. Recent studies such as Other-tech [18], Gitcoin [19], Majuri [20] analyze the "`gasPrice`" of Ethereum while several Ethereum websites (e.g., Etherscan[6], Etherchain[7]) provide tools to predict the "`gasPrice`" in a short time. However, those tools are essentially *black boxes* and the accuracy of them cannot be assured. In summary, the prediction of "`gasPrice`" has great economic value such that the user of Ethereum can save the money or shorten waiting time through the "`gasPrice`" prediction while it is worthwhile to conduct an in-depth study in the future.

3) Performance Benchmark: Performance is crucial to blockchain systems. There are a number of studies on blockchain performance optimizations, such as Omniledger [21], Algorand [22] and RapidChain [23]. Meanwhile, some optimized blockchain systems (e.g., Monoxide [24]) adopt the realistic blockchain transaction data to conduct performance evaluation for blockchain systems. To compare the performance of different optimization methods, a common benchmark of real-world user cases for blockchain systems is needed. Zheng et al. [25] and BlockBench [26] propose the performance evaluation of blockchain systems. The performance benchmark requires simulating the user behaviors and obtaining data similar to real-world blockchain systems. In this aspect, the XBlock-ETH framework can be regarded as a benchmark since the source data is generated exactly by the real-world users.

B. Smart Contract Analysis

As one of the most popular smart contract platforms, Ethereum has attracted a large number of software developers as well a huge number of smart contracts. Therefore, Ethereum has a more active developer community compared with other smart contract platforms such as EOS and Tron, which claim to have the higher throughput and lower latency than Ethereum. Consequently, our XBlock-ETH framework (on top of Ethereum) can be used in the studies of smart contracts. We summarize the potential applications of XBlock-ETH as follows.

1) Contract Similarity and Recommendation: As indicated in Section IV, there is a great similarity between the smart contract codes and call of smart contracts. Code similarity evaluation is a traditional research topic in software engineering as a number of studies concentrate on code similarity detection [27] [28] [29]. Several recent studies focus on similarity analysis of smart contracts. In particular, Etherscan[6] provides the query system based on similar contracts. Finding the similar contracts is beneficial to the developers during developing new contracts. For example, developers can estimate the user behaviors before the publishing the contract. Meanwhile, Huang et al. [30] propose the method to recommend differentiated codes to update smart contracts based on the existing codes of smart contracts. In addition, in the aspect of users, recommending the similar smart contract will help users to find the contracts suitable for themselves.

[6]http://etherscan.io
[7]http://etherchain.org

2) Contract Developer Analysis: Developer analysis that is another traditional research topic in software engineering includes developer network analysis [31], behavior analysis [32], fault prediction [33], and so on. With respect to developer analysis, XBlock-ETH also includes a large network of smart contract developers. For example, there some on-chain libraries deployed and provided by different developers; these libaries can be invoked by others. Each developer can be identified by his/her own Ethereum address. Thus, the contract calling network can be also regarded as the collaboration network of contract developers. The network and structure of developer collaboration may inform us about the reliability of the contract codes. For example, the developer who develops a smart contract with vulnerabilities will have a higher risk to develop new contracts with vulnerabilities than others. In this sense, our XBlock-ETH can be beneficial to the developer analysis after analyzing smart contracts of developers.

3) Contract Vulnerability Detection: The security of smart contracts has been a hot research topic in blockchain research community. In particular, the vulnerability of smart contracts has attracted extra attentions. A number of malicious attacks on Ethereum (e.g., TheDAO attack) have already resulted in huge loss (in terms of tens of millions of dollars) [34]. To prevent smart contracts from malicious attacks, the vulnerability detection on contracts is a critical step. There are some recent attempts in vulnerability detection. For example, Oyente [7], Zeus [35], teEther [36], S-gram [37], ContractFuzzer [38] propose the tools of vulnerability detection on smart contracts. In some cases, the vulnerability detection methods of smart contracts can be inspired and motivated by traditional software vulnerability detection methods as they are essentially equivalent to the verification of the codes. In this aspect, several studies focus on verifying contract codes on blockchains; these contract codes are also called "bytecode" or "opcode". Our XBlock-ETH that essentially includes the data of contract codes can be applied to contract vulnerability detection.

4) Fraud Detection: Due to the huge economic value and the popularity of smart contracts, smart contracts can be exploited by malicious users as scams. For example, crowdfunding contracts with a promised huge return to attract victims for investment. It is reported in [39] that Ponzi scam contracts can defraud others' cryptocurrencies. Several approaches [39]–[42] have been proposed to detect the fraud contracts on Ethereum. Most of the methods are mainly based on the codes and transaction records of smart contracts while they are included in XBlock-ETH data. Thus, XBlock-ETH data can be further leveraged in fraud detection.

C. Cryptocurrency Analysis

Blockchain-based cryptocurrency has become a hot topic recent years due to the decentralization and the reduced cost. There are a large amount of cryptocurrencies in Ethereum, including the Ether, ERC20 tokens and ERC721 tokens. It is shown in the CoinMarketCap[8] that more than 2,000 kinds of tokens can be used in third-party exchange. Therefore, cryptocurrency analysis based on blockchain data can bring huge financial values. We roughly categorize the cryptocurrency analysis into cryptocurrency transferring analysis, cryptocurrency price analysis and fake user detection, which are explained as follows.

1) Cryptocurrency Transferring Analysis: Analysis on cryptocurrency transactions is a preliminary step to conduct cryptocurrency transferring analysis. Regarding Ether transferring, Chen et al. [6] propose the graph analysis on Ether transactions and derive some insights from graph analysis. With regard to ERC20/ERC721 tokens, Victor et al. [43] and Somin et al. [44] propose the analysis of the token trading network. After the analysis on cryptocurrency transactions, the further analysis on user behaviours can be done. For example, the users of tokens may form different communities. The community discovery can be conducted through analyzing cryptocurrency transactions. Moreover, the anonymity of blockchain-based cryptocurrency can result in money-laundering behaviors, which can be essentially identified and detected via cryptocurrency transaction analysis. Our XBlock-ETH data offers the potential solutions to these issues.

2) Cryptocurrency Price Analysis: The price of blockchain-based cryptocurrencies has been affected by multiple different factors such as government policies, technology innovations, social sentiment and business activities. Several recent studies focus on the price analysis and prediction of cryptocurrencies [45]–[47]. The typical cryptocurrency price analysis consists of three steps: (i) collect price data form the cryptocurrency exchanges, (ii) identify the relevance between cryptocurrency prices and other factors, (iii) forecast the future prices and predict the potential profits. However, the price of cryptocurrencies can sometimes be maliciously controlled by some parties. Thus, the data cleaning process is necessary to obtain the accurate and normal cryptocurrency price data. Our XBlock-ETH also contains cryptocurrency price data, which can be used for cryptocurrency price analysis while the raw receipt data may require the further preprocess to benfit the future analysis.

3) Fake User Detection: Fake user detection [48]–[50] is a traditional research topic in social networks. The cryptocurrency users in blockchain systems also form social-network like communities, in which there are also some fake users controlled by the developers to improve the DApps activity rankings. Because the DApp (or cryptocurrency) ranking is based on some metrics related to the user activities, such as Daily Active Users (DAU). Therefore, many developers exploit the loophole to fabricate some fake users to improve activities so as to gain higher rankings. Although some DApp websites, such as DAppReview[9] mark the cryptocurrencies with fake users, this kind of fake user detection is almost done in a black box or manually. In addition, there are few studies on fake user detection on cryptocurrency. The permission-less blockchain systems which are often free of charge may advocate more

[8]https://coinmarketcap.com/all/views/all/

[9]http://dapp.review

frequent fake user activities than permissioned blockchain systems. Our XBlock-ETH will be further improved to support the fake user detection in the future.

VI. Related Work and Discussion

Some previous studies on Ethereum data will be described and discussed in this section. We categorize the state-of-the-art literature into two types: *Data tools* and *Data analysis*.

Regarding Ethereum data tools, some studies provide open-source tools or APIs with users to obtain the data. For example, EtherQL [51] offers a query layer for Ethereum. Blocksci [52] constructs a platform for researchers to analyze the blockchain data. DataEther [53] is a tool to obtain the data from Ethereum, with code modification of the Ethereum clients. Google BigQuery [54] imports the data of Bitcoin and Ethereum and enables researchers to analyaze the data online while updating Ethereum data has been stopped for a long time. Meanwhile, it is pretty challenging for researchers to download, update and analyze the blockchain data. There are also some websites offering data APIs for developers to use or analyze, including Amberdata[10]. However, these third-party APIs always restrict the usage rating so that it is difficult for researchers to crawl all the data. In summary, most of these studies only offer tools or APIs to researchers while failing to offer well-processed up-to-date datasets.

Some recent studies provide the analysis on the Ethereum data. For example, studies of [39]–[41] propose the contract classification methods to detect Ponzi schemes. Moreover, Chen et al. [6] analyze the transactions and construct three graphs to observer the behaviors on Ethereum. Furthermore, the work of [55] analyzes the ERC20 tokens on Etherem and find un-standard token. Another popular research area on Ethereum data is the smart contracts security. For example, Oyente [7], Zeus [35] propose the security analysis tools for Ethereum smart contracts to find the vulnerable codes. Although some of these studies release some datasets, most of them are only suitable for specific research questions. Furthermore, most of them are difficult to be updated.

It is worth mentioning that XBlock-ETH does not contain the off-chain data such as the price data in exchanges, the source code of verified smart contracts, the behavior on Github of the DApps even if they are also crucial for the analysis. Since those data are not generated by the Ethereum, we only concentrate on the on-chain data in this paper.

VII. Conclusion and Future work

This paper introduces a well-processed up-to-date on-chain dataset of Ethereum, namely XBlock-ETH, which includes the data of the Ethereum blockchain, smart contracts and cryptocurrencies. Moreover, comprehensive statistics and exploration of the datasets are presented. The XBlock-ETH datasets have been released on XBlock.pro website. Furthermore, the research opportunities of the XBlock-ETH datasets are also outlined.

[10]http://amberdata.io

Our XBlock-ETH is promising to promote the studies on Ethereum. The future improvements are listed as following: **(1) More features:** The exploration of the basic features of the datasets are given in this paper. Ethereum is a complex ecosystem that includes decentralized finance, stable coin, and so on. More features of the Ethereum data will be explored in the future. **(2) More data from exchanges and open-source communities:** The off-chain data is also important since it provides the information of off-chain behaviors of both developers and users. In the future, the off-chain data will be collected. **(3) Combined analysis with other blockchain systems:** There are some other blockchain systems that have also attracted a large number of users and developers. The combined analysis between Ethereum and other permissionless blockchains will be conducted in the future.

References

[1] Z. Zheng, S. Xie, H.-N. Dai, and H. Wang, "Blockchain challenges and opportunities: A survey," *International Journal of Web and Grid Services*, 2016.

[2] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[3] V. Buterin *et al.*, "Ethereum white paper," 2013.

[4] V. Buterin and F. Vogelsteller, "Erc20 token standard," *URL: https://theethereum. wiki/w/index. php/ERC20 Token Standard*, 2015.

[5] H.-N. Dai, Z. Zheng, and Y. Zhang, "Blockchain for internet of things: A survey," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8076 – 8094, 2019.

[6] T. Chen, Y. Zhu, Z. Li, J. Chen, X. Li, X. Luo, X. Lin, and X. Zhange, "Understanding ethereum via graph analysis," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1484–1492.

[7] L. Luu, D.-H. Chu, H. Olickel, P. Saxena, and A. Hobor, "Making smart contracts smarter," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS*. ACM, 2016, pp. 254–269.

[8] N. Szabo, "The idea of smart contracts," 1997.

[9] W. Entriken, D. Shirley, J. Evans, and N. Sachs, "Erc-721 non-fungible token standard," *Ethereum Foundation*, 2018.

[10] O. Kharif, "Cryptokitties mania overwhelms ethereum networks processing," *Bloomberg, 2017*.

[11] Y. V. L. Luu, "Kybernetwork: A trustless decentralized exchange and payment service," *URl: https://home. kyber. network/assets/KyberNetworkWhitepaper. pdf*.

[12] T. Chen, X. Li, Y. Wang, J. Chen, Z. Li, X. Luo, M. H. Au, and X. Zhang, "An adaptive gas cost mechanism for ethereum to defend against under-priced dos attacks," in *International Conference on Information Security Practice and Experience*. Springer, 2017, pp. 3–24.

[13] S. T. Howell, M. Niessner, and D. Yermack, "Initial coin offerings: Financing growth with cryptocurrency token sales," National Bureau of Economic Research, Tech. Rep., 2018.

[14] P. van Valkenburgh, "A token airdrop may not spare you from securities regulation," 2017.

[15] R. K. Merton, "The matthew effect in science: The reward and communication systems of science are considered," *Science*, vol. 159, no. 3810, pp. 56–63, 1968.

[16] C. Wang, X. Chu, and Q. Yang, "Measurement and analysis of the bitcoin networks: A view from mining pools," *arXiv preprint arXiv:1902.07549*, 2019.

[17] A. E. Gencer, S. Basu, I. Eyal, R. Van Renesse, and E. G. Sirer, "Decentralization in bitcoin and ethereum networks," *arXiv preprint arXiv:1801.03998*, 2018.

[18] Jin.S, "Ethereum gas price analysis," 2018.

[19] K. Owocki, "A brief history of gas prices on ethereum," 2018.

[20] Y. Majuri, "Simply explained: Ethereum gas," 2018.

[21] E. Kokoris-Kogias, P. Jovanovic, L. Gasser, N. Gailly, E. Syta, and B. Ford, "Omniledger: A secure, scale-out, decentralized ledger via sharding," in *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2018, pp. 583–598.

[22] Y. Gilad, R. Hemo, S. Micali, G. Vlachos, and N. Zeldovich, "Algorand: Scaling byzantine agreements for cryptocurrencies," in *Proceedings of the 26th Symposium on Operating Systems Principles*. ACM, 2017, pp. 51–68.

[23] M. Zamani, M. Movahedi, and M. Raykova, "Rapidchain: Scaling blockchain via full sharding," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2018, pp. 931–948.

[24] J. Wang and H. Wang, "Monoxide: Scale out blockchains with asynchronous consensus zones," in *16th USENIX Symposium on Networked Systems Design and Implementation, NSDI)*, 2019, pp. 95–112.

[25] P. Zheng, Z. Zheng, X. Luo, X. Chen, and X. Liu, "A detailed and real-time performance monitoring framework for blockchain systems," in *Proceedings of the 40th International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP*. ACM, 2018, pp. 134–143.

[26] T. T. A. Dinh, J. Wang, G. Chen, R. Liu, B. C. Ooi, and K.-L. Tan, "Blockbench: A framework for analyzing private blockchains," in *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017, pp. 1085–1100.

[27] M. Chilowicz, E. Duris, and G. Roussel, "Syntax tree fingerprinting for source code similarity detection," in *2009 IEEE 17th International Conference on Program Comprehension*. IEEE, 2009, pp. 243–247.

[28] L. Luo, J. Ming, D. Wu, P. Liu, and S. Zhu, "Semantics-based obfuscation-resilient binary code similarity comparison with applications to software plagiarism detection," in *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM, 2014, pp. 389–400.

[29] K. Lemhöfer and T. Dijkstra, "Recognizing cognates and interlingual homographs: Effects of code similarity in language-specific and generalized lexical decision," *Memory & Cognition*, vol. 32, no. 4, pp. 533–550, 2004.

[30] Y. Huang, Q. Kong, N. Jia, X. Chen, and Z. Zheng, "Recommending differentiated code to support smart contract update," in *Proceedings of the 27th International Conference on Program Comprehension*. IEEE Press, 2019, pp. 260–270.

[31] A. Meneely, L. Williams, W. Snipes, and J. Osborne, "Predicting failures with developer networks and social network analysis," in *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*, 2008, pp. 13–23.

[32] L. Layman, L. Williams, and R. S. Amant, "Toward reducing fault fix time: Understanding developer behavior for the design of automated fault detection tools," in *First International Symposium on Empirical Software Engineering and Measurement, ESEM*. IEEE, 2007, pp. 176–185.

[33] E. J. Weyuker, T. J. Ostrand, and R. M. Bell, "Using developer information as a factor for fault prediction," in *Proceedings of the Third International Workshop on Predictor Models in Software Engineering*, 2007, p. 8.

[34] M. I. Mehar, C. L. Shier, A. Giambattista, E. Gong, G. Fletcher, R. Sanayhie, H. M. Kim, and M. Laskowski, "Understanding a revolutionary and flawed grand experiment in blockchain: the dao attack," *Journal of Cases on Information Technology*, vol. 21, no. 1, pp. 19–32, 2019.

[35] S. Kalra, S. Goel, M. Dhawan, and S. Sharma, "Zeus: Analyzing safety of smart contracts." in *NDSS*, 2018.

[36] J. Krupp and C. Rossow, "teether: Gnawing at ethereum to automatically exploit smart contracts," in *27th USENIX Security Symposium, Security)*, 2018, pp. 1317–1333.

[37] H. Liu, C. Liu, W. Zhao, Y. Jiang, and J. Sun, "S-gram: towards semantic-aware security auditing for ethereum smart contracts," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. ACM, 2018, pp. 814–819.

[38] B. Jiang, Y. Liu, and W. Chan, "Contractfuzzer: Fuzzing smart contracts for vulnerability detection," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. ACM, 2018, pp. 259–269.

[39] W. Chen, Z. Zheng, J. Cui, E. Ngai, P. Zheng, and Y. Zhou, "Detecting ponzi schemes on ethereum: Towards healthier blockchain technology," in *Proceedings of the 27th International Conference on World Wide Web, WWW*. ACM, 2018.

[40] M. Bartoletti, S. Carta, T. Cimoli, and R. Saia, "Dissecting ponzi schemes on ethereum: identification, analysis, and impact," *Future Generation Computer Systems*, 2019.

[41] W. Chen, Z. Zheng, E. C.-H. Ngai, P. Zheng, and Y. Zhou, "Exploiting blockchain data to detect smart ponzi schemes on ethereum," *IEEE Access*, vol. 7, pp. 37 575–37 586, 2019.

[42] C. F. Torres, M. Steichen, and R. State, "The art of the scam: Demystifying honeypots in ethereum smart contracts," in *Proceedings of the 28th USENIX Conference on Security Symposium*, ser. SEC'19. Berkeley, CA, USA: USENIX Association, 2019, pp. 1591–1607. [Online]. Available: http://dl.acm.org/citation.cfm?id=3361338.3361449

[43] F. Victor and B. K. Lüders, "Measuring ethereum-based erc20 token networks," in *International Conference on Financial Cryptography and Data Security*, 2019.

[44] S. Somin, G. Gordon, and Y. Altshuler, "Network analysis of erc20 tokens trading on ethereum blockchain," in *International Conference on Complex Systems*. Springer, 2018, pp. 439–450.

[45] C. Lamon, E. Nielsen, and E. Redondo, "Cryptocurrency price prediction using news and social media sentiment," *SMU Data Sci. Rev*, vol. 1, no. 3, pp. 1–22, 2017.

[46] J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, "Cryptocurrency price prediction using tweet volumes and sentiment analysis," *SMU Data Science Review*, vol. 1, no. 3, p. 1, 2018.

[47] W. Mensi, K. H. Al-Yahyaee, and S. H. Kang, "Structural breaks and double long memory of cryptocurrency prices: A comparative analysis from bitcoin and ethereum," *Finance Research Letters*, vol. 29, pp. 222–230, 2019.

[48] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2012, pp. 15–15.

[49] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Eleventh international AAAI conference on web and social media*, 2017.

[50] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[51] Y. Li, K. Zheng, Y. Yan, Q. Liu, and X. Zhou, "Etherql: a query layer for blockchain system," in *International Conference on Database Systems for Advanced Applications*. Springer, 2017, pp. 556–567.

[52] H. Kalodner, S. Goldfeder, A. Chator, M. Möser, and A. Narayanan, "Blocksci: Design and applications of a blockchain analysis platform," *arXiv preprint arXiv:1709.02489*, 2017.

[53] T. Chen, Z. Li, Y. Zhang, X. Luo, A. Chen, K. Yang, B. Hu, T. Zhu, S. Deng, T. Hu *et al.*, "Dataether: Data exploration framework for ethereum," in *Proceedings of the 39th IEEE International Conference on Distributed Computing Systems*, 2019.

[54] J. Tigani and S. Naidu, *Google BigQuery Analytics*. John Wiley & Sons, 2014.

[55] T. Chen *et al.*, "Tokenscope: A system for detecting inconsistent behaviors of cryptocurrency tokens." 2019.

Peilin Zheng is a student at Sun Yat-sen University, Guangzhou, China. His research interests include performance monitoring and evaluation on blockchain, optimization of smart contracts, and blockchain-based decentralized applications.

Zibin Zheng is a professor at Sun Yat-sen University, Guangzhou, China. He received Ph.D. degree from The Chinese University of Hong Kong in 2011. He received ACM SIGSOFT Distinguished Paper Award at ICSE' 10, Best Student Paper Award at ICWS' 10, and IBM Ph.D. Fellowship Award. His research interests include services computing, software engineering, and blockchain.

Hong-Ning Dai is an Associate Professor in Faculty of Information Technology at Macau University of Science and Technology. He obtained his PhD in Computer Science and ngineering from the Department of Computer Science and Engineering at the Chinese University of Hong Kong in 2008. His research interests include wireless networks, mobile computing, and distributed systems