IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received September 26, 2021, accepted October 5, 2021, date of publication October 11, 2021, date of current version October 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3119291

The 51% Attack on Blockchains: A Mining Behavior Study

FREDY ANDRES APONTE-NOVOA[1,2], **ANA LUCILA SANDOVAL OROZCO**[1,3], **RICARDO VILLANUEVA-POLANCO**[1], **AND PEDRO WIGHTMAN**[4], **(Senior Member, IEEE)**
[1]Department of Computer Science and Engineering, Universidad del Norte, Puerto Colombia, Barranquilla 081001, Colombia
[2]Department of Systems Engineering, Universidad Santo Tomás, Tunja 150009, Colombia
[3]Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Office 431, Universidad Complutense de Madrid (UCM), 28040 Madrid, Spain
[4]Escuela de Ingeniería, Ciencia y Tecnología, Universidad del Rosario, Bogotá 111711, Colombia

Corresponding author: Fredy Andres Aponte-Novoa (faponte@uninorte.edu.co)

This work was supported in part by the Program "Convocatoria 779 de 2017—Doctorado Nacional Convocatoria para la Formación de Capital Humano de Alto Nivel para el Departamento de Boyacá–2017," and in part by the Universidad del Norte.

ABSTRACT The applications that use blockchain are cryptocurrencies, decentralized finance applications, video games, and many others. Most of these applications trust that the blockchain will prevent issues like fraud, thanks to the built-in cryptographic mechanisms provided by the data structure and the consensus protocol. However, blockchains suffer from what is called a 51% attack or majority attack, which is considered a high risk for the integrity of these blockchains, where if a miner, or a group of them, has more than half the computing capability of the network, it can rewrite the blockchain. Even though this attack is possible in theory, it is regarded as hard-achievable in practice, due to the assumption that, with enough active members, it is very complicated to have that much computing power; however, this assumption has not been studied with enough detail. In this work, a detailed characterization of the miners in the Bitcoin and Crypto Ethereum blockchains is presented, with the aim of proving the computing distribution assumption and creating profiles that may allow the detection of anomalous behaviors and prevent 51% attacks. The results of the analysis show that, in the last years, there has been an increasing concentration of hash rate power in a very small set of miners, which generates a real risk for current blockchains. Also, that there is a pattern in mining among the main miners, which makes it possible to identify out-of-normal behavior.

INDEX TERMS 51% attack, bitcoin, blockchain, double-speding, ethereum, hash rate.

I. INTRODUCTION

Blockchain technology promises to become a great opportunity to provide different solutions for society's problems, "…Like the internet reinvented communication, blockchain may similarly disrupt transactions, contracts, and trust – the underpinnings of business, government, and society" [1]. It also has been defined as "a perpetually updated record of transactions independently saved by users across the internet"; in other words, it is an immutable distributed ledger [2]. The basic operation of a blockchain consists of the secure administration of a shared ledger, where transactions are verified and stored in a network of anonymous nodes that does not have a central authority. A blockchain can be public or private, where read or write permissions can be configured. Some mathematical tools, like cryptographic hash functions, as well as computational ones, like a p2p network and consensus algorithms, allow the blockchain to work, not only to execute transactions but also to protect the integrity and anonymity of the users. However, blockchains, despite their strong data structure and other benefits, has some shortcomings, like the computational cost to run the blockchain's consensus algorithms, which requires the solution of complex mathematical problems in parallel by a large number of users, all competing to finish first in a global race [1].

Despite the effort required to solve the problems is high, there are users with enough computing power that could not only solve them quickly and in a distributed manner, but also try to take over the network by generating a new version of the blockchain that would allow them to spend a particular number of coins at least two times, violating one of the design principles of digital currencies. This is called the double-spending attack, and it poses a high risk for the security of the blockchain. In particular, this attack can be

The associate editor coordinating the review of this manuscript and approving it for publication was Diana Gratiela Berbecaru.

carried out by a miner if this miner has more than 51% of all the mining power. This is also known as the 51% attack, which can be defined as a hash-based attack that occurs in a blockchain when one or more miners take control of at least 51% of all the mining of hash or the computation in the blockchain network. With this computational power, a miner may alter transactions in a blockchain network and hence hinder the process of storing a new block [3].

By executing a 51% attack, a miner can arbitrarily manipulate and modify the information on the blockchain. Specifically, an attacker can exploit this vulnerability to carry out the following attacks: a) reverse transactions and initiate a double-spending attack; that is, spending the same coins multiple times; b) exclude and modify the order of transactions; c) hamper the normal mining operations of other miners; and d) prevent the confirmation operation of normal transactions [4].

If a few miners gather the mining power in a blockchain that uses the Proof of Work (PoW) consensus mechanism, then fear of an inadvertent situation may occur, such as one group controlling more than 50% of the computing power hash [5]. In January 2014, the mining group ghash.io reached 42% of the total hash power in bitcoin, which caused several miners to voluntarily leave the group, while ghash.io in a press release assured the Bitcoin community that would avoid reaching the 51% hash power threshold [6]. In this case, there was a self-control mechanism based on honor; however, this kind of issue cannot be left to chance if the blockchain would like to become a more widely accepted infrastructure for transactions.

To the best of our knowledge, the literature lacks a work making a deep analysis of the behavior of the miners to identify patterns that could help detect early signs of a 51% attack taking place. This paper presents a characterization of the principal miners in Bitcoin and Ethereum. In particular, we focus on miners with a hash rate of more than 1% per defined period to create profiles that may allow the detection of anomalous behaviors. In addition, a validation of the theoretical work of [7] is performed, based on the actual transactions dataset of these blockchains.

This article is organized as follows: Section II presents some consensus algorithms. Section III presents some works related to techniques for preventing 51% attacks. In Section IV, the methodology used in its data selection, preprocessing and mining characterization phases are detailed. In Section V, we present the results obtained organized in the number of miners, hash rate/share of the miner, percentage of mined consecutively, the profile of miners, and analysis of double-spending. Finally, the conclusions and future works are presented in Section VI

II. CONSENSUS ALGORITHMS

This section presents the consensus algorithms Proof of Work, Proof of Stake, and Hybrid form of these algorithms, also other proof-based consensus algorithms.

A. ORIGINAL PROOF OF WORK

In a blockchain, when a new block is added, an agreement between the nodes is required. For this, the Proof of Work (PoW) algorithm requires that each node solves a puzzle to which the difficulty can be adjusted, so that the first node that solves the puzzle, will get the right to add a new block to the current chain. The effort made by the node for the solution of the puzzle, is called PoW and is payed to the node that calculated the hash right. This node is called a mining node or miner, and the action of solving the puzzle is called mining [8].

In the PoW, a search for the puzzle solution is made, such that when the hash is created, usually using the SHA-256 hash function, it must start with a number of zero bits. The average work required is exponential in relation to the number of required zero bits and can be verified by executing a simple hash operation [9]. In PoW the difficulty of the puzzle is adjusted every time that 2016 blocks are added, so that the average speed to add one new block in the chain is one (1) block every ten minutes [8].

When a new block is created, the header information is combined and sent as an input parameter to the SHA-256 hash function [10]. If the output of this function is below a threshold T (which depends on the difficulty), then the value sought is accepted. Otherwise, the node must continue calculating the secret value until the output of the SHA-256 function is accepted. The difficulty of the puzzle increases as the value T becomes smaller [8].

B. PoS-BASED

The Proof of Work algorithm is not fair for all miners, because not all have the same hardware. Some have modern equipment and other very basic equipment to process data and information which, given that solving the puzzle is very computing-intensive, the first ones will have an advantage. The algorithms based on Proof of Stake (PoS) seek to deal with this inequality. The basic principle of the PoS algorithms is to use the idea of a bet or participation magnitude, to define which node will have the opportunity to mine the next block in the chain. Using participation as evidence has an advantage: any node that has had a lot of previous participation is more reliable, and thus it is expected that this node will not perform any fraudulent activity to attack the chain that contains a large part of its profits. Also, the use of PoS implies that there has to be at least 51% of all bets in the network, to perform a double-cost attack, which is very difficult. There are currently two popular types of consensus that use PoS: those that use pure participation to obtain consensus and the hybrids that combine PoS and PoW [8].

C. HYBRID FORM OF PoW AND PoS

Sunny and Scott [11] proposed a new concept called the coin age of each miner, which is calculated by his bet multiplied by the time the miner owns it. For a node to get the right to add a new block to the chain, it creates a special block called coin stake, which contains many transactions, but also includes a special one from that miner to itself. The amount of money spent on the transaction gives the miner more possibilities to mine a new block, then solve the puzzle, as in PoW. The more money is spent on the transaction, the easier it is to solve the puzzle. When the puzzle is solved, the mining node gets 1% of the amount of the coins that they have spent in the transaction, but the accumulated coin age by these coins is reset to zero (0) [8].

Unlike the previous proposal, Vasin [12] does not use the coin age in his blockchain, because it is assumed that, by making use of the coin age, the attacker can be given the possibility to accumulate enough value to deceive the network. Another problem is the possible existence of some miners who keep their bet until they have a large number of coins, while they remain outside the verification system; therefore, the proposal by Vasin [12], is to use pure participation in exchange for the age of the currency to offer miners the possibility of mining a new block. This may encourage more nodes to be online to obtain profits. When the existence of offline miners is untied, Ren [13] proposes to use an exponential decay function with the coin age, in which, when the miner waits for the increase in the coin age, less is the speed of increase. References [14] propose a method that combines PoW and PoS, which will be explained in the next section as a way to also mitigate the 51% attacks.

D. OTHER KINDS OF PROOF-BASED CONSENSUS ALGORITHM

One of the main problems of PoW is the excessive energy demands required to find the nonce, besides the fact that this calculation is disposable and does not provide any long-term benefit to the users. This was presented by Blocki and Zhou in [15] and by Sunny [16]. To address this issue, Blocki and Zhou [15] proposed the use of some types of puzzles for education and social activities, which were easy to solve for computers but difficult for people to solve; thus, the effort to solve the puzzle to undermine a new block corresponds to people and not in using hardware. This is fairer for all because not all miners can invest in modern hardware [8].

Different authors have proposed other evidence-based consensus algorithms that do not use the idea of PoW and PoS. Examples of these are: Proof of Burn in [17], Proof of Space in [18], Proof-of-QoS (PoQ) in [19] and A fair selection protocol [20]. In Proof of Burn, the miners send their coins to a direction to be burned, in this way these coins cannot be used by others, so the miner who burns most coins, earns the right to mine a new block. On the other hand, the miners of Proof of Space must invest in hard disks for their computers, which in comparison with the hardware required in PoW is much cheaper. The Proof of Space algorithm generates large data sets called plots on the hard disk, so the more data a node has, the more likely it is that it will mine a new block. In PoQ the network is divided into small regions. Each of these chooses a node based on its QoS. Then, a deterministic Byzantine Fault Tolerance (BFT) consensus is run between all the chosen nodes. The goal of PoQ is to achieve very high transaction throughput as a permissionless protocol and to provide a fairer environment for participants. Fair selection protocol is composed of two main phases: the mining process and the confirmation of the new nodes list. More consensus techniques can be found in [21], where a new classification method was proposed.

III. TECHNIQUES FOR 51% ATTACK PREVENTION

In this section, we present some works focused on mitigating the 51% attack on blockchain networks.

The double-spending attack considers a high risk for the security of a blockchain, when the miners own more than 51% of the mining power, generating an alert. To mitigate this problem, [14] proposed a method combining PoW and PoS (2-hop Blockchain). The objective of this method is to make sure that, even if a miner owns more than 51% of the mining power, he will not have many possibilities to carry out a fraudulent action. To achieve this, the authors propose using a PoW first to choose a winning node, which is the first to solve the puzzle. Next, this node, in addition to adding a block called PoW Block to the chain, provides a basis for choosing another miner who has a bet. If the return value of the hash function that has input parameters of the newly added PoW Block and the private key of the owner of the bet is below a threshold, the chosen miner will have the possibility to add the PoS Block to the chain.

This research paper [22] proposes an encoder-decoder deep learning model to detect anomalies in the use of blockchain systems. The contributions of this work are the following: a) the identification of a relevant set of characteristics calculated in blockchain registers that describe the state of the network in certain time steps; and b) The use of a sequence-by-sequence neural network model to recognize anomalous changes in the blockchain network.

Due to the uniqueness of the attacks and how this makes them hard to identify and detect, [22] adopt an unsupervised approach to address this problem. Also, they propose as future work to study the use of hybrid architectures based on the combination of Recurrent Neural Networks (RNN) with convolutional neural networks [23] to perform the feature-selection process and assess eventual improvements. In addition, they consider defining models capable of predicting attacks before they occur, improving network security.

The paper by [24] proposes a method to mitigate 51% attack on proof-of-work blockchains based on weighted history information. In this approach, the authors use the frequency rate of miners in historical blocks and calculate the total weighted historical difficulty to establish if a branch change is required. The proposed protocol is called ''Proof of work based on historical weighted difficulty'' (HWD-PoW), and its analysis indicates that the cost of the attack increases by two orders of magnitude when the new technique is implemented.

In [25], Dey proposes a methodology in which intelligent software agents can be used to monitor stakeholder activity in blockchain networks to detect anomalies, such as collusion, by making use of a supervised machine learning algorithm and algorithmic game theory, to mitigate the majority or the 51% attack.

Horizon proposes a delayed block sending penalty system [26]. This proposal suggests modifying the Satoshi consensus protocol (PoW) to secure a network against 51% attack. The sanction applied is determined based on the time the attacking node is hidden from the network. This technique notifies the entire network about the fork, and during that period, participants, miners, and exchanges cannot transact until the delay period is removed. The penalty system is a research prototype technique that has not yet been implemented in a real network, and it also includes several limitations. According to Rosenfeld in [7], when an attacker owns 51% of the network's hash, he will always succeed regardless of the imposed delay. Consequently, the possibility of carrying out the 51% attack when this security mechanism is in place is very large. Also, the delay process slows down the general transactions of the network and strongly impacts the usual transactions because the delay blocks will not be confirmed until the penalty is lifted. This fact makes this technique not very appropriate to be implemented in a real network, and it is not completely effective against a 51% attack.

In [27], Chainzilla proposes a Kodomo security solution called ''delayed proof of work'' (dPoW). This solution is implemented for cryptocurrencies based on UTXO. This security technique is already implemented in some blockchains to safeguard against double-spend attacks. The main attribute of dPow is that it does not recognize the rule of the longest chain; consequently, attacks that are intended to be carried out in private cannot gain an advantage to double spend. For its operation, dPow chooses 64 special nodes each year to acquire information from Komodo and store it on the bitcoin blockchain. The strengthened security orientation of this proposal intends for the attacker to rewrite the Komodo chain and bitcoin checkpoints. Likewise, the attacker must also be able to influence the majority of the notary network. This makes the technique robust. However, the limited number of special nodes makes the security technique centralized, which leads to the known problem of a ''single point of failure'', where attackers know exactly what to attack. Another limitation that dPoW presents is that it is only Implementable in cryptocurrencies based on UTXO and it is not profitable since it requires an implementation fee. In addition, the participating nodes of the network must wait an explicit amount of time for the notarization process to be completed, which can discourage certain participants who intend to make a faster transaction. The notarization process is carried out every 10 minutes, which gives attackers a window of time to carry out the 51% attack in cryptocurrencies given that the block confirmation time only needs a few seconds [28].

PirlGuard is a security protocol developed to mitigate the 51% attack, this approach modifies the consensus algorithm to protect itself from a 51% attack [29]. This protocol is based on the attributes of the Horizen penalty protocol (system delay block send penalty), but it is built primarily for Ethash. When the network detects longer blocks extracted privately, PirlGuard abandons the node instantly by penalizing the extraction of x number of blocks, based on the total number of blocks extracted secretly. The PirlGuard approach employs notarial contracts that are controlled by master nodes; these master nodes are in charge of notarizing the blockchain and penalizing malicious nodes by regaining legitimate consensus on the Pirl blockchain. As in the Komodo solution, PirlGuard also employs master nodes, a feature that makes the security technique centralized, leading to the known problem of a ''single point of failure''. Another limitation of this solution derives from the fact that the penalty is not a final solution to protect against the 51% attack, as there is a probability that attackers with a hash rate of 51% will be able to overcome that penalty.

ChainLocks in [30] is another security technique developed to protect DASH, based on the implementation of ''long-living master node quorums'' (LLMQs) to mitigate the 51% attack. This technique includes a network-wide voting process that comprises a ''first-look'' policy. For each of the blocks, an LLMQ of a large number of master nodes is approved. All participating nodes in the network are required to sign the designated block to extend the active chain. While at least 60% of the network participants verify a block, they generate a P2P message (CLSIG) to notify all other nodes about the event. This CLSIG message cannot be generated unless enough members of the network comply, so it implies a valid signature of authenticity and verifiable by the nodes within the network. Because only one confirmation is required for the publication of a block, attackers with at least 51% hashing power have a chance to double-spend, making the Dash blockchain vulnerable. In addition, the main disadvantage of ChainLocks is that it is designed only for the Dash cryptocurrency, which has a low network hash. This feature, in addition to the master node approach, makes it a weak security approach allowing the possibility of a 51% attack simply by renting the required hashing power [31].

Merged Mining or merged mining is not a security technique, but a method that allows merging several cryptocurrencies with mining at the same time. Low hashing power cryptocurrencies that share the same consensus can benefit from merged mining to improve their security [32]. The merged mining process makes it possible to increase the hashing power by starting in the other coin that comprises a higher hashing power. Although cryptocurrencies take advantage of merged mining, transactions on both networks can run in sequence. Blockchains are classified as main and auxiliary. In addition to improving security, another benefit is the ability for miners to mine more than one block simultaneously. Although merged mining increases security on blockchains, the process is not straightforward and is very often neglected by miners. The main limitation of this method is that the cryptocurrencies that take advantage of this approach must be in the same consensus protocol and mining algorithms [33].

Another limitation is that, if two low-hashed cryptocurrencies are combined, it is possible to exploit them as long as the attacker achieves the required hashing power. Consequently, merged mining is only a process to increase the cost of the attack by merging hashing power, and does not provide an effective solution to the 51% attack.

In [34], Sayeed & Marco-Gisbert present a novel technique called "Proof of Adjourn" (PoAj), whose objective is to mitigate the main blockchain attacks and the problem of delay in the processing of transactions with large transactions in cryptocurrencies passed in UTXO. This proposal does not recognize the longest chain to verify the authenticity of the chain, instead, a deferral period is imposed to regulate the verification of the block, although miners with high hashing power could have an advantage in the mining process, the transmission of more than one block will disqualify the block from being included in the chain by abstaining from mining activities for some time. The security of this proposal lies in eliminating the possibility of block reversion. PoAj confirms transactions with just one confirmation eliminating the waiting time of six confirmations brought by PoW. This leads to a much faster transaction confirmation rate compared to many existing consensus protocols. Similarly, PoAj introduces a unique approach that is activated when there is more than one block transmitted within a predefined period of time. This approach is unique and, so far, it is the first approach to fully solve the problem. This proposal is not found in any cryptocurrency; the authors implemented a proof of concept of the PoAj consensus protocol in the Python programming language.

The table 1 presents the advantages, risks, vulnerabilities, implementation cost, and the working of the techniques for 51% attack prevention presented in the analyzed works.

All these techniques assume a wide decentralization of the nodes, but the reality of this assumption has not been evaluated in the literature and will be addressed in this work.

IV. METHODOLOGY

In this section, each of the phases of the methodology is detailed, at first, we explained how the data were obtained and selected, then we explain the process of pre-processing the data sets, and finally the process of characterization of the miners.

A. DATA SELECTION

In this study, the complete historical data generated by the nodes of the Bitcoin and Ethereum networks were acquired and processed.

On the one hand, the Bitcoin dataset is hosted at [35]. This data have been progressively collected through Web Scraping from the website [36] using a script written in Python, which is available hosted at [37]. The corresponding file has a current size of 243 MB, and it contains the data of each bitcoin block ranging from the Genesis block to the block number 682676 mined on May 9, 2021. The data stored in the file is organized into the following fields: `Hash`, `Confirmations`, `Timestamp`, `Height`, `Number of Transactions`, `Difficulty`, `Merkle root`, `Version`, `Bits`, `Weight`, `Size`, `Nonce`, `Transaction Volume`, `Block Reward`, `Fee Reward`, `Miner Name`, `Date`, `URL Miner`, `Year`, `Month` and `Day`.

On the other hand, the Crypto Ethereum dataset is hosted on the Google cloud platform. It was queried and downloaded by using the Google BigQuery tool from the `blocks` table of the `crypto_ethereum` dataset

(bigquery-public-data:crypto_ethereum.blocks) [38].

All data in this dataset was extracted, transformed and loaded through a set of Python scripts available on [39]. The dataset hosted on Google Cloud contains data on each Crypto Ethereum block ranging from the Genesis block to current date. This because it gets updated daily, and its size is of more than 13 Gb. It is organized into the following fields (columns): `Timestamp`, `number`, `hash`, `parent_hash`, `nonce`, `sha3_uncles`, `size`, `transactions_root`, `state_root`, `receipts_root`, `miner`, `logs_bloom`, `total_difficulty`, `difficulty`, `extra_data`, `gas_limit`, `gas_used`, `transaction_count` and `base_fee_per_gas`.

For the analysis, a subset of data was taken from the main dataset, which corresponds to the records from the Genesis block until April 27, 2021, when the block with identifier 12322990 was mined. This subset only contains the fields `timestamp`, `number`, `miner`, `difficulty` and `gas_used` fields. This subset of data was extracted using the following BigQuery command

```
SELECT
timestamp,number,miner,difficulty,gas_used
FROM
bigquery-public-data.crypto_ethereum.block
```

of which result was saved into a file with a size of 1.2 Gb. Both Bitcoin and crypto Ethereum data sets are saved in a flat comma-separated CSV file for later loading and preprocessing with a Python script.

B. PREPROCESSING

The next stage after selecting data was preprocessing each dataset. This task was carried out with Google Colaboratory [40], which is a cloud service offered by Google that offers a virtual machine environment to run Jupyter-based Notebooks. On one hand, on the preprocessing of the Bitcoin data, the columns other than Height and Miner Name were removed. Additionally, the columns `Date`, `Year`, and `Year_Month` were created from the `Timestamp` field. On the other hand, for the preprocessing of the crypto Ethereum data, the column `miner` was renamed as `Miner Name` and the column number by `Height`. As similar as in the bitcoin file, the columns `Date`, `Year` and `Year_Month` were created from the field `Timestamp`, and the remaining columns were removed.

TABLE 1. Main characteristics of the analyzed techniques for 51% attack prevention.

Technique	Advantage	Risks	Vulnerability Identification	Cost	Underlying Blockchain
2-hop Blockchain - [14]	Even if a malicious node manages to control more than 50 hash rate, honest nodes may still defend the blockchain through honest participation	higher centralization and resource consumption	May not identify the vulnerability in advance	Its implementation may involve a cost	TwinsCoin
Encoder-decoder deep learning mode - [22]	Use of a sequence-to-sequence neural network model to recognize anomalous changes in the Blockchain network	Wrong detection	Detects anomalous situations and trigger an alert	Its implementation may involve a cost	Any Blockchain with sequential data
Proof of work based on historical weighted difficulty (HWD-PoW) [24]	Increase the cost of the attack by two orders of magnitude	Not efective against a slow gradual increase of hash rate and may affect the privacy and security of miners	May not identify the vulnerability in advance	Its implementation may involve a cost	Blockchain based on Proof of Work (PoW)
Methodology with intelligent software agents - [25]	Implement an intelligent agent in the application layer of the blockchain network system	Wrongly deducts the level of importance of the product	May identify the vulnerability before it is adopted by the main chain	It is just a proposed methodology	Blockchain based on Proof of Work (PoW)
Delayed block sending penalty - [26]	It forces the attacker to privately mine a large number of consecutive blocks before being able to join the main chain, consequently carrying out the attack is much more expensive	Not enough to completely mitigate the attack, it is a very weak solution for a cryptocurrency with low hashing	May identify the vulnerability before it is adopted by the main chain	It does not require, the technique is a prototype	Blockchain based on Proof of Work (PoW)
Delayed proof of work (DPoW) - [27]	Ignore the longest chain rule, the use of notarial nodes adds security to the protocol	A malicious node with at least 51% hash rate could execute an attack within the notarial period	May not identify the vulnerability in advance	There is a cost to adopt this technique	Any blockchain based on Unspent transaction output (UTXO)
PirlGuard [29]	Employ a penalty system for attack nodes	Using master nodes for notarization weakens the network due to a single point of failure	Identify vulnerability as soon as it is recognized	Its implementation may involve a cost	Blockchain based on Ethash Algorithm
ChainLocks - [30]	Very fast transaction confirmation	Confirmation with a single block may enable double-spending with a low hash rate	Lock the first block as a legitimate block, discarding any other blocks	Require a cost for its implementation	Dash CryptoCurrency
Merged Mining - [32]	Merge the hashing power of two blockchains, making the attack more costly	A malicious node with a high hash rate may perform an attack	May not identify the vulnerability in advance	No cost for implementation	Auxiliary Proof of Work (PoW)
Proof of Adjourn - [34]	offer sufficient protection to the network regardless of the hash rate of the attackers. shorter confirmation time for very large transactions	Participants with ample hash rate may get an advantage through the mining process	May identify the vulnerability before it is adopted by the main chain	It is a proposal, a proof of concept was made	Any blockchain based on Unspent transaction output (UTXO)

C. MINING CHARACTERIZATION

In this section, we provide an analysis of mining behavior on two popular blockchains: Bitcoin and Crypto Ethereum. This analysis is based on real collected data that has been examined methodically and in detail using data sciences techniques.

V. RESULTS

In the information analysis process, different algorithms were designed and coded in the Python programming language in order to identify aspects such as the number of miners and number of blocks in different periods of time, the hash rate of miners, and the percentage of blocks mined consecutively. In addition, own algorithms and clustering algorithms were used to create profiles of the miners. On the other hand, the model presented by [7] was also coded in the Python programming language. All of the above was done for the bitcoin and crypto Ethereum dataset.

A. NUMBER OF MINERS

1) BITCOIN

A first analysis of the Bitcoin data indicates that 682676 blocks were mined by 73 different miners in the analyzed period. In particular, Fig. 1 shows the number of mined blocks per year, and Fig. 2 shows the number of miners per year in the analyzed period. In this first analysis, it can be seen how the Unknown miners, those with Miner Name 'Unknown', are treated as a single miner, which differs from reality. So we individualize the miners by creating a new field in the data set from concatenating the fields 'Miner Name' and 'URL Miner', this dataset is published in Mendeley Data, [41]. As a result of this individualization process on the miners, 198892 different miners were identified in the entire analyzed period in contrast to the 73 miners found first. Fig.3 shows the number of known miners and the number of unknown miners for each of the years in the analyzed period.

2) CRYPTO ETHEREUM

A similar analysis was carried out on the 'Crypto Ethereum' dataset. In particular, this analysis shows that 12322991 blocks were mined by 5430 different miners in the analyzed period. Fig. 4 shows the number of mined blocks, and Fig. 5 shows the number of miners per year in the same time interval.

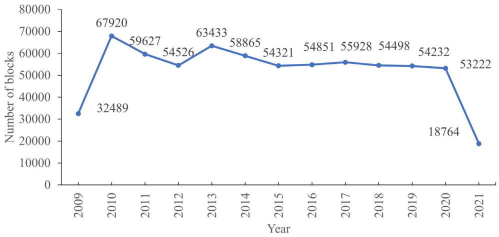

FIGURE 1. Number of bitcoin blocks mined by year.

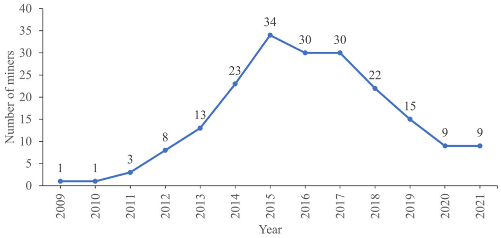

FIGURE 2. Number of unidentified unknown bitcoin miners by year.

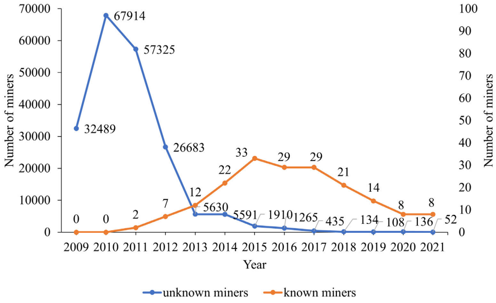

FIGURE 3. Number of bitcoin miners by year, after individualizing the unknown miners.

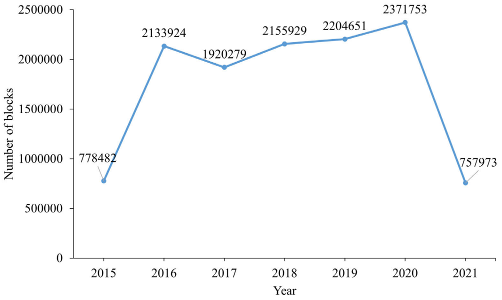

FIGURE 4. Number of crypto Ethereum blocks mined by year.

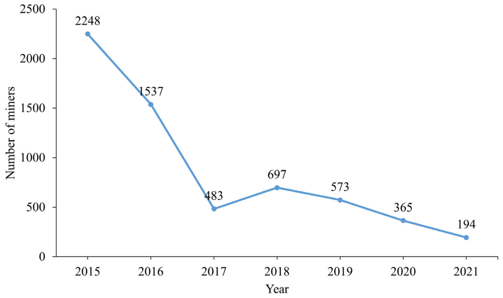

FIGURE 5. Number of crypto ethereum miners by year.

B. HASH RATE/SHARE OF MINERS

1) BITCOIN

For the 682676 bitcoin blocks mined in the analyzed period, the hash rate distribution is made for that same period. Fig. 6 shows this distribution in which the unknown miners are taken as one. Because the number of miners presented is very large, only the miners that have a hash rate greater than or equal to 0.1% in the entire period are shown, resulting in 29 miners that represent 99.49% of the total hash rate in the observed period.

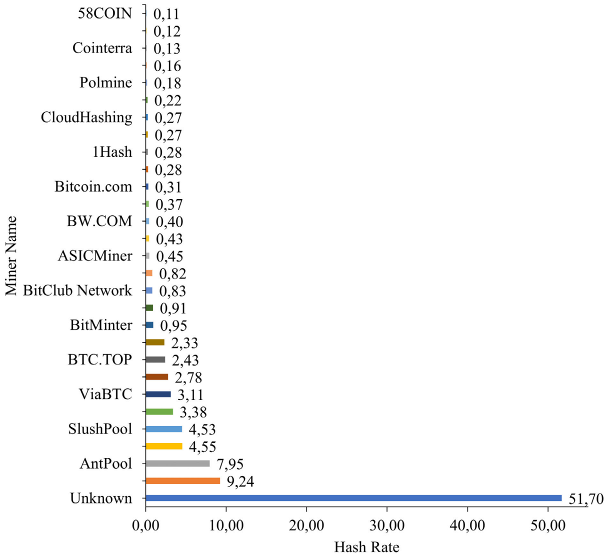

FIGURE 6. Bitcoin hash rate distribution throughout the period, miners with hash rate of at least 0.1%.

In the process of individualization of the unknown Bitcoin miners, it goes from having 73 miners to 198892 in the entire time. To analyze the hash rate distribution with all the miners already individualized in the period, this distribution is verified with the miners that present a minimum hash rate of 0.1%, 0.3%, 0.5%, 0.7% and 1% throughout the period. Table 2 presents these results.

TABLE 2. Hash rate distribution of bitcoin nodes throughout the observed period.

Participation hash rate throughout the time period	Number of miners in the period	Percentage of hash rate represented
less than 0.1%	198824	35.88%
at least 0.1%	68	64.13%
at least 0.3%	33	57.33%
at least 0.5%	22	53.37%
at least 0.7%	18	51.01%
at least 1%	11	45.22%

Based on the data in the table 2 and Fig. 7, it can be seen that only 11 miners that represent 0.0053064% of their total number of nodes have a hash rate of at least 1% and together represent 45.22% of the hash rate in throughout the period. It is also identified that, of 198824 miners that represent 99.965% of the total miners, present less than 0.1% of total hash rate in the entire period; in other words, only 68 miners, that represent 0.0341% of the total of miners have at least 0.1% hash rate in the entire period, and together represent 64.13% of the hash rate in the entire period analyzed. Fig. 7 presents the hash rate distribution for the entire period with the miners that have at least 1% total hash rate in the analyzed period.

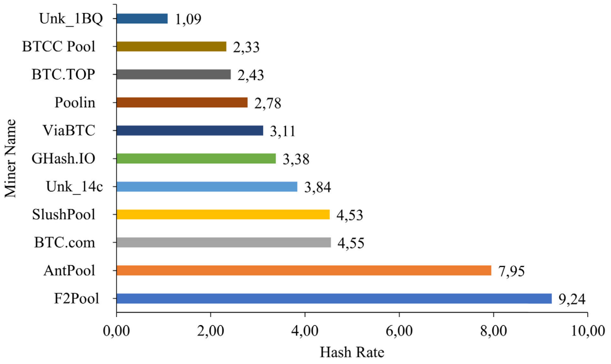

FIGURE 7. **Bitcoin hash rate distribution throughout the period with hash rate of at least 1% after individualizing the unknown miners.**

Fig. 7 shows that of the 11 miners that have at least 1% of the total hash rate, two of those are unknown and the remaining nine individually represent mining pools. Taking as a starting point these 11 miners who individually possess at least 1% of the total hash rate, the hash rate distribution of the entire period is plotted year by year in Fig. 8.

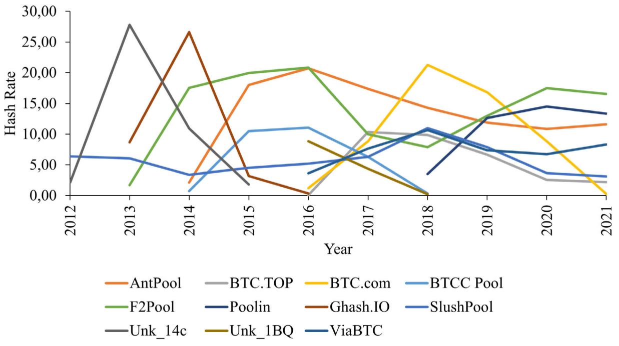

FIGURE 8. **Bitcoin hash rate distribution over time with hash rate of at least 1%, after individualizing the unknown miners.**

In Fig. 8 it can be seen that all the miners that have a hash rate of at least 1% in the entire period are present in the Bitcoin network since the year 2012, only one of them has remained active since 2012 to the present (May 2021), and 7 of the 11 miners in question are active in May 2021.

To verify the current status of the Bitcoin network in terms of the hash rate distribution, and taking into account the number of active miners since 2017, the hash rate distribution of the entire Bitcoin network is generated from January 1, 2017, until May 8, 2021 (Fig. 9).

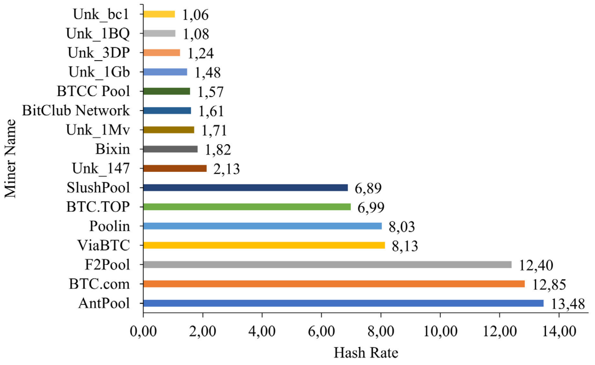

FIGURE 9. **Bitcoin hash rate distribution with hash rate of at least 1% after individualizing the unknown miners from January 1, 2017 until May 8, 2021.**

Fig. 9 indicates that for the last four years (2017-2020) in the bitcoin network only 16 miners that have a total hash rate of at least 1% for that period, represent 82.47% of the hash rate of the entire network. From these 16 miners, 10 have been identified and the remaining 6 correspond to unidentified miners. The Fig. 9 indicates that only 7 of the 11 miners exceed a hash rate of 5%, accumulating together 68.77% of the total hash rate.

Fig. 10 shows the hash rate distribution for these 7 miners in the period from January 1, 2017, to May 8, 2021, seen by years. In this figure, it can be seen that, of the 7 miners in question, 6 has been present on the network since 2017, and one of them since 2018.

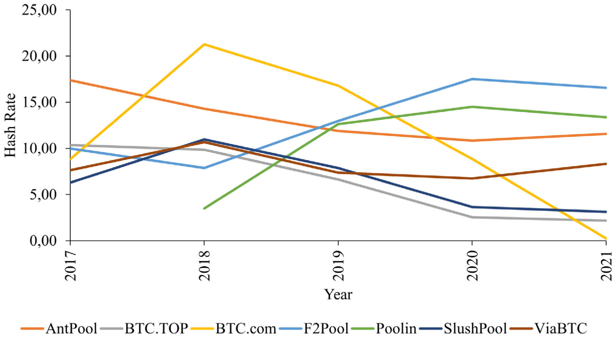

FIGURE 10. **Bitcoin hash rate distribution over time with hash rate of at least 5% after individualizing the unknown miners from January 1, 2017 until May 8, 2021.**

The hash rate distribution is generated throughout the period for the 7 miners that have presented the most hash rate in the last four years, and it is shown in Fig.11. The results show that, of the 7 miners, "SlushPool" is the only one that has been present on the network since 2012, 'F2Pool' since 2013, 'AntPool since 2014' and miners the 'BTC.com', 'BTC.TOP', 'ViaBTC' since 2016 and for their part 'Poolin' since 2018.

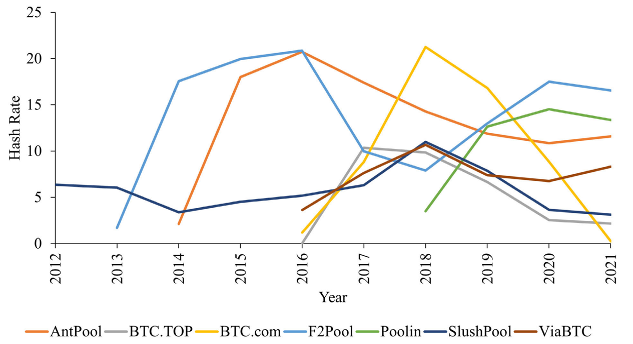

FIGURE 11. Bitcoin hash rate distribution of the top 7 miners over time after individualizing the unknown miners from January 1, 2009 until May 8, 2021.

2) CRYPTO ETHEREUM

For the 12322990 Crypto Ethereum blocks mined in the analyzed period, the share distribution is made for that same period. Fig. 12 shows this distribution. Because the number of miners presented is very large, only the miners that have a share greater than or equal to 1% in the entire period, resulting in 13 miners that represent 78.18% of the total share in the observed period, are analyzed.

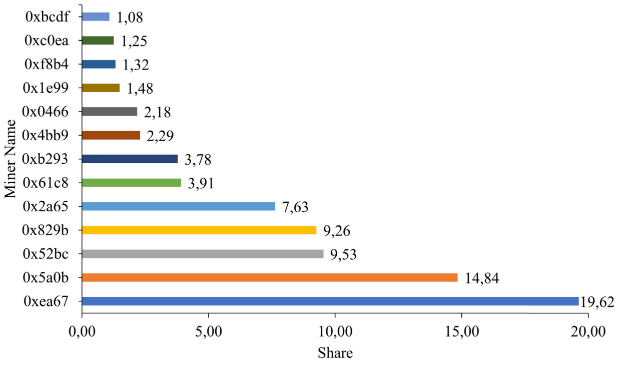

FIGURE 12. Crypto ethereum share distribution throughout the period with share of at least 1%.

To further analyze the share distribution of all the miners of the Crypto Ethereum network in the study period, this distribution is verified with the miners that present a minimum share of 0.1%, 0.3%, 0.5%, 0.7%, and 1% over the entire period, Table. 3 shows these results.

Based on the results in Table 3, it can be said that only 13 miners that represent 0.23% of their total number of nodes and have a total share of at least 1% (Fig. 12) represent 78.18% of the share in the entire period. It is also identified that 5366 miners, which represent 98.82% of the total miners, present less than 0.1% of the total share in the entire period; in other words, only 64 miners, that represent the 1.17% of all miners that have at least 0.1% share in the entire period, represent 93.90% of the share in the entire period analyzed.

To identify which miners are currently more important, the share of the miners that represent at least 1% share is generated in the period between January 1, 2019, and April 27, 2001, as shown in Fig. 13. The figure indicates that only 11 miners comply with having at least 1% share in the analyzed period and together present an 82.25% total share. Of these 11 miners, it is identified that 2 of these are not present in 2020 or 2021. Of the nine miners that have participated in 2021, 8 have been present since 2019, and one of them since 2020. For the 8 miners that are present in the years 2019, 2020, and 2021, their share is analyzed in the period between January 1, 2019, and April 27, 2001, Fig. 14. Each of these miners presents individually by at least 1% share and together they add up to 77.59% share.

TABLE 3. Share distributions of some crypto ethereum nodes throughout the period.

Participation share throughout the time period	Number of miners in the period	Percentage of share represented
less than 0.1%	5366	6.1%
at least 0.1%	64	93.90%
at least 0.3%	37	89.25%
at least 0.5%	19	82.21%
at least 0.7%	15	79.99%
at least 1%	13	78.18%

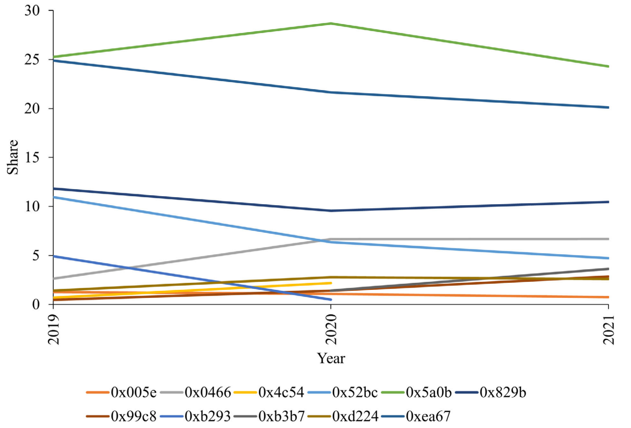

FIGURE 13. Crypto ethereum share distribution with share of at least 1% from January 1, 2019 until April 27, 2021.

C. PERCENTAGE OF BLOCKS MINED CONSECUTIVELY

For each of the data sets, an algorithm was applied to identify the percentage of blocks mined consecutively in the entire period of observed time, and for each of the years in the period analyzed.

1) BITCOIN

In the Bitcoin data set, if the probability of mining at least one block in a consecutive manner is 100%, the longest

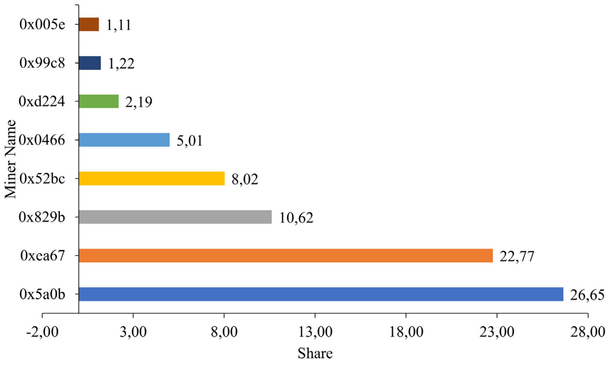

FIGURE 14. Crypto ethereum share distribution from January 1, 2019, to April 27, 2021.

mining chain was 11 blocks, and it had a calculated probability of 0.0001%, in the observed time. Fig. 15 shows the percentage of blocks mined consecutively in the whole period with chains of 2 blocks up to 11 consecutive blocks. Fig. 15 shows how the probability of mining consecutive blocks decreases as the number of blocks is greater; this trend is presented for each of the years. In this analysis, it can be seen that, for the first four years analyzed (2009-2012), the longest chain occurs in 2012 with 5 blocks and a mining probability of 0.0073%. From the year 2013 to the year 2021, mining chains greater than 5 blocks have been presented, with the year 2013 presenting chains of 11 blocks and the year 2014 of 10 blocks with probabilities of 0.0016% and 0.0017%, respectively. The year 2016 presents mining chains of up to 8 consecutive blocks with a probability of 0.0018%, In the years 2015, 2018, and 2019, there were mining chains of up to 7 blocks, and up to 6 blocks in the years 2017, 2020 and 2021, with average probabilities of 0.0030% and 0.0066% respectively.

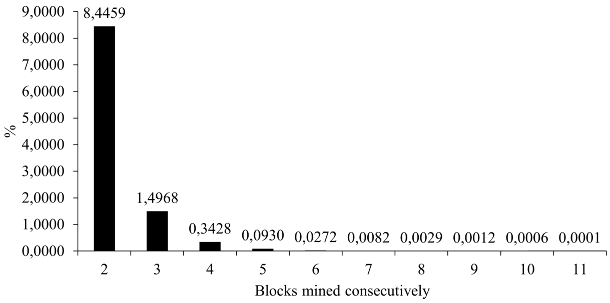

FIGURE 15. Percentage of bitcoin blocks mined consecutively.

2) CRYPTO ETHEREUM

For the Crypto Ethereum data, it is identified that the longest chain of mined blocks corresponds to 17 blocks, which occurs with a probability of 0.000008% throughout the period, and 0.000128% in 2015 that was when this event happened. Fig.16 shows the percentage of blocks mined consecutively in the whole period, with chains from 2 blocks to17 consecutive blocks. It can be seen that the probability of mining consecutive blocks decreases as the number of blocks is greater, behavior that is presented for the entire period analyzed as well as in each of the years of that period. It is identified in this analysis that, every years during the observed period, there are mining chains of 10 blocks with an average probability of 0.00126%. Likewise, for the years 2019 and 2020, there are mining chains of 11 blocks with a probability of 0.000044%, and for the year 2016 and 2018 chains of 12 and 14 blocks respectively which present 0.00018745% and 0.00004638% probability of mining.

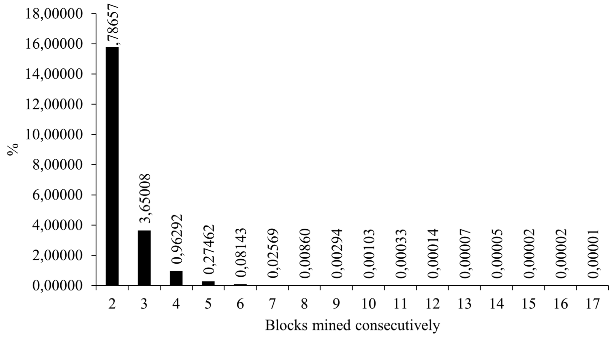

FIGURE 16. Percentage of crypto ethereum blocks mined consecutively.

D. PROFILE OF MINERS

1) BITCOIN

To create profiles of the current most representative miners of the Bitcoin network, the 10 miners that make a presence on the Bitcoin network in 2021 and have at least a 1% hash rate in the period from 1%, are considered from January 1, 2019, to May 8, 2021. They have a combined hash rate of 76.27% for that period and 73.2% for the period of January 1, 2017, to May 8, 2021, and of 36.13% in 2009-2021. In Fig. 17 the percentage of active days for each of the 10 miners in the 2019-2021 period is shown together with the hash rate of each of the miners throughout the period, years 2017 to 2021 and years 2019 to 2021. It is identified that the hash rate of each of the miners is higher in recent periods compared to the entire period of the presence of the Bitcoin network (2009-2021). Likewise, it can be seen that 7 of the 10 miners have an active presence ratio of more than 87% in the analyzed period. Only one of those miners has less than 50% presence in the analyzed period with 35.3%

The day-to-day hash rate of each of the miners is taken in the period from January 1, 2019, to March 8, 2021 [42], and it is represented in a vector of 859 positions corresponding to the each days in that period. The value of each position of the vector corresponds to the hash rate of the particular miner on each day. The clustering algorithms K-Means, DBScan, and Birch are applied to the set of hash rate vectors of the miners in order to identify a patterns among them. After applying these algorithms, 3 and 4 groups are generated. For

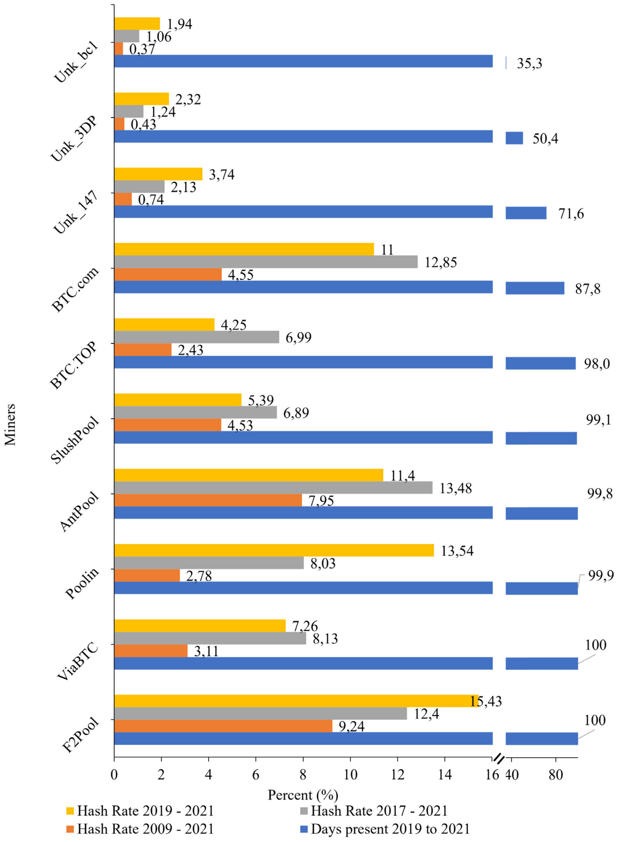

FIGURE 17. Best bitcoin miners in the period January 1, 2019 to May 8, 2021.

the case of 3 groups, the parameters for the algorithms are K-Means (k = 3), DBScan (eps = 1.65; min_sample = 1.0; metric = "euclidean") and Birch (branching_factor = 23, threshold = 1.05). In this case, the results of the application of the different algorithms generated the same 3 groups, as shown in Table 4.

TABLE 4. Groups formed for the best bitcoin miners with the K-Means, DBScan, and Birch algorithms to generate 3 groups.

Group 0		Group 1		Group 2	
Miner	% presence	Miner	% presence	Miner	% presence
AntPool	60.65	BTC.TOP	83.35	BTP.com	NA
		SlushPool	75.09		
F2Pool	66.94	ViaBTC	58.21		
		Unk_147	86.26		
Poolin	78.11	Unk_3DP	76.25		
		Unk_bc1	48.20		

To create 4 groups with the clustering algorithms, the parameters for the algorithms are K-Means (k = 4), DBScan (eps = 1.6; min_sample = 1.0; metric = "euclidean"), and Birch (branching_factor = 23, threshold = 1.25). In this case, the algorithms created different groups, as it can be seen in Table 5. The names of the unknown miners in tables 4 and 5 were truncated for better visualization. For each of the different groups formed by the clustering algorithms, it was verified that their elements (miners) are in the range of a mean +/− 1 standard deviation.

2) CRYPTO ETHEREUM

To create profiles of the most representative miners currently in the Crypto Ethereum network, the 8 miners that have an active presence in the years 2019 - 2021 and that individually have at least 1% share in the period, are taken from January 1, 2019, to April 27, 2021. They have a combined share of 77.59% for that period and 69.63% for the period of January 1, 2017, to April 27, 2021, and 57.45% between 2015-2021. Fig. 18 shows the percentage of days present for each of the 8 miners, in the 2019-2021 period, together with the share of each of the miners throughout the period, years 2017 to 2021 and years 2019 to 2021. It is identified that the share of each one of the miners is higher in recent periods compared to the entire period of the presence of the Crypto Ethereum network (2015-2021). Likewise, it can be seen that 7 of the 8 miners have an active presence of the 100% during the analyzed period and the remaining 99.4% presence in the analyzed period.

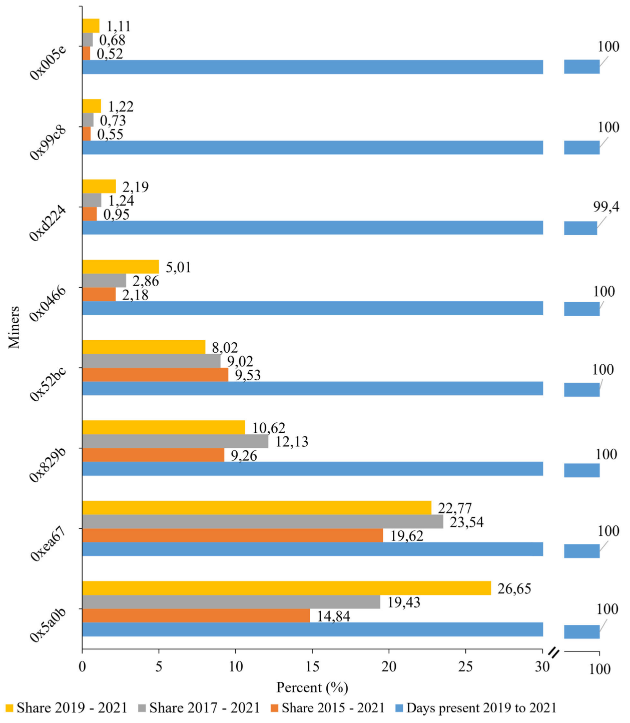

FIGURE 18. Best crypto ethereum miners in the period January 1, 2019 to April 27, 2021.

The day-to-day share of each of the miners is taken in the period from January 1, 2019, to April 27, 2021 [42], and it is represented by a vector of 859 positions corresponding to the days in that period. The clustering algorithms K-Means, DBScan, and Birch are applied to the set of hash rate vectors of the miners. After applying these algorithms, 3 and 4 groups were generated. The parameters used for the algorithms are

TABLE 5. Groups formed for the best bitcoin miners with the K-Means, DBScan, and Birch algorithms to generate 4 groups.

Kmeans		DbScan		Birch	
Group 0					
Miner	% presence	Miner	% presence	Miner	% presence
AntPool F2Pool Poolin	60.65 66.94 78.11	AntPool	NA	AntPool F2Pool Poolin	60.65 66.94 78.11
Group 1					
Miner	% presence	Miner	% presence	Miner	% presence
Unk_147 Unk_3DP Unk_bc1	76.37 88.71 58.09	BTC.TOP SlushPool ViaBTC Unk_147 Unk_3DP Unk_bc1	83.35 75.09 58.21 86.26 76.25 48.20	BTC.TOP SlushPool ViaBTC Unk_147 Unk_3DP	74.97 76.72 57.74 65.42 50.99
Group 2					
Miner	% presence	Miner	% presence	Miner	% presence
BTP.com	NA	BTP.com	NA	BTP.com	NA
Group 3					
Miner	% presence	Miner	% presence	Miner	% presence
BTC.TOP SlushPool	75.90 80.21	F2Pool	NA	Unk_bc1	NA

K-Means (k = 3), DBScan (eps = 1.9; min_sample = 1; metric = ''euclidean'') and Birch (branching_factor = 2, threshold = 1.05) and * (branching_factor = 2, threshold = 1.15). The final grouping is shown in Table 6. For the case of 3 groups, Birch generated two different results options.

For the generation of 4 groups with the clustering algorithms, the parameters for the algorithms are K-Means (k = 4), DBScan (eps = 1; min_sample = 1), and Birch (branching_factor = 2, threshold = 1). In this case, the results of the 3 different algorithms presented the same results, which are presented in Table 7. The names of the miners in tables 6 and 7 were truncated for better visualization. For each of the different groups formed with the clustering algorithms, it was verified that their elements (miners) are in the range of mean +/− 1 standard deviation.

E. ANALYSIS OF DOUBLE-SPENDING

The research article by [7] presents an analytical solution to model the probability of a double-spending attack via a stochastic process. A double-spending attack happens when an attacker persuades a seller that a transaction has been confirmed and subsequently convinces the entire network to accept other transactions that make the first transaction invalid. If such an attack occurs, then the merchant is left without the product and the payment, and thus the attacker keeps the product and the value of the payment.

Recall that a transaction included in a block within the valid chain has n confirmations if there are n blocks that follow the block containing the transaction. The model proposed in [7] assumes that there is a block B_r within a branch known to the honest miners (normally the longest branch) and that such block B_r contains the transaction T_r that credits the payment to the seller and has n confirmations. To perform the attack, the attacker has to construct a branch with additional m blocks starting from the block to which the block B_r points. Both the honest miners and the attacker are in the task of extending their respective branches. This model for the double-spending attack is inspired by a catching up game, in which the attacker's goal is to make its branch longer than the valid chain.

This model also supposes that the hash rate of the honest network and the attacker is constant. Specifically, if the complete hash rate is H, then $p \cdot H$ is the portion that corresponds to the honest miners, and $q \cdot H$ is the remaining portion that corresponds to the attacker, where $p+q = 1$. Also, it supposes that the mining difficulty is unchanging for the hash rate H, and that the average time to mine a block is T_0.

Let a_r be the probability that the attacker will be able to catch up when he is currently r blocks behind. Following the analysis in [7], $a_r = min(q/p, 1)^{\max(z+1,0)}$. Additionally, m is regarded as a negative binomial variable; it represents the number of successes (blocks mined by the attacker) before n failures (blocks mined by the honest network), where q is the success probability. Therefore, the probability for a particular value of m is given by:

$$P(m) = \binom{m+n-1}{m} p^n q^m \tag{1}$$

Combining both probabilities, we can compute the probability of successfully carrying out a double spending attack, r, as $r := \sum_{m=0}^{\infty} P(m) a_{n-m-1}$. From this equation, it follows that

$$r = \begin{cases} 1 - \sum_{m=0}^{n} \binom{m+n-1}{m} (p^n q^m - p^m q^n) & if\ q < p \\ 1 & if\ q \geqslant p \end{cases} \tag{2}$$

Following these results, an analysis of the success probability of a double-spending attack for the historical data of Bitcoin and Ethereum is presented in this work. For both Bitcoin and Ethereum, the mining data for the year 2020 is taken from the best miners presented in Fig. 17 and Fig. 18. For each of the analyzed blockchain networks, 3 miners with different hash rate/share values are selected. The results are shown in Table 8.

For these miners, the success probability of a double-spending attack, r, is calculated for each month of the

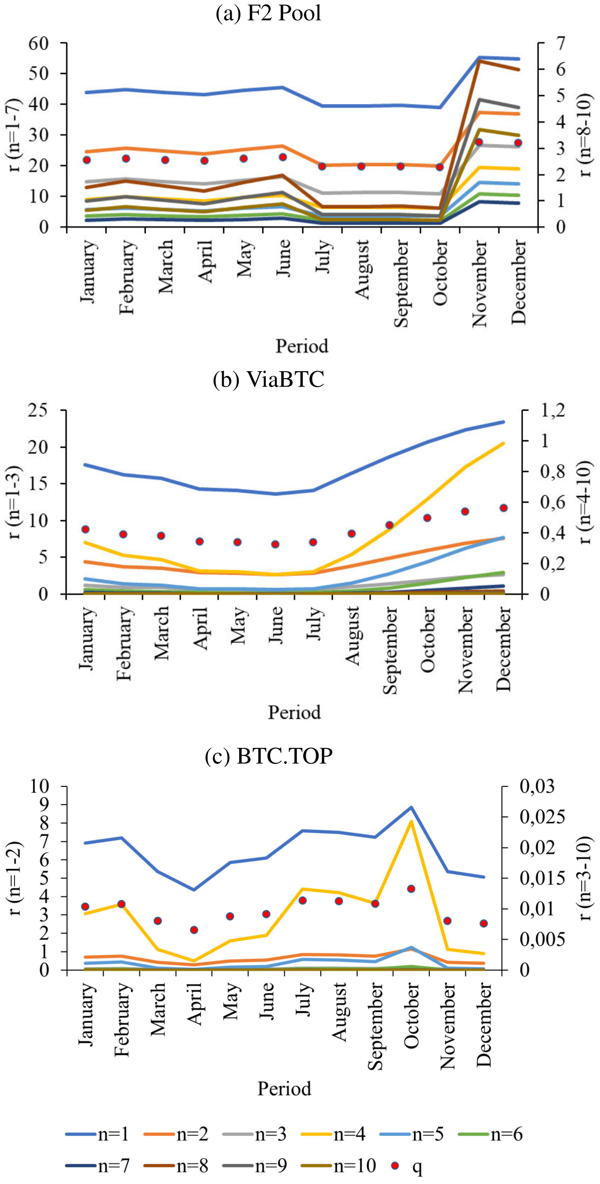

FIGURE 19. **Probability of success of a double spending attack on bitcoin miners.**

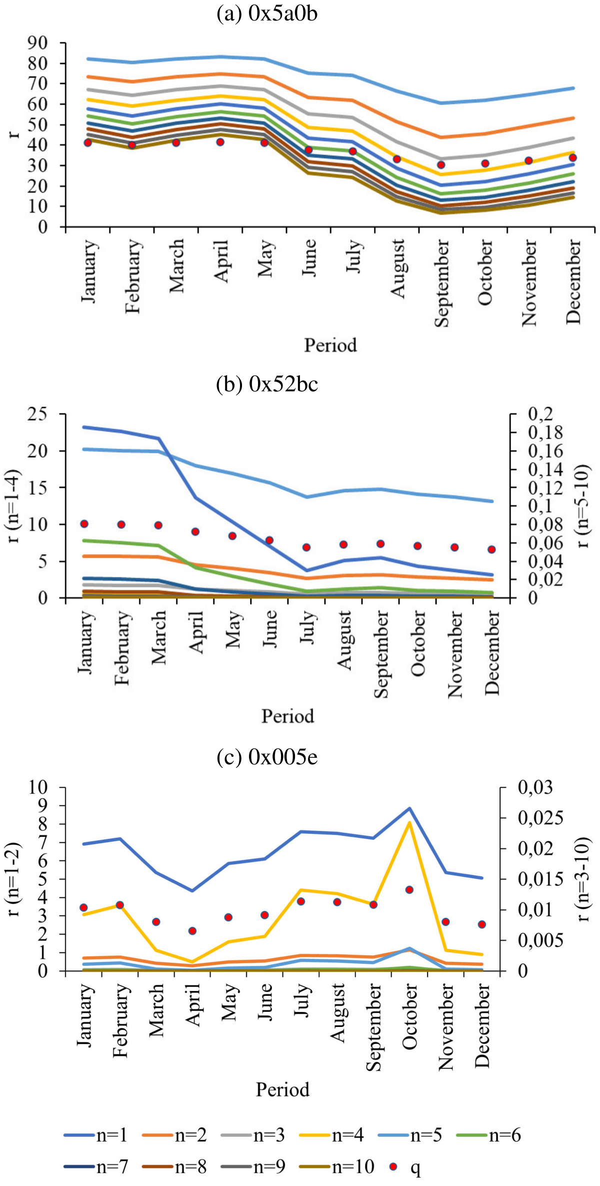

FIGURE 20. **Probability of success of a double spending attack on crypto ethereum miners.**

year 2020. For a miner, r is computed for each $q \in \{q_1, q_2, \ldots, q_{12}\}$ and $n \in \{1, 2, \ldots, 10\}$, where q_i denotes the hash/share ratio per month for the miner and n the number of confirmations. Particularly, q_i is calculated by computing the number of blocks mined by the miner in the month i, divided by the number of blocks mined in the month i. Fig. 19 and Fig. 20 show the result of this calculation for the Bitcoin and Crypto Ethereum networks, respectively.

In the case of the Bitcoin network, it can be seen that, as the hash rate for a miner grows higher and n assumes the lowest values, the success probability of carrying out a double-spending attack also grows higher, as expected. In particular, the highest success probability for carrying out a double-spending attack was registered for the miner F2Pool, since it has the highest hash rate and $n = 1$, as shown by Fig. 19 (a). This probability was 0.55 in November 2020.

For the other cases, for all miners, it is observed that the probability of performing a double-spending attack is less than 0.4. In fact, most of the computed probabilities are less than 0.2. In Fig. 19 (c) it is observed that, in the specific case of the BTC.TOP Miner, which presents a lower hash rate, the probabilities to carry out the attack are very low, being 0.0886 for one confirmation, 0.0114 for two confirmations, and less than 0.0017 for $n \geq 3$.

For the Crypto Ethereum network, the general trend for the success probability of carrying out the attack is similar to that of the Bitcoin network. In particular, the success probability is directly proportional to the share of the miner and inversely

TABLE 6. Groups formed for the best crypto ethereum miners with the K means, DBscan, and Birch algorithms together with the percent of the presence of their elements in the range "mean + / − one standard deviation" for 3 groups.

Kmeans / Birch		DbScan		Birch	
Group 0					
Miner	% presence	Miner	% presence	Miner	% presence
0x0466 0xd224 0x99c8 0x005e	18.04 94.92 77.24 92.09	0x5a0b	NA	0xd224 0x99c8 0x005e	58.72 55.66 85.61
Group 1					
Miner	% presence	Miner	% presence	Miner	% presence
0x5a0b 0xea67	100 100	0xea67	NA	0x5a0b 0xea67	100 100
Group 2					
Miner	% presence	Miner	% presence	Miner	% presence
0x829b 0x52bc	100 100	0x829b 0x52bc 0x0466 0xd224 0x99c8 0x005e	2.83 65.56 88.67 99.64 85.84 61.32	0x829b 0x52bc 0x0466	73.34 77 50

TABLE 7. Groups formed for the best crypto ethereum miners with the Kmeans, DBscan, and Birch algorithms together with the percent of the presence of their elements in the range "mean +/− one standard deviation" for four groups.

kmeans / birch							
Group 0		Group 1		Group 2		Group 3	
Miner	% presence	Miner	% presence	Miner	% presence	Miner	% presence
0x5a0b	NA	0xea67	NA	0x829b 0x52bc	100 100	0x0466 0xd224 0x99c8 0x005e	18.04 94.92 77.24 92.09

TABLE 8. Comparison of the probability of success of a double-spend attack based and real data from bitcoin and crypto ethereum.

Bitcoin			
Miner	Max hash rate	Min hash rate	Avg hash rate
F2Pool	27.67	19.53	22.24
ViaBTC	11.71	6.81	8.63
BTC.TOP	4.43	2.18	3.22
Ethereum			
Miner	Max share	Min share	Avg share
0x5a0b	41.57	30.20	36.68
0x52bc	10.09	6.6	8.12
0x005e	1.85	0.84	1.38

proportional to the number of confirmations. It should be noted that for the miner 0 × 5a0b Fig. 20 (a), which has the greatest hash power, the probabilities of performing the double-spending attack are above 0.65, sometimes reaching values above 0.7.

For the miner 0 × 52bc Fig. 20 (b), it was found that it had success probabilities of performing the attack greater than 0.01 for $n \leq 2$. However, for the other cases, these probabilities are less than 0.01, reaching down to the order of 0.1×10^{-7}. In the case of the miner with the lowest share power, such as 0 × 005e Fig. 20 (c), the success probability of a double-spending attack is greater than 0.01 only for $n = 1$. For $n \geq 2$, on the other hand, the probabilities are less than 0.002, reaching down to the order of 6.2×10^{-16}.

VI. CONCLUSION AND FUTURE WORK

The decrease in the number of miners along with the centralization of the hash rate/share are threats to the security of these blockchains (Bitcoin and Ethereum), therefore understanding the behavior of the miners becomes a relevant research topic. In this paper, we further analyze the behavior of the miners by following a different approach and present a miner behavior characterization for both the bitcoin blockchain and crypto Ethereum blockchain.

We conclude that according to the conducted analysis, the centralization of the hash rate seems to be a real threat. On the one hand, for bitcoin, only 18 miners representing 0.00905014% of the total number of miners have 51.01% of the hash rate in the entire period. In other words, 99.9658% of the total miners only reach 35.88% of the total hash rate. On the other hand, for the crypto Ethereum network, only 13 miners representing 0.23% of the total number of miners collectively achieve 78.18% of the share, i.e. 98.82% of miners collectively achieve 6.10% of the total share.

In both scenarios, there is a real possibility that a 51% attack could take place if the most powerful miners get together, which violates the main general assumption of the blockchains. Now, there is a real negative incentive to perform such an attack because the credibility of the blockchain network will go to zero, as well as the value of the crypto assets; thus, a self-protection policy takes place in these public networks by its members. However, there are plenty of other blockchains, public and private, in which the value loss maybe not be too critical to discourage such attacks to obtain a specific asset. This centralization of hash rate/share and its risks imply that all new prevention mechanisms to address this kind of vulnerability must take this new situation seriously in their considerations.

As future work, based on the conclusion of this work, a tool to identify anomalous behavior can be implemented, to detect a possible attack being performed by a miner, or group of miners, and generate a general alert to protect the integrity of the blockchain. Another way to address this issue may be to design a new consensus protocol that will prevent the possibility of overpowering the blockchain by the miners with the highest hash rate while distributing the hash rate and preventing its centralization in a small set of members.

REFERENCES

[1] E. Piscini, J. Guastella, A. Rozman, and T. Nassim. (2016). *Innovating in the Digital Era*. [Online]. Available: https://bit.ly/3toYpqE

[2] D. Wessel and P. Olson. (2016). *The Hutchins Center Explains: How Blockchain Could Change the Financial System (Part 1) | Brookings Institution*. [Online]. Available: https://www.brookings.edu/blog/up-front/2016/01/11/the-hutchins-center-explains-how-blockchain-could-change-the-financial-system-part-1/

[3] N. Anita and M. Vijayalakshmi, "Blockchain security attack: A brief survey," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 6–11.

[4] Dean. (2015). *51% Attack*. Accessed: Jul. 14, 2019. [Online]. Available: http://cryptorials.io/glossary/51-attack/

[5] X. Li, P. Jiang, T. Chen, X. Luo, and Q. Wen, "A survey on the security of blockchain systems," *Future Gener. Comput. Syst.*, vol. 107, pp. 841–853, Jun. 2020.

[6] N. Hajdarbegovic. (2014). *Bitcoin Miners Ditch Ghash.io Pool Over Fears of 51% Attack*. Accessed: May 2, 2021. [Online]. Available: https://www.coindesk.com/bitcoin-miners-ditch-ghash-io-pool-51-attack

[7] M. Rosenfeld, "Analysis of hashrate-based double spending," 2014, *arXiv:1402.2009*. [Online]. Available: http://arxiv.org/abs/1402.2009

[8] G. T. Nguyen and K. Kim, "A survey about consensus algorithms used in blockchain," *J. Inf. Process. Syst.*, vol. 14, no. 1, pp. 101–128, 2018.

[9] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *J. Gen. Philosophy Sci.*, vol. 39, no. 1, pp. 53–67, 2008. [Online]. Available: http://www.bitcoin.org

[10] Bitcoinwiki. (2016). *SHA-256*. Accessed: Jan. 20, 2021. [Online]. Available: https://en.bitcoin.it/wiki/SHA-256

[11] K. Sunny and N. Scott, "Ppcoin: Peer-to-peer crypto-currency with proof-of-stake," Stichting Peercoin Found., The Hague, The Netherlands, White Paper, 2012.

[12] P. Vasin. (2014). *BlackCoin's Proof-of-Stake Protocol V2 Pavel*. [Online]. Available: https://blackcoin.co/blackcoinpos-protocol-v2-whitepaper.pdf

[13] L. Ren. (2014). *Proof of Stake Velocity: Building the Social Currency of the Digital Age*. Accessed: Jan. 20, 2021. [Online]. Available: https://www.cryptoground.com/storage/files/1528454215-cannacoin.pdf

[14] T. Duong, L. Fan, and H.-S. Zhou, "2-hop blockchain: Combining proof-of-work and proof-of-stake securely," in *Proc. 25th Eur. Symp. Res. Comput. Secur. (ESORICS)*, Guildford, U.K., Sep. 2020, pp. 697–712. [Online]. Available: https://eprint.iacr.org/2016/716.pdf

[15] J. Blocki and H.-S. Zhou, "Designing proof of human-work puzzles for cryptocurrency and beyond," in *Theory of Cryptography*. Berlin, Germany: Springer, 2016, pp. 517–546.

[16] K. Sunny. (2013). *PrimeCoin: Cryptocurrency With Prime Number Proof-of-Work*. [Online]. Available: http://primecoin.io/bin/primecoin-paper.pdf

[17] P4Titan. (2014). *Slimcoin. A Peer-to-Peer Crypto-Currency With Proof-of-Burn*. Mining Without Powerful Hardware. Accessed: Jan. 20, 2021. [Online]. Available: http://www.slimcoin.org

[18] S. Park, A. Kwon, G. Fuchsbauer, P. Gaži, J. Alwen, and K. Pietrzak, "SpaceMint: A cryptocurrency based on proofs of space," in *Financial Cryptography and Data Security*. Berlin, Germany: Springer, 2018, pp. 480–499.

[19] B. Yu, J. Liu, S. Nepal, J. Yu, and P. Rimba, "Proof-of-QoS: QoS based blockchain consensus protocol," *Comput. Secur.*, vol. 87, Nov. 2019, Art. no. 101580.

[20] Y. Liu, J. Liu, Z. Zhang, and H. Yu, "A fair selection protocol for committee-based permissionless blockchains," *Comput. Secur.*, vol. 91, Apr. 2020, Art. no. 101718.

[21] F. Aponte, L. Gutierrez, M. Pineda, I. Merino, A. Salazar, and P. Wightman, "Cluster-based classification of blockchain consensus algorithms," *IEEE Latin Amer. Trans.*, vol. 19, no. 4, pp. 688–696, Apr. 2021.

[22] F. Scicchitano, A. Liguori, M. Guarascio, E. Ritacco, and G. Manco, "A deep learning approach for detecting security attacks on blockchain," in *Proc. CEUR Workshop*, vol. 2597, 2020, pp. 212–222.

[23] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*. [Online]. Available: https://arxiv.org/abs/1511.08458

[24] X. Yang, Y. Chen, and X. Chen, "Effective scheme against 51% attack on proof-of-work blockchain with history weighted information," in *Proc. IEEE Int. Conf. Blockchain (Blockchain)*, Jul. 2019, pp. 261–265.

[25] S. Dey, "Securing majority-attack in blockchain using machine learning and algorithmic game theory: A proof of work," 2018, *arXiv:1806.05477*. [Online]. Available: http://arxiv.org/abs/1806.05477

[26] A. Garoffolo, P. Stabilini, R. Vigliano, and U. Stav. (2018). *Proposal to Modify Satoshi Consensus to Enhance Protection Against 51% Attack*. [Online]. Available: https://www.horizen.global/assets/files/A-Penalty-System-for-Delayed-Block-Submission-by-Horizen.pdf

[27] ChainZilla. (Jan. 2019). *Blockchain Security and How to Mitigate*. Accessed: Feb. 17, 2021. [Online]. Available: https://medium.com/chainzilla/solutions-to-51-attacks-and-double-spending-71526be4bb86

[28] Komodo. (2018). *Komodo: An Advanced Blockchain Technology, Focused on Freedom*. Accessed: Jun. 18, 2021. [Online]. Available: https://cryptorating.eu/whitepapers/Komodo/2018-02-14-Komodo-White-Paper-Full.pdf

[29] R. Minchev. (2018). *PirlGuard—Innovative Solution Against 51% Attacks*. [Online]. Available: https://medium.com/pirl/pirlguard-innovative-solution-against-51-attacks-87dd45aa1109

[30] A. Block. (Nov. 2018). *Mitigating 51% Attacks With LLMQ-Based Chainlocks | by Alexander Block | Dash Blog*. [Online]. Available: https://blog.dash.org/mitigating-51-attacks-with-llmq-based-chainlocks-7266aa648ec9

[31] E. Ng. (2018). *Dash to Mitigate 51% Attacks With Chainlocks*. Accessed: Jun. 20, 2021. [Online]. Available: https://blockchainreporter.net/dash-to-mitigate-51-attacks-with-ChainLocks/

[32] Cryptocompare.com. (2015). *What is Merged Mining—Bitcoin & Namecoin—Litecoin & Dogecoin?* [Online]. Available: https://www.cryptocompare.com/mining/guides/what-is-merged-mining-bitcoin-namecoin-litecoin-dogecoin/

[33] BiXBiT. (2018). *Merged Mining—Collective Benefit and a Panacea for 51% Attack?* Accessed: Jun. 20, 2021. [Online]. Available: https://medium.com/bixbit.official/merged-mining-collective-benefit-and-a-panacea-for-51-attack-373404106a9

[34] S. Sayeed and H. Marco-Gisbert, "Proof of adjourn (PoAj): A novel approach to mitigate blockchain attacks," *Appl. Sci.*, vol. 10, no. 18, p. 6607, Sep. 2020.

[35] J. Ventrone. (2021). *Bitcoin Blockchain Data | Kaggle*. Accessed: May 9, 2021. [Online]. Available: https://bit.ly/3DVOr5b

[36] Blockchain.com. (2021). *Blockchain Explorer—Search the Blockchain | BTC | ETH | BCH*. Accessed: Apr. 9, 2021. [Online]. Available: https://www.blockchain.com/explorer

[37] J. Ventrone. (2021). *GitHub-JuanVentrone/Bitcoin_Blockchain_Scrapper*. Accessed: May 9, 2021. [Online]. Available: https://bit.ly/2Vtch72

[38] Google Cloud Platform. (2021). *BigQuery—Google Cloud Platform—Crypto Ethereum Dataset*. Accessed: May 9, 2021. [Online]. Available: https://bit.ly/2Yl61iI

[39] B. ETL. (2021). *GitHub—Blockchain-Etl/Ethereum-Etl*. Accessed: May 9, 2021. [Online]. Available: https://github.com/blockchain-etl/ethereum-etl#readme

[40] Colab.research.google.com. (2021). *Colaboratory*. Accessed: May 9, 2021. [Online]. Available: https://colab.research.google.com/

[41] F. Aponte, R. Villanueva, and P. Wightman, "Bitcoin miners with individualized unknown miners, mendeley data, v1," Datasets, 2021, doi: 10.17632/25gx6pbc6s.1.

[42] F. Aponte, R. Villanueva, and P. Wightman, "Daily computing power bitcoin and crypto ethereum, mendeley data, v1," Datasets, 2021, doi: 10.17632/cfw9d9cvrj.1.

FREDY ANDRES APONTE-NOVOA received the B.Sc. degree in computer and systems engineering from the Universidad Pedagógica y Tecnológica y Colombia, Colombia, in 2006, and the M.Eng. degree in free software from the Universidad Autónoma de Bucaramanga, Colombia, in 2011. He is currently pursuing the Ph.D. degree in computer and systems engineering with the Universidad del Norte, Colombia. His research interests include blockchain and virtual learning environments.

ANA LUCILA SANDOVAL OROZCO received the B.Sc. degree in systems engineering from the Universidad Autónoma del Caribe, Colombia, in 2001, and the M.Sc. and Ph.D. degrees in computer science from the Universidad Complutense de Madrid, Spain, in 2009 and 2014, respectively. She received a Specialization Title in computer networks from the Universidad del Norte, Colombia, in 2006. She is currently a Postdoctoral Researcher with the Universidad Complutense de Madrid. Her main research interests include coding theory, information security, and its applications.

RICARDO VILLANUEVA-POLANCO received the B.Eng. degree in computer science and engineering from the Universidad del Norte, in 2008, the M.Eng. degree in computer science and engineering from the Universidad de Los Andes, in 2010, and the Ph.D. degree in information security from the Royal Holloway, University of London, in 2018. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Universidad del Norte, Barranquilla, Colombia. His research interests include cyber-security, in particular, reliability of distributed systems, such as peer-to-peer systems and cloud systems, and applied cryptography, such as post-quantum cryptography, cryptographic protocols, and side-channel attacks on cryptographic implementations.

PEDRO WIGHTMAN (Senior Member, IEEE) received the B.Sc. degree in systems engineering from the Universidad del Norte and the Ph.D. degree in computer science and engineering from the University of South Florida. He is currently a Principal Professor with the School of Engineering, Science and Technology, Universidad del Rosario. He is the author of three technical books and several publications in indexed journals and international events. His research interests include location data privacy, blockchain, and medical information systems, communication infrastructure for the Internet of Things, and industry 4.0.

• • •