IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 26, 2019, accepted December 16, 2019, date of publication December 19, 2019,
date of current version December 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2961065

New Mathematical Model to Analyze Security of Sharding-Based Blockchain Protocols

ABDELATIF HAFID[1], (Member, IEEE), ABDELHAKIM SENHAJI HAFID[2], AND MUSTAPHA SAMIH[1]

[1]Team of EDA—Mathematical Laboratory and Their Applications, Department of Mathematics, Faculty of Sciences, University of Moulay Ismail, Meknes 50050, Morocco
[2]Montreal Blockchain Lab, Department of Computer Science and Operational Research, University of Montreal, Montreal, QC H3T 1J4, Canada

Corresponding author: Abdelatif Hafid (a.hafid@edu.umi.ac.ma)

This work was supported by the Montreal Blockchain Lab.

ABSTRACT In recent years, the scalability issue of blockchain protocols has received huge attention. Sharding is one of the most promising solutions to scale blockchain. The basic idea behind sharding is to divide the blockchain network into multiple committees where each committee processes a separate set of transactions. In this paper, we propose a mathematical model to analyze the security of sharding-based blockchain protocols. Moreover, we analyze well-known sharding protocols including RapidChain, OmniLedger, and Zilliga to validate our model. The key contribution of our paper is to bound the failure probability for one committee and so for each epoch using probability bounds for sums of upper-bounded hypergeometric and binomial distributions. In addition, this paper contribution answers the following fundamental question: ''how to keep the failure probability, for a given sharding protocol, smaller than a predefined threshold?''. Three probability bounds are used: Chebyshev, Hoeffding, and Chvátal. To illustrate the effectiveness of our proposed model, we conduct a numerical and comparative analysis of the proposed bounds.

INDEX TERMS Blockchain, failure probability, hypergeometric distribution, probability bounds, sharding.

I. INTRODUCTION

Blockchain is a technology that, when used, can have a great impact in almost all industry segments including banking, healthcare, supply chain, and government sector [9], [13]. It can be simply defined as a distributed digital ledger that keeps track of all the transactions (e.g., asset transfer, storage) that have taken place in a secure, chronological and immutable way using peer-to-peer networking technology. It does not rely on any trusted central entity (e.g., bank) to validate transactions and extends the blockchain; the network nodes (aka miners), using a consensus protocol, agree on which node can create (i.e., mine) a valid block and append it to the blockchain. For example, when Proof-of-Work consensus protocol [5] is used, the node that first solves a mathematical puzzle, adds the block to the blockchain and gets rewarded (by the network and transaction fees). More specifically, a transaction is broadcasted to all the nodes in the network (9459 nodes in the case of Bitcoin [32] and 8083 for Ethereum [33]); upon receipt of the transaction, a node that receives the transaction, it checks whether the transaction is valid. If the response is yes, it sends the transaction to its neighbors; otherwise, it drops the transaction. Periodically (e.g., each 10 minutes in Bitcoin [5]), a block (includes a list of transactions; e.g., up to 4000 transactions in Bitcoin) is created/mined; the node that mined the block (first to solve the mathematical puzzle), appends the block to its copy of the blockchain and broadcasts it to its neighbors. A node that receives a block, it validates the block. If valid, it appends the block to its copy of the blockchain and broadcasts to its neighbors; otherwise, it drops it. Thus, in general, all nodes have the same copy of the blockchain; if not, nodes build on the longest chain. One of the key limitations of proof-of-work based blockchains is scalability; indeed, the number of transactions that can be processed per second is small (e.g., up to 7 for Bitcoin [5] and 15 for Ethereum [6]). This is unacceptable for most payment applications that require 1000s of transactions per second (e.g., Visa and PayPal). The objective of blockchain scalability is to process a high number of transactions per second (i.e., throughput) without sacrificing security and decentralization [23]. Indeed, we can easily considerably increase the throughput but we will lose in terms of decentralization (which is a key characteristic of blockchain) [23], [34].

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

A number of solutions to scale blockchain have been proposed; we can classify them into two categories: (1) On-chain solutions: they propose modifications to the blockchain protocols, such as sharding (e.g., [2], [14]) and block size increase (e.g., [15]); and (2) off-chain solutions (aka layer 2 solutions): these are built on the blockchain protocols; they process certain transactions (e.g., micro-payment transactions) outside the blockchain and only record important transactions (e.g., final balances) on the blockchain. Examples of layer 2 solutions include Lightning Network [12], Raiden Network [16], Plasma [8], and Atomic-swap [24]. Security and decentralization should be taken into account while solving the scalability issue in public blockchains. This is called the scalability trilemma; indeed, finding a balance between scalability, security, and decentralization is very challenging. In this paper, we focus on analyzing the security of scalability solutions that use the concept of sharding; this is motivated by the fact that sharding is one of the most promising solutions to the scalability problem. The basic idea behind sharding is to divide the network into subsets, called shards/committees; throughout the paper, we will use the terms shard and committee interchangeably. Each committee will be working on different set of transactions rather than the entire network processing the same transactions. Several sharding protocols have been proposed in the literature; they include Elastico [4], OmniLedger [3], RapidChain [2], Zilliga [11], PolyChard [7] and Harmony [10]. Generally, sharding is used in non-byzantine settings (i.e., settings not able to resist the class of failures derived from Byzantine generals' problem [35], e.g., RSCoin [17]); Elastico [4] is the first sharding-based protocol that assumes the presence of byzantine adversaries. Elastico, divides the network into multiple committees where each committee handles a separate set of transactions. The number of shards grows nearly linearly with the size of the network. When the network grows up to 1, 600 nodes, Elastico succeeds at increasing the throughput (e.g., up to 40 transactions per second (tx/s)). However, Elastico has shortcomings that include: (1) the randomness used in each epoch (i.e, in each fixed time period; e.g., once a week) of Elastico can be biased by malicious nodes; and (2) it can only tolerate up to 25% of malicious/faulty nodes (total resiliency) and 33% of malicious nodes in each committee (committee resiliency). OmniLedger [3] has been proposed to fix some of the shortcomings of Elastico. In particular, it uses a bias-resistant public-randomness protocol to ensure security. The OmniLedger consensus protocol uses a variant of ByzCoin [18], to handle and achieve faster transactions (e.g., up 500 tx/s when the network grows up to 1, 800 nodes). OmniLedger claims the same resiliency, for global and committee, as Elastico. Recently, Zamani and Movahedi in [2] proposed RapidChain as a sharding-based public blockchain protocol which succeeds at outperforming existing sharding algorithms (e.g., [3], [4]) in terms of scalability and security. Indeed, RapidChain can tolerate up to 33% of malicious/faulty nodes and 50% of malicious nodes in each committee. RapidChain claims a high throughput (e.g., up to 4, 220 tx/s when the network grows up to 1, 800 nodes).

TABLE 1. Resiliency bound.

Class	Protocol	Total Resiliency	Committee Resiliency
Class A	Elastico [4]	1/4	1/3
	Omniledger [3]	1/4	1/3
	Harmony [10]	1/4	1/3
	Zilliqa [11]	1/4	1/3
Class B	Rapidchain [2]	1/3	1/2

In industry, Harmony and Zilliga have also succeeded in increasing transaction throughput [10], [11]. Zilliga's sharding design allows the network to process transactions in parallel and reach high throughput (e.g., at Ethereum's present mining network, which is over 8083 nodes, Zilliqa is expected to process about a thousand times the transaction rate of Ethereum [11], [33]). However, Zilliqa has shortcomings that include: (1) It does not divide the storage of blockchain data (i.e., state sharding); and (2) Zilliqa's sharding process is susceptible to a single-shard takeover attack [10]. Harmony [10] has been proposed to fix some of the shortcomings of Zilliqa. Harmony claims that is fully scalable (i.e., Harmony shards not only the network communication and transaction validation like Zilliqa, but also shards the blockchain state); in addition, Harmony proved that its sharding process ensures high security thanks to its distributed randomness generation process [10]. Harmony and Zilliqa claim the same local and global resiliency as Elastico and OmniLedger [10], [11].

The table bellow summarizes common characteristics of related protocols used in our analysis.

Probability bounds (aka, tail inequalities) are one of the most basic and versatile tools in the life of theorical computer science, with apparently endless amount of applications. Almost every modern publication on algorithms or complexity theory contains a statement and a proof of the bounds [25]–[28]. There are several books in computer sciences that discuss its various applications in great detail [28]–[30]. Additionally, many articles (e.g., [36], [37], [46], [47]) appeared that use the probability bounds to address some of the algorithmic issues, such as decision analysis problems and structural engineering design decisions.

In the literature, there are several sharding-based blockchain protocols that use the binomial distribution to refer to the sampling without replacement (e.g., OmniLedger [3], Ethereum-sharding [14]); this is a problem that can compromise the security of protocols since the binomial distribution is not accurate when the sampling is done without replacement. We use the binomial distribution to refer to the sampling with replacement; when the sampling is done without replacement, we use the hypergeometric distribution [22]. We note that, we can use the binomial distribution to refer to the sampling without replacement only when the sample size is smaller than 10% of the size of the blockchain network [22].

The key contribution of our work is the use of probability bounds as an alternative solution when the simulations and

computations are complex [29], especially when the size of the committee gets larger. Indeed, if we are to calculate the failure probability reported in Equation (23) (see Section III), we often need to execute a simple expression function that approximates this failure probability rather than perform complex computations using Equation (23). Particularly, using Chvátal's bound this failure probability can be approximated by function $x \rightarrow F(x)$ reported in Equation (35) (see Section III), which is very simple and easier to compute.

Generally, the paper contribution consists of a solution to analyze security of sharding-based blockchain protocols, i.e., bound the failure probability and how to keep this failure probability smaller than a given threshold. For example, in the case of RapidChain, our solution allows us to find out that when the network grows up to 4000 nodes, each committee must contain more than 275 nodes to ensure that the failure probability for one committee is smaller than 8.64E-08; therefore, we have a very reasonable risk of security (i.e., our network will fail on average every 2242 years or more).

In this paper, we select three probability bounds to analyze the security of sharding-based blockchain protocols; Chebyshev [47], Chvátal [21], and Hoeffding [19]. Generally, there are several scenarios in which we resort to compute probability bounds. For instance, if one aims to ensure that the failure probability is smaller than a given threshold (e.g., 10^{-3}), there is no need to compute the exact failure probability which may be very difficult to calculate; indeed, computing a probability bound (e.g., 10^{-3}) in this case is much simpler and sufficient (e.g., Chvátal's bound).

To the best of our knowledge, there is no existing work that analyzes security of Blockchain protocols using Hoeffding and Chvátal inequalities except our contribution [1]. In this contribution [1], we presented a probabilistic security analysis that is specific to Elastico, OmniLedger, and RapidChain. However, in this paper, we generalize the model to analyze any sharding-based blockchain protocol. Furthermore, we consider and implement more bounds (than in [1]) in addition to a more comprehensive evaluation of the proposed model. The contributions of this paper can be summarized as follows:

- We develop a general model to analyze the security for any sharding-based blockchain protocol.
- We analyze Elastico [4], OmniLedger [3], and RapidChain [2] to validate the proposed model. We also analyze Harmony [10] and Zilliqa [11] protocols.
- We implement the classical bound of Chebyshev (that is commonly used in the litterature [31], [37], [39], [47]) together with Hoeffding and Chvátal bounds to evaluate our model and compare the three bounds.
- We propose an approach that determines the conditions that need to be satisfied, by a sharding-based blockchain protocol, in order to keep the failure probability smaller than a given threshold.

The rest of the paper is organized as follows. Section II presents the general proposed probabilistic model. Section III validates the proposed model. Section IV evaluates the performance of the proposed model. Finally, Section V concludes the paper and presents future work.

II. ANALYTICAL MODEL

In sharding-based blockchain protocols, we propose to use the hypergeometric distribution instead of the binomial distribution; this is because the process of assigning nodes to shards can be defined as a sampling without replacement (committees do not overlap); in this case, hypergeometric distribution yields better approximation compared to binomial's [22]. Thus, the formation of committees/shards (or partition of the network into shards/committees) will be modeled by using hypergeometric distribution. In this paper, we use also the binomial distribution to model the formation of committees/shards when the sample size is smaller than 10%. Generally, we use binomial distribution when the sample is drawn with replacement [22].

There are several probability bounds that can be applied in computer sciences; Markov and Chebyshev bounds are the most common inequalities used in probability theory [37], [38]. Chernouff's bound is also very used in literature [45], [46]. Markov's bound is only applied to non-negative random variables [38], whereas Chebyshev's bound can be applied to any random variable, e.g, see Equation (11) for binomial random variable and see Equation (10) for hypergeometric random variable. In addition, for any independent random variable we can apply Chernouff's bound [46]. Chvátal's bound [21] and Hoeffding's bound [19] have analogue tail bounds for both binomial and hypergeometric distributions [19], [20]. Hush and Scovel [41] and Bardenet and Maillard [40] bounds are used for hypergeometric distribution. For the binomial bounds, we can use Leon and Perron [42] and Talagrand [43].

The three bounds, namely Chebyshev, Chvátal, and Hoeffding, can be applied for both binomial and hypergeometric distributions. We can also apply Markov's bound; however, Markov's bound is the weakest (in terms of accuracy) because it is constant and does not change as the number of nodes in a committee increases. Chernouff's bound is not reported in this work due to the difficulties to obtain the moment generating function for the hypergeometric distribution [48].

In this section, we present the details of our probabilistic model. More specifically, we present the details of (1) hypergeometric and binomial distributions; and (2) Chebychev's bound that defines a classical bound which is applicable for any random variable, Hoeffding's bound that is applicable for binomial and hypergeometric random variable, and Chvátal's bound that defines an exponential bound which is applicable for binomial and hypergeometric random variables.

A. NOTATIONS AND DEFINITIONS

Table 2 shows the list of symbols/variables that are used to describe the proposed model. Note that the cumulative

TABLE 2. Notations.

Notation	Description
N	Total number of nodes
n	Committee size
K	Total number of malicious nodes
r	Committee resiliency
R	Total resiliency
n_c	Number of committees
p_c	Committee failure probability
p_e	Epoch failure probability
$p_{bootstrap}$	Bootstrap probability
$h(K, N, n, k)$	Hypergeometric distribution with parameters K, N and n
$H(K, N, n, k)$	Cumulative hypergeometric distribution with parameters K, N and n
$\mathrm{B}(n, p, k)$	Cumulative binomial distribution with parameters n and p
A	Average number of years to failure
E_s	Expected number of sharding rounds until failure
N_s	Number of sharding rounds per year

hypergeometric distribution $H(K, N, n, k)$ is the sum of the probability distribution function $h(K, N, n, i)$ for all $i \geq k$.

Definition 1: Failure Probability or error probability. Failure probability is defined as the probability that the number of malicious nodes exceeds the malicious nodes limit (i.e., maximum percentage of nodes/users that can act in a malicious manner, e.g., in case of Elastico [4], the limit is 25% of the nodes in the network) in the network/committee.

Definition 2: Failure Probability Bound. Given a sharding-based blockchain protocol, the failure probability bound is an upper bound function that estimates the failure probability.

Definition 3: Committee Resiliency. The maximum number of malicious nodes that the committee can contain whereas still being secure.

Definition 4: Total Resiliency. The maximum number of malicious nodes that the whole network can contain whereas still being secure.

B. PROBABILITY DISTRIBUTION

Let X and $P(X = k)$ denote the random variable corresponding to the number of malicious nodes in the sampled committee and the probability that a committee contains k malicious nodes out of n draws without replacement from a finite population (entire network) of size N with K total malicious, respectively.

1) HYPERGEOMETRIC DISTRIBUTION

Let X follows the hypergeometric distribution with parameters K, N and n. We have:

The mean (expected value) is:

$$\mathrm{E(X)} = n\frac{K}{N} = np \tag{1}$$

and the variance is as follows:

$$Var(X) = \frac{np(1-p)(N-n)}{N-1} \tag{2}$$

The probability that a committee contains k malicious nodes ($P(X = k)$) is expressed as follows:

$$h(K, N, n, k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{3}$$

The failure probability for a committee with resiliency r is defined as follows:

$$H(K, N, n, nr) = \sum_{k=\lfloor nr \rfloor}^{n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{4}$$

2) BINOMIAL DISTRIBUTION

Let X follows the binomial distribution with parameters n and p where $p = \frac{K}{N}$ is the probability that a node is malicious. The mean is expressed as follows:

$$\mathrm{E(X)} = np \tag{5}$$

and the variance is expressed as follows:

$$Var(X) = np(1-p) \tag{6}$$

In general, when the hypergeometric distribution is used, a comparison is performed with the binomial distribution [22]. More specifically, it is reported that if n is small relative to the population size N, then X could be approximated by a binomial distribution [22]. Practically, we approximate hypergeometric distribution by a binomial distribution when the sample size is smaller than 10% of the population [22]. However, when the sample size gets larger relative to the population size, it is recommended to use the hypergeometric distribution (the hypergeometric distribution yields a better approximation in this case) [22]. If the sampling is done with replacement, we use the cumulative geometric distribution [20] or cumulative binomial distribution [22] instead of the cumulative hypergeometric distribution to calculate the failure probability [22]. Now, if we assume that $X \sim \mathrm{B}(n, p)$ (i.e., X follows the binomial distribution with parameters n and p) where $p = \frac{K}{N}$ is the probability that a node is malicious. Thus, the failure probability of one committee with resiliency r using the cumulative binomial distribution can be expressed as follows:

$$P(X \geq nr) = \sum_{k=\lfloor nr \rfloor}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \tag{7}$$

C. TAIL INEQUALITIES

The main contribution of our work is to bound the failure probability for one committee and so for one epoch using three bounds functions. The tail inequalities are powerful results that can compute these bounds [19]–[21], [31], [44]. Let X follows the hypergeometric distribution with parameters K, N, and n. Firstly, we bound the failure probability for one committee as well as for each epoch using Chebyshev's bound [47].

1) CHEBYCHEV'S BOUND [47]

If X is a random variable, then for any $a \geq 0$ we have

$$P(\mid X - E(X) \mid \geq a) \leq \frac{Var(X)}{a^2} \tag{8}$$

a) Chebyshev's bound corresponding to the hypergeometric random variable.

Using Chebyshev's bound, we propose to find a bound on $P(X \geq nr)$.

$$\begin{aligned} P(X \geq nr) &= P(X - \mathrm{E(X)} \geq nr - \mathrm{E(X)}) \\ &= P(X - np \geq nr - np) \\ &\leq P(\mid X - np \mid \geq nr - np) \\ &\leq \frac{\mathrm{Var(X)}}{(nr - np)^2} \\ &= \frac{np(1-p)(N-n)}{(N-1)(nr-np)^2} \end{aligned} \tag{9}$$

Thus, we bound the failure probability of one committee with resiliency r as follows:

$$P(X \geq nr) \leq \frac{np(1-p)(N-n)}{(N-1)(nr-np)^2} \tag{10}$$

b) Chebyshev's bound corresponding to the binomial random variable.

We propose to bound the failure probability of one committee with resiliency r, i.e., to find a bound on $P(X \geq nr)$, by using Equation (6) and Equation (9):

$$P(X \geq nr) = \frac{np(1-p)}{(nr-np)^2} \tag{11}$$

Now, we bound the failure probability of each epoch; we calculate the union bound over n_c committees, where each committee can fail with probability p_c. When the sample size is smaller than 10% of the size of the blockchain network, p_c is calculated using cumulative binomial distribution. Otherwise, we use the cumulative hypergeometric distribution. In the first epoch for each protocol, the committee election procedure fails with bootstrap probability (e.g., for RapidChain $p_{bootstrap} \leq 2^{-26.36}$, see [2]). Thus, the failure probability for one epoch, by using Chebyshev's bound corresponding to the hypergeometric random variable, is bounded as follows:

$$p_{bootstrap} + n_c p_c \leq p_{bootstrap} + n_c \frac{np(1-p)(N-n)}{(N-1)(nr-np)^2}, \tag{12}$$

2) HOEFFDING'S BOUND

Hoeffding proposes another bound [19], which is expressed as follows:

$$H(K, N, n, k) \leq G(x), \tag{13}$$

where

$$G(x) = \left(\left(\frac{p}{p+x} \right)^{p+x} \left(\frac{1-p}{1-p-x} \right)^{1-p-x} \right)^n, \tag{14}$$

$p = \frac{K}{N}$ and $k = (p + x)n$ with $x \geq 0$.

Hence, we can bound the failure probability of one committee with resiliency r as follows:

$$H(K, N, n, nr) \leq G(x), \tag{15}$$

where

$$x = r - p, \quad (p \leq R).$$

The binomial distribution coincidentally has an analogous tail bound [21]; thus,

$$\mathrm{B}(n, p, nr) \leq G(x), \tag{16}$$

where

$$\mathrm{B}(n, p, nr) = \sum_{k=\lfloor nr \rfloor}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

The failure probability for one epoch (p_e) is bounded as follows:

$$p_{bootstrap} + n_c p_c \leq V(x), \tag{17}$$

where

$$V(x) = p_{bootstrap} + n_c G(x), \quad n_c = \frac{N}{n}.$$

3) CHVÁTAL'S BOUND

Chvátal proposes another tail bound [21]; it is simple and elegant (i.e., exponential function), but weaker bound compared to Hoeffding's [19]. We obtain the following bound:

$$H(K, N, n, k) \leq F(x), \tag{18}$$

where

$$F(x) = \exp^{-2x^2 n}.$$

Thus, the failure probability for one epoch is bounded as follows:

$$p_{bootstrap} + n_c p_c \leq U(x), \tag{19}$$

where

$$U(x) = p_{bootstrap} + n_c F(x), \quad n_c = \frac{N}{n}.$$

D. YEARS TO FAILURE

In this subsection, we propose to quantify/measure the security of the network. More specifically, we compute the average number of years to failure using the failure probability of committee/epoch per sharding round. The average number of years to failure is given by:

$$A = \frac{E_s}{N_s}, \tag{20}$$

where

$$E_s = \frac{1}{p_e} \tag{21}$$

III. APPLICATION

In this section, we will apply our model on two classes of protocols; class A that contains Elastico [4], OmniLedger [3], Zilliga [11], and Harmony [10] (protocols that claim the same committee resiliency and total/global resiliency, which are $\frac{1}{3}$ and $\frac{1}{4}$ respectively), and class B that contains RapidChain [2] (that claims $\frac{1}{2}$ of committee resiliency and $\frac{1}{3}$ of total resiliency). The aim of this section is to validate our model by applying it to protocols of classes A and B.

A. PROBABILITY DISTRIBUTIONS

For all these protocols (i.e., Elastico [4], OmniLedger [3], Zilliga [11], Harmony [10], RapidChain [2]), the sampling is drawn without replacement (i.e., the committees can not overlap). This is the reason we chose the hypergeometric distribution to calculate the failure probability for one committee and then for each epoch. The failure probability for one committee for the class A protocols (they have the same committee resiliency 33%, see Table 1) using the cumulative hypergeometric distribution is expressed as follows:

$$H(K,N,n,\frac{n}{3}) = \sum_{k=\lfloor \frac{n}{3} \rfloor}^{n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}. \tag{22}$$

Similarly, we can express the failure probability for class B using the hypergeometric distribution as follows:

$$H(K,N,n,\frac{n}{2}) = \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}. \tag{23}$$

Now, the failure probability of one committee for class A using the cumulative binomial distribution is expressed as follows:

$$P(X \geq \frac{n}{3}) = \sum_{k=\lfloor \frac{n}{3} \rfloor}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \tag{24}$$

Likewise, the cumulative binomial distribution for class B is given by:

$$P(X \geq \frac{n}{2}) = \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \tag{25}$$

B. TAIL INEQUALITIES

As stated earlier, our aim is to analyze security by computing failure probability bounds. We will apply probability bounds, presented in Section II, in order to bound the failure probability. Firstly, we apply Chebyshev's bound [47].

a) Chebyshev's bound corresponding to the hypergeometric random variable.

The failure probability of one committee for protocols in class A using Equation (9) is bounded as follows:

$$H(K,N,n,\frac{n}{3}) \leq \frac{np(1-p)(N-n)}{(N-1)(\frac{n}{3}-np)^2}, \tag{26}$$

Likewise, we bound the failure probability of one committee for class B, as:

$$H(K,N,n,\frac{n}{2}) \leq \frac{np(1-p)(N-n)}{(N-1)(\frac{n}{2}-np)^2}, \tag{27}$$

We consider that all protocols in class A have the same bootstrap probability as in RapidChain [2]. Using Equation (12), we can bound the failure probability for each epoch as follows:

For class A

$$p_{bootstrap} + n_c H(K,N,n,\frac{n}{3}) \leq 2^{-26.36} + n_c \frac{np(1-p)(N-n)}{(N-1)(\frac{n}{3}-np)^2}, \tag{28}$$

and for class B

$$p_{bootstrap} + n_c H(K,N,n,\frac{n}{2}) \leq 2^{-26.36} + n_c \frac{np(1-p)(N-n)}{(N-1)(\frac{n}{2}-np)^2}, \tag{29}$$

b) In the same way, we apply Chebyshev's bound corresponding to binomial random variable.

The failure probability for each epoch using Equation (11) is expressed as follows:

For class A

$$p_{bootstrap} + n_c H(K,N,n,\frac{n}{3}) \leq 2^{-26.36} + n_c \frac{np(1-p)}{(\frac{n}{3}-np)^2}, \tag{30}$$

and for class B

$$p_{bootstrap} + n_c H(K,N,n,\frac{n}{2}) \leq 2^{-26.36} + n_c \frac{np(1-p)}{(\frac{n}{2}-np)^2}, \tag{31}$$

Secondly, we can bound the failure probability for one committee for class A protocols using Hoeffding's bound (see Equation (15)) as follows:

$$H(K,N,n,\frac{n}{3}) \leq G(x), \tag{32}$$

where

$$x = \frac{1}{3} - p, \quad (p \leq \frac{1}{4}).$$

By using Equation (16), the binomial distribution has an analogous tail bound; thus,

$$\mathrm{B}(n,p,\frac{n}{3}) \leq G(x), \tag{33}$$

where

$$\mathrm{B}(n,p,\frac{n}{3}) = \sum_{k=\lfloor \frac{n}{3} \rfloor}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

We conclude that the failure probability bound for one epoch for class A protocols can be computed as follows:

$$p_0 + n_c p_c \leq V(x), \tag{34}$$

where

$$V(x) = 2^{-26.36} + n_c G(x), \quad n_c = \frac{N}{n}.$$

TABLE 3. Parameter settings.

Parameter	Class	
	A	B
r	0.333	0.499
R	0.250	0.333
x	0.083	0.167
N_s	365	365

In the same way, we bound the failure probability for each committee/epoch for class B.

Finally, by using Chvátal's bound, the failure probability bound for one committee for class B can be computed as follows:

$$H(K, N, n, \frac{n}{2}) \leq F(x), \tag{35}$$

where

$$F(x) = \exp^{-2x^2 n}.$$

Hence, the failure probability bound for an epoch for class B can expressed as follows:

$$p_0 + n_c p_c \leq U(x), \tag{36}$$

where

$$U(x) = 2^{-26.36} + n_c F(x), \qquad n_c = \frac{N}{n}.$$

Similarly, we can bound the failure probability for each committee/epoch for class A using Chvátal's bound.

IV. PERFORMANCE EVALUATION

In this section, we present a simulation-based evaluation of our proposed model by applying it to well-known sharding blockchain protocols, i.e., class A ([3], [4], [10], [11]), and class B ([2]). The main aim of the evaluation is to bound the failure probability and how to keep this probability smaller than a given threshold.

A. SIMULATION SETUP

To implement our model, we use the `hypergeom` and `binom` functions imported from `scipy.stats` Python module [49]. We use `binom.cdf()` and `hypergeom.cdf()` to compute the cumulative binomial and hypergeometric distributions respectively. Table 3 shows the values of the parameters used in the simulations.

B. RESULTS AND ANALYSIS

Figures 1 and 2 show Hoeffding and Chvátal bounds of the failure probability when varying the size of the committee ($10 - 260$ nodes) in a network of $4,000$ nodes, whereas Figure 3 shows Chebyshev's bound when varying the size of the committee ($40 - 300$ nodes) in a network of $4,000$ nodes. We observe that the three bounds of class B decrease rapidly compared to class A; this can be explained by the fact that class B has high resiliency compared to class A.

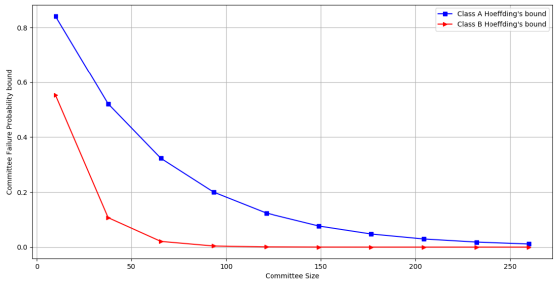

FIGURE 1. A comparison of failure probability between class *A* and class *B* protocols using Hoeffding's bound.

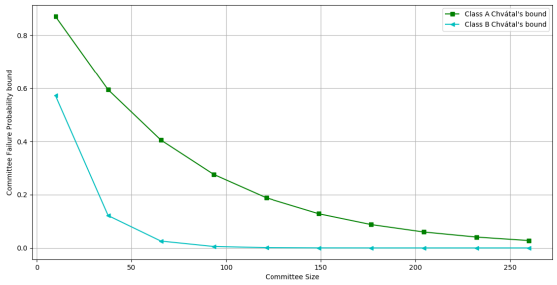

FIGURE 2. A comparison of failure probability between class *A* and class *B* protocols using Chvátal's bound.

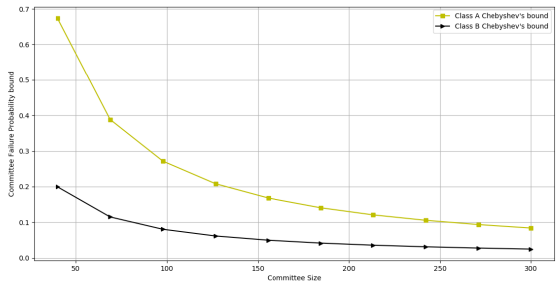

FIGURE 3. A comparison of failure probability between class *A* and class *B* protocols using Chebyshev's bound.

More specifically, Figure 1 shows Hoeffding's bound for both classes A and B; we observe that Hoeffding's bound for class B is more precise than Hoeffding's bound class A thanks to the high resiliency of RapidChain's protocol (class B). But, both bounds look similar when the committee sizes' get larger. Precisely, when the committee includes more than 200 nodes. Figure 2 shows the plot of Chvátal's bound of the failure probability for one committee in the class A and B versus the committee sizes'. We observe that Chvátal's bound for both classes both decreases when the committee size increases. We observe also that Chvátal's bound for class B gets more precise and allows less failure probability. In addition, the bounds get closer starting from 200 nodes. Finally, Figure 3 illustrates the plot of the Chebyshev's bound for classes A and B. We observe also that the Chebyshev's bound of class B decreases faster than Chebyshev's bound of class A.

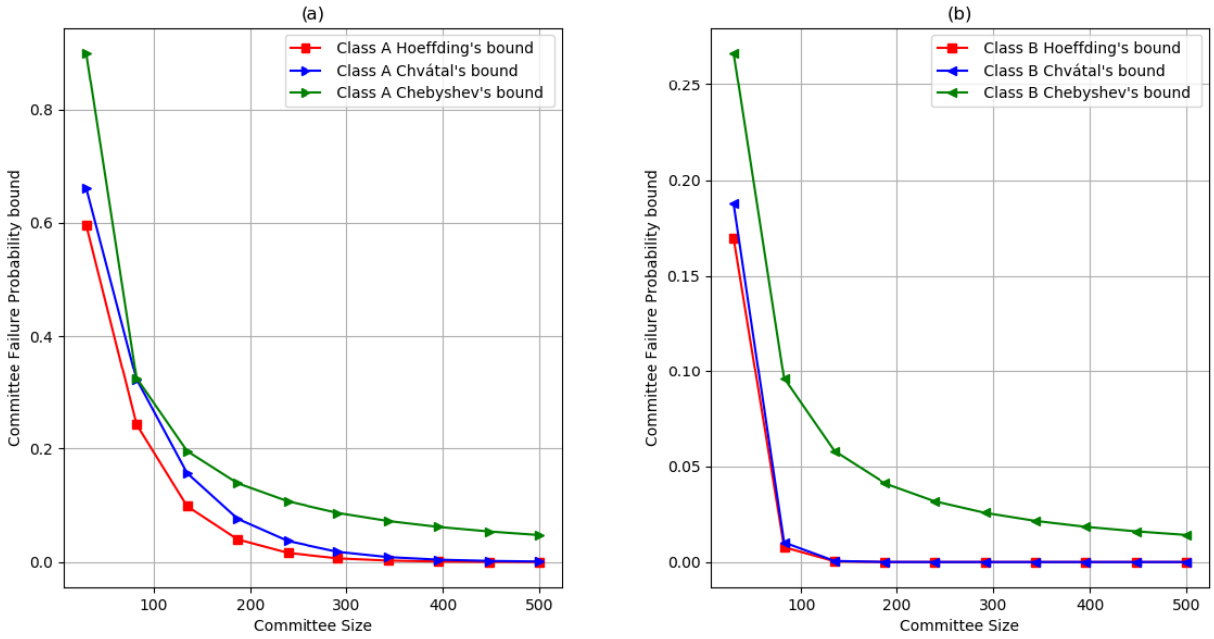

FIGURE 4. A comparison between bounds: Hoeffding, Chvátal, and Chebyshev; (a) for class *A*, and (b) for class *B*.

Figure 4 illustrates a comparison of the three bounds using classes A and B. Both figures 4(a) and 4(b) show that Hoeffding's bound provides better approximation followed successively by Chvátal's and Chebyshev's bounds.

In summary, Figures 1, 2, 3, and 4 show that Hoeffding's bound is the best one. Hence, the feasibility of the Hoeffding's bound to better analyze the security of sharding-based blockchain protocols.

Figure 5 shows the Hoeffding's bound and failure probability of classes A and B using the hypergeometric and binomial distributions to sample a committee without replacement with various sizes ($30 - 500$ nodes) from a pool of $4,000$ nodes. In particular, Fig. 5(a) shows the plot of the failure probability for one committee as well as Hoeffding's bound of failure probability for class A. We observe that Hoeffding's bound looks similar to the plot of failure probability when the committee size increases (when it approaches 225 nodes). Hence, we get a good approximation bound when the size of the committee gets larger. Fig. 5(b) illustrates the plot of Hoeffding's bound of the failure probability; we observe that the failure probability and the failure probability bound decrease when the size of the committee increases. We also observe that when the size of the committee is bigger than 100, the probability bound and the probability assume similar values.

Figure 6 shows Hoeffding's bound of the failure probability for one epoch (i.e., union bound over the number of committees) as well as the failure probability simulations for for classes A and B whereas varying the size of the committee ($100 - 560$ nodes) from a pool of $4,000$ nodes. More specifically, Figure 6(a) shows the Hoeffding's bound of the failure probability for one epoch for class A and the failure probability for one epoch when the size of the committee gets larger. We observe that Hoeffding's bound and the failure probability get closer when the number of nodes is bigger than 200. Figure 6(b) illustrates the Hoeffding's bound of the failure probability for one epoch for class B as well as the failure probability for one epoch when the committee size gets larger.

As shown in the Figures 5 and 6, Hoeffding's bound achieves better failure probability estimation especially when the size of the committee gets larger.

The key questions we want to answer with the evaluation is how to keep the failure probability smaller than a given threshold for the purpose of achieving a very reasonable risk of security. To do this, we can use failure probability simulations or probability bounds. For example, based on Figure 5(b), If $k < \frac{n}{2}$ (as in case of class B) each committee must contain more than 80 nodes to keep the failure probability negligible, i.e., $P(X \geq \frac{n}{2}) < 1.4\text{E-}03$. Using Hoeffding's bound, if we want to keep the failure probability smaller than 1.4E-04 the committee size must contain more than 150 nodes. For class A, the committee size can tolerate up to $\frac{n}{3}$; thus, the committee size will be significantly bigger compared to class B; based on Figure 5(a), if we want to keep the failure probability smaller than 1.34E-02, the committee size must contain more than 250 nodes for $P(X \geq \frac{n}{3}) < 1.34\text{E-}02$.

Now, based on Figure 6, if we want to keep the failure probability for one epoch for class A smaller than 1.44E-02, the committee size must contain more than 260 nodes. By using Hoeffding's bound, the committee must contain more than 260 nodes for $P(X \geq \frac{n}{3}) < 0.17$.

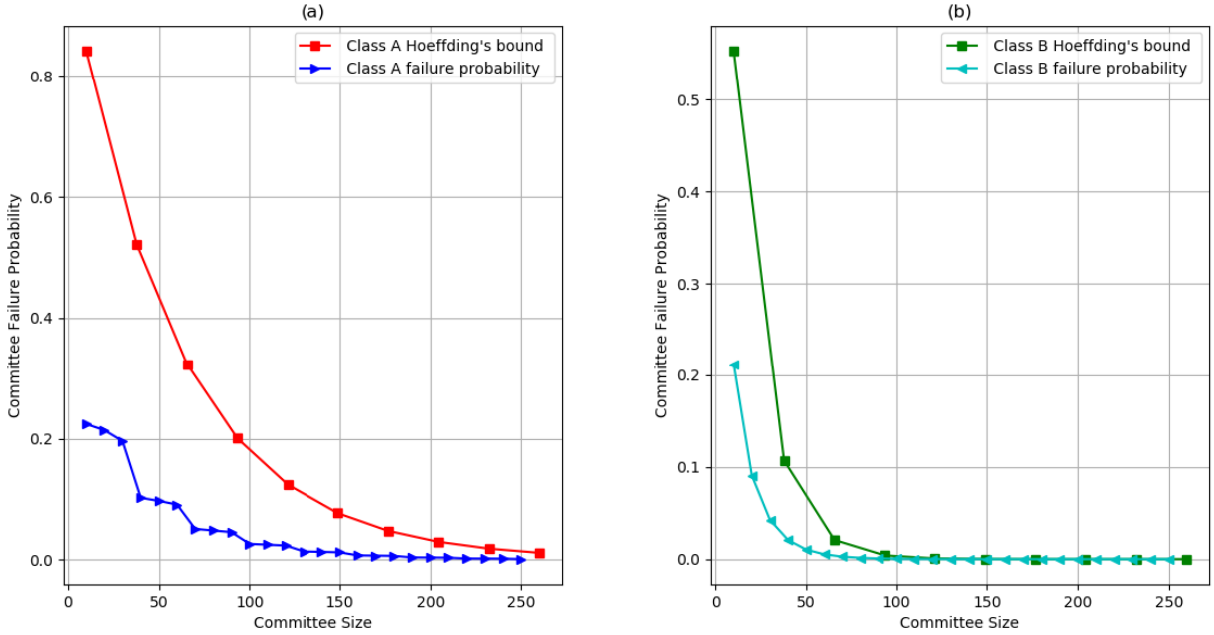

FIGURE 5. Plot of Hoeffdings' bounds, as well as the failure probability vs. committee sizes; (a) for one committee for class *A* and (b) for one committee for class *B*.

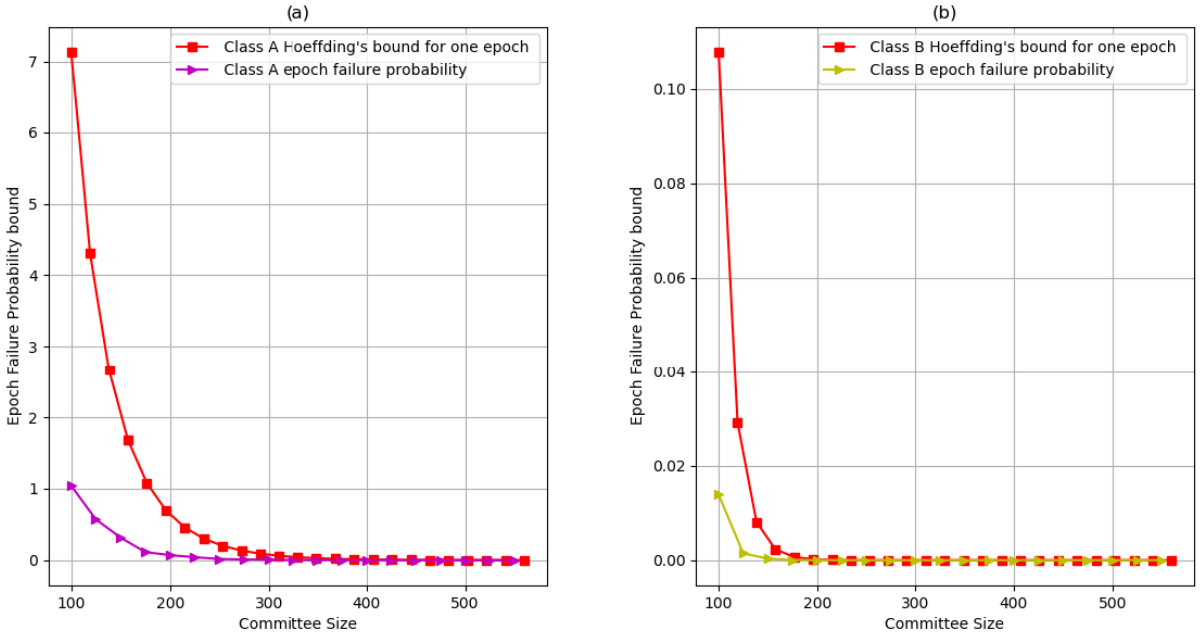

FIGURE 6. Plot of Hoeffdings' bounds, as well as the failure probability vs. committee size's; (a) for one epoch of class *B*, and (b) for one epoch of class *A*.

We can keep the failure probability smaller than a given threshold by carefully configuring the size of the committee. Also, it turns out that we have a trade-off between security and throughput; the bigger the committee size the higher the security and the smaller the throughput, and the smaller the committee size the lower the security and the bigger the throughput (the smaller the committees' size the bigger the number of committees in the network and hence the bigger the throughout). In other words, the smaller committee size leads to better throughput but can compromise security.

To quantify/measure the security of the network, we propose to compute the average number of years to failure. To perform this computation, we need to determine the failure probability of committee/epoch (see Equation (20)). There will be always a non-zero probability that a shard will be compromised, but we can make this risk level reasonable. For example, a network that fails every 1000 years on average has

TABLE 4. Years to fail.

Class	Committee Size	Value[a]	Years to Fail
A	250	1.34E-02	0.013
	800	1.03E-06	533
B	250	3.79E-07	450
	800	2.83E-23	235970

[a] Estimated value of the committee failure probability by using Hoeffding's bound [19].

an acceptable level of security, whereas a network that fails once a year is not secure enough. For class B, if we assume a network that contains 4000 nodes and we have to achieve a very reasonable risk of security, e.g., network will fail on average every 1672 years or more, using Hoeffding's bound, the failure probability for one committee must be smaller than 1.16E-07; this means that the committee must contain more than 270 nodes. Table 4 shows more details about time to fail in years when varying the size of the committee from a pool of 4000 nodes using Hoeffding's bound [19].

V. CONCLUSION

In brief, this paper provides a new model to analyze security of sharding-based blockchain protocols. More specifically, we proposed three probability bounds in order to estimate/bound the failure probability for one committee, thereafter for each epoch when we use the hypergeometric and the binomial distributions. Furthermore, we proposed an approach that determines the conditions that need to be satisfied, by a sharding-based blockchain protocol, in order to keep the failure probability smaller than a given threshold. Finally, given a failure probability threshold we propose to compute the average number of years for the network to fail. Thus, to achieve a given level of security (in terms of number of years to failure), our proposal allows to compute the minimum size of committee to consider by sharding-based blockchain protocols.

REFERENCES

[1] A. Hafid, A. S. Hafid, and M. Samih, "A methodology for a probabilistic security analysis of sharding-based blockchain protocols," in *Proc. Int. Congr. Blockchain Appl.* Springer, 2019, pp. 101–109.
[2] M. Zamani, M. Movhedi, and M. Raykova, "Rapidchain: Scaling blockchain via full sharding," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2018, pp. 931–948.
[3] E. Kokoris-Kogias, P. Jovanovic, L. Gasser, N. Gailly, E. Syta, and B. Ford, "OmniLedger: A secure, scale-out, decentralized ledger via sharding," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2018, pp. 583–598.
[4] L. Luu, V. Narayanan, C. Zheng, K. Baweja, S. Gilbert, and P. Saxana, "A secure sharding protocol for open blockchains," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 17–30.
[5] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," in *Proc. Working Paper*. 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf
[6] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum Project Yellow Paper*, vol. 151, pp. 1–32, Apr. 2014. [Online]. Available: https://gavwood.com/paper.pdf
[7] S. Li, M. Yu, S. Avestimehr, S. Kannan, and P. Viswanath, "PolyShard: Coded sharding achieves linearly scaling efficiency and security simultaneously," 2018, *arXiv:1809.10361*. [Online]. Available: https://arxiv.org/abs/1809.10361
[8] J. Poon and V. Buterin, "Plasma: Scalable autonomous smart contracts," White Paper, 2017, pp. 1–47. [Online]. Available: https://plasma.io/plasma.pdf
[9] J. A. Jaoude and R. G. Saade, "Blockchain applications—Usage in different domains," *IEEE Access*, vol. 7, pp. 45360–45381, 2019.
[10] Harmony Team, "Harmony," Tech. White Paper. [Online]. Available: https://harmony.one/whitepaper.pdf
[11] ZILLIQA Team and others, "The ZILLIQA," Tech. White Paper, vol. 16, p. 2019, Sep. 2017.
[12] J. Poon and T. Dryja, "The bitcoin lightning network: Scalable off-chain instant payments," DRAFT Version 0.5.9.2, Tech. Rep., 2016, pp. 1–59. [Online]. Available: https://lightning.network/lightning-network-paper.pdf
[13] K. Salah, M. Rehman, N. Nizamuddin, and A. Al-Fuqaha, "Blockchain for AI: Review and open research challenges," *IEEE Access*, vol. 7, pp. 10127–10149, 2019.
[14] H.-W. Wang. (2017). *Ethereum Sharding: Overview and Finality*. Accessed: Sep. 8, 2019. [Online]. Available: https://medium.com/@icebearhww
[15] J. Garzik, "Block size increase to 2MB," *Bitcoin Improvement Proposal*, vol. 102, 2015.
[16] Raiden Network-Fast. (2018). *Cheap, Scalable Token Transfers for Ethereum*. [Online]. Available: https://raiden.network/
[17] G. Danezis and S. Meiklejohn, "Centrally banked cryptocurrencies," 2015, *arXiv:1505.06895*. [Online]. Available: https://arxiv.org/abs/1505.06895
[18] E. Kokoris-Kogias, P. Jovanovic, N. Gailly, I. Khoffi, L. Gasser, and B. Ford, "Enhancing bitcoin security and performance with strong consistency via collective signing," in *Proc. 25th USENIX Secur. Symp.*, 2016, pp. 279–296.
[19] W. Hoeffding, "Probability inequalities for sums of bounded random variables," in *Collected Works Wassily Hoeffding*. Springer, 1994, pp. 409–426.
[20] M. Skala, "Hypergeometric tail inequalities: Ending the insanity," 2013, *arXiv:1311.5939*. [Online]. Available: https://arxiv.org/abs/1311.5939
[21] V. Chvátal, "The tail of the hypergeometric distribution," *Discrete Math.*, vol. 25, no. 3, pp. 285–287, 1970.
[22] J. Wroughton and T. Cole, "Distinguishing between binomial, hypergeometric and negative binomial distributions," *J. Statist. Educ.*, vol. 21, no. 1, 2013.
[23] S. Kim, Y. Kwon, and S. Cho, "A survey of scalability solutions on blockchain," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2018, pp. 1204–1207.
[24] Komodo, "Advanced blockchain technology, focused on freedom," in *Working Paper*. 2018. [Online]. Available: https://docs.komodoplatform.com/
[25] S. Arora and B. Barak, *Computational Complexity: A Modern Approach*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[26] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 2009.
[27] O. Goldreich, "Computational complexity: A conceptual perspective," *ACM SIGACT News*, vol. 39, no. 3, pp. 35–39, 2008.
[28] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[29] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2017.
[30] M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Jorge, *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 2013.
[31] J. E. Cohen, "Markov's inequality and Chebyshev's inequality for tail probabilities: A sharper image," *Amer. Statistician*, vol. 69, no. 1, pp. 5–7, 2015.
[32] *BITNODES*. Accessed: Jul. 31, 2019. [Online]. Available: https://bitnodes.earn.com
[33] *Network Number 1*. Accessed: Jul. 31, 2019. [Online]. Available: https://www.ethernodes.org/network/1
[34] V. Buterin. Accessed: Oct. 25, 2019. [Online]. Available: https://github.com/vbuterin/scalability_paper
[35] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," *ACM Trans. Program. Lang. Syst.*, vol. 4, no. 3, pp. 382–401, Jul. 1982.
[36] I. Blanes, M. Hernández-Cabronero, J. Serra-Sagristà, and M. W Marcellin, "Lower bounds on the redundancy of Huffman codes with known and unknown probabilities," *IEEE Access*, vol. 7, pp. 115857–115870, 2019.
[37] K. B. Rao, M. Anoop, and N. R. Iyer, "Application of Chebyshev-and Markov-type inequalities in structural engineering," *Rel., Theory Appl.*, vol. 8, no. 1, 2013, Art. no. 28.

[38] S. Samuels, "The Markov inequality for sums of independent random variables," *Ann. Math. Statist.*, vol. 40, no. 6, pp. 1980–1984, 1969.
[39] S. Malamud, "Some complements to the Jensen and Chebyshev inequalities and a problem of W. Walter," in *Proc. Amer. Math. Soc.* Providence, RI, USA: AMS, 2001, pp. 2671–2678.
[40] R. Bardenet and O.-A. Maillard, "Concentration inequalities for sampling without replacement," *Bernoulli*, vol. 21, no. 3, pp. 1361–1385, 2015.
[41] D. Hush and C. Scovel, "Concentration of the hypergeometric distribution," *Statist. Probab. Lett.*, vol. 75, no. 2, pp. 127–132, 2005.
[42] C. León and F. Perron, "Extremal properties of sums of Bernoulli random variables," *Statist. Probab. Lett.*, vol. 62, no. 4, pp. 345–354, 2003.
[43] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *Ann. Probab.*, vol. 22, no. 1, pp. 28–76, 1994.
[44] E. Greene and J. A. Wellner, "Exponential bounds for the hypergeometric distribution," *Bernoulli, Off. J. Bernoulli Soc. Math. Statist. Probab.*, vol. 23, no. 3, p. 1911, 2017.
[45] J. P. Schmidt, A. Siegel, and A. Srinivasan, "Chernoff–Hoeffding bounds for applications with limited independence," *SIAM J. Discrete Math.*, vol. 8, no. 2, pp. 223–250, 1995.
[46] M. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 368–372, Jul. 1970.
[47] J. E. Smith, "Generalized Chebychev inequalities: Theory and applications in decision analysis," *Oper. Res.*, vol. 43, no. 5, pp. 807–825, 1995.
[48] K. Janardan, "On an alternative expression for the hypergeometric moment generating function," *Amer. Statistician*, vol. 27, no. 5, p. 242, 1973.
[49] E. Bressert, *SciPy and NumPy: An Overview for Developers*. O'Reilly Media, 2012.

ABDELATIF HAFID received the B.Sc. degree in mathematics and applications from the University of Moulay Ismail, Meknes, Morocco, and the M.Sc. degree in mathematical engineering from the University of Abdelmalek Essaâdi, Tangier, Morocco. He is currently pursuing the Ph.D. degree with the University of Moulay Ismail. His current research interests include applied probability, statistics, and blockchain.

ABDELHAKIM SENHAJI HAFID spent several years, as senior research scientist, at Bell Communications Research (Bellcore), NJ, USA, working in the context of major research projects on the management of next generation networks. He was also an Assistant Professor with Western University (WU), Canada, the Research Director of the Advance Communication Engineering Center (venture established by WU, Bell Canada and Bay Networks), Canada, a Researcher with CRIM, Canada, a Visiting Scientist at GMD-Fokus, Germany, and a Visiting Professor with the University of Evry, France. He is currently a Full Professor with the University of Montreal. He is also the Founding Director of the Network Research Lab and Montreal Blockchain Lab. He is also a Research Fellow with CIRRELT, Montreal, Canada. He has extensive academic and industrial research experience in the area of the management and design of next generation networks. His current research interests include the IoT, fog/edge computing, blockchain, and intelligent transport systems.

MUSTAPHA SAMIH received the Ph.D. degree in fundamental and applied mathematics from the University of Montpellier, France. He is currently a Full Professor with the University of Moulay Ismail, Meknes, Morocco. His current research interests include applied probability, statistics, and blockchain.