Truncating Blockchains with Tangly Statistics

Nathan Stouffer advised by Dr. Mike Wittie

Abstract

Blockchains have applications in cryptocurrencies, supply chain tracking, and providing data integrity. The essential service provided by a blockchain is keeping an immutable, decentralized ledger. To provide immutability, some blockchains use Proof of Work to construct an ever growing chain of blocks in a peer-to-peer network. Miners are expected to expend computational resources to create new blocks. Since blocks are intentionally difficult to create, users of a blockchain can trust the information held in the longest blockchain.

By design, miners have collective, but not individual, control over a blockchain. When miners are sufficiently decentralized, it is infeasible for a coalition of miners to gain explicit control of a blockchain. Explicit control would allow a coalition to decide which blocks are added to the chain, reducing the integrity of the system. Thus a blockchain performs better when control of the chain is decentralized. Decentralization can be difficult to achieve when blockchains grow to an unmanageable length. To begin mining, one must download and process the entire chain. This is not a problem when the chain is relatively small, but a larger chain (such as Bitcoin) can take days to process. Such a long chain prevents lightweight devices from becoming miners and incredibly long bootstrapping time deters participation from users who do have sufficient space. Together, these issues make joining a blockchain more difficult.

Over the summer, I worked with collaborators to devise a method that bypasses the need for every block. At a high level, our solution has miners of recent blocks vote for a blockchain summary. If sufficiently many votes agree, bootstrapping nodes can trust the blockchain summary and drastically reduce their bootstrapping time. For my capstone project, I will complete the solution from this summer and implement a model of the solution. Using the model, I will run simulations and extract experimental results.

1 Introduction

Blockchain was introduced in order to provide distributed consensus without relying on a central party [1]. Bitcoin was released to the general public in 2009 as the first implementation of a blockchain. Bitcoin garnered a lot of support and quickly became one of the most prominent cryptocurrencies. Since Bitcoin's inception, blockchains have found other applications. They are used in other cryptocurrencies, verify supply chains, and increase device autonomy in the Internet of Things [2]. Blockchains are even being used to provide security and privacy in the medical field [3]. Blockchain technology has the potential to enhance data security in a wide range of diverse applications.

By design, blockchains are decentralized. Avoiding a central party provides two benefits. First, there is no central location for an attacker to compromise, giving blockchain a robustness that centralized systems struggle to achieve. Second, there is no single party with complete control over a system, preventing a monopoly of power. An example of where lacking a central party is useful is international currency exchange. Banks have the power to charge high transaction fees for international monetary transfers and customers are forced to pay the fees because there is no alternative. Instead, if one transfers money over a cryptocurrency, there is too much competition between parties (miners) for a monopoly to emerge. Since there is no monopoly, the market finds an fair equilibrium between miners and users.

Blockchain's services are best realized when control of the chain is decentralized. Decentralized control can become difficult to achieve when blockchains grow to an unmanageable size. Joining a well-established blockchain can take days, excluding lightweight devices and reducing the likelihood a large device joins. This means a blockchain is less decentralized and prevents Internet of Things devices from participating in a blockchain network. This proposal provides and analyzes a new, off-chain, efficient, and secure method for joining a blockchain. If successful, this project contributes towards making blockchains more effective and increasing their decentralization. More effective blockchains can extend data reliability and user privacy [2] [3].

The remaining sections provide background information, our solution, our methodology, results, and a discussion of the results.

Nathan Stouffer

email: nathanstouffer1999@gmail.com
phone: 208.293.5883
github: github.com/nathanstouffer

A motivated student graduating in Spring 2021 with a double major in computer science and mathematics. Hoping to join a fast-paced and exciting work environment where I can combine both my majors to develop software.

EDUCATION

Computer Science Major and Data Science Minor — *Bachelor of Science*

GRADUATING SPRING 2021

Mathematics Major — *Bachelor of Science*

GPA: 3.94

Montana State University (MSU) – Bozeman, MT

Relevant coursework: advanced algorithms, networks, software engineering, computer security, computer graphics, machine learning, computer science theory, topology, dynamical systems

EMPLOYMENT

Math Tutor — *MSU*

FALL 2018 – PRESENT

Working as a tutor in the Math Learning Center at MSU
Identifying difficult areas for students and explaining problem solving techniques
Assisting students with algebra, pre-calculus, calculus I/II, and linear algebra

Research Experiences for Undergraduates — *MSU*

SUMMER 2020

Participated in a Research Experiences for Undergraduates (REU) program
Communicated with collaborators about complex technical problems and ideas
Worked towards truncating a blockchain network's ever-growing chain

Computer Science Course Assistant — *MSU*

FALL 2019

Instructed a lab section of 24 students in an introductory computer science course
Encouraging students to think through problems and own their solution
Hosted weekly office hours
Graded lab assignments and larger programs

PROJECTS

Senior Capstone — *MSU*

FALL 2020 – PRESENT

Implementing and extending the research performed in my REU program
Working with collaborators to produce a new protocol that prunes blockchains
Building and interpreting results from a large-scale model of our solution

Emergent Behavior — *MSU*

SUMMER 2020 – PRESENT

Exploring the emergent behavior of agents acting on local rules
Building an agent-based computer simulation
Analyzing bifurcations in emergent behavior based on initial conditions
Comparing results with a partial differential equation model

Directed Reading Program — *MSU*

SPRING 2020

Independently studied a book about abstract algebra
Discussed thoughts and questions with two graduate students once per week

SKILLS

Java, Python, C++, C, OpenGL, MATLAB, LaTeX, Git

HONORS AND AWARDS

MSU Math Department's Outstanding Scholar Award

MSU's Achievement Scholarship

Milton F. Chauner Mathematics Excellence Scholarship

Mary C. Griffin Scholarship

President's List (MSU): S18, F18, F19, S20

Dean's List (MSU): F17, S19

1st place team at MSU's 2019 programming contest

2nd place team at MSU's 2018 programming contest

Ran a marathon

Voted "Most Influential (2018–2019)" by MSU's National Residence Hall Honorary

ADDITIONAL EMPLOYMENT

Resident Assistant — MSU
Construction Laborer — ID

INTERESTS

skiing (downhill/xc)
ultimate frisbee
running
cribbage

2 Background

2.1 Blockchain

At a high level, a blockchain is a sequence of blocks that is immutable in practical applications. Immutability in blockchains is often provided through Proof of Work. Proof of Work was introduced in [4] and first used for blockchain in the Bitcoin whitepaper [1]. In a Proof of Work blockchain, a block is valid if its cryptographic hash is below a preset threshold. Each block primarily consists of data that cannot change (financial transactions, medical records, etc). However, every block contains a field for a fixed-length string of bits called a nonce [1]. If someone finds a nonce such that the hash of the block is below the threshold, they have "mined" a block. Since the hash function is cryptographic, it is difficult to reverse. A miner's best strategy is to guess and check nonces until finding a sufficiently low output. Figure 1 displays the typical structure of a blockchain.

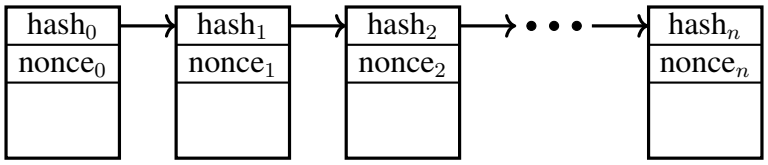

Figure 1: Prototypical Proof of Work Blockchain

A broader perspective of a blockchain is an implementation of a finite state machine. In the state machine, blocks are transitions between the states of applications running on top of a blockchain. Abstractly, the k^{th} block B_k is a transition from state S_{k-1} to state S_k with certain validity requirements. See Figure 2 for a visual. A tangible example is a cryptocurrency. Cryptocurrencies realize state as a record of the amount of currency each public key controls. Blocks are a set of transactions that transfer currency between public keys. In state S_0, no one controls any currency. Then block B_1 moves the blockchain to S_1 where its miner controls some newly created currency. Then block B_2 grants cryptocurrency to its miner and contains currency transfers between users of the currency. This process repeats until state S_n is reached, which contains the current account-balance pairs of all the current users of a cryptocurrency.

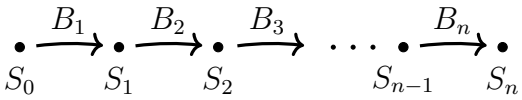

Figure 2: Blocks as transitions between states

2.2 Technical Problem

To understand the technical problem, we must understand two properties of blockchains. First, miners must obtain every block before mining new blocks. And, second, blockchains grow in

perpetuity. These properties mean that new miners must download and process an ever increasing amount of data as time passes. Even now, when Bitcoin is just over a decade old, it can take days to become a miner. Such a barrier can deter large machines and entirely prevent small machines from mining.

The root of the problem is that mining requires all the information in a blockchain. Our goal is to produce an off-chain, efficient, and secure method to bootstrap miners in a blockchain. An off-chain solution is one that does affect the underlying blockchain. This is important because many on-chain solutions to this problem cause a hard fork in a blockchain. A hard fork is when a blockchain splits because miners disagree on the format of a block [5]. While a blockchain can survive a hard fork, the mining power splits and weakens the blockchain. We want an efficient solution so that our idea could be useful in an actual blockchain. Finally, security is important because blockchains often record important information such as medical records. Our solution should not introduce vulnerabilities to such important information.

To frame this project, we use the Goal Question Metric approach presented in [6]. This results in "the specification of a measurement system targeting a particular set of issues and a set of rules for the interpretation of the measurement data [6]." The model has three levels. At the conceptual level, a goal is defined. A goal consists of a purpose, an issue, a process, and a viewpoint. At the operational level, a set of questions are asked that characterizes the goal. At the quantitative level, data is associated with questions in an attempt to answer them. The data can be either objective or subjective. Table 1 displays the Goal Question Metric Approach we will use for this project.

Goal	Purpose	Decrease
	Issue	required time
	Process	bootstrap a node to a blockchain network
	Viewpoint	the bootstrapping node
Question	Q1	Is the solution off-chain?
Metrics	M1	Will there be a hard fork?
Question	Q2	Is the solution efficient?
Metrics	M2	Is the solution affected by chain length?
	M3	Bytes of network traffic generated
Question	Q3	Is the solution secure?
Metrics	M4	Theoretical probability that a malicious actor fools a bootstrapping node
	M5	Empirical probability that a malicious actor fools a bootstrapping node

Table 1: The Goal Question Metric approach for this project

2.3 Related Work

Truncating the required bootstrapping information in a blockchain is an active area of research. The remainder of Section 2 is devoted to summarizing existing ideas and solutions.

The Bitcoin whitepaper [1] introduces the concept of lightweight nodes. Instead of becoming a full node, lightweight nodes only store the header chain and query full nodes to see if transactions are valid. This is a step in the right direction because the header chain downloads quickly, but it does not grant nodes the ability to mine new blocks. Nodes can only verify old transactions.

Another idea is to periodically place a summary block in the original blockchain [7] [8]. A summary block is responsible for reporting the net change caused by a certain amount of previous blocks. This process can be made recursive to any depth by creating a summary block responsible for a set of summary blocks (a summary of summaries) [7] [8]. Trust in the summary blocks is built in the same fashion as trust in the regular blockchain. A node joining the network would download the header chain and relevant summary blocks. If a summary block has been accounted for by another summary block, there is no need to download it. This idea is nice because it relies on the same security principles as blockchain, but the implementation would require a hard fork.

A different idea would be to use mini-blockchain [9]. A mini-blockchain requires that a block B_k includes the hash of S_k, cryptographically tying the state to the blockchain. A bootstrapping node can then request the header chain, a recent state, and blocks following the recent state. The new node can verify that the received state corresponds with the blockchain and then compute the current state with the information in the newest blocks. While this solution is quite clean, it would cause a hard fork.

Another way to provide state would be to create a new blockchain that records state [10]. The new blockchain records the state of the original chain every so often and bootstrapping nodes can eliminate the need for blocks prior to the most recent summary. Since each state exists independently of all other states, the size of the second blockchain will not bloat like the original chain [10]. While this solution does operate off-chain, it is not necessarily secure. It is unlikely that a miner will give up valuable computing power to contribute to the second blockchain. This could allow a powerful, malicious actor to sabotage the second blockchain, making it unreliable.

The final related idea holds an election to verify state [11]. Miners are allowed to vote for a state and then bootstrapping nodes trust the majority of votes. Each vote is recorded in a block and each block has the capacity to store a single vote [11]. Storing votes in blocks makes them immutable, so a bootstrapping node is able to collect every vote. The issues with this idea are two-fold. First, a vote can only be generated as quickly as new blocks. In Bitcoin, this averages to about 10 minutes per block [1], meaning a bootstrapping node might have to wait a week for there to be sufficiently many votes to trust a state. Second, [11] relies on implementation details in the Bitcoin protocol and does not apply to blockchains in general.

3 Solution

3.1 Idea

Recall that our goal is to devise an off-chain, efficient, and secure bootstrapping method. The primary barrier to this goal is that blockchains require that a miner obtain every block in the chain. Viewing blockchain as a finite state machine allows us to reduce the amount of information that a bootstrapping miner must obtain. To compute the current state S_n, one need only know is some state S_k and every block B_j with $j > k$. In standard blockchains, S_k is S_0 and the bootstrapping node needs every block in the chain. If S_k is close to the end of the chain, then the miner only needs the blocks following S_k, which will download much faster!

At this point we have reduced the problem of bootstrapping to providing some recent state S_k of the blockchain. How does a bootstrapping miner obtain the state S_k? They can't just ask a blockchain node, because the node might lie and give a false state. To solve this, we draw from [11] and host an election at regular intervals (say every 1000 blocks). We allow recent miners to vote (miners are identified by the public keys in the blocks they mined). If the election is for state S_k, then a miner computes the hash of their version of S_k. The miner submits a digital signature of the hash; the signature should use the private key from the block the miner created. The digital signature prevents a malicious actor from forging the vote.

The election is decided based on the relative vote distribution. Let V denote the total number of voters (not the number of votes cast). For a tuneable $\beta \in (0.5, 1]$, we will decide an election if greater than $\lceil \beta V \rceil$ votes agree. However, we do not have every vote since voting is not required and votes can be deleted by malicious actors. Instead, we have a sample of votes. How can we figure out whether the sample was drawn from the required distribution? The χ^2 goodness of fit test is a way to measure the probability that a sample comes from an expected distribution. If the probability returned from the χ^2 test is above a tuneable threshold α, we accept the state with the most votes.

The votes are submitted to a Distributed Hash Table (DHT). Storing votes in a DHT is a compromise: it is convenient because miners can vote at their leisure, but a powerful, malicious actor can delete votes. To combat vote deletion, we draw on the idea of a tangle [12]. A tangle is a directed acyclic graph where the vertices are added to tangle by selecting the two most recently added vertices as parents. In this context, vertices are realized as votes.

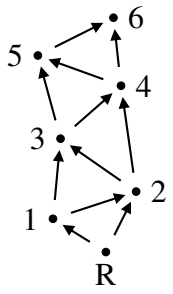

Figure 3: Example of a tangle

A set of DHT nodes are designated as tangle managers, each given complete control of their own tangle. A voter is expected to submit their vote to a deterministic subset of the tangles. A tangle manager receives votes and inserts each of them to their tangle by attaching them to parent votes. If a vote is not in every tangle of the deterministic subset, then it is considered invalid. The children of an invalid vote are also invalid. Then, deleting a vote causes a chain reaction that invalidates all the descendant of the deleted vote. The distribution of states that the descendants support matches the distribution of cast votes, so a malicious actor cannot swing an election by selectively deleting votes. A bootstrapping node can request the vote tangles and then make a decision based on valid votes.

3.2 Implementation

This capstone project will include a model implementation of the solution. The implementation will run a full-fledged DHT running on top of a network simulator. Scripted miners will submit votes to the DHT and bootstrapping nodes will request the tangles and decide the election. Using the simulation, we can gather empirical results indicating the security of our solution. The network simulator will allow us to gather information about the amount of network traffic required to facilitate an election.

4 Methodology

Much of this project is devising a method to bootstrap blockchain nodes. Since this is a conceptual task, it cannot be replicated. However, the experimental verification of our protocol can be replicated. Section 4 is devoted to providing the necessary information for someone to replicate our implementation.

In order to match our model as closely as possible with our implementation, we chose to implement our solution on a network simulator using a full-fledged DHT. This is not strictly necessary but we felt that it made our results more applicable to a real implementation of our idea.

A neglected detail of the solution is the process of selecting parents in the tangle. The tangle manager has complete control over *which* votes are selected as parents, but we require that the parents are digitally signed in the vote. This means that the manager must tell the voter which votes

are the parents and the voter includes that information in their vote. While this is some overhead communication for our solution, it prevents a malicious actor from changing the structure of a tangle to their advantage.

The most important aspect of the implementation is the intelligence of the attackers. We expect the constraints of our solution to prevent an attacker from *successfully* fooling a bootstrapping node. But the only way we can verify this is to simulate intelligent attackers. We have divided attacks into two categories: deleting votes and reordering votes. An accurate replication of this project should at least implement these strategies. Implementing more strategies is welcome, since it provides more information on our system!

We used the χ^2 test as follows. Consider some sample of votes S and our expected proportion of agreeing votes β. Our distribution comes in the form $(majority, other)$ where $majority$ references the number of valid votes that supported the most popular state and $other$ is the number of valid votes that did not support the most popular state. We have an expected distribution $(\beta|S|, (1-\beta)|S|)$ and an actual distribution of (S_{maj}, S_{oth}). Then we can use the χ^2 test to determine the probability that our sample comes from the expected distribution.

4.1 Design

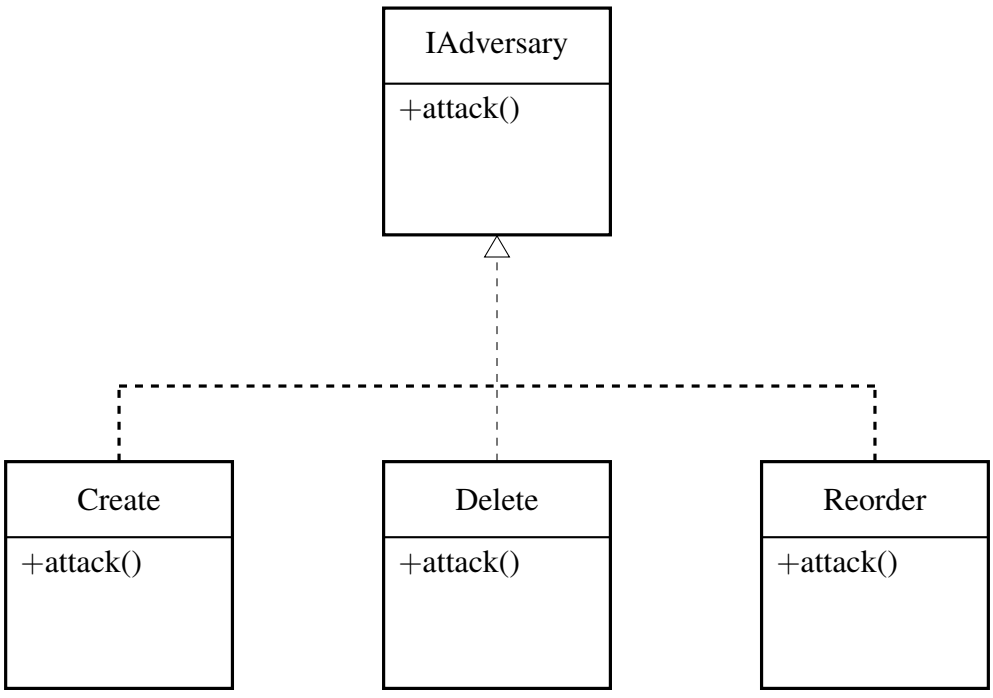

Figure 4: Strategy Pattern for attack types

5 Results

6 Discussion

7 Conclusion

References

[1] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system, 2009.

[2] Wei Cai, Zehua Wang, Jason B Ernst, Zhen Hong, Chen Feng, and Victor CM Leung. Decentralized applications: The blockchain-empowered software system. *IEEE Access*, 6:53019–53033, 2018.

[3] Asad Ali Siyal, Aisha Zahid Junejo, Muhammad Zawish, Kainat Ahmed, Aiman Khalil, and Georgia Soursou. Applications of blockchain technology in medicine and healthcare: Challenges and future perspectives. *Cryptography*, 3(1):3, 2019.

[4] Cynthia Dwork and Moni Naor. Pricing via processing or combatting junk mail. In *Annual International Cryptology Conference*, pages 139–147. Springer, 1992.

[5] Iuon-Chang Lin and Tzu-Chun Liao. A survey of blockchain security issues and challenges. *IJ Network Security*, 19(5):653–659, 2017.

[6] Victor R Basili, Gianluigi Caldiera, and H Dieter Rombach. The goal question metric approach. *Encyclopedia of software engineering*, pages 528–532, 1994.

[7] A. Palai, M. Vora, and A. Shah. Empowering light nodes in blockchains with block summarization. In *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, pages 1–5, 2018.

[8] Ulfah Nadiya, Kusprasapta Mutijarsa, and Cahyo Y Rizqi. Block summarization and compression in bitcoin blockchain. In *2018 International Symposium on Electronics and Smart Devices (ISESD)*, pages 1–4. IEEE, 2018.

[9] J. D. Bruce. The mini-blockchain scheme (a.k.a purely p2p crypto-currency with finite mini-blockchain), Jul 2014. Accessed: 2014-04-05.

[10] A. Marsalek, T. Zefferer, E. Fasllija, and D. Ziegler. Tackling data inefficiency: Compressing the bitcoin blockchain. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 626–633, 2019.

[11] Roman Matzutt, Benedikt Kalde, Jan Pennekamp, Arthur Drichel, Martin Henze, and Klaus Wehrle. How to securely prune bitcoin's blockchain. *ArXiv*, abs/2004.06911, 2020.

[12] Serguei Popov. The tangle. *cit. on*, page 131, 2016.