Mining the Web with Webcoin

Uri Klarman
Northwestern University
uri.klarman@u.northwestern.edu

Marcel Flores
Northwestern University
marcel-flores@u.northwestern.edu

Aleksandar Kuzmanovic
Northwestern University
akuzma@northwestern.edu

ABSTRACT

Four major search engines, Google in particular, hold a unique position in enabling the use of the Internet, as they alone direct over 98% of Internet users to the content they seek, using proprietary indices. While the contribution of these companies is undeniable, their design is necessarily affected by their economic interests, which may or may not align with those of the users, raising concerns regarding their effect on the availability of information around the globe. While multiple academic and commercial projects aimed to distribute and democratize the Web search, they failed to gain much traction, mostly due to inferior results and lack of incentives for participation. In this paper, we show how complex networking-intensive tasks can be crowdsourced using Bitcoin's incentive model. We present Webcoin, a novel distributed digital-currency which utilizes networking resources rather then computational, and can only be mined through Web indexing. Webcoin provides both the incentives and the means to create Google-scale indices, freely available to competing services and the public. Webcoin's design overcomes numerous unique challenges, such as index verification, scalability, and nodes' ability to actively manipulate webpages. We deploy 200 fully-functioning Webcoin nodes and demonstrate their low bandwidth requirements.

KEYWORDS

Blockchain, incentives, crowdsourcing, Web indexing.

1 INTRODUCTION

Networking tasks, such as Web crawling, webpage scraping, Web indexing, routing measurements, as well as latency and bandwidth measurements, are required by CDNs, cloud services, search engines, ISPs, and researchers. Traditionally, networking tasks are considered difficult to perform and hard to estimate, since their results depend heavily on an observer's network characteristics. At the same time, *incentivizing* users to leverage their positions in the network to perform these networking tasks remains a challenging open problem. In this paper we provide an alternative and comprehensive solution to this problem; namely, we show how Bitcoin's incentive model can be used to crowdsource complex networking tasks. Considering the large space of networking tasks which can be outsourced, we necessarily limit ourselves to one particular subproblem in this domain, and focus on continuous Web indexing, a necessary step to enable Web search, and arguably first among these networking tasks, in terms of real-world influence.

The Web search market, for both mobile and desktop clients, is extremely consolidated and dominated by Google, which holds a share of over three-quarters of the market [16]. Furthermore, the 4 largest search engines, Google, Bing, Baidu and Yahoo, together control over 98% of the market. Each one of these search engines maintains an index of the Internet's webpages and their content in order to return appropriate responses to their clients' queries. Google, which research had shown to hold the largest index, maintains an index of approximately 46.5 billion webpages [24, 26].

One of the key barriers that hinders the proliferation of a larger number of search engines is the daunting, multi-billion-dollar-worth task of crawling and indexing the exabyte-scale Web space[1]. Hence, this task is carried out by a small number of powerful players capable of committing substantial resources to conduct continuous 24/7 Web crawling and indexing. Despite efforts to democratize this space, *i.e.*, promote the use of distributed, objective and unbiased index instead, *e.g.*, [7, 9, 14, 53], the use of such distributed systems had not become common. This is mostly due to inferior results associated with small Web indices used by such systems, induced by the lack of economic incentives for participation in distributed Web crawling, or due to the small-scale nature of the communities of users interested in collaborative Web indexing.

In this paper we design Webcoin, a novel distributed digital currency which crowdsources networking tasks. Specifically, Webcoin rewards Web crawling, scraping, and indexing, by *producing money*, for those who participate in these tasks. Systems with a similar economic model, which produce money for participation, had already been successfully implemented and are gaining traction. For example, Steem [13] rewards users with money for the blogs, posts, comments and other social media content they produce, and has a market cap of over 300 million USD. However, Webcoin's novelty, focus, and unique contribution lies in being the first to incentivize networking tasks, rather than computational tasks or personal content, to *produce Google-scale Web indices within days.*

The starting point in designing Webcoin is Bitcoin [48], the first peer-to-peer currency to gain considerable traction globally. Bitcoin is a mechanism to maintain and update a ledger of transactions in a distributed manner. Its security and integrity are achieved through a process called *mining*, where every node (*miner*) conducts intensive computations in an attempt to add a valid block of transactions to the ledger. To incentivize mining, each block produces new Bitcoins which are awarded to its creator, an incentivizing mechanism which Webcoin utilizes to perform non-computational tasks. Unlike Webcoin, which aims to democratize the search engine industry, the unparalleled computation power involved in Bitcoin mining, currently at 15–20 exahashes per second, serves no other purpose beyond Bitcoin. For comparison, Bitcoin is currently consuming

[1] as disclosed to us by an executive of a major search engine

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CoNEXT '18, December 4–7, 2018, Heraklion, Greece
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6080-7/18/12…$15.00
https://doi.org/10.1145/3281411.3281415

roughly 43.7 TWh annually, similarly to Hong Kong's annual energy consumption [3].

Webcoin's goal is to utilize Bitcoin's principles, yet force miners to crawl and index the Web instead of conducting purposeless computations. However, whereas creating a valid Bitcoin block is computationally hard, and thus the production of such a block is a proof-of-work of the computational resources invested, it is challenging to use the content of indexed webpages as a proof-of-work, while providing the same security guarantees as Bitcoin. This is because both legitimate and fraudulent webpages are susceptible to manipulation. To overcome this challenge, we decouple the webpage indexing proof-of-work from the nodes' success, remodeling it as a necessary but insufficient requirement for successful Webcoin mining. We thus remove any incentive for miners to manipulate webpages, as such efforts will ***not*** affect their chances to be rewarded.

A second major challenge in Webcoin's design is the requirement for Web indices to be collected from *all* Webcoin miners, not only from the few which successfully mined a Webcoin block. Unlike Bitcoin, where only the miner which had successfully mined a new block publishes it, Webcoin must force every single Webcoin miner to publish its index, so that the entire product of the indexing effort is made available to all. We accomplish this by requiring miners to publish a hash of their Web index as a way to qualify for mining Webcoin, and to upload their compressed Web index to ensure their competitiveness.

The third challenge is associated with Web index validation, *i.e.*, how do we ensure that the Web indices submitted by miners are accurate? Contrary to Bitcoin, where a proof-of-work validation lasts several nanoseconds, validating all indices does not scale. We thus provide a mechanism that incentivizes miners to submit valid Web indices. In particular, once a miner mines a new Webcoin, its webpage indexing proof-of-work is validated. As a result, while only a small fraction of submitted indices is actually verified, the scheme ensures that nodes have no chance of winning Webcoins if they do not properly index the Web. This guarantees the validity of the Web index collectively produced by the miners.

Our main contributions are the following:

- We present a novel primitive to incentivize the crowdsourcing of complex *networking* tasks.
- We present the first digital currency to enable and incentivize the distributed creation of a verifiable Google-scale Web indexing.
- We present the first digital currency to utilize *networking* resources for its mining process, rather than processing, memory, or storage resources.
- We detail the key challenges of migrating cryptographic mining primitives to the networking domain, and provide novel techniques to address them.
- We deploy a network of 200 fully-functioning Webcoin nodes to analyze and prove their practicality and scalability. We release the Webcoin source code.
- While it is possible for resource-rich entities to amass resources (*e.g.*, acquire a large number of IPs) to participate in Webcoin's mining, they do ***not*** jeopardize Webcoin's security model (see Section 5).

We note that the creation of a competitive real-world search engine requires significant amounts of additional work, in fields ranging from Human-Computer Interaction to Information Retrieval. However, we argue that no such additional work cannot ***begin*** without an access to a global Web index, which is limited today to the employees of the largest search engines. We further note that reviewing all technical aspects of a crypto-currency designed for a new domain exceeds the scope of this paper. Our goal is thus to present the principals which enable and incentivize the crowdsourcing of networking tasks, rather than to review all of Webcoin's underlying low-level details.

2 FROM BITCOIN TO WEBCOIN

2.1 Bitcoin's Principles

Bitcoin [21, 48] is the first crypto-currency to gain considerable traction globally, with a market capitalization of over 100 billion USD. While other crypto-currencies exist, and more are suggested by academia, in general they each offer an incremental variation of the principles on top of which Bitcoin is built [46].

Bitcoin is a mechanism to maintain and update a ledger of transactions in a distributed manner. The ledger, called the *blockchain*, consists of blocks, each containing a group of transactions, with the order of blocks determining the order of the transactions. Each person or entity wishing to possess bitcoins must create a wallet, consisting of a private key and a public key, and Bitcoin transactions detail the transfer of bitcoins between these wallets.

The security and integrity of the system is achieved through a process called *mining*. Against common belief, mining is not simply the mechanism used to produce bitcoins. Rather, it produces bitcoins in order to incentivize nodes to create more blocks, and through that, to create a single secure history of all transactions. In order to mine a new block, nodes participate in a unique kind of contest. Every mining node (*miner*) exhaustively searches for a binary value (*nonce*), which, when hashed along with the previous block and the transactions of the new block, produces a value (*digest*) which must be smaller than Bitcoin's *target*.

Once found, the miner will publish the block, which includes the nonce, and it will be accepted by all nodes as the next block in the blockchain. The incentive for nodes to mine blocks arrives from the block's single *coinbase transaction*, which creates new Bitcoins and transfers them to the miner's wallet. Note that the nonce cannot be used with any other coinbase transaction, *e.g.*, one which transfers the reward to another wallet. This is because such a change will change the digest the nonce yields, which would almost certainly invalidate the block. The nonce is thus used as the miner's proof-of-work; the investment of computational resources can be proven, statistically, by presenting a nonce, for which the miner is being rewarded.

Bitcoin's security is derived from the resources invested in mining, and from the definition of the longest blockchain as the single legitimate blockchain, if multiple versions are presented. If a dishonest miner were to replace a block with a modified block, including or excluding transactions in its favor, the modified block's hash value will change, and the dishonest node will have to exhaustively search for a new nonce. More importantly, the change will also

invalidate any consecutive block, as each block's hash value depends on the block preceding it, and new nonces will need to be found sequentially. The dishonest node will have to search for a new nonce for every invalidated block, in order to make the modified blockchain as legitimate as the original one. At the same time, all other nodes invest their resources in finding the next nonce and lengthening the blockchain. Thus, the dishonest node will have to control enough resources to produce nonces at a higher rate then all other nodes *combined*, *i.e.*, the majority of computational power in the system. The likelihood of a dishonest node's success is statistical in nature, and diminishes with the number of blocks it is required to re-validate. Thus, while the transactions of the last several blocks are considered less secure, transactions included in earlier blocks are considered immutable.

Bitcoin is capable of overcoming a wide variety of scenarios which might have compromised the blockchain integrity and its nodes consensus, by incorporating the concept of *forks*. A fork is defined to be the scenario where multiple legitimate versions of the blockchain exist, *e.g.*, when two blocks containing different transactions have been mined simultaneously, and have both propagated through the nodes network. The blockchain will experience temporary ambiguity regarding which transactions have indeed taken place. However, once additional blocks are mined, one version of the blockchain will become longer then the other, and therefore legitimate. At such time, miners will switch from mining the shorter version to the longer, as there is a higher likelihood for the blockchain to converge to the longer version, and blocks mined for the shorter version and their rewards will be discarded.

2.2 Proof-of-Work Wastefulness

Bitcoin is the first blockchain-based crypto-currency, and as noted above, its mining consumes growing amounts of resources. The energy consumed in the mining process can be viewed as energy invested in the security of Bitcoin. However, the increased investment does not translate to additional functionality, which bears asking how much energy should be invested for additional security.

One approach to reduce crypto-currencies' energy footprint is to shift from Bitcoin's Proof-of-Work (*PoW*) to the use of Proof-of-Stake (*PoS*) [39]. In PoS, the exhaustive search for a nonce is eliminated, and the eligibility to produce a new block is distributed among those who hold crypto-currency funds. The more coins one holds, the higher its probability to produce the next block. However, the security guarantees and economical model of PoS systems remains to be proven. A different approach aims to replace the computational intense hashing with alternatives which require frequent memory access [52] or storage access [47, 50].

Webcoin takes a different approach, aiming to crowdsource a beneficial work that has significant real-world implications, to miners. In addition, the footprint is considerably reduced by networking tasks, as they consist of long waiting times and infrequent computations. It is worth noting that Permacoin [47] suggests it is possible to use its storage-lookup PoW to crowdsource the archiving of large data sets. However, it lacks mechanisms to make the archived information available to all.

2.3 Webcoin's Goals

Our ultimate goal in the creation of Webcoin is to incentivize distributed Web indexing, in a manner which enables unrestricted access to the resulting indices. To achieve this goal, Webcoin migrates Bitcoin's principles from the computational domain to the networking domain, and rewards miners for performing Web indexing, unlike any other digital currency, to the best of our knowledge. To achieve the above goal in practice, Webcoin must satisfy the following goals:

(1) **Security.** Webcoin must provide the same security and integrity guarantees as Bitcoin to achieve practicality.
(2) **Scalability.** Webcoin must scale to the Bitcoin network size and beyond, without increasing the load on the miners as the index size reaches Google-scale. Webcoin must also not increase the load on Web servers as the number of miners increases.
(3) **High Participation.** Webcoin's design must allow and incentivize the participation of miners with limited networking capacities, *i.e.*, with low bandwidth and high latency. Large user participation is essential for gaining momentum with system's popularity, particularly in early days of the system adoption. We discuss the economic effects of larger mining operations in Section 5.
(4) **Indices Availability and Reliability.** Unlike Bitcoin, where only the first node to mine a block publishes it, to be propagated to the entire Bitcoin network, Webcoin must not only allow, but force every single Web-coin miner to publish its index, so the entire product of the indexing effort is made available to all. Moreover, Webcoin aims to verify the indices' correctness, and to incentivize miners to reliably report the results of their indexing.

2.4 Webcoin's Non-Goals

While Webcoin is designed with real-world deployment in mind, some aspects of its operation go beyond the scope of this paper. We define these items as non-goals, as listed below:

(1) **Bitcoin's Imperfections.** Bitcoin's design is not necessarily optimal. The concerns regarding Bitcoin include the threat which mining pools pose to Bitcoin's security [29], vulnerability of wallets to common attacks, *e.g.*, phishing attacks [37], and Bitcoin's time interval between blocks, which could be reduced to incorporate transactions at a higher rate [21, 58]. In designing Webcoin, we do not attempt to address these inherited concerns, and we focus on the utilization of Bitcoin's principles in the networking domain.
(2) **Indexing and Crawling Frameworks.** Webpage indexing, and document indexing in general, can be performed in multiple fashions, depending on the desired features of the resulting index, and numerous indexing frameworks are available [6, 14, 22, 33, 34, 44, 45, 51, 53, 54]. However, the design of an index depends on the prioritization among its features, requires ongoing adjustments, and is outside the scope of this paper. We thus refrain from integrating Webcoin with any of the above frameworks, and implement an inverted indexing procedure [19]. Future implementation could integrate with any of the above frameworks.

Similarly to indexing, multiple frameworks could be utilized for scalable and distributed web crawling [23, 36, 43]. Webcoin utilizes Scrapy [12], as in [57], a highly popular and powerful open-source framework for web-scraping. Alternative frameworks, such as [43], might be even better suited for Webcoin, as their capacity to filter on-the-fly content-generating webpages, spam, and "crawler traps" exceeds Scrapy's. Such improvements to the crawling and indexing procedures are outside the scope of this paper.

3 THE WEBCOIN PROTOCOL

The Webcoin protocol is identical to Bitcoin on most aspects; it uses the same messages, the same gossip protocol, the same transactions, and the same blockchain mechanism. Webcoin only differs from Bitcoin in the requirements a newly-mined block must fulfill in order to be accept by nodes. In Bitcoin, a node will only accept a new block if hashing the new block with its predecessor (including its timestamp, nonce, target, and its transactions' Merkle root) yields a small enough value. In Webcoin, each new block is accompanied by an index, and the block will be considered valid only if:

(1) Block is valid.
(2) Index is valid.
(3) Index reported ahead of time.
(4) Index includes webpages of a specific crawl path, determined by miner's IP address, also included in block's header.
(5) The hashing of some fields in the new block, including said IP address, with the fields of previous blocks yields a value smaller than Webcoin's target.

We explain these requirements and the mining process, depicted in Figure 1, below.

3.1 Webcoin's Proof-of-Work

In earlier versions of Webcoin, we have made attempts to use the index as a naïve Proof-of-Work, *i.e.*, to mine a new block, a miner must simply present a valid index, where the hashing of a new block, its predecessor, and the index, produces a value smaller than Webcoin's target. Unfortunately, this approach has two significant drawbacks. First, it only incentivizes the miner who successfully mines a new block to publish the result of his efforts, rather than incentivizing all miners to publish their indices. Second, it opens the door to numerous types of index and block manipulation; rather than invest networking resources to crawl and index the Web, a dishonest miner might produce valid block and index much faster by altering webpages under its control, or by manipulating the transactions included in the block.

To overcome these challenges, the Webcoin protocol uses its Proof-of-Work as a necessary but inefficient condition. In every *round* where miners attempt to mine a new block, each miner must produce a valid index and publish its hash to be eligible to mine a block in the following round (*not* the present round). However, the index by itself does not determine whether a miner will successfully mine the following block, it is merely a prerequisite. In addition to this prerequisite, miners compete on who will be the first to mine the next block and receive its reward, as in Bitcoin, however their competition is quiescent. Each miner waits until such time that the hashing of the current time (in seconds) and the IP address it used to construct its index, as they appear in the header of the new block, and a new header field *anybit*, and the hash of past transactions, as they appear in the headers of past blocks, produce a digest which is smaller than Webcoin's target. We continue to outline here the mining process in its entirety, and provide a detailed explanation of these parameters, their effect on security, and the implications of utilizing multiple IP addresses in Section 5.

Simply put, a Webcoin miner first produces a new Web index and publishes its hashing along with its IP address. Then, in the next round, it awaits until such time it can successfully mine a new block, or more likely, it hears of a new block mined by another miner, accompanied by the index from its miner. Thus, in every round the miner both quiescently tries to mine a new block and actively produces a Web index to be eligible to mine a block in the following round.

Upon receiving a new block, which is accompanied by its miner's index, a miner will verify that the block is valid, that the index is accurate, and that it had been accurately reported in the previous round. To validate the block itself, the miner will validate all transactions are valid, as in Bitcoin. It will also verify the block contains an IP address which matches the index's IP address, and that it had received the index's digest and IP address during the previous mining round. The miner will then test the miner's control over said IP address by sending it a single-packet request for the block's hash, and finally it will statistically validate the index by sampling a small subset of its webpages, a process we detail in Section 4. If block and index are both valid, the miner will accept the block.

Note that the blockchain does not contain any indices. The miner of a new block will send its index alongside its newly-mined block, and they will both propagate throughout the Webcoin network, however the index is discarded soon after the block is accepted.

3.2 Index Collectors

The prerequisite to publish the indices' digests ahead of time incentivizes all miners to produce Web indices and publish their digests. While it is clearly possible to require all miners to publish their indices, rather than the digests, such undiscriminating flood of data is wasteful. Most miners have *no* interest in collecting the indices for their own purposes, and are only interested to receive the index of the miner who successfully mined a new block, in order to validate it. However, Webcoin provides both the means and the incentives to allow interested parties to collect all the indices from all miners.

Nodes interested in collecting indices, *collectors*, *e.g.*, new search engines, existing search engines aiming to cut costs, companies performing analysis of big-data from the Web, and researchers, can gain knowledge of all the indices produced through the propagation of their digests. Collectors may request miners to provide them with their indices, and such collaboration is facilitated based on their respective interests, in a tit-for-tat manner. While collectors have an interest in receiving indices for their own goals, miners have an interest to minimize the time required for their block and index to reach the majority of nodes. Miners are interested in to shorten their propagation time since it reduces the chances for a competing block, mined by a different miner at a similar time, to reach and be accepted by the majority of nodes first, thus to prevent their block from being included in the blockchain. A miner can thus

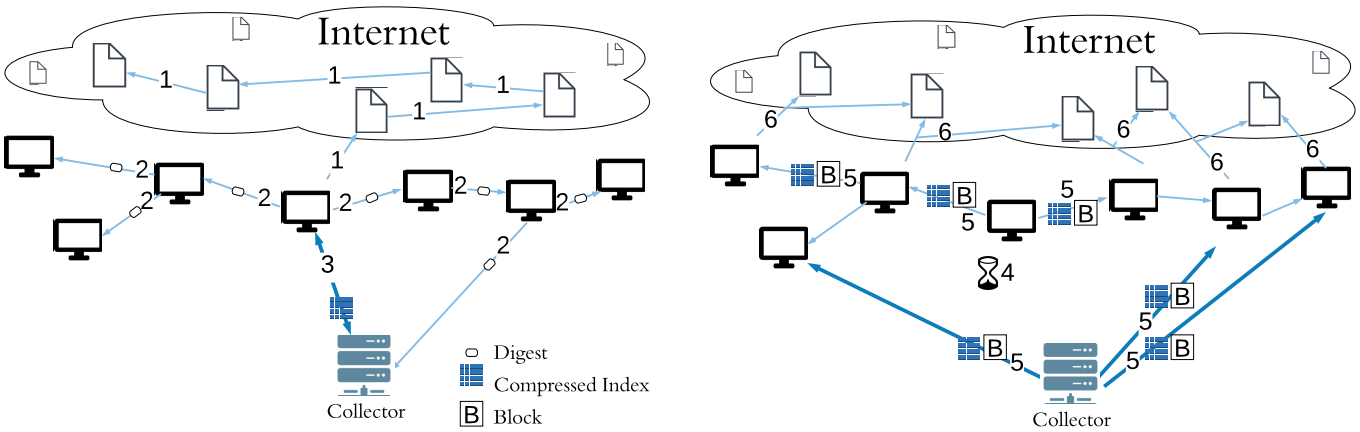

Figure 1: Webcoin mining process: (left) (1) Miner crawls webpages along a directed path, (2) miner indexes webpages, hashes the index, propagates digest to peers, (3) collectors request and receive compressed indices from miners. (right) (4) Miners wait until some miner becomes eligible to mine a new block, (5) eligible miner propagates block and index, accelerated by collectors, (6) each miner verifies the block, compares index to past digest, and verifies a small subset of webpages.

increase its profitability by sending its index before the next round begins.

It is therefore in the best interest of a miner to comply with requests for its index it receives from collectors, so in the event it successfully mines a new block in the following round, the index and block he must present would propagate faster through the network, assisted by the upload bandwidth of the collectors. The cost for a miner to comply with such requests, and send his index in a compressed form to collectors is minimal. Collectors, on the other hand, do not earn Webcoins by participating, the way miners do. However, by assisting with the propagation of a single index and block in each round, collectors get access to all indices, which they can use for their own goals. To be useful to miners, collectors secure sufficient upload capacity and maintain a long list of outbound connections.

4 WEB INDEXING AND INDEX VALIDATION

Each Webcoin miner is required to crawl 1,000 webpages every 10 minutes, a value experimentally determined in Section 6. As the number of miners increases, so does the likelihood of different miners to index the same webpages. To increase Webcoin's efficiency, and to allow for a truly scalable Web indexing, Webcoin limits the possible webpages each miner can crawl and index. Miners must comply with the following restrictions for their index to be considered valid.

First, miners must create their indices using continuous crawls, *i.e.*, their indices must contain one or more series of webpages, each referencing those following it. Each index must specify its crawls' order in a designated data structure. Second, the first webpage of each crawl must be selected from among the indices accompanying the previous two blocks in the blockchain. An additional crawl is only valid if the current crawl had exhausted its webpages. Third, in order for a webpage to be included in a crawl, the hashing of the miner's IP address, the previous block's digest, the webpage's URL, and the URLs of all the webpages which led to it via crawl must be smaller than $\lfloor \frac{2^{256}-1}{10} \rfloor$. Simply put, only 10% of the webpages a miner encounters are valid for it to index, and their validity differs across miners and for different crawl paths. Necessarily, the eligible webpages are *unknown* to a miner ahead of time.

Each index produced by a miner contains the URLs crawled to produce it, the crawl path used, the miner's IP address (which had determined the crawl path), and a hashtable mapping between every word encountered and its appearances in the different URLs. Given the power-law distribution of the Web [40], a collective "Brownian crawling motion" will effectively cover the most prominent portions of the Web, assuming crawlers have the ability to systematically avoid, on-the-fly, content-generating webpages, spam, and "crawler traps" [43].

We note that the Web indexing which Webcoin incentivizes contains a computational aspect, namely, mapping between each word and its occurrences. However, we consider Webcoin's Web indexing to reside mostly in the network domain, rather than the computational, since additional computational resources would not improve a miner's success rate, while additional networking resources would (as we explain in Section 5.3). Additionally, while Web crawling requires 1–2 minutes to perform, the time required for the computational aspect is shorter by an order of magnitude.

4.1 Statistical Index Validation

Validating that the miner of a new block had indeed performed Web crawling and indexing prior to its mining raises several main challenges. One challenge is overcoming the Internet's natural churn. Webpages must remain stable enough to allow their validation for at least several minutes after they have been indexed. Otherwise, an index might be found accurate by the first miners to validate it, and inaccurate by those following. A second challenge is to avoid placing a heavy load on the webpages included in the index, as a large number of clients, *e.g.*, 100K miners, will attempt to contact them over a short period of time, a scenario known as a "flash event" [38]. A third challenge is to reduce the validation time to a minimum, so that even low-bandwidth miners will be able to perform the validation in a timely manner.

To overcome these challenges, Webcoin uses statistical index validation, which requires only the validation of a very small subset of webpages, yet can accurately predict the index validity. To assess webpages churn, we have measured the change rate of over 100, 000 webpages, arbitrarily selected from among the webpages indexed by Webcoin miners in our experiments, detailed in Section 6. We define a change to be any change in the HTML body (or in text for non-HTML webpages) which causes the Web indexing to yield a different result. Thus, changes of HTML tags, whitespaces, and metadata, are not considered to be a change in the webpage. We have found that only 1.97% of webpages change within 4 minutes, and of the remaining, only 0.231% change within the following 30 minutes, when excluding dates and timestamps fields, as well as regular expression string-matching. By allowing indices for inaccuracies in up to 3.5% of their webpages without deeming them invalid, not a single index had failed its validation in our experiments.

We thus relax the accuracy requirements, and require every miner to randomly sample 30 (out of 1000) webpages from the index, download them, and validate they were properly indexed. In the event that up to 1 webpage is inaccurate, the miner will deem the index valid. If more than 6 webpages are inaccurate, it is deemed invalid. If 2–6 inaccurate webpages are found, the process (of randomly sampling new 30 webpages) will be repeated, up to 4 times, after which it will be deemed invalid if 12 or more inaccurate webpages were encountered, and valid otherwise.

4.2 Statistical Validation Accuracy

The use of statistical validation to assess the accuracy of the index accompanying a newly-mined block, rather than requiring every miner to validate every webpage included in said index, reduces the load from both miners and servers. At the same time, it remains infeasible for a valid index to be rejected by the Webcoin network, nor for an invalid block to be accepted. We compute that there is a 71.7% probability that a valid index will be labeled correctly in the first iteration, relieving most miners from doing additional iterations, and thus reducing both the time required and the load on webpages. The probability increases to 92% by the second iteration, and to 99.36% by the fourth. The resulting load for each webpage indexed, for a network of 100K miners, is only 6.9 requests per second, on average, and for a 1 Mbps miner, the validation requires 6.1 seconds, on average.

There is 0.0001% probability that a valid index will be rejected by a miner, and the erroneous miner would amend its mistake when the consecutive block is mined. For a network of 100K miners, the probability for an index to be wrongly rejected by even 0.2% of the network is approximately 10^{-11}, and is expected to occur once every hundreds of thousands of mining years. The probability for an inaccurate index to be validated is very small; for an index which half of its webpages contain inaccuracies, it is infeasible to be accepted by the network, or even by 0.01% of the network. An almost accurate index, which 85% of its webpages are accurately indexed, has a probability of 95.8% to labeled invalid by a miner. This translates into a probability of 10^{-8} to be accepted by 5% of the network. Thus, Webcoin ensures that valid indices are accepted by the *network*, while invalid indices are rejected. Any forks induced at individual nodes are quickly resolved by the mining of the next block.

5 WEBCOIN'S SECURITY

The transition of the cryptographic principles upon which Bitcoin is built into the networking domain is non-trivial. A major difference between the domains is that while the first is governed by a priori mathematical and statistical principles, the latter depends on real-world state, and is thus susceptible to manipulation.

In earlier versions of Webcoin, where we have tested the usability of Web indices as proof-of-work, we have found that the miners' ability to affect the content of webpages opens the door to the manipulation of their indices, in an attempt to improve their mining success rate. Despite numerous counter-measurements tested, we have found it infeasible to guarantee that such manipulation does not take place, as the nodes' level of control over arbitrary Web domains is unbounded. At the same time, miners may attempt to manipulate the blocks they produce, *i.e.*, to add, remove and reorder a block's transactions, to increase their mining success rate. To that end, miners might produce new public keys and transactions between their own wallets.

The manipulation of either webpages or blocks allows dishonest miners to move the task of mining Webcoins back to the computational domain, where they can explore the space of possible adjustments to webpages and transactions permutations. This space is larger by several orders of magnitude, and can be explored faster by several orders magnitude, than the space explored by a miner in the networking domain, *i.e*, the space of crawled webpages. A dishonest miner capable of such manipulations voids all of Webcoin's security guarantees, as it is capable of producing blocks at a higher rate than all other nodes combined, and can thus alter the blockchain at will. Webcoin's mining process thus must answer a novel challenge, consisting of six requirements:

- Webcoin's mining must require the publication of valid Web indices.
- A block's validity must ***not*** rely on said indices.
- A block's validity must not rely on the transactions it contains, nor on any element under its miner's control.
- A block's transactions cannot be selected or manipulated to affect its miner's future success rate.
- A block's validity must rely on transactions of previous blocks, as per Bitcoin.
- Transactions of previous blocks, even if no block's validity is yet to rely on them, must not be susceptible to manipulation.

To address all six requirements stated above, Webcoin employs three unique principles. First, it requires a valid Web index for a block to be mined, yet the success of the process is not determined by its content. Second, the ability of miners to produce new block relies on elements outside their control. Third, Webcoin employs a novel hashing mechanism which breaks the bidirectional relation between blocks; while it allows a block's validity to rely on its predecessors, the miners of said predecessors are prevented from directly affecting the success in mining the blocks succeeding them.

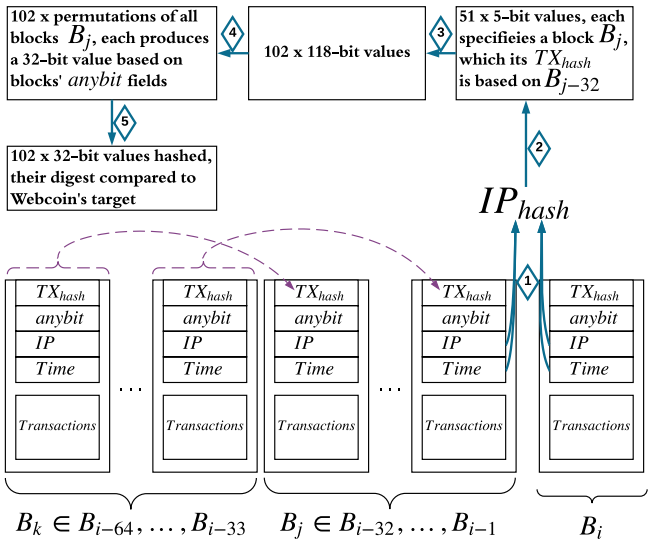

Figure 2: The mining process of a new Webcoin block, performed every second.

5.1 Block Mining

Figure 2 depicts Webcoin's block mining process, which is repeated every second. Webcoin utilizes Bitcoin's SHA-256 double-hashing to hash 4 elements. First, every block B_i contains an additional field, which contains the digest yielded by hashing of the entire block preceding it by 32 blocks (B_{i-32}), denoted TX_{hash}. Second, every block B_i contains an additional 1 bit field, denoted *anybit*, which its value is randomly selected by the miner, and is introduced to counter pre-computation of future blocks. Third, Webcoin utilizes the current time, in seconds, based on the Coordinated Universal Time (UTC) [20]. Fourth, the success of each miner also depends on the hashing of its IP address, and it must prove its ability to communicate through it in order to validate its block. The mining of a new block B_i is as follows.

(1) The IP address of the miner and the current time, in seconds, are hashed with the mining time of B_{i-1} and with its miner's IP address, both specified in B_{i-1}. Denote the result IP_{hash}.
(2) The resulting 256 bit digest is divided into 51 values of 5 bits, discarding the last bit, and each of these values is used as a pointer to a block B_j in the range B_{i-32} to B_{i-1}.
(3) For each block B_j specified by such a 5 bit value, the block's TX_{hash}, which holds the digest yielded from the hashing of block B_{j-32}, is divided into two 118 bit binary values, and the remaining 20 bits are discarded.
(4) Each such 118 bit binary value is translated to a permutation of the 32 blocks in the range between B_{i-32} and B_{i-1}. Using each permutation, the *anybit* fields are extracted from the blocks according to their order, to create a single 32 bit binary value per permutation. Hence, 102 32-bit values are created.
(5) Lastly, these 102 values are hashed according to their order, and the resulting digest is compared to Webcoin's target, which determines whether the miner can produce a new block.

Webcoin's hashing procedure serves multiple purposes. First, from the perspective of B_i's miner, its success depends exclusively on its IP address and the current time. It is thus unable to manipulate the block in order to affect its success rate. Second, from the perspective of every miner of a block

$$B_j \in B_{i-32}, \ldots, B_{i-1},$$

i.e, one of the 32 blocks preceding B_i, it must accurately include the hashing of block B_{j-32} in B_j's TX_{hash} field, otherwise B_j will be rejected by the network. Thus, B_j's miner can only affect the mining of B_i through its selection of *anybit*. This effect is very limited, as it is bounded to two possible values, 0 and 1, without any guarantees that neither will better serve the miner.

In the edge case, a miner M of infinite computational power and perfect system knowledge, attempting to improve its likelihood to mine block B_{j+1}, is twice as likely to select the value to better serve it, in comparison to a coin toss. This does *not* mean the miner will be able to arbitrarily set itself as the miner of B_{j+1}, but, for example, that it will set *anybit* to 1 if it will shorten the period of time until M is eligible to mine B_{j+1}, in comparison to setting it to 0. However, it is extremely likely that a different miner will become eligible prior to M, *for both values*, a fact M cannot change. Furthermore, M's ability to determine the effect of its selection decreases exponentially for blocks beyond B_{j+1}, as their mining depends on *anybit* values not yet determined, until the value of no *anybit* is known for the 32nd consecutive block. From the perspective of the miners of blocks

$$B_k \in B_{i-64}, \ldots, B_{i-33},$$

they hold the largest possibility space to affect the mining of B_i, as their hashing will define the permutations of *anybit* values to be used in the mining process. However, due to Webcoin's novel mining scheme, this possibility space only affects the permutations of value to be used, *prior to the selection of these values, i.e.*, they dictate permutations of *anybit* fields of blocks which have not been mined yet. By design, a block's hashing can only affect the permutation of values of future blocks.

While the mining procedure does not allow manipulation to affect the mining of B_i, it is worth noting that the possibility space it provides through the 102 permutations of 32 bits values exceeds the possibility space of the hash function by more than a hundred orders of magnitude, and thus does not negatively affect the security guarantees. In addition, although a block B_i's transactions do not affect its hashing by design, their hashing is included in B_{i+32}, and affects all consecutive blocks in the blockchain. The requirement for 32 blocks for transactions security is not disproportional, in comparison for the 100 to 120 blocks required before Bitcoin's reward can be used.

In the event that a miner can indeed produce a new block, the miner will report its IP address in a designated field, the timestamp which had produced the new block, the hashing of B_{i-32} in the $TXhash$ field, the resulting digest, and a random bit in the *anybit* field. The miner will then include transactions in the block, and will publish the newly-mined block to the Webcoin network, along with its Web index. While the entire procedure can be done in advance,

the block and index will be rejected by peer miners until the time used for the successful mining had been reached.

5.2 Block Validation

Upon receiving a new block, a miner will ensure that it is indeed valid. Blocks can be easily validated by repeating the above process, using the time and IP address specified in the new block. However, in order to assure that the block's miner had indeed produced the block using an IP address under its control, the validating miner sends a single packet to the IP address over a *designated* port number (Webcoin's current implementation supports one miner per NAT). Furthermore, in order to assure the transactions contained in the block had not been altered, the packet will include the hashing of the entire block, to which the block's miner will reply with a single packet to confirm both that it indeed controls the IP address, and that the transactions had not been altered.

For a block to be accepted as valid, it must be accompanied by a valid index which its miner had reported in advance, as detailed in Section 4. Webcoin thus requires miners to hash their indices and publish only the resulting 256 bit digests and their IP addresses, in order for their blocks to be accepted in the following round. Miners are prevented from reporting a digest in hopes of creating an index to match it at a later time, as finding an input to produce a desired digest is computationally infeasible, an infeasibility upon which Bitcoin's security model relies.

5.3 Resource Aggregation and Security

It is clearly possible for a miner to obtain more than a single IP address, whether legally from ISPs, or illegally, through the use of malicious bots [30]. It is also possible for numerous miners to form large mining pools, and for resource-rich entities, *e.g.*, search engines, to participate in the mining of Webcoin. Here, we consider how these phenomenas affect Webcoin's security model.

Similarly to Bitcoin, the reward of a Webcoin miner is proportional to the portion of resources it controls, and any miner to control over 50% of the resources voids all of Webcoin security guarantees and controls the blockchain. However, unlike Bitcoin, the resources required for Webcoin mining are two-fold: an IP address, and a bandwidth of approximately 1 Mbps. The latter of the two makes Webcoin *more* resilient to resources aggregation than Bitcoin.

Any Webcoin miner wishing to utilize multiple IP addresses to increase its mining capacity will need to control enough bandwidth to support crawling the different paths required for each IP address, consuming approximately 1 Mbps per IP address. However, the bandwidth resource does not scale similarly to Bitcoin's computational resources. For comparison, a Bitcoin miner using a home desktop can increase its capacity by 8 orders of magnitude, by purchasing a single Antminer S9 [1] for $2400, and these machines can be easily stacked. For a similar investment of capital, a Webcoin miner using a residential internet connection can only increase it capacity by 2 orders of magnitude, and its ability to scale it much further is uncertain.

While it is expected for resource-rich entities and large mining pools to participate in Webcoin's mining, as is the case with Bitcoin, they do not jeopardize Webcoin's security model any more than

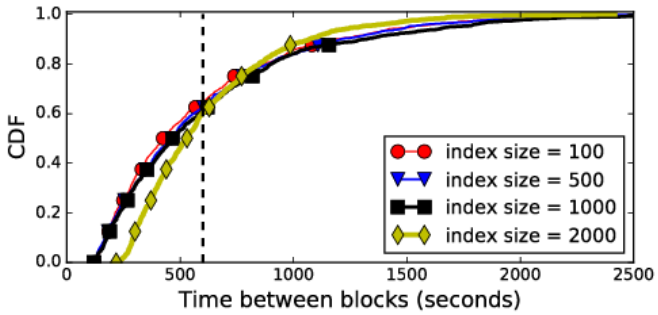

Figure 3: CDF of the number of seconds between the mining of consecutive blocks. Vertical line represents target at 600 **seconds.**

they jeopardize Bitcoin's. All miners are constantly competing over the mining of the next block, and must invest ever-growing resources to remain competitive. The investment of considerable resources by such entities will be met by the investment of other miners, which aim to increase their own share of the reward, which prevents the formation of a monopoly [27].

5.4 Minimal Inter-Block Time

In our experiments, we have found that miners controlling more bandwidth require less time to crawl webpages and to produce a web indices, as to be expected. Since despite a new block is mined every 10 minutes on average, the exact inter-block time varies due to its statistical nature, the inter-block time between some blocks is very short. In such cases, since miners must produce a new Web index prior to the mining of the next block to be eligible to participate in the next round, limited-capacity miners often fail to produce a Web index before the block is mined, and are thus excluded from participating in the following round. To mitigate this phenomena, we add an additional requirement to the quiescent mining competition among miners: a new block's timestamp must exceed its predecessor by at least 180 sec. This requirements prevents a new block to be mined very shortly after its predecessor, which would exclude limited-capacity miners from participation in the next round. We provide additional insight as to this design choice in Section 6.

6 EVALUATION

To evaluate Webcoin's ability to crawl and index the Web, we deploy a network of 200 Webcoin miners on PlanetLab [25]. Our implementation of a Webcoin miner is built on top of Pyminer, a python implementation of a Bitcoin miner [11]. We set miners to commence mining Webcoins using random transactions, using blocks with an average size of 700 KB. Miners utilize download bandwidths ranging between 0.5 and over 300 Mbps, with a median of 19.68 Mbps. Like in Bitcoin, each miner has 8 outbound connections, *i.e.*, 8 peer-miners to which it propagates blocks and indices, which we select randomly from among all other nodes.

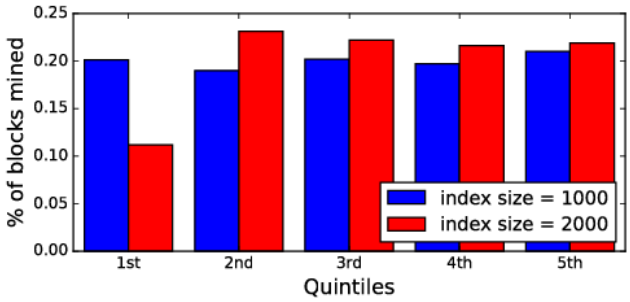

Figure 4: The partitioning of miners to quintiles (20% units) according to their download bandwidth, and the percentage of blocks mined by nodes in each quintile

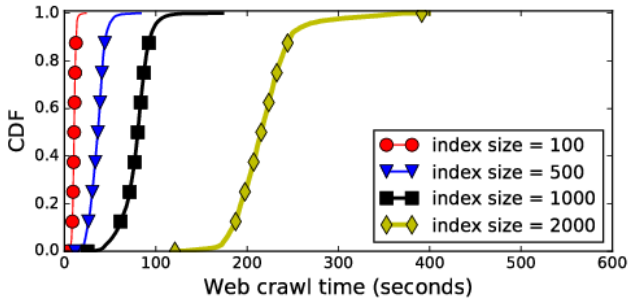

Figure 5: CDF of the number of seconds required for miners to crawl webpages for indexing, per index size.

6.1 Crawling Feasibility

Here, we evaluate miners' ability to crawl the Web and to produce indices of sufficient size in a timely manner, as to be useable in the context of block mining time scales. We utilize 4 index sizes, of 100, 500, 1000 and 2000 webpages per index, and we measure the performance of the miners over a period of one week for each of the index sizes.

The above index sizes were chosen with the desire to allow even miners of limited capacity, *i.e.*, miners utilizing a bandwidth in the order of 1 Mbps, to participate, while still producing Google-scale indices. Indeed, if 100K miners, which is the estimated number of Bitcoin miners [10], were to participate using an index of 100 webpages, the resulting index will reach Google's index size approximately within a month, and if an index of 1000 webpages were to be used, approximately within 3 days.

Figure 3 presents a CDF of the time intervals between the mining of new blocks for each of the index sizes we have tested. Note that identically to Bitcoin, Webcoin adjusts its difficulty target in an attempt to reach an average of 10 minutes, *i.e.*, 600 seconds. It is evident that the index sizes we have tested allow for mining in the desired time scales. The average time intervals between consecutive blocks are 582.75, 608.4, 608.2 and 610.1 seconds, for indices of 100, 500, 1000 and 2000 webpages, respectively. Webcoin thus had missed its target of 600 seconds by up to 17.25 seconds. For comparison, Bitcoin which its large miners community makes its mining time more statistically reliable, regularly misses its target of 600 seconds by similar and larger scales.

Another phenomena which can be observed is the similar distribution of mining times experienced by smaller index sizes, in contrast to that of indices consisting of 2000 webpages. As the larger indices require longer times to produce, mining time as a whole lengthen. To compensate for this lengthening, Webcoin increases its target value, *i.e.*, making mining easier by relaxing the validity requirements, in order to maintain an average mining time of 600 seconds. The combined effect is of a denser distribution function, as the same number of blocks are mined during shorter time interval.

Figure 4 presents the distribution of mined blocks over all miners, which are partitioned to quintiles (20% units) based on their bandwidth, using a time span of 180 seconds in which block mining is prohibited. The distribution of mined blocks of size 1000 (blue) is almost equally distributed, ranging between 19% of blocks mined by 2nd quintile miners, and a maximum of 21% is mined by 5th quintile miners. Similar results are achieved when using smaller index sizes of 100 and of 500 webpages. In contrast, when using an index size of 2000 webpages (red), the 1st quintile miners mine only 11.2% of the blocks, while the miners of the other quintiles mine 23.1%, 22.2% 21.6% and 21.9% of blocks, respectively.

The low success-rate of mining the 1st quintile miners when using an index size of 2000 webpages is explained by the longer time they require to produce such indices, which often causes them to be ineligible to mine the following block. Figure 5 shows the CDF of the crawl times for the 200 miners as a function of the number of webpages crawled. Necessarily, as the number of pages increases, so does the time required for nodes to download them. In the case of an index size of 2000 webpages, we observe that the low-bandwidth miners can spend between 250 and 400 seconds in crawling. This corresponds to the tail shown in the figure for index size of 2000. Hence, such timescales overlap with inter-block timescales shown previously in Figure 3. Whenever the crawling time exceeds the inter-block times, a miner fails to qualify for mining a block in the following round. This explains the reduced performance of low-bandwidth miners for the index size of 2000 in Figure 4. We thus conclude that an index of size 2000 is too large to allow for fair mining. We further conclude that miner equality and performance can be balanced through the use of an index size of 1000, combined with a time span of 180 seconds in which mining is invalid.

6.2 Network Usage

Figure 6 presents the networking usage of a single Webcoin miner during the mining of two consecutive blocks, using an index size of 1000 webpages. The miner's bandwidth, as measured periodically, is approximately 96 Mbps in the download direction and 4.1 Mbps in the upload direction. The figure presents both its download (red) and upload (blue) bandwidth utilization, as well as the events occurring during the mining process. The networking usage plot provides several insights regarding the dynamics of Webcoin mining, and the resources it consumes. First, we note the seven events which consists the mining of a single block, which are divided into 3 phases, with very different characteristics.

The first phase starts with the mining of a new block by a peer miner (denoted $Line1$), which indicates the beginning of the 180

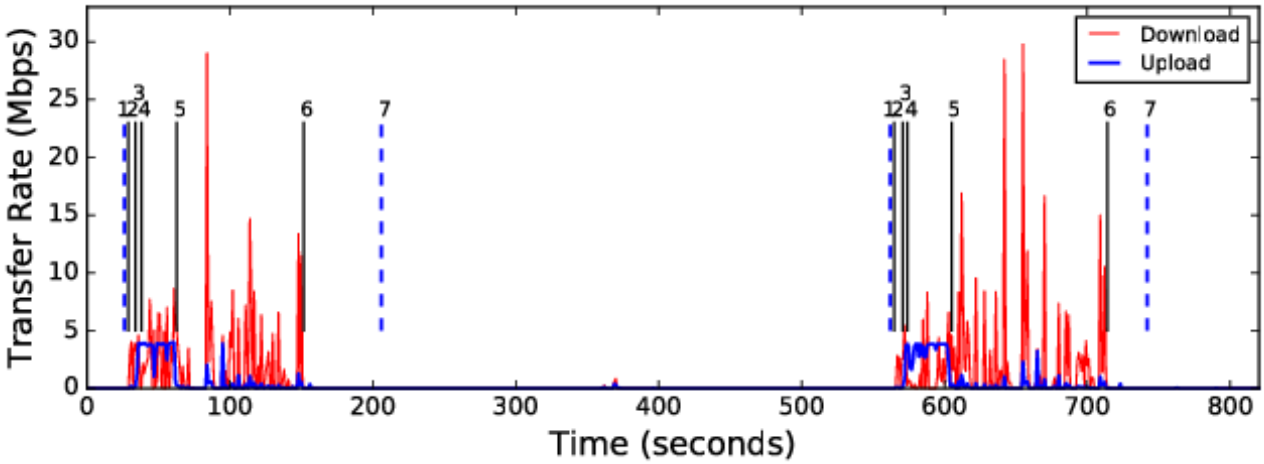

Figure 6: Network usage of a Webcoin miner deployed on a PlanetLab node, during the mining of two consecutive blocks utilizing an index size of 1000 webpages. Dashed lines mark the periods in which mining is prohibited (between 1 and 7). Numbers mark the events of the mining process: (1) block is mined by another miner, (2) block and index download from peers, (3) block and index propagation on the upload direction, index validation on the download direction, (4) Web crawling and indexing, (5) block and index propagation completion, (6) Web crawling completion, (7) block mining begins.

seconds time span in which blocks cannot be mined. The block's mining is closely followed by the arrival of the block and its index (denoted *Line*2), and upon the block's successful validation, the propagation of both to the miner's peers, using its upload bandwidth (denoted *Line*3). At the same time, the miner utilizes its download bandwidth to validate the index he had received, its completion denoted *Line*4. These events, which make up the first phase of the mining, occur in a rapid succession, and are completed within 12 seconds from the mining of the new block. Note that the bandwidth requirements for this phase, including the statistical validation of the new index, are very limited, under 5 Mbps.

In the second phase, with the successful validation of the index (*Line*4), the miner begins its Web crawling and indexing on the download direction, while in parallel it continues to propagate the block and the index it had received to its peers. Said propagation is completed at *Line*5, and the Web crawling and indexing at *Line*6, marking the end of the second phase. The second phase is considerably longer than the first, and it is there where the miner invests its networking resources to produce a new Web index. However, even in this more network-intensive phase the miner utilizes only a very small portion of its bandwidth. Indeed, despite a download bandwidth of almost 100 Mbps, the miner surpasses the use of 10 Mbps only thrice during the mining of the first block presented, and similar characteristics can be observed during the mining of the consecutive block. This pattern allows for miners of low capacity to compare on even footing with miners of much greater capacity.

In the third phase, upon completing the Web crawl (*Line*6), the miner propagates the hashing of the index to its peers, and then awaits for the mining to resume with the passing f 180 seconds from the mining of the previous block, denoted *Line*7. Following this, the miner will quiescently await until such time that a new block is mined, which occurs at 563 seconds, denoted *Line*1. This phase requires very little resources, as it consists mostly of idle waiting for a new block to be mined.

It is essential to understand that the index hash, published by the miner at *Line*6, at time 152 seconds, makes the miner eligible to compete in the *next* round, the one for which mining starts at *Line*7 at 743 seconds. A miner thus has a sufficient time to crawl the Web and propagate its index hash. Observing the network usage as a whole, it can be seen that the networking resources needed for participation are meager.

6.3 Mining at Large Scales

We consider the scale's effect on Webcoin's ability to produce a new block every 10 minutes, for a Bitcoin-scale network of 100K miners, and the ability of a miner with a capacity of 1 Mbps bandwidth (both upload and download) to participate on equal grounds, despite increased volume of traffic.

Upon the mining of a new block, said block is propagated to all miners, followed by its index. According to Bitcoin measurements in [4, 5], the average time required for a block to reach the majority of nodes, over a 9 week period, was 5.9 seconds, and the average block size had been 700.7 KB. As Webcoin's blocks are almost identical to Bitcoin's, and are of the same size, their propagation across a large network of 100K miners is expected to be identical. As Webcoin's blocks are followed by their indices, which are approximately a half of a block's size on average, yet may reach as much as twice the blocks' size, we estimate their combined propagation to the majority of nodes to be less than 20 seconds, on average.

During the propagation, every miner to receive the block queries the block's miner, to verify it indeed controls the IP address used for mining. In the event that the block's miner has a bandwidth of 1 Mbps, it is likely to become a bottleneck. Each such request contains 16 Bytes of the block's digest, yet together with TCP/IP

headers the load totals to 2.67 MB for the majority of requests. A low capacity miner will require approximately 21.3 seconds to respond to them. However, as such requests start to arrive to the miner almost immediately after the new block's mining, this time is incurred in parallel to the propagation time, and the completion of both is expected within approximately 25 seconds from the block's publishing time.

Following the block validation and the propagation of both block and index, each miner statistically validates the index, which for a low capacity miner will require 6.1 seconds, on average. Thus, a new block and its Web index can be propagated and validated by the majority of the system, within approximately 30 seconds from the new block's mining, for a network of 100K miners.

The last factor which might affect scalability is the propagation of all the indices' digests, sent for every Web index produced by the system. Whereas the digests are only 32 Bytes long, they incur an overhead which doubles their size when sent individually over TCP. They are thus aggregated by miners prior to their transmission. For a network of 100K miners, each miner must download 3.05 MB of hashes during a time span of 10 minutes, on average, which consumes only 0.04 Mbps of the miner's bandwidth. We conclude that, in addition to empirical evidence, the analysis of the resources required to mine Webcoins in a network of 100K miners shows that even limited-capacity miners can participate on equal grounds in Webcoin's mining.

6.4 Collectors' Properties.

For a collector to be capable of collecting all the Web indices produced by a network of 100K miners, it must be capable of downloading them in the timespan required to mine a new block, *i.e.*, 10 minutes on average. This is because miners have an incentive to publish their index only as long as there is a chance they will need to propagate it, in the event of successful mining in the next round. While miners must download 20.45 MB of data, on average, to create an index of 1000 webpages, the compressed index itself is approximately 2 orders of magnitude smaller. Thus, for a collector to collect all the indices produced by a network of 100K miners, it must download 20 GB in the span of 10 minutes, *i.e.*, it requires only a consumer-grade bandwidth of 266 Mbps.

Collectors may use the indices they collect from miners for any purpose, yet in the context of Webcoin's effort to democratize Web search, it is worth noting that the indices they collect must be merged and aggregated to enable a competitive real-world web search. Identifying the best methods to construct a large-scale index from these indices is outside the scope of this work.

7 RELATED WORK

Guided Tour Puzzle [17] is a proof-of-work in the networking domain, aiming to filter the originators of spam emails from legitimate senders, a task for which traditional proof-of-work had been proved inadequate [42]. Rather than requiring computational power, the puzzle requires a node to sequentially retrieve information from multiple servers in order to successfully create its proof-of-work. While successfully reducing the performance variance among nodes, the Guided Tour Puzzle (*i*) is centrally managed, (*ii*) utilizes dedicated servers, (*iii*) does not support any additional purpose such as Web search, and (*iv*) does not support any distributed consensus similar to crypto-currencies.

The authors of [35] suggest incorporating Internet transactions, *e.g.*, BGP advertisements, in a blockchain to distribute the Internet management, and a similar suggestion to utilize a blockchain to store DNS entries is suggested in [18]. Yet, they do not change any of the Bitcoin mechanisms, but simply use Bitcoin for a different type of transactions. Webcoin fundamentally alters Bitcoin mechanisms by utilizing networking resources.

The authors of [59] suggest DDoSCoin, a blockchain which uses a malicious Proof-of-Work, where miners are rewarded for participating in a DDoS attack on websites using Transport Layer Security (TLS), and must present a signed response from the server which is smaller than DDoSCoin's target. While this work provides the incentive to perform a malicious network task, the DDoSCoin's Proof-of-Work is computational in nature, generated by the targeted website. Webcoin, in contrast, can be used for constructive purposes, and its Proof-of-Work does not require an objective computational component to fairly reward miners for the resources they have invested.

Several "alt-coins" aimed to become ASIC-resistant. First among these solutions, is the use of Scrypt [52] as a hashing function, rather than Bitcoin's SHA-256 [32]. Scrypt requires frequent memory access, instead of relying solely on computational power [28]. Unfortunately, the first generation of scrypt-mining ASIC machines have been introduced by multiple vendors during 2014 [2, 55]. Webcoin utilizes a novel block mining algorithm which enables miners with moderate network capacities to participate.

8 DISCUSSION

Practical Challenges in Crawling and Indexing. While Webcoin is the first primitive designed to crowdsource complex networking tasks, and the first digital currency to incentivize open Web indexing, it must surmount several significant challenges in order to disrupt the Web search market.

First, Modern webpages are complex; they contain dependancies among their components, and consists of dynamic content, images, video clips, and code scripts. The challenge in crawling and indexing them in a consistent manner across all miners is significant, and it must be resolved for Webcoin to enable new high-quality search engines. Second, Webcoin must overcome crawling permissions conflicts. Webpage administrators aim to allow search engines to access to their content, while denying access from others who might take advantage of their content, *e.g.*, making their content freely available and harm their revenues. While known search engines' crawlers can be identified based on their IP addresses, the same cannot be easily done for Webcoin. For Webcoin to gain traction, it should incentivize webpage administrators to allow miners to crawl their webpages, possibly by improving webpage visibility and by providing proof of past Webcoin blocks they produced.

Another real-world challenge is to prevent firewalls from interfering with Webcoin's operation, as is often the case with peer-to-peer communication patterns. This matter would have to be addressed in order for Webcoin to gain traction. The fashion in which webpages to be crawled are selected can also be improved upon, as more popular webpages should be crawled and indexed more

frequently than less popular ones. It's likely possible to improve the indices Webcoin produces by utilizing Webcoin's blockchain to reach a consensus regarding the popularity of domains and webpages, which would affect the frequency at which they are crawled.

IP Addresses. Webcoin uses IP addresses to direct miners on different crawl paths. While miners can use multiple IP addresses, and they may have some level of control over their IP address, each address would direct the miner to crawl a different path, and a miner would require a bandwidth of roughly 1 Mbps per IP address, which does not easily aggregate by the same orders of magnitude. Thus, the use of IP addresses does not jeopardize Webcoin's security. However, We believe Webcoin would benefit from avoiding any reference to miners' IP addresses in the mining process altogether, as they make miners susceptible to DDoS, spoofing, and BGP hijacking attacks. In addition, the usage of IP addresses allows for only a single miner behind each NAT, which is a significant limitation. We are actively exploring changes to the mining process which would remove this dependency on miners' IP addresses.

Additional Networking Tasks. Webcoin's goal is to incentivize miners to perform Web crawling and indexing to democratize the Web search market. As such, its attributes were chosen to allow even miners with limited capacity to perform these tasks, and to minimize the load on miners and servers. However, Bitcoin's incentive mechanism can be similarly utilized for a wide range of networking tasks. For example, miners can be incentivized to measure latencies and routing paths towards a changing subset of miners, where each miner requests for signed proofs from the miners he's targeting, and the "winner" miner must present these proofs in order for its block to be considered valid. Combining traceroute and reverse-traceroute, and utilizing a concept similar to collectors for entities interested in such measurements, *e.g.*, CDNs, VPNs, and ISPs, can provide additional accuracy and reliability.

Diverging from Bitcoin's Proof-of-Work. While we have made every attempt to keep Webcoin as close to Bitcoin as possible, the need to prevent Webcoin's Proof-of-Work from affecting a miner's success rate had forced us to make adjustments. While some adjustments, *e.g.*, miners' quiescent mining competition, have equivalences in the Proof-of-Stake domain, others, such as determining the "winner" miner based on arbitrary attributes of multiple past blocks, are unique to Webcoin. Most notably, while the correctness of Bitcoin's Proof-of-Work is not time-sensitive, the correctness of Webcoin's Proof-of-Work is temporal due to the Web's ever-changing nature. As such, Webcoin's consensus requires witnessing blocks as they come, which is a substantial divergence from Bitcoin's design. It is worth noting that Webcoin still requires the majority of mining power to alter the blockchain, however it appears that introducing "check-points" as in PoS, where the state of the blockchain is finalized periodically, will improve Webcoin's security and robustness.

Additional Network Load. As Webcoin miners crawl the Web and report their indices to collectors, they increase the load on the servers hosting said webpages, and on the networks which connect them. This load is similar to the load of every other search engine's crawler, centralized or decentralized, and we thus consider such load reasonable, similarly to the legitimacy of an additional search engine. We acknowledge the need for additional real-world measurements at a larger scale to assess how the size of the Webcoin network affects the load on hosting servers, and possibly to optimize the parameters of the index statistical validation accordingly.

Economical Incentives. Webcoin's economic model follows the footsteps of other single-purpose blockchain-based ecosystems; the value of Webcoins is derived both from its usage and from its exchange rate [31, 49, 56, 59]. As the importance of online privacy and the drawbacks of ad-based revenue models become apparent in mainstream media [8, 15], the use of Webcoins can appeal to many who value their privacy. Incorporating "hot wallets" in Web browsers, similarly to Google's recent patent [41], can seamlessly make micro-payments to online services in Webcoins, which in turn increases their demand, increases their price, and incentivizes Webcoin miners. Users can thus attain Webcoins through mining or exchange, and search engines and similar online services can offer their services either under the existing ad-based model, or as seamless "almost free" service.

More radical and complex models, where online services can sign long-term contract with miners to buy Webcoins at a discount, and sell these Webcoin to users, allowing both miners and services to profit, or forgoing user payments altogether and require users to provide a cryptographic proof of holding some amount of Webcoins instead, which similarly increases Webcoin's demand, are also possible. While we have designed Webcoin as a system in and by itself, it might be expanded to other networking tasks to create a more robust ecosystem of online services.

9 CONCLUSIONS

We presented the first primitive to use Bitcoin's incentive model to crowdsource complex networking tasks, which are difficult to perform or estimate, and through that, to disrupt key online industries, such as Web search, cloud services, CDNs, and ISPs. We presented Webcoin, a novel distributed digital currency which can only be mined through Web indexing, and which provides both the means and incentives for *open* large-scale Web mining. While Webcoin provides the same security guarantees as Bitcoin, it is the first primitive to use networking resources rather than computational. This transition had introduced numerous challenges with respect to feasibility, scale, and security, which we addressed.

We showed that a Bitcoin-size Webcoin network, which requires its miners to index 1000 webpages in every block cycle, creates Google-scale indices within approximately three days. We further showed that parties interested in Web indices can download them from *all* Webcoin miners, using a consumer-grade bandwidth. Webcoin ensures the index *quality* via statistical index verification that minimizes the Webcoin network's validation traffic. Webcoin not only removes the catastrophic energy footprint associated with Bitcoin, but incentivizes miners to conduct useful work, thus opening novel sustainable crowdsourcing avenues.

We conclude by emphasizing that Webcoin does not disrupt the Web search industry by itself. Rather, it provides the key prerequisite for such disruption to occur: an open access to a high-quality global Web index for all to study, experiment, and use.

ACKNOWLEDGMENTS

This project is supported by the National Science Foundation (NSF) via grant CNS-1810582.

REFERENCES

[1] Antminer S9 ASIC Bitcoin Miner. http://www.bitmain.com/.
[2] ASICS for Litecoin. Here They Come. https://bitcoinmagazine.com/articles/asics-litecoin-come-1394826069/.
[3] Bitcoin Energy Consumption Index - Digiconomist. https://digiconomist.net/bitcoin-energy-consumption.
[4] Bitcoinity.org. http://www.data.bitcoinity.org.
[5] Bitcoinstats. http://www.bitcoinstats.com.
[6] Compass. www.compass-project.org/.
[7] Faroo Search Engine. http://www.faroo.com/index.en.html.
[8] GOOGLE'S PRIVACY WHIPLASH SHOWS BIG TECH'S INHERENT CONTRADICTIONS. https://www.wired.com/story/googles-privacy-whiplash-shows-big-techs-inherent-contradictions/.
[9] Majestic 12 Search Engine. https://www.majestic12.co.uk/about.php.
[10] Neighborhood Pool Watch. http://www.organofcorti.blogspot.mx/.
[11] Pyminer. http://www.github.com/jgarzik/pyminer.
[12] Scrapy. http://www.scrapy.org.
[13] Steem - An incentivized, blockchain-based, public content platform. www.steem.io/SteemWhitePaper.pdf.
[14] YaCy Search Engine. http://www.yacy.net.
[15] Yes Facebook is using your 2FA phone number to target you with ads. https://www.techcrunch.com/2018/09/27/yes-facebook-is-using-your-2fa-phone-number-to-target-you-with-ads/.
[16] Operating system market share, 2017. http://www.netmarketshare.com/operating-systemmarket-share/.
[17] Abliz, M., and Znati, T. A Guided Tour Puzzle for Denial of Service Prevention. In *Computer Security Applications Conference, 2009. ACSAC'09. Annual* (2009), IEEE, pp. 279–288.
[18] Ali, M., Nelson, J. C., Shea, R., and Freedman, M. J. Blockstack: A global naming and storage system secured by blockchains. In *USENIX Annual Technical Conference* (2016), pp. 181–194.
[19] Anh, V. N., and Moffat, A. Inverted Index Compression using Word-Aligned Binary Codes. *Information Retrieval 8*, 1 (2005), 151–166.
[20] Arias, E., and Guinot, B. Coordinated Universal Time UTC: Historical Background and Perspectives. In *Journees systemes de reference spatio-temporels* (2004).
[21] Barber, S., Boyen, X., Shi, E., and Uzun, E. Bitter to Better – How to Make Bitcoin a Better Currency. In *International Conference on Financial Cryptography and Data Security* (2012), Springer, pp. 399–414.
[22] Bender, M., Michel, S., Triantafillou, P., Weikum, G., and Zimmer, C. Minerva: Collaborative P2P Search. In *Proceedings of the 31st international conference on Very large data bases* (2005), VLDB Endowment, pp. 1263–1266.
[23] Boldi, P., Codenotti, B., Santini, M., and Vigna, S. Ubicrawler: A Scalable Fully Distributed Web Crawler. *Software: Practice and Experience 34*, 8 (2004), 711–726.
[24] Bosch, A., Bogers, T., and Kunder, M. Estimating Search Engine Index Size Variability: a 9-Year Longitudinal Study. *Scientometrics 107*, 2 (2016), 839–856.
[25] Chun, B., Culler, D., Roscoe, T., Bavier, A., Peterson, L., Wawrzoniak, M., and Bowman, M. Planetlab: an Overlay Testbed for Broad-Coverage Services. *ACM SIGCOMM Computer Communication Review 33*, 3 (2003), 3–12.
[26] De Kunder, M. The Size of the World Wide Web. *WorldWideWebSize* (2012).
[27] Dimitri, N. Bitcoin mining as a contest. *Ledger 2* (2017), 31–37.
[28] Dziembowski, S. Introduction to Cryptocurrencies. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (2015), ACM, pp. 1700–1701.
[29] Eyal, I., and Sirer, E. G. Majority is Not Enough: Bitcoin Mining is Vulnerable. In *International Conference on Financial Cryptography and Data Security* (2014), Springer, pp. 436–454.
[30] Feily, M., Shahrestani, A., and Ramadass, S. A Survey of Botnet and Botnet Detection. In *Emerging Security Information, Systems and Technologies, 2009. SECURWARE'09. Third International Conference on* (2009), IEEE, pp. 268–273.
[31] Freund, A., and Stanko, D. The wolf and the caribou: Coexistence of decentralized economies and competitive markets. *Journal of Risk and Financial Management 11*, 2 (2018), 26.
[32] Gilbert, H., and Handschuh, H. Security Analysis of SHA-256 and Sisters. In *International Workshop on Selected Areas in Cryptography* (2003), Springer, pp. 175–193.
[33] Gormley, C., and Tong, Z. *Elasticsearch: The Definitive Guide.* " O'Reilly Media, Inc.", 2015.
[34] Grainger, T., Potter, T., and Seeley, Y. *Solr in Action.* Manning, 2014.
[35] Hari, A., and Lakshman, T. The Internet Blockchain: A Distributed, Tamper-Resistant Transaction Framework for the Internet. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks* (2016), ACM, pp. 204–210.
[36] Isele, R., Umbrich, J., Bizer, C., and Harth, A. LDspider: An Open-Source Crawling Framework for the Web of Linked Data. In *Proceedings of the 2010 International Conference on Posters & Demonstrations Track-Volume 658* (2010), CEUR-WS. org, pp. 29–32.
[37] Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. Social Phishing. *Communications of the ACM 50*, 10 (2007), 94–100.
[38] Jung, J., Krishnamurthy, B., and Rabinovich, M. Flash Crowds and Denial of Service Attacks: Characterization and Implications for CDNs and Web Sites. In *Proceedings of the 11th international conference on World Wide Web* (2002), ACM, pp. 293–304.
[39] King, S., and Nadal, S. Ppcoin: Peer-to-Peer Crypto-Currency with Proof-of-Stake. *self-published paper, August 19* (2012). www.wallet.peercoin.net/assets/paper/peercoin-paper.pdf.
[40] Kong, J. S., Sarshar, N., and Roychowdhury, V. P. Experience versus talent shapes the structure of the web. *Proceedings of the National Academy of Sciences 105*, 37 (2008), 13724–13729.
[41] Langschaedel, J., Armstrong, B. D., and Ehrsam, F. E. Hot wallet for holding bitcoin, Sept. 17 2015. US Patent App. 14/660,418.
[42] Laurie, B., and Clayton, R. "Proof-of-Work" Proves Not to Work; version 0.2. In *Workshop on Economics and Information, Security* (2004).
[43] Lee, H.-T., Leonard, D., Wang, X., and Loguinov, D. IRLbot: Scaling to 6 Billion Pages and Beyond. *ACM Transactions on the Web (TWEB) 3*, 3 (2009), 8.
[44] Leela, K., and Haritsa, J. SphinX: Schema-Conscious XML Indexing, 2001. http://dsl.cds.iisc.ac.in/pub/TR/TR-2001-04.pdf.
[45] McCandless, M., Hatcher, E., and Gospodnetic, O. *Lucene in Action: Covers Apache Lucene 3.0.* Manning Publications Co., 2010.
[46] Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G. M., and Savage, S. A Fistful of Bitcoins: Characterizing Payments Among Men with No Names. In *Proceedings of the 2013 Conference on Internet Measurement Conference* (2013), ACM.
[47] Miller, A., Juels, A., Shi, E., Parno, B., and Katz, J. Permacoin: Repurposing Bitcoin Work for Data Preservation. In *2014 IEEE Symposium on Security and Privacy* (2014), IEEE, pp. 475–490.
[48] Nakamoto, S. Bitcoin: A Peer-to-Peer Electronic Cash System, 2008. www.bitcoin.org.
[49] of Money Research Collaborative:, F., Nelms, T. C., Maurer, B., Swartz, L., and Mainwaring, S. Social payments: Innovation, trust, bitcoin, and the sharing economy. *Theory, Culture & Society 35*, 3 (2018), 13–33.
[50] Park, S., Pietrzak, K., Alwen, J., Fuchsbauer, G., and Gazi, P. Spacecoin: A Cryptocurrency Based on Proofs of Space. Tech. rep., IACR Cryptology ePrint Archive, 2015: 528, 2015.
[51] Parreira, J. X., Donato, D., Michel, S., and Weikum, G. Efficient and Decentralized Pagerank Approximation in a Peer-to-Peer Web Search Network. In *Proceedings of the 32nd international conference on Very large data bases* (2006), VLDB Endowment, pp. 415–426.
[52] Percival, C. Stronger Key Derivation via Sequential Memory-Hard Functions. *Self-published* (2009), 1–16. www.bsdcan.org/2009/schedule/attachments/87_scrypt.pdf.
[53] Pujol, J., and Rodriguez, P. Porqpine: a Distributed Social Search Engine. In *In Proc. of WWW'09* (Madrid, Spain, Apr. 2009).
[54] Sankaralingam, K., Sethumadhavan, S., and Browne, J. C. Distributed Pagerank for P2P Systems. In *High Performance Distributed Computing, 2003. Proceedings. 12th IEEE International Symposium on* (2003), IEEE, pp. 58–68.
[55] Serapiglia, A., Serapiglia, C. P., and McIntyre, J. Crypto Currencies: Core Information Technology and Information System Fundamentals Enabling Currency without Borders. *Information Systems Education Journal 13*, 3 (2015), 43.
[56] Swartz, L. Blockchain dreams: Imagining techno-economic alternatives after bitcoin. *Another economy is possible* (2017), 82–105.
[57] Wang, J., and Guo, Y. Scrapy-Based Crawling and User-Behavior Characteristics Analysis on Taobao. In *Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2012 International Conference on* (2012), IEEE, pp. 44–52.
[58] Wood, G. Ethereum: A Secure Decentralised Generalised Transaction Ledger. *Ethereum Project Yellow Paper* (2014).
[59] Wustrow, E., and VanderSloot, B. Ddoscoin: Cryptocurrency with a malicious proof-of-work. In *Proceedings of the 10th USENIX Workshop on Offensive Technologies* (2016), USENIX Association.