Information Services & Use 38 (2018) 141–148
DOI 10.3233/ISU-180017
IOS Press

Stable and decentralized? The promise and challenge of a shared citation ledger

Richard Ford Burley[a,b,*]
[a]*Managing Editor, Ledger, University of Pittsburg Press, Pittsburg, USA*
[b]*Boston College, Boston, USA*

Abstract. Current citation indexes such as Scopus and the Web of Science rely on centralized curation to function. This creates a series of potentially negative incentives, both financial and academic, which could be addressed through the creation of an unpermissioned, distributed ledger of citations. This paper explores this possibility using Bitcoin as a model.

Keywords: Citation index, unpermissioned ledger, Garfield's Law, Bradford's Law, Bitcoin, Blockchain technology

1. Introduction

This is a brief paper exploring stability, decentralization, and the potential benefits and challenges of using blockchains - or, more generally, of using blockchain technology as a model - for creating a shared, distributed citation ledger [1]. The question mark in the title is a giveaway of sorts: "stable" and "decentralized" are two things that do not tend to go hand-in-hand. The more voices one adds to a discussion, the harder it becomes to have a coherent conversation. Furthermore, this paper does not end in the release of a fully-functional citation ledger; rather than being an exploration of how to set one up, it is about whether we might want to - and if we do, whether we even can.

First, I will discuss some of the features and drawbacks of centralized databases of citations, such as Clarivate Analytics's "Web of Science", Elsevier's "Scopus", or Google Scholar. I will address the double-edged sword of curation and the way it impacts the overall look and usefulness of these services. Next, I will turn to incentives. The "real genius" of Bitcoin was not, as many think, in getting away from the big banks to create a radically libertarian economy; rather, it was in the realignment and leveraging of incentives to create a system that was both distributed and stable. I will therefore look at the incentives structure that Bitcoin uses, and the way it uses that incentive structure to counter some of the more well-known attack vectors on the distributed system. Lastly, I will return to the idea of a shared citation ledger and discuss whether it is possible to leverage - or at least mitigate - the incentives already present within academia to stabilize a decentralized ledger of citations - with or without leaning on another, pre-existing shared ledger.

*E-mail: richardfordburley@gmail.com.

2. Citation indexes

In 1964, Eugene Garfield returned to the idea of the citation index that he had first proposed a decade earlier. "Whether or not citation indexes are useful", he wrote, "is a question that has now been answered…However, a citation index must meet the same economic test that all products in our society must meet: Does the cost justify the benefits?" [2]. Even in the age of the internet, as we edge closer and closer to the "world brain" that he imagined at that time, the question of economics still shapes the choices that we make.

Citation indexing is not an inexpensive venture in 2018. Elsevier, responsible for "Scopus", is a roughly $3.3 billion dollar company that publishes nearly half a million articles a year in 2,500 journals worldwide [3]. Clarivate Analytics - responsible for the "Web of Science" - was valued at over $3.5 billion when Onex and Baring Asia bought it from Thomson Reuters last year [4]. When the *up-and-comer* in an industry is owned and operated by Google (i.e. Google Scholar), one can be fairly certain that the industry in question has become "big business".

But despite the resources available, or perhaps because of them, important decisions on what is and is not indexed are based in many ways on economic considerations - though this often goes without saying. In the ironically-titled "Why Be Selective?" section of James Testa's "Journal Selection Process" essay for Web of Science, for example, he cites both Bradford's Law and Garfield's Law to justify the index's judicious selection *process*, rather than the act of being selective itself [5]. Bradford's Law of Scattering [6], when combined with Garfield's Law of Concentration [7], does indicate that a majority of a given index's subscribers' needs can be met with a solid core of journals; but even so, Testa never comes out and actually says why being selective is a choice they are making. Thankfully Garfield outlines it himself when he writes that "any abstracting or indexing service that ignores Bradford's Law in attempting to realize the *myth* of complete coverage does so at its great financial peril" (emphasis mine) [8]. Because of financial reasons, "complete coverage" is a "myth". Infrastructure, hosting services, curation, sales - the centralized citation index business model requires funding for all of them. This not only means they have to be selective, it means they are forced to charge for access.

This system leaves the act of curation in the hands of a small number of people, who themselves are subject to some not insignificant market demands (and ones that may not always align with the interests of the academy). Take the Emerging Sources Citation Index (ESCI), which Web of Science uses to expand its coverage to newer sources. It has as one of its criteria for consideration "recommendation or request for coverage by Web of Science users. Journals of particular importance to Web of Science users", it continues, "are given the highest priority in evaluation and selection for ESCI". This is to say that a market demand for a new journal by those paying for the index (the subscribers) drives the decision to include or not include new publications. This, unfortunately, pushes financial pressure downstream onto startup and independent publishers who, in addition to finding funding sources to sustain operations, must also advertise in order to generate the demand to be indexed. It also effectively writes off short-run and more obscure publications, whose quality may be scholarly, but whose reach is too limited to create effective market demand for indexing. It is also worth pointing out at this juncture that this is likely to reproduce existing power dynamics in academia: in situations where funding for advertising is short, word-of-mouth networks tend to be used in their stead, and these tend to privilege straight white men.

This system of incentives reveals a challenge that appears as indexes play greater and greater roles as gatekeepers to scholarship: an index which relies upon Garfield's Law of Concentration too much will end up making that concentration greater over time, by driving academic contributions to larger, more- or better-indexed publications. In effect, it turns Garfield's Law into a kind of self-fulfilling prophecy.

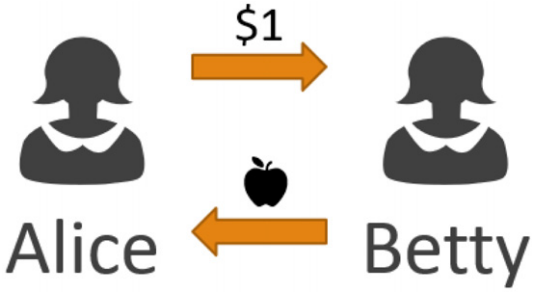

Fig. 1. Alice wants to buy an apple from Betty, Alice gives Betty a dollar and Betty gives Alice an apple.

This is not to say that these issues cannot be mitigated from within a given centralized citation indexing system, but rather to merely point out that the centralized system itself is expensive and subject to market forces that may diverge from the academy's interest in a greater plurality of scholarship from more diverse voices.

3. Incentives

The question thus arises of whether a decentralized citation ledger could be used to address these challenges. As the title of this paper suggests, the most pressing issue would be stability. Regardless of any fault we might find with the centralized citation ledgers, they have the very real benefit of being stable. We do this often, as humans: when there are a large number of us with competing interests, we decide on (or have decided for us) a small group of people who we are going to trust, and hope that they will adequately represent our interests. This is how democracy works, how banks work (though they have more- or less-democratic checks and balances in place), and how centralized citation networks work: we trust someone to be the custodian of the collective interest.

This is where Bitcoin offers a solution, because it is a *trustless* decentralized system. Unlike a traditional currency, Bitcoin does not have physical or even digital objects. Among the greatest of ironies concerning Bitcoin is the fact that there are no actual "coins". Instead, there is a shared ledger, a single, continually-updating document that everyone has a copy of that says what Bitcoin balances are at what accounts. With traditional currencies (i.e. dollars), if Alice wants to buy an apple from Betty, Alice gives Betty a dollar and Betty gives Alice an apple (Fig. 1).

In Bitcoin and other shared ledgers, there are no coins to give in exchange for the apple. Instead, Alice adds an entry to the shared ledger saying some of the Bitcoin she used to control is now controlled by a different account, an account controlled by Betty. Everybody updates their copy of the ledger, and Alice gets an apple (Fig. 2).

The problem is, of course, making sure everyone updates their ledgers in the same way at the same time, which is where "miners" and their "proof of work" come in.

When Alice and Betty create a new transaction to add to the shared ledger, it goes into a list that can be seen by a group of people called miners (Fig. 3). Miners are the ones responsible for adding new transactions to the shared ledger. These miners spend a great deal of computational power - which is to say electricity, and therefore money - to guess a number. Guessing that number is called "solving a block", and the first miner to guess it "wins". The winning miner then takes all the transactions that took place

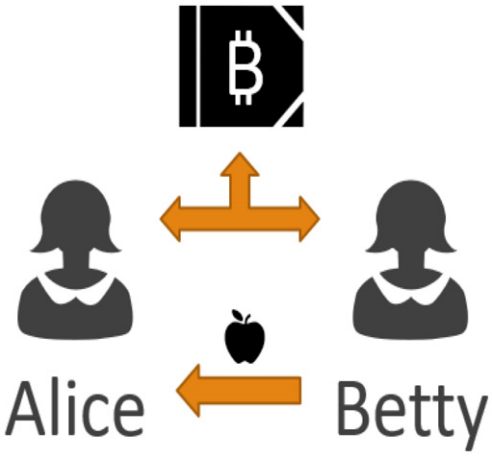

Fig. 2. In Bitcoin and other shared ledgers, there are no coins to give in exchange for the apple.

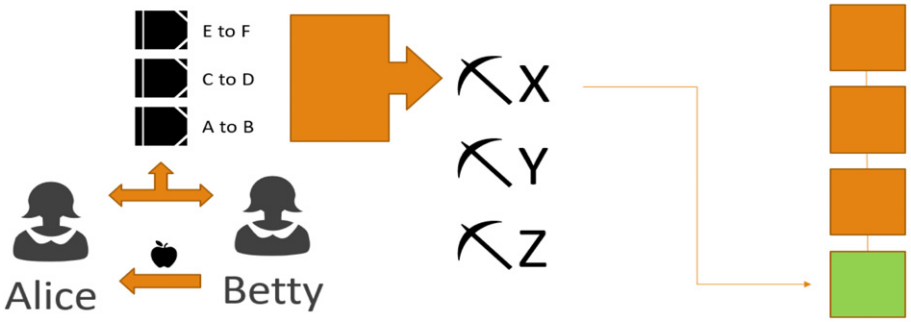

Fig. 3. When Alice and Betty create a new transaction to add to the shared ledger, it goes into a list that can be seen by a group of people called miners.

since the last "block" was mined (since the last time a miner guessed the right number, which is roughly every ten minutes at present) and has the privilege of adding them to the shared ledger [9].

Of course the question thus arises of what keeps this system stable and trustless, especially when it appears that Alice and Betty would have to trust that the miners will add their transaction to the ledger. In reality, trust is unnecessary because the miners have an incentive: they will add the transaction because it is profitable for them to do so. First, solving blocks comes with a financial reward. As of 2018, a miner gets 12.5 Bitcoins just for solving the block. Second, Alice and Betty offer a fee to the miners to add the transaction (at their own discretion, usually a tiny percentage of a Bitcoin). The miner that "wins" receives those fees in addition to the mining reward (Fig. 4).

At first it would seem like this system over-privileges the miners. Those responsible for updating the ledger, after all, could also attack the ledger, or potentially falsify it: add transactions that assign Bitcoins to themselves, or "double spend" their Bitcoins. This would allow them to spend the same set of Bitcoins in one set of blocks, then process a *different* set of blocks in which the Bitcoins *were not* spent. This creates a "branch" off the blockchain: two alternate realities in which the Bitcoins were and were not spent.

Bitcoin, however, uses something called "proof of work" to verify which blocks are the "real" blocks, and this very quickly collapses any ambiguities. Essentially, the most computation power that has gone

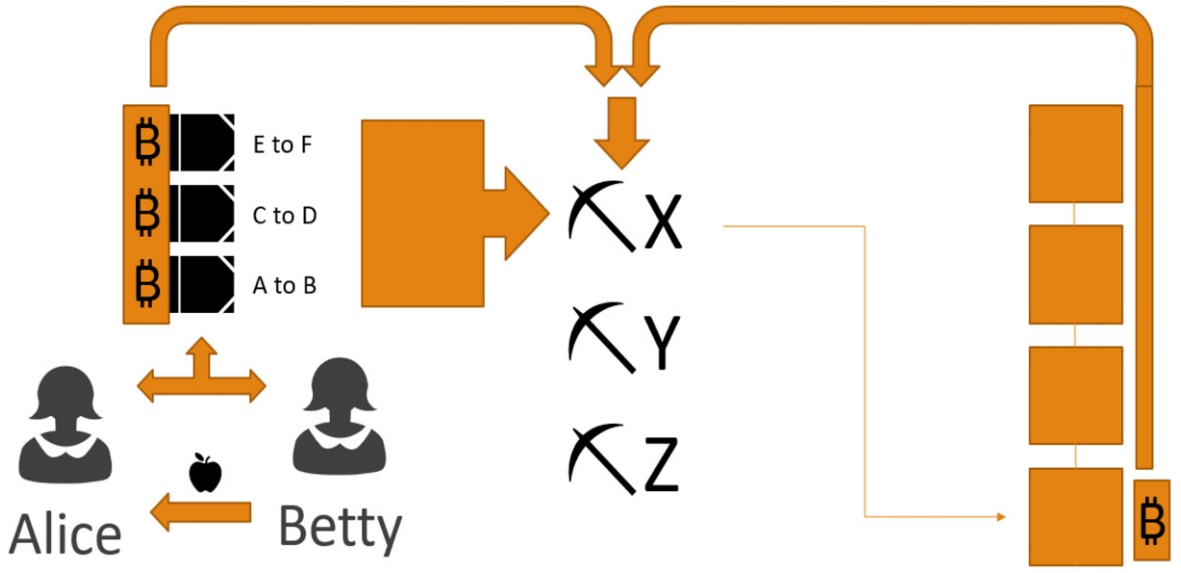

Fig. 4. Alice and Betty offer a fee to the miners to add the transaction (at their own discretion, usually a tiny percentage of a Bitcoin). The miner that "wins" receives those fees in addition to the mining reward.

into solving a branch - that is, the most investment in the form of electricity that has gone into solving a branch - is the one that is agreed by all the miners to be the "right" one. This bears repeating: the branch that has the most investment in it, in the form of computational power (meaning investment in electricity and investment in computers to turn that electricity into processing power) is the one that is "right".

What all this means is that a miner needs to execute something called a "51% Attack" to falsify the ledger. They need to control more processing power than *all the other miners combined* to create false entries in the ledger. And because "proof of work" really means "proof of financial investment" that is a very expensive endeavor. Furthermore, if a single miner did manage such a feat, or if group of miners colluded to reach a 51% share of the mining power, they would be unable to hide it. Soon, all concerned would know that Bitcoins were being double spent, and that the shared ledger was no longer reliable. This would, in essence, destroy Bitcoin. It would cause the system to fall apart, and the price of Bitcoin would tumble.

At this point it is worth remembering also that the rewards from this attack would all be *in Bitcoin*. The immense financial outlay to falsify the ledger would therefore result in any moderate gains a miner might receive from doing so becoming immediately worthless. On the other side of this incentive structure, it is clear that having a large investment in mining Bitcoin is a profitable venture in itself. So on the one hand, a miner can invest, operate, and follow the rules and make a lot of money, or on the other, they can invest a great deal more and, for their trouble, lose all their money trying to falsify the ledger. What keeps Bitcoin stable, as a decentralized system, is its incentives structure: *it is more profitable to play by the rules than to try to break them*. This is also, incidentally, why it is almost impossible to have a stable blockchain without a cryptocurrency to back it: the tie to value baked into the investment in the blockchain itself is what creates the series of incentives that stabilize it. The secret to the stability of Bitcoin's shared ledger, its blockchain, is the way it arranges incentives to drive investment in stability.

4. Bitcoin as a model

Knowing the potential problems with a centralized citation database, and seeing how something like Bitcoin plays with incentives to stabilize a decentralized shared ledger, we are left with three questions: How might we use this as a potential model for a citation index? What might that look like? And how might it help to alleviate some of the problems with centralized citation databases? These are all intrinsically linked.

Let us return to Bradford's Law for a moment, and the logic that makes complete coverage a myth. The logic goes that the work of indexing, in order to be stable, requires a centralized authority to do the work of indexing. Because of Bradford's Law of Scattering, complete coverage for any field requires coming up against a series of diminishing returns: it takes more and more effort to achieve completeness. Moreover, because this is centralized, more and more effort equates to greater and greater cost. Now, instead, imagine redistributing that work in a way that leverages incentives. Essentially there are three major shifts that would take place in who would "do the work" with a decentralized versus centralized citation index.

First, in this system, the act of indexing is performed not by a centralized team, nor even by publishers, but by the authors of the articles themselves. Like Alice and Betty buying the apple, the action of data entry becomes a simple transaction, which is validated and logged onto an ever-growing blockchain of who cites whom. From an incentives perspective, it makes a great deal of sense to put the onus on being included in the index on the scholars themselves. Researchers top the list of those who benefit most from an accurate and up-to-date citation index, as it not only helps them to produce more and better research, but also helps them to further their careers in terms of raising their respective h-index numbers. Given the role that h-index numbers play in everything from the hiring process to grants, and given how important reputation is in academia, it makes sense to leverage that in the service of a system that would ultimately benefit its users.

However, offloading data entry to authors would also fundamentally change the nature of the curation process. Accuracy of data entry is going to be a concern. The client software for interacting with the citation ledger could of course be programmed in such a way as to limit the numbers of (and ways in which) duplicate or incomplete entries are entered into the ledger, but ultimately will result in a messier database overall than one that is well-curated by a centralized source.

Second, when it comes to curation, it would still be required (perhaps even more so); however, where that curation takes place in the process would ultimately shift. Instead of taking place *prior* to inclusion in the database, as happens in a centralized database, it would take place *afterward.* An open and public citation blockchain would therefore create a new business opportunity for "big data" companies. Third parties could freely access the shared ledger, and compete in respect to the kinds of access they provide. While one might prove to be more conservative - writing off all but a set, approved list of sources and/or authors populated from those sources - another might take a more liberal approach, and allow all entries save those specifically *dis*allowed. One could imagine a variety of data-mining tools being created by a variety of companies and nonprofits to process the raw, unfiltered citation data into tools that are more or less useful to a variety of scholarly audiences. These processing layers would also, one presumes, take on the duty of eliminating and/or combining duplicate entries.

This kind of data "cleanup" might seem daunting at first, but in fact it appears to be the route Google is going with its centralized database, Google Scholar. Google's process since its inception has been to find the links between published sources and use them to extrapolate the most searched-for ones. Its Google Scholar database is no different: Google's bots crawl through publicly-available scholarly papers to see

who cites whom. It is messy, but it takes advantage of at least one of the same decentralization ideas discussed here: it relies upon scholars themselves to interact with and "tidy up" their Google Scholar accounts. This could be imitated with a fully- decentralized citation ledger, or a third-party curation solution could perform similar work in a more automated way.

The third shift would take place in hosting, and once again we turn to incentives to answer the question of who would provide that hosting for this ever-expanding raw-data blockchain. As with Bitcoin devotees, there are a range of people into whose interest this would fall. Some scholars, feeling that it is in their personal interest for the chain to be backed up in multiple places, would likely set up personal, lab- or department-based servers. The companies whose software is based on processing the blockchain would also have a strong financial interest in maintaining the citation blockchain in its entirety. And if the blockchain were (as it ought probably to be, for stability) based on a cryptocurrency and mined much like Bitcoin, then the miners would of course have an interest in keeping an up-to-date copy. But, as with Bitcoin, the ever-growing blockchain would be quite large, and so scholars might, once it reaches a certain size, only keep on hand a shortened, "hashed" version of the blockchain, to allow themselves to interact with it and add new data without having the full functionality of local data processing that the full copy would allow.

5. Conclusion

This is a fairly fundamental reimagining of the form a citation ledger could take, but it would have certain benefits. While the argument that "disruptive" technologies should be implemented for the disruption alone leaves much to be desired, this citation ledger would have the effect of disrupting the business. The idea that it would include any and all authors who want their work to be indexed would mitigate some of the pressures on smaller, startup journals, while the shift in curation to a more data-management approach would allow for new businesses and nonprofits to compete in ways of "serving up" the data in ways that are more tailored to more targeted audiences. As just one example, self-advocacy could come into play in the indexing world, with for instance women's organizations working together to create an index that highlights both women scholars and scholars who cite women's work. It would provide a new ecosystem and opportunities for new voices to be heard.

That said, as mentioned above, this paper does not end with the release of a fully-functional citation ledger, and challenges yet remain. Like the Bitcoin blockchain, there would be those who would try to attack it. Fake and disreputable journals are currently kept out of indexes fairly well by the curatorial processes of centralized citation indexes. In a distributed ledger of citations, there would be incentives on the part of dishonest scholars to artificially inflate the number of papers that cite them, and incentives on the part of disreputable journals to do what they currently do: take money to artificially advance the careers of disreputable authors through the publication of poor or wholly spurious scholarship. While "spam" could likely be mitigated by implementing "proof of humanity" into the client software - spam attacks could still potentially bog down the shared citation ledger, inflate its size and slow down the work of the honest users - there are likely other, more targeted attacks that may need to be imagined in order to implement client software that can mitigate this.

There is also the question of whether this would rely on its own blockchain (and therefore its own cryptocurrency) or if it would tie itself to an existing one, for example Ethereum or Bitcoin Cash, which as of May 15, 2018, has implemented new functionality. There remains much work that would need to be done toward creating a stable, decentralized citation ledger, but I hope that this paper has begun a

conversation about doing so, and that it has furthermore demonstrated at least some of the thoughts that would have to go into such a venture, some of the potential challenges it might face, and some of the very real promise it could present in the future.

About the Author

Richard Ford Burley holds a PhD in English from Boston College and is the Deputy Managing Editor of *Ledger*, the first peer-reviewed journal dedicated to the study of cryptocurrency and shared ledger technologies (see: https://ledgerjournal.org/ojs/index.php/ledger). E-mail: richardfordburley@gmail.com.

References

[1] This paper is a lightly-edited revision of a presentation given at the NFAIS Conference on Blockchain for Scholarly Publishing in Alexandria, VA on May 16, 2018.
[2] E. Garfield, 'Science citation index' - A new dimension in indexing, *Science*, 144.3619 (1964) 649–654, 649. Referring to Eugene Garfield, Citation indexes for science a new dimension in documentation through association of ideas, *Science*, 122.3159 (1955) 108–111.
[3] Elsevier, *Wikipedia*. Accessed 10 August 2018. https://en.wikipedia.org/wiki/Elsevier.
[4] K. Falconer, Onex, Baring Asia unveil Clarivate Analytics with close of $3.55 bln deal, *The PE Hub Network* (2016) https://www.pehub.com/canada/2016/10/onex-baring-asia-launch-close-3-55-bln-buy-of-thomson-reuters-rename/.
[5] J. Testa, "Journal Selection Process", Clarivate Analytics updated 18 July 2016, accessed 4 May 2018. https://clarivate.com/essays/journal-selection-process/.
[6] Articles of interest to a specialist must occur not only in the periodicals specializing in his subject, but also, from time to time, in other periodicals, which grow in number as the relation of their fields to that of his subject lessens, and the number of articles on his subject in each periodical diminishes". From Bradford, S. C. *Documentation* (Washington, DC: Public Affairs Press: 1950) 156. Cited in Garfield, Eugene. The Mystery of Transposed Journal Lists—Wherein Bradford's Law of Scattering is Generalized According to Garfield's Law of Concentration, *Essays of an Information Scientist* 1. // *Current Comments* **17** (1971) 222–223.
[7] In opposition to scattering, a basic concentration of journals is the common core or nucleus of all fields. Garfield, E., The Mystery of Transposed Journal Lists - Wherein Bradford's Law of Scattering is Generalized According to Garfield's Law of Concentration, *Essays of an Information Scientist* 1. // *Current Comments* **17** (1971) 222–223.
[8] E. Garfield, The mystery of transposed journal lists - Wherein Bradford's law of scattering is generalized according to Garfield's law of concentration, *Essays of an Information Scientist 1. // Current Comments* **17**: (1971), 222–223.
[9] As a side note, this is the reason the shared ledger is called a "blockchain" - because the shared ledger is just a series (a "chain") of "blocks" of transactions, one after the other, added by miners.