LEDGER
ledgerjournal.org

ISSN 2379-5980 (online)
associated article DOI
10.5915/LEDGER.2018.101

Data Insertion in Bitcoin's Blockchain: Open Review

Andrew Sward,[†] Ivy Vecna,[‡] Forrest Stonedahl[§*]

Reviewers: Reviewer A, Reviewer B

Abstract. The final version of the paper "Data Insertion in Bitcoin's Blockchain" can be found in Ledger Vol. 3 (2018) 1-23, DOI 10.5915/LEDGER.2018.101. There were two reviewers who responded, neither of whom have requested to waive their anonymity at present, and are thus listed as A and B. After initial review (1A), the submission was accepted on the recommendations of the reviewers with minor revisions required. The authors responded (1B), and the changes were accepted, thus completing the peer-review process. Authors' responses are bulleted for clarity.

1A. Review, Initial Round

Reviewer A:

Feedback on the paper:

This paper addresses and adds clarity on an issue of growing importance in blockchain research.

The paper is generally well organized and well-written.

I would like to see more examples of data storage to contextualize the discussion. The authors use fairly early stage trivial examples (*e.g.*, crypto graffiti), but a growing number of business-critical uses of blockchain rely on data storage. It would be good to cite some of these to provide an indication of the scale of both the opportunity and the challenge. Moreover, each data storage technique should provide an example of where and when it has been used, since the author's wish to aid "data archeological" or digital records forensics analyses.

The paper mentions testing, and provides scripts used, but does not go into detail in a separate methods section. It would be helpful to add this in a separate section to make it clear how the

[†]A. P. Sward (andrewsward@augustana.edu) is Assistant Professor of Applied Mathematics at Augustana College, IL
[‡]I. Vecna (ivyvecna15@augustana.edu) is an undergraduate researcher at Augustana College, IL
[§]F. Stonedahl (forreststonedahl@augustana.edu) is Assistant Professor of Computer Science at Augustana College, IL
[*]Augustana College Cryptocurrency Analytics Lab: 1Data2DYNE9ajurrnDmqZ8xNU1GzXsa9E9

conclusions were reached. The section should identify how you identified data storage methods *e.g.*, literature or sources surveyed, and how you went about testing the methods.

The factors considered in the paper concerning data storage are limited in number and not necessarily comprehensive. For example, in many business uses of data storage, there is a need to ensure a persistent link back to the business context of the transaction because 1) transactions provide evidence or proof of business transactions and 2) because this context is necessary for semantic interpretability of data stored on chain. This factor was not considered. No doubt there are others. The authors should include a statement that they begin their research by analyzing a limited set of factors relating to storage of data as a starting point.

Section 4.6 - P2FMS - you might mention that a user may accidentally spend their outputs, thereby deleting the data they stored before they actually intended to.

Issues not considered that are important from the standpoint of organizations storing data:

- Impact on privacy *e.g.*, storing data in clear text and discoverable data which may contain sensitive, personal, or confidential information
- Authenticity *e.g.*, link to business process mentioned above

Section 6.2 Transaction Malleability para. 3, p. 9 - the discussion considers an attacker motivated by economic gain e.g., value of dust versus value of transaction fees. However, other attack motivation might be the desire to record "fake news" or "false facts" in the blockchain for purposes of propaganda, etc. You mention this in fn. 38, but I think the point needs to be promoted to the main text.

In the discussion on long-term archiving issues in the final section of the paper, it should be noted that a protocol for long-term archiving of full copies of the ledger has yet to emerge and remains an open, but necessary, area of research.

A table as an appendix comparing all the methods across various parameters would really add to the paper. You should expand the factors you consider in Table 2 to include permanence (not just malleability), retrievability, and cost which are issues you discuss in your paper. This way the reader can see what the trade-offs are quite clearly.

In the appendix providing definitions of key terms, add definitions for: Malleability, Provably unspendable, Snipeable

Reviewer B:

Referee report – data insertion in bitcoin's blockchain'

A significant reason for the growth and development of the bitcoin ecosystem is the technical ability to write data onto the bitcoin blockchain that then becomes part of the permanent

ISSN 2379-5980 (online)
Associated article DOI
10.5915/LEDGER.2018.101

ledger. This 'feature' of the bitcoin blockchain is one of its most controversial but ultimately significant aspects. There are multiple ways of inserting data, however. The paper lists 5 standard script types.

This paper presents a technical survey of the current methods for inserting arbitrary data in bitcoin's blockchain. The paper lists and explains the standard methods, and provides a technical and economic analysis of their relative costs and benefits. This is a thorough and useful survey and analysis and merits publication in I think essentially its current form.

The paper is well written, comprehensive, and useful (and correctly identifies a problem of much current 'confusion and misinformation about the variety of methods by which data can be stored' and a 'dearth of academic work studying the publication and storage of arbitrary data).

The paper is particularly valuable for focusing on the question of 'whether the value of the data being stored is a worthwhile use of the bitcoin network's resources…'. I think it might be useful to draw this out further. Clearly this is endogenous to bitcoin price (and data compression technology or protocols), but what would be interesting and instructive to see is an estimate or presentation of the types of data (and applications) that actually do pass (or do not) that test. *e.g.* big gambling bets, property titling, alibi, etc.…

I would not make that a condition of revision, just a point to consider.

The broader issue here (which my well be a separate paper) is the economic externality associated with the storage of that data, and the extent to which that data is a private good, with benefits internalized by the parties that pay the fees etc to store the data, versus data that has strong public good aspects, and therefore might be undersupplied. The issue here is that its hard to think clearly about this problem (public goods value of data on the bitcoin blockchain) until we have an understanding of the efficiencies associated with the different private good aspects, which this paper provides.

Minor point: (top of page 2) 'the Blockchain' … maybe need to either specify bitcoin blockchain or clarify how general these claims are.

Minor point: (top of page 5) maybe update prices again.

Minor point (fig 5) –the vertical (cost) scale of the two graphs predicts we should see different methods at the 200 byte or so transition. Do we?

2A. Authors' Responses

Reviewer A:

This paper addresses and adds clarity on an issue of growing importance in blockchain research.

ISSN 2379-5980 (online)
Associated article DOI
10.5915/LEDGER.2018.101

The paper is generally well organized and well-written.

I would like to see more examples of data storage to contextualize the discussion. The authors use fairly early stage trivial examples (*e.g.*, crypto graffiti), but a growing number of business-critical uses of blockchain rely on data storage. It would be good to cite some of these to provide an indication of the scale of both the opportunity and the challenge. Moreover, each data storage technique should provide an example of where and when it has been used, since the author's wish to aid "data archeological" or digital records forensics analyses.

- We decided to focus more on the how than the why of data storage. Thus, we have decided not to include more examples beyond those we have already included. Each basic method of data storage has been exemplified in at least one endnote and/or figure.

The paper mentions testing, and provides scripts used, but does not go into detail in a separate methods section. It would be helpful to add this in a separate section to make it clear how the conclusions were reached. The section should identify how you identified data storage methods e.g., literature or sources surveyed, and how you went about testing the methods.

- We added a brief subsection under Background: The Bitcoin Script Language discussing our methods, as well as an appendix with sample code.

The factors considered in the paper concerning data storage are limited in number and not necessarily comprehensive. For example, in many business uses of data storage, there is a need to ensure a persistent link back to the business context of the transaction because 1) transactions provide evidence or proof of business transactions and 2) because this context is necessary for semantic interpretability of data stored on chain. This factor was not considered. No doubt there are others. The authors should include a statement that they begin their research by analyzing a limited set of factors relating to storage of data as a starting point.

- We feel this is outside the scope of the paper. Protocols for organizing the relevant data content is a subject for continued research.

Section 4.6 - P2FMS - you might mention that a user may accidentally spend their outputs, thereby deleting the data they stored before they actually intended to.

- Spending the P2FMS output would not delete the data, but it would remove it from the UTXO set. We feel that the paper already sufficiently addresses this point.

Issues not considered that are important from the standpoint of organizations storing data:
impact on privacy - *e.g.*, storing data in clear text and discoverable data which may contain sensitive, personal, or confidential information

ISSN 2379-5980 (online)
Associated article DOI
10.5915/LEDGER.2018.101

- We added an endnote reminding the reader that publishing data using the Blockchain makes it public, and that information intended to be private should either not be published or be published in an encrypted state or as a hash.

Authenticity *e.g.*, link to business process mentioned above

- We agree this would be interesting, but we feel that it is beyond the scope of the paper.

Section 6.2 Transaction Malleability para. 3, p. 9 - the discussion considers an attacker motivated by economic gain e.g., value of dust versus value of transaction fees. However, other attack motivation might be the desire to record "fake news" or "false facts" in the blockchain for purposes of propaganda, etc. You mention this in fn. 38, but I think the point needs to be promoted to the main text.

- This is an important point, and we have promoted the footnote to the main text, as the reviewer suggested.

In the discussion on long-term archiving issues in the final section of the paper, it should be noted that a protocol for long-term archiving of full copies of the ledger has yet to emerge and remains an open, but necessary, area of research.

- We added a note about this in the conclusion.

A table as an appendix comparing all the methods across various parameters would really add to the paper. You should expand the factors you consider in Table 2 to include permanence (not just malleability), retrievability, and cost which are issues you discuss in your paper. This way the reader can see what the trade-offs are quite clearly. In the appendix providing definitions of key terms, add definitions for: Malleability, Provably unspendable, Snipeable

- We added Transaction Malleability, Provably Unspendable, and Snipeable to Appendix A, and we added a table of all the methods comparing additional factors. We determined here that the real question of "permanence" is whether or not the data is stored in the UTXO set. Retrievability is too complicated an issue to address in a simple table, so we did not include it. In effect, all methods allow the data to be retrieved; the real questions involve whether or not a standard already exists for doing so, how easy the data is to find and identify, etc.

Reviewer B:

Referee report – data insertion in bitcoin's blockchain'

ISSN 2379-5980 (online)
Associated article DOI
10.5915/LEDGER.2018.101

A significant reason for the growth and development of the bitcoin ecosystem is the technical ability to write data onto the bitcoin blockchain that then becomes part of the permanent ledger. This 'feature' of the bitcoin blockchain is one of its most controversial but ultimately significant aspects. There are multiple ways of inserting data, however. The paper lists 5 standard script types.

This paper presents a technical survey of the current methods for inserting arbitrary data in bitcoin's blockchain. The paper lists and explains the standard methods, and provides a technical and economic analysis of their relative costs and benefits. This is a thorough and useful survey and analysis and merits publication in I think essentially its current form.

The paper is well written, comprehensive, and useful (and correctly identifies a problem of much current 'confusion and misinformation about the variety of methods by which data can be stored' and a 'dearth of academic work studying the publication and storage of arbitrary data).

The paper is particularly valuable for focusing on the question of 'whether the value of the data being stored is a worthwhile use of the bitcoin network's resources…'. I think it might be useful to draw this out further. Clearly this is endogenous to bitcoin price (and data compression technology or protocols), but what would be interesting and instructive to see is an estimate or presentation of the types of data (and applications) that actually do pass (or do not) that test. *e.g.* big gambling bets, property titling, alibi, etc…

I would not make that a condition of revision, just a point to consider.

- We intentionally avoided exploring this issue in-depth, instead opting for a more straightforward, neutral presentation of the methods of data storage. Certainly this would be a point of further discussion. We hope our paper will help provide a basis from which to further explore this topic.

The broader issue here (which my well be a separate paper) is the economic externality associated with the storage of that data, and the extent to which that data is a private good, with benefits internalized by the parties that pay the fees etc to store the data, versus data that has strong public good aspects, and therefore might be undersupplied. The issue here is that its hard to think clearly about this problem (public goods value of data on the bitcoin blockchain) until we have an understanding of the efficiencies associated with the different private good aspects, which this paper provides.

Minor point: (top of page 2) 'the Blockchain' … maybe need to either specify bitcoin blockchain or clarify how general these claims are.

- Per Ledger journal style standards, the standalone word Blockchain is capitalized in this paper when referring to Bitcoin's blockchain. This has been added as an endnote following the first instance of the word Blockchain in the main text.

Minor point: (top of page 5) maybe update prices again.

ISSN 2379-5980 (online)
Associated article DOI
10.5915/LEDGER.2018.101

- This paper was written in July 2017, and accurately reflects the state of the Blockchain at that time. Since this paper does not address the influence of SegWit, the BCH hard fork, or other changes to the Bitcoin protocol, we feel that it is most appropriate to leave the original price quotes in the paper. Given the price volatility of Bitcoin, any prices are likely to be outdated in a matter of weeks, if not months, and readers can easily scale our results to match current prices.

Minor point (fig 5) –the vertical (cost) scale of the two graphs predicts we should see different methods at the 200 byte or so transition. Do we?

- We have not seen a trend of opting for different methods when they become more cost-efficient. To obtain a thorough analysis is beyond the scope of this paper, as it is difficult to determine exactly what is and is not used for data storage. For example, P2FKH UTXOs are effectively indistinguishable from legitimate P2PKH UTXOs, unless the data is understood (e.g. printable ASCII). This said, the number of interesting uses of P2SH (most uses of P2SH simply encapsulate a Multi-Signature script) is relatively small. What we have observed suggests that the main method used is P2FKH, regardless of the amount of data to be stored. This is an interesting point to be discussed, but not one that we felt should be addressed so explicitly in the paper.

ISSN 2379-5980 (online)
Associated article DOI
10.5915/LEDGER.2018.101