# Resolving Review Comments on the IJDC Paper

RBLandau 20210430

[Comments from the reviewer are normal text. My comments on the comments are all indented and bulleted.]

The framework used for the simulations is open source, the authors say, and they invite the community to use it. Yet surprisingly, the paper does not provide a link to the framework.

* RBL: Serious oops, sorry. We had the github reference in the slides but failed to put it in the paper? Duh.
* RBL: Science, and other journals, I think, usually specifies the author(s) who are the contacts for questions and comments on the papers. The github README does list both our email addresses as contacts. I should probably specify mine as the contact for computer code and yours as the contact for curation discussions.
* **RBL: DONE**

In the Methodology section, the authors explain that, in their model, any error in a document makes the document unusable. They might hint that this assumption is not as rigid as it looks because a less fragile document can be modeled as multiple smaller documents, as they do in their discussion about compression later on.

* RBL: Who gets to justify this basic assumption we made in 2014-15?
* **RBL: NO CHANGE. This is addressed later on.**

In table 1, the column “Distribution” contains Poisson events and an exponential lifetime. Isn’t this exponential lifetime a consequence of a Poisson event too (server failure)? And “Poisson event of some duration” sounds strange because the event itself has no duration. As I understand, it is really “Poisson event *within some time span*” or “Poisson *process* of some duration”.

* RBL: I think this is the glitch-duration wording that might be confusing. Yes, the event itself is instantaneous, but we allowed its aftereffects to persist for a fixed or exponentially decaying time. The wording needs fixed.
* **RBL: DONE. Added word in Table 1 to say that the Poisson event triggers an environmental change of some duration.**
* **RBL: BUT this created a bad page break that separated the caption from the table. I moved Table 1 within the text to fix that. Still not great.**

Figure 2: the article text and the caption say that it is about the likelihood that a single document is lost (in a collection of 10,000, note 4) but the vertical axis is labelled “permanent document losses (% of collection)” which is not the same. I stick to the latter for my interpretation below.

* RBL: Well oops on the labels. I will redo, but it may be done with Photoshop rather than correcting the R code that drew the pictures.
* **RBL: DONE. The labels are correct. I added wording to clarify that unaudited collections lose documents permanently even with large numbers of copies and high quality storage.**

As the authors themselves explain, you need to know document sizes to calculate document loss percentage from the number of bad sectors. To arrive at Figures 2 and 3, showing document loss, some assumption was necessarily made about document sizes. It should be made clear what this assumption was.

* RBL: We have figures and graphs to demonstrate that the choice of document size was not relevant, easily extrapolated linearly. We probably need to justify the choice of a fixed document size, though. It just simplified the interpretation of the numbers a lot.
* **RBL: DONE. I added Appendix A containing the table of document sizes, along with explanation of the purpose and results of the table; and a sentence pointing to it under the picture of the red Xs.**

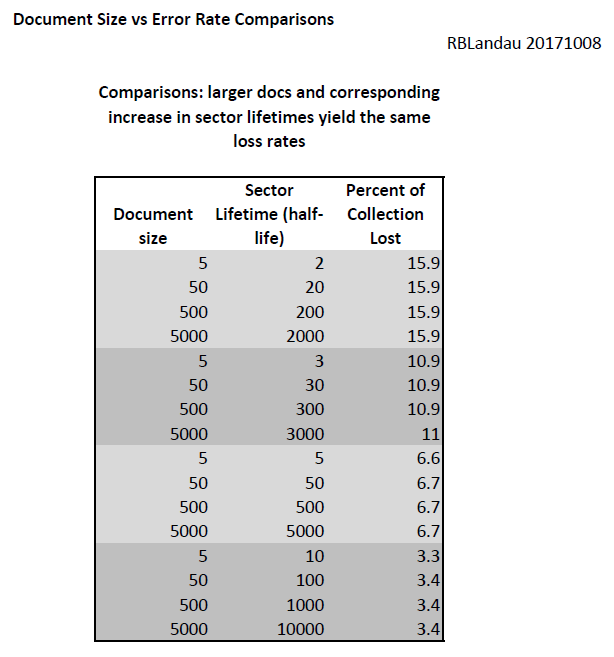
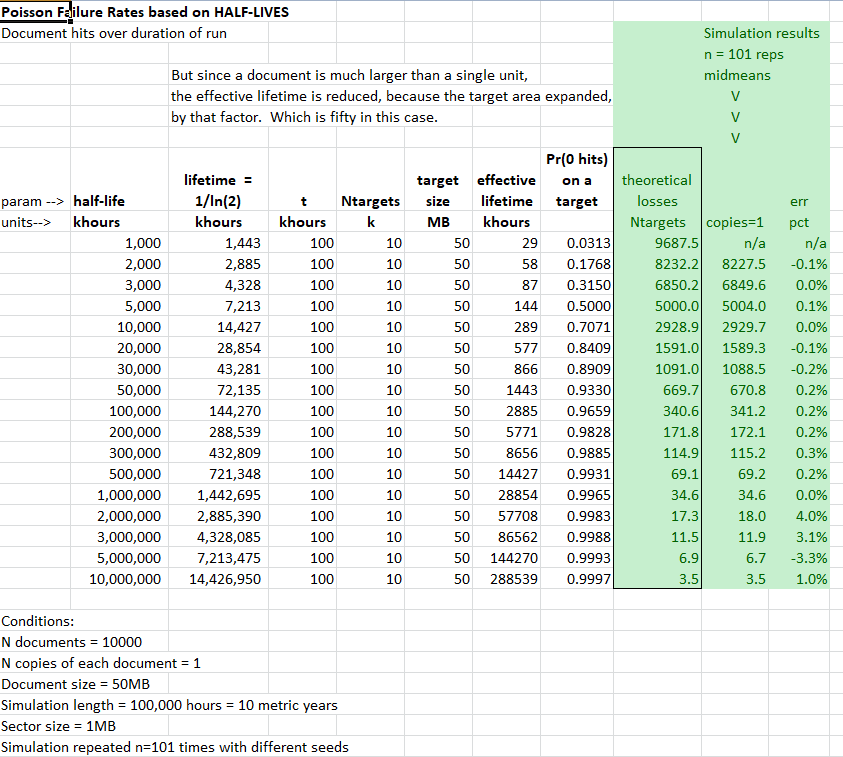


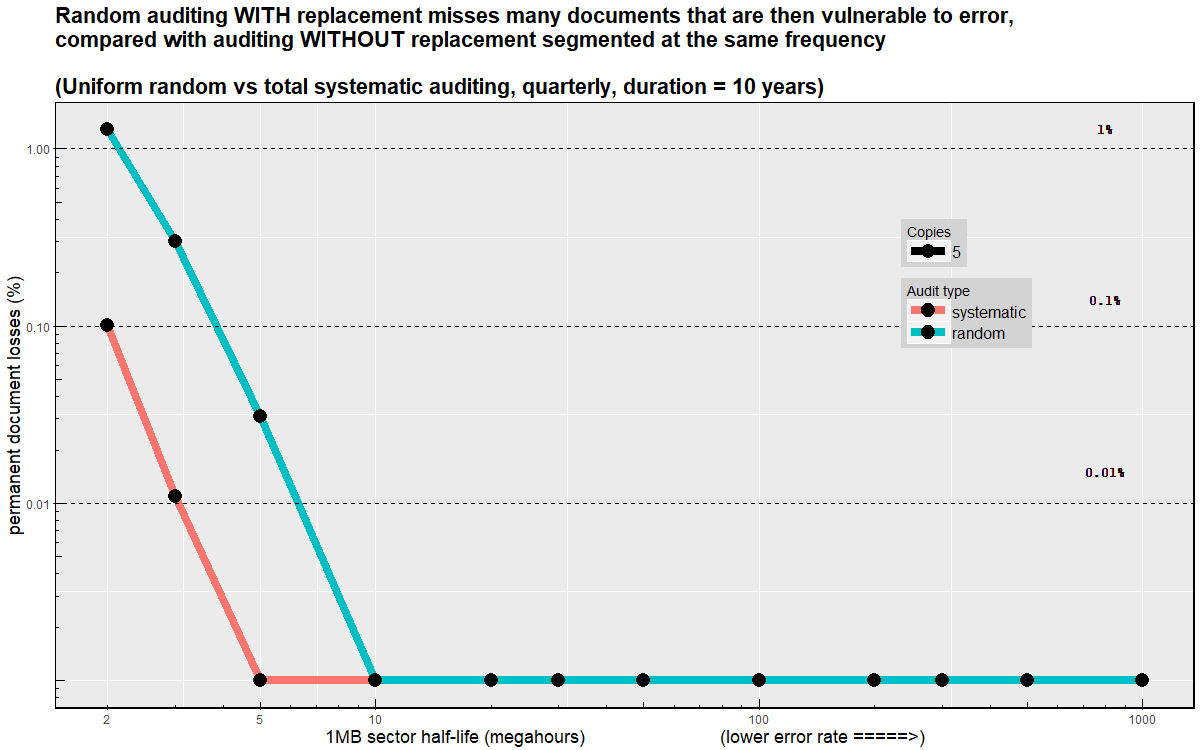
Figure 2A, the simplest case, can also be calculated in closed mathematical form if all documents were assumed to have the same size (but I don’t know if they were). In that case, it would be useful to see this theoretical curve along with the simulation results to confirm the quality of the simulation. On the bright side: the simulations seem to reach an asymptote with slope -1 on log-log scale, as would be expected.

* Discuss: should we add something to the picture, which might muddy it, or can we include the little table that I did showing the accuracy of the simulation for that case. Something like a subset of this old (2018) picture.
* **RBL: DONE. I added Appendix B containing this table, and a sentence under figure 2A mentioning it.**



The authors make a strong case that auditing should occur by sampling without replacement (i.e. systematic). It would be illustrative to quantify this statement, e.g. by including a curve for “sampling with replacement” (i.e. random) in Figure 3A.

* RBL: should we add a picture? Would a small table of figures do? I think we actually did a picture like that looong ago.
* **RBL: DONE. Added Appendix C containing this picture, and a sentence referring to it in the paragraph about the perils of replacement.**



* RBL: But we need prettier pictures anyway, right? Pls suggest types of improvements.

In Figure 3A, the shaded area (I presume this is the yellow part) representing the range of plausible disk quality covers 30 – 1000 Mh while it is estimated at the bottom of page 7 to be 20 – 30 Mh.

* RBL: oops, typo?
* RBL: he means figure 2B not 3A.
* **RBL: DONE. Minor change to clarify that we extended the range under consideration. The discussion said that the \*plausible\* manufactured range is 100-1000, but that poor environmental conditions, glitches, caused us to \*extend\* the \*realistic\* range to 20 or 30. Was perhaps easy to misread.**

In many of the figures, y-values below 0.001% are shown as exactly 0.001%. This is probably clear from inspection for most readers, but it should be stated explicitly in the text (just omitting the number from the axis is not enough).

* RBL: ah, this one might be a pain. To get ggplot (and other plotting packages) to use log scales, one has to edit out the zeros. Well, oops on that, because most of the interesting data are actually zeros. So what I did (and we discussed this a couple times) was bias the zeros by adding 10ppm. The values that \*say\* they are 0.001% are actually zero. Well, oops that someone noticed.

Can we explain this in the text? Is what I did an acceptable workaround? Do you know how to convince R not to choke on zeros if I ask for a log scale?

Redoing all the pictures to take the logs manually and then force the axis labels to look like log scales might be a biggish deal. And we don’t have MIT’s purse anymore to hire a consultant, do we?

* **RBL: DONE. I added a footnote that very small values are grouped together, but that all of them are in fact zeros. Feel free to improve the wording, please.**

The figures would benefit from a slightly bigger font size for readability.

* RBL: What do you think? This might require redoing every picture in that icky R code.
* **RBL: No change. Unless this is \*really\* important, I don’t wish to spend the time or energy on it.**

In the sections about the loss of encryption keys and file-format obsolescence, the authors make specific recommendations about the number of copies and auditing frequencies of keys and formats, respectively. Unlike in most sections about other risks, it is not clear how these numbers arise from the model. Maybe the authors can shine some light on that without making the paper too long, e.g. by quantifying some key parameters that went into the simulations.

* RBL: tag, you’re it, I think.

References: some have no link but I found them online by search. Can links be added?

* RBL: tag, you’re it, I think.

Typo: [Gallinger] surve = survey

* **RBL: DONE.**

Dr Angus Whyte

Associate Editor, IJDC