Response to Reviewer Comments

September 1, 2022

We thank the Editor and reviewers for their encouraging assessment and constructive feedback.

Editor

> *The paper has been reviewed by a referee and an associate editor and I have read the manuscript independently, before looking at the reports. When I confront my notes and overall impression with the reviewers' comments, I find myself in substantial agreement with the reviewers.*
>
> *As indicated by both reviewers, the work expands on previous work by Sadinle, introducing computational advances that are achieved by relaxing some requirements in the original formulation of the problem. The close connections with the previous work and the extent of the novel contributions are not adequately explained. This makes it hard for the reader to understand what is new and why the new contributions are valuable. A careful editing of the presentation is needed to address this shortcoming.*

E.1: We thank the Editor, the AE, and the reviewer for their feedback, and pointing out our strengths of the paper and areas that we can improve the quality of our paper. In summary, we have revised our paper in the abstract and introduction as to make our contributions more clear to that of prior work. Furthermore, we have carefully edited the paper to address issues that were pointed out by the Editorial Board. [TODO]

> *Regarding the presentation, the AE points out a number of problems with notation, typos, and various inconsistencies. Independently, I found many similar problems. The intersection between what I found and what the AE found is non-empty and so is the symmetric difference. Also, the AE found some issues that I did not find and vice versa. So, I am almost certain that an additional review would uncover more problems. Now, the results presented in the paper seem plausible as do the broad strokes of the derivations, but I must admit that I was not able to follow all the details. This, in large part, was because of the issues that I just mentioned, and I believe that most readers would find it difficult to follow the developments, as these issues are exceedingly distracting.*
>
> 1. *There are problems with the notation, which in places is not defined, in others is used before being defined, and in others yet is used inconsistently.*
> 2. *For example, $I_obs()$ is never defined, $n_12(Z)$ first appears on p. 5, but is not defined (in passing) until the bottom of p. 7, and it seems to become D on p. 8.*
> 3. *Throughout the manuscript, n_1 and n_2 are used interchangeably with n_A and n_B, sometimes in the same section, as it happens, for example, in Section 3.1.*
> 4. *The loss function on p. 9 is out of Sadinle's paper and it inherits the typo therein $(\theta_1 1, if Z_j, \hat{Z}_j, \dots)$.*
> 5. *In the statement of Lemma 1 on p.13, B comes out of nowhere. (Thinking I had missed something, I went back to the previous pages only to find out that B would be then defined in the proof.)*
> 6. *Formatting of all displayed math must be checked and fixed, especially as far as punctuation is concerned, as commas seem to be missing in multiple places.*
> 7. *For example, in Equation (1), there should be a comma after 0 and after 1, and this is by no means an isolated occurrence.*

E.2: We thank the Editor, the AE, and the reviewer for their feedback. We have made an effort to make it more clear our contributions in this work as opposed to previous work. Furthermore, we have carefully edited the paper to address issues that were pointed out by the Editorial Board. [TODO]

> *Confusion can also arise from lack of clarity in the exposition. Take comment 4 from the AE, for example. Whose marriage certificates and whose birth records are involved? What records are in X_1 and what records are in X_2? The reader should not be expected to go back to Newcombe's paper to clarify the issue.*

E.3: Go fix this! [TODO]

> *Other comments that were raised are as follows:*

> 1. *This is not an exhaustive list, but only a set of examples meant to point out what makes the paper hard to read and how it can be improved. Many more specific points are raised in the AE's report and here are a couple of other minor issues I noticed:*
> 2. *p. 12, second paragraph: insert "is" after Section 3.1*
> 3. *p. 15: the first sentence of the last full paragraph is either missing a verb ("we observe"?) or it includes an extra "that."*
> 4. *As I said, I do not believe that we, as reviewers, were able to uncover all the problems, and I strongly encourage the authors to do their part, as they should, to improve the presentation and eliminate all typos and inconsistencies.*
> 5. *The AE finds the simulation studies to be incomplete and gives detailed suggestions on how those should be improved. The AE also gives important suggestions about other aspects of the manuscript that must be carefully considered and addressed. All of the AE's comments are right on the mark.*

E.4: Go fix these! [TODO]

To summarize, the paper contains an interesting algorithmic contribution that can speed up calculations at the expense of relaxing some of the modeling conditions, without much adverse impact on the resulting inferences. This aspect should be emphasized in a revised presentation. The presentation should also make clearer the close connections to the relevant work by Sadinle, and all the issues that I mentioned above should be ironed out.

The authors should prepare a careful and substantive revision that remains within the editorial limit of 25 pages and that answers the various comments satisfactorily. The authors should upload the revised manuscript together with a document detailing how they addressed the reviewers' comments.

E.5: Thank you for the opportunity to submit a revision and improve the quality of our paper. We have provided detailed responses to the reviewers' comments below, which prompted significant changes in our paper.

Suggestions/Tips for Brian to Discuss: Let's work on the easy fixes first. Go fix all the notation, check the grammar, and fix the point by point responses of the Editor as these can be done hopefully quickly. Let's use this more organized format so we can more quickly talk about what comment the Editor is making. It also makes it easier for the Editor/AE to review, so they like this (just a pointer). Suggest putting lines in the paper so we can also know where we are and it's often helpful to put changes in color so we can check every detail before resubmission. These are just some tips that have worked in the past that I hope that will help. I typically like to handle the big changes last as they take a lot of time and more attention to detail. Antidote is a nice tool for checking grammar or asking someone who is really great at proofing to find these issues and mark it up for you. Olivier is pretty good at spotting such things or Ted, so they might be good choices. Finally, given that we have some sloppy stuff going on, I'd recommend having this really polished before sending it to HRDAG or Patrick Ball will think it's a mess. I personally don't like sending things that aren't polished out the door, and this my fault for not catching this prior to submission.

Reviewer 2

Subtle comment to Brian: The Editor told us that this review is the AE, so we want to satisfy all these comments if possible since the Editor thinks the comments are ”spot on”

> *This manuscript addresses the bipartite matching problem in data linkage and proposes a new fast computation version of “Beta Linkage” to allow such matching to take place for pairs of large databases without resorting to ad hoc blocking. It is built upon the Felegi-Sunter model of record linkages, where linkages are made within strata of agreement between records across the two databases, and relies heavily on advances made by Sadinle (2017). Overall, the paper presents a practical implementation of a modified Beta Linkage approach to matching. It demonstrates that even though the new method suffers from the limitations of not seamlessly enforcing one-to-one matching, it is feasible for large datasets with modest numbers of features on which to match. Though the novelty of the presented methodology is primarily in the computational approach, I believe it to be a real advance for practical implementation of record linkage in modern applications. I have no major concerns about the content or accuracy of the presentation. However, substantial improvements are warranted to confirm my impression. In addition to the recommendation made by Reviewer 1 to include more clarity of the novel contributions of this manuscript, I believe that the explanation of the method should be written more clearly, the simulation comparisons should be repeated/expanded, and various features of the simplification should be fleshed out a bit more. Finally, more practical advice would be welcome. All of these points are more specifically described in the itemized list below*

R1.1: We thank the reviewer (AE) for noticing our computational novelties that expand upon the prior limitations of the prior work of Sadinle (2017). Our review paper clarifies our novel contributions, improves upon the exposition of the writing, provides a practical guidance for users, and expands the simulations. We expand on these points further below. [TODO]

> *Overall, the notation is, somewhat by necessity, quite extensive. However, I’m not convinced that it’s entirely consistent. A glossary in the appendix may help the reader navigate the paper more easily.* ***R1.2:*** *I think this is an excellent idea and it could make sure that we don’t miss anything. It would really help the reader. You could make a table of the notation, and Neil does this quite well in the Marchant 2021 paper as an example.*

> *Section 2, line 1: It may help to explicitly define X1 and X2 as being vectors of indices as implied by equations (1) and (2).*

I'm not sure I agree with this point The reviewer is stating that equations 1 and 2 are vectorized, so everything is else needs to be vectorized as well. In short, all notation needs to be consistent.

> *Section 2.2, beta distribution: the $n_{12}(.)$ notation needs to be defined.*

R1.3: We thank the reviewers for identifying this oversight, and have defined the notation.

> *Section 3, 4th full paragraph: My understanding of the example of matching birth records and marriage certificates (presuming monogamy) may not quite be consistent with the algorithm as presented, as I would think that one certificate (element in X1) would match multiple birth records (elements in X2) which violates the size ordering. Can the algorithm be easily adjusted for this reverse sizing?*

I need some help addressing this in the paper. The answer is yes, the algorithm can easily be adapted for the reverse sizing. The issue is mostly notational. It seems as though Sadinle chose to label the datasets X_1 and X_2 such that $n_1 > n_2$ for mostly notational reasons. He uses the notation $Z_j = n_1 + j$ when record $j \in X_1$ has no match in X_2, which works because when $n_1 > n_2$, $n_1 + j$ never refers to any record in X_2. Jody uses the simplified notation $Z_j = n_1 + 1$ to refer to unmatched records, and this works for the same reason. However, when we relax the one-to-one constraint, or even encourage a modeller to allow multiple linkages in one dataset, we shouldn't denote the datasets according to simple notational convenience. This might require clever edits of the "Review of Prior Work" section, and other areas of the paper. Notationally, we can use $Z_j = 0$ or $Z_j = -j$ as a way to unambiguously denote a record as nonmatching. Additionally, we could just strike the paragraph about the advantages of many-to-one matchings, and keep this strictly a paper about bipartite matching

This is a good point. I wonder if we could perhaps remove a lot of the notation by Sadinle/FS and redo the notation such that the ordering wouldn't matter. Also, it would make more sense to follow with the examples that we use in the paper – NLTCS and EL Salvador so that we don't confuse the reader. Let's talk through this point as this would be a big change. If the algorithm and notation can be easily modified to reverse things, this would be a nice addition to the paper and improvement over Sadinle.

> *5. Section 3, full model: $I_obs(.)$ seems to be taken from a different paper dealing with missing fields, which is not at all discussed in this manuscript. This consideration of missing fields and the accompanying discussion of missing data mechanism should be included if such data is to be considered for this manuscript. This comment also applies to the Gamma^{obs} notation in the following subsection.*

R1.4: I can add these definitions, no problem. But is it worth including mention of missing data? Given that we have missing data (or could have missing data), this needs to be added into the paper. Otherwise, the authors are going to complain that we don't have it. If you don't have this in your code, then remove it. Make sense?

> *Section 3, full model: The Phi notation needs to be defined.*

R1.5: We thank the reviewers for identifying this oversight, and have defined the notation. [TODO]

> *Section 3, full model: The square brackets seem to be misplaced, as the exponent includes indices of the sum contained in the brackets.*

R1.6: We thank the reviewers for identifying this oversight, and have corrected the expression. [TODO]

> *Section 3.1, I found the motivational descriptions in this section to be somewhat confusing, in that they seem to mix the posterior behavior (e.g., parameters are updated though standard multinomial-Dirichlet...) and model construction (as a function of pi, Z is a series of successes...). This subsection might be easier to understand if its purpose was more bluntly described in an opening sentence or paragraph and streamlined to focus only on the full conditional distributions for a Gibbs Sampler. In addition, the mixed motivation may have led to errors in intermediate conclusions. In particular, the derivation of the probability of Gamma$_{.j}$ seems to be missing a factor of $u^I_{fl}(gamma^f_{ij} = l)$. This is not important to the eventual full conditional of Z, but does not seem to be correct for the distribution of Gamma as claimed.*

R1.7: Let's talk through why the reviewer find this to be confusing. I think that the AE is asking for the joint posterior to be written in such as way that we can easily decompose the full conditionals more naturally. So, we can start with what you have and then simplify the joint posterior such the full conditionals fall out more easily. The point below makes me think this is his/her suggestion.

> *Section 3.1, full joint posterior distribution: I was surprised to see that the contribution of the conditional distribution of Z was included using the summation-style notation rather than the binomial/multinomial-style multiplicative notation that naturally motivates the full conditional presented on the first line of page 8.*

R1.8: This comment goes with the one above.

> *Section 3.1, first equation on page 8: There is an extra parenthesis in the exponent of (1-pi).*

R1.9: We thank the reviewers for identifying this oversight, and have corrected the expression. [TODO]

> *Section 3.1, pmf for Gamma$_{.j}$ and full conditional for Z: The notation in the final equation may be more clear if you replace i with z_j, as in $w_{z_j j}$. I believe that "$n_1 + 1$" should be "$n_1 + j$" in several places.*

R1.10: I would recommend fixing this to make it more clear since this is the AE.

> *Section 3.1, last equations on page 8: There seems to be some lack of specificity in describing these equations as full conditional distributions vs. steps in the Gibbs Sampler algorithm. It would be helpful to provide a brief justification of integrating out pi.*

I'm not sure how to justify this step, but also this step isn't actually necessary. We could just sample π and use it, as in a standard Gibbs sampler. Should I just remove this step? The AE makes a good point. We have given the full conditionals, but have not written down the Gibbs algorithm. The AE is expecting something like this for the algorithm: https://arxiv.org/pdf/1403.0211.pdf

> *Section 4.2, SEI procedure. I'm not sure I follow the description of the SEI procedure. My understanding of SEI is that for patterns with lots of possible pairs, no single pair is likely to (should) be identified in the posterior distribution. Thus, you take a small subsample of the records of size $S < H_{j_p}$, and store these in R^{SEI} rather than the complete index of pairs in each R. Can you quantify the computational savings due to this method? Do you have advice for how small S can be? In your simulations, how does the use of this method or choice of S affect the results?*

Do you think you could write the computational complexity of this out Brian as this would help? In addition, how does the SEI come into play in the Gibbs sampler? If it comes into play, it will need to be in the algorithm.

All computational savings occur from conducting the Gibbs sampler over the P patterns instead of the n_2 records. SEI is for memory savings. I think I need help articulating this clearly.

> *Section 4.4, Assumptions and definitions should be included in the statement of this lemma.*

> ***R1.11:*** *Do I really need this to be a Lemma? It seems overly clunky to me, but if its convention to provide computational complexity as a lemma, thats OK with me. Yes, this is convention.*

> *Section 5.1 simulations – there is a detailed comparison for one set of simulation settings, including a binary definition of a "match". Do these results differ as these settings change?*

R1.12: Yes, results would be the same. To me, this is clear, I'm not sure how to clarify this in the paper. Did you run them on other settings? What were they? Don't include this in the paper, but explain this here and we'll work on this together.

> *Section 5.2 and 5.3 simulations – are these a single simulated datasets? Do these same results hold over multiple simulated datasets, Or was this just a chance result for each setting?*

R1.13: In the caption under Figure 3, we explain that we have 100 pairs of sets of records for each level of overlap and error. We have added this description in the body of the paper as well for clarity. We need to more clearly answer the question. Was this one data set? This isn't clear to me from the answer.

> *Section 5.3 simulation – RR results are not presented.*

I will redo simulations in order to get this

> *Section 6.1 last sentences. I believe the last sentence should be removed, as it is redundant.*

R1.14: This has been removed.

> *Section 6.2, second sentence. I believe the word "survey" is missing, as in "the initial SURVEY began..."*

R1.15: This has been fixed.

> *Section 6.2. In the NLTCS, are participants added in later cohorts, or may we assume that every member of the smaller dataset should also be included in the earlier one?*

Participants can be added in later cohorts. This has been clarified in the paper. [TODO]

> *Figures 5 and 7 are not referenced in the text.*

R1.16: This has been fixed.

Reviewer 3

The paper presents a computational variant of the method introduced by Sadinle (2017). Simulation results are presented and the variant is applied on a new, significantly larger data set. A significant part of the paper is a reproduction of the paper of Sadinle. The variant is introduced in section 4. It is simple but consequential. The simulations and applications show how the variant accelerates the computation relative to the original method and reveals the trade-offs. The value of the paper resides in that computational complexity continues to be a difficult obstacle in the application of Bayesian statistics. The proposals of the paper enable the application of Sadinle's approach to an extent not possible before. It is important to present the research for what it is: An extension of the method of Sadinle motivated by computational considerations. Here Sadinle supersedes Fellegi-Sunter in the sense while the starting point of Sadinle's research is Fellegi-sunter, the starting point of this paper is Sadinle's work. This should be made clearer, beginning in the abstract and reiterated in the conclusion.

We thank the reviewer for their comments, and we have revised the clarity of our paper. [TODO]