Response to Reviewer Comments

September 9, 2022

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

Editor

E.0: 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.

> ***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]*

E.1: 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.

> ***We appreciate these comments immensely, and have done our utmost to address the ones listed by yourself and the AE. Furthermore, we have carefully gone through the paper ourselves, where we have corrected other issues with notation, typos, and various inconsistencies.*** *[TODO]*

E.2: 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. 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.

> ***The definition of*** $I_{obs}()$ ***now defined. We have defined*** $n_{12(Z)}$ ***in the correct place (and removed the error with*** D***). We have checked for other notational inconsistencies. Additionally, we have defined*** $n_{12}(Z)$ ***when it first appears.***

E.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.

> ***We have fixed the inconsistency, using*** n_1 ***and*** n_2 ***throughout the paper.***

E.4: The loss function on p. 9 is out of Sadinle's paper and it inherits the typo therein (θ_{11}, if $Z_j,\ \hat{Z}_j, \ldots$).

> ***The previous expression*** $Z_j, \hat{Z}_j \leq n_1$ ***was meant to be understood as*** $Z_j \leq n_1, \hat{Z}_j \leq n_1$. ***We agree that the latter expression is more clear. This has been updated in our revision.***

E.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.)

> ***We have updated the Lemma to include the definition of B as well as other variables in the Lemma.***

E.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. For example, in Equation (1), there should be a comma after 0 and after 1, and this is by no means an isolated occurrence.

> *BEKA/JERRY: This exact reference seems purely stylistic. We need to decide how to respond to this. The Editor is unhappy as the displayed equations are not consistent regarding punctuation. They must all be checked and be the same. This could be due to the fact that Brian/myself edited the paper separately. I would follow the notation that Mauricio uses (as one example) or that of Marchant et al. (2021). The idea is to be very consistent in every single equation. Yes, this is very nitpicky, but we need to satisfy the Editor. Brian: I'm happy to work with you if this is still unclear or you need some additional help with it.*

.

E.7: 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.

> ***We agree with this comment, and we have removed this example as it does not add to the paper. Furthermore, we have gone through the paper to improve the overall exposition and clarity of the writing.***

E.8: Other comments that were raised are as follows: 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. We summarize these below.

> 1. *p. 12, second paragraph: insert "is" after Section 3.1.* ***This has been fixed.***
> 2. *p. 15: the first sentence of the last full paragraph is either missing a verb ("we observe"?) or it includes an extra "that."* ***We are not able to find this error. Go check this.***
> 3. *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.* ***This has been done by all authors.*** *[TODO]*
> 4. *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 consid-*

> *ered and addressed. All of the AE's comments are right on the mark.* ***We have provided detailed comments to the AE.***

E.9: 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.

> ***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.*** *[TODO]*

Reviewer 1

R1.0: 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

> ***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]*

R1.1: 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.

> *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. [TODO]*

R1.2: 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).

> *We thank the reviewer for the suggestion, which has been incorporated into the paper in the following way: "Consider two databases X_1 and X_2 containing records $\{x_{1i}\}_{i=1}^{n_1}$ and $\{x_{2j}\}_{j=1}^{n_2}$ respectively. Without loss of generality, denote files such that $n_1 \geq n_2$. We follow the convention established by Sadinle (2017) and say "record $i \in X_1$" rather than the more compact x_{1i} in order to avoid double subscripts." [TODO]*

R1.3: Section 2.2, beta distribution: the $n_{12}(.)$ notation needs to be defined.

> ***This has been defined.***

R1.4: 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?

> ***Thank you for the interesting comment/question. In this paper, our main goal and applications deal with bipartite record linkage (such that we can make fair comparison with prior work), which is consistent with the algorithm that we present in the paper. We have removed the confusing text mentioned by the reviewer. We are working on extensions beyond bipartite record linkage in other work, where the reverse sizing is not an issue, however, this is beyond the scope of this paper.***

R1.5: 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.

> ***Our missingness assumptions and the definition of $I_{obs}()$ are near the end of Section 3.***

R1.6: Section 3, full model: The Phi notation needs to be defined.

> ***This has been defined.***

R1.7: Section 3, full model: The square brackets seem to be misplaced, as the exponent includes indices of the sum contained in the brackets.

> ***The expression is now corrected.***

R1.8: 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 $\Gamma_{\cdot j}$ seems to be missing a factor of $u_{fl}^{I}(gamma_{ij}^{f} = l)$. 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.

> *We have revised our presentation of the Gibbs sampler with this feedback in mind. The main body of the text now presents the final full conditionals, with revised derivations moved now in the appendix. In particular, we have included an extra step in the derivation of $p(\Gamma_{\cdot j})$ that shows exactly how we divide by $\prod_{i=1}^{n_1}\prod_{f=1}^{F}\prod_{\ell=1}^{L_f} u_{fl}^{I(\gamma_{ij}^{f}=l)}$ and use proportionality to arrive that the distribution presented.*

R1.9: 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.

> ***The contribution of Z in the full conditional is just the prior distribution, which we have expressed with summation notation to reflect the piecewise nature of the prior. The binomial-looking expression at the top of page 8 is the posterior distribution for π. We note that the posterior $p(\pi|\Gamma, Z, \Phi)$ in fact only depends on Z, and is thus shortened to $p(\pi|Z)$. We have updated our explanation to make this more clear.***

R1.10: Section 3.1, first equation on page 8: There is an extra parenthesis in the exponent of (1-pi).

> ***This has been corrected.***

R1.11: Section 3.1, pmf for $Gamma_{\cdot j}$ 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.

> ***We have changed w_{ij} to $w_{z_j,j}$ in the pmf for $\Gamma_{\cdot j}$ and the full conditional for Z_j. We have been sure that "$n_1 + 1$" has been changed to "$n_1 + j$"***

> ***throughout the paper.***

R1.12: 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.

> ***We have revised the presentation of the Gibbs sampler based on this feedback. The main body of the paper presents the results, and revised derivations (including justification for integrating out π) have been placed in the appendix.***

R1.13: Section 4.2, SEI procedure. I'm not sure I follow the description of the SEI procedure. (1) 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. (2) 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. (3) Can you quantify the computational savings due to this method? (4) Do you have advice for how small S can be? (5) In your simulations, how does the use of this method or choice of S affect the results?

> *Brian: There are five questions here that should be answered. The first two should be yes, the understanding is correct or no, it's not correct. (If the reviewer doesn't understand, then provide an explanation of your approach. For (3), we quantify the SEI procedure in Section 4.2, where we state: "Rather than storing $n_1 \times n_2$ record labels, SEI allows us to store at most $n_2 \times P \times S$ labels, regardless of how large n_1 might be." SEI does not affect the computation time of the Gibbs sampler, since the complexity of the sampler is $O(n_2 P)$, regardless of how many comparison vectors are of pattern P. Turning to (4), in practice, we recommend $S = 10$, as this reduces the number of stored indices for highly unlikely record pairings, but does is not likely to eliminate any of the indices for likely matches. Choosing S too low can concentrate undue mass on unlikely matches and distort linkage results. I don't see an answer to the question (5).*

R1.14: Section 4.4, Assumptions and definitions should be included in the statement of this lemma.

> ***The lemma now includes assumptions/definitions for all variables used within the statement.***

R1.15: 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?

> *Regardless of the specific construction of the comparison vectors, the computational complexity the Gibbs sampler under BRL is $O(n_1 n_2)$, where as it is $O(n_2 P)$ for fabl. The exact shape of the lines will be different, but in all cases, fabl removes dependence on n_1 in the computational complexity. To illustrate this, we have included in the appendix an additional simulation study under different comparison vector settings.*

R1.16: 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?

> *In the caption under Figure 3, we explain that we have 100 pairs of sets of records for each level of overlap and error. Thus the table represents results over 900 pairs of datasets. We have added this description in the body of the paper as well for clarity.*

R1.17: Section 5.3 simulation – RR results are not presented.

> *We have updated the Figure XXXX with these results. Since its is presented along side NPV and PPV, for which values close to 1 indicate strong performance, we have presented decision rate (DR) results.*

R1.18: Section 6.1 last sentences. I believe the last sentence should be removed, as it is redundant.

> ***This has been removed.***

R1.19: Section 6.2, second sentence. I believe the word "survey" is missing, as in "the initial SURVEY began. . . "

> ***This has been fixed.***

R1.20: 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?

> *In each iteration of the study, some participants are added to the lists, and others are lost due to loss of contact or death. This has been clarified in the paper. [TODO]*

R1.21: Figures 5 and 7 are not referenced in the text.

> ***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]**