# Response to Reviewer Comments

July 30, 2023

We thank the Editor and Associate Editor for their assessment and constructive feedback.

## 1 AE Comments

**AE.0:** I must add that reading the paper still feels like a chore, in part because the notation is so heavy, which may be hard to avoid, but also, in part, because some wording choices make certain sentences hard to parse, which can certainly be improved.

> *We have made several revisions to the paper, as described in our responses below.*

**AE.1:** Line 80 on p. 3 says: "For ease of readability, we follow the convention established by Sadinle (2017) and say "record $i \in X_1$" rather than the more compact $x_{1i}$." Taking this at face value implies that, in any sentence, one could swap "record $i \in X_1$" for "$x_{1i}$," which clearly is not true. So, as it stands, the sentence does little to improve ease of readability. In fact, if the point is that "$x_{1i}$" will never be used again, why is the notation introduced in the first place?

> *We now refer to the two data files as $A$ and $B$. The updated material now reads:*
>
> *"Consider two data files $A$ and $B$ comprising $n_A$ and $n_B$ records, respectively, and including $F$ linkage variables measured in both files. For $i = 1, \ldots, n_A$, let record $i$ be given by $A_i = (A_{i1}, \ldots, A_{iF})$, so that $A = (A_i : i = 1, \ldots, n_A)$. Similarly, for $j = 1, \ldots, n_B$, let record $j$ be given by $B_j = (B_{j1}, \ldots, B_{jF})$, so that $B = (B_j : j = 1, \ldots, n_B)$. Without loss of generality, denote files such that $n_A \geq n_B$.". See Section 2.*

## 2 Associate Editor Comments

**R.0:** For the missing data treatment now included just before Section 2.1, is the assumption truly missing at random, or missing completely at random? Or, does this distinction not matter because of the independence assumed across elements of the comparison vector?

> *We have included the following explanation of missingness in Section 2.*
>
> *"In the construction of comparison vectors, it is common to encounter missing information in record $A_i$ or $B_j$. As a result, the comparison vector $\gamma_{ij}$ will have missing values. We assume that this missingness occurs completely at random (MCAR, per Little and Rubin (2002)). To notate a missing value in any $\gamma_{ij}^f$, we use $I_{obs}(\gamma_{ij}^f) = 1$ when $\gamma_{ij}^f$ is observed and $I_{obs}(\gamma_{ij}^f) = 0$ otherwise. With the MCAR assumption, we can marginalize over the missing data, and do all computation simply using the observed data."*

**R.1:** Appendix 8.2: I appreciate the streamlining of the discussion of the proposed algorithms and think it was a good choice to move the full derivation to an appendix. However, I do not follow the re-expression of the pmf for $\Gamma_{\cdot j}$. First, the opening square brackets are still misplaced throughout. In the second line of this derivation, the authors divide the expression by a product of the element-wise conditional probabilities of a match in the comparison vector (u), raised by an indicator that the element of the comparison vector equals a particular value. Perhaps I misunderstand, but I believe this product is not constant in $\Gamma_{\cdot j}$ or $u$, and so the total expression is not proportional to the line above, as interpreted as a pmf. I think that the end effect is that the last line of the newly expressed pmf is missing a factor of $\prod_i \prod_f \prod_l u_{fl}^{I(\gamma_{ij}^f = l) I_{obs}(\gamma_{ij}^f)}$ regardless of the value of $z_j$.

> *The square brackets have been corrected.*
>
> *We have revised the derivation of the full conditional distribution of $Z_j$, separately handling the case when $B_j$ has a match and when it does not. Refer to Appendix A for this extended derivation.*

**R.2:** Appendix 8.2: Thank you for including the details on integrating out pi from the full conditionals. They surprised me. I had assumed that the authors had integrated out pi in the prior for Z, as this is what Sadinle (2017) had done to form the "beta prior for bipartite matchings". (1) Does the alternate approach presented in this paper provide a different algorithm than directly integrating the prior distribution? (2) Is the presented alternative approach justified?

> *We realize that the integration of $\pi$ out of the full conditionals is not necessary for our sampler. Therefore, we have updated the paper with full conditionals for the standard Gibbs sampler.*
>
> *We have redone all simulations and case studies using a Gibbs sampler that samples $\pi$ from its full conditional, and then samples $\boldsymbol{Z}$ from its full conditionals. We obtained equivalent results.*

**R.3:** I do not understand the statement just below Equation (10): "When j has no match in $X_1$, we write $(n_1 + j, j) \in h_{P+1}$" My understanding of these patterns is that they are based on observed comparison vectors without consideration of Z (matches). In the second paragraph of Section 4.3, the H notation includes the matches (Z), in notation and definition that seems to conflict with the statement just below Equation (10). This also comes into play in Equation (16).

> *Yes, the comparison vectors are created (and therefore, patterns are assigned) without regard to $\boldsymbol{Z}$. We have removed that sentence, and updated the equations in the Section 4.3 to avoid confusion.*

**R.4:** Third paragraph of Section 4.1: The authors claim they are computing "sufficient statistics". What exactly are these statistics sufficient for?

> *We have revised the text to state "summary statistics" instead of "sufficient statistics."*

**R.5:** First paragraph of Section 4.2: The authors state: "Posterior calculations still attribute the appropriate weight to all records through the summary statistics..." What is meant by the term "weight"? Which records are appropriately weighted – those in X2?

> *Our intended meaning was "All record pairs contribute to the likelihood through the summary statistics."*
>
> *In our revised submission, we have re-expressed the likelihood from (6a) through the perspective of agreement patterns in (14). This form of the likelihood is then used to express contributions to the likelihood for the $\boldsymbol{m}$ and $\boldsymbol{u}$ parameters in (15) and for $\boldsymbol{Z}$ in (17), both of which use the summary statistics. We hope that this revision makes this point more clear.*

**R.6:** Second paragraph of Section 4.2: "and delete those comparison vectors". Which are "those" vectors?

> *In the original manuscript, "those comparison vectors" refer to removing the larger (and expensive)* $\Gamma^{ab}$ *from memory and continuing our calculations with the compressed comparison vectors* $\tilde{\Gamma}^{ab}$.
>
> *The revised text states: "Then, we conduct hashing, obtain the compressed comparison matrix,* $\tilde{\gamma}^{ab}$*, and remove the memory-intensive comparison matrix,* $\gamma^{ab}$*, before continuing with the next batch of data. See Section 4.3.*

**R.7:** Where does $R^{SEI,cd}$ come into play in the partitioned algorithm presented in Equations (13) and (14)? I recommend that the authors either refrain from suppressing the SEI notation or further explain how the SEI algorithm has changed the quantities in these equations.

> *To make the SEI algorithm more clear, we have revised the ordering of Sections 4, so that we have 1) hashing, 2) posterior inference, 3) chunkwise computation of the comparison matrix, and 4) SEI. With this new ordering, and with the newly provided equations (15) through (19), we hope it is clearer that that the contribution of* all *record pairs is recorded through the summary statistics in* $\mathcal{N}$*. In particular, (16a), (16b), and (18) depend only* $\mathcal{N}$*. The SEI algorithm only affects the step shown in (19).*

**R.8:** Section 4.2: I appreciate the practical advice about choosing S for the SEI method. However, this choice seems arbitrary in the absence of further discussion/evidence. Given that the primary novelty of the manuscript is in methods to speed and otherwise improve computation, I am surprised that this aspect of computational innovation is presented with virtually no theoretical or empirical exploration. Presumably the SEI method has some sort of accuracy trade-off, as the authors warn that linkage results may be "distorted" if S is low. However, this trade-off is not quantified or even discussed in practical terms beyond the terse recommendation to choose S=10.

> *We have included an additional simulation in Section 5.3, which explores the trade-offs regarding different choices of* $S$*.*

**R.9:** Page 5, 2 sentences before equation (4): I believe the sum should be of $I(Z_j \leq n_1)$, not $I(Z_j \leq n_1 + 1)$.

> *This has been corrected.*

**R.10:** Equation (6a): The indices do not match the subscripts in the indicator function in each summand, or their standard meaning in table 1.

> *The typo has been fixed.*

**R.11:** Generally, the authors seem to arbitrarily use upper and lower case z interchangeably in function definitions.

> *We use $\boldsymbol{Z}$ when discussing a random quantity. In all instances where we discuss the random variable $Z_j$ taking on a particular value, we now use $q$ to avoid confusion and to follow notation used by Sadinle (2017).*

**R.12:** Equations (8) and (9): Should the weights have superscript (s) (as the Zs do)?

> *This has been fixed. See Equations (9) and (10).*

**R.13:** Gamma is in some places described as a set and in others as a matrix (particularly in sections 4.2 and 5.1).

> *The comparison matrix $\gamma$ is comprised of the comparison vectors for all record pairs in $A \times B$. All references to $\gamma$ as a set have been removed. See Section 2, where we define both the comparison vector and comparison matrix.*

**R.14:** Section 4.3, second paragraph, the definitions of the concatenated vectors $\alpha_0$ and $\beta_0$ should have final elements subscripted by $L_f$, where the sub-subscript is capitalized.

> *This has been revised; thank you. See Section 4.2,*

## References

Little, R. and Rubin, D. (2002), *Statistical Analysis with Missing Data*, Wiley, Hoboken, New Jersey.

Sadinle, M. (2017), "Bayesian Estimation of Bipartite Matchings for Record Linkage," *Journal of the American Statistical Association*, 112, 600–612.