Response to Reviewer Comments

July 20, 2023

We thank the Editor and reviewers 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.

> *Thank you for the opportunity to submit a revision. We have provided detailed responses to the reviewers' comments below, which prompted significant changes in our article. Specifically, we have attempted to improve notation and wording choices, at the reviewers' suggestion.*

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?

> *Thank you for the suggestion. We now refer to the two data files as A and B, where the updated material reads:*
>
> *"Consider a data files A and B, consisting of records A_i and B_j respectively, where $i \in \{1, \ldots n_A\}$ and $j \in \{1, \ldots n_B\}$". See Section 2, page 3.*

2 Reviewer 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?

> *The distinction does not matter due to assuming that the comparison vectors are conditionally independent given the coreference matrix (or matching label). We now state the following in Section 3, page 5:*
>
> *"We now explain how we accommodate missing fields, which leads to missing comparison vectors. To denote a missing value in any γ_{ij}^f, we use $I_{obs}(\gamma_{ij}^f) = 1$ when γ_{ij}^f is observed and $I_{obs}(\gamma_{ij}^f) = 0$ otherwise. Assume the comparison vectors are conditionally independent given the coreference matrix (or matching label) and that missing comparisons are missing at random (MAR) or missing completely at random (MCAR). It is important to note that for likelihood based inference, both MAR and MCAR lead to ignorability, or rather, being able to consider the likelihood for the observed data (Little and Rubin, 2002, Section 6.2). Therefore, we can marginalize over the missing data and perform all computation using the observed data, leading to (6a). For further details on these connections, refer to (Little and Rubin, 2002, Section 6.2), (Sadinle, 2014, Section 3.1), and (Sadinle, 2017, Section 4.2)."*

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 $\Gamma_{.j}$. 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 $\Gamma_{.j}$ 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, pages 26 – 27 regarding 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?

> *No, in fabl, we do not integrate out π as done in Sadinle (2017) because this leads to a sequential sampler.*
>
> *(1) Yes, our alternative approach is different than directly integrating over the prior distribution π. In short, our prior distribution over Z leads to a Gibbs sampler that leads to parallel updates, which contrasts that of Sadinle (2017). Of course, we could integrate out π in our approach, however, this leads to sequential updates, which would be slow in practice. See Appendix A, page 27.*
>
> *(2) Yes, our alternative approach is justified. We provide a justification in Appendix A, page 27 regarding why the updates are independent and not sequential. Thank you for the excellent questions, which have greatly improved our paper.*

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, in the original manuscript, the comparison vectors are created and therefore, patterns are assigned) without regard to Z. Given the confusion, we have removed the sentence.*

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

> *Thank for pointing out our typo. 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?

> *In the original manuscript, we consider all records in the first file (X_1) that share an agreement pattern with the second data file (X_2). In this situation, all these*

> *records have the same Fellegi-Sunter likelihood ratio weight, w_{ij}. Given this result, we denoted the weight w_p since these records map to the same agreement pattern.*
>
> *In the revised manuscript, we define m_p and u_p, which are the probabilities that records A_i and B_j form agreement pattern p given that they are a match and non-match, respectively. For each pattern p, we define $w_p = m_p/u_p$. We hope that this will be more clear than our previous presentation of the material. See Section 4.1, pages 8–9.*

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) Γ^{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, γ^{ab}, before continuing with the next chunk of data. See Section 4.3, page 10.*

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.

> *In order to attempt to make the SEI algorithm more clear, we have revised the ordering of Sections 4.2 and 4.3, so that 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 -19, we hope that the contribution of* all *record pairs is recorded through the summary statistics in $\mathcal{N}$ is now more clear. In addition, we hope that the SEI algorithm updates will now be more clear that the updates in (16a), (16b), and (18) only depend on $\mathcal{N}$. The SEI algorithm 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.

> *Refer to Section 5.3, which explores the trade-offs regarding different choices of S and discusses this in practical terms.*

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; thank you.*

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 Z when discussing a random quantity, and z to reference a realized value. We have revised all instances of the inconsistency.*

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

> *This has been fixed. Refer to equations (12) and (13), page 7.*

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).

> *Note that γ is a comparison matrix comprised by comparison vectors. All references to γ as a set have been removed. See Section 2, page 3, where we define both the comparison vector and comparison matrix.*

R.14: Section 4.3, second paragraph, the definitions of the concatenated vectors α_0 and β_0 should have final elements subscripted by L_f, where the sub-subscript is capitalized.

> *This has been revised; thank you. Refer to Section 4.2, page 10.*

References

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

Sadinle, M. (2014), "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach," *The Annals of Applied Statistics*, 8, 2404–2434.

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