

Response to Reviewer Comments

July 17, 2023

Thank you for the thoughtful feedback. We have made several broad changes to address the Associate Editor's overall critiques, and have addressed each specific comment from the reviewer.

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.

Suggestion: Write a strong paragraph to the Editor after the paper is finalized. Come back to this. Explain that we have tried to simplify the notation when possible, however, in some places, the heavy notation is not possible. Thus, we have worked on improving more precise wording and removing sections or wording that is distracting or not needed. Thank the Editor and Associate Editor for helping us improve upon what we believe to be an important paper and breakthrough in the literature.

I would not make a large list of all the notation we have changed as it may cause the Editor to go picking over the paper more.

We thank the Editor for the opportunity to revise our paper and for the suggestions. We have improved notation, when possible, however, as noted by the Editor in certain places, such as in the computational speeds ups of Section 4, this is difficult to avoid. We have done our utmost to simplify notation (when possible) and improve the parsing of sentences. Finally, we have attempted to make it more clear the general contributions of our paper, which we highlight for the Editor below.

We propose fast beta linkage (*fabl*), which extends the BRL model for increased efficiency and scalability. We use independent priors for the matching status of each record, inducing a dependency that is stronger than the record pair independence in the original Fellegi and Sunter (1969) but weaker than the one-to-one requirement of Sadinle (2017). This independence assumption allows for a scalable update of the Gibbs sampler, contrasting sequential updates of Sadinle (2017). We employ the technique of Steorts et al. (2016) in order to ensure our method after record linkage is bipartite. This discovery allows us to (1) employ hashing techniques that speed up calculations and reduce computational costs, (2) compute the pairwise record comparisons over large data files via parallel computing, and (3) reduce memory costs through what we call storage efficient indexing. These contributions allow *fabl* to perform record linkage on much larger data files than previous Bayesian Fellegi-Sunter models at significantly increased speed with similar levels of accuracy. In particular, computation time under BRL grows quadratically, with the size of each data file, while computation time under *fabl* grows linearly, only with the size of the smaller data file.

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 the Editor, explain the change. I have removed repetition and shortened this bit.

We thank the Editor for bringing up this point. In order to simplify notation, we 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 $i \in \{1, \dots, n_A\}$ and $j \in \{1, \dots, n_B\}$ ”.

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?

See two comments below regarding MAR versus MCAR.

Answer for MAR: Yes, the assumption is truly missing at random. This distinction between truly missing at random and missing completely at random does not matter due to two assumptions. Following Sadinle (2017), we assume that the fields in the comparison vectors are conditionally independent given the coreference matrix and we assume that the missing comparisons are ignorable (Sadinle (2017)). Therefore, our likelihood is the same as in Sadinle (2017) (see equations (3) and (4) of his paper). As stated by Sadinle (2017), “we work under the assumption of ignorability of the missingness mechanism for the comparisons so that we can base our inferences on the marginal distribution of the observed comparisons (Little and Rubin, p. 90).” We refer to Sadinle (2017), Section 4.2, page 604 for further details.

Answer for MCAR: The assumption is missing completely at random due to fact that we assume the fields in the comparison vectors are conditionally independent given the coreference matrix. The assumption is not missing at random as we do not assume that the missing comparisons are ignorable.

The assumption is missing completely at random. You are correct that this is a result of the independence assumed across elements of the comparison vector. This has been clarified.

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 .

Don’t apologize!

The square brackets have been corrected. We have clarified 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.

The square brackets have been corrected, and we apologize for the oversight. We have clarified derivation of the likelihood for Z_j , separately handling the case when B_j has a match and when it does not.

R.3: Appendix 8.2: Thank you for including the details on integrating out π from the full conditionals. They surprised me. I had assumed that the authors had integrated out π 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?

Brian: I think it’s very important that we answer this carefully and in a strong manner. See below. Note: The edits in Appendix 8.2 regarding the justification that I wrote have been removed.

No, we do not integrate out π as done in Sadinle as this would result in a sequential sampler as opposed to the parallel one that we propose. This is one of our contributions to our paper. (1) Yes, our alternative approach is different than directly integrating over the prior distribution. We have attempted to make this more clear in our revised version. In short, our prior distribution over Z leads to a Gibbs sampler that leads to parallel updates, which contrasts that of Sadinle (2018). Of course, we could integrate out π in our approach, however, this would lead to sequential updates, which would be slow in practice. (2) Yes, our alternative approach is justified. We provide a justification in Appendix 8.2 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)

Comments: Tell the AE they are correct and mention where you correct the confusion. Then move on. (No need to state what you intended). Mention what equations numbers have been updated to help the AE now given the major revision. **Please redo this one to answer the question more directly.**

You are correct that the comparison vectors are created (and therefore, patterns are assigned) without regard to Z . This notation was created purely to be able to denote when a record was left unlinked during the Gibbs Sampler as in Equation 16 in the submitted draft.

I can see why this is confusing. Therefore, we have removed that line, and changed the notation for record B_j being unmatched to just be “otherwise”.

R.4: Third paragraph of Section 4.1: The authors claim they are computing “sufficient

statistics”. What exactly are these statistics sufficient for?

Comment: Make it more brief and no need to provide notation when it’s not needed.

Thank for pointing out our typo. We have revised the text to state “summary statistics” instead of “sufficient statistics.”

We show in the revised Section 4.2 that we can write conditional likelihoods and posterior updates for \mathbf{m} , \mathbf{u} , and \mathbf{Z} using the statistics in $\tilde{\Gamma} = \{\mathcal{P}, \mathcal{R}, \mathcal{N}\}$. However, we cannot express the full likelihood in (14) through these statistics. Therefore, we have revised to text to say “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 \mathbf{X}_2 ?

Comment: Brian, you have not answered the reviewer’s question regarding what is meant by weight and what records are appropriately weighted. **Please revise to directly answer the question in the revised text with equation numbers.**

We have revised the ordering of Sections 4.2 and 4.3, so that we first discuss hashing, then posterior inference, then chunkwise computation of the comparison matrix, and then 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.

We have switched the order of Sections 4.2 and 4.3, so that we first discuss hashing, then posterior inference, then chunkwise computation of the comparison matrix, and then SEI. With this new ordering, and with the newly provided equations (15 -19) it is clearer that the contribution of all record pairs is recorded through the summary statistics in \mathcal{N} .

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

“Those comparison vectors” referred 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 vectors $\tilde{\gamma}^{ab}$, and remove the memory-intensive comparison vectors 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.

We have reordered Section 4 so that all posterior inference is presented before SEI. This makes it clear that the posterior updates in (16a), (16b), and (18) depend only on \mathcal{N} . SEI 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 provided an additional simulation study in Section 5.3 to show the sensitivity to 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, 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.

This 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 inconsistency.

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

This has been fixed.

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).

γ is a matrix comprised by comparison vectors. All references to γ as a set have been removed.

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

References

- Fellegi, I. P. and Sunter, A. B. (1969), “A Theory for Record Linkage,” *Journal of the American Statistical Association*, 64, 1183–1210.
- Sadinle, M. (2017), “Bayesian Estimation of Bipartite Matchings for Record Linkage,” *Journal of the American Statistical Association*, 112, 600–612.
- Steorts, R. C., Hall, R., and Fienberg, S. E. (2016), “A Bayesian Approach to Graphical Record Linkage and Deduplication,” *Journal of the American Statistical Association*, 111, 1660–1672.