Review of: BA2112-029RA0

Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl)

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.

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.
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).
3. Section 2.2, beta distribution: the n_{12}(.) notation needs to be defined.
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?
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.
6. Section 3, full model: The Phi notation needs to be defined.
7. Section 3, full model: The square brackets seem to be misplaced, as the exponent includes indices of the sum contained in the brackets.
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_{.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.

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.
10. Section 3.1, first equation on page 8: There is an extra parenthesis in the exponent of (1-pi).
11. 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.
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.
13. 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?
14. Section 4.4, Assumptions and definitions should be included in the statement of this lemma.
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?
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?
17. Section 5.3 simulation – RR results are not presented.
18. Section 6.1 last sentences. I believe the last sentence should be removed, as it is redundant.
19. Section 6.2, second sentence. I believe the word "survey" is missing, as in "the initial SURVEY began…"
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?
21. Figures 5 and 7 are not referenced in the text.