General Comments

1 Suggestions

1. Let us read through the paper extremely carefully (everyone, including myself and Jerry) to make recommendations and formulate a plan.

2. I think it would be helpful to make comments outside the main document before making major changes in the document to make sure we address everything and check that we address every comment that was made by the Editor/AE.

3. Should try and make clear and crisp answers to them (and cite existing literature to avoid opening any holes or new issues).

4. I will do my best to serve as a "reviewer/editor" and find other comments to fix.

5. If you find yourself getting frustrated, etc. please reach out/let me know. One idea is to work on another project that might be more fun or start formulating the rest of your thesis in case this might be weighing you down. I'd be happy to sketch this out with you and help you as much as you might like, Brian.

2 General Comments

1. The connection between the relaxation of Sadinle (2018) of the bipartite linkage structure may be getting lost due to AE comment 3 as they stated "they surprised me" regarding how the fabl prior was defined. This should be looked at more carefully and introduced with more care regarding the contribution of the prior as a generative (Bayesian) prior with post-processing after due to the computational issues faced with that of Sadinle. (This is AE comment 3).

2. Section 2 could be structured to be more clear to a general reader regarding notation overall, prior work (FS and Sadinle) and then the relaxation of the Sadinle (2018) prior. (Editor comment).

3. The overall paper needs extensive checking regarding consistent notation, consistent a, no repetitive comments, no typos, etc. (Editor comment).

4. Appendix 8.2 needs to be provided without any typos and all details provided. (Don't skip any steps and do not use any short hand notation). I think that this was the main point of confusion for the reviewer. You may wish to refer to Aleshin and Guendel and Sadinle (2022) and follow their notation as it appears this may be what you want. See Appendix B.2 page 27. (In short, they use the notation $\mathcal{L}_{ij}$ to make it clear that this is a factor coming from the likelihood). I think that you could make your proof more clear for those working on this topic, which make this an opportunity and opening as it wasn't clear to the AE (and it's probably not obvious to others).

5. Let's walk through each part of AE comment 3. I think it would be helpful to refer to Appendix B.2 of Aleshin and Guendel and Sadinle (2022) regarding this discussion. Page 25–26.

   "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."

   "Does the alternate approach presented in this paper provide a different algorithm than directly integrating the prior distribution."

   Yes, as mentioned, our approach is a relaxation of that of Sadinle (2017) as we relax the one-to-one assumption. Our algorithm provided is not the same, however, empirically we find the the results are similar due to a pre-processing step that has been utilized in the literature by Jaro (??).

   "Is the presented algorithm justified?"

   It will not be justified unless the full conditionals of the Gibbs sampler contain a sequential update or one proves that this is a justified Gibbs sampler. This was the point that Serge mentioned in his email. Currently, it is not a valid Gibbs sampler.

   It would need to be something like this, which comes from Aleshin-Guendel and Steorts (2023). The partition is equivalent to the co-reference matrix or linkage structure. The partition approach is more widely understood by the Bayesian non-parametrics community (or statistics community), which is why we are using this terminology. (They will not easily understand the linkage structure or coreference matrix without connections being made for them).

   > " The most common way to update the partition, $\mathcal{C}$, in Bayesian entity resolution is the Gibbs update [**?** **?** **?** **?** **?** **?** **?**], which originated in the Bayesian clustering literature [see e.g. **?**]. We emphasize that this is only referring how the *partition* is being updated.
   >
   > The Gibbs update works by sampling the cluster assignment for each record $i \in [R]$ sequentially. To describe the update we will

represent a partition $\mathcal{C}$ by a vector of labels $\boldsymbol{Z} = (Z_1, \ldots, Z_R)$, where $Z_i \in \{1, \ldots, R\}$ such that $Z_i = Z_j$ if records i and j are in the same cluster in $\mathcal{C}$. Suppose that we have already sampled the cluster assignment for the first $i-1$ records. Sampling the cluster assignment for record i proceeds as follows:

(a) Let $\boldsymbol{Z}_{-i}^{[t]} = (Z_1^{[t]}, \ldots, Z_{i-1}^{[t]}, Z_{i+1}^{[t-1]}, \ldots, Z_R^{[t-1]})$ denote the current partition of $[R]$ after removing record i.

(b) Let $\boldsymbol{c}$ denote the union of the unique labels occupied in $\boldsymbol{Z}_{-i}^{[t]}$ and one unoccupied label. Each of these labels correspond to a cluster to which record i could be assigned, with the unoccupied cluster representing the creation of new cluster with only record i.

(c) Sample record i's cluster assignment, by sampling the cluster $c \in \boldsymbol{c}$ with probability proportional to $p(Z_i^{[t]} = c \mid f(\boldsymbol{X}), \boldsymbol{Z}_{-i}^{[t]}, \boldsymbol{\theta}^{[t]})$.

"

6. I have comments about the hashing section, which are largely that I think the AE/Editors are having a difficult time with this section. I think their main concerns are with the following **notation, precise wording, understanding of the paper, correctness of the appendix**. I would suggest that we all carefully read through the paper and have a plan, where we are all on the same page regarding our response and revisions.

3 Language/Writing

1. There is repetition present in the manuscript, such as the first two lines of the opening paragraph of the introduction. For example, analysis, analysts, and analyses are used three times in the first two sentences. Tasks is used twice. Furthermore, the language could be written to be more precise and condensed. "In many applications, users seeks to identify duplications across two data files." (This removes the repetition, and I think the remainder can stay).

2. This should be checked in an editor checked such as Antidote closer to the time of resubmission.

4 Notation

1. The notation is not clear or consistent throughout.

2. The notation is confusing at times when reading through the paper, where one has to flip back and forth between pages to re-reference previously defined notation. The authors should more carefully define (perhaps following that of Sadinle (2018) to avoid typos) and use referencing of equations

so that the reader doesn't have to constantly flip back and forth throughout the manuscript.

3. The usage of lower and upper case greek letters should be consistent. The usage of subscripts should be consistent. It appears that Sadinle (2018) and Aleshin and Guendel and Sadinle (2022) make use of notation consistently, so this could be followed. It might be worth noting that they utilize X_1 and X_2 in order to more effectively extend to any number of databases, so this might be more natural than A and B.

5 AE Comments with responses

The comments to the AE need to be extremely precise and accurate. I have attempted to write one in painful detail to provide guidance and an example.

1. Line 111. Is this missing at random or something else? This needs to be clarified (AE Comment 1). This needs to be addressed with extreme care regarding the precise wording.
   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.

2. 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
   Is the new proof, Brian, missing an indicator over the cluster assignment? I believe that it should match those of Sadinle (2017), Aleshin and Guendel (2022), and Wortman (2019). Where is the indicator regarding the cluster assignment that you condition on? Specifically, the update for each gamma will depend upon your cluster assignment, so I believe there is a typo for this reason. (To help with notation, Jody calls this S_k, Sadinle calls it Z_{-j}, and Serge calls it $\mathcal{C}_{-j}$). Could you carefully check this please? I think this question is related to the next point/comment by the AE.

3. 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". Does the alternate approach presented in this paper provide a different algorithm than directly integrating the prior distribution? Is the presented

alternative approach justified?
Thank you for noting your surprise regarding our proposal. We do not extend the approach of Sadinle (2017), but relax the one-to-one assumption for computational reasons. This motivates the choice of our prior (instead of the beta prior for bipartite matchings). Note: I do not believe that we should be integrating out the prior distribution. I think we did this at the confusion from the first revision of the AR. We should instead just update the conditionals to form a Gibbs update, which will make the algorithm justified. If you believe that the algorithm is correct, we will need to convince the AE why it is a valid Markov chain or rather Gibbs update of the partition. It will not be justified unless the full conditionals of the Gibbs sampler contain a sequential update or one proves that this is a justified Gibbs sampler. This was the point that Serge mentioned in his email. Currently, it is not a valid Gibbs sampler. I attempt to explain this more clearly below.

It would need to be something like this, which comes from Aleshin-Guendel and Steorts (2023). The partition is equivalent to the co-reference matrix or linkage structure. The partition approach is more widely understood by the Bayesian non-parametrics community (or statistics community), which is why we are using this terminology. (They will not easily understand the linkage structure or coreference matrix without connections being made for them).

> " The most common way to update the partition, $\mathcal{C}$, in Bayesian entity resolution is the Gibbs update [**? ? ? ? ? ? ? ?**], which originated in the Bayesian clustering literature [see e.g. **?**]. We emphasize that this is only referring how the *partition* is being updated.
>
> The Gibbs update works by sampling the cluster assignment for each record $i \in [R]$ sequentially. To describe the update we will represent a partition $\mathcal{C}$ by a vector of labels $\boldsymbol{Z} = (Z_1, \ldots, Z_R)$, where $Z_i \in \{1, \ldots, R\}$ such that $Z_i = Z_j$ if records i and j are in the same cluster in $\mathcal{C}$. Suppose that we have already sampled the cluster assignment for the first $i-1$ records. Sampling the cluster assignment for record i proceeds as follows:
>
> (a) Let $\boldsymbol{Z}_{-i}^{[t]} = (Z_1^{[t]}, \ldots, Z_{i-1}^{[t]}, Z_{i+1}^{[t-1]}, \ldots, Z_R^{[t-1]})$ denote the current partition of $[R]$ after removing record i.
>
> (b) Let $\boldsymbol{c}$ denote the union of the unique labels occupied in $\boldsymbol{Z}_{-i}^{[t]}$ and one unoccupied label. Each of these labels correspond to a cluster to which record i could be assigned, with the unoccupied cluster representing the creation of new cluster with only record i.
>
> (c) Sample record i's cluster assignment, by sampling the cluster $c \in \boldsymbol{c}$ with probability proportional to $p(Z_i^{[t]} = c \mid f(\boldsymbol{X}), \boldsymbol{Z}_{-i}^{[t]}, \boldsymbol{\theta}^{[t]})$.

”