Population Sized Graphical Record Linkage

Rebecca C. Steorts
Departments of Statistical Science
and Computer Science
Duke University
beka@stat.duke.edu

Abstract

1 Introduction

do later.

1.1 Prior work

do later.

2 Bayesian Record Linkage

We first introduce a very general way of performing graphical record linkage, and then introduce a very natural approach for providing the uncertainty of the record linkage process into subsequent analysis.

2.1 Notation

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ represent the data, with k databases, indexed by i. The ith list has n_i observed records, indexed by j. Each record corresponds to one of N latent entities, indexed by j'. Assume $N = \sum_{i=1}^{k} n_i$ without loss of generality. Each record or latent entity has values on p fields, indexed by ℓ, and are assumed be categorical and the same across all records and entities [**?**, **?**]. M_ℓ denotes the number of possible categorical values for the ℓth field. In both models, $X_{ij\ell}$ denotes the observed value of the ℓth field for the jth record in the ith list, and $Y_{j'\ell}$ denotes the true value of the ℓth field for the j'th latent entity. Then Λ_{ij} denotes the latent entity to which the jth record in the ith list corresponds, i.e., $X_{ij\ell}$ and $Y_{j'\ell}$ represent the same entity if and only if $\Lambda_{ij} = j'$. Then $\boldsymbol{\Lambda}$ denotes the Λ_{ij} collectively. Distortion is denoted by $z_{ij\ell} = I(X_{ij\ell} \neq Y_{\Lambda_{ij}\ell})$, where $I(\cdot)$ denotes the indicator function. As usual, I represents the indicator function (e.g., $I(x_{ij\ell} = m)$ is 1 when the ℓth field in record j in file i has the value m), and let δ_a denote the distribution of a point mass at a (e.g., $\delta_{y_{\Lambda_{ij}\ell}}$).

Remark: In order to take into the record linkage uncertainty exactly, we require looking at the entire graphical space, or rather looking at duplication within and across lists.

Question: What would the bound be on the record linkage and speed up if we simply did de-duplication? This would be interesting to look.

3 A Generalized Bayesian Graphical Record Linkage Model

Let us assume the following model:

$$\begin{aligned} X_{ij\ell} \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell} &\overset{\text{ind}}{\sim} G && (1)\\ Y_{j'\ell} &\overset{\text{ind}}{\sim} H \\ z_{ij\ell} \mid \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_{i\ell}) \\ \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Beta}(a, b) \\ \Lambda_{ij} &\overset{\text{ind}}{\sim} \text{KP}, && (2) \end{aligned}$$

where all distributions are also independent of each other; assume that a, b, N are assumed known. We refer to the resulting class of marginal distributions over $\boldsymbol{\Lambda}$ as Kolchin partition (KP) models [**?**, **?**] because the form of

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

these is closely related to Kolchin's representation theorem for Gibbs-type partitions ([**?**, theorem 1.2]). [[It would probably be good to show that the EB model exhibits the microclustering property as was done in the NIPS paper]].

Remark: observe that record linkage and de-duplication are both simply a question of whether $\Lambda_{i_1,j_1} = \Lambda_{i_2,j_2}$, where $i_1 \neq i_2$ for record linkage and $i_1 = i_2$ for de-duplication.

1. We should talk about advantages of such models.
2. We should give an example that we will work with and talk about why.
3. Anything else?

3.1 The Record Linkage Uncertainty

In most record linkage tasks, one is interested in perform record linkage as a pre-processing tool, such that other analyses can be performed afterward that may include performing linear regression, capture recapture, or other types of statistical analyses. It's of great importance to assess the record linkage uncertainty after the record linkage task in finished and propagate this error into these subsequent analyses. With such motivations, suppose that post-linkage, we are interested in estimating a parameter about a population $\boldsymbol{\eta}$.

It is most natural to quantify the uncertainty of the record linkage process, which arises from the posterior distribution of $p(\boldsymbol{\Lambda} \mid \boldsymbol{X})$. We will assume without loss of generality that this posterior, such as the one, we have presented is a proper distribution. Then assuming we knew the true value of the linkage structure $\boldsymbol{\Lambda}$, we could represent our beliefs about the population size $\boldsymbol{\eta}$ by the posterior denoted by $p_C(\boldsymbol{\eta} \mid f(\boldsymbol{\Lambda}))$, which is just the posterior obtained from a capture-recapture model C with some likelihood and prior distribution. Given this setup, it is most natural to account for the record linkage uncertainty in the population size by the following quantity:

$$U(\boldsymbol{\eta}) =: E_{\boldsymbol{\Lambda}|\boldsymbol{X}}[p_C(\boldsymbol{\eta} \mid f(\boldsymbol{\Lambda})] = \sum_{\boldsymbol{\Lambda}} p_C(\boldsymbol{\eta} \mid f(\boldsymbol{\Lambda})p(\boldsymbol{\Lambda} \mid \boldsymbol{X}), \tag{3}$$

where f is some function of $\boldsymbol{\Lambda}$. Namely, $U(\boldsymbol{\eta})$ is the marginal posterior distribution of $\boldsymbol{\eta}$ assuming a linkage model and a joint prior $p(\boldsymbol{\eta} \mid f(\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$.

Remark: We can think of $U(\boldsymbol{\eta})$ as the expected posterior distribution of the population parameter of interest, where we have averaged over the posterior of the linkage structure. In addition, it is quite easy to see that $U(\boldsymbol{\eta})$ is a proper distribution. Finally, the total variability represented by $U(\boldsymbol{\eta})$ can also be easily decomposed.

Recall $U(\boldsymbol{\eta}) = p(\boldsymbol{\eta} \mid \boldsymbol{X})$. Then

$$\text{Var}(\boldsymbol{\eta} \mid X) = \text{Var}_{\boldsymbol{\Lambda}|\boldsymbol{X}}[E[N \mid \boldsymbol{\Lambda}]] + E_{\boldsymbol{\Lambda}|\boldsymbol{X}}[\text{Var}[\boldsymbol{\eta} \mid \boldsymbol{\Lambda}]], \tag{4}$$

where $\text{Var}_{\boldsymbol{\Lambda}|\boldsymbol{X}}[E[\boldsymbol{\eta} \mid \boldsymbol{\Lambda}]]$ corresponds to record linkage uncertainty due to population size estimation and $E_{\boldsymbol{\Lambda}|\boldsymbol{X}}[\text{Var}[\boldsymbol{\eta} \mid \boldsymbol{\Lambda}]]$ correspond to the variability associated from a Bayesian methods due to estimating N. (In practice, both $U(N)$ and both quantities above must be estimated by Markov chain Monte carlo).

3.1.1 Example to Population Sized Estimation

As an example, suppose we wish to perform population sized estimation (capture recapture) after performing record linkage. Let N denote the unknown size of the population and let $\boldsymbol{h} = (h_1, \ldots, h_K) \in \{0,1\}^K$ denote the inclusion pattern, where $h_k = 1$ indicates inclusion in database k. Let $n_{\boldsymbol{h}}$ denote the number of entities with inclusion pattern $\boldsymbol{h}$, where the inclusion pattern's frequencies are in a contingency table $n^* = \{n_{\boldsymbol{h}}\}$ where $\boldsymbol{h} \in \{0,1\}^K$. Of course, the number of entities missed by all databases is unobserved and missing, $n_{00\cdots}$, and thus, the observed counts are denoted by $\{n_{\boldsymbol{h}}\}$ where $\boldsymbol{h} \in \{0,1\} \setminus \{0\}^K$.

Given a linkage structure, we first describe how to compute the contingency table (incomplete) $\boldsymbol{n}$. Let n denote the total number of labelings that occur in the linkage structure $\boldsymbol{\Lambda}$. Without loss of generality, we can write the labelings of $\boldsymbol{\Lambda}$ as being $1, \ldots, n$. Then for each label $\lambda = 1, \ldots, n$ let

$$H_{\lambda,k} = \begin{cases} 1, \text{ if there exists a record } (i,j) \in \boldsymbol{X} \text{ such that } \lambda_{i,j} = \lambda \\ 0, \text{otherwise.} \end{cases}$$

That is, $H_{\lambda,k}$ contains inclusion indicators of where each of the n individuals are included in database k, and then we obtain the $\boldsymbol{n}(\boldsymbol{\Lambda})$ contingency table using these inclusion indicators. Note that we write $\boldsymbol{n}(\boldsymbol{\Lambda})$ since each contingency table is a function of the linkage structure. Suppose that the prior on the linkage structure is $p(\boldsymbol{\Lambda} \mid \boldsymbol{X})$. Now conditioning on the linkage structure, assume one can calculate the posterior of $p_C(N \mid \boldsymbol{n}(\boldsymbol{\Lambda}))$.

In order to calculate the expected distribution of the population size, while averaging with respect to the posterior of the linkage structure, we simple use equation 3. Under the estimation of population sized estimation,

$$U(N \mid \boldsymbol{X}) = E_{\boldsymbol{\Lambda}\mid\boldsymbol{X}}[p_C(N \mid \boldsymbol{n}(\boldsymbol{\Lambda})] = \sum_{\boldsymbol{\Lambda}} p_C(N \mid \boldsymbol{n}(\boldsymbol{\Lambda})p(\boldsymbol{\Lambda} \mid \boldsymbol{X}).$$

I need to clarify this for AK and the reader in general. This is a marginal distribution above and this is not clear as written. I will add in some further details about this.

3.2 Empirically Motivated Bayesian Graphical Model

For the rest of the paper, we will work with the empirically motivated Bayesian graphical model [**?**] , due it's ability to work with both categorical and textual features and it's shown use in the literature to be superior to semi-supervised methods. In addition, the method is available on `CRAN` via the `blink` packing, making it easy to work with. We briefly review the method, and then work with this method for the remainder of the paper.

The work of [**?**] assumes fields $1, \ldots, p_s$ are string-valued, while fields $p_s + 1, \ldots, p_s + p_c$ are categorical, where $p_s + p_c = p$ is the total number of fields. They assume an empirical Bayesian distribution on the latent parameter. For each $\ell \in \{1, \ldots, p_s + p_c\}$, let S_ℓ denote the set of *all* values for the ℓth field that occur anywhere in the data, i.e., $S_\ell = \{X_{ij\ell} : 1 \leq i \leq k, 1 \leq j \leq n_i\}$, and let $\alpha_\ell(w)$ equal the empirical frequency of value w in field ℓ. Let G_ℓ denote the empirical distribution of the data in the ℓth field from all records in all databases combined. So, if a random variable W has distribution G_ℓ, then for every $w \in S_\ell$, $P(W = w) = \alpha_\ell(w)$. Hence, let G_ℓ be the prior for each latent entity $Y_{j'\ell}$. The distortion process changes such that

$$\begin{aligned} &P(X_{ij\ell} = w \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell}) \\ &\quad = \frac{\alpha_\ell(w) \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]}{\sum_{w \in S_\ell} \alpha_\ell(w) \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]}, \end{aligned}$$

where $c > 0$ is a fixed normalizing constant corresponding to an arbitary distance metric $d(\cdot, \cdot)$. Denote this distribution by $F_\ell(Y_{\Lambda_{ij}\ell})$. The model becomes

$$\begin{aligned} X_{ij\ell} \mid \Lambda_{ij},\, Y_{\Lambda_{ij}\ell},\, z_{ij\ell} &\overset{\text{ind}}{\sim} \begin{cases} \delta(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1, \ell \leq p_s \\ G_\ell & \text{if } z_{ij\ell} = 1, \ell > p_s \end{cases} \\ Y_{j'\ell} &\overset{\text{ind}}{\sim} G_\ell \\ z_{ij\ell} \mid \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_{i\ell}) \\ \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Beta}(a, b) \\ \Lambda_{ij} &\overset{\text{ind}}{\sim} \text{Uniform}(1, \ldots, N), \end{aligned} \tag{5}$$

where all distributions are also independent of each other; assume that a, b, N are assumed known. Finally, observe that record linkage and de-duplication are both simply a question of whether $\Lambda_{i_1,j_1} = \Lambda_{i_2,j_2}$, where $i_1 \neq i_2$ for record linkage and $i_1 = i_2$ for de-duplication.

Figure 2 contains a graphical representation of model 6.

Remark: This framework was shown to work well in applications and simulation studies, however, it was quite sensitive to the choice of the hyperparameters. This method beat supervised methods, such as random forests when the amount of training data input into the supervised methods was $< 10\%$.

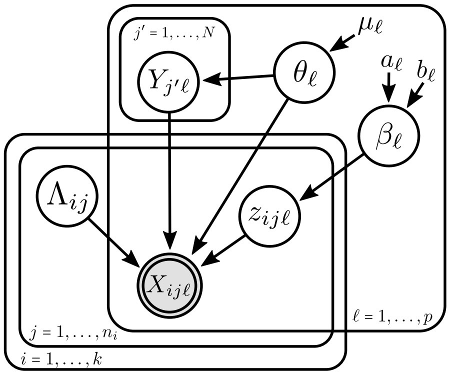

Figure 1: Graphical representation of model 6.

4 Post-linkage Analysis

There are many post-linkage analysis situations that arise in real data applications, including regression, capture recapture, networks analysis, among others. We present a general Bayesian approach to these, where the error can be propagated and then illustrate an example using capture recapture and networks analysis. [[We may just want to do one for ICML and then do a second one for the longer journal version]].

ATTN: Andee: can you try writing out the post-linkage approach in a general way for tomorrow and then fill in the sections for CRC and a networks approach.

5 Experiments

Perhaps we should look at some simulated data first to fully understand

For each data set, we consider four statistics: (a) the number of singleton clusters, (b) the maximum cluster size, (c) the mean cluster size, and (d) the 90^{th} percentile of cluster sizes. We compare each statistic's true value to its posterior distribution according to each of the models. For each model and data set combination, we also consider five entity-resolution summary statistics: (a) the posterior expected number of clusters, (b) the posterior standard error, (c) the false negative rate, (d) the false discovery rate, and (e) the computational run time.

[[What else might we want to look at]]. What is standard for CRC. We should report nice summary metric here too and state why we are doing these.

5.1 Data Sets

We consider XXX data sets for each of the proposed methods and the evaluation metrics.

Restaurant: This data set comes from XXX. There are 864 total records and 4 fields, including XXXX. Using the unique identifiers we find that there are 112 pairs of records that point the the same entity and there are 752 unique entities. Ground truth is available based upon XXX; roughly XX of the clusters are singletons.

CD: This data set comes from XXX. There are 9,763 total records and 106 fields, including XXXX. Using the unique identifiers we find that there are 299 pairs of records that point the the same entity and there are 9,508 unique entities. Ground truth is available based upon XXX; roughly XX of the clusters are singletons.

Cora Add the Cora data set.

5.2 Results

We report the results of our experiments in table **??** and figure **??**.

6 Summary

old material for paper.

7 Bayesian Record Linkage

We assume a graphical model for record linkage that can be viewed as a hit and miss model [[CITE]], and incorporates both both categorical and noisy string data, such as names, addresses, etc. [?]. We extend this approach to a general population sized estimation framework and [[develop some other results]].

7.1 Notation

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ represent the data, with k databases, indexed by i. The ith list has n_i observed records, indexed by j. Each record corresponds to one of N latent entities, indexed by j'. Assume $N = \sum_{i=1}^{k} n_i$ without loss of generality. Each record or latent entity has values on p fields, indexed by ℓ, and are assumed be categorical and the same across all records and entities [?, ?]. M_ℓ denotes the number of possible categorical values for the ℓth field. In both models, $X_{ij\ell}$ denotes the observed value of the ℓth field for the jth record in the ith list, and $Y_{j'\ell}$ denotes the true value of the ℓth field for the j'th latent entity. Then Λ_{ij} denotes the latent entity to which the jth record in the ith list corresponds, i.e., $X_{ij\ell}$ and $Y_{j'\ell}$ represent the same entity if and only if $\Lambda_{ij} = j'$. Then $\boldsymbol{\Lambda}$ denotes the Λ_{ij} collectively. Distortion is denoted by $z_{ij\ell} = I(X_{ij\ell} \neq Y_{\Lambda_{ij}\ell})$, where $I(\cdot)$ denotes the indicator function. As usual, I represents the indicator function (e.g., $I(x_{ij\ell} = m)$ is 1 when the ℓth field in record j in file i has the value m), and let δ_a denote the distribution of a point mass at a (e.g., $\delta_{y_{\Lambda_{ij}\ell}}$).

7.2 Empirical Bayesian Record Linkage

The work of [?] assumes fields $1, \ldots, p_s$ are string-valued, while fields $p_s + 1, \ldots, p_s + p_c$ are categorical, where $p_s + p_c = p$ is the total number of fields. They assume an empirical Bayesian distribution on the latent parameter. For each $\ell \in \{1, \ldots, p_s + p_c\}$, let S_ℓ denote the set of *all* values for the ℓth field that occur anywhere in the data, i.e., $S_\ell = \{X_{ij\ell} : 1 \leq i \leq k, 1 \leq j \leq n_i\}$, and let $\alpha_\ell(w)$ equal the empirical frequency of value w in field ℓ. Let G_ℓ denote the empirical distribution of the data in the ℓth field from all records in all databases combined. So, if a random variable W has distribution G_ℓ, then for every $w \in S_\ell$, $P(W = w) = \alpha_\ell(w)$. Hence, let G_ℓ be the prior for each latent entity $Y_{j'\ell}$. The distortion process changes such that

$$\begin{aligned} &P(X_{ij\ell} = w \mid \Lambda_{ij}, Y_{\Lambda_{ij}\ell}, z_{ij\ell}) \\ &\quad = \frac{\alpha_\ell(w)\, \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]}{\sum_{w \in S_\ell} \alpha_\ell(w)\, \exp[-c\, d(w, Y_{\Lambda_{ij}\ell})]}, \end{aligned}$$

where $c > 0$ is a fixed normalizing constant corresponding to an arbitary distance metric $d(\cdot, \cdot)$. Denote this distribution by $F_\ell(Y_{\Lambda_{ij}\ell})$. The model becomes

$$\begin{aligned} X_{ij\ell} \mid \Lambda_{ij},\, Y_{\Lambda_{ij}\ell},\, z_{ij\ell} &\overset{\text{ind}}{\sim} \begin{cases} \delta(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\Lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1, \ell \leq p_s \\ G_\ell & \text{if } z_{ij\ell} = 1, \ell > p_s \end{cases} \\ Y_{j'\ell} &\overset{\text{ind}}{\sim} G_\ell \\ z_{ij\ell} \mid \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Bernoulli}(\beta_{i\ell}) \\ \beta_{i\ell} &\overset{\text{ind}}{\sim} \text{Beta}(a, b) \\ \Lambda_{ij} &\overset{\text{ind}}{\sim} \text{Uniform}(1, \ldots, N), \end{aligned} \tag{6}$$

where all distributions are also independent of each other; assume that a, b, N are assumed known. Finally, observe that record linkage and de-duplication are both simply a question of whether $\Lambda_{i_1,j_1} = \Lambda_{i_2,j_2}$, where $i_1 \neq i_2$ for record linkage and $i_1 = i_2$ for de-duplication.

Remark: This framework was shown to work well in applications and simulation studies, however, it was quite sensitive to the choice of the hyperparameters. This method beat supervised methods, such as random forests when the amount of training data input into the supervised methods was $< 10\%$.

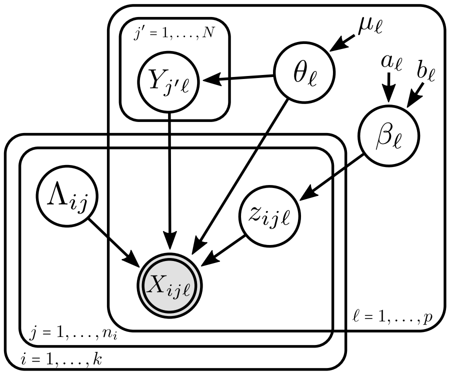

Figure 2: Graphical representation of model 6.

Figure 2 contains a graphical representation of model 6.

Theorem: Show that this model is a special case of the microclustering models from my NIPS paper. This is not too complicated. I just need to write out the proof.

7.3 Record Linkage Uncertainty

It is most natural to quantify the uncertainty of the record linkage process, which arises from the posterior distribution of $p(\mathbf{\Lambda} \mid \boldsymbol{X})$.

We will assume without loss of generality that this posterior, such as the one, we have presented is a proper distribution. [[We could extend this to Gibbs models actually and these will still be proper]]. Then assuming we knew the true value of the linkage structure $\mathbf{\Lambda}$, we could represent our beliefs about the population size N by the posterior denoted by $p_C(N \mid n(\mathbf{\Lambda}))$, which is just the posterior obtained from a capture-recapture model C with some likelihood and prior distribution. Given this setup, it is most natural to account for the record linkage uncertainty in the population size by the following quantity:

$$U(N) =: E_{\mathbf{\Lambda}|\boldsymbol{X}}[p_C(N \mid \boldsymbol{n}(\mathbf{\Lambda})] = \sum_{\mathbf{\Lambda}} p_C(N \mid \boldsymbol{n}(\mathbf{\Lambda})p(\mathbf{\Lambda} \mid \boldsymbol{X}). \tag{7}$$

Remark: We can think of $U(N)$ as the expected posterior distribution of the population size, where we have averaged over the posterior of the linkage structure. In addition, it is quite easy to see that $U(N)$ is a proper distribution. Finally, the total variability represented by $U(N)$ can also be easily decomposed.

Recall $U(N) = p(N \mid \boldsymbol{X})$. Then

$$\mathrm{Var}(N \mid X) = \mathrm{Var}_{\mathbf{\Lambda}|\boldsymbol{X}}[E[N \mid \mathbf{\Lambda}]] + E_{\mathbf{\Lambda}|\boldsymbol{X}}[\mathrm{Var}[N \mid \mathbf{\Lambda}]], \tag{8}$$

where $\mathrm{Var}_{\mathbf{\Lambda}|\boldsymbol{X}}[E[N \mid \mathbf{\Lambda}]]$ corresponds to record linkage uncertainty due to population size estimation and $E_{\mathbf{\Lambda}|\boldsymbol{X}}[\mathrm{Var}[N \mid \mathbf{\Lambda}]]$ correspond to the variability associated from a Bayesian methods due to estimating N. (In practice, both $U(N)$ and both quantities above must be estimated by Markov chain Monte carlo).

8 Population Size Estimation

We first introduce a general framework for inclusion pattens for population sized estimation. Next, we derive a way of measuring the record linkage uncertainty of Bayesian graphical models. In addition, we look at a general way the variability for record linkage and capture recapture can be assessed. Finally, we work out this formulation under the framework of Madigan and York (1992) and [[let's try and go for that of Johndrow and Valllier]].

8.1 Capture Recapture Inclusion Patterns

Let $\boldsymbol{h} = (h_1, \ldots, h_K) \in \{0, 1\}^K$ denote the inclusion pattern, where $h_k = 1$ indicates inclusion in database k. Let $n_{\boldsymbol{h}}$ denote the number of entities with inclusion pattern $\boldsymbol{h}$, where the inclusion pattern's frequencies are in a contingency table $n^* = \{n_{\boldsymbol{h}}\}$ where $\boldsymbol{h} \in \{0,1\}^K$. Of course, the number of entities missed by all databases is unobserved and missing, $n_{00\cdots}$, and thus, the observed counts are denoted by $\{n_{\boldsymbol{h}}\}$ where $\boldsymbol{h} \in \{0,1\} \setminus \{0\}^K$.

For a particular entity, their inclusion pattern $\boldsymbol{h}$ can be written such that $Pr(\boldsymbol{h} \mid \boldsymbol{\eta}) = \boldsymbol{\eta}_{\boldsymbol{h}}$, where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_K)$ gives the probability of each inclusion pattern. Each model will dictate different inclusion patterns; let $\boldsymbol{\eta}_m$ denote the inclusion probabilities under model m. Assuming $N = \sum_{\boldsymbol{h}} n_h$ entities in the population and assuming that each entity is indendently and identically distributed, we find that

$$P(\boldsymbol{n}^* \mid N, \boldsymbol{\eta}_m, m) = N! \prod_{\boldsymbol{h}\in\{0,1\}} \frac{(\boldsymbol{\eta}_{\boldsymbol{h},m})^{n_{\boldsymbol{h}}}}{n_{\boldsymbol{h}}!} \tag{9}$$

Since in practice N and $\boldsymbol{n}$ are fixed, we can write $n_{00\cdots0} = N - \sum_{\boldsymbol{h}\in\{0,1\}^K\setminus\{0\}^K} n_h$. It then follows that $P(\boldsymbol{n} \mid N, \boldsymbol{\eta}_m, m) = P(\boldsymbol{n}^* \mid N, \boldsymbol{\eta}_m, m)$. Now given any model m and a prior on the population size p(N), our main goal is to obtain

$$P(N \mid \boldsymbol{n}, m) = \frac{P(\boldsymbol{n} \mid N, m)p(N)}{\sum_N P(\boldsymbol{n} \mid N, m)p(N),} \tag{10}$$

where $P(\boldsymbol{n} \mid N, m) = \int_{\boldsymbol{\eta}_m} P(\boldsymbol{n} \mid N, \boldsymbol{\eta}_m, m)p(\boldsymbol{\eta}_m \mid m)\, d\boldsymbol{\eta}_m$.

The above formulation only requires that N and $\boldsymbol{\eta}_m$ are independent. We could consider model that should lead to a closed form solution such as those in Madigan and York https://pdfs.semanticscholar.org/7aed/b2eb2345c1d6ef844be7a58bcca3f7f9cb78.pdf, or perhaps https://arxiv.org/pdf/1606.02235.pdf.

The Madigan and York paper is super cool because it allows us to do Bayesian model averaging, so we could write out the Bayesian partition models and average over all of the priors that we might believe in and then incorporate the record linkage error.

Recall that our modeling framework can be viewed as a bi-partite graph, which has strong connections to probabilistic graphical models [[CITE]], where in this literature one encodes the set of conditional independencies of a multivariate distribution into a graph using edges and nodes. In a graphical model (and in our framework), a node is represented by a random variable, and two nodes are joined if they are conditionally dependent given all the other nodes or random variables. Thus, a graphical model captures conditional dependencies between two nodes of inclusion patterns intrinsically. Madigan and York (1992, 1997) restrict to working with decomposable graphical models, meaning that the independence graph is triangulated (chordal). One reason we prefer to work with this approach is that the conditionals are available in closed form and we show the details of this below.

Given a model m and a set a parameters $\boldsymbol{\eta}_m$, we can use the hyper-Dirichlet distribution (Dawid and Lauritzen, 1993 and Madigan and York (1992) for the parameters $\boldsymbol{\eta}_m$, which lead to a closed form solution for $P(\boldsymbol{n} \mid N, m)$. (We refer to Dawid and Lauritzen, 1993 regarding a review of the hyper-Dirchlet distribution). Thus, given a hyper-Dirichlet prior on $\boldsymbol{\eta}_m$ and assuming that N and $\boldsymbol{\eta}_m$ are independent, Madigan and York (1992) show that

$$P(\boldsymbol{n} \mid N, m) = \int_{\boldsymbol{\eta}_m} P(\boldsymbol{n} \mid N, \boldsymbol{\eta}_m, m)p(\boldsymbol{\eta}_m \mid m)\, d\boldsymbol{\eta}_m \tag{11}$$

$$= \frac{N!}{\prod_{\boldsymbol{h}\in\{0,1\}^K} n_{\boldsymbol{h}}!} \frac{\Phi_m(\boldsymbol{\alpha} + \boldsymbol{n}^*)}{\Phi_m(\boldsymbol{\alpha})} \tag{12}$$

.

Note that

$$\Phi_m(\boldsymbol{\alpha}) = \frac{\prod_{\ell=1}^{L} \prod_{\boldsymbol{h}_{C\ell}} \Gamma(\alpha_{\boldsymbol{h}_{C_\ell}})}{\Gamma(\sum_{\boldsymbol{h}\in\{0,1\}^K} \alpha_{\boldsymbol{h}})^Q \prod_{\ell=2}^{L} \prod_{\boldsymbol{h}_{S\ell}} \Gamma(\alpha_{\boldsymbol{h}_{S_\ell}})} \tag{13}$$

where $\{C_\ell\}$ denotes the set of maximal cliques, $\{S_\ell\}$ the set of separators, and Q the set of connected components. In addition, for a subset of nodes O, $\boldsymbol{h}_O$ represents an inclusion pattern constrained to the subset of variables in O. Also,

$$\alpha_{\boldsymbol{h}_O} = \sum_{\boldsymbol{h}':\boldsymbol{h}'_O = \boldsymbol{h}_O} \alpha_{\boldsymbol{h}'}.$$

Within this framework, it is natural to take into account the uncertainty regarding the population sized estimation, which can be written as

$$P(N \mid n) = \frac{P(N)\sum_m P(\boldsymbol{n} \mid N, m)p(m)}{\sum_N P(N)\sum_m P(\boldsymbol{n} \mid N, m)p(m)}$$

for prior distributions $p(m)$.

(Perhaps we want to look at uniform priors to start and adjust from there).

9 Population Sized Estimation Post Record Linkage

We should be able to write this down as mixture models using BNP. Given a linkage structure, we first describe how to compute the contingency table (incomplete) $\boldsymbol{n}$. Let n denote the total number of labelings that occur in the linkage structure $\boldsymbol{\Lambda}$. Without loss of generality, we can write the labelings of $\boldsymbol{\Lambda}$ as being $1, \ldots, n$. Then for each label $\lambda = 1, \ldots, n$ let

$$H_{\lambda,k} = \begin{cases} 1, & \text{if there exists a record } (i,j) \in \boldsymbol{X} \text{ such that } \lambda_{i,j} = \lambda \\ 0, & \text{otherwise.} \end{cases}$$

That is, $H_{\lambda,k}$ contains inclusion indicators of where each of the n individuals are included in database k, and then we obtain the $\boldsymbol{n}(\boldsymbol{\Lambda})$ contingency table using these inclusion indicators. Note that we write $\boldsymbol{n}(\boldsymbol{\Lambda})$ since each contingency table is a function of the linkage structure.

9.1 Details under particular models, like Madigan and York, Johndrow and Vallier

Sketch out the details next for Madigan and York. Then do so for the Johndrow paper.

10 Things to look at next

Now, all we need to do is write down how to perform this in practice and how to evaluate the uncertainty. We also need to try this out in practice. It might be interesting to look at graphs that aren't decomposable or see if the framework of Johndrow and Manrique-Vallier fits in at all.

It seems that we could work with a very large class of capture-recapture models, starting with a stick-breaking process, a MFM, a DPM, and potentially others. The record linkage model can be viewed at some level as a stick breaking process.

1. The Madigan and York model will be limiting but in closed form.

2. I think we can also write all the Kolchin partition models down in closed form as well. We can at least start with the approach of Johndrow and Vallier, but special cases that follow are a DP, PYP, NBD, NBNB. This would be quite nice. Could the Madigan and York be a special case of a KP model? This would be quite interesting.

3. It would be nice to tie some of this together because it would make a very general framework.

11 Issues with the approach

[[There is an issue with Madigan and York as they require that models be decomposable or rather that the graph be a triangle. We want graphs to be able to have any type of form, so this is a weakness of their proposal]].

This deals with graphs that should be able to be non-decomposable http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.1333&rep=rep1&type=pdf

12 Small Area Estmation

A small area estimation model can be written as

$$y = \theta + X\beta + u + e$$

where the error terms are independent and Gaussian. Again, this is a mixture model. If you think about the approach to SAE and CRC, we can view them both as mixture models. We should try and write the formulation very general.

13 Networks

Notes regarding the network problem.

Think for example you want to reconstruct a relational network among some actors (may be countries, or famous people) using news media reports. I.e. anytime some news media comes out with a news report involving say actor A and actor B you add an edge among them. Now, as we know news media reports are very noisy and computer systems processing them have additional noise. This creates lots of duplication problems in two directions:

1] Duplication in the nodes: For example two media reports may refer to the same actor, but the way the report is written may lead you to think they are two different nodes. Of course in this case record linkage will be important to solve these duplication problems in the two nodes. I.e. your network is indeed much smaller (in terms of real actors) than it seems, because sometimes you are adding nodes that indeed refer to the same actor. 2] Duplication in the edges: Suppose news media 1 said that A and B had a conflict, whereas in the same day news media 2 said that A and B had a fight. This refers of course to the same relation (the only difference is that news media 1 used term ?conflict? and news media 1 used term ?fight?). So you should add only 1 edge instead of 2 among actors A and B. These duplications in the edges are quite common and an issue, since you end up over-counting the relations among actors.

I think you can easily find these issues in this dataset: https://www.gdeltproject.org/data.html. But I can think about many other examples in citation networks, and also brain networks (especially on the edge duplication and also nodes one when you go to voxels).

I tried to address those in https://www.gdeltproject.org/data.html with some ad-hoc choice when I was using these data some years ago. It worked well, but I am sure that your methods here would do a great job. This is quite important in networks, since such pre-processing can provide more high-quality relational data.

One quick idea from my email exchange with Dany:

Build up the network using the raw data (without record-linkage on nodes and edges)?call it network X1? and the build up another network using data that have been de-duplicated (both nodes and edges) ? call it network X2, and then check how different the two networks are.

I expect that:

1] They will be quite different 2] The ?cleaned? network X2, will have a structure that better fits social processes (i.e. has better communities, has small world structure etc?).

Applying your methods to these data shouldn?t be hard. Indeed, each row in your dataset is a different news media report for which you have many informations (actors, time of publication, type of news, text data,). So your record linkage problem will have news media reports as statistical units (instead of people). If you can solve the duplication problem for the news reports, then you are jointly solving both the node and edge duplication issue. In fact, you will end up with a set of unique news (in terms of actors and type of relation among them).

Notes for paper. Cannot get our PSD paper to compile.

AK: We'll want to talk about notation here.

It can be shown that a special case of [[CITE]] is equivalent to the linkage structure prior [**?**, **?**], which we now explicitly formalize. We review the model of [[CITE]] in the special case and then show the equivalence of the models.

Let K be the number of potential clusters. Then following [[CITE]], we have the following model:

$$\begin{aligned} K &\sim \text{NegBin}\,(a, q) \\ N_1, \ldots, N_K \,|\, K &\overset{\texttt{iid}}{\sim} \text{NegBin}\,(r, p), \end{aligned} \tag{14}$$

for $a, r > 0$ and $q, p \in (0, 1)$. Note that K and some of $N_1, \ldots, N_K$ may be zero. We then define $N = \sum_{k=1}^{K} N_k$ and, given $N_1, \ldots, N_K$, generate a set of cluster assignments $z_1, \ldots, z_N$ by drawing a vector uniformly at random from the set of permutations of

$$(\underbrace{1, \ldots, 1}_{N_1 \text{ times}}, \underbrace{2, \ldots, 2}_{N_2 \text{ times}}, \ldots\ldots, \underbrace{K, \ldots, K}_{N_K \text{ times}}).$$

The cluster assignments $z_1, \ldots, z_N$ induce a random partition C_N of $[N]$, where N is itself a random variable—i.e., C_N is a random partition of a random number of elements. The resulting marginal distribution of C_N the NegBin–NegBin (NBNB) model. If $\mathscr{C}_N$ denotes the set of all possible partitions of $[N]$, then $\bigcup_{N=1}^{\infty} \mathscr{C}_N$ is the set of all possible partitions of $[N]$ for $N \in \mathbb{N}$. If one replaces the negative binomials in equation **??** with Poissons, then we obtain a limiting special case of the NBNB model. We will refer to this model the permuted Poisson sizes (PERPS) model.

We now show the equivalence of the model of PERPS to that in equation **??**. Let $Z = (z_1, \ldots, z_n)$. According to the model defined by PERPS(α, λ),

$$P(Z \,|\, K) = P(Z, N \,|\, K) = P(Z \,|\, N, K)\, P(N \,|\, K) \tag{15}$$

since the value of N is completely determined by the value of Z. From the Poisson superposition principle, we know that if $N = \sum_{k=1}^{K} N_k$ and $N_1, \ldots, N_K \,|\, K \overset{\texttt{iid}}{\sim} \text{Poisson}(\lambda)$, then $N \,|\, K \sim \text{Poisson}(K\lambda)$ and

$$P(N \,|\, K) = \frac{(K\lambda)^N}{N!} \exp\,(-K\lambda). \tag{16}$$

We also know that $P(Z \,|\, N, K) = \left(\frac{1}{K}\right)^N$. Therefore,

$$P(Z \,|\, K) = \frac{\lambda^N}{N!} \exp\,(-K\lambda). \tag{17}$$

We show that equation 16 holds by writing $P(Z \,|\, K) = P(Z, N, S_1, \ldots, S_K \,|\, K)$ since N and $S_1, \ldots, S_K$ are completely determined by Z. It then follows that

$$\begin{aligned} &P(Z \,|\, K) && (18)\\ &= P(Z, N, S_1, \ldots, S_K \,|\, K) && (19)\\ &= P(Z \,|\, N, S_1, \ldots, S_K, K)\, P(N, S_1, \ldots, S_K \,|\, K) && (20)\\ &= P(Z \,|\, S_1, \ldots, S_K, K)\, P(N \,|\, S_1, \ldots, S_K, K) && (21)\\ &\qquad \times P(S_1, \ldots, S_K \,|\, K) && (22)\\ &= \frac{\prod_{k=1}^{K} S_k!}{N!} \cdot 1 \cdot \frac{\lambda^N}{\prod_k = 1 S_k!} \exp\,(-K\lambda) && (23)\\ &= \frac{\lambda^N}{N!} \exp\,(-K\lambda). && (24) \end{aligned}$$

Thus, conditioning on N, or rather $P(Z \,|\, N, K) = \left(\frac{1}{K}\right)^N$ leads exactly [**?**, **?**] to the uniform prior on the linkage structure, which clusters records to unknown latent entities in the framework of [**?**, **?**], where N, K are both fixed.

Remark: Due to this, $P(Z \mid K)$ is the same for [**?**, **?**] and PERPS and it follows that $P(K)$ is the same for the PERPS and Poisson-Uniform. This means that $P(Z)$ is the same since $P(Z \mid K)$ and $P(K)$ are the same for both models. This implies that $P(C_N \mid N)$ will be the same under our proposed framework and that of [**?**, **?**].