Scalable Bayesian Record Linkage

Brendan S. McVeigh
Jared S. Murray

Carnegie Mellon University, Department of Statistics & Data Science
University of Texas at Austin, Department of Information, Risk, and Operations Management

Introduction

Probabilistic record linkage (PRL) is the task of merging two or more databases that have entities in common but no unique identifier. Matching must be done based on incomplete information, since features for records may be incorrectly or inconsistently recorded.

Label	Last Name	Sex	Education
a_1	Smith	M	High School
a_2	Meyer	M	**College**
a_3	**Jonnes**	F	High School
a_4	Zimmerman	F	Graduate

Label	Last Name	Sex	Education
b_1	Meyer	M	**Graduate**
b_2	Zimmerman	F	Graduate
b_3	**Jones**	F	High School
b_4	Smith	M	High School

Model for Record Linkage

Sets of records, A and B with n_A and n_B entries. We use a matrix C to represent which records are linked (correspond to the same entity).

$$C_{ab} = \begin{cases} 1 \; a \sim b \\ 0 \; a \nsim b \end{cases}$$

Comparisons between record pairs used to estimate links

$$\gamma_{ab} = (\gamma_{ab}^1, \gamma_{ab}^2, \ldots, \gamma_{ab}^d)$$

Mixture model with two components, corresponding to comparisons between linked records and non-linked records [1]

$$m(g) = \Pr\left(\gamma_{ab} = g \mid C_{ab} = 1\right); \quad u(g) = \Pr\left(\gamma_{ab} = g \mid C_{ab} = 0\right)$$

Component membership determines likelihood of comparison vectors

$$m_{jh} = \Pr\left(\gamma_{ab}^j = h | C_{ab} = 1\right); \quad u_{jh} = \Pr\left(\gamma_{ab}^j = h | C_{ab} = 0\right),$$

for $1 \leq j \leq d$ and $1 \leq h \leq k_j$, where comparison j has k_j possible levels.

Conditional independence assumed between comparisons

$$m(\gamma_{ab}) = \prod_{j=1}^{d} \prod_{h=1}^{k_i} m_{jh}^{\mathbb{1}(\gamma_{ab}^j = h)}; \quad u(\gamma_{ab}) = \prod_{j=1}^{d} \prod_{h=1}^{k_i} u_{jh}^{\mathbb{1}(\gamma_{ab}^j = h)}$$

MAP Parameter Estimation

Relative likelihood of component membership used to weight links

$$w_{ab} = \log\left(m(\gamma_{ab}) / u(\gamma_{ab})\right)$$

Penalized log likelihood discourages over-linking

$$\begin{aligned} \ell(C, m, u | \Gamma) &= \sum_{a,b \in A \times B} \left[C_{ab} \log(m(\gamma_{ab})) + (1 - C_{ab}) \log(u(\gamma_{ab}))\right] - \theta \sum_{ab} C_{ab} \\ &= \sum_{a,b \in A \times B} \log(u(\gamma_{ab})) + C_{ab}(w_{ab} - \theta) \end{aligned}$$

Equivalent prior over C is $p(C) \propto \exp(-\theta L)$ where $L = \sum_a \sum_b C_{ab}$ [2].

Compute a local mode by alternating maximization of parameters

(1) Maximize C holding m, u constant
(2) Maximize m, u holding C constant

Maximize C using linear sum assignment problem

$$\max_C \sum_{a,b \in A \times B} C_{ab}(w_{ab} - \theta)$$

$$\text{subject to } \; C_{ab} \in \{0, 1\}; \; \sum_{b \in B} C_{ab} \leq 1 \; \forall a \in A; \; \sum_{a \in A} C_{ab} \leq 1 \; \forall b \in B$$

Maximization of m and u can be done in closed form

$$m_{jh} = \frac{n_{mjh} + \sum_{ab} C_{ab} \mathbb{1}(\gamma_{ab}^j = h)}{\sum_h n_{mjh} + \sum_{ab} C_{ab}}; \qquad u_{jh} = \frac{n_{ujh} + \sum_{ab} (1 - C_{ab}) \mathbb{1}(\gamma_{ab}^j = h)}{\sum_h n_{ujh} + \sum_{ab} (1 - C_{ab})}$$

Post-hoc Blocking

Blocking techniques, restrict the set of record pairs considered (e.g. only considering record pairs from the same geographic area).

Post-hoc blocking: New data-driven approach to unsupervised blocking

(1) Generate $\hat{m}$, $\hat{u}$ and $\hat{C}$ via MAP parameter estimation
(2) Compute G, the adjacency matrix with edges $G_{ab} = \mathbb{1}(\hat{w}_{ab} > w_0)$
(3) Use the connected components of G as *post-hoc blocks*

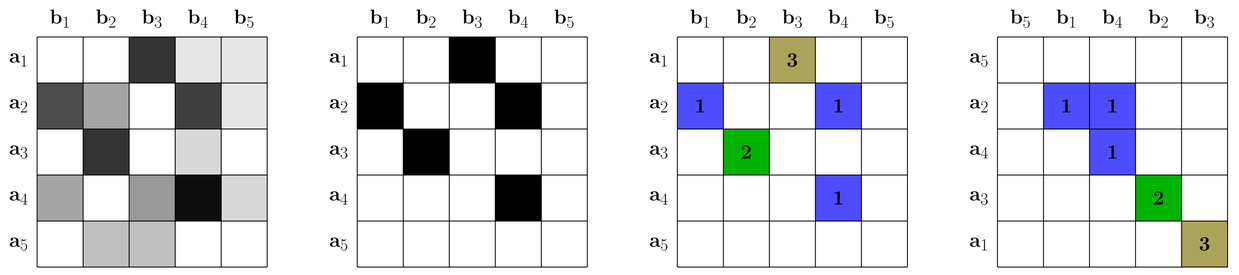

Figure 1: An example of post-hoc blocking.

Restricted MCMC

Approximate posterior distribution estimated using a MCMC algorithm

- Fix all record pairs outside of post-hoc blocks to non-link
- Add/drop/swap link updates performed on each post-hoc block
- m and u parameters updated via a Gibbs step

Full Bayesian model need not match model for MAP estimates

Simulation Study

Synthetic dataset with four linking fields with two induced errors [3]

- Construct two files with 50,000 records each
- Initially 2.5 billion comparisons with 25,000 true matches
- Initial blocking on postal code (no induced errors)

Comparison vectors utilize string similarity and binary comparisons

- First Name (string) Compared with 4 levels of string similarity
- Last Name (string) Compared with 4 levels of string similarity
- Age (categorical) Binary comparison
- Occupation (categorical) Binary comparison

Selection of w_0 can be based on computational budget

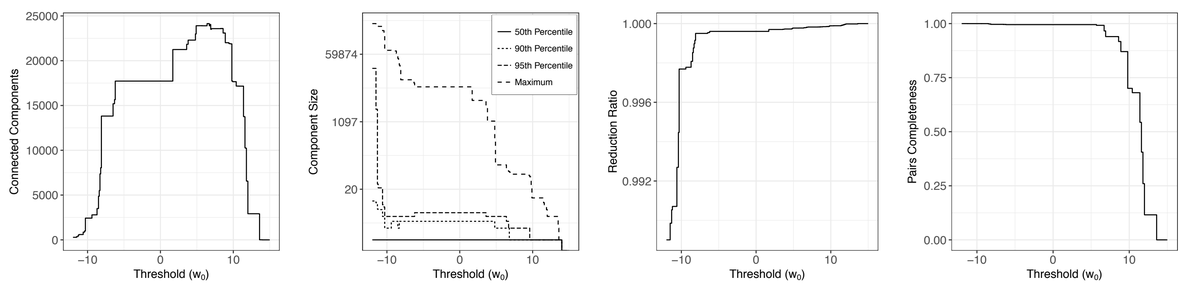

Figure 2: Post-hoc blocks resulting from varying threshold value w_0.

Post-hoc blocks contained almost all linked record pairs

	Record Pairs	Blocks	True Matches	Pairs completeness	Reduction ratio
Postal code blocking	160,207,943	29	25,000	-	-
Post-hoc blocking ($w_0 = 4.9$)	36,356	23,903	24,877	99.5%	99.98%

Feasible to estimate an approximate posterior with post-hoc blocking

	Recall	Precision
MAP Parameter Estimation ($\theta = 7.5$)	92.7%	94.6%
Restricted MCMC ($\theta = 7.5$, $w_0 = 4.9$)	95.5%	94.6%

Convergence not achieved using only postal code blocking

Conclusions

- Threshold w_0 allows trade-off between bias introduced into the estimate and computational constraints.
- Post-hoc blocking allows Bayesian inference to be to be performed on larger problems than was previously possible.

Future Work

- Apply to a broader range of PRL models.
- Expand approach to de-duplication problems.

References

[1] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.
[2] P. J. Green and K. V. Mardia. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93(2):235–254, 2006.
[3] M. Sadinle. Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518):600–612, 2017.

Corresponding Author: Brendan McVeigh, bmcveigh@stat.cmu.edu