Performance Bounds for Graphical Record Linkage

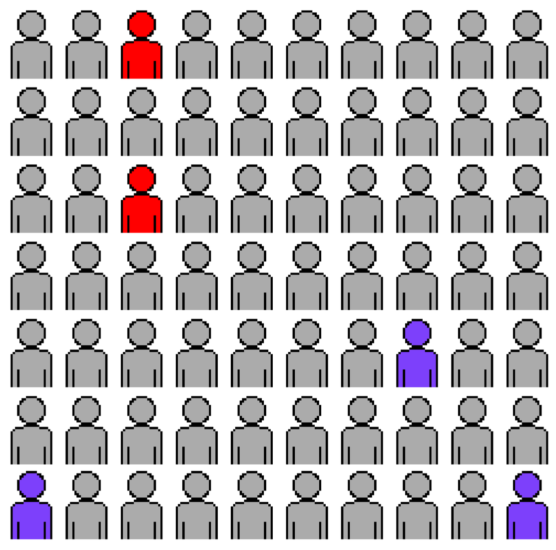
Rebecca C. Steorts[1], Matt Barnes[2], Willie Neiswanger[2]

[1]Duke University, Durham, NC

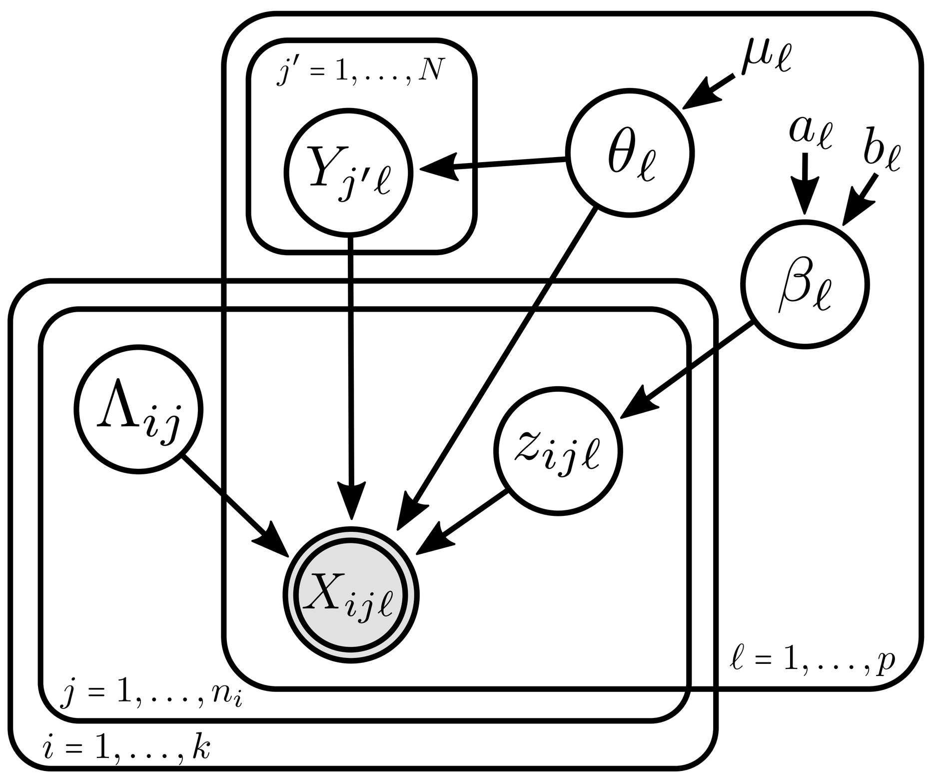
[2]Carnegie Mellon University, Pittsburgh, PA

Record Linkage

Record linkage (entity resolution or de-duplication) is the process of removing duplicate entities from large noisy databases.



Graphical Record Linkage



Kullback-Leibler (KL) divergence

For any two distributions P and Q, the maximum power for testing P versus Q is

$$\exp\{-nD_{\text{KL}}(P||Q)\}.$$

- A low value of D_{KL} means that we need many samples to distinguish P from Q.
- How does changing $\boldsymbol{Y}$ (latent entity) or $\boldsymbol{\Lambda}$ (linkage structure) change the distribution of $\boldsymbol{X}$ (observed records)?
- We search for both meaningful upper and lower bounds.

Assuming the conditions of [1, 2], let

$$\mathcal{P} = \{f(X \mid \boldsymbol{Y}, \Lambda_{ij}, \boldsymbol{\theta}, \boldsymbol{\beta}) : \forall \Lambda_{ij} \in \{1, \ldots, N\}.\}$$

- $X_1, X_2, \ldots, X_N$ are all independent given $(\boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta})$ under both $P, Q \in \mathcal{P}$.
- This implies that $D_{X_1,X_2,\ldots,X_N}(P||Q) = \sum_i D_{X_i}(P||Q)$.

Performance Bounds

Theorem 1. *This result finds an upper bound on the KL divergence and a lower bound for the probability that the categorical model in [1] gets the linkage structure incorrect. Let*

$$\gamma = \max_{\Lambda_{ij} \neq \Lambda'_{ij}} 2 \sum_{ij\ell} I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell})(1-\beta_\ell)\ell_n\left\{\frac{1}{\min_m \theta_{\ell m}\beta_\ell}\right\}.$$

i) The KL divergence is bounded above by γ. That is, $D_X(P||Q) \leq \gamma \;\; \forall P, Q \in \mathcal{P}$.

ii) The minimum probability of getting a latent entity wrong is $Pr(\Lambda_{ij} \neq \Lambda'_{ij}) \geq 1 - \dfrac{\gamma + \ell_n 2}{\ell_n r}, \;\; \forall i, j$

That is, as the latent entities become more distinct, γ increases. On the other hand, as the latent entities become more similar, $\gamma \to 0$.

Remark: Consider Theorem 1 (i). Suppose $\beta_\ell \to 1$. Then $D_{\boldsymbol{X}} \geq 0$. If instead $\beta_\ell \to 0$, then $D_{\boldsymbol{X}} \geq 1$. The lower bound is only informative when $\beta_\ell \to 0$. We have more information when the latent entities are separated.

Theorem 2. *Assume string and categorical data $\boldsymbol{X}$ as in [2] and distributions $P, Q \in \mathcal{P}$. Assume two distinct linkage structures, denoted by $Y_{\Lambda_{ij}\ell}, Y_{\Lambda'_{ij}\ell}$.*

i) There is an upper bound on the KL divergence between any $P, Q \in \mathcal{P}$ given by κ, that is $D_X(P||Q) \leq \kappa$.

ii) $Pr(\Lambda_{ij} \neq \Lambda'_{ij}) \geq 1 - \dfrac{\kappa + \ell_n 2}{\ell_n r}$, where

$$\kappa = \max_{\Lambda_{ij} \neq \Lambda'_{ij}} \Bigg[2 \sum_\ell (1-\beta_\ell) I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell}) + \sum_{\ell m} I(Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell})\left(1 - e^{-cd(Y_{\Lambda_{ij}\ell}, Y_{\Lambda'_{ij}\ell})}\right) \times E[e^{-cd(m, Y_{\Lambda_{ij}\ell})}]\Bigg] \ell_n\{(\min Q)^{-1}\}$$

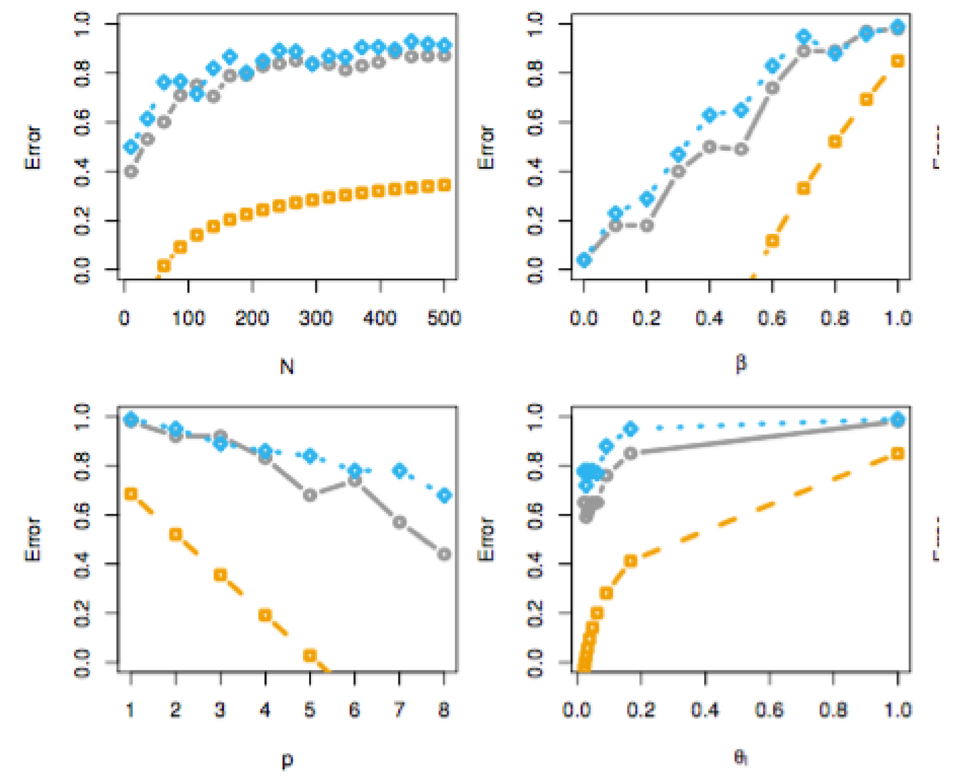
and $r+1$ is the cardinality of $\mathcal{P}$.

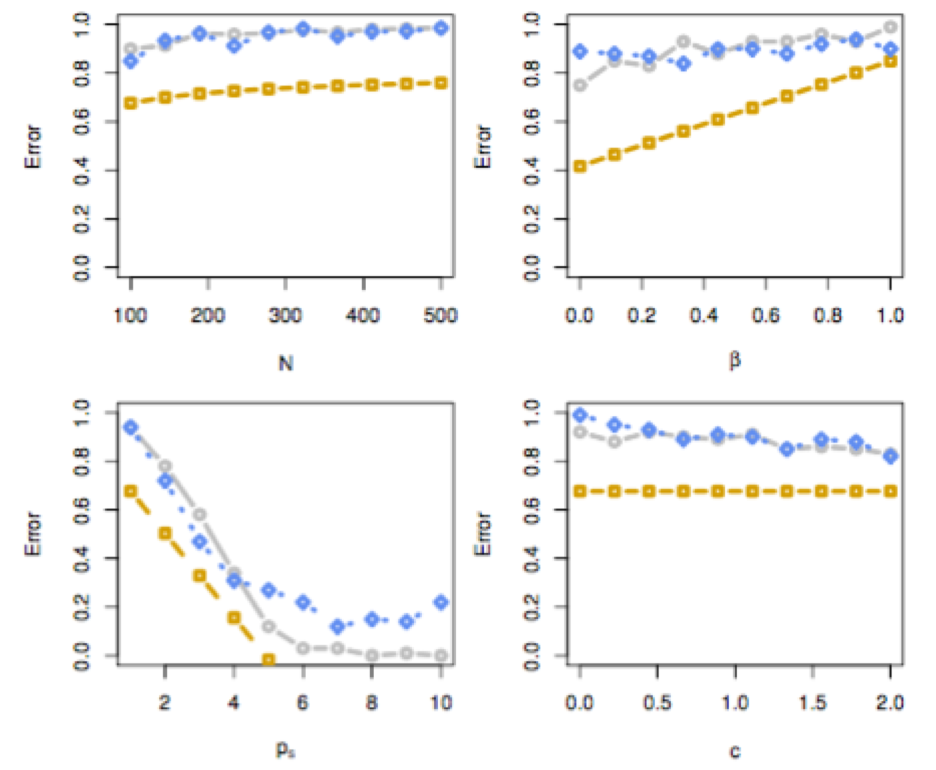
Priors on the Linkage Structure

- Above we a specific discrete uniform prior on $\boldsymbol{\Lambda}$.
- We extend this to include other discrete uniform priors on $\boldsymbol{\Lambda}$ including those that are informative.
- Special cases include the work of [3, 4, 5, 6].
- The theorem on performance bounds generalizes naturally, allowing comparisons to be made in future work.

Experiments

In our experiments (**Experiment I** and **Experiment II**), synthetic categorical data are generated according to the Steorts, Hall Fienberg (2014, 2016) or Steorts (2015) using the parameters in the figures below.





Conclusions and Acknowledgements

- We have proposed the first performance bounds, to our knowledge, for record linkage models.
- Is it possible to prove tighter bounds?
- Is it possible to compare to models outside of Gibbs partition prior models?

Acknowledgements: This work was supported in part by NSF CAREER Award SES-1652431 and SES-1534412. This poster is based upon the original open source work of Sofia Jijon (https://sjijon.github.io).

References

[1] R. C. Steorts, R. Hall, and S. E. Fienberg. SMERED: A Bayesian approach to graphical record linkage and de-duplication. *Journal o* 2014.

[2] R. C. Steorts. Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4):849–875, 2015.

[3] Giacomo Zanella, Brenda Betancourt, Jeffrey W Miller, Hanna Wallach, Abbas Zaidi, and Rebecca Steorts. Flexible models for with application to entity resolution. In *Advances in Neural Information Processing Systems*, pages 1417–1425, 2016.

[4] A. Tancredi and B. Liseo. A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, 5(2B):1553–1585, 2011.

[5] Mauricio Sadinle. Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *The Annals of Applied Statistics*, 8(4):2404–2434, 2014.

[6] Jim Pitman. *Combinatorial Stochastic Processes: Ecole D'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer, 2006.

[7] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Contact information:
Rebecca C. Steorts
resteorts.github.io