Performance Bounds for Graphical Record Linkage

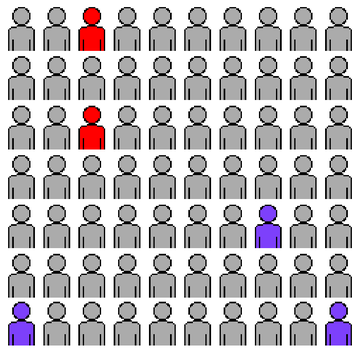
Rebecca C. Steorts,[1] Matt Barnes[2] and Willie Neiswanger[2]

Duke University[1] and Carnegie Mellon University[2]

Record Linkage

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

- Three data sets have duplicated data and require record linkage.



Graphical Record Linkage Models

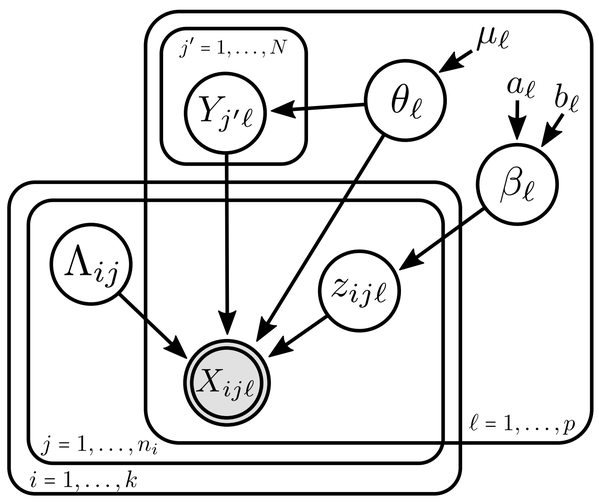


FIGURE 1: Graphical representation of models in Steorts (2015); Steorts et al. (2014).

Kullback-Leibler (KL) divergence

For any two distributions P and Q, the maximum power for testing P versus Q is $\exp\{-nD_{\mathrm{KL}}(P||Q)\}$.

- 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.

Performance Bounds

Assuming the conditions of Steorts (2015); Steorts et al. (2014), 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 $D_{X_1,X_2,\ldots,X_N}(P\|Q) = \sum_i D_{X_i}(P\|Q)$.

We provide two theorems under recent record linkage model, finally providing a general theorem.
Finally, we illustrate how the bounds hold on simulated data.

Theorem 1. *This result finds an upper bound on the KL divergence and a lower bound for the probability that the categorical model in Steorts et al. (2014) gets the linkage structure incorrect. Let*

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

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 $Pr(\Lambda_{ij} \neq \Lambda'_{ij}) \geq 1 - \dfrac{\gamma + \ln 2}{\ln r}, \ \ \forall i, j$

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 Steorts (2015) 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) $Pr(\Lambda_{ij} \neq \Lambda'_{ij}) \geq 1 - \dfrac{\kappa + \ln 2}{\ln r}$, *where*

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

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

Other 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 Pitman (2006); Sadinle (2014); Tancredi and Liseo (2011); Zanella et al. (2016).
- 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 Figures 2 and 3.

- In order to consider a realistic set of strings for S, we consider the set of 20 most popular female baby names from 2014, according to the United States Census. Then for the distance d, we consider the generalized Levenshtein edit distance.
- For each experiment, we vary exactly one of the parameters to demonstrate its impact of the linkage error rate $Pr((\hat{\Lambda}_{ij}, \boldsymbol{Y}) \neq (\Lambda_{ij}, \boldsymbol{Y}))$.
- We choose the other values such that the performance is neither extremely low nor extremely high. We set the distortion parameter β_ℓ to the same value for each ℓ, i.e. $\beta_\ell = 0.6$ denotes a distortion probability of 0.6 for every field. $\beta_\ell = 0.0$ to 1.0 means we started with $\beta_\ell = 0$ for all ℓ and swept the values until $\beta_\ell = 1$ for all ℓ.
- Recall p is the number of fields, and thus the maximum value of ℓ.
- We also set each $\theta_{\ell m}$ to the same value, i.e. $\theta_{\ell m} = 0.1$ denotes $\theta_{\ell m} = 0.1$ for all ℓ and all m. This further implies each field ℓ takes on exactly $M_\ell = 1/\theta_{\ell m}$ values in order for θ_ℓ to be a valid probability distribution.

Experiment	N	β_ℓ	$p = p_c$	$\theta_{\ell m}$
Fig. 1(a)	10 to 500	0.6	3	0.1
Fig. 1(b)	100	0 to 1	3	0.1
Fig. 1(c)	100	0.6	1 to 8	0.25
Fig. 1(d)	100	0.8	5	$\frac{1}{46}$ to 1

FIGURE 2: Categorical Experiments

Experiment	N	β_ℓ	$p = p_s$	c
Fig. 2(a)	100 to 500	0.6	1	1.0
Fig. 2(b)	100	0.2 to 1	1	1.0
Fig. 2(c)	100	0.6	1 to 10	1.0
Fig. 2(d)	100	0.6	1	0 to 2

FIGURE 3: String Experiments

Comparisons

- Exact sampler: samples directly from $Pr(\boldsymbol{\Lambda}|X, \boldsymbol{Y}, \boldsymbol{z})$
- Gibbs sampler: empirically motivated priors proposed by Steorts (2015). In order to compute the empirical probability $Pr(\hat{\Lambda}_{ij} \neq \Lambda_{ij})$, we hold $\boldsymbol{Y}$ fixed during Gibbs sampling to ensure errors in $\hat{\boldsymbol{\Lambda}}$ are not due to arbitrary changes in the ordering of the labels of $\boldsymbol{Y}$.

Results of Experiment I and II

- Not surprisingly, the bounds are not tight for the categorical model (Figure 4).
- However, we categorical data and string data are both used, the bounds are tight (Figure 5).
- The effects of parameter variation is less noticeable in the string experiments due to the fact that linking string fields is easier than ones that have been anonymized, i.e., categorical fields.

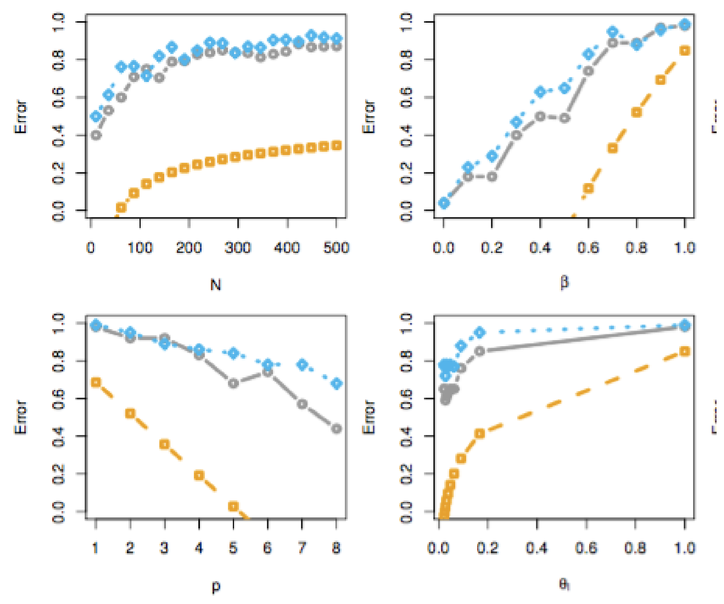


FIGURE 4: Theorem 1 (gold squares) holds on simulated categorical records compared to exact sampling (grey circles) and Gibbs sampler (blue diamonds).

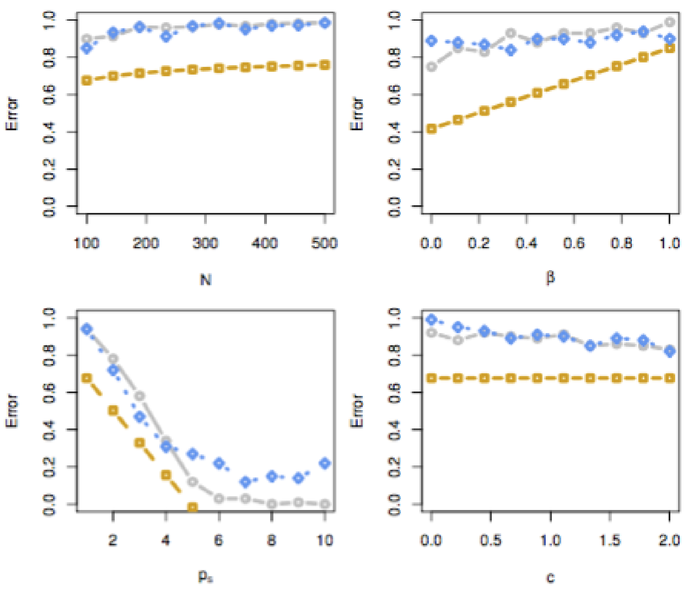


FIGURE 5: Theorem 2 (gold squares) holds on simulated noisy string records compared to exact sampling (grey circles) and Gibbs sampler (blue diamonds).

Discussion

- Is it possible to prove tighter bounds?
- Is it possible to compare to models outside of Gibbs partition prior models?
- Can we avoid the label switching issue to make the performance bounds practical for real data?

Acknowledgements: This work was supported in part by NSF CAREER Award SES-1652431 and SES-1534412.

References

PITMAN, J. (2006). *Combinatorial Stochastic Processes: Ecole D'Eté de Probabilités de Saint-Flour XXXII-2002*. Springer.

SADINLE, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *The Annals of Applied Statistics*, **8** 2404–2434.

STEORTS, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, **10** 849–875.

STEORTS, R. C., HALL, R. and FIENBERG, S. E. (2014). SMERED: A Bayesian approach to graphical record linkage and de-duplication. *Journal of Machine Learning Research*, **33** 922–930.

TANCREDI, A. and LISEO, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, **5** 1553–1585.

ZANELLA, G., BETANCOURT, B., MILLER, J. W., WALLACH, H., ZAIDI, A. and STEORTS, R. (2016). Flexible models for microclustering with application to entity resolution. In *Advances in Neural Information Processing Systems*. 1417–1425.