Supplementary Material for "Performance Bounds for Graphical Record Linkage"

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

Matt Barnes
The Robotics Institute
Carnegie Mellon University
mbarnes1@cs.cmu.edu

Willie Neiswanger
Machine Learning Department
Carnegie Mellon University
willie@cs.cmu.edu

A Example of the Record Linkage Process

We provide a toy illustration of the general record linkage process in figure 1. Consider three databases D_1, D_2, D_3 and the notation already introduced, where here $k = 3$. Suppose the "population" entities have four members, where name and address are stripped for anonymity and they are listed by state, age, and sex, as is often the case with de-identified data.

For instance, assume the true latent entity vector $\boldsymbol{y}$ is *known*:

$$\boldsymbol{y} = \begin{bmatrix} \text{NC, 72, F} \\ \text{SC, 73, F} \\ \text{PA, 91, M} \\ \text{VA, 94, M} \end{bmatrix}.$$

The observed records $\boldsymbol{X}$ are given in three separate databases (k=3), which would combine into a three-dimensional array. We write this here as three two-dimensional arrays for notational simplicity:

$$D_1 = \begin{bmatrix} \text{NC, 72, F} \\ \text{SC, 70, F} \\ \text{PA, 91, M} \end{bmatrix}, D_2 = \begin{bmatrix} \text{SC, 37 , F} \\ \text{VA, 93, M} \\ \text{PA, 92, M} \end{bmatrix},$$

$$D_3 = \begin{bmatrix} \text{NC, 72 , F} \\ \text{NC, 72, F} \\ \text{SC, 72, F} \\ \text{VA, 94, M} \end{bmatrix}.$$

Here, for the sake of keeping the illustration simple, only age is distorted. Comparing $\boldsymbol{X}$ to $\boldsymbol{y}$, the intended linkage and distortions are

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 2 & 3 & \\ 2 & 4 & 3 & \\ 1 & 1 & 2 & 4 \end{bmatrix},$$

$$\boldsymbol{z_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \boldsymbol{z_2} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \boldsymbol{z_3} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

In this linkage structure, every entry of $\boldsymbol{\Lambda}$ with a value of 2 means that some record from $\boldsymbol{X}$ refers to the latent entity with attributes "SC, 73, F." Here, the age of this entity is distorted in all three databases, as can be seen from $\boldsymbol{z}$. (Note that $\boldsymbol{z}$, like $\boldsymbol{X}$, is also really a three-dimensional array.) Looking at $\boldsymbol{z}_1$ and $\boldsymbol{z}_3$, we see that there is only a single record in either list that is distorted, and it is only distorted in one field. In list 2, however, every record is distorted, though only in one field.

Figure 1 illustrates the interpretation of the linkage structure as a bipartite graph in which each edge links a record to a latent entity. For clarity, figure 1 shows that X_{11} and X_{22} are the same entity and shows that X_{13}, X_{21}, and X_{34} correspond to the same entity. The rest are non-matches (or singleton entities).

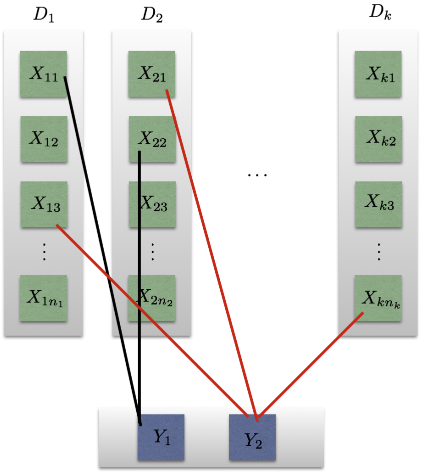

Figure 1: A general illustration of the record linkage process. We assume databases $D_1, \ldots D_k$. We assume records $\boldsymbol{X}$ that we cluster to latent entities $\boldsymbol{Y}$. Records that belong to the same same latent entity are kept track of using the linkage structure $\boldsymbol{\Lambda}$.

B Derivation of Theorem 1

We briefly restate the theorem, and then provide its derivation.

Theorem 1. *Assume data $\boldsymbol{X}$, 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

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

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

Proof. We first prove (i). Consider

$$\begin{aligned}
&Pr(X_{ij\ell} = m \mid \boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta}) \\
&= Pr(X_{ij\ell} = m \mid \boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta}, Z_{ij\ell} = 1) \\
&\times Pr(Z_{ij\ell} = 1 \mid \boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta}) \\
&\quad + Pr(X_{ij\ell} = m \mid \boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta}, Z_{ij\ell} = 0) \\
&\times Pr(Z_{ij\ell} = 0 \mid \boldsymbol{Y}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{\beta}) \\
&\propto I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \alpha_\ell(X_{ij\ell})\beta_\ell \\
&\times \Big[\exp\{-c\, d(X_{ij\ell}, Y_{\Lambda_{ij}\ell})\} \Big]. \qquad (1)
\end{aligned}$$

Suppose that $Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell}$ Equation 1 implies that

$$D_{X_{ij\ell}}(P\|Q) \propto \sum_{m=1}^{M_\ell} I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \alpha_\ell(m)\beta_\ell \times \Big[e^{-c\, d(X_{ij\ell}, Y_{\Lambda_{ij}\ell})} \times \phi \Big], \qquad (2)$$

where $\phi =$

$$\log \left[\frac{I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) + \alpha_\ell(m)\beta_\ell \Big[e^{-c\, d(m, Y_{\Lambda_{ij}\ell})} \Big]}{I(Y_{\Lambda'_{ij}\ell} = m)(1-\beta_\ell) + \alpha_\ell(m)\beta_\ell \Big[e^{-c\, d(m, Y'_{\Lambda'_{ij}\ell})} \Big]} \right].$$

We now consider $\|P - Q\|_1$ and by equation 2, we find

$$\begin{aligned}
\|P - Q\|_1 = \sum_{m \in M_\ell} \Big| & I(Y_{\Lambda_{ij}\ell} = m)(1-\beta_\ell) \\
&+ \alpha_\ell(m)\beta_\ell \exp\{-c\, d(m, Y_{\Lambda_{ij}\ell})\} \\
&- I(Y_{\Lambda'_{ij}\ell} = m)(1-\beta_\ell) \\
&- \alpha_\ell(m)\beta_\ell \exp\{-c\, d(m, Y_{\Lambda'_{ij}\ell})\} \Big|. \qquad (3)
\end{aligned}$$

Then by equation 3, it is clear that

$$\begin{aligned}
\|P - Q\|_1 \leq & \sum_m (1-\beta_\ell) \Big| \Big[I(Y_{\Lambda_{ij}\ell} = m) - I(Y_{\Lambda'_{ij}\ell} = m) \Big] \Big| \\
&+ \sum_m \alpha_\ell(m)\beta_\ell \\
&\times \Big| \exp\{-c\, d(m, Y_{\Lambda_{ij}\ell}) - \exp\{-c\, d(m, Y_{\Lambda'_{ij}\ell})\} \Big| \\
&\leq 2(1-\beta_\ell) + \beta_\ell \sum_m \alpha_\ell(m) \\
&\times \Big| \exp\{-c\, d(m, Y_{\Lambda_{ij}\ell}) - \exp\{-c\, d(m, Y_{\Lambda'_{ij}\ell})\} \Big|.
\end{aligned}$$

Now assume that two field attributes are different. That is, suppose there exists an $m \neq m'$. Then we assume that there exists a $\delta > 0$ such that $d(m, m') \geq \delta$. By the reverse triangle inequality, for any m, m', m'',

$$\begin{aligned}
|d(m, m') - d(m, m'')| \leq d(m', m'') \implies \\
e^{-c[d(m,m') - d(m,m'')]} \geq e^{-cd(m', m'')}. \qquad (4)
\end{aligned}$$

Equation 4 in turn implies that

$$\begin{aligned}
&\sum_m \Big[\Big(1 - e^{-c[d(m,m') - d(m,m'')]}\Big) e^{-cd(m',m'')} \alpha_\ell(m) \Big] \\
&\geq \sum_m \Big(1 - e^{-c[d(m',m'')]}\Big) e^{-cd(m',m'')} \alpha_\ell(m).
\end{aligned}$$

Then

$$\begin{aligned}
&\sum_m \alpha_\ell(m) \Big[e^{-cd(m,m')} - e^{-cd(m,m'')} \Big] \\
&= \sum_m \alpha_\ell(m) e^{-cd(m,m')} \Big(1 - e^{-cd(m',m'')}\Big) \\
&= \Big(1 - e^{-cd(m',m'')}\Big) \sum_m \alpha_\ell(m) e^{-cd(m,m')} \\
&= \Big(1 - e^{-cd(m',m'')}\Big) E[e^{-cd(m,m')}],
\end{aligned}$$

where $M \sim \alpha_\ell$.

That is,

$$\sum_m \alpha_\ell(m) e^{-cd(m,m')}$$

is the moment generating function of $d(M, m')$ (evaluated at c), where $M \sim \alpha_\ell$. This implies that

$$\begin{aligned} &\|P - Q\|_1 \\ &\leq 2(1-\beta_\ell) + \beta_\ell \sum_m \\ &\qquad \times \left(1 - e^{-cd(Y_{\Lambda_{ij}\ell}, Y_{\Lambda'_{ij}\ell})}\right) E[e^{-cd(m, Y_{\Lambda_{ij}\ell})}]. \end{aligned}$$

Then by reverse Pinker's inequality [1], we can write

$$\begin{aligned} &\max_{P,Q \in \mathcal{P}} D_{\boldsymbol{X}}(P||Q) \\ &\qquad \leq \max_{\Lambda_{ij} \neq \Lambda'_{ij}} \Bigg[2 \sum_{ij\ell} (1-\beta_\ell) I(Y_{\Lambda_{ij}\ell} \neq Y'_{\Lambda_{ij}\ell}) \\ &\qquad + \sum_{ij\ell m} I(Y_{\Lambda_{ij}\ell} \neq Y'_{\Lambda_{ij}\ell}) \left(1 - e^{-cd(Y_{\Lambda_{ij}\ell}, Y_{\Lambda'_{ij}\ell})}\right) \\ &\qquad \times \left[E[e^{-cd(m, Y_{\Lambda_{ij}\ell})}] \right] \times \ln\{(\min Q)^{-1}\} \Bigg] =: \kappa, \end{aligned}$$

where

$$Q = I(Y_{\Lambda'_{ij}\ell} = m)(1-\beta_\ell) - \alpha_\ell(m)\beta_\ell \exp\{-c\, d(m, Y_{\Lambda'_{ij}\ell})\}.$$

Thus, (i) is established. Using Fano's inequality, we find that

$$Pr(\hat{\Lambda}_{ij} \neq \Lambda_{ij}) \geq 1 - \frac{\kappa + \ln 2}{\ln r}.$$

We have established that for any $Y_{\Lambda_{ij}\ell} \neq Y_{\Lambda'_{ij}\ell}$, the minimum probability of getting a latent entity wrong is governed by the constant c. That is, the lower bound grows as c goes to ∞, and its rate of growth is determined by the moment generating function of the distances. We have now established (ii). $\square$

References

[1] Daniel Berend, Peter Harremoës, and Aryeh Kontorovich. Minimum KL-divergence on complements of l_1 balls. *IEEE Transactions on Information Theory*, 60(6):3172–3177, 2014.