Connecting Bipartite Record Linkage Models

Connecting Bipartite Record Linkage Models

Edoardo Redivo

Department of Statistical Sciences, University of Bologna,
edoardo.redivo@unibo.it

Abstract. Bipartite record linkage aims to link observations of the same individual across two distinct non-duplicated datasets. The two main approaches to solve this task are the Fellegi-Sunter model, which is based on comparing all pairs of observations; and the graphical record linkage model, which explicitly considers the data generating process and links observations to latent entities. In this work, we explore the similarities between these two methods. Specifically, we show that their parameters can be directly related under a common data model; and that they can be estimated within the same framework using a classification expectation-maximization algorithm, taking into account the problem constraints and allowing for the introduction of prior information.

Keywords: record linkage, entity resolution, CEM algorithm, clustering

1 Introduction

Record linkage, or entity resolution, aims at identifying observations coming from the same statistical unit, in settings where a unique identifier or key is missing. This field has two main settings: deduplication, where a single dataset contains duplicate observations of the same entity; and Bipartite Record Linkage (BRL), where two datasets, without duplicates within, share some entities. Statistical methods for entity resolution define a probabilistic model to detect the unique entities taking into account the fact that observations from the same entity are not necessarily equal as different variable representations, temporal variation, noise or error mechanisms, and missing data might affect the datasets. The main goals of entity resolution are integrating databases, and improving data quality and the reliability of subsequent analyses. Among its different applications there are those in the fields of official statistics and social sciences, where it can be used to link datasets containing different sets of variables, or creating longitudinal datasets from panel data. For a recent review on entity resolution, see [1].

The first statistical framework for record linkage, proposed by [2], is based on assessing the similarity between pairs of observations and remains a foundational method in probabilistic record linkage, known as the Fellegi-Sunter model. More recently, graphical entity resolution models have been proposed, which link observations to latent entities mimicking a data-generating process with an error mechanism.

In this work we assess the similarity between comparison-based model from [2] and graphical models in the case of BRL. More precisely, we show that they

can be seen as two different ways of estimating the same probabilistic model, leading to a direct relationship between their parameters. Moreover, they can be dealt within a common framework for the latent linkage structure and its prior information. Furthermore, we propose a common way for estimating both models via a classification expectation-maximization algorithm, which provides a new and efficient method for estimating graphical models for BRL.

2 Models for Bipartite Record Linkage and Their Connections

Denote the two datasets as $\mathbf{X}_1$ and $\mathbf{X}_2$, with n_1 and n_2 observations respectively, and $n_1 \geq n_2$, so that the maximum possible number of coreferent pairs (pairs of observations referring to the same entity) is n_2. The two datasets have p common variables used in the linkage process, which are called the matching or key variables. For each of the $n_1 \times n_2$ pairs made up of one observation from $\mathbf{X}_1$, $\mathbf{x}_{1i}$ with $i = 1, \ldots, n_1$, and one from $\mathbf{X}_2$, $\mathbf{x}_{2j}$ with $j = 1, \ldots, n_2$, we compare the values across the matching variables, resulting in comparison data $\boldsymbol{\Gamma}$, made of vectors $\boldsymbol{\gamma}_{ij} = (\gamma_{ij1}, \ldots, \gamma_{ijh}, \ldots, \gamma_{ijp})$, whose generic element is defined as:

$$\gamma_{ijh} = \begin{cases} 1 & \text{if } x_{1ih} = x_{2jh} \\ 0 & \text{otherwise.} \end{cases}$$

The linkage structure information is stored in a latent binary matrix $\Delta \in \mathbb{R}^{n_1 \times n_2}$, called coreference matrix [3], where:

$$\Delta_{ij} = \begin{cases} 1 & \text{if } \mathbf{x}_{1i} \text{ and } \mathbf{x}_{2j} \text{ are coreferent} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The Fellegi-Sunter model makes the following distributional assumptions:

$$\begin{aligned} \gamma_{ijh} \mid \Delta_{ij} = 1 &\overset{\text{iid}}{\sim} \text{Bernoulli}(m_h) \\ \gamma_{ijh} \mid \Delta_{ij} = 0 &\overset{\text{iid}}{\sim} \text{Bernoulli}(u_h) \\ \Delta_{ij} &\overset{\text{iid}}{\sim} \text{Bernoulli}(\lambda), \end{aligned} \tag{2}$$

where m_h and u_h are the probabilities of showing the same value for variable h conditional on the pair being respectively coreferent or not, while parameter λ is the probability of a pair being coreferent.

The main issue with this model is its assumption of independence in coreference status among pairs, which can lead to a many-to-many relationship without even ensuring transitive closure. By transitive closure, it is meant that if an observation a matches with b, and b matches with c, then a must also match with c; this property applies to any kind of entity resolution task. In the BRL setting, since no duplicates are present within the two datasets, each observation in one

dataset should be linked to at most one observation in the other; in terms of the coreference matrix this translates to the constraints:

$$\begin{aligned} \sum_{j=1}^{n_2} \Delta_{ij} \leq 1 \quad i = 1, \ldots, n_1 \\ \sum_{i=1}^{n_1} \Delta_{ij} \leq 1 \quad j = 1, \ldots, n_2. \end{aligned} \tag{3}$$

From a Bayesian perspective, the problem can be solved by defining a prior on the coreference matrix following the constrains, as in [4–6].

In the graphical model for record linkage proposed in [7, 8], instead of working with comparisons data, the observed data,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times p},$$

is modeled according to:

$$\begin{aligned} x_{ih} \mid z_i = k, y_{kh}, s_{ih}, \boldsymbol{\theta}_h &\sim \begin{cases} \delta_{y_{kh}} & s_{ih} = 0 \\ \text{Categorical}(L_h, \boldsymbol{\theta}_h) & s_{ih} = 1 \end{cases} \\ s_{ih} \mid \beta_h &\sim \text{Bernoulli}(\beta_h) \\ y_{kh} \mid \boldsymbol{\theta}_h &\sim \text{Categorical}(L_h, \boldsymbol{\theta}_h), \end{aligned} \tag{4}$$

with $i = 1, \ldots, n_1 + n_2$, $h = 1, \ldots, p$, $k = 1, \ldots, K$. Observed data entries x_{ih} are assumed to be categorical, with variable h taking values in $\{1, \ldots, L_h\}$, and to be either equal to their latent entity value y_{kj}, or to have been distorted according to the binary variable s_{ih}. The number of latent entities K is unknown and equal to the number of unique values in the allocation vector $\mathbf{z} = (z_1, \ldots, z_{n_1+n_2})$. This framework highlights the fact that record linkage is fundamentally a clustering problem. In [9], the likelihood conditional on $\mathbf{z}$ is derived as:

$$f(\mathbf{X} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{h=1}^{p} f(\mathbf{x}_{[h]}^{(k)} \mid \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{k=1}^{K} \prod_{h=1}^{p} \left[\prod_{i:z_i=k} \beta_h \theta_{h x_{ih}} \right] g(\mathbf{x}_{[h]}^{(k)}, \beta_h, \boldsymbol{\theta}_h), \tag{5}$$

where $\mathbf{x}_{[h]}^{(k)} = \{x_{ih} : z_i = k\}$, and the function $g(\cdot)$ is defined as:

$$g(\mathbf{x}_{[h]}^{(k)}, \beta_h, \boldsymbol{\theta}_h) = \sum_{l=1}^{L_h} \theta_{hl} \prod_{i:z_i=k} \frac{\beta_h \theta_{h x_{ih}} + (1 - \beta_h) \mathbb{1}[x_{ih} = l]}{\beta_h \theta_{h x_{ih}}}.$$

3 Linking the Two Approaches

The likelihood in Equation (5) is the starting point for the comparison between the Fellegi-Sunter model and the graphical record linkage model. In BRL the admissible partitions for clustering are limited to those containing sets of size two,

linking two observations from the two datasets, and singletons for all remaining observations. This domain is captured by the coreference matrix (Equation 1), which allows us to describe any such BRL partition and to rewrite the likelihood as:

$$f(\mathbf{X} \mid \Delta, \boldsymbol{\beta}, \boldsymbol{\theta}) = \left[\prod_{i=1}^{n_1}\prod_{j=1}^{n_2}\left[\prod_{h=1}^{p}\left(\beta_h(2-\beta_h) + (1-\beta_h)^2 \prod_{l=1}^{L_h} \frac{\mathbb{1}[x_{1ih}=l]\mathbb{1}[x_{2jh}=l]}{\theta_{hl}}\right)\right]^{\Delta_{ij}}\right] \left[\prod_{i=1}^{n_1}\prod_{h=1}^{p}\theta_{hx_{1ih}}\right]\left[\prod_{j=2}^{n_2}\prod_{h=1}^{p}\theta_{hx_{2jh}}\right]. \tag{6}$$

The corresponding conditional likelihood for the comparison-based Fellegi-Sunter model depending on parameters $\mathbf{m} = (m_1, \ldots, m_h, \ldots, m_p)$ and $\mathbf{u} = (u_1, \ldots, u_h, \ldots, u_p)$ (Equation 2) can be recast in terms of pairwise entry comparisons in the original data as:

$$f(\boldsymbol{\Gamma} \mid \Delta, \mathbf{m}, \mathbf{u}) = \prod_{i=1}^{n_1}\prod_{j=1}^{n_2}\left[\prod_{h=1}^{p}\left[\frac{m_h}{u_h}\right]^{\mathbb{1}[x_{1ih}=x_{2jh}]}\left[\frac{1-m_h}{1-u_h}\right]^{\mathbb{1}[x_{1ih}\neq x_{2jh}]}\right]^{\Delta_{ij}} \left[\prod_{h=1}^{p} u_h^{\mathbb{1}[x_{1ih}=x_{2jh}]}(1-u_h)^{\mathbb{1}[x_{1ih}\neq x_{2jh}]}\right]. \tag{7}$$

Equations (6) and (7) present some similarities in their factorization structure depending on coreferent pairs, but they are not equivalent, because they specify the distribution for two different things: the data itself and all pairwise comparisons respectively.

If we assume the data generating process of the graphical model also for the Fellegi-Sunter then, we can write parameters $\mathbf{m}$ and $\mathbf{u}$ in terms of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$:

$$m_h = (1-\beta_h)^2 + 2(1-\beta_h)\beta_h \sum_{l=1}^{L_h}\theta_{hl}^2 + \beta_h^2 \sum_{l=1}^{L_h}\theta_{hl}^3 + \beta_h^2 \sum_{l=1}^{L_h}\theta_{hl}\sum_{l^*\neq l}\theta_{hl^*}^2$$

$$u_h = \sum_{l=1}^{L_h}\theta_{hl}^2$$

Expressing m and u probabilities based on an error mechanism is exactly one of the methods proposed in [2] for estimating these parameters, although a full data generating mechanism such as that in Equation (4) is not given explicitly. However, using formulas such as those in [2], or the ones just derived, for parameter estimation requires a priori knowledge and quantification of the possible distortion processes and for this reason it is not the standard estimation procedure.

4 A Common Estimation Framework

The Fellegi-Sunter model is usually estimated via an Expectation-Maximization (EM) algorithm. This has the disadvantage of having to explicitly consider the Bernoulli prior distribution for the entries of the coreference matrix (Equation 2). As a result, independent posterior distributions must be computed for each entry, which may lead to a maximum a posteriori estimate of the coreference matrix that does not follow the bipartite structure.

In [10] a post-processing step is presented ensuring an at-most one-to-one match between the observations in the two datasets. It consists in solving the following linear programming optimization problem:

$$\max_{\Delta} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Delta_{ij} \log \frac{f(\boldsymbol{\gamma}_{ij} \mid \Delta_{ij} = 1)}{f(\boldsymbol{\gamma}_{ij} \mid \Delta_{ij} = 0)},$$

subject to the constraints in Equation (3), where the conditional distributions are evaluated based on estimates $\hat{\mathbf{m}}$ and $\hat{\mathbf{u}}$ from the EM algorithm. In [5] it has been shown that the solution of the optimization problem maximizes the conditional likelihood of Equation (2) with respect to Δ. This fact enables the development of a Classification Expectation-Maximization (CEM) algorithm [11], guaranteeing the convergence to a stationary value for the complete data likelihood, which in the case of the Fellegi-Sunter model is given by:

$$f(\boldsymbol{\Gamma}, \Delta \mid \mathbf{m}, \mathbf{u}) = f(\boldsymbol{\Gamma} \mid \Delta, \mathbf{m}, \mathbf{u}) f(\Delta),$$

where the first term is the conditional likelihood of Equation (7), and the second is the prior on the coreference matrix. The previously mentioned linear optimization problem corresponds to performing together the expectation and maximization steps for the CEM algorithm under a uniform prior over the coreference matrices. The CEM is more computationally expensive than the EM as it requires to solve a linear program at each iteration, but it has the advantage of being able to include the bipartite constraints in the estimation itself.

Sharing the same structure relating to the coreference matrix, also the graphical record linkage model can be estimated via a CEM algorithm. Moreover, this common framework allows for the interchangeable use of the same latent distributions on the coreference matrix between the two models, provided their logarithm is a linear function of matrix entries. One such case is the `fabl` prior [6], originally proposed for the Fellegi-Sunter model, which sets a distribution on the proportion of observations in $\mathbf{X}_2$ having a match, called the overlap proportion.

In Figure 1 we assess the estimates from the two models, in a simple scenario with $n_1 = 100$, $n_2 = 70$, and $\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Delta_{ij} = 40$, with data simulated according to the graphical model of Equation (4), with $\beta = 0.05$ and a uniform distribution over $L = 10$ categories for $\boldsymbol{\theta}$ for all $p = 5$ variables. Both models provide the same estimated coreference matrix, shown on the left panel, with a false positive rate of $1/(7000 - 40) \approx 0.01\%$ and a false negative rate of $2/40 = 5\%$.

The exploration of the trade-offs between the two approaches needs to be investigated further via theoretical and computational arguments.

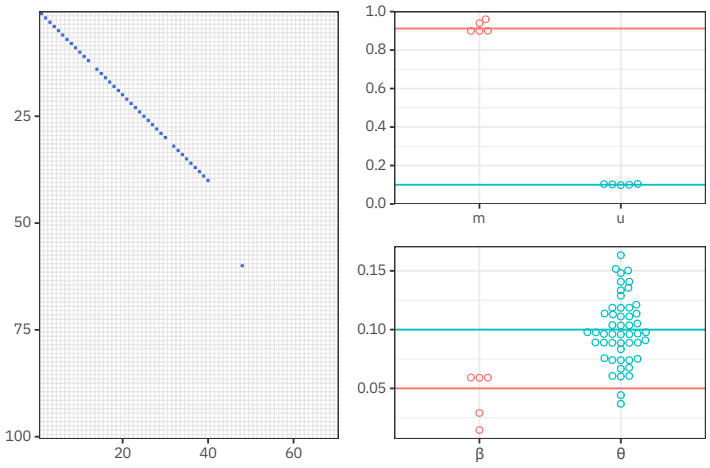

Fig. 1. Estimated parameters from the two BRL models. On the left panel the coreference matrix $\hat{\Delta}$, on the right panel the other parameters from the two models with horizontal lines indicating their true value.

References

1. Binette, O., Steorts, R.C.: (Almost) all of entity resolution. Sci. Adv. 8, eabi8021 (2022). doi:10.1126/sciadv.abi8021
2. Fellegi, I.P., Sunter, A.B.: A Theory for Record Linkage. J. Am. Stat. Assoc. 64(328), (1969). doi:10.1080/01621459.1969.10501049
3. Sadinle, M.: Detecting duplicates in a homicide registry using a Bayesian partitioning approach. Ann. Appl. Stat. 8(4), (2014). doi:10.1214/14-AOAS779
4. Fortini, M., Liseo, B., Nuccitelli, A., Scanu, M.: On Bayesian Record Linkage. Research in Official Statistics, 4, (2001).
5. Sadinle, M.: Bayesian Estimation of Bipartite Matchings for Record Linkage. J. Am. Stat. Assoc. 112(518), (2017). doi: 10.1080/01621459.2016.1148612
6. Kundinger, B., Reiter, J.P., Steorts, R.C.: Efficient and Scalable Bipartite Matching with Fast Beta Linkage (fabl). Bayesian Anal. (2024). doi:10.1214/24-BA1427
7. Steorts, R.C., Hall, R., Fienberg, S.E.: A Bayesian Approach to Graphical Record Linkage and Deduplication. J. Am. Stat. Assoc. 111(516), (2016). doi:10.1080/01621459.2015.1105807.
8. Steorts, R.C.: Entity Resolution with Empirically Motivated Priors. Bayesian Anal. 10(4) 849-875, (2015). doi:10.1214/15-BA965SI
9. Betancourt, B., Zanella, G., Steorts, R.C.: Random Partition Models for Microclustering Tasks. J. Am. Stat. Assoc. 117(539), (2020). doi:10.1080/01621459.2020.1841647
10. Jaro, M.A.: Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J. Am. Stat. Assoc. 84(406), (1989). doi:10.1080/01621459.1989.10478785
11. Celeux G., Govaert G.: A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data Anal. 14(3), (1992). doi:10.1016/0167-9473(92)90042-E