Improving record linkage via the application of Occam's razor

Duncan Smith and Mark Elliot

University Of Manchester

January 15, 2018

Abstract

Probabilistic record linkage is used to identify records in distinct datasets that correspond to the same entities. Classical probabilistic record linkage only considers the information contained in the variables common to two datasets A and B. Information contained in the variables that appear only in A or only in B is ignored. In recent years attempts have been made to exploit this information.

Approaches have either been *ad hoc*, or have been computationally expensive Markov Chain Monte-Carlo approaches that (to some degree) have moved away from the classical approach. Here we present a theoretically grounded approach that is an extension of the classical approach. The motivating idea is to improve record linkage by indirectly applying Occam's razor by requiring a set of chosen links to be consistent with a parsimonious generating process. Other approaches have been motivated by goals of improving the estimation of model parameters or accounting for matching constraints. Although our goal is different, we show (as others have before) that linkage performance can be improved by combining the record linkage procedure with statistical modelling.

Keywords: Bayesian methods; expectation maximization; graphical models

1 Introduction

Data are often contained in distinct databases held by distinct organisations. It is sometimes desirable to identify records in such databases that relate to the same entities. This is termed record linkage. The entities in question are often individuals. Record linkage might be used to replace high cost surveys, fill in missing values in survey data, or to produce combined data sets that can address research questions that cannot be addressed using existing data sets. For instance, we might combine historic survey data with a database of offenders to identify factors that might influence the onset of offending. There is great interest in finding improved approaches to linkage. Better linkage should generally result in more robust research conclusions. For a comprehensive review of record linkage approaches see Christen (2012).

In the absence of unique identifiers probabilistic approaches to linkage must be employed. These assign match probabilities to each possible match (record pair). The classical approach to probabilistic record linkage (Fellegi and Sunter, 1969) compares values on the variables that are common to a pair of datasets A and B. These are termed the *key* variables. Comparisons are for equality, and the results of the comparisons on the key variables for a record pair are assumed to be mutually independent given the match status of the pair. Record pairs that cannot be confidently classified as matches or non-matches are sent for clerical review – manual classification by a human.

There have been many subsequent attempts to improve on the classical approach. Some have sought to remove the binary comparison assumption in order to exploit similarities between values (e.g. Winkler, 1990; Smith and Shlomo, 2014). Others have sought to eliminate the conditional independence assumption (e.g. Tancredi and Liseo, 2011). Performance can also be improved by incorporating matching constraints. In particular, the assumption that neither file contains duplicated records is often reasonable. In this case a record in A can match at most a single record from B, and vice versa. The Hungarian algorithm (Kuhn, 1955) can be used as a post hoc procedure to find a maximal 1 to 1 matching that maximizes the product of the match probabilities for the included links. Some approaches have sought to more adequately account for the dependencies implied by 1 to 1 matching by incorporating the constraint into the linkage procedure itself (Tancredi and Liseo, 2011; Gutman et al., 2013; Sadinle, 2017; Steorts et al., 2016).

Here we seek to improve linkage performance by exploiting the relationships between non-key variables. Within the classical linkage approach this source of information is ignored, although prior knowledge of these relationships would be exploited in the clerical review process. For example, a record pair that (if merged) would imply a 6 year old CEO of a large company would likely be (manually) classified as a non-match.

It has been previously recognised that the "link then analyse" approach can be improved upon – "It is important to conceptualize the linkage and analysis steps as part of a single statistical system", Scheuren and Winkler (1993). Scheuren and Winkler (1997) demonstrated the benefits of a more integrated approach for linkage and linear regression modelling. The authors specified a simple straight line model and used it to generate data for the response variable Y and the independent variable X. Y and X were adjoined to datasets A and B with known match status. A and B were initially linked using the standard Fellegi-Sunter approach. Then Y was regressed against X using the linked data (see Scheuren and Winkler (1993) for details). Record pairs corresponding to large residuals were removed. The model was then re-fitted, and the X-values for record pairs corresponding to large residuals were imputed from the fitted model. The model was then re-fitted using the retained record pairs (including those with imputed values). The authors showed that both the fitted model and linkage could be improved. Furthermore, the process did not need to be iterated many times to achieve significant benefits.

Other approaches which have exploited the non-key variables tend to be based on Markov Chain Monte-Carlo methods (Gutman et al., 2013; Hof et al., 2017). Our approach is more similar to that of Scheuren and Winkler (1997) in that we retain much of the classical linkage framework, and combine linkage and modelling. However,

our approach does not require prior information regarding the relationships between variables – "A crucial practical assumption for the work of this paper is that analysts are able to produce a reasonable model (guesstimate) for the relationships between the noncommon quantitative items", Scheuren and Winkler (1997). Our motivation stems directly from Occam's razor – we should prefer a set of links that is consistent with a parsimonious generating process. Thus the modelling component seeks models with simple structure, rather than just improved parameter estimates. That is the essential difference between our approach and that of Scheuren and Winkler (1997). We manipulate the model structure rather than the data. We show that our approach can be effective at improving both record linkage and the quality of the resulting model.

In Section 2 we provide an overview of the classical (Fellegi-Sunter) approach to record linkage, and a commonly used method for parameter estimation. In Section 3 we present our extended approach. In Section 4 we describe our modelling approach. In Section 5 we demonstrate the effectiveness of the approach via experiments using real-world data. We summarize our results in Section 6.

2 Classical linkage background

Classical record linkage (Fellegi and Sunter, 1969) is an unsupervised learning approach which only considers comparisons for equality on the key variables. It also adopts the naïve Bayes assumption that the results of comparisons on key variables are mutually independent given the match status.

Assume we have two datasets A and B and seek to identify the records in A and B that correspond to the same population entities.

There is a set of all possible matches,

$$A \times B = \{(a, b); a \in A, b \in B\}.$$

This can be partitioned into sets of matched and unmatched pairs,

$$M = \{(a, b); a = b, a \in A, b \in B\}$$

$$U = \{(a, b); a \neq b, a \in A, b \in B\}$$

and the goal of record linkage is to classify the possible matches as either members of M or members of U.

Bayes theorem leads to the following expression for the posterior odds that a and b correspond to the same population unit,

$$\frac{Pr((a,b) \in M|(a,b))}{Pr((a,b) \in U|(a,b))} = \frac{Pr((a,b)|(a,b) \in M)}{Pr((a,b)|(a,b) \in U)} \frac{Pr((a,b) \in M)}{Pr((a,b) \in U)}. \quad (1)$$

Assume that the data are aligned so that each index $i \in \{1, \ldots, n\}$ corresponds to the same variable in A or B (so $n = |X_{A \cap B}|$). Then, Fellegi-Sunter assumes that the posterior odds can be factorized as,

$$\frac{Pr((a,b) \in M | (a,b))}{Pr((a,b) \in U | (a,b))} = \left(\prod_{i=1}^{n} \frac{Pr((a_i, b_i) | (a,b) \in M)}{Pr((a_i, b_i) | (a,b) \in U)} \right) \frac{Pr((a,b) \in M)}{Pr((a,b) \in U)}. \tag{2}$$

Fellegi and Sunter (1969) initially allocate possible matches to one of three sets,

A_1 – a set of positive links

A_2 – a set of possible links

A_3 – a set of positive non-links

Pairs of records allocated to A_2 are subjected to clerical review and manually allocated to either A_1 or A_3.

Fellegi and Sunter (1969) show that to maintain error rates $Pr(A_1|U) = \mu$ and $Pr(A_3|M) = \lambda$ while minimising the number of record pairs allocated to A_2 it is usually necessary to allocate some possible links probabilistically. Their rule is based on the link probabilities and, as we only have estimates of these, the effectiveness of the rule depends on the quality of estimation.

2.1 Expectation-Maximization

Jaro (1989) details an Expectation-Maximization (EM) approach for estimating the Fellegi-Sunter parameters. Expectation-Maximization (Dempster et al., 1977) is an iterative approach to maximum likelihood estimation. When we have a likelihood (or log likelihood) function that is difficult to maximize we can sometimes produce a function that is much easier to maximize by treating a subset of the parameters as missing data and replacing them with fixed values. The EM approach works by iteratively generating and maximizing such functions. Initial parameter values are specified (for the non-missing parameters) and these are used to calculate the expected values of the missing data (E-step). The expected values are then substituted for the missing data to produce a new function that can be more easily maximized (on the M-step). The parameter estimates generated on the M-step are used on the subsequent E-step. The E-step and M-step are repeated until convergence to some (not necessarily global) maximum of the underlying likelihood function. The use of expected values guarantees that the value of the underlying likelihood function is increased on each iteration (until convergence).

In Jaro (1989) the unobserved match statuses are treated as missing data. This generates functions that can be trivially maximized on the M-steps.

We have $i = 1, \ldots, n$ key variables and $j = 1, \ldots, N$ record pairs. We have a set M of correct matches and a set U of incorrect matches.

$\gamma_i^j = 0$ if attribute i differs for record pair j, and $\gamma_i^j = 1$ if attribute i matches for record pair j.

$$m_i = Pr(\gamma_i^j = 1 | r_j \in M)$$

$$u_i = Pr(\gamma_i^j = 1 | r_j \in U)$$

$$p = \frac{|M|}{|M \cup U|}$$

$$Pr(\gamma^j | M) = \prod_{i=1}^{n} m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j}$$

$$Pr(\gamma^j | U) = \prod_{i=1}^{n} u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j}$$

The last two equations reflect the naïve Bayes assumption.

We want to estimate the unknown parameters $\Phi = (m, u, p)$.

Let x be the complete data vector equal to $\langle \gamma, g \rangle$, where g_j = (1,0) iff the jth record $r_j \in M$ and g_j = (0,1) iff $r_j \in U$. Then the log likelihood for the complete data is,

$$\ln(f(x|\Phi)) = \sum_{j=1}^{N} g_j \cdot \left(\sum_{i=1}^{n} \ln(m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j}), \sum_{i=1}^{n} \ln(u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j}) \right)^T + \sum_{j=1}^{N} g_j \cdot (\ln(p), \ln(1-p))^T .$$

In the expectation step we calculate the expectation of the missing data, g_j given pragmatic starting values for the m_i, u_i and p. The expectations for the g_j are simply the posterior probabilities that the r_j are members of M or U given the current values for m_i, u_i and p.

$$g_m(\gamma^j) = \frac{p \prod_{i=1}^{n} m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j}}{p \prod_{i=1}^{n} m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j} + (1-p) \prod_{i=1}^{n} u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j}}$$

$$g_u(\gamma^j) = \frac{(1-p) \prod_{i=1}^{n} u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j}}{p \prod_{i=1}^{n} m_i^{\gamma_i^j} (1 - m_i)^{1-\gamma_i^j} + (1-p) \prod_{i=1}^{n} u_i^{\gamma_i^j} (1 - u_i)^{1-\gamma_i^j}}$$

So each g_j is equal to $(g_m(\gamma^j), g_u(\gamma^j))$.

In the maximization step we find the maximum likelihood parameters for the m_i, u_i and p given the complete data – the γ_i^j and the g_j.

$$\widehat{m}_i = \frac{\sum_{j=1}^{N} \gamma_i^j \cdot g_m(\gamma^j)}{\sum_{j=1}^{N} g_m(\gamma^j)}$$

$$\widehat{u}_i = \frac{\sum_{j=1}^{N} \gamma_i^j \cdot g_u(\gamma^j)}{\sum_{j=1}^{N} g_u(\gamma^j)}$$

$$\widehat{p} = \frac{\sum_{j=1}^{N} g_m(\gamma^j)}{N}.$$

These estimates are used as fixed parameters for the next expectation step. This generates new $(g_m(\gamma^j), g_u(\gamma^j))$ which are used to generate the subsequent estimates for m_i, u_i and p, and so on, until convergence.

The final solution can be sensitive to the chosen starting values. The u probabilities can be estimated directly from the record pairs if the proportion of correct matches is very low (Jaro, 1989). These estimates can be used as the initial values. The m probabilities are related to data quality, and starting values are often close to 1. In many cases an upper bound for p can be easily derived from the sizes of A and B. If neither A nor B contains duplicates, then the maximum number of matches is equal to the minimum of $|A|$ and $|B|$. In practice the initial value for p would usually be set at a value much lower than this bound (an obvious exception being the case where, say, we knew that every entity in A is also represented in B).

Record linkage performance can often be substantially improved by *blocking*. If there are variables which are considered to be reliably recorded, then we can require that any record pair to be allocated to M must match on these blocking variables. Restricting probabilistic linkage to such pairs can exclude many non-matches, increasing the overall proportion of matches, and improve record linkage performance. Jaro (1989) suggests fixing u at values estimated from all the record pairs when using a blocking strategy in order to reduce bias. In fact any of the parameters can be fixed, and the remainder estimated within the EM framework.

The computational cost of the EM approach can be significantly reduced by constructing a mapping of distinct comparison vectors to frequencies (Jaro, 1989). Under the usual Fellegi-Sunter assumptions each record pair associated with a given comparison vector has the same posterior probability of a correct match. Thus we only need to calculate the expectations of a maximum of 2^n distinct comparison vectors (or 3^n if we have missing values on the key variables) on the E-step. The frequencies of these comparison vectors are used to weight the expectations on the M-step.

3 Extended Fellegi-Sunter record linkage

Consider a partitioning of the set of variables X contained in A or B,

$$X_{A\setminus B} = (X_z)_{z\in A\setminus B}$$

$$X_{B\setminus A} = (X_z)_{z\in B\setminus A}$$

$$X_{A\cap B} = (X_z)_{z\in A\cap B}$$

.

It is the variables in $X_{A\setminus B}$ and $X_{B\setminus A}$ that are ignored in the standard Fellegi-Sunter approach.

The set of matching pairs will follow some (unknown) probability distribution $f(X)$. But for the set of non-matching pairs we have the conditional independence relationship $X_{A\setminus B} \perp\!\!\!\perp X_{B\setminus A}|X_{A\cap B}$. So the joint distribution for the non-matching pairs can be factorized as $f(X_{A\setminus B})f(X_{B\setminus A})/f(X_{B\cap A})$ Thus we have distinct factorizations for the joint distribution over X under M and U, and their ratio provides an additional Bayes factor. This is essentially the same Bayes factor that is used by Smith (2016) and Gutman et al. (2013). The difference here is that we need to handle missing values due to non-matches on key variables. This leads to the latent model given by Equation 3,

$$\begin{aligned}\frac{Pr((a,b)\in M|(a,b))}{Pr((a,b)\in U|(a,b))} =& \frac{f_{A\cup B}(a\oplus b)}{f_{A\setminus B}((a\oplus b)_{A\setminus B})f_{B\setminus A}((a\oplus b)_{B\setminus A})/f_{A\cap B}((a\oplus b)_{A\cap B})}\times \\ & \left(\prod_{i=1}^{n}\frac{Pr((a_i,b_i)|(a,b)\in M)}{Pr((a_i,b_i)|(a,b)\in U)}\right)\times \\ & \frac{Pr((a,b)\in M)}{Pr((a,b)\in U)}\end{aligned} \tag{3}$$

where f_Z denotes the marginal distribution of Z, $a\oplus b$ is the record formed from merging a and b into a single record (with missing values for unequal comparisons on the key variables), and $(a\oplus b)_Z$ is the merged record for the values corresponding to the variables in Z.

The model implies that the probability of observing a given comparison vector is the same for all configurations of X given the match status. We might suspect that certain typographical errors will be more likely for commonly misspelled or longer words, so it is a questionable assumption. Nevertheless, we might still expect some improvement in linkage performance through being able to distinguish between record pairs implying e.g. 6 year old and 50 year old company CEOs.

We can estimate $f(X_{A\setminus B})$, $f(X_{B\setminus A})$ and $f(X_{B\cap A})$ directly from the unlinked data. The issue is how to estimate the marginal distribution $f(X)$. An intriguing possibility is to estimate this from using the results of standard Fellegi-Sunter linkage. After all, a standard use case would be to use the results of standard Fellegi-Sunter linkage to estimate some statistical model. So why not estimate $f(X)$ and use this to generate improved linkage using the extended latent model in 3? If this improves record linkage, then we could generate an improved estimate of $f(X)$ and iterate the process. We might also be able to exploit prior information and training data.

It might seem counterintuitive that we could achieve improvements in performance without either prior information regarding $f(X)$ or training data. But in this case we can exploit the most fundamental of scientific principles - Occam's razor. In the classical record linkage framework there are no steps to ensure that links will be consistent with a parsimonious generating process. Here we can introduce such a step by generating a parsimonious model for $f(X)$ from the linked data and feeding this back into the linkage process via the additional Bayes factor. This general approach of combined linkage and modelling was set out in Smith and Elliot (2016) and the methodology

builds on that in Smith (2016).

3.1 Extended EM

It is clear that all the required marginal distributions for the additional Bayes factor could be generated from the single marginal distribution $f(X)$. But for the moment let us assume that the numerator and denominator terms of the Bayes factor are based on distinct full probability models with parameter vectors Φ_m and Φ_u. Let the configuration of the evidence on the values of the variables for the jth record pair be denoted δ^j. Then,

$$g_m(\gamma^j, \delta^j) = \frac{pPr(\gamma^j|M)Pr(\delta^j|\Phi_m)}{pPr(\gamma^j|M)Pr(\delta^j|\Phi_m) + (1-p)Pr(\gamma^j|U)Pr(\delta^j|\Phi_u)} \tag{4}$$

$$g_u(\gamma^j, \delta^j) = \frac{(1-p)Pr(\gamma^j|U)Pr(\delta^j|\Phi_u)}{pPr(\gamma^j|M)Pr(\delta^j|\Phi_m) + (1-p)Pr(\gamma^j|U)Pr(\delta^j|\Phi_u)}. \tag{5}$$

If we apply the same mathematical treatment as for the standard EM algorithm (Jaro, 1989) we find that the estimation of the Fellegi-Sunter parameters on the M-step is unaffected. The estimated match probabilities generated on the E-step now also depend on the δ^j (via the additional Bayes factor) as shown in Equation 3. Thus for a given set of Bayes factors we can estimate the Fellegi-Sunter parameters via a minimal extension of the usual EM algorithm (Jaro, 1989). We simply need to incorporate the numerator and denominator terms of the additional Bayes factor when calculating expectations, as shown in Equations 4 and 5.

The log likelihood for the complete data is,

$$\begin{aligned}\ln(f(x|\Phi, \Phi_m, \Phi_u)) = &\sum_{j=1}^{N} g_j \cdot \left(\sum_{i=1}^{n} \ln(m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j}), \sum_{i=1}^{n} \ln(u_i^{\gamma_i^j}(1-u_i)^{1-\gamma_i^j}) \right)^T + \\ &\sum_{j=1}^{N} g_j \cdot (\ln(p), \ln(1-p))^T + \\ &\sum_{j=1}^{N} g_j \cdot \left(\ln(Pr(\delta^j|\Phi_m)), \ln(Pr(\delta^j|\Phi_u))\right)^T\end{aligned} \tag{6}$$

This still leaves open the question of the modelling approach. Φ_u can be estimated directly from the unlinked data and fixed. But we still need to deal with the estimation of $f(X)$ from the linked data and match probabilities, and how many times we might iterate the process to generate incremental improvements in linkage performance.

4 Modelling

For our full probability modelling we choose decomposable graphical modelling. A decomposable graphical model consists of a decomposable graph and the parameters for a corresponding collection of probability tables. Parsimonious models are characterised by a sparse graph. We assume that all variables are categorical. In practice, many

of the variables we meet will be categorical, or will have been categorised to limit the risks of statistical disclosure. Variables that are not categorical can be categorised for our purposes. This allows us to exploit the standard results presented in Section 4.2 and identify good models with sparse structure.

4.1 Decomposable graphical models

A decomposable graph is an undirected graph $G = (V, E)$ that contains no unchorded cycles of length greater than three. The node set represents a set of variables $X = (X_v)_{v \in V}$, and the absence of an edge $\{v,w\}$ implies that X_v is conditionally independent of X_w given the variables in $(X_u)_{u \in V \setminus \{v,w\}}$. A decomposable graph can also be represented as a cluster tree. Each maximal pairwise connected subgraph of G is a cluster, and clusters are connected into a tree (or forest in the case of statistically independent components) so as to respect the running intersection property (Lauritzen and Spiegelhalter, 1988):

> *If a node is contained in two clusters, C_1 and C_2, then it is contained in all clusters on the unique path between C_1 and C_2.*

Each edge in the cluster tree is associated with a *sepset* – the intersection of the node sets associated with the clusters that it connects. A cluster tree implies a factorization over the joint distribution of the variables in X,

$$Pr(X) = \frac{\prod_{C \in \mathcal{C}} Pr(C)}{\prod_{S \in \mathcal{S}} Pr(S)}$$

where $\mathcal{C}$ is the set of clusters in the cluster tree (or forest) and $\mathcal{S}$ is the multiset of sepsets.

Posterior beliefs over clusters given observed evidence can be generated via message passing in a cluster tree (Lauritzen and Spiegelhalter, 1988). This exploits conditional independencies and avoids calculating $Pr(X)$. Posterior beliefs over sets of variables not contained in a single cluster can be generated via variable firing (Jensen, 1996) or, at least as efficiently, by manipulating the tree (or forest) so that there exists a cluster containing all the relevant variables (Smith, 2001).

4.2 Model determination

Model determination algorithms for decomposable graphical models generally depend on two important results. The first result is that it is possible to move between any pair of decomposable graphs, G and G', by iteratively adding or removing only a single edge at a time while remaining within the class of decomposable graphs (Frydenberg and Lauritzen, 1989).

The basic rules for edge addition / deletion in decomposable graphs are:

> *An edge $\{v,w\}$ can be added if, and only if, it is not already present, and v and w are either in adjacent clusters or in distinct connected components*

An edge {v,w} can be deleted only if, and only if, it is present in exactly one cluster

The second important result is that the Bayes factor for two neighbouring models (differing in only a single edge) involves only four terms which can be calculated locally (Dawid and Lauritzen, 1993).

Assume the variables in X are categorical, taking values in finite sets $(I_v)_{v\in V}$. Let $I = \underset{v\in V}{\times} I_v$ denote the possible configurations of X. Assume we have a random sample of X contained in a contingency table of counts $n = (n(i))_{i\in I}$. Let n_Z denote the counts $n(i_Z)$ in the marginal table I_Z over the variables in Z. If we also specify a hyper Dirichlet prior as a contingency table of parameters $\lambda = (\lambda(i))_{i\in I}$, then:

For any complete set C the marginal likelihood is,

$$p_C(x_C) = \frac{\Gamma(\lambda)}{\Gamma(\lambda+n)} \prod_{i_C\in I_C} \left(\frac{\Gamma(\lambda_{i_C}+n_{i_C})}{\Gamma(\lambda_{i_C})} \right)$$

and under the hyper multinomial-Dirichlet law (Dawid and Lauritzen, 1993) the marginal likelihood for the full dataset is,

$$p(x) = \frac{\prod_{C\in\mathcal{C}} p_C(x_C)}{\prod_{S\in\mathcal{S}} p_S(x_S)}.$$

If we have graphs $G = (V, E)$ and $G' = (V, E')$ where E' contains the edges in E and an additional edge $\{v, w\}$, then the Bayes factor (ratio of marginal likelihoods) is given by,

$$\frac{p(x|G')}{p(x|G)} = \frac{p_C(x_C)p_S(x_S)}{p_A(x_A)p_B(x_B)}$$

where C is the unique cluster in G' containing $\{v, w\}$ and $A = C \setminus \{v\}$, $B = C \setminus \{w\}$ and $S = C \setminus \{v, w\}$.

These results have been exploited by various model determination algorithms. Markov Chain Monte Carlo (MCMC) algorithms (e.g. Madigan and York, 1995) generate a posterior distribution over the model space. Averaging over this distribution takes into account uncertainty in the model structure and generally provides improved predictive performance (Hoeting et al., 1999). Madigan and Raftery (1994) use an alternative model selection strategy where they reject any models that are sufficiently poorer than the best model(s). Their *Occam's razor* strategy is based on comparisons of models differing by only a single edge. If the evidence favours the larger model to a sufficient degree (decided by a threshold on the Bayes factor), then the smaller model and all its submodels are rejected. A model M_0 is defined as a submodel of M_1 if all the edges in M_0 are also in M_1. Search can start from an arbitrary set of candidate models. Their *up* algorithm considers only edge additions. Their *down* algorithm considers only edge deletions. The up and down algorithms can be run in turn to identify a set of candidate models. Models that are sufficiently poorer than the best candidate(s) are then removed to produce a final set of models that can be used for model averaging purposes.

In common with other approaches we base our full probability modelling on adding and removing single edges

while remaining within the class of decomposable graphs. We also choose to work with categorical variables – categorizing continuous variables as necessary. This allows us to exploit the hyper multinomial-Dirichlet law (Dawid and Lauritzen, 1993) presented earlier.

Here we use a greedy algorithm based on single edge additions / deletions (Smith, 2016). It has some similarities with the Occam's razor strategy. The main differences are that we only seek a single model, and we add / delete an edge that maximizes the increase in marginal likelihood on each iteration of an upwards / downwards search. In order to ensure that our final (and locally optimal) model is parsimonious we start with a full independence model (the model containing no edges) and perform an upwards search followed by a downwards search. This greedy approach is much more computationally efficient than the MCMC and Occam's razor approaches, particularly for larger numbers of variables. For smaller numbers of variables (such as the examples in Madigan and Raftery (1994)) it tends to find the best models found under their Occam's razor approach.

4.3 Missing data

In the current application we have two distinct sources of missing data.

1. Values that are missing in the underlying data

2. Non-matches on key variables

The latter arise under the hypothesis of a match when a pair of records are merged into a single record but do not agree on all the key variables.

We use an EM algorithm due to Fuchs (1982) to generate a maximum likelihood joint distribution over X under the assumption that values are missing at random. We generate an initial estimate for $f(X)$ using only the merged records with no missing values, and using the corresponding match probabilities as weights. We also add some additional weight to each possible configuration so that out initial $f(X)$ contains no zero probabilities.

On the E-step we generate tables of pseudocounts from merged record pairs and match probabilities. There is a distinct table for each 'pattern of missingness'. For example, if all the missing values were due to non-matches on key variables, then we would have one table for each observed comparison vector. We then extend each table to full dimension using the relevant conditional distributions derived from joint distribution. Table totals remain unchanged.

We use tables for computational convenience. The correctness of the E-step is easier to appreciate if we consider individual records. We might have a merged record with a single missing value for gender and with weight w given by the corresponding match probability. The conditional distribution for gender given the observed values for other variables (and the current estimate for $f(X)$) might be (male=0.4, female=0.6). So we generate two copies of the original record and assign gender=male and weight $0.4w$ to one, and gender=female and weight $0.6w$ to the

other. Extending a table to full dimension is equivalent to treating each weighted record in a similar manner, and aggregating the generated records and weights into a table of pseudocounts.

On the M-step we estimate the joint distribution by (elementwise) summation of the full-dimensional tables and normalisation to sum to 1. Details of the algorithm are contained in (Fuchs, 1982).

Once the algorithm has converged we use the last contingency table generated by (elementwise) summation for model determination purposes. The maximum likelihood estimates for the model parameters are generated by marginalisation to the model's clusters and sepsets from the maximum likelihood distribution over X.

4.4 Training data

Larsen and Rubin (2001) present an approach based on the iterative re-estimation via EM after clerical review of uncertain links. This requires that labelled training data can be incorporated on the M-step. After each run of EM some record pairs are reviewed, labelled, removed from the set of pairs to be classified, and added to the labelled training data. Thus their approach can be used to incorporate training data into the estimation of Fellegi-Sunter parameters.

In our approach we simply incorporate training data by including the data as an additional table of counts in the EM algorithm described in Section 4.3.

4.5 Linkage strategies

There are a number of potential strategies for modelling under this general scheme. Thus far we have advocated a 2-model approach, where we have distinct models for the numerator and denominator terms of the additional Bayes factor. The potential issue here is that we can end up with numerator and denominator models that have inconsistent marginal distributions. This is symptomatic of over-fitting the numerator model.

An alternative is to estimate $f(X)$ from the record pairs and both the match probabilities and non-match probabilities. The non-match probabilities would be associated with the marginal distributions $f(X_{A\setminus B})$, $f(X_{B\setminus A})$ and $f(X_{B\cap A})$. The approach for model determination would be unaltered, except for having to deal with additional marginal tables relating to the denominator term. The problem is that the high number of large non-match probabilities places great weight on the conditional independence relationship that characterises the denominator term. We have investigated this approach and found that it does not generally improve linkage performance.

There is also the issue that we have to estimate both structure and probability tables. We can always maximize the (log) likelihood by adopting a full dependence model for both the numerator and denominator. Of course, this would frustrate our approach, which is based on the idea of introducing Occam's razor to the linkage process by requiring consistency with a parsimonious generating process. But for fixed model structure(s) we could still re-estimate parameters for the probability tables within the EM framework.

The result of these considerations is that we estimate model structure outside the EM framework, using the

standard approach for model determination based on marginal likelihood and the results due to Frydenberg and Lauritzen (1989) and Dawid and Lauritzen (1993). In any case, we do not suggest that decomposable graphical modelling is the only means of generating the additional Bayes factor.

The modelling approach is relatively *ad hoc*, and for the reasons given above does not fit easily within the EM framework. In essence we have described a stepwise EM approach where model determination and Bayes factor generation takes place outside the framework, but subsequent esimation of the Fellegi-Sunter parameters is by EM as described in Section 3.1. However, future work will consider alternative methods that could potentially bring model determination wholly within an EM framework.

Of course we also have all the options available under the standard Fellegi-Sunter EM approach. We can fix parameters (such as the u_i), use priors and maximum a posteriori (MAP) estimation, exploit matching constraints, use similarity scores etc.

In the following section we present results for the following pragmatic strategy.

1. Estimate parameters for u_i from the unlinked data
2. Estimate $f(X_{A\setminus B})$, $f(X_{B\setminus A})$ and $f(X_{B\cap A})$ from the unlinked data
3. Estimate parameters for m_i and p using the standard EM algorithm (Jaro, 1989) with fixed u_i
4. Estimate $f(X)$ using the match probabilities and linked data
5. Re-estimate m_i using the minimally extended EM algorithm with fixed u_i, fixed p and fixed Bayes factors

Steps 4 and 5 can be iterated a small number of times.

5 Simulations

5.1 Data

Available generators of test data for record linkage were found to be unsuitable, as it is difficult to generate the dependencies between variables that we find in practice and which we exploit here. For this reason we chose to construct a dataset from two real-world datasets. Data for key variables was sampled from a file available from the North Carolina Voter Registration Database (NCRDB) [1]. Data for the non-key variables was generated from the 1991 Sample of Anonymised Records [2] (SAR). The variables from the SAR were recoded (some categories were merged) to reduce the numbers of variable levels. This ensured that the data could be held in memory and reduced the model determination costs.

Equal sized samples ($n = 2,500$) from the NCRDB and the SAR were merged together to create a single dataset[3].

[1] https://s3.amazonaws.com/dl.ncsbe.gov/data/ncvoter36.zip

[2] The 1991 SARs are provided through the Census Microdata Unit, at the Cathie Marsh Centre for Census and Survey Research (University of Manchester), with the support of the ESRC/JISC/DENI. All tables containing Census data, and the results of analysis, are reproduced with the permission of the Controller of Her Majesty's Stationery Office (Crown Copyright).

[3] We did attempt to use only the NCRDB data, but found few dependencies that could be exploited

This dataset was then partitioned to create population datasets corresponding to the variables chosen for A and B. These were then independently sub-sampled ($n_A = 1,500$, $n_B = 1,000$), and the key variables in B perturbed to ensure that matching record pairs did not necessarily match on all key variables.

The SAR variables were, by construction, independent of the NCRDB variables. So model determination was restricted to the SAR variables, and the NCRDB variables were ignored for the purposes of generating the additional Bayes factors.

The SAR variables were randomly partitioned into sets {SEX, TENURE, ETHGROUP, FAMTYPE, AGE} and {CARS, LTILL, ECONPRIM, QUALEVEL, MSTATUS, SOCLASS} which were allocated to A and B respectively. The key variables selected from the NCRDB and allocated to both A and B were 'first_name' and 'middle_name'. This choice of key variables was designed to produce a 'difficult' linkage problem where we would have some non-matches that would agree on both key variables.

In order to illustrate different aspects of the approach this initial configuration was varied, but only by the addition of extra key variables and / or training data. Training data was a sample ($n = 500$) from the SAR data taken after removal of the initially sampled 2,500 records.

In each case the u_i were estimated from the unlinked data using the approach in Jaro (1989) and fixed. The m_i and p were then estimated via the standard Fellegi-Sunter EM algorithm. We then iterated the extended approach 3 times, retaining the fixed u_i and also fixing p at the value estimated by the standard Fellegi-Sunter run. Fixing these parameters was a pragmatic decision based on the idea of reducing the number of parameters to be estimated, and reducing the risk of over-fitting. The decision was not a response to any specific issue with the linkage results.

Results are presented in the form of Precision Recall curves. As with any binary classification exercise we will have counts of false positives *fp*, false negatives *fn*, true positives *tp* and true negatives *tn*. These will not generally be observed, but we can generate these counts here because we know the true match status for each record pair.

Precision and recall are defined as,

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}.$$

A plot of precision against recall (for all possible thresholds) allows the comparison of classification approaches. Better approaches will tend to produce curves with greater area under the curve (AUC).

Fellegi-Sunter linkage associates a match probability with each observed comparison vector. Thus we have an ordering of equivalence classes, and under the Fellegi-Sunter decision rule we will usually have to allocate the members of one of these equivalence classes to M or U according to a Bernoulli random variable. Rather than perform this random allocation (which would result in one of many possible outcomes) we plot expected precision

against expected recall.

5.2 Initial configuration

Figure 1 demonstrates that we have achieved a general improvement in linkage performance by applying the extended approach. Although there are thresholds for which FS would produce better classification than the extended approach (greater precision for a given recall or greater recall for a given precision), the opposite is the case over large ranges of thresholds. It is notable that under the extended approach the highest ranked equivalence class contains only matches. This is significant if we were to use record linkage to assess the risk of re-identification in a statistical disclosure control exercise.

There is little evidence of improvement in performance through iterating the process. We achieve improvements quickly, mirroring the finding in Scheuren and Winkler (1997). But the most notable fact is that we have achieved this improvement simply by adjusting the m_i so that the resulting match probabilities are more consistent with a parsimonious generating process. The improvement is solely due to the application of Occam's razor, and this was achieved via the integration of linkage and analysis.

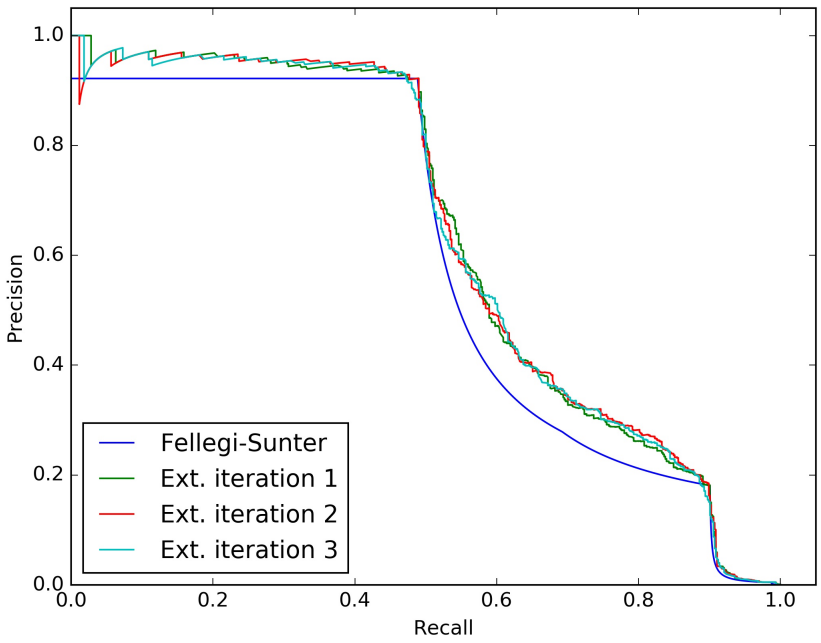

Figure 1: Precision-recall for the initial configuration

Figure 2 shows the decomposable graphical model generated after the initial FS run and used to generate Bayes factors for the 1st iteration. The nodes are coloured to distinguish the variables in A from those in B. It is a very sparse model, with only a single edge between variables only in A and only in B. Yet it has been enough to generate a significant overall improvement in linkage performance.

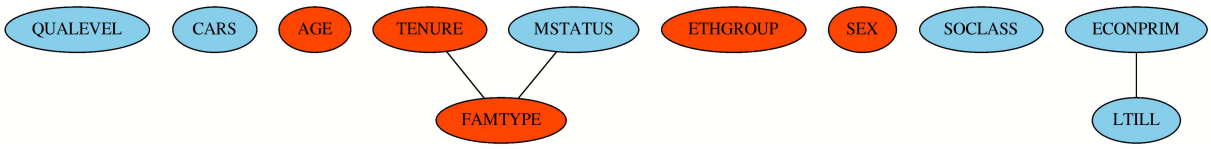

Figure 2: Model for 1st iteration of extended approach

We should note that without any such edges we would have conditional independence between the variables only in A and those only in B given the key variables. This is the relationship we assume under the hypothesis of a non-match and with consistent numerator and denominator distributions the additional Bayes factors would all equal 1. Given the 2-model approach presented here we would generally have some inconsistency between numerator and denominator models, and therefore some small difference in linkage outputs. But this illustrates an important point.

> *We can only expect performance to be improved when we have sufficient statistical power to generate a numerator model that contains an edge between at least one non-key variable in A and at least one non-key variable in B.*

Of course, when we have training data we can boost the power considerably. The impacts of adding training data are shown in Figures 3 and 4. Linkage performance is further improved, and we have some evidence that iterating the process has had a positive impact. Encouragingly there are commonalities in structure with the model in Figure 2. Structurally the models are very similar, and we still only have a single edge between the variables only in A and those only in B. The main impact of the training data appears to have been the generation of better parameter estimates for the probability tables.

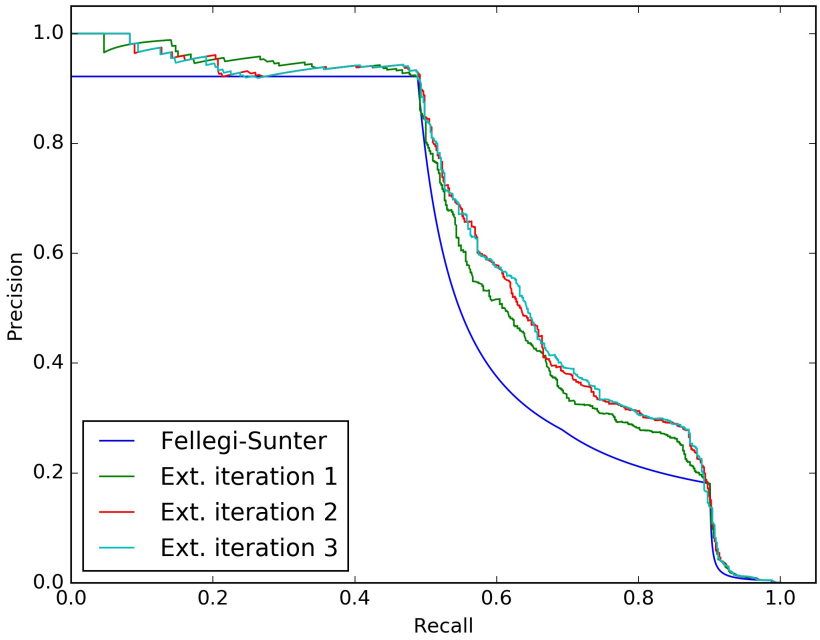

Figure 3: Precision-recall for the initial configuration with training data

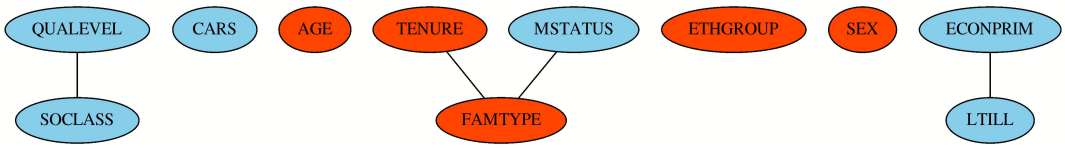

Figure 4: Model for 1st iteration of extended approach with training data

5.3 Initial configuration with additional key variables

The addition of the extra key variable 'res_city_desc' (from NCRDB) produced similar improvements to those previously presented. Improvements were consistent, and largely achieved after only a single iteration. We then also introduced CARS (from the 1991 SAR) as a key variable (adding it to dataset A). So we now had a key variable that was not (by construction) independent of all the non-key variables. The resulting precision recall curve is shown in Figure 5, with the corresponding model shown in Figure 6 (with the modelled key variable coloured white).

We see a small increase in linkage performance, with some evidence that iterating the process helps. We also see that the model shares much structure with the previous models, and that we are still relying on the single edge between MSTATUS and FAMTYPE to generate meaningful additional Bayes factors. The addition of the extra key variables has improved the initial Fellegi-Sunter linkage, perhaps leaving us with less scope for improvement. However, Figures 7 and 8 show that further improvements are possible via improved model determination through the addition of training data. We also see that increased statistical power, through the additional key variables and training data, has generated a more densely connected, and (potentially) useful model.

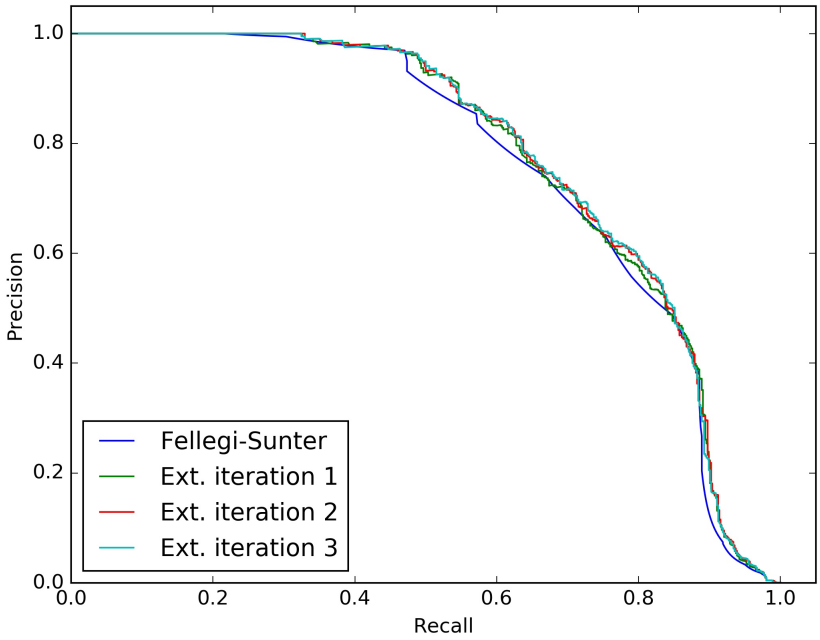

Figure 5: Precision-recall for the initial configuration plus additional key variables 'res_city_desc' and CARS

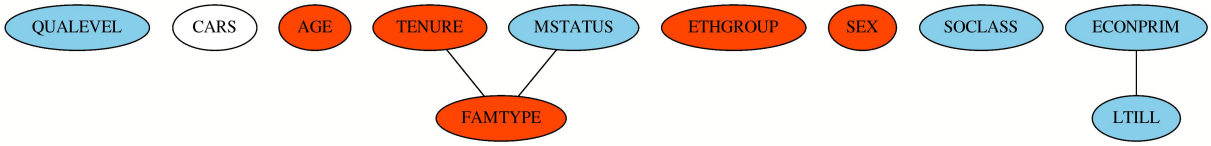

Figure 6: Model for 1st iteration of the extended approach with additional key variables 'res_city_desc' and CARS

Although these results seem to demonstrate a consistent improvement, this cannot be guaranteed. The extended approach works by modifying the underlying likelihood function (via the additional Bayes factors). The initial solution that was locally optimal for the original likelihood function will not generally be locally optimal for the modified likelihood function. We hope that the initial solution will be close to a local maximum of the modified likelihood function that is consistent with our parsimonious model, and that the subsequent EM run will find that maximum. If initial linkage is poor then the model will be poor and subsequent linkage will be more consistent with a parsimonious (but poor) model.

> *We can only expect performance to be improved if the initial Fellegi-Sunter run produces reasonable parameter estimates.*

We can increase the chances of finding good parameter estimates using relatively standard approaches, such as those based on the use of similarity scores or matching constraints. We can also check that the EM converges quickly and smoothly, and that parameter estimates are plausible. Of course, we can also check that the model used to generate the Bayes factors is plausible. It is also notable that when the extended EM approach performs

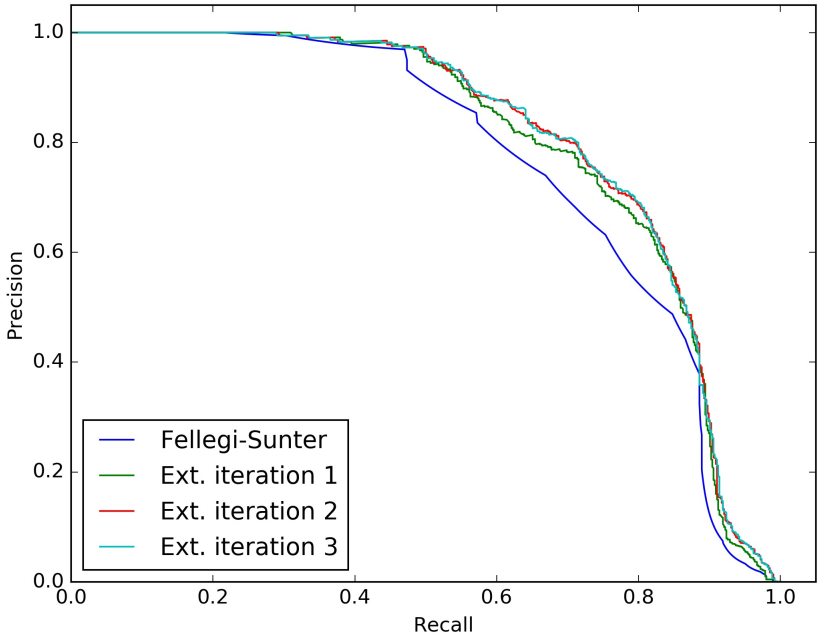

Figure 7: Precision-recall for the initial configuration plus additional key variables 'res_city_desc' and CARS and with training data

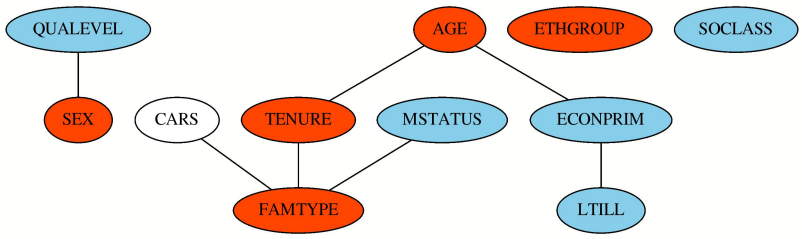

Figure 8: Model for 1st iteration of the extended approach with additional key variables 'res_city_desc' and CARS and with training data

well it does not generally change the Fellegi-Sunter parameter estimates a great deal (they remain plausible).

For the simulations presented here we were sufficiently lucky to achieve reasonable linkage on the initial Fellegi-Sunter run and see improvements in linkage via the extended EM approach. But generally the extended EM approach must be applied with care, with due attention paid to available diagnostics. Of course this is generally true of all linkage approaches, including those that seek to iteratively improve linkage such as the methods of Larsen and Rubin (2001) and Scheuren and Winkler (1997).

5.4 Model determination performance

The question that we have not so far addressed is whether analysis can also be improved by combining linkage with analysis. As we have the population data we can compare the probability distributions over X implied by our models with the probability distribution over X resulting from the normalisation (to sum to 1) of our population. The results corresponding to Figures 1, 3, 5 and 7 are shown in Figure 9.

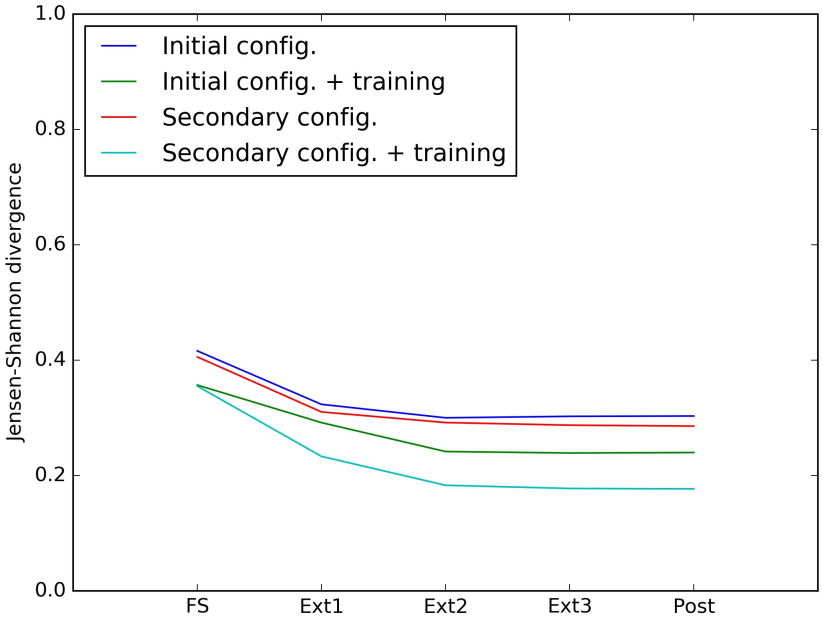

Figure 9: Jensen-Shannon divergences for models under various strategies

FS corresponds to the distribution generated by applying Fuch's algorithm to the initial linkage (and training data if available) without any decomposable graphical model determination. Ext1, Ext2 and Ext3 correspond to the decomposable graphical models generated prior to the runs of the extended EM linkage algorithm. 'Post' is the result of performing an additional model determination exercise with the results of the 3rd extended EM run.

We can see that improved linkage is accompanied by improved models. But for each combination of key variables and training data there is little improvement over the model generated prior to the 2nd iteration of the extended

model. This is consistent with the evidence of improved linkage from the precision-recall plots.

6 Conclusions

We have presented an extension of the usual Fellegi-Sunter record linkage approach that can result in improved linkage and model determination performance. It is based on the basic scientific principle that we should prefer sets of links that are consistent with a parsimonious generating process. We apply Occam's razor by integrating record linkage with full probability modelling. A variety of approaches could be used for modelling; we chose decomposable graphical models. The only requirement is that the model can be used to generate our additional Bayes factors. Accommodating these additional Bayes factors requires only a minimal extension to the standard EM algorithm for estimating Fellegi-Sunter parameters.

Improvement in linkage performance is by no means guaranteed. Firstly, we require sufficient statistical power to generate a model that produces distinguishing Bayes factors. Secondly, we require that the parameters estimated via an initial Fellegi-Sunter run (or previous iteration of the extended approach) are plausible. If not, we may move to a solution that is consistent with a poor model. It is important to check diagnostics, just as it is for standard Fellegi-Sunter linkage.

As the methodology closely mirrors Fellegi-Sunter linkage most of the existing techniques for improving FS linkage can still be applied. We have already highlighted the possibility for exploiting similarity scores. The extended approach also constitutes a convenient means of introducing prior information and / or training data.

Essentially we have presented a proof of concept along the same lines as in Scheuren and Winkler (1997). But we have not yet fully investigated the potential of the approach. It is possible to specify priors on the Fellegi-Sunter parameters and use MAP estimation rather than maximum likelihood. Again, this requires only a minor extension to the standard EM algorithm. It also provides a mechanism by which we could penalize model complexity (rather than relying on marginal likelihood). This could be used to bring the model determination within the EM framework. Although a naive implementation would be computationally costly, we could potentially bring the costs down by adopting a generalised EM approach. That is, we could seek an increase in the likelihood on each iteration, rather than requiring maximization. This may go some way towards limiting the chances of poor models / linkage when iterating the modelling. Future work will address this possibility.

References

P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer, 2012.

A. Dawid and S. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317, 1993.

A. P. Dempster, N. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

I. Fellegi and A. Sunter. A theory for record linkage. *JASA*, 64(238):1183–1210, 1969.

M. Frydenberg and S. Lauritzen. Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika*, 76(3):539–555, 1989.

C. Fuchs. Maximum likelihood estimation and model selection in contingency tables with missing data. *JASA*, 77 (378):270–278, 1982.

R. Gutman, C. Afendulis, and A. Zaslavsky. A Bayesian procedure for file linking to analyze end-of-life medical costs. *JASA*, 108(501):34–47, 2013. doi: 10.1080/01621459.2012.726889. URL http://dx.doi.org/10.1080/01621459.2012.726889. PMID: 23645944.

J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14 (4):382–417, 1999.

M. H. Hof, A. C. Ravelli, and A. H. Zwinderman. A probabilistic record linkage model for survival data. *JASA*, 0(ja): 0–0, 2017. doi: 10.1080/01621459.2017.1311262. URL http://dx.doi.org/10.1080/01621459.2017.1311262.

M. A. Jaro. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *JASA*, 84 (406):414–420, 1989.

F. V. Jensen. *An Introduction to Bayesian Networks*. Springer-Verlag, New York, 1996.

H. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, (2):83–97, 1955.

M. Larsen and D. Rubin. Iterative automated record linkage using mixture models. *JASA*, 96(453):32–41, 2001.

S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *JRSS series B*, 50(2):157–224, 1988.

D. Madigan and A. Raftery. Model selection and accounting for model uncertainty in graphical models using Occam's window. *JASA*, 89(428):1535–1546, 1994.

D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, (63):215–232, 1995.

M. Sadinle. Bayesian estimation of bipartite matchings for record linkage. *JASA*, 112(518):600–612, 2017. doi: 10.1080/01621459.2016.1148612. URL http://dx.doi.org/10.1080/01621459.2016.1148612.

F. Scheuren and W. Winkler. Regression analysis of computer files that are computer matched. *Survey Methodology*, 19(1):39–58, 1993.

F. Scheuren and W. Winkler. Regression analysis of computer files that are computer matched - Part II. *Survey Methodology*, 23(2):157–165, 1997.

D. Smith. The efficient propagation of arbitrary subsets of beliefs in discrete-valued bayesian belief networks. In T. Jaakkola and T. Richardson, editors, *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, pages 292–297, 2001.

D. Smith. Re-identification in the absence of common matching variables. Technical report, University of Manchester, 2016. URL http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2016/2016-02.pdf.

D. Smith and M. Elliot. Towards a general record linkage framework for statistical disclosure control. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*, PrAISe '16, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4304-6. doi: 10.1145/2970030.2970037. URL http://doi.acm.org/10.1145/2970030.2970037.

D. Smith and N. Shlomo. Report for the data without boundaries project. Technical report, University of Manchester, 2014. URL http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/reports/2014-01-Data_without_Boundaries_Report.pdf.

R. C. Steorts, R. Hall, and S. E. Fienberg. A Bayesian approach to graphical record linkage and deduplication. *JASA*, 111(516):1660–1672, 2016. doi: 10.1080/01621459.2015.1105807. URL http://dx.doi.org/10.1080/01621459.2015.1105807.

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

W. E. Winkler. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, pages 354–359, 1990.