Bayesian Estimation of Bipartite Matchings for Record Linkage[1]

Mauricio Sadinle
Duke University, Durham, North Carolina, USA - msadinle@stat.duke.edu

Abstract

In this article we are concerned with the most traditional scenario of record linkage, which consists of linking two disparate datafiles containing overlapping information on a set of entities, and it is assumed that each entity is recorded maximum once in each datafile. This is an important task with a wide variety of applications, given that it needs to be solved whenever we have to combine information from different sources. Most statistical techniques currently in use are derived from a seminal paper by Fellegi and Sunter (1969) who formalized procedures that had been used earlier by other researchers. These techniques usually assume independence in the matching status of record pairs to derive estimation procedures and optimal point estimators (e.g. Fellegi-Sunter decision rule). We argue that this independence assumption is unreasonable and target instead a bipartite matching between the two sets of records coming from the two files as our parameter of interest. The Bayesian implementation presented here allows us to incorporate prior information on the quality of the fields in the datafiles, which in turn helps to obtain better results when the datafiles do not share a large amount of identifying information. We demonstrate the improvements of our approach over traditional methodologies in a number of realistic simulation studies.

Keywords: bipartite matching; data integration; data matching; Fellegi-Sunter.

1 Introduction

Linking different datafiles that contain information on the same population is important in a wide variety of applications, including merging post-enumeration surveys and census data for census coverage evaluation (e.g., Winkler, 1988; Jaro, 1989), and linking health-care databases for epidemiological studies (e.g., Bell et al., 1994). This task is not trivial when unique identifiers are not available, and it is especially difficult when the records are subject to errors and missing values. In this paper we focus on bipartite record linkage, which consists of linking two datafiles containing overlapping information on a set of individuals or entities, and it is usually assumed that each entity is recorded maximum once in each datafile. Most of the statistical literature on record linkage deal with this scenario (Fellegi and Sunter, 1969; Winkler, 1988; Jaro, 1989; Winkler, 1993, 1994; Larsen and Rubin, 2001; Herzog et al., 2007), and most of the statistical techniques currently in use are derived from a seminal paper by Fellegi and Sunter (1969) who formalized procedures that had been used earlier by other researchers (e.g. Newcombe et al., 1959; Newcombe and Kennedy, 1962, and references therein). These statistical techniques have a number of drawbacks that we discuss and address in this paper. In the next section we review the Fellegi-Sunter methodology for linking two files, its modern implementation using mixture models, and we describe the limitations of this approach. Later on we present a Bayesian approach that solves some of the limitations of the mixture model approach to record linkage, and we compare these methodologies in an extensive simulation study.

2 The Fellegi-Sunter Approach to Record Linkage and its Mixture Model Implementation

Fellegi and Sunter (1969) considered two datafiles $\mathbf{X}_1$ and $\mathbf{X}_2$ that record information on overlapping sets of individuals. These files originate from two record-generating processes that may induce errors and missing values. The record linkage task is to determine which record pairs in $\mathbf{X}_1 \times \mathbf{X}_2$ refer to the same entities. We denote $\Delta_{ij} = 1$ if record pair (i, j) refers to the same entity (pair (i, j) is a match), $\Delta_{ij} = 0$ otherwise. The input of the method are pairwise comparison data, where for each pair of records (i, j) we construct vectors $\boldsymbol{\gamma}_{ij}$ containing comparisons of their information.

[1]Supported by NSF grants SES-11-30706 to Carnegie Mellon University and SES-11-31897 to Duke University.

Comparison Data

We compare pairs of records with the goal of finding evidence of whether two records refer to the same entity. Intuitively, two records referring to the same entity should be very similar. The way of constructing the comparisons depends on the information contained by the records. Following Winkler (1990) we can take into account partial agreements as follows. We compare the field f of records i and j by computing some similarity measure $\mathcal{S}_f(i,j)$ (e.g. the Levenshtein distance to compare names). The range of this similarity measure is then divided into L_f+1 intervals $I_{f0}, I_{f1}, \ldots, I_{fL_f}$, that represent different levels of disagreement. By convention, the interval I_{f0} represents the highest level of agreement, which includes no disagreement, and the last interval, I_{fL_f}, represents the highest level of disagreement. We can then build ordinal variables from these intervals. For records i and j, and field f, we define $\gamma_{ij}^f = l$, if $\mathcal{S}_f(i,j) \in I_{fl}$. The larger the value of γ_{ij}^f, the larger the disagreement between records i and j with respect to field f. These different field comparisons are collected in a vector for each record pair. $\boldsymbol{\gamma}_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^f, \ldots, \gamma_{ij}^F)$ denotes the comparison vector for records i and j, where F is the number of fields being compared.

Although in principle the number of record comparisons grows quadratically with the file sizes, notice that the number of matches can only grow linearly, and so in practice most record pairs are non-matches that can be trivially detected using some simple criteria, thereby avoiding the computation of the complete set of comparisons. One approach to reduce the number of comparisons is called *blocking*, which consists on declaring record pairs as non-matches whenever they disagree in a reliable categorical field. A more robust although more computationally expensive approach is to declare a priori record pairs as non-matches when they have strong disagreements in a combination of fields. The advantage of this filtering step is that it reduces the set of pairs that we need to resolve to a small number to which we can apply a more sophisticated procedure.

Mixture Model

In this approach we assume that the comparison vectors $\{\boldsymbol{\gamma}_{ij}\}$ are realizations of random vectors $\{\boldsymbol{\Gamma}_{ij}\}$. Since we expect matches to largely agree in the information that they contain, we assume that the distribution of $\boldsymbol{\Gamma}_{ij}$ is the same for all record pairs that refer to the *same* entity (regardless the entity), and that the distribution of $\boldsymbol{\Gamma}_{ij}$ is the same for all record pairs that refer to *different* entities (regardless the pair of entities). This intuitive description can be formalized into a model for the comparison data as

$$\boldsymbol{\Gamma}_{ij}|\Delta_{ij} = 1 \overset{iid}{\sim} G_1, \qquad \boldsymbol{\Gamma}_{ij}|\Delta_{ij} = 0 \overset{iid}{\sim} G_0, \tag{1}$$

where G_1 and G_0 represent the models of the comparison data for matches and non-matches, respectively. These models may change depending on the comparison data at hand. The key component of the mixture model implementation is that in addition to the model of Equation (1), the Δ_{ij}'s are modeled as i.i.d. Bernoulli(p). The Δ_{ij}'s, along with the parameters of the complete model are usually estimated using the EM algorithm. The idea of finding matched pairs by modeling comparison data as two disparate distributions for matches and non-matches goes back at least to the work of Newcombe et al. (1959); Newcombe and Kennedy (1962); Fellegi and Sunter (1969); Du Bois (1969), and some examples of its modern implementations using the EM algorithm are due to Winkler (1988); Jaro (1989); Larsen and Rubin (2001).

Despite the popularity of this methodology it has certain caveats. In terms of modeling the comparison data as a mixture, there is an implicit "hope" that the clusters that we obtain are associated with matches and non-matches. In practice, however, the mixture components may not correspond with these groups of record pairs. In particular, the mixture model will identify two clusters regardless of whether the two files have an actual overlap. Winkler (2002) mentions conditions for the unsupervised mixture model to give good results based on experience working with large administrative files at the US Census Bureau: the proportion of matches should be greater than 5%, the classes of matches and non-matches should be well separated, typographical error must be relatively low, and there must be redundant fields that overcome errors in other fields. In many practical situations these conditions may not hold.

Furthermore, even if the mixture model is relatively successful at separating matches from non-matches, given that this approach outputs independent decisions on the matching status of pairs of records it may violate coherent constraints in the problem. For example, each record in the first file may be linked to maximum one other record in the second file, and viceversa, as a consequence of the assumption of no-duplicates within file.

When solving the record linkage problem using this approach there is nothing that enforces this maximum-one-to-one requirement in the model itself, and therefore some post processing steps are required, such as in Jaro (1989) who proposed to solve an assignment problem to enforce the maximum one-to-one constraint. We believe that a more sensible approach is to incorporate this constraint into the model, as done for example by Larsen (2005), rather than forcing it in a post-processing step.
In the next section we show an adaptation of the work of Sadinle (2014) to the problem of bipartite record linkage, and later we explore its performance compared with the mixture model approach.

3 A Bayesian Approach to Bipartite Record Linkage

We saw that treating the decisions on the matching statuses of record pairs as independent of one another can lead to inconsistencies in bipartite record linkage. In the closely related context of duplicate detection, Sadinle (2014) proposed a methodology that avoids these inconsistencies by targeting the coreference partition of the datafile as the parameter of interest, that is, the partition that groups records according to the entities that they refer to. The coreference partition can be represented by a matrix $\mathbf{\Delta}$ such that its $(i,j)th$ entry $\Delta_{ij} = 1$ if records i and j refer to the same entity, $\Delta_{ij} = 0$ otherwise. We can adapt that methodology to deal with bipartite record linkage by creating a concatenated datafile $\mathbf{X}_1 \cup \mathbf{X}_2$ and solving the problem of duplicate detection for this datafile under the constraints that $\Delta_{ij} = 0$ if $i, j \in \mathbf{X}_k$, $i \neq j$, $k = 1, 2$. These constraints come from the no-duplicates within file assumption. Given this restriction, $\mathbf{\Delta}$ represents a matching matrix which is nothing but a matrix representation of a bipartite matching between the two files.
The model used by Sadinle (2014) uses the structure presented in Equation (1), except that the matching statuses $\{\Delta_{ij}\}$ are regarded as entries of a matrix that represents a partition, or in the current case, a bipartite matching. $\mathcal{D}$ denotes the set of possible bipartite matchings, which may have been reduced after blocking or other non-match pair filtering. We use a uniform prior on the space $\mathcal{D}$. Regarding the model for the comparison data, we use a simple parametrization for G_1 and G_0. Our model assumes that the comparison fields are independent for both matches and non-matches. If comparison Γ_{ij}^f takes $L_f + 1$ values corresponding to levels of disagreement, its distribution among matched records can be modeled according to a multinomial distribution, and we choose to parameterize it in terms of sequential conditional probabilities m_{fl}, $m_{f0} = \mathbb{P}(\Gamma_{ij}^f = 0 | \Delta_{ij} = 1)$ and $m_{fl} = \mathbb{P}(\Gamma_{ij}^f = l | \Gamma_{ij}^f > l - 1, \Delta_{ij} = 1)$ for $0 < l < L_f$. This parameterization for G_1 facilitates prior specification. Analogously, the distribution of Γ_{ij}^f among non-matched pairs uses parameters u_{fl} with $u_{f0} = \mathbb{P}(\Gamma_{ij}^f = 0 | \Delta_{ij} = 0)$, $u_{fl} = \mathbb{P}(\Gamma_{ij}^f = l | \Gamma_{ij}^f > l - 1, \Delta_{ij} = 0)$ for $0 < l < L_f$. Notice that if $L_f = 1$, that is, if comparison f is binary, we obtain the traditional model used in record linkage for binary comparisons (e.g., Winkler, 1988; Jaro, 1989).
We now explain our selection of the priors for m_{fl} and u_{fl}, $l = 0, \ldots, L_f - 1$. $m_{f0} = \mathbb{P}(\Gamma_{ij}^f = 0 | \Delta_{ij} = 1)$ represents the probability of observing the level zero of disagreement in the comparison f among matches. This level represents no disagreement, or a high degree of agreement, so if we believe that field f contains no error, m_{f0} should be, a priori, a point mass at one, but as the error in field f increases, the mass of m_{f0}'s prior should move away from one. We therefore take a priori m_{f0} to be distributed uniformly in some interval $[\lambda_{f0}, 1]$ for some $0 < \lambda_{f0} < 1$. If we believe that the field used to compute comparison f is fairly accurate, then we should set the threshold λ_{f0} to be close to one. A similar reasoning as for m_{f0} leads us to take the prior of m_{fl} as uniform in some interval $[\lambda_{fl}, 1]$ (see Sadinle, 2014, 2015, for further details). Also, we take $u_{fl} \sim \text{Uniform}(0, 1)$. From this formulation we obtain a posterior distribution on bipartite matchings, which can be approximated using Markov chain Monte Carlo (MCMC). For further details on the estimation procedure we refer the reader to Sadinle (2015).
To determine a point estimate from the posterior distribution on bipartite matchings we use the procedure proposed by Jaro (1989), which under our model and Bayesian approach gives us the mode of the posterior distribution on bipartite matchings conditioning on values of the m_{fl} and u_{fl} parameters (see Sadinle, 2015, for a proof). To use this procedure we condition on the posterior means of the m_{fl} and u_{fl} parameters as obtained from a posterior MCMC sample.
We now compare this methodology with the traditional mixture model approach in a realistic simulation study.

4 A Simulation Study

We now present a simulation study to compare the performance of the Bayesian approach and the mixture model approach in the bipartite record linkage context. We generated pairs of datafiles using a synthetic data generator developed by Christen and Vatsalan (2013). Each datafile has 250 records and five fields: given and family names, age, gender and occupation. To create the datafiles we randomly generated a population of 500 individuals from which we selected 250 to be in the first datafile. The second datafile was created from n_{12} individuals that were included in the first datafile, and $250 - n_{12}$ individuals not included in the first datafile. The records in the datafiles were generated and corrupted as described in Sadinle (2015). We refer to the n_{12} individuals included in both files as their *overlap*.

We are interested in comparing the performance of these record linkage methodologies across different scenarios of files' overlap and measurement error. We generated 100 pairs of datafiles for each combination of 100%, 50%, and 10% files' overlap, and 1, 3, and 5 erroneous fields per record. For each pair of files we computed comparison data as in Sadinle (2015), and the pairs that had the level three of disagreement in given and family name were declared as non-matches a priori. Similarly, age, gender or occupation were used for blocking. The comparison data used to train the record linkage models correspond to the comparisons of given and family names. Notice that this is a challenging linkage scenario given that decisions have to be made based on only a small amount of information.

We implemented the mixture model approach using the model for the comparison data presented in the previous section and the EM algorithm. For the Bayesian approach, we used two different sets of priors corresponding to prior truncation points for the m_{fl} parameters being 0.5 and 0.95. For simplicity, each set of priors has the same prior truncation point for all the m_{fl} parameters. These priors correspond to one scenario where we believe that the amount of error in each field can be anywhere between 0% and 50%, and one where we believe that it may only go up to 5%. In this section we refer to these priors as *weak* and *strong*, respectively. For each pair of datasets, and for each set of priors, we ran 10,000 iterations of the MCMC presented in Sadinle (2015), and discarded the first 1,000 as burn-in. The average runtime using an implementation in R with parts written in C language was of 54, 49 and 44 seconds for files with overlap 100%, 50%, and 10%, respectively, including the computation of the comparison data, on a laptop with a 2.80 GHz processor.

For each pair of files and for each approach we computed point estimates of the bipartite matching using the maximum one-to-one assignment proposed by Jaro (1989). For each estimated bipartite matching we computed the measures of precision and recall. To summarize the performance of the methods under each scenario of overlap and measurement error we computed the median, the first and 99th percentiles of these measures across the 100 pairs of datafiles.

In Figure 1 we present the results of the simulation study. The first row shows the results of the mixture model approach, and the second and third rows show the results of the Bayesian approach (BPA, after *Bayesian partitioning approach*) under the weak and strong priors, respectively. The columns of Figure 1 show the results for different amounts of overlap between the files, and in each panel black lines refer to recall, gray lines to precision, solid lines show medians, and dashed lines show first and 99th percentiles.

We can see from the first row of Figure 1 that the mixture model approach works well when the files have a large overlap and a small number of errors, but its performance deteriorates rapidly when the number of errors increase or when the overlap of the files decrease. These findings agree with the observations made by Winkler (2002) in the sense that the mixture model approach leads to poor results when there is a large amount of error, when the overlap of the files is small, and when the files do not contain a lot of identifying information. Under these scenarios the mixture model approach is not able to identify the clusters of record pairs associated with matches and non-matches, and instead outputs two clusters revealing other structures in the comparison data.

In record linkage we have an idea of how the two types of records that we care about should look like under normal circumstances: matching pairs should be somewhat similar, and non-matching pairs should be somewhat dissimilar. The Bayesian approach allows us to incorporate this information in the form of priors and therefore further constrain the set of probable bipartite matchings. We can see from the second row of Figure 1 that the Bayesian partitioning approach under weak priors has a better performance in this simulation study than the mixture model approach. By constraining the model parameters m_{fl} to be over 0.5 we simply specify our prior belief that the set of matching pairs should be composed of record pairs agreeing

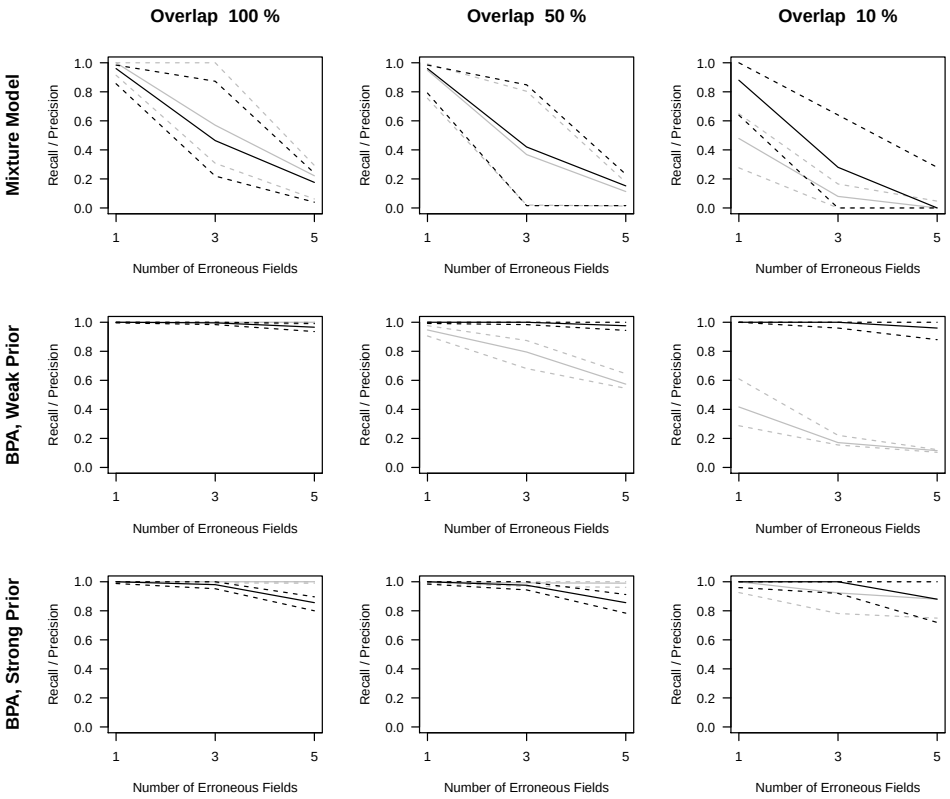

Figure 1: Comparing the performance of two methodologies for record linkage in the simulation study of Section 4. The first row shows the results of the mixture model approach with the Jaro constraint. The second and third rows show the results of the conditional posterior mode estimator of the bipartite matching using the Bayesian partitioning approach (BPA) with a weak and a strong prior, corresponding to prior truncation points $\lambda_{fl} = 0.5$ and $\lambda_{fl} = 0.95$. Black lines refer to recall, gray lines to precision, solid lines show medians, and dashed lines show first and 99th percentiles.

in most of their information, and this allows the method to identify the right bipartite matching in a large number of scenarios. Under this prior the Bayesian approach is usually able to identify the existing matches, which can be seen from the recall (black) lines in the second row of Figure 1 being very close to one. The precision of this approach, however, deteriorates when the number of errors increase and when the overlap of the files decrease, which means that under those circumstances this approach is prone to identify a large proportion of false-matches.

Finally, from the third row of Figure 1 we can see that if we use a strong prior on the expected amount of error in the files, the precision of the Bayesian approach stays close to one across all scenarios, but the recall falls when the amount of error is too high. These results indicate that under this strong prior the Bayesian approach may miss some true matches although it ensures that those pairs declared as matches are very likely to be actual matches. Although the performance under this strong prior is the best in this simulation study, in other scenarios it may lead to miss many more true matches.

5 Conclusions and Future Work

The mixture model approach to record linkage along with Jaro's constraint works well when there is not a lot of error in the files and their overlap is large. This approach is also appealing given that it is fast. In our R implementation of the mixture model approach the run time was less than one second for each pair of datafiles, whereas the Bayesian approach took between 40 seconds and one minute per pair of files.

Although the mixture model approach is fast and it gives good results in some linkage scenarios, in more challenging ones it can be outperformed by a Bayesian approach that provides more guidance on the desired bipartite matching. In addition, having a posterior distribution on the bipartite matchings allows us to use

different point estimators, including those that allow a rejection option (see Sadinle, 2015). Decision rules with rejection option allows us to leave unresolved parts of the bipartite matching that are very uncertain. Fellegi and Sunter (1969) proposed a decision rule designed for this purpose but its optimality relies on the assumption that the linkage decision for a record pair is determined only by its comparison vector, which does not hold in the traditional record linkage scenario. Furthermore, given that the mixture model assumes that the matching statuses of the record pairs are independent of one another, the estimated matching probabilities $\hat{P}(\Delta_{ij} = 1|\boldsymbol{\gamma}_{ij})$ obtained from the EM algorithm lead to conflicting decisions if used in general decision rules for bipartite matching.

There are a number of areas that can be further explored. Improvements can be made to the Bayesian approach, in particular by developing informative priors for the bipartite matching since they can be useful in cases where we have prior knowledge on the overlap of the files. Furthermore, in this document we focused on unsupervised record linkage, but further comparisons with supervised and semi-supervised approaches are also important. In particular, in a context where we are willing to leave some decisions to be resolved by clerical review, it may pay off to reallocate some of those resources to build hand-matched training data to be used in semi-supervised training of record linkage models, thereby improving the models' fit, and perhaps leading to an overall smaller number of pairs that have to be matched by hand (see Larsen and Rubin, 2001, for related ideas).

References

Bell, R. M., Keesey, J., and Richards, T. (1994). The Urge to Merge: Linking Vital Statistics Records and Medicaid Claims. *Medical Care*, 32(10):1004–1018.

Christen, P. and Vatsalan, D. (2013). Flexible and Extensible Generation and Corruption of Personal Data. In *Proc. of the ACM International Conference on Information and Knowledge Management (CIKM 2013)*.

Du Bois, Jr., N. S. D. (1969). A Solution to the Problem of Linking Multivariate Documents. *JASA*, 64(325):163–174.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *JASA*, 64(328):1183–1210.

Herzog, T. N., Scheuren, F. J., and Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer, New York.

Jaro, M. A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *JASA*, 84(406):414–420.

Larsen, M. D. (2005). Advances in Record Linkage Theory: Hierarchical Bayesian Record Linkage Theory. In *Proc. Sec. on Survey Research Methods*, pages 3277–3284. ASA.

Larsen, M. D. and Rubin, D. B. (2001). Iterative Automated Record Linkage Using Mixture Models. *JASA*, 96(453):32–41.

Newcombe, H. B. and Kennedy, J. M. (1962). Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information. *Communications of the ACM*, 5(11):563–566.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. (1959). Automatic Linkage of Vital Records. *Science*, 130(3381):954–959.

Sadinle, M. (2014). Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach. *Annals of Applied Statistics*, 8(4):2404–2434.

Sadinle, M. (2015). *A Bayesian Partitioning Approach to Duplicate Detection and Record Linkage*. PhD thesis, Carnegie Mellon University.

Winkler, W. E. (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. In *Proc. Sec. on Survey Research Methods*, pages 667–671. ASA.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proc. Sec. on Survey Research Methods*, pages 354–359. ASA.

Winkler, W. E. (1993). Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of Survey Research Methods Section*, pages 274–279. ASA.

Winkler, W. E. (1994). Advanced Methods for Record Linkage. In *Proc. Sec. on Survey Research Methods*, pages 467–472. ASA.

Winkler, W. E. (2002). Methods for Record Linkage and Bayesian Networks. In *Proc. Sec. on Survey Research Methods*, pages 3743–3748. ASA.