Approaches to Multiple Record Linkage

Mauricio Sadinle, Rob Hall, and Stephen E. Fienberg
Department of Statistics, Heinz College, and Department of Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213-3890, U.S.A.
E-mail: msadinle@stat.cmu.edu; rjhall@cs.cmu.edu; fienberg@stat.cmu.edu

We review the theory and techniques of record linkage that date back to pioneering work by Fellegi and Sunter on matching records in two lists. When the task involves linking $K > 2$ lists, the most common approach consists of performing all $\binom{K}{2}$ possible pairs of lists using a Fellegi-Sunter-like approach and then somehow reconciling the discrepancies in an ad hoc fashion. We describe some important uses of the methodology, provide a principled way of accomplishing the reconciliation and we finally present some key parts of the generalization of Fellegi and Sunter's method to $K > 2$ lists.

Introduction

Record linkage is a family of techniques for matching two data files using names, addresses, and other fields that are typically not unique identifiers of entities. Most record linkage approaches are based or emulate a method presented in a pioneering paper by Fellegi and Sunter (1969). Winkler (1999) and Herzog et al. (2007) document the basic two-list methodology and several variations. We begin by reviewing some common applications of record linkage techniques, including the linking $K > 2$ lists.

Data Integration. Synthesizing data on a group of individuals from two or more files in the absence of unique identifiers requires record linkage, e.g., to create a data warehouse to be used for querying such as credit checks, e.g., see Talburt (2011), or to assess enrollment in multiple government programs, e.g., see Cole (2003).

Multiple Systems Estimation. Estimating the size of a population often requires the use of multiple samples from the population itself. This problem is widely known as capture–recapture estimation in biological settings and multiple systems estimation in social settings. These methods have the intrinsic assumption that the individuals simultaneously recorded by different samples can be identified (e.g., see Fienberg, 1972; Darroch et al., 1993). Record linkage becomes important in many social contexts where unique identifiers are unavailable, e.g., census correction and human rights violations (e.g., see Anderson and Fienberg, 2001; Guberek et al., 2010; Fienberg and Manrique-Vallier, 2009).

Analysis Using Linked Data. When the goal is a statistical analysis of the computer-matched files, care must be taken to propagate the uncertainty in the record linkage into the analysis. This line of research is referred to as "analytic linking" by Winkler (1999). For example, Lahiri and Larsen (2005) propose ways to perform regression on the record-linked files in a way that attempts to account for bias introduced by the matching error. The techniques hinge on the availability of well-calibrated probability models of record linkage, i.e., good estimates of the probability of a particular record-pair being a match. These techniques have applications in drug safety and surveillance, for example in long term vaccine surveillance, where the goal is to perform logistic regression predicting some side effect using patients' vaccine histories. In this case the side effect data may reside in medical records, whereas the data regarding vaccine exposure may reside in an insurer's database, e.g., see Brown et al. (2010).

In the next section, we describe a modern variation of the basic Fellegi–Sunter methodology. When the task involves linking $K > 2$ lists, the most common approach consists of performing all $\binom{K}{2}$ possible pairs of lists using a Fellegi–Sunter–like approach and then somehow reconciling the discrepancies in an ad hoc fashion. In a subsequent section, we provide a principled way of accomplishing the reconciliation and later we present an alternative which involves the generalization of the Fellegi–Sunter ideas. Finally we provide a comparison of the computational complexity of both approaches.

Record Linkage of Two Files

Following Fellegi and Sunter (1969), we let A and B denote the two overlapping subpopulations of individuals from some larger population whose data is recorded in files. We assume the existence of a pair of record-generating processes, α, β which produce the actual data recorded in the files, $\alpha(A) = \{\alpha(a); a \in A\}$ and $\beta(B) = \{\beta(b); b \in B\}$, where $\alpha(a)$ and $\beta(b)$ represent vectors of information of individuals a and b, respectively. The set of ordered record pairs

$$\alpha(A) \times \beta(B) = \big\{\big(\alpha(a), \beta(b)\big); a \in A, b \in B\big\}$$

is the union of the set of *matched* record pairs M, with the set of *unmatched* record pairs U, i.e.,

$$M = \big\{\big(\alpha(a), \beta(b)\big); a = b, a \in A, b \in B\big\} \text{ and } U = \big\{\big(\alpha(a), \beta(b)\big); a \neq b, a \in A, b \in B\big\}.$$

Our aim is to identify M and U, a process that is non-trivial in the absence of unique identifiers. Below we present a modern version of the Fellegi–Sunter approach.

Reduction of Data by Comparison Functions. We apply a vector of comparison functions, γ, to each record pair, $r_j \in \alpha(A) \times \beta(B)$, yielding $\gamma^j = \gamma(r_j) = (\gamma_1(r_j), \ldots, \gamma_\iota(r_j)) \in \{0,1\}^\iota$, where $\gamma_i^j = \gamma_i(r_j)$ is the i^{th} comparison function applied to pair r_j. A simple choice of comparison function, when the files record the same set of variables, is to set $\gamma_i(r_j) = 1$ whenever the records have the same value for field i, and zero otherwise. This information alone is insufficient for the determination of whether $r_j \in M$, since the variables being compared are random in nature. Therefore we estimate $\mathbb{P}(r_j \in M|\gamma^j)$ and $\mathbb{P}(r_j \in U|\gamma^j) = 1 - \mathbb{P}(r_j \in M|\gamma^j)$. In the absence of correctly matched files, there is no easy way to obtain these probabilities, but we can estimate them using the EM algorithm.

EM Estimation. Applying Bayes' rule to $\mathbb{P}(\gamma^j|r_j \in M)$ and $\mathbb{P}(\gamma^j|r_j \in U)$ we get

$$(1) \quad \mathbb{P}(r_j \in M|\gamma^j) = \frac{\mathbb{P}(\gamma^j|r_j \in M)\mathbb{P}(r_j \in M)}{\mathbb{P}(\gamma^j|r_j \in M)\mathbb{P}(r_j \in M) + \mathbb{P}(\gamma^j|r_j \in U)(1 - \mathbb{P}(r_j \in M))}$$

Next, we let

$$g_j = \begin{cases} 1 & \text{if } r_j \in M, \\ 0 & \text{if } r_j \in U, \end{cases}$$

and define $x_j = (g_j, \gamma^j)$ as the "complete data" vector for r_j. Winkler (1988) and Jaro (1989) model the complete data, x_j, via some vector of parameters Φ, as:

$$\begin{aligned} \mathbb{P}(x_j; \Phi) &= \Big[\mathbb{P}(\gamma^j, r_j \in M)\Big]^{g_j} \Big[\mathbb{P}(\gamma^j, r_j \in U)\Big]^{1-g_j} \\ &= \Big[\mathbb{P}(\gamma^j|r_j \in M)\mathbb{P}(r_j \in M)\Big]^{g_j} \Big[\mathbb{P}(\gamma^j|r_j \in U)(1 - \mathbb{P}(r_j \in M))\Big]^{1-g_j}, \end{aligned}$$

and thus they obtain the log–likelihood for the sample $\mathbf{x} = \{x_j; j = 1, \dots, n\}$ as

$$\ell \;=\; \sum_{j=1}^{n} g_j \log\big[\mathbb{P}(\gamma^j|r_j \in M)\mathbb{P}(r_j \in M)\big] + \sum_{j=1}^{n}(1-g_j)\log\big[\mathbb{P}(\gamma^j|r_j \in U)\mathbb{P}(r_j \in U)\big].$$

In order to estimate $\mathbb{P}(\gamma^j|r_j \in M)$ and $\mathbb{P}(\gamma^j|r_j \in U)$, Fellegi and Sunter (1969) use the simplifying assumption that the components of the vector γ^j are conditional independent with respect to the state of the indicator g_j, i.e.,

$$\mathbb{P}(\gamma^j|r_j \in M) = \prod_{i=1}^{\iota} m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j} \text{ and } \mathbb{P}(\gamma^j|r_j \in U) = \prod_{i=1}^{\iota} u_i^{\gamma_i^j}(1-u_i)^{1-\gamma_i^j},$$

where $m_i = \mathbb{P}(\gamma_i^j = 1|r_j \in M)$ and $u_i = \mathbb{P}(\gamma_i^j = 1|r_j \in U)$. Defining $p = \mathbb{P}(r_j \in M)$, we obtain the log-likelihood as

$$\ell \;=\; \sum_{j=1}^{n} g_j \log\big[p \prod_{i=1}^{\iota} m_i^{\gamma_i^j}(1-m_i)^{1-\gamma_i^j}\big] + \sum_{j=1}^{n}(1-g_j)\log\big[(1-p)\prod_{i=1}^{\iota} u_i^{\gamma_i^j}(1-u_i)^{1-\gamma_i^j}\big]$$

Since the values of g_j are unknown, we do the estimation of the parameters $\Phi = (p, m_1, \dots, m_\iota, u_1, \dots, u_\iota)$ via maximum likelihood estimation using the EM algorithm following Jaro (1989).

Weights. Once we estimate the parameters Φ, our next step is to determine the matched pairs of records using estimates of the likelihood ratios:

$$w^j = \log\frac{\hat{P}(r_j \in M|\gamma^j)}{\hat{P}(r_j \in U|\gamma^j)} \propto \log\frac{\hat{P}(\gamma^j|r_j \in M)}{\hat{P}(\gamma^j|r_j \in U)} = \sum_{i=1}^{\iota} w_i^j \quad \text{where} \quad w_i^j = \begin{cases} \log(\frac{\hat{m}_i}{\hat{u}_i}) & \text{if } \gamma_i^j = 1, \\ \log(\frac{1-\hat{m}_i}{1-\hat{u}_i}) & \text{if } \gamma_i^j = 0. \end{cases}$$

The Assignment Problem. Having obtained the weights for each record pair, we can treat the assignment of record pairs as a linear sum assignment problem, following Jaro (1989). Setting $c_{ab} = w_j$ for some $a \in A$, $b \in B$ and where j is the index associated with the pair (a, b), we take:

$$y_{ab} = \begin{cases} 1 & \text{if record } (\alpha(a), \beta(b)) \in M, \\ 0 & \text{otherwise.} \end{cases}$$

We then solve the maximization problem:

$$\max_y \sum_{a=1}^{|A|}\sum_{b=1}^{|B|} c_{ab}y_{ab} \text{ subject to } y_{ab} \in \{0,1\}, \quad \sum_{a=1}^{|A|} y_{ab} \le 1, \quad b = 1, 2, \dots, |B|, \quad \sum_{b=1}^{|B|} y_{ab} \le 1, \quad a = 1, 2, \dots, |A|.$$

The first constraint ensures that the variables represent a discrete structure, and the second and third constraints assure that each element of A is matched with at most one element of B. This is a maximum-weight bipartite matching problem, for which efficient algorithms exist. Note that this step is convenient only if there are not intra list duplicates.

Cutoff Values. The process thus far yields matching that maximizes the sum of the weights among the declared matches, but the possibility exists that the matching will include some pairs with a very low matching weight. Thus Fellegi and Sunter (1969) propose to compute cutoff values of the weights, to declare a pair as a link or as a non-link. We order the 2^ι possible values of γ^j by their weights in decreasing order indexing by the subscript (j) and determine two values, (j') and (j''), such that

$$\sum_{(j)=1}^{(j')-1} \mathbb{P}(\gamma^j|r_j \in U) < \mu \leq \sum_{(j)=1}^{(j')} \mathbb{P}(\gamma^j|r_j \in U) \text{ and } \sum_{(j)=(j'')}^{2^\iota} \mathbb{P}(\gamma^j|r_j \in M) \geq \lambda > \sum_{(j)=(j'')+1}^{2^\iota} \mathbb{P}(\gamma^j|r_j \in M)$$

where $\mu = \mathbb{P}(\text{assign } r_j \text{ as link}|r_j \in U)$ and $\lambda = \mathbb{P}(\text{assign } r_j \text{ as non–link}|r_j \in M)$ are two admissible error levels. Finally, we divide the record pairs assigned with configurations of $\gamma^{(j)}$ into three groups: (1) those for $(j) \leq (j') - 1$ are links, (2) those for $(j) \geq (j'') + 1$ are non–links, and (3) those with configurations between (j') and (j'') require clerical review.

Blocking. When the sizes of the data files to be linked are moderate (e.g., tens of thousands of records or more) then applying the above theory may be too inefficient, since there would be hundreds of millions of pairs under consideration. A common way to deal with this problem is to perform “blocking” whereby we remove “obvious” non-matches from consideration, leaving blocks of potential links. The terminology goes back in some sense to the census uses where the population is divided into physical blocks, but also reflects the experimental design notion of “blocking” to remove heterogeneity. The idea is that a “reliable” field such as zip code or gender may be used to quickly label some of the non-links. See Herzog et al. (2007) for discussion. The result is a tradeoff of computational efficiency versus accuracy in the final linkage; however, the impact on the accuracy is usually fairly mild.

Extensions of Fellegi-Sunter for More than Two Data Files

We now turn to the possibility that multiple files may be matched together at the same time—an important practical problem heretofore explored only via ad hoc approaches. Two possibilities suggest themselves: (1) estimate the matching weights for each pair of files, and then perform matching of all the files at once (reconciliation), or (2) estimate the matching weights for the direct product of all the files, and then perform the matching. We explore these alternatives in turn.

Let $A_1, A_2, \ldots, A_K$ correspond to K overlapping subpopulations with recorded data files and suppose that, for each data file, there exists one different record generating process

$$\alpha_k(A_k) = \{\alpha_k(a_k); a_k \in A_k\}, k = 1, \ldots, K,$$

where the member $\alpha_k(a_k)$ represents a vector of information for individual a_k, who belongs to subpopulation A_k. This information could be erroneous or incomplete.

Define the K-ary cartesian product $\bigotimes_{k=1}^{K} \alpha_k(A_k) = \big\{\big(\alpha_1(a_1), \alpha_2(a_2), \ldots, \alpha_K(a_K)\big); a_k \in A_k, k = 1, \ldots, K\big\}$ composed by all the possible record K-tuples in which the kth entry corresponds to the information recorded for some unit a_k in the subpopulation k. When we dealt with two files, this cartesian product was simply partitioned into the set of matches and non-matches. With multiple files a tuple can contain multiple individuals and therefore we must re-define the goal of record linkage.

It is possible that certain record K-tuple contain information about K different individuals, i.e. for some $\big(\alpha_1(a_1), \alpha_2(a_2), \ldots, \alpha_K(a_K)\big)$, $a_i \neq a_j$ for all pairs i, j. At the other extreme, the same individual can appear in all K data files, i.e., in the record K-tuple $\big(\alpha_1(a_1), \alpha_2(a_2), \ldots, \alpha_K(a_K)\big)$, $a_1 = a_2 = \cdots = a_K$. In general, we aim to classify the elements of each record K-tuple into subsets that record information about the same individual. In order to establish formally this idea, let Π_K denote the set of partitions of the set $\mathbb{N}_K = \{1, 2, \ldots, K\}$. Thus, each record K-tuple is characterized by an element of Π_K.

Let S_p be the set of record K-tuples corresponding to the pattern of agreement $p \in \Pi_K$. It is clear that

$$(2) \quad \bigotimes_{k=1}^{K} \alpha_k(A_k) \quad = \quad \bigcup_{p \in \Pi_K} S_p$$

Thus the goal of multiple record linkage is to partition the product of the files into the disjoint sets S_p corresponding to the $p \in \Pi_K$. The number of ways a set of K elements can be partitioned into nonempty subsets is the Kth *Bell number*, B_K, found using the recurrence relation $B_K = \sum_{k=0}^{K-1} B_k \binom{K-1}{k}$, where $B_0 = 1$. There are B_K subsets S_p of record K-tuples. Let n denote the cardinality of (2). Also, for $j = 1, \ldots, n$, let $r_j = \big(\alpha_1(a_1), \alpha_2(a_2), \ldots, \alpha_K(a_K)\big)$, for some $a_k \in A_k, k = 1, \ldots, K$, be the jth record K-tuple of the K-ary product (2).

As before, in the absence of unique identifiers for the individuals, the partitioning cannot be done exactly. Therefore we take the same approach as above, namely to reduce each tuple to a set of matching variables, then to estimate a model of the match variables given the partition. By using this information our goal is to estimate the probability that the record K-tuple belongs to each subset S_p. We proceed by first demonstrating the possibility of using the above two-file record linkage approach as a module in a larger procedure, then give a second approach which instead generalizes the Felligi–Sunter method for multiple files.

Approach 1: Reconciling Bipartite Record Linkages

Suppose we had estimated the $\binom{K}{2}$ Fellegi–Sunter models described above, one for each pair of files under consideration. We may then naturally compute the probability that a particular K-tuple belongs to a certain block of the partition S_p. For notational convenience define the relation $a \equiv_p b$, for $a, b \in \{1, \ldots, K\}$ whenever a, b are in the same block of the partition $p \in \Pi_K$. Then a tuple $r = \big(\alpha_1(a_1), \alpha_2(a_2), \ldots, \alpha_K(a_K)\big)$ is in S_p, if for all pairs i, j we have $i \equiv_p j$ if and only if $a_i = a_j$, in other words elements are in the same partition iff they correspond to the same individual. For each pair i, j we have a Fellegi-Sunter model which gives:

$$\mathbb{P}((a_i, a_j) \in M_{i,j} | \gamma_{i,j}(r)) \quad \text{and} \quad \mathbb{P}((a_i, a_j) \in U_{i,j} | \gamma_{i,j}(r))$$

i.e., the probability that elements (a_i, a_j) are matches or non matches, based on the requisite match variables (here represented as functions of the entire tuple). We therefore may take:

$$\mathbb{P}(r \in S_p | \gamma(r)) \propto \prod_{i \neq j} \mathbb{P}((a_i, a_j) \in M_{i,j} | \gamma_{i,j}(r))^{1\{i \equiv_p j\}} \mathbb{P}((a_i, a_j) \in U_{i,j} | \gamma_{i,j}(r))^{1\{i \not\equiv_p j\}}$$

The normalizing constant is the summation over the B_K possible partitions. Then the final partitioning of the tuples may be done via a method described below. This method is conceptually appealing since it uses the original Fellegi–Sunter method as a sub-routine to estimate model parameters, and then finally performs a joint record linkage of all the files at once.

Approach 2: Fellegi–Sunter Extension for $K > 2$ Data Files

The alternative is to modify the Fellegi–Sunter approach so that it directly applies to multiple files. We describe the changes needed to the original method below.

Comparison Data. In order to model the probability that a certain record K-tuple belongs to some subset S_p, we determine the pattern of agreement for each field of the information recorded. If we search for agreement for a certain field, we can associate each component of the record K-tuple with a number in $\{1, 2, \ldots, K\}$ and a partition would describe the pattern of agreement of the record K-tuple, grouping in the same element of the partition all the files that agree in the field being compared. Now, let $\gamma_p^{j_f} = 1$ if the record K-tuple r_j has the pattern of agreement p in the field f. Then, for each field $f = 1, \ldots, F$, of each record K-tuple r_j, we obtain a vector $\gamma^{j_f} = (\gamma_{1/2/\ldots/K}^{j_f}, \ldots, \gamma_p^{j_f}, \ldots, \gamma_{12\ldots K}^{j_f})$. The length of the vector γ^{j_f} is B_K, since that is the number of patterns of agreement for each field. Finally, the comparison data for r_j is the vector containing the information of all the F fields, and we write it as $\gamma^j = (\gamma^{j_1}, \ldots, \gamma^{j_f}, \ldots, \gamma^{j_F})$, which represents $(B_K)^F$ possible patterns of agreement γ^j for each record K-tuple.

Matching Probabilities. Let $g^j = (g_{1/2/\ldots/K}^j, \ldots, g_{12\ldots K}^j)$ be the vector that indicates the subset S_p that contains the record K-tuple r_j, such that $g_p^j = 1$ if $r_j \in S_p$ and it is 0 otherwise. Thus, it is clear that $\sum_{\Pi_K} g_p^j = 1$. Now, let $x^j = (g^j, \gamma^j)$ be the (unobserved) complete data vector for r_j. In order to model the matching probabilities we consider

$$\mathbb{P}(x^j|\Phi) \quad = \quad \prod_{p \in \Pi_K} \Big[\mathbb{P}(\gamma^j|S_p)\mathbb{P}(S_p)\Big]^{g_p^j}.$$

Each γ^{j_f} represents the pattern of agreement of r_j in the field f, which corresponds to categorical information that we can model using a multinomial distribution with just one trial as

$$\mathbb{P}(\gamma^{j_f}|S_p) \quad = \quad \prod_{p' \in \Pi_K} (\pi_{p'|p}^f)^{\gamma_{p'}^{j_f}},$$

where $\pi_{p'|p}^f = \mathbb{P}(\gamma_{p'}^{j_f} = 1|S_p)$, and p' is just another indicator of the patterns of agreement in Π_K. Under conditional independence for the sample $\mathbf{x} = \{x^j; j = 1, \ldots, n\}$ and the assumption of independence for the comparison data of each field, we obtain the complete log–likelihood as

$$L \quad = \quad \sum_{j=1}^{n} \sum_{p \in \Pi_K} g_p^j \Big[\log \mathscr{s}_p + \sum_{f=1}^{F} \sum_{p' \in \Pi_K} \gamma_{p'}^{j_f} \log \pi_{p'|p}^f\Big],$$

where $\mathscr{s}_p = \mathbb{P}(S_p)$. We can use a standard EM algorithm for estimation here, e.g., see Sadinle (2011).

A Generalized Fellegi–Sunter Decision Rule. Each record K-tuple is potentially declared to belong to the subset S_p if and only if p is the pattern for which $\hat{\mathbb{P}}(S_p|\gamma^j)$ is maximum among all possible patterns in Π_K. Thus, the set of record K-tuples is partitioned in B_K subsets and for each tuple in one of these partitions we consider only two possibilities, whether to declare it belongs to the subset S_p or to keep it undeclared. For the record K-tuples in each partition, we order the possible values of γ^j using weights $w_p^j = \text{logit}[\hat{\mathbb{P}}(S_p|\gamma^j)]$ in non-increasing order indexing by the subscript $(j)_p$. See Sadinle (2011) for a detailed justification and a comparison with the traditional weights for bipartite record linkage. Later, we find one value $(j')_p$ for each set of weights related to each subset, in order to determine the record K-tuple membership. The value $(j')_p$ satisfies

$$\sum_{(j)_p=1}^{(j')_p-1} \hat{\mathbb{P}}(\gamma^{(j)_p}|S_p^c) < \mu_p \leq \sum_{(j)_p=1}^{(j')_p} \hat{\mathbb{P}}(\gamma^{(j)_p}|S_p^c)$$

where $\mu_p = \mathbb{P}(\text{assign to } r_j \text{ the membership of } S_p | r_j \in S_p^c)$ is an admissible error level. We can compute each $\hat{\mathbb{P}}(\gamma^{(j)} | S_p^c)$ as

$$\hat{\mathbb{P}}(\gamma^{(j)_p} | S_p^c) = \frac{\hat{\mathbb{P}}(\gamma^{(j)_p}) - \hat{\mathbb{P}}(\gamma^{(j)_p} | S_p)\hat{\mathscr{s}}_p}{1 - \hat{\mathscr{s}}_p}.$$

Finally, for those record K-tuples with configurations of $\gamma^{(j)_p}$, $(j)_p = 1, \dots, (j')_p - 1$, we allocate them to the subset S_p. For those record K-tuples with configurations $\gamma^{(j)_p}$ with $(j)_p \geq (j')_p$, we keep them undeclared. This decision rule minimizes the probability of assigning the tuples to the wrong subset S_p or keeping them undeclared, subject to a set of admissible error levels μ_p, $p \in \Pi_K$ (Sadinle, 2011). We could then send the undeclared tuples to clerical review or iterate between clerical review and refitting the mixture model (Larsen and Rubin, 2001).

Computational Issues

When the number of files and their sizes grow large, then one or both of the above techniques may be infeasible from a computational perspective. Thus we briefly describe the growth of the computational complexity. We use K to be the number of files and n, F to be the maximum number of elements and fields respectively in any file. The first step of either approach is the construction of the match variables. For the first approach there are $\binom{K}{2} n^2 F = O(K^2 n^2 F)$ pairs to consider (modulo blocking). For the latter approach, entire K-tuples are considered at once, of which there are $O(n^K F)$ modulo blocking, therefore the latter approach may be vastly more computationally expensive as the number of files is allowed to grow.

The second phase is the estimation of the parameters. Both procedures resort to the use of EM, for which the number of iterations depend on the data themselves and are impossible to characterize in terms of the input sizes. Nevertheless we may consider the computational cost of an iteration of EM in either case. In the reconciliation approach, there are $\binom{K}{2}$ independent instances of EM, in which there are two models to update in the M-step, one for links and one for non-links. In the E-step there are at most 2^F matching patterns for which the class membership indicator expectations are computed. Thus the overall complexity of an iteration scales as $O(K^2 2^F)$. For the second approach, there are essentially $|\Pi_k|$ models to update in the M-step, and in the E-step there are $|\Pi_K|^F$ matching patterns (recall that $|\Pi_K| = B_K$ is the K^{th}-Bell number which is exponentially large in K). Thus the overall complexity of the EM phase of this approach is $O(B_K^F)$ which may grow astronomically large as K increases. In the final phase, the approaches seem to be on an equal footing if the goal is to optimize exactly the probability of the output partitioning, so both require computing $O(B_K n^K)$ partition membership probabilities for the tuples. We speculate that this could be instead rapidly approximated, if we use the pairwise Felligi–Sunter models to efficiently filter out those tuples in which some elements are either clearly matches or clearly non-matches; however, this is a goal for future work.

While the second approach may be more expensive for moderate or large numbers of files, when the number is small, e.g., $K = 3$, then either approach will be computationally tractable, and we may anticipate superior performance from the second model, since it encodes a richer model of the matching variables, i.e., it does not make a false independence statement with respect to the pairs of files.

Acknowledgments

This research was partially supported by Army contract DAAD19-02-1-3-0389 to Cylab at Carnegie Mellon University, and by a grant from the Interactive Digital Media R&D Programme Office and the Singapore

National Research Foundation to the Living Analytics Research Centre, Singapore Management University and Carnegie Mellon University.

References

Anderson, M. and Fienberg, S. (2001). *Who Counts? The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation, New York, revised paperback edition.

Brown, J., Holmes, J., Shah, K., Hall, K., Lazarus, R., and Platt, R. (2010). Distributed health data networks: A practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Medical Care*, 48(suppl):S45–S51.

Cole, N. (2003). *Feasibility and Accuracy of Record Linkage To Estimate Multiple Program Participation. Volume I, Record Linkage Issues and Results of the Survey of Food Assistance Information Systems*. Electronic Publications from the Food Assistance & Nutrition Research Program. Economic Research Service, Washington, DC.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *JASA*, 88(423):1137–1148.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *JASA*, 64(328):1183–1210.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, 59(3):591–603.

Fienberg, S. E. and Manrique-Vallier, D. (2009). Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. *AStA Advances in Statistical Analysis*, 93(1):49–60.

Guberek, T., Guzmán, D., Price, M., Lum, K., and Ball, P. (2010). To count the uncounted: An estimation of lethal violence in Casanare. A Report by the Benetech Human Rights Program.

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.

Lahiri, P. and Larsen, M. (2005). Regression analysis with linked data. *JASA*, 100(469):222–230.

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

Sadinle, M. (2011). Multiple record linkage: Generalizing the Fellegi–Sunter theory. Unpublished manuscript.

Talburt, J. R. (2011). *Entity Resolution and Information Quality*. Morgan Kaufmann, Burlington, MA.

Winkler, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 667–671. American Statistical Association.

Winkler, W. E. (1999). The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census.