Valid Statistical Inference on Automatically Matched Files

Rob Hall and Stephen Fienberg

Department of Statistics and Machine Learning Department,
Carnegie Mellon University, Pittsburgh, PA, 15213 USA
rjhall@cs.cmu.edu, fienberg@stat.cmu.edu

Abstract. We develop a statistical process for determining a confidence set for an unknown bipartite matching. It requires only modest assumptions on the nature of the distribution of the data. The confidence set involves a set of linear constraints on the bipartite matching, which permits efficient analysis of the matched data, e.g., using linear regression, while maintaining the proper degree of uncertainty about the linkage itself.

1 Introduction

Record linkage is an historically important statistical problem arising when data about some population of individuals is spread over several files. Most of the literature focuses on the two file setting. The record linkage goal is to determine whether a record from one file corresponds to a record of a second file, in the sense that the two records describe the same individual. Winkler and others describe application areas, computational techniques and statistical underpinnings in detail in [1, 5, 9, 10]. The typical purposes of record linkage are:

- data integration.
- as an intermediate step in performing a computation on the integrated data.
- to create a public use file that will allow others to analyze the integrated data.

There are also other purposes which are especially relevant in the field of privacy-preserving data mining. First, when one or more datasets have been modified in order to respect the privacy of the individuals, then linking them without using identifiers in order to carry out a statistical analysis requires special care by the analyst. Second there is a more sinister reason for undertaking record linkage, which is to permit re-identification of individuals in supposedly anonymized data releases (see e.g., [2]). Any method which is intended to facilitate statistical analyses of data for which the true linkage structure is unknown may also be used by an intruder to attack supposedly anonymized datasets via linkage to records of known individuals. Thus, a privacy protection scheme should be judged based on how difficult it makes the latter task.

We aim to admit all types of statistical inferences, by generating a "confidence set" of linkage structures which has some requisite coverage probability. This way, our technique may in principle be useful for any analysis which takes as input a linked file. For example, suppose the problem is to determine the number of matching records. By taking the minimum and maximum number of links among all linkage structures in the confidence set, we can obtain a valid confidence interval for the number of matching records. This is useful for "capture-recapture" problems, but also may in principle lead to a leakage of information regarding certain populations, e.g., by linking some anonymized private data against a file of known criminals. Our method is also useful for the regression of a variable in one file against covariates in another file, by considering the maximum and minimum values of the coefficients that are found when computed on all the matchings in the confidence set. Of course, since we deal with an exponentially large space of structures we may anticipate that the confidence set we produce will itself be exponentially large, which would preclude exhaustive enumeration of the set. Therefore we demonstrate that the set may be represented by a small number of constraints, which means that the maximization of a statistic over the set may be achieved by some form of constrained optimization.

Our contributions are:

- We propose a nonparametric model for record linkage.
- We give a nonparametric hypothesis test which rejects an assignment on the basis that it contains too few of the true links.
- We demonstrate how this test may be relaxed so that a confidence set of assignments may be rapidly constructed.
- We demonstrate that rejecting an assignment on the basis that it contains false links is infeasible under our model, so we construct a parametric test for this purpose.

Related Work

The problem of performing a valid statistical analysis between two files which require matching was considered in [6]. There the setting was that one file contained a response variable while the other contained predictors. Their goal was to perform regression accurately without requiring human intervention to resolve the matching. They use a record linkage model similar to the model of Felligi and Sunter [5], estimate the parameters using EM, and then (supposing that model to be correct) use it to unbias a least squares regression estimate.

Another related work is [7]. There the analysis they are interested in is determining the size of the matched set of records (i.e., the number which appear in both files). This is useful for estimating the population size via a capture-recapture approach. They obtain Bayesian credible intervals for the size of the matched set.

What we propose is similar in spirit to these techniques but perhaps more versatile. The construction of valid frequentist confidence sets for the matching allows the computation of confidence intervals for several statistics of interest. These range from e.g., the size of the matching as considered in [7] to intervals for regression coefficients, among others.

Finally we note that there is a history of record linkage use for statistical analyses when the data are private. An overview of techniques are given by [4]. Although we focus on the statistical methodology in this paper, in principle the ideas could be applied in a privacy-preserving way (e.g., by instantiating our method inside a cryptographic framework).

2 Nonparametric Model

We now introduce the model which we use for the remainder of this paper. We consider the problem in an abstract fashion in which the records are envisioned as nodes in a graph, and the linkage or "assignment" is considered as a subset of edges. The observations constitute two sequences of data points, $\boldsymbol{x} = (x_1, \ldots, x_m)$, $\boldsymbol{y} = (y_1, \ldots, y_n)$, forming a combined data sample

$$x_1 \ldots, x_m, y_1, \ldots, y_n \in \mathcal{X}^{n+m},$$

where without loss of generality, $m \leq n$, and where $\mathcal{X}$ is some abstract space in which the records lay. For example, in the case of records containing several measurements of the individuals, $\mathcal{X}$ may be considered as a product space of the ranges of the measurements (e.g., $\mathbb{R}^p$ in the case of real valued measurements). We consider these two sets of observations as nodes of a graph, and our goal is to determine a bipartite matching between the sets. A matching Π is a set of (x_i, y_j) pairs such that each element x_i, y_j may appear in at most one pair. In the case when $|\Pi| = m$ we say the matching is "maximal" in the sense that it is impossible to add more pairs without first removing some. When $m = n = |\Pi|$ then Π is called a "perfect matching." We consider Π to be a subset of the edges of the complete bipartite graph formed from the x_i, y_j.

We denote by S_X the set of elements x_i that do not appear in a pair in Π, and likewise define S_Y (these elements are the "singletons"). We propose a model for the data in which the density factorizes according to the bipartite matching

$$dP(\boldsymbol{x}, \boldsymbol{y}) = \prod_{(x_i, y_j) \in \Pi} f(x_i, y_j) \prod_{x_i \in S_X} g(x_i) \prod_{y_j \in S_Y} g(y_j), \tag{1}$$

in which f, g are density functions. We only place the following restriction on these functions

$$f(a, b) = f(b, a), \forall a, b \in \mathcal{X},$$

and

$$\int_{\mathcal{X}} f(a, x) \; dx = \int_{\mathcal{X}} f(x, a) \; dx = g(a), \forall a \in \mathcal{X}.$$

Thus we may consider f to be a symmetric bivariate density on the linked pairs, and g to be the marginal. An example which fits into this regime is

$$f(a, b) = \int_{\mathcal{X}} p(a|c) p(b|c) q(c) \; dc, \;\; g(a) = \int_{\mathcal{X}} p(a|c) q(c).$$

In this example c may be some underlying element of the population due to q, and p represents some "distortion model." For example, in the case that x_i, y_j are elements of databases about individuals, then q may represent some sampling distribution over the population (which is assumed to be shared by both databases), whereas p may represent a model of typographical errors or measurement errors that corrupt the records. This above model encodes the assumption that the errors in the records are equally likely in either data sequence.

The requirement that f be symmetric is the lynchpin of the hypothesis test we present below, because it constrains the sufficient statistics have a particular structure. This type of assumption is reasonable when e.g., the same agency is responsible for taking all the measurements. In the case that the two files arise from different agencies, however, the distributions of measurement errors may be different between the two files. This latter situation may be handled for example if the distributions of measurement errors were known or could be estimated, by e.g., sampling new values for each measurement from the posterior distribution over the non-errorful measurement. We leave extensions such as this for future exploration.

We require that $\mathcal{X}$ be equipped with an ordering denoted by $<$. Then we consider the sample put in order with respect to the assignment. We take the pairs of Π, and order the elements in each pair according to $<$. Then, these ordered pairs are put into the lexicographic ordering corresponding to $<$. We call this sequence $A(\Pi)$. Likewise we may consider the sequence of "singletons" $S_X \cup S_Y$ having been put into the proper order. We call this sequence $B(\Pi)$. It is clear that $A(\Pi), B(\Pi)$ constitute the sufficient statistics for the above model. Consider the case when the elements take different values almost surely (e.g., the case of real valued measurements). Then we have

$$dP_\Pi(\boldsymbol{x}, \boldsymbol{y}) = \left(2^{-|A|} \binom{|B|}{S_x}^{-1} \right) \prod_{(a_1,a_2)\in A} f(a_1, a_2) \prod_{b \in B} g(b), \tag{2}$$

where the first term is the reciprocal of the number of ways in which the sufficient statistics may be re-arranged into a sample $\boldsymbol{x}, \boldsymbol{y}$, and the remaining terms are functions of the sufficient statistics and the unknown density functions. Note that each pair in A is in one of two configurations, depending on which element is assigned to $\boldsymbol{x}$ and which to $\boldsymbol{y}$. The singletons of B comprise two sets of size S_X, S_Y. and hence the binomial coefficient appearing in expression 2. Considering $\boldsymbol{x}, \boldsymbol{y}$ as a rearrangement of the sufficient statistics A, B is the heart of the hypothesis testing approach described below.

3 A "Permutation Test" for False Non-Links

We now present a fairly general scheme for testing the hypothesis $\Pi = \Pi_0$ against the alternative $\Pi \neq \Pi_0$. Note that the null hypothesis specifies the sufficient statistics $A(\Pi_0), B(\Pi_0)$. The overarching strategy is to choose a test statistic for which we can evaluate the conditional distribution given the sufficient statistics for the model. This way, we can calculate rejection regions despite our lack of knowledge of the density f from which the sample is generated. Such a statistic is one which depends on the particular arrangement of the sufficient statistics. Thus a test statistic which is constant with respect to these rearrangements would not be useful.

The overarching strategy is this:

1. Choose a set of pairs of records $D = D(A(\Pi_0), B(\Pi_0))$ as a function of the sufficient statistics.
2. Let $T(\Pi_0) = T(\Pi_0, D, \boldsymbol{x}, \boldsymbol{y})$ be the number of edges in D which cross between an x_i and a y_j. Note that this depends on the observed arrangement of the sufficient statistics into $\boldsymbol{x}, \boldsymbol{y}$.
3. Compute the distribution of T by summation over the arrangements
$$\begin{aligned} &P_{\Pi_0}(T(\Pi_0) = t | A(\Pi_0), B(\Pi_0)) = \\ &\qquad \sum_{\boldsymbol{x},\boldsymbol{y}} \mathbf{1}\left\{T(\Pi_0, D, \boldsymbol{x}, \boldsymbol{y}) = t\right\} P_{\Pi_0}(\boldsymbol{x}, \boldsymbol{y} | A(\Pi_0), B(\Pi_0)) \end{aligned} \tag{3}$$
4. Reject Π_0 whenever $T(\Pi_0) > T_{1-\alpha}(\Pi_0)$ the latter being the $1 - \alpha$ quantile of the distribution of T given the sufficient statistics as in (3).

This technique will give a valid hypothesis test with size α, due to the sufficiency of $A(\Pi_0), B(\Pi_0)$. Since we achieve the correct false rejection rate for each value of A, B (by the definition of the test), we also achieve the correct overall false rejection rate, since the latter is nothing more than the expectation of the former, where the distribution in question is that of the A, B which depends on the unknown densities. This way we construct a test which achieves the correct size even though we do not know the form of f.

In implementing the above approach we consider the complete graph with $G = (V, E)$, where

$$V = (x_1, \ldots, x_m, y_1, \ldots, y_n), \quad E = \{(u, v) : u, v \in V\}.$$

We define the set of edges which "cross" between an x_i and a y_j: $C = \{(x_i, y_j) : 1 \leq i \leq m, 1 \leq j \leq n\}$. The graph (V, C) is the complete bipartite graph. We choose a subset of edges $D \subset E$ in a way which depends only on the sufficient statistics of the model (note that the sufficient statistics are different for each null hypothesis Π_0). This restriction is not necessary to construct a valid hypothesis test, however it becomes easier to evaluate (3) when D does not change inside the sum. We also further restrict D in the interest of alleviating the computational and mathematical burden of determining the distribution of T. Namely we require that the edges in D are disjoint in the sense that no two edges are incident on the same node. In the interest of having a concrete running example, we take the following choice of D

$$D_{d,\epsilon} = \{(u, v) \in E : d(u, v) \leq \epsilon, d(u, w) > \epsilon, d(v, w) > \epsilon, \forall w \in V \setminus \{u, v\}\}, \tag{4}$$

i.e., those pairs for which the elements are close under the distance function $d(\cdot, \cdot)$, and for which no other elements are close (thus rendering the determination of the set unambiguous). Although d is written as operating on nodes of the graph, it would be based on the fields of the corresponding record, and may consist of e.g., string edit distances between fields of the two records, etc. The edges in this set are disjoint by definition. We take the statistic

$$T(\Pi_0) = |D \cap C \setminus \Pi_0|,$$

which measures the number of edges in D which are "crossing edges," and which are not contained in the assignment Π_0. We may reject the null hypothesis $\Pi = \Pi_0$ whenever $T(\Pi_0)$ is too large. This idea is conceptually similar to the permutation test. We construct the distribution of T, based on A, B by inspecting every re-labeling of the points which is consistent with Π_0. Since the definition of D did not depend on the labeling of the data it is the same set in each case, however C changes depending on whether the data are labelled as x or y, and therefore T also changes. Each re-labeling has equal probability under the null hypothesis, and therefore we take $T_{1-\alpha}$ so that the fraction of the re-labelings having $T > T_{1-\alpha}$ is at most α. A re-labeling of the data corresponding to Π_0 constitutes setting the orders of the $|\Pi_0|$ links (i.e., deciding for each pair, which element is the x_i and which is the y_j), and then assigning the remaining $m + n - 2|\Pi_0|$ points into sets of size $m - |\Pi_0|, n - |\Pi_0|$. We thus concentrate on labeling each sample as an x_i, or a y_j. We do not care about the ordering within each of these sets since it does not impact the test statistic and therefore the terms due to these rearrangements will cancel out. See figure 1 for an illustration of the principle of our proposed test.

3.1 Implementation and Inversion of the Test

In principle we may inspect every configuration of the data, evaluate T, and thereby construct the distribution of T (conditional on the sufficient statisics); however, this is not computationally efficient. Due to the restriction on D, that the edges be disjoint, we can compute the rejection region without resorting to such a full enumeration of the configurations of the data. What is more, we may relax the test so that we produce a valid critical value which holds for each Π_0 simultaneously. Due to space constraints we omit many details which may be found in a longer version of this document. [add ref to your ADA paper]

We begin by noting that the number of crossing edges from the different connected components formed by the edges $D \cup \Pi_0$ can be considered as independent random variables, therefore the distribution of T can be computed as the convolution of these. Whats more, for each connected component of two or more edges, the distribution is sub-Gaussian, that is, it is majorized by a certain Gaussian distribution. See [8] for details. Likewise for the edges which are incident on the singletons under the hypothetical matching, the distribution of the total number of crossing edges is also sub-Gaussian. The convolutions of these random variables is therefore majorized by an appropriate normal distribution,

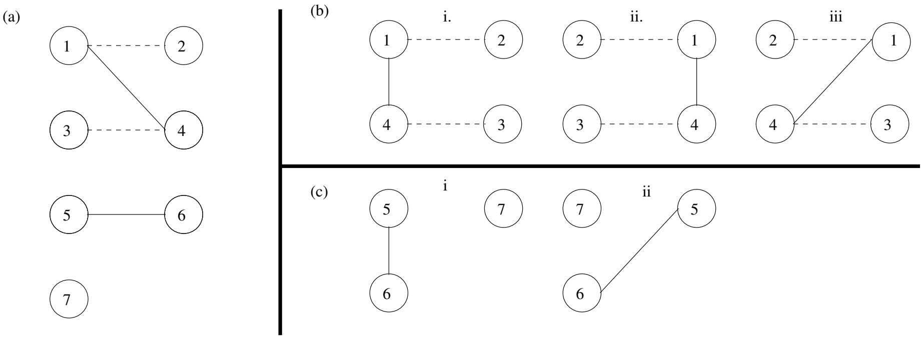

Fig. 1. (a) An example of a subset of edges D (shown as solid lines) and pairs due to a hypothetical assignment Π_0 (shown as dashed lines). The two columns correspond to the two databases. (b) The rearrangements of the linked pairs. (c) The rearrangements of the singletons. The vertices are numbered in order to make the rearrangement clear. Note that the number of edges which cross between the sides of the graph depends on the rearrangement. Only four out of the twelve possible rearrangements would have two crossing edges, thus the assignment in (a) may be rejected for $\alpha \geq 1/3$.

with mean and variance given by the sums of the means and variances of the individual distributions. These parameters depend on the number of edges in the components in question as well as their particular structure (whether they are chains, contain cycles etc). In essence each edge is crossing in approximately half of the rearrangements, but since the graph structure leads to dependence between the edges, the variance of the total number of crossing edges is on the order of $|D|/2$ (as opposed to the $|D|/4$ if they were independent).

We next propose a conservative relaxation of the test, namely one with a rejection region not depending on Π_0, and which only rejects Π_0 when the above test would. We consider the threshold

$$T^{\star}_{1-\alpha} \overset{\text{def}}{=} \max_{\Pi} T_{1-\alpha}(\Pi) \text{ s.t. } T(\Pi) < T_{1-\alpha}(\Pi),$$

thus the rejection region

$$T(\Pi_0) \geq T^{\star}_{1-\alpha}$$

leads to a conservative test. The reason is that for a specific null hypothesis Π_0, either $T^{\star}_{\alpha} \geq T_{\alpha}(\Pi_0)$ in which case it is immediate that this rejection region is a subset of the former, or if the opposite inequality holds we must have that $T(\Pi_0) \geq T_{\alpha}(\Pi_0)$ from the constraint in the definition, and so Π_0 would be rejected under both tests, since $T(\Pi_0) \geq T_{\alpha}(\Pi_0) > T^{\star}_{\alpha}$.

Before describing how $T^{\star}_{\alpha}$ is calculated, we remark that this relaxation of the test yields an appealing representation for the associated confidence set:

$$C_{1-\alpha} = \{\Pi : |D \cap C \setminus \Pi| < T^{\star}_{1-\alpha}\} \tag{5}$$

In other words, those assignments which include “enough” of the crossing edges of D. This confidence set may be seen as a constraint on the set of bipartite matchings, and this representation is useful when computing the extreme values of statistics which depend on the matching (e.g., as a constrained optimization problem).

Computation of $\boldsymbol{T^\star_{1-\alpha}}$ Consider maximization of the $1-\alpha$ quantile among all assignments which have $T(\Pi)=t$, i.e., those which include all but t of the crossing edges. We find

$$T^t_{1-\alpha} \overset{\text{def}}{=} \max_{\Pi : T(\Pi)=t} T_{1-\alpha}(\Pi) \approx \mathcal{N}_{1-\alpha, \frac{t+s+1}{2}, \frac{t+s}{2}},$$

in which the last term is the α quantile of a Normal distribution with the parameters arising from the sub-gaussian bounds. Here $s \overset{\text{def}}{=} |D \cap C^C|$ is the number of non-crossing edges of D. We may take

$$\tilde{T}^\star_{1-\alpha} \overset{\text{def}}{=} \max_t T^t_{1-\alpha} \text{ s.t. } t < T^t_{1-\alpha}.$$

Note that this is in essence an even further relaxed version of $T^\star_{1-\alpha}$, in which resulted from an unconstrained maximization over the Π having $T(\Pi)=t$, followed by a constrained maximization over t. The result is that $T^\star_{1-\alpha} \leq \tilde{T}^\star_{1-\alpha}$ and so the use of the latter quantity as the threshold still yields a test of the correct size. Which leads to

$$T^t_{1-\alpha} \approx \frac{s+t+1}{2} + \sqrt{\frac{s+t}{2}} Z_{1-\alpha},$$

where $Z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal. Performing the maximization of this quantity gives

$$\tilde{T}^\star_{1-\alpha} \approx s + 1 + Z^2_{1-\alpha} + \sqrt{\frac{Z^2_{1-\alpha}}{2} + 4Z^2_{1-\alpha}s}. \tag{6}$$

In summary we have a scheme to compute a confidence set of bipartite matchings which proceeds as follows:

1. Choose a set D of pairs, which are disjoint and in a way which is blind to the partitioning of the data into the two sets.
2. Count s, the number of these pairs which are not crossing, and compute $\tilde{T}^\star_{1-\alpha}$ using (6).
3. Construct the confidence set (5).

4 Testing for False Links

The test we have constructed thus far lacks the ability to reject an assignment on the grounds that it contains false links. Specifically, if $\Pi \subseteq \Pi_0$, then $T(\Pi) \geq$

$T(\Pi_0)$, and so clearly such Π_0 is never rejected with probability greater than α. As we demonstrated above, in order to successfully reject assignments on the grounds that they contain false links, we must leave the fully nonparametric setup.

We offer a solution which is not as general as the permutation test above was, but we tailor it in such a way as to yield an efficient algorithm. We restrict attention to case in which the data represent vectors of measurements about individuals. In addition, we suppose to have first constructed the above confidence set using the choice of D given in (4). The goal of this step is to reduce the size of the confidence set further. The main idea is to suppose some parametric form for the distances between true pairs, namely the distribution of $d(x, y)$ when $(x, y) \sim f$. When one or more datasets have been deliberately anonymized prior to their use, then knowledge of the anonymization procedure (e.g., differential privacy) will lead to the form for this distribution. Otherwise, we could estimate some model from a different sample of linked data. Finally if this is not possible then an arbitrary model may be proposed based on domain knowledge (e.g., distances between true pairs being bounded almost surely, and probability decreasing with distance), however this may come at the expense of a valid test. With such a model in hand, we construct a test by taking the set of edges with high distance

$$F = \{(u, v) \in E : d(u, v) \geq \tau\},$$

and constructing the test statistic

$$S(\Pi_0) = |\pi_0 \cap F|.$$

Then our goal is to now determine the critical value. Let τ be the ξ quantile of the distribution of $d(x, y)$ with $(x, t) \sim f$, then $S(\Pi_0)$ is binomially distributed with Π_0 trials and success probability ξ, and thus we may reject Π_0 whenever

$$S(\Pi_0) > B_{1-\alpha,|\Pi_0|,\xi},$$

the latter being the $1 - \alpha$ quantile of said binomial distribution. Once again this rejection region depends on Π_0 and we may again relax this into a conservative test by instead rejecting when

$$S(\Pi_0) > S^{\star}_{1-\alpha} \stackrel{\text{def}}{=} \min\{b \in \mathbb{N} : b > B_{1-\alpha,b+c,\xi}\},$$

where

$$c = \max\{|\Pi_0| : S(\Pi_0) = 0\}.$$

The reason is that the binomial quantile increases with the size of the matching, and any Π_0 having $S(\Pi_0) = b$ must have $|\Pi 0| < b + c$, therefore there is some minimal b above which any such Π_0 may be rejected. We can do this computation using a binary search. Finally we have the confidence set of matchings as:

$$C'_{1-\alpha-\beta} = \left\{\Pi : T(\Pi) < \tilde{T}^{\star}_{1-\alpha}, S(\Pi) < S^{\star}_{1-\beta}\right\}, \tag{7}$$

with coverage probability at least $1 - \alpha - \beta$ due to the subadditivity of probabilities.

5 Statistical Analysis Over the Confidence Set

Finally we describe a class of analyses we can carry out efficiently over this confidence set, to obtain the extreme values of a statistic of interest. Suppose our goal is to determine e.g., a regression of a variable in one file against a predictor in the other. If we knew the true matching we would take $\hat{\beta}(\varPi)$, the typical least squares estimator on the matched data. Since this is unknown, however, we can compute e.g.,:

$$\hat{\beta}^L_{1-\alpha} = \min_{\varPi \in C'_{1-\alpha}} \hat{\beta}(\varPi), \quad \hat{\beta}^U_{1-\alpha} = \max_{\varPi \in C'_{1-\alpha}} \hat{\beta}(\varPi)$$

So we may obtain a confidence interval for the regression coefficient by taking the maximum and minimum value that the regression coefficient reaches as the bipartite matching ranges over the confidence set. Evidently the coverage probability for the resulting confidence interval will be the same as the coverage probability for the set of bipartite matchings, since for the regression coefficient under the true matching to fall outside the interval would require that the true matching not appear in the confidence set. We stress that what we propose to obtain here are confidence intervals for a statistic one would normally compute, not for a parameter of interest (namely β, the true regression parameter). Therefore such a confidence interval would have to be dilated (e.g., convolved with a gaussian confidence interval) in order to obtain a valid interval for the parameter itself.

We concentrate on statistics of the form:

$$\hat{\beta}(\varPi) = \frac{1}{|\varPi|} \sum_{(u,v)\in\varPi} h(u,v),$$

where h is some function of the linked data elements under the matching $\varPi$. Examples that fit into this framework include estimation of covariance, for which h gives the product of the regressor and response variable, and estimation of a two dimensional histogram in which case h is the indicator of some set. To obtain a regression estimator for a particular coordinate k we can take

$$h(x_i, y_j) = \left(n(X^T X)^{-1} x_i y_j\right)_k .$$

Then $\frac{n}{|\varPi|}(X^T X)^{-1}$ approximates the restriction of $(X^T X)^{-1}$ to those elements which appear in the bipartite matching, a result which is heuristically appealing.

To rapidly compute the maximum and minimum of such sample means across the set of bipartite matchings, we seek to find the smallest set of the most extreme values that h can take, so that the corresponding matching is in C. Since $\varPi$ must contain at least $|D \cap C| - \tilde{T}^{\star}_{1-\alpha}$ of the edges of $D \cap C$, we begin by finding this number of edges which have the smallest (resp largest) values of h. Next, we add additional edges whenever h is less than (rep greater than) the average over the current set—until no more edges may be added (due to the constraint on the number of non-matching edges). This is a greedy algorithm and thus is not guaranteed to find the true maximum or minimum. It does, however, result in a factor of two approximation.

6 Experiment

We use data from the National Longterm Healthcare Survey to illustrate the methodology (a thorough description of the data is found in [3]). This survey involves data gathered on a sample of individuals age 65 and above over 6 waves at roughly 5 year intervals, beginning in 1982 with around 20000 subjects. In subsequent iterations, some subjects had died, and thus new individuals were replace them. Therefore for the individuals captured in the survey, there exists a non-maximal matching between each pair of consecutive waves. We first describe the variables used to construct our confidence set for the matchings, then give some experimental results.

We take four variables from the files: date of birth, sex, state of residence, and the number of the regional office which interviewed the subject. Evidently typographical errors may occur in any of these fields, and people may move between states etc. We take the set D given above in (4), in which d is the hamming distance (i.e., the number of fields in which the records disagree) and $\epsilon = 1$. Thus D consists of those pairs which match exactly and for which there is a unique match. For testing false links we propose a model for the hamming distance between true links which has $\mathbb{P}(d \geq 1) = 0.15$ and $\mathbb{P}(d < 4) = 1$. This means that pairs which disagree on every field are not considered for the linkage. This choice of parameters is meant to be a conservative estimate, since in principle we do not anticipate the error rate to be this high. We take the survey from 1989 and 1994, and discard any subjects which are missing more than two measurements. We thus have files having 17,483 and 19,171 records respectively. Evaluating D, we find that it consists of 9000 pairs, of which 8798 are crossing, 65 are wholly within the 1989 file, and 137 are within the 1994 file. We find c (the number of possible candidate links having zero distance) to be: 9273. We calculate:

$$\tilde{T}^{\star}_{0.975} = 262, \quad S^{\star}_{0.975} = 1723.$$

Hence we obtain a 0.95 confidence set from (7). This is a set which contains at least 8536 of the identified unique pairs, and at most 1723 other edges which do not disagree on every field, and any number of other pairs which do agree on every field. Thus a 0.95 confidence interval for the size of the bipartite matching is $(8536, 10996)$. Examining the keys in the data reveals the size of the true matching to be: 10074, and that the true matching is an element of the confidence set in this case.

7 Conclusion

To summarize, we propose a method for constructing a set of bipartite matchings which (under modest assumptions on the data generating process) contains the true matching with some prescribed probability. The set involves a pair of constraints on the matching which may be useful for convex optimization of functions of the matching constrained to this set.

In the context of privacy the method has at least two possible uses,. First, it permits inference of combined databases in the absence of unique identifiers, e.g., when the data is somehow anonymized. Second, we can use the method to decide whether data has been suitably anonymized, by examining the confidence set obtained by matching the anonymized data to some second dataset which contains identifiers.

What remains to be seen is how domain knowledge may be useful in extending our method. For example, suppose we know that certain individuals in the data may never participate in links, e.g., in the context of the experiment above, the cohort brought in to replace the dead individuals will have no match in the prior wave. We can also incorporate other constraints such as blocking. Finally there is the prospect of extending the approach to the situation involving matching multiple files simultaneously.

References

1. Mikhail Bilenko, Raymond J. Mooney, William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. Adaptive name matching in information integration. *IEEE Intelligent Systems*, 18(5):16–23, 2003.
2. Cynthia Dwork. Differential privacy: A survey of results. In Manindra Agrawal, Ding-Zhu Du, Zhenhua Duan, and Angsheng Li, editors, *Theory and Applications of Models of Computation, 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings*, volume 4978 of *Lecture Notes in Computer Science*, pages 1–19, Berlin, 2008. Springer.
3. Elena Erosheva, Stephen E. Fienberg, and Cyrill Joutard. Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics*, 1(2):502–537, 2007.
4. Rob Hall and Stephen E. Fienberg. Privacy-preserving record linkage. In J. Domingo-Ferrer and E. Magkos, editors, *Privacy in Statistical Databases (PSD 2010)*, volume 6344, pages 269–283, Berlin, 2010. Springer.
5. Thomas N. Herzog, Fritz J. Scheuren, and William E. Winkler. *Data Quality and Record Linkage Techniques.* Springer, 1st edition, 2007.
6. Partha Lahiri and Michael Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2002.
7. Andrea Tancredi and Brunero Liseo. A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, 5(2B):1553–1585, 2011.
8. R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing Theory and Applications*, pages 1–64. Cambridge University Press, 2010.
9. William E. Winkler. Matching and record linkage. In *Business Survey Methods*, pages 355–384. Wiley, 1995.
10. William E. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census, 1999.