# Valid Statistical Inference on Automatically Matched Files

Rob Hall

November 22, 2011

**Abstract**

A process for determining a confidence set for an unknown bipartite matching is developed. It requires only modest assumptions on the nature of the distribution of the data. The confidence set is relaxed as a set of linear constraints on the bipartite matching, which permits efficient analysis of the matched data (e.g., linear regression), while maintaining the proper uncertainty about the linkage itself. An experiment using the National Long-Term Care Survey is performed to demonstrate the validity of the approach.

## 1 Introduction

Record linkage is n 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 [3, 1, 8, 9]. 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.

Here we focus on the second purpose, where the goal is to obtain a valid statistical inference in the face of the unknown linkage structure between the files. 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 was to determine the number of matching records. By taking the minimum and maximum number of links among all linkage structures in the confidence set, a valid confidence interval for the number of matching records is obtained. If the problem is to regress some variable in one file against covariates in another file, then by iterating over every linkage in the confidence set, computing the regression on the linked data, and then taking the maximum and minimum, a confidence interval for the regression coefficient is obtained. 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.

### 1.1 Related Work

The problem of performing a valid statistical analysis between two files which require matching was considered in [4]. 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 [3], 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 [6]. 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 [6] to intervals for regression coefficients, among others.

## 2 Nonparametric Model

We next give the model which we use for the remainder of this work. 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, we thus consider having the 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). In the interests of space we denote: $\vec{x} = (x_1, \ldots, x_m)$, $\vec{y} = (y_1, \ldots, y_n)$. We consider these two sets of observations as nodes of a graph, in which the goal is to determine a bipartite matching between the sets. A matching $\Pi$ 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 $\Pi$ is called a "perfect matching." We consider $\Pi$ 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 $\Pi$, 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.

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

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

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

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

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 presented below. Specifically because it means that the sufficient statistics have a particular structure explained below. This type of assumption is reasonable when e.g., the same agency is responsible for taking all the measurements. However in the case that the two files arise from different agencies then 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-erroneous measurement. However, extensions such as this are left for future exploration.

We require that $\mathcal{X}$ be equipped with a ordering denoted by $<$. Then we consider the sample put in order with respect to the assignment. We take the pairs of $\Pi$, 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 to the above model. Consider the case when the elements take different values almost surely (e.g., the case of real valued measurements), we have:

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

Where the first term is the reciprocal of the number of ways in which the sufficient statistics may be re-arranged into a sample $\vec{x},\vec{y}$, 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 $\vec{x}$ and which to $\vec{y}$. The singletons of $B$ are divided up into sets of size $S_X, S_Y$ hence the binomial coefficient appearing in the expression. Considering $\vec{x},\vec{y}$ as a rearrangement of the sufficient statistics $A, B$ is the heart of the hypothesis testing approach described below.

# 3 Testing a Bipartite Matching

Before proceeding we note that in essence there are two types of null hypotheses we could consider for the problem of testing a bipartite matching:

(a) $H_0 : \Pi \subseteq \Pi_0$ vs $A : \Pi \not\subseteq \Pi_0$.

(b) $H_0 : \Pi_0 \subseteq \Pi$ vs $A : \Pi_0 \not\subseteq \Pi$.

The rejection of these null hypotheses correspond to the case that (a) $\Pi_0$ does not contain all the links (i.e., there are false non-links) and (b) $\Pi_0$ contains false links. Evidently, if we were to have valid tests for both of these hypotheses, then we could construct a test for the null $\Pi = \Pi_0$ by taking the union of the rejection regions. We next show that under the above model, there is no "good" test for (b) (in the sense that the power of any test is bounded near its size).

First suppose that there exists a non-randomized test for (b), which is defined as:

$$\Psi_{\Pi_0}(\vec{x},\vec{y}) = \begin{cases} 1 & H_0 \text{ is rejected} \\ 0 & \text{o/w} \end{cases}$$

Although we use a non-randomized test for the purposes of convenience, the below argument would hold equally well for a randomized test. Further, suppose that there exist some density $f$ for which the power of said test is uniformly bounded from below:

$$\forall \Pi \not\supseteq \Pi_0 : \int_{\mathcal{X}^{n+m}} \Psi_{\Pi_0}(\vec{x},\vec{y})\ dP_\Pi(\vec{x},\vec{y}) \geq \beta$$

Where $dP_\Pi(\vec{x},\vec{y})$ is the density over the vectors $x, y$ which factorizes according to $\Pi$ as in (1). We define:

$$\tilde{f}(x,y) = \lambda f(x,y) + (1-\lambda) g(x) g(y)$$

Further define $d\tilde{P}_\Pi(\vec{x},\vec{y})$ to be the joint density on the vectors $\vec{x},\vec{y}$ arising from $\tilde{f}$ and the factorization according to $\Pi$. We define the conditional distribution over assignments for the purposes of this argument:

$$q(\Pi'|\Pi) = \prod_{(x,y)\in\Pi'} \lambda \prod_{(x,y)\in\Pi\setminus\Pi'} (1-\lambda)\mathbf{1}\left\{\Pi' \subseteq \Pi\right\}$$

We then have that:

$$\sum_{\Pi'\subseteq\Pi} dP_{\Pi'}(x,y) q(\Pi'|\Pi) = d\tilde{P}_\Pi(x,y)$$

Therefore:

$$\begin{aligned}
&\sup_f \max_{\Pi\subseteq\Pi_0} \int_{\mathcal{X}^{n+m}} \Psi_{\Pi_0}(\vec{x},\vec{y})\ dP_\Pi(\vec{x},\vec{y}) \\
&\quad\geq \int_{\mathcal{X}^{n+m}} \Psi_{\Pi_0}(\vec{x},\vec{y})\ d\tilde{P}_{\Pi_0}(\vec{x},\vec{y}) \\
&\quad= \sum_{\Pi\subseteq\Pi_0} q(\Pi|\Pi_0) \int_{\mathcal{X}^{n+m}} \Psi_{\Pi_0}(\vec{x},\vec{y})\ dP_\Pi(\vec{x},\vec{y}) \\
&\quad> \sum_{\Pi\subset\Pi_0} q(\Pi|\Pi_0) \int_{\mathcal{X}^{n+m}} \Psi_{\Pi_0}(\vec{x},\vec{y})\ dP_\Pi(\vec{x},\vec{y}) \\
&\quad\geq \left(1-\lambda^{|\Pi_0|}\right)\beta
\end{aligned}$$

Thus as $\lambda$ approaches zero, we find that the size of the test is bounded from below by a quantity approaching $\beta$. Therefore all the tests will have a power very close to their size, which suggests that such tests will be essentially useless in determining $\Pi$. Whats more, note that this result could be strengthened, so that $\lambda$ is chosen in order that $q$ puts more mass on the $\Pi$ where the power of the test is the greatest (for example in the case when the power depends on $|\Pi_0 \setminus \Pi|$).

Clearly, the above proof would fail whenever it was known that $f$ was bounded away from its own product of marginals (e.g., in terms of KL-divergence). Therefore if there is any hope of rejecting the assignments on the basis of false links then such a restriction must be made to the space of $f$. This is problematic since as the example demonstrates, the size of the test will depend on how such a restriction is made. This result is reminiscent of negative results for detecting edges of graphical models (see e.g., theorem 1 of [5]), where the same problem arises: that the joint distribution may come arbitrarily close to a product of marginals.

In the interest of conceptual simplicity, and practical applicability, we propose a restriction to the model which yields a parametric hypothesis test for this problem. However we first proceed and show that we may reject an assignment on the basis of false "non-links" without needing to deviate from the above non-parametric setting.

# 4 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 the conditional distribution given the sufficient statistics to the model may be readily evaluated. This way, rejection regions may be calculated despite the lack of knowledge of the density $f$ from which the sample is generated. Such a statistic is one which depends on which rearrangement of the sufficient statistics is chosen to yield the observations. A statistic which is e.g., 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 $T(\Pi_0) = T(\Pi_0, D, \vec{x}, \vec{y})$ 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 $\vec{x}, \vec{y}$.
3. Compute the distribution of $T$:

$$P_{\Pi_0}(T(\Pi_0) = t | A(\Pi_0), B(\Pi_0)) = \sum_{\vec{x},\vec{y}} \mathbf{1}\left\{T(\Pi_0, D, \vec{x}, \vec{y}) = t\right\} P_{\Pi_0}(\vec{x}, \vec{y} | A(\Pi_0), B(\Pi_0)) \tag{3}$$

4. Reject $\Pi_0$ 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 $\alpha$, due to the sufficiency of $A(\Pi_0), B(\Pi_0)$. Since we evidently achieve the correct false rejection rate for each value of $A, B$ (by the definition of the test), we must 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 $f$.

First we consider the complete graph with $G = (V, E)$, $V = (x_1, \ldots, x_m, y_1, \ldots, y_n)$, $E = \{(u, v) : u, v \in V\}$. 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 evidently 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 $\Pi_0$). 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 $D$:

$$D = \{(u, v) \in E : u = v, u \neq w, v \neq w \ \forall w \in V\} \tag{4}$$

i.e., those pairs which have equal values, and for which no other element has the same value. The edges in this set are therefore all 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 $\Pi_0$. 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 $\Pi_0$. 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 labeled 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-labellings having $T > T_{1-\alpha}$ is at most $\alpha$. A re-labeling of the data corresponding to $\Pi_0$ constitutes setting

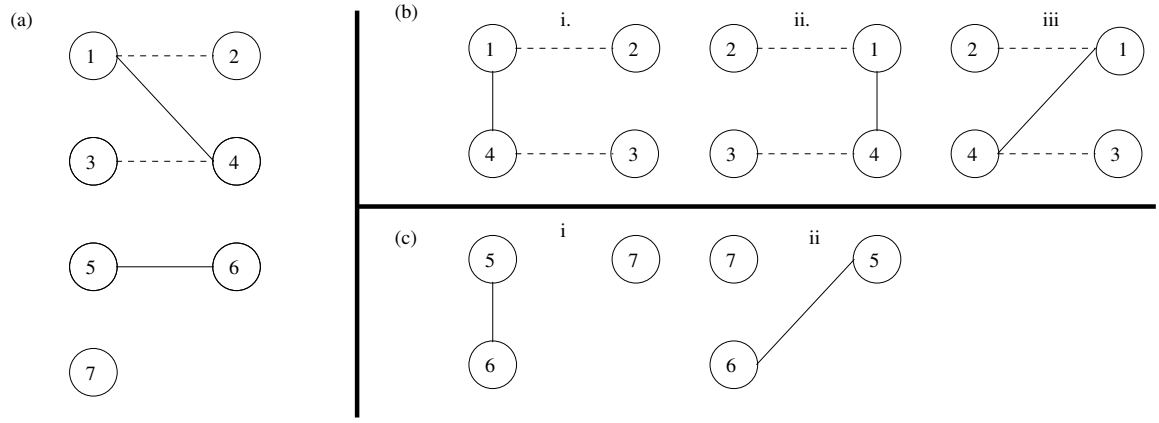


Figure 1: (a) An example of a subset of edges (shown as solid lines) and pairs due to a hypothetical assignment (shown as dashed lines). (b) The rearrangements of the linked pairs. (c) The rearrangements of the singletons. The vertices are numbered in order to make the rearrangement clear. Note that 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 if $\alpha \geq 1/3$.

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 the proposed test.

## 4.1 Computation of the Rejection Region

We note that in principle we may inspect every configuration of the data, evaluate $T$, and construct the distribution of $T$ (conditional on the sufficient statistics). However this is not a computationally efficient method. Due to the restriction put on $D$, that the edges be disjoint, we may compute the rejection region without resorting to full enumeration of the configurations of the data. We first demonstrate how this is done in the case when $\Pi_0$ is a maximal matching (i.e., that each $x_i$ participates in a link), then proceed for general bipartite matchings.

### 4.1.1 Maximal Matchings

We begin by showing that the number of crossing edges from the different connected components formed by the edges $D \cap \Pi_0$ can be considered as independent random variables. We find the distribution for each one, and conclude that the distribution of $T$ conditioned on the sufficient statistics can be computed as the convolution of these random variables. Finally we give a normal approximation to the distributions for which the quantiles are readily available.

First note that for $e \in D$, we have that either:

1. $e \in \Pi_0 \cap D$.
2. $e$ is in a cycle which alternates edges of $\Pi_0$ and those of $D$.
3. $e$ is in a path which alternates edges of $\Pi_0$ and those of $D$, which is not a cycle.
4. $e$ is incident on no edge of $\Pi_0$.

Since $\Pi_0$ is maximal, all the singletons are elements of $\vec{y}$. Therefore an edge $e$ which is incident to no edges of $\Pi_0$ is an edge between singletons which are always labeled as $y_j$. Therefore these edges never cross and never contribute to $T$. Edges in $\Pi_0 \cap D$ are also disjoint from the other edges of $\Pi_0$ and $D$ due to these sets containing only disjoint edges. Since both sets have this property, the only connected components in the graph formed by $(V, \Pi_0 \cup D)$ are those that alternate edges of $\Pi_0$ and $D$, and these connected components are either cycles or "paths" of edges (i.e., a chain of edges which is not a cycle), since each vertex may participate in at most two edges of $\Pi_0 \cup D$. Since the connected components contain disjoint sets of edges of $\Pi_0$, we see that re-ordering the data within one component does not affect the number of crossing edges in any other component. Therefore the number of crossing edges in each one may be treated as independent random variables.

We find the distributions of the number of crossing edges from these kinds of connected components below, where we use $k$ to mean the number of edges of $D$ contained in the component, and $t$ to be the number which cross:

$$f_{\text{path}}(t;k) = 2^{-k}\binom{k}{t}$$

$$f_{\text{cycle}}(t;k) = \begin{cases} 2^{-k+1}\binom{k}{t} & k-t \text{ is even} \\ 0 & \text{o/w} \end{cases}$$

Thus for any hypothetical matching $\Pi_0$ and choice of $D$, one may compute the distribution of $T$ as the convolution of the random variables resulting from the connected components in the graph, and thereupon compute the rejection region. However, we note that both of these random variables fall into a class known as "sub-gaussian" (see e.g., [2]), therefore valid (conservative) rejection regions may be found via a normal approximation to the above distributions. We first give the definition of sub-gaussian and then demonstrate the sub-gaussian nature of the two types of random variables described above. Finally we appeal to previous results about the sum of independent sub-gaussians in order to approximate the distribution of $T$.

**Definition 4.1** (Sub-gaussian)**.** A random variable $X$ having $\mathbb{E}X = 0$ and:

$$\mathbb{E}e^{tX} \leq e^{\sigma^2 t^2/2} \tag{5}$$

Is called sub-gaussian with parameter $\sigma^2$.

Following [2], we use the notation $\sigma^2(X)$ to mean the value of $\sigma^2$ required to fulfill (5) for a sub-gaussian random variable $X$. We first demonstrate the sub-gaussian nature of the binomial distribution after shifting it to have mean zero.

**Proposition 4.2.** *For $Y \sim$ Binomial(n, 1/2) the random variable $X = Y - n/2$ is sub-gaussian with $\sigma^2(X) = n/4$.*

*Proof.* First note that $\mathbb{E}X = \frac{n}{2} - \frac{n}{2} = 0$ also:

$$\begin{aligned} \mathbb{E}e^{tX} &= \mathbb{E}e^{t(Y-n/2)} \\ &= \left(\frac{1}{2} - \frac{e^t}{2}\right)^n e^{-nt/2} \\ &= \left((\frac{1}{2} - \frac{e^t}{2})e^{-t/2}\right)^n \\ &= \left(\frac{e^{-t/2}}{2} - \frac{e^{t/2}}{2}\right)^n \\ &\leq \left(e^{t^2/8}\right)^n \\ &= e^{nt^2/8} \end{aligned}$$

□

Evidently this covers the case of paths. For the cycles, we consider two cases, depending on whether the cycle is of even or odd size. For odd size cycles the proof is dramatically easier since we may appeal to a symmetrized version, which corresponds exactly to the binomial distribution as seen above:

**Proposition 4.3.** *Let $n$ be odd and let $Y$ have the pmf:*

$$f_Y(y) = \begin{cases} 2^{-n+1}\binom{n}{y} & y \text{ is odd} \\ 0 & o/w \end{cases}$$

*Then $X = Y - \mathbb{E}Y$ has $\sigma^2(X) \leq \frac{n}{4}$.*

*Proof.* Note that for all $p \geq 0$:

$$\mathbb{E}|X|^p = \frac{1}{2}\left(\mathbb{E}|X|^p + \mathbb{E}|-X|^p\right) = \mathbb{E}|B|^p$$

Where $B = B' - \mathbb{E}B'$, $B' \sim \text{Binomial}(n, \frac{1}{2})$. Thus lemma 5.5 of [7] gives that $\sigma^2(X) \leq \sigma^2(B) \leq \frac{n}{4}$. □

**Proposition 4.4.** *Let $n > 2$ be even and let $Y$ have the pmf:*

$$f_Y(y) = \begin{cases} 2^{-n-1}\binom{n}{y} & y \text{ is even} \\ 0 & o/w \end{cases}$$

*Then $X = Y - \mathbb{E}Y$ has $\sigma^2(X) \leq \frac{n}{4}$.*

*Proof.* We use the following fact about binomial coefficients which we state without proof:

$$\forall n, \exists k_0 : \binom{2n}{2k} \leq 2^{n-1}\binom{n}{k} \text{ for } k < \frac{n}{2} - k_0 \text{ or } k > \frac{n}{2} + k_0$$

Where the inequality is reversed for $k \in [\frac{n}{2} - k_0, \frac{n}{2} + k_0]$. Take $t_0 \geq k_0$:

$$\begin{aligned} P(|X| > t_0) &= 2\sum_{t=t_0}^{n} \binom{n}{t} 2^{-n+1} 1\{t \text{ is even}\} \\ &\leq 2\sum_{t=t_0}^{n} \binom{n/2}{t/2} 2^{-n/2} 1\{t \text{ is even}\} \\ &= 2\sum_{t=t_0}^{n/2} \binom{n/2}{t} 2^{-n/2} \\ &= P(|Z| > t_0) \end{aligned}$$

Where $Z = Z' - \mathbb{E}Z'$ and $Z' \sim \text{Binomial}(\frac{n}{2}, \frac{1}{2})$. Likewise for $t_0 < k_0$ we may prove the same inequality holds, since $P(|X| < t_0) > P(|Y| < t_0)$. Therefore theorem 5.5 of [7] gives that $\sigma^2(X) \leq \sigma^2(Z)$, the latter which we have already demonstrated to be smaller than $\frac{n}{4}$. □

Finally the case of a cycle of size 2 is handled exactly by example 1 of [2]. In which it is found to have $\sigma^2 = 1$. We make use of two important properties of sub-gaussians which are found in [2]:

**Proposition 4.5.** *For independent sub-gaussian random variables $X_1, \ldots X_n$ we have:*

$$\sigma^2\left(\sum_{i=1}^{n} X_i\right) \leq \sum_{i=1}^{n} \sigma^2(X_i)$$

**Proposition 4.6.** *A sub-gaussian random variable $X$ is majorized by $\mathcal{N}(0, \sigma^2(X))$.*

From the first proposition we find that the sub-gaussian parameter of the total number of crossing edges from all the components is upper bounded:

$$\sigma^2 \leq \frac{|D \setminus \Pi_0|}{2}$$

In which the upper bound is achieved when the components are all cycles of size two. Note that this quantity is also the mean number of crossing edges irrespective of $\Pi_0$ (since for each type of component considered, the mean is the number of edges of $D$ in that component divided by two). Therefore we may use the $1 - \alpha$ quantile of:

$$\mathcal{N}(\frac{|D \setminus \Pi_0|}{2}, \frac{|D \setminus \Pi_0|}{2})$$

as a valid threshold for the above test.

### 4.1.2 General Matchings

In a non-maximal matching there is the prospect that singletons are on both sides of the bipartite graph. Two new kinds of components present themselves in this scenario. First there may exist paths of nodes, in which one or both endpoints are singletons. Secondly there may be edges of $D$ which lay between two singletons. In this section we demonstrate that both such structures have the sub-gaussian property described above, and sketch the proof that neither have sub-gaussian constants larger than half the number of edges of $D$.

Let $P(t|k, a, b)$ denote the probability of $t$ crossing edges out of $k$ edges among the singletons, where $a$, $b$ are the number of singletons among $\vec{x}, \vec{y}$ respectively. We find the recurrence:

$$P(t|k, a, b) = P(t-1|k-1, a, b)\lambda(t, k, a, b) + P(t|k-1, a, b)(1 - \lambda(t, k, a, b))$$

In which $\lambda$ is a function which gives the fraction of the number of pairs consisting of an $x_i, y_j$ among all those pairs of nodes of which neither element is incident on any edge, and when there are $k$ edges of which $t$ are crossing. We thus find that as $k$ increases then the distribution of $t$ gains more central tendency, however its range also increases. The mean of the number of crossing edges is found to be no more than $(n+1)/2$. Letting $Z_n$ denote the random number of crossing edges after having its mean subtracted we have:

$$P(|Z_n| > z) \leq P(|Z_{n-1}| > z - 1)$$

Therefore if $Z_n$ is sub-gaussian for some $n_0$, it will be for all $n \geq n_0$, where the constant obeys:

$$\sigma^2(Z_n) \leq \sigma^2(Z_{n-1})\frac{n}{n-1}$$

This may be observed from theorem 5.5 part (i) of [7]. We find that for $n = 2$ irrespective of $a, b$ we have $\sigma^2(Z_2) \leq 1$ and so conclude that $\sigma^2(Z_n) \leq n/2$ for all $n$.

## 4.2 Efficient Inversion of the Test

So far we have described a method to test a single null hypothesis of the form $\Pi = \Pi_0$, in which the test statistic $T(\Pi)$ was compared against the $1 - \alpha$ quantile of its distribution, which could be computed as shown. When the goal is analysis of the linked data, then it is useful to invert this test, namely to produce a set of all the $\Pi_0$ which are not rejected this way. The above test is time consuming to invert (as stated) since the rejection region for each assignment is potentially different. Therefore we first propose a conservative relaxation of the test, namely one for which the rejection region does not depend on $\Pi_0$, and which only rejects $\Pi_0$ when the above test would. We consider the threshold:

$$T^\star_{1-\alpha} \stackrel{\text{def}}{=} \max_{\Pi} T_{1-\alpha}(\Pi) \text{ s.t. } T(\Pi) < T_{1-\alpha}(\Pi)$$

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 $\Pi_0$, 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 $\Pi_0$ 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:

$$C_{1-\alpha} = \{\Pi : |D \cap C \setminus \Pi| < T^\star_{1-\alpha}\} \tag{6}$$

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).

#### 4.2.1 Computation of $T^\star_{1-\alpha}$

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. From the above considerations we have:

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

In which the last term is the $1-\alpha$ quantile of a Normal distribution with the parameters arising from the sub-gaussian bounds given above in addition to the bounds on the expectation. Here $s \stackrel{\text{def}}{=} |D \cap C^C|$ is the number of non-crossing edges of $D$. Therefore we may take:

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

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 $\Pi$ 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. This 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:

$$\tilde{T}^\star_{1-\alpha} \approx s + 1 + Z^2_{1-\alpha} + \sqrt{\frac{Z^2_{1-\alpha}}{2} + 4Z^2_{1-\alpha} s} \tag{7}$$

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 contained entirely within one of the two sets, and compute $\tilde{T}^\star_{1-\alpha}$ using (7).
3. Construct the confidence set (6).

# 5 Testing for False Links

Note that the above test 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 $\Pi_0$ is never rejected with probability greater than $\alpha$. As 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 general as the above permutation test was, but is tailored 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. Whats more, we suppose to have first constructed the above confidence set using the choice of $D$ given in (4). We consider the hamming distance $d_H(x_i, y_j)$. This is simply the number of measurements on which the records disagree. Suppose that there are $p$ measurements (i.e., "fields or "columns" in the case of a database), then $0 \leq d_H \leq p$ for all $x, y$. Note that $f$ induces a measure over the hamming distance. We propose the following restriction to $f$:

$$\mathbb{P}_f(d_H(x,y) = d) = \begin{cases} 0 & d > d_{\max} \\ \theta_d & 1 \leq d \leq d_{\max} \\ 1 - \sum_{i=1}^{d_{\max}} \theta_d & d = 0 \end{cases}$$

Where $\theta_1, \ldots, \theta_{d_{\max}}, d_{\max}$ are the parameters, and $(x, y) \sim f$. This model states that links have a distance almost surely below $d_{\max}$, and that the probabilities for varying degrees of mismatches are specified. Assuming knowledge of all these parameters appears to be unrealistic, therefore we propose that $d_{\max}$ and $\xi \stackrel{\text{def}}{=} \sum_{i=1}^{d_{\max}} \theta_i$ are known. These correspond to the maximum distance between links, and the probability that linked elements are not exactly the same. These quantities could be determined either from knowledge about the instrumentation that collected the measurements, or from a sample of data in which the matching was known. We will propose a test which will be honest (in the sense that it will deliver the correct $\alpha$-level) whenever an upper bound for these parameters is used (i.e., if one was not certain of either quantity he could use a conservative estimate and anticipate the correct level).

Under the above model, an assignment $\Pi_0$ may be rejected whenever it contains an edge $(x_i, y_j)$ having $d_H(x_i, y_j) > d_{\max}$, since such an edge has zero probability of occurring under the restriction to $f$. let $S(\Pi_0)$ count the number of edges for which the elements have non-zero hamming distance. Under the null hypothesis this is binomially distributed with $|\Pi_0|$ trials and success probability $\xi$. Thus $\Pi_0$ may be rejected whenever:

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

Let $c$ be the size of the largest matching which consists of only pairs which match exactly. Note that this may consist of edges not in $D$ (namely those pairs which match exactly but have multiple potential matches). Then any $\Pi_0$ having $S(\Pi_0) = b$ may be rejected if:

$$b > B_{1-\alpha, b+c, \xi}$$

Since $|\Pi_0| \leq b + c$ and the quantile increases with $|\Pi_0|$. Let $S^\star 1 - \alpha$ be the minimal $b$ for which this inequality holds. Then we have the rejection region: $S(\Pi_0) > S^\star_{1-\alpha}$. The latter quantity can be computed directly using a subroutine to compute binomial quantiles (we do not give a normal approximation since for $\xi$ close to zero the approximation breaks down). Nevertheless this quantity can be computed efficiently.

Finally we have the confidence set of matchings:

$$C'_{1-\alpha-\beta} = \left\{ \Pi : T(\Pi) < \tilde{T}^\star_{1-\alpha}, S(\Pi) < S^\star_{1-\beta} \right\} \tag{8}$$

With coverage probability at least $1 - \alpha - \beta$ due to the sub-additivity of probabilities.

# 6 Statistical Analysis Over the Confidence Set

We next describe a class of analyses which may be carried out efficiently over this confidence set, to obtain extreme values of a statistic of interest. Suppose the 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 $\hat{\beta}(\Pi)$, the typical least squares estimator on the matched data. However since this is unknown we can compute e.g.,:

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

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 $\beta$, 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:

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

Where $h$ is some function of the linked data elements under the matching $\Pi$. Examples that fit into this framework are e.g., estimation of covariance (in which $h$ gives the product of the regressor and response variable), and estimation of a histogram (in which case $h$ is the indicator of some set). This regime also permits linear regression estimation thorugh the function:

$$h(x_i, y_j) = \frac{n}{|\Pi|}(X^T X)^{-1} x_i y_j$$

Then note that $\frac{n}{|\Pi|}(X^T X)^{-1}$ approximates the restriction of $(X^T X)^{-1}$ to those elements which appear in the bipartite matching. In this case the function maps to a vector, and so each coordinate would be maximized and minimized in order to determine a confidence hypercube for the parameters.

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 $C$. There appear to be two viable approaches to this problem. First, for linear statistics of the data as considered above, and when the size of the matching is fixed, the optimization becomes a linear program. Therefore we may consider solving a linear program for each size of the matching (or rather to binary search on the size of the matching for the optimum size). Although this would be an exact method it may be time consuming. An alternative approach would be to greedily build the matching, starting with those edges which are required by the constraints of the confidence set, then adding whatever other edges may increase or decrease the average weight. Although this technique is conceptually appealing it unfortunately leads to only a 2-approximation, meaning that the resulting interval must be dilated by a factor of two in order to make it valid.

# 7 Experiment

We use data from the National Long-Term Care Survey [1]. This is a survey taking measurements of elderly subjects at five year intervals. It began in 1982 with around 20000 subjects included in the first survey. In subsequent iterations, some subjects had died, and so new people were surveyed to replace them. Therefore there exists a

[1] see e.g., `http://www.nltcs.aas.duke.edu`

non-maximal matching between each pair of consecutive waves of the survey. This matching is given in the data files, however we ignore it for the purposes of estimating the set of matchings, and use it only to evaluate our method to assess its correctness. 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 individuals may move between states. Therefore there are true links for which the records contain differing values. We also find pairs of different individuals which have equal values for all attributes. 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.

We take the set $D$ given above in (4), and take $\xi = 0.15$, $d_{\max} = 3$. 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. 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:

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

Whence we obtain a 0.95 confidence set from (8). 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.

# 8 Conclusion

We have demonstrated a pair of techniques which when used in tandem result in an appealing confidence set for the bipartite matching parameter in a two-file record linkage problem. Approximations to the rejection regions were given in order that they may be computed analytically. Remarks were given about proper techniques to perform statistical analysis with respect to this confidence set. Finally the approach was demonstrated to be useful on a moderate size problem using real data.

In moving forwards we plan to extend this work to allow the statistician to inject his own domain knowledge into these hypothesis tests. For example it may be known that certain individuals will never participate in links (e.g., the new cohorts brought in in each wave of the NLTCS data), and these individuals may be identified by their attributes. We also plan to compare our technique with techniques which produce point estimates of the bipartite matching structure, to find whether they tend to produce answers within the confidence set or not.

# 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] V. V. Buldygin and Yu. V. Kozachenko. Sub-gaussian random variables. *Ukrainian Mathematical Journal*, 32:483–489, 1980. 10.1007/BF01087176.

[3] Thomas N. Herzog, Fritz J. Scheuren, and William E. Winkler. *Data Quality and Record Linkage Techniques*. Springer, 1 edition, May 2007.

[4] P. Lahiri and M. Larsen. Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230, 2002.

[5] James M. Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90:491–515, 2003.

[6] 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.

[7] R Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed sensing theory and applications Cambridge University Press Submitted*, pages 1–64, 2010.

[8] William E. Winkler. Matching and record linkage. In *Business Survey Methods*, pages 355–384. Wiley, 1995.

[9] William E. Winkler. The state of record linkage and current research problems. Technical report, Statistical Research Division, U.S. Bureau of the Census, 1999.