Census Linking: A Bounds Approach

Arkadev Ghosh, Sam Il Myoung Hwang, Munir Squires*

February 19, 2019

Abstract

Linking historical data at scale typically requires substantial human effort and subjective individual judgement on the quality of links. We propose a method to identify bounds on statistics of interest that requires minimal assumptions. This method is complementary to state-of-the-art approaches to linking census records. We implement our method to compute an upper and lower bound on the migration rate out of Arkansas between the 1850 and 1860 US Census. We implement this both with objective criteria for limiting possible links, and by using a machine learning model trained with limited human RA decisions. We find a lower bound of 38.2% and an upper bound of 49% on outmigration rate from Arkansas between 1850 and 1860, which is higher than existing estimates of inter-state migration in the literature. We discuss why our estimate is larger than past estimates that are typically on smaller samples. We also discuss simulations that mimic the census data to explore sensitivity of bounds under a set of conditions likely to be encountered in applying this method.

*VSE, University of British Columbia. Ghosh: arkadev@mail.ubc.ca; Hwang: hwangii@mail.ubc.ca; Squires: munir.squires@ubc.ca

1 Introduction

The availability of large linked datasets have transformed empirical work in social sciences. With individuals being tracked over time, researchers can now break new ground on topics such as migration, intergenerational mobility of income and education, short and long run effects of different policy measures, and returns to education. Several very high quality longitudinal datasets are now available for many different countries, which can be used to study these topics. However, much less is known about these measures historically, primarily due to the unavailability of such data in the past. Consequently, we lack historical benchmarks against which present estimates can be compared with.

Recent work in historical record linkage has aimed to mitigate this to a large extent. Machine linkage has been critical to many such projects, especially those aiming to link people across U.S. Censuses. Using a combination of machine learning techniques and text analysis algorithms, researchers have been successful in linking records across censuses, which are now being widely used for analysis (see Feigenbaum, 2016, IPUMS, 2015). However, very little is known about the performance of these methods thus far. In a recent paper, Bailey et al. [2018a] evaluate a number of widely used automated record linkage methods. Basing performance on match rate (proportion of people matched), representativeness (unbiased sample), Type I errors (False matches) and Type II errors (True matches not found), they find modest match rates for all methods, none of the methods to consistently produce representative and unbiased samples, and find false match rates of up to 32 percent for certain algorithms.

Researchers are generally interested in matching a subset of the population based on the particular research question. We find that following a consistent set of rules to identify links in a random sample of individuals from the census to be very difficult. One can therefore imagine that the issues related to machine linkage are likely to be exacerbated

significantly, when one is interested in population estimates or in general as the baseline population grows. Such selection effects could strongly bias estimates of the statistic of interest.

In this paper we develop a methodology that estimates upper and lower bounds on a statistic of interest, rather than focus on identifying which links are correct. It allows us to learn about the statistic of interest without using human judgement at a scale significantly greater than present linking methods do. Our method imposes no restriction on sample choice and is easily replicable with sufficient computation power.

Given a set of possible links, our method uses bipartite graphs and a greedy algorithm to find the lower and upper bound of the statistic of interest. We require matchings to be *feasible* and *non-wasteful*. *Feasibility* requires that everyone is either unmatched or matched to one of their potential links and *non-wastefulness* implies that unmatched individuals have no potential links that are themselves unmatched. At the outset, the algorithm can be run with minimal assumptions on identifying candidate links and therefore is not subject to any sample selection and can be performed without human judgement. This also makes our method easily replicable without significant monetary or time costs. We also document the benefits of human intervention, by using RAs to train a model that reduces the size of candidate links allowing us to obtain narrower bounds on the statistic of interest. The extent to which human intervention can help narrow bounds would vary across projects, but the two step process we use can help researchers evaluate the benefit of human intervention against its costs.

We apply our method to estimate migration rate out of the state of Arkansas between 1850 and 1860. As with every other paper in this literature, we restrict attention to males, since tracking females over time is complicated by the high frequency with which female surnames change at marriage.[1] We choose Arkansas because it was a populous state in

[1]The only exception to this that we know of is the IPUMS one percent linked sample, where fewer than

1850 with significant outward migration, and therefore it allows a good test of the performance of our method. (If overall migration is low, bounds are mechanically narrower.) Using our method, estimates of migration rates can be obtained using sample sizes that are, to the best of our knowledge, many times larger than has been possible in the past. We report results both without and with RA input, which allows an evaluation of the value they add.

Because we do not restrict ourselves to finding unique matches for given individuals, we significantly reduce sample selection relative to traditional methods of linking. These methods typically ignore individuals with links that are "ties" (i.e. two equally good candidate links for one person), or randomly declare one of the links as a match, neither of which is desirable especially when the count of such events is high. Furthermore, in being agnostic about the difference in quality of links that are not ruled out, we are able to mitigate concerns of false matches.[2]

To identify potential links between individuals across census rounds, we first use a set of criteria that are meant to rule out obvious non-matches. We do this by following Feigenbaum [2016], where links are ruled out if they have sufficiently dissimilar names, year of birth differences of more than three, or different state of birth.

Using the remaining links as potential matches, we find that the lower bound on migration out of Arkansas between these two census rounds is 39.2%, and the upper bound is 55.3% in a baseline population of 119,200 men. Existing estimates of inter-state migration (for the entire US, not restricted to Arkansas) range from 19 to 28%.

To refine the set of possible matches, we use research assistants to identify links that are less likely to be matches ("non-matches") amongst existing links, in a training dataset.[3]

10% of women are linked, and this is restricted to women whose marital status does not change.

[2]Bailey et al. [2018b] report false match rates of up to 32% in the matching algorithms that they evaluate.

[3]The agreement rate between RAs assigned to identify non-matches (66%) is much higher than when identifying a single match amongst possible candidates (49%)

Using this input and standard machine learning techniques, we train a model that eliminates links out of sample. This allows us to obtain narrower bounds of the outcome variable than would be the case without the input of RAs. One advantage of our method is that we can explicitly quantify the improvement in precision by using these RAs. This would allow researchers to decide whether the gains from hiring RAs to train the model further would justify the cost in time and money. We show that the bounds can be narrowed significantly by using this trained model to refine the set of possible links. We find the lower bound of migration rate out of Arkansas to be 38.3% and the upper bound to reduce significantly to 49% after the refinement.

We also test the performance of our model in simulated data with perfectly known links and show how different factors affect estimate bounds. Specifically we explore how bounds are affected by changes in population size, true share of movers, frequency of common names, measurement error in the recording of names, and cutoff criteria for links being considered obvious non-matches. These simulations provide insight into the overall performance of our method for calculating upper and lower bounds, and suggest in which settings these bounds are likely to be most informative.

The rest of the paper is organized as follows. In section 2, we briefly review the literature on census linking methods and historical migration estimation in the U.S. and discuss the contribution of our paper. In section 3 we describe the model and algorithm used to estimate upper and lower bounds of the statistic of interest. Section 4 discusses the two step linking procedure used to identify potential links for individuals. Section 6 presents simulations documenting how different factors affect estimate bounds. Finally, section 7 concludes.

2 Literature Review

2.1 Census Linking Methods and Contribution

Record linkage is important for many projects, across academic disciplines and within many sub fields of economics, such as economic history, health and development economics. This is complicated when using historical data by the absence of unique identifiers which can be used to track individuals over time. In this section, we briefly review and comment on different methods of historical record linking. Our key contribution relative to this literature is that our method, which estimates bounds of the statistic of interest (instead of a point estimate), suffers from significantly less sample selection issues. It therefore is likely to be more informative of the true value (population estimate) of the statistic, especially as the baseline population grows.

Ferrie's (1996) approach of linking men between the 1850 and 1860 U.S. Censuses is foundational for the literature on historical record linking. He matched individuals with uncommon names between the two census rounds. The choice of uncommon names was made to reduce problems of "ties" while linking (i.e. there could be multiple links that satisfy all the criteria required to be a "good" match, leading to ambiguity over the correct choice). This lead to a linked sample of 4938 men. Using this sample Ferrie estimates a range of factors including migration rates in the U.S. between 1850-1860. However, to the extent that individual names are correlated with the propensity to migrate (for e.g. uncommon names could be determined by current place of residence, which in turn affects migration propensity), Ferrie's estimates may be biased. Furthermore the degree of bias itself is hard to quantify.

Abramitzky et al. [2012a] and Abramitzky et al. [2014] essentially automate Ferrie's method to the full census count. They relax Ferrie's uncommon name selection to the extent that they try to identify links that are unique in name and age combination. Abramitzky

et al. [2012a] and Abramitzky et al. [2012b] link boys aged 3 to 15 with unique name-age combinations in the 1865 Norwegian Census, standardize first and last names phonetically and look for exact and unique matches. They use an algorithm that relaxes the age requirement by an absolute value of 1 iteratively, if they do not find any exact match. The algorithm is terminated at $\pm$ 2 years.

Their method therefore does not link a record if the phonetic (NYSIIS) name codes do not exactly match, if there are two or more links that satisfy every criteria, or if there are no potential links that fall with the age band of $\pm$ 2 years. They obtain a match rate of roughly 29% using this method. In Abramitzky et al. [2014] the authors use the same method, to link men aged between 18-35 from the 1900 U.S. Census to the 1910 and 1920 U.S. Censuses achieving modest match rates of 16 and 12 percent respectively.

IPUMS (2015) and Feigenbaum [2016] take a probabilistic approach to record linking. The key feature of these methods are the fact that the best link does *not* have to match exactly on first or last names and date of birth, but should "score" high when a combination of matching factors are simultaneously considered. IPUMS uses matches identified by humans to train a Support Vector Machine (SVM), that links the publicly available 1850-1930 one percent US Census samples. Their method, using a number of tuning parameter choices, classifies all links as either true or false. Their match rates range between 6-13 percent depending on individual characteristics (native-born whites, foreign born whites, African Americans etc). In a similar vein, Feigenbaum [2016] uses a probit model, trained on clerically reviewed data to quantify the empirical importance of key variables in identifying matches. The model is then tuned to flag a link as a match only if its probability (probit score) of being a match is high enough and the score of the second best link is low enough (conditional on the fact that it exists). He achieves a high matching rate of 57%.

Bailey et al. [2018a] provide an assessment of how well these methods perform on four

different historical datasets with known links. They assess performance based on match rate (proportion of baseline population that could be matched), representativeness (comparing characteristics of linked sample to unlinked sample), false link rate (Type I error) and false negative rate (fraction of true links not found, Type II error rate). They find that none of the linking methods consistently produce representative samples (including hand-linking), automated linking methods produce false match rates of up to 32 percent, and that false links are systematically correlated with different baseline characteristics (selection). They further validate their findings in simulated data with known links, to account for the fact that the datasets they use might not be free of error. [4]

Our paper starts from the idea that identifying true links for all individuals in a given population (or sub-population) is difficult if not impossible. Our primary contribution to this literature is to propose a model to calculate the lower and upper bounds of the statistic of interest, under the assumption that each of the potential links are "equally good"[5]. This method can be applied at a scale that to the best of our knowledge has not been possible before.

2.2 Historical Migration in the U.S.

Longitudinal evidence on 19th century mobility in the U.S. is surprisingly lacking. Economists working with contemporary data stress on the importance of geographic mobility on occupational and financial mobility. They particularly emphasize how selection into migration and the time taken by migrants to adjust in their new community are best observed at the individual level rather than at the aggregate. While longitudinal data now exist for the U.S. population, such as National Longitudinal Surveys (NLS) or the Panel Study of Income Dynamics (PSID), the lack of historical panel data makes it difficult to do this for

[4] They find some linking algorithms attenuate measured intergenerational income elasticity by 50%.

[5] We take different approaches when it comes to restricting the set of links that we consider to be "equally good", as described below

anything but the recent past.

Steckel [1988] studies migration patterns of 1600 rural families matched in manuscript schedules of the 1850 and 1860 Censuses. They use a child's state of birth reported in 1860 as a pointer for the residence of the family in 1850. Therefore, they restrict attention to families with at least one native born child aged 10 years or older in 1860. They generate a mobility matrix classifying U.S. states into four main regions (Northeast, North Central, South and West, according to ICPSR definitions) and estimate across- and within-region migration. Ferrie [1996] uses his linked sample of 4938 men between 1850-1860 to estimate an inter-state migration rate of 28%, which is higher than 19% estimated by Steckel [1988]. Schaefer [1985] also estimates inter-county migration rates in portions of the South using a linked sample of men between 1850-1860.[6] He estimates an inter-county migration rate of 27.7%, significantly lower than 55% estimated by Ferrie. Ferrie argues that by focusing on already-established families Schaefer and Steckel grossly underestimate migration rates. However, as mentioned earlier Ferrie's method of restricting attention to uncommon names is likely to lead to selection problems of its own.

Owing to limitations with respect to linking individuals across censuses, more recent works that study historical migration patterns at the population level focus simply on the cross-section of individuals in a given census. Rosenbloom and Sundstrom [2003] document long-run trends (1850-1990) in inter-state migration using two main measures. First, they consider an individual to have moved if they reside in a different state than where they were born. The second considers a family to have moved if they reside in a different state than the state of birth of one of its young children. The first measure is silent about the timing of moves, as well as intermediate moves. While the second

[6]He draws individuals from the Parker-Gallman sample to establish links. By identifying the head of a household in the Parker-Gallman sample, he locates the family in the 1860 Census of Population. (Done by locating the person's farm in the 1860 Manuscript Census of Agriculture.) Thereafter, using computerized census indexes and family characteristics, he matches the family back to the 1850 Census schedule.

measure captures the timing of the move, it restricts the sample to families with young children, leading to potential selection bias.

Precisely studying historical migration rates over time helps address a range of important questions: Did internal migration redistribute labor from low to high wage states? How did differences in government transfers across locations affect migration rates? Overall, how important was internal migration in overall economic convergence in the 19th century? Using our proposed method, reasonably narrow bounds on migration rates can be obtained at any geographical unit of individual residence, which can help answer these questions with more precision than has been historically possible.

3 Theoretical framework

In this section we propose and justify the use of algorithms to compute the upper and lower bounds of statistics of interest given many possible matches. We do this by relying on two key assumptions: feasibility and non-wastefulness, without which we believe we could not obtain non-trivial bounds. While the algorithm for obtaining the upper bound is exact, the one for the lower bound relies on an approximation. We show in simulations that this approximation is likely to be very precise.

3.1 Model

Let I denote the set of people living in a country (e.g., the U.S.) in Census year y and J denote the set of people living in the same country in Census year y'. We can partition the set I into the following two subsets:

1. I_1 : those who survive the years between the two Censuses, hence $I_1 \subseteq J$.

2. I_2 : those who either died or moved out of the country between the two Censuses,

hence $I_2 \not\subset J$.

Similarly, we can partition the set J into the following two subsets:

1. J_1 : those who are present in both Censuses, hence $I_1 = J_1$.
2. J_2 : those who either are born or moved into the country between the two Censuses, hence $J_2 \not\subset I$.

Let $\mu^* : I \cup J \to I \cup J \cup \{\emptyset\}$ denote the true matching function between I and J. That is, for each $i \in I$, if $\mu^*(i) \in J$, it indicates that person i in Census year y is the same person as $\mu^*(i)$ in Census year y'. If $\mu^*(i) = \emptyset$ for some i, person i either died or moved out of the country between the two Censuses. Similarly, for some $j \in J$, $\mu^*(j) \in I$ means that $\mu^*(j)$ and j are the same person in the two Censuses, and $\mu^*(j) = \emptyset$ indicates that either j was born between the two Census years or migrated into the country.

Suppose that researchers are interested in whether an observable characteristic of people in the country changed between the two Census years. For example, they may be interested in the number of people who changed state of residence between the two Census years. More precisely, let s_i and s_j denote the state of residence of person $i \in I$ and $j \in J$, respectively. Then researchers may be interested in the following quantity:

$$\sum_{i \in I_1} \mathbb{1}\{s_i \neq s_{\mu^*(i)}\} \tag{1}$$

where $\mathbb{1}\{\cdot\}$ denotes an indicator function equal to 1 if the statement inside the bracket is true, 0 otherwise. Other quantities of interest analogous to (1) may be the number of people who change occupations or marital status. Throughout the paper, we use the quantity (1) as our primary example.

In many cases, researchers do not observe the true matching function μ^*. They instead have access to some variables that help them partially identify the true match of each

person in I. Section 4 describes such a method, which can be used to generate a set of potential links in J for each person i. Let $J(i)$ denote the set of people in J that are potential matches for person i; conversely, let $I(j)$ denote the set of people in I for whom j may be a potential match. Note that for all $i \in I$ and $j \in J$, $j \in J(i)$ iff $i \in I(j)$.

We first restrict the set of matching functions with which we compute the bounds on (1), using two key assumptions on how matches are chosen. Given $J(i)$ and $I(j)$ for all $i \in I$ and $j \in J$, there may be many matching functions that map each person in I either to one of his/her potential matches or an empty set (if they died or moved out of the country). We call such a matching *feasible*, i.e., a matching function $\mu : I \to J \cup \{\emptyset\}$ is feasible if for all $i \in I$, $\mu(i) \in J(i) \cup \{\emptyset\}$. Our first assumption about the true matching is that it is feasible:

Assumption 1. *The true matching is one of the feasible matchings.*

The set of feasible matchings include ones under which some people in I are unmatched [7] even though one or more of their potential matches in J remain unmatched. In fact, even the most trivial matching in which everyone in I is matched to an empty set (and the same for everyone in J) is feasible. We refer to matchings with such a property as *wasteful*, i.e., if under a matching function μ, there exists an $i \in I$ such that $\mu(i) = \emptyset$ while there exists $j \in J(i)$ such that $\mu(j) = \emptyset$, then we call μ wasteful. Without an assumption about how wasteful the true matching is, the lower bound on the quantity of interest cannot be identified. Therefore, we make the following assumption about the non-wastefulness of the true matching:

Assumption 2. *The true matching is non-wasteful.*

Note that Assumption 2 allows death or emigration of people in I and birth or immigration of people in J. For example, suppose there are two people in I, say i_1 and i_2,

[7] That is, they are treated as having died or migrated out of the country.

whose $J(i_1)$ and $J(i_2)$ contain one and only one person in J, denoted by j', (and an empty set), and they happen to be the same person, i.e., $J(i_1) = J(i_2) = \{j', \emptyset\}$. Suppose further that j' is not a potential match for any other person in I. Then at any feasible matching, at least one of them will be matched to the empty set, i.e., he/she will be treated as dead or emigrated in between two Census years.

Under Assumptions 1 and 2, the true matching μ^* is contained in the set of feasible and non-wasteful matchings. Let $\mathbb{M}$ denote the set of all feasible and non-wasteful matchings. Then we can construct the bounds on the number of cross-state migrations as follows:

$$\min_{\mu \in \mathbb{M}} \sum_{i \in I, \mu(i) \in J} \mathbb{1}\{s_i \neq s_{\mu(i)}\} \leq \sum_{i \in I_1} \mathbb{1}\{s_i \neq s_{\mu^*(i)}\} \leq \max_{\mu \in \mathbb{M}} \sum_{i \in I, \mu(i) \in J} \mathbb{1}\{s_i \neq s_{\mu(i)}\} \tag{2}$$

The upper bound in (2) can be computed via the following algorithm:

Algorithm 1. *Upper bound*

1. *Create a bipartite graph $G = (I, J, E)$ where E denotes the set of edges between I and J. For each $i \in I$ and $j \in J$, there is an edge (i, j) between them iff $i \in I(j)$ and $s_i \neq s_j$.*

2. *Find a maximum matching [8] of bipartite graph G using an existing algorithm such as the one proposed in Hopcroft and Karp [1973]. Denote the maximum matching with $M^{max} = (I, J, E^{max})$, where E^{max} denotes the set of edges in the matching M^{max}. Then assign the size of the matching M^{max} (i.e.,$|E^{max}|$) to the upper bound in (2) and terminate the algorithm.*

Algorithm 1 results in the desired upper bound in the inequality (2), as we formally state and prove below:

Proposition 1. *The size of M^{max} is equal to the upper bound in the inequality (2).*

See Appendix A for proof.

[8] In graph theory, a matching in graphs (bipartite or otherwise) is defined to be a set of edges without common vertices. A Maximum matching is a matching that contains the largest possible number of edges.

As for computing the lower bound in the inequality (2), a naïve way to compute it may be to use an algorithm similar to 1 to find the maximum number of those who did *not* migrate across states, or *stayers*, then subtract it from the number of people in I. More specifically, the following algorithm may be considered:

Algorithm 2. *Lower bound*

1. *Create a bipartite graph $G = (I, J, E)$ where E denotes the set of edges between I and J. For each $i \in I$ and $j \in J$, there is an edge (i, j) between them iff $i \in I(j)$ and $s_i = s_j$.*

2. *Find a maximum matching of bipartite graph G using an existing algorithm such as the one proposed in Hopcroft and Karp [1973]. Denote the maximum matching with $M^{min} = (I, J, E^{min})$, where E^{min} denotes the set of edges in the matching M^{min}. Then assign the size of the set I minus the size of the matching M^{min} (i.e.,$|E^{min}|$) to the lower bound in (2) and terminate the algorithm.*

However, Algorithm 2 may not result in the exact lower bound. For example, suppose that there are two states, "State 1" and "State 2". In this example, the primitives of the model, namely, $I, J, J(\cdot), (s_i)_{i \in I}$, and $(s_j)_{j \in J}$ are specified as follows:

- $I = \{i_1, i_2, i_3\}, J = \{j_1, j_2, j_3\}$
- $J(i_1) = J(i_2) = \{j_1, \emptyset\}$,$J(i_3) = \{j_3, \emptyset\}$
- $s_{i_1} = 1, s_{i_2} = 1, s_{i_3} = 2$
- $s_{j_1} = 1, s_{j_2} = 2, s_{j_3} = 2$

The maximum number of stayers in this example is two; such maximum is achieved at either of the following two matching functions, denoted by μ_1 and μ_2:

1. $\mu_1(i_1) = j_1, \mu_1(i_2) = \emptyset, \mu_1(i_3) = j_3, \mu_1(j_2) = \emptyset$

2. $\mu_2(i_1) = \emptyset, \mu_1(i_2) = j_1, \mu_1(i_3) = j_3, \mu_2(j_2) = \emptyset$

The minimum number of movers in this example is 0; however, Algorithm 2 would assign a value of 1 to the lower bound, since the maximum number of stayers is two, and the algorithm assigns the size of the set I less the maximum number of stayers ($|I| - 2 = 1$) to the lower bound. Therefore, this example shows that Algorithm 2 does not necessarily result in the lower bound of interest.

As far as we know, there exists no known existing algorithm to compute the exact lower bound, other than exhaustive search. (That is, finding the minimum by computing the number of movers at each matching function in the set of all feasible and non-wasteful matching functions $\mathbb{M}$.) However, Census data are generally too large for such a brute-force approach. Instead, we propose an alternative greedy algorithm to approximate a lower bound as follows:

Algorithm 3. *Lower bound approximation*

Step 1 *Initialize the algorithm by setting a set* $U = \{(i, j) \in I \times J : i \in I(j) \text{ and } j \in J(i)\}$ *and two sets* $C = E = \emptyset$. *In addition, for each edge* $(i, j) \in U$, *assign "cost" as follows:*

- $w_{ij} = 1$ *if* $s_i \neq s_j$
- $w_{ij} = 0$ *if* $s_i = s_j$

Step 2 *For each edge* $(i, j) \in U$, *let* S_{ij}^1 *denote the set of edges in* U *that are incident to* i *or* j *or both and whose cost is equal to 1, i.e.,*

$$S_{ij}^1 = \{(i', j') \in U : i' = i \text{ or } j' = j \text{ or both and } w_{i'j'} = 1\}$$

In addition, let S_{ij}^0 *denote the set of edges in* U *that are incident to* i *or* j *or both and whose*

cost is 0, i.e.,

$$S_{ij}^0 = \{(i', j') \in U : i' = i \text{ or } j' = j \text{ or both and } w_{i'j'} = 0\}$$

Step 3 *Choose the edge in U with the following characteristics:*

1. *The edge with the lowest value of cost*
2. *If there are multiple such edges in U, then choose the edge with the largest cardinality of the cardinality of S_{ij}^1*
3. *If there are multiple such edges in U, then choose the edge with the smallest cardinality of S_{ij}^0*
4. *If there are multiple such edges in U, then break the ties randomly.*

Denote the chosen edge with (i^*, j^*).

Step 4 *Update the sets U, C, and E as follows:*

$$\begin{aligned} U &\leftarrow U - \{(i^*, j^*)\} \cup S_{i^*j^*}^1 \cup S_{i^*j^*}^0 \\ C &\leftarrow C \cup \{(i^*, j^*)\} \cup S_{i^*j^*}^1 \cup S_{i^*j^*}^0 \\ E &\leftarrow E \cup (i^*, j^*) \end{aligned}$$

Step 5 *If $U = \emptyset$, then assign to the lower bound of the inequality (2) the cardinality of the set E and Terminate the algorithm Otherwise, go back to Step 2.*

We can construct a matching function, call it μ^{min}, associated with the resulting set of edges E at the termination of Algorithm 3: for each $i \in I$, $\mu(i) = j$ iff $(i, j) \in E$; for each $i \in I$ and $j \in J$, $\mu^{min}(i) = \emptyset$ and $\mu^{min}(j) = \emptyset$ if there exists no edges in E that is incident

to i or j. It can be shown that μ^{min} is a well-defined matching function, and a feasible and non-wasteful one. We explain these facts below in turn.

μ^{min} is well-defined matching function, i.e., none of people in I or J are matched to more than on person in J or I under the matching function μ^{min}. The reason is that each element in the set E is always chosen from the set U in Algorithm 3; whenever an edge, say (i^*, j^*), is chosen from U to be included in E, all edges in U that are incident to i^* or j^* or both are removed (see Step 4 of Algorithm 3); therefore, once an edge (i^*, j^*) is chosen, it is impossible for an edge incident to i^* or j^* can be chosen again to be included in E.

The matching function μ^{min} is feasible since the resulting set of edges E is a subset of U, which is a set of all edges $(i, j) \in I \times J$ such that $j \in J(i)$ and $i \in I(j)$. Lastly, the fact that μ^{min} is non-wasteful can be shown by way of contradiction. Suppose, by way of contradiction, that μ^{min} is wasteful. Then there exist a person $i \in I$ that is matched to an empty set under μ^{min} and a person $j \in J(i)$ that is also matched to an empty set. But the fact that Algorithm 3 terminated means that (i, j) was excluded from the set U after an edge incident to either i or j (but not both) was chosen to be included in E. Therefore, either i is matched to some $j' \in J(i) - \{j\}$ or j is matched to some $i' \in I(j) - \{i\}$, which is a contradiction.

Intuitively, Algorithm 3 is likely to result in a small number of movers because of the way it chooses the edge to include in the set E: at any iteration in the algorithm it greedily chooses the low cost edges (i.e., edges that increase the number of stayers) that are incident to as many high cost edges (i.e., edges that increase the number of movers) as possible. Once an edge is chosen, edges that are incident to one of the end points of the chosen edge will be removed from consideration in the next iteration. Therefore, when all the low-cost edges are depleted in the set U and therefore we are forced to choose a high cost edge, there will not be many such edges to begin with.

Unfortunately, we were not able to prove the approximation ratio of Algorithm 3,

i.e., an upper bound on the ratio of the lower bound obtained by Algorithm 3 to the actual lower bound (i.e., left hand side of the inequality (2)). Instead we measure how well Algorithm 3 approximates the lower bound in an experiment with simulated data. These simulated data are constructed in such a way that the exact bound can be easily calculated. Therefore, we can measure the performance of our algorithm by comparing its output with the exact lower bound.

To simulate the data sets for our experiments, we consider the following setup:

- There are more people in the latter Census than in the previous Census, i.e., $n \equiv |I| < |J| \equiv m$
- There are two states, "state 1" and "state 2". In Census year y, everyone lives in state 1, i.e., for all $i \in I$, $s_i = 1$; in Census year $y' > y$, a fraction $\alpha \in (0,1)$ of people in J live in state 1, and the rest live in state 2.
- Each person in I can be matched to anyone in J, i.e., for each $i \in I$, $J(i) = J \cup \{\emptyset\}$ (consequently, for each $j \in J$, $I(j) = I$).

Under this specification, the minimum number of movers over the set of all feasible and non-wasteful matching functions is equal to $n - \alpha m$ (ignoring the constraint that αm is an integer). To see why, note first that everyone in I is matched with someone in J in any non-wasteful matching. Then we can minimize the number of people in I who are matched to people in J who live in state 2 by assuming that there are $m - n$ immigrants into state 2 between the two Census years (or alternatively, we can assume that there are $m - n$ births in state 2). Such assumption would leave $(1 - \alpha)m - (m - n) = n - \alpha m$ people in J who live in state 2 and who needs to be matched to someone in I.

We experiment with different values of the parameters n, m, and α and compare the resulting lower bound from Algorithm 3 to the exact lower bound, i.e., $n - \alpha m$. We find that for all parameter values we experiment with, Algorithm 3 produces the same lower

bounds as the exact one (see Table 1 and Table 2). While we cannot generalize our results to cases where we do not know the exact lower bounds, the outcome of these experiments leads us to believe that Algorithm 3 would produce at least a reasonable approximation of the true lower bound in more general cases. Obtaining the approximation ratio for Algorithm 3 is left for future research.

	Exact lower bound	Output from Algorithm 3
α=0.1	85	85
α=0.2	70	70
α=0.3	55	55
α=0.4	40	40
α=0.5	25	25

Table 1: Comparision of exact and calculated lower bounds (n=100,m=150)

	Exact lower bound	Output from Algorithm 3
α=0.1	170	170
α=0.2	140	140
α=0.3	110	110
α=0.4	80	80
α=0.5	50	50

Table 2: Comparision of exact and calculated lower bounds (n=200,m=300)

4 Extracting Links and Algorithm to Identify Non-matches

In this section, we first describe how we obtain potential links for each individual in the baseline population. This provides us with the sets $J(i)$ and $I(j)$ which we then use to find the upper and lower bounds of migration out of Arkansas. We also describe how we train a model using RA input to eliminate "non-matches" which reduces the size of the sets $J(i)$ and $I(j)$ allowing us to obtain narrower bounds.

4.1 Step 1: Extracting Possible Links

Suppose that we are interested in estimating the out-migration rate from state s between census years y and y'. The first step is to identify possible links in census year y' for each individual living in state s in census year y. Using notation used in Feigenbaum [2016] we denote the first set of records as $X1$ and the second set to be $X2$. We restrict possible candidates for each individual in y to those in y' with sufficiently similar first and last names, a year of birth distance of less than or equal to 3 years and the same state of birth following Feigenbaum [2016]. String similarity of names is measured by Jaro-winkler score, a continuous measure of similarity between two string variables that weighs strings that match from the beginning (in terms of individual letters) more heavily. We require scores to be greater than .8 for both first and last names. The state blocking condition reduces the complexity of finding potential matches to a great extent. This is because calculating string distance measures and taking matrix Cartesian products (which is required to compare potential links) are computationally challenging, whereby restricting the set of candidates to those with the same birth state at the outset is helpful.[9]

Therefore, in estimating out-migration from state s in census year y, for each individual i we first restrict attention to those in census year y' born in the same state as i and

[9]State of birth for the same individual could only change between two census years due to enumerator error which is extremely rare [Feigenbaum, 2016]

within a year of birth distance of 3. After eliminating links that do not satisfy the criteria above, we calculate Jaro-winkler similarity scores between first and last names for all remaining links and drop those with scores less than .8 for either first or last name. Since calculating Jaro-winkler similarity scores require the most amount of computation time, following these steps in order speeds up the process greatly. Let us denote this set of all possible matched data as XX. Each record $xx \in XX$, matches an individual from X1 to X2. The number of observations in XX can vary significantly depending on the state of origin under consideration.

At this stage, researchers typically use a subset of XX (say XX_s), where matches have been hand-picked, in order to train a model (matching function) which then identifies matches in the rest of the data. As discussed earlier, instead of training a model to identify matches, we train a model that eliminates links for each individual that are unlikely to be the true match. We discuss this in greater detail in the next section.

4.2 Step 2: Non-Match Function

Constructing a matching function by manually matching a random sample from the Census to train a model is often not straightforward. Table 3 provides an illustrative example. For each individual in 1850 the set of candidates in Table 3 is restricted to males born in the same state (Arkansas in this case) with first and last name jaro-winkler scores greater than .8 and a year of birth difference of less than or equal to 3. At this stage, "human intelligence" is used to select potential matches in a training dataset which can then be used to construct a matching function which identifies out of sample matches. As can be

observed in Table 3, it is not always obvious which (if any) should be the correct match.

Table 3: A sample of potential matches for individuals from 1850 to 1860

Given1850	Surname1850	Yob1850	Given1860	Surname1860	Yob1860
William H	Fowler	1844	William H	Fawler	1844
William H	Fowler	1844	William	Fowler	1844
William H	Fowler	1844	William	Fulmer	1842
William H	Fowler	1844	William	Fowler	1844
Andrew J	Jones	1841	Andrew J	Jones	1844
Andrew J	Jones	1841	Andrew	Jones	1840
Andrew J	Jones	1841	Andrew	Jones	1839
Andrew J	Jones	1841	Andrew J	Jones	1838
Andrew J	Jones	1841	Andrew B	Johnston	1839

Notes: Yob1850 and Yob1860 denote the year of birth provided by each individual in the respective censuses.

It is thus difficult to follow a well-defined rule when deciding matches. On average, our RAs were unable to identify matches for over 30% of individuals in the training data. If individual names are correlated with outcomes of interest, this could lead to substantial selection effects.

However, as can be observed, it is comparatively easier to identify non-matches in this dataset. We should be fairly comfortable in declaring Andrew B Johnston as a non-match for Andrew J Jones and William Fulmer as a non-match for William H Fowler. This is where we use input from research assistants to train a model which identifies non-matches amongst potential links, which we then eliminate in order to obtain narrower bounds for the estimates for migration. Note that non-matches in our setting are defined as links that are comparatively unlikely to be matches.

We use input from three research assistants and use their intersection to train our model i.e. we only declare a non-match if all three RAs agree on it. There could be significant heterogeneity in the rate at which different individuals eliminate links as probable non-matches. Using the combination of inputs from multiple RAs is likely to ensure that the trained model is unaffected by individual idiosyncrasies.

4.2.1 Training the Algorithm

We train a probit model using bootstrap resampling with replacement to obtain the final model coefficients. Table 4 contains the list of variables used as predictors of a non-match. How exactly were these variables chosen? We start with the entire set of predictors used in Feigenbaum [2016] and then select the optimal model using Recursive Feature Elimination method[10] to retain the set of predictors that provide the best performance. Why did we not retain all the predictors used by Feigenbaum [2016]? The process of identifying matches is inherently different from identifying non-matches. For example, one might be worried about the number of links for each individual that are identical in terms of first and last names, when declaring a match[11]. However, in the case of identifying non-matches, since each link is evaluated on its own (de)merit, this variable turns out to be much less important and only adds noise to the model and was eliminated by the process. In Table 5 results from the final trained model that we use to predict non-matches are reported.

We split the training data into two equal halves, train our model on one half (training set) and test its performance on the other (test set). Note that the outcome variable in the regressions are 0 if the RAs coded it as a non-match and 1 otherwise. As expected, coefficients on different measures of string similarity are positive. We then use the trained

[10]We use the caret package in R to do this.

[11]Not being able to declare matches for people with names that are fairly common therefore induces selection, which could potentially bias the outcomes of interest.

Table 4: Variable Description

Variable Name	Description
jwsim_first	Jaro-winkler similarity score of first name
jwsim_last	Jaro-winkler Similarity score of last name
soundex_first	Indicator for first name Soundex code match
soundex_last	Indicator for last name Soundex codes match
hits	Number of links per individual
f_start	Indicator for first letter of first name match
l_start	Indicator for last letter of first name match
f_end	Indicator for first letter of last name match
l_end	Indicator for last letter of last name match
minmatch	Indicator for middle initials match (if any)
yob_diff_i	Indicator for birth of year difference $= i$

model to make out of sample predictions with links that remain after Step 1, and eliminate those that are unlikely to be true matches.

The difference between the performance of the probit and logit models in terms of out of sample prediction accuracy is small. Confusion Matrices in Table 6 show how the models perform on the test sets. In both cases, out of sample prediction accuracy is over 95%. Using the model coefficients, predicted scores are calculated and used to classify links out of sample. We use the probit model, whereby the scores are simply predicted probabilities.

How do we define a threshold value for the predicted scores in order to make our

Table 5: Probit and Logit Models to Identify Non-Matches

	Probit	Logit
jwsim_first	11.14***	22.02***
	(1.05)	(2.11)
jwsim_last	12.03***	22.15***
	(0.90)	(1.73)
soundex_first	0.39***	0.77***
	(0.10)	(0.19)
soundex_last	0.76***	1.51***
	(0.12)	(0.22)
hits	−0.01***	−0.01***
	(0.00)	(0.00)
f_start	0.80***	1.30***
	(0.19)	(0.36)
l_start	0.97***	1.61***
	(0.15)	(0.27)
f_end	−0.02	−0.03
	(0.10)	(0.20)
l_end	0.23**	0.48**
	(0.08)	(0.16)
minmatch	0.77***	1.49***
	(0.13)	(0.24)
yob_diff_1	−8.40	−22.50
	(90.66)	(437.31)
yob_diff_2	−8.60	−22.84
	(90.66)	(437.31)
yob_diff_3	−8.86	−23.37
	(90.66)	(437.31)
Constant	−14.10	−20.28
	(90.67)	(437.32)
AIC	1346.93	1315.84
BIC	1432.71	1401.62
Log Likelihood	-659.46	-643.92
Deviance	1318.93	1287.84
Num. obs.	3386	3386

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

classification? There are several ways to do this. The standard procedure is to strike a balance between two machine learning assessment measures, the True Positive Rate (TPR) and Positive Prediction Value (PPV). Note that in our model, the higher the probit score, the less likely it is that a link is coded as 0 (non-match or negative prediction). Hence, the TPR in our case is the number of links correctly retained over the total number of actual links retained $\frac{TruePositive}{TruePositives+FalseNegatives}$. PPV is the total number of links correctly retained over the sum of all retained links, $\frac{TruePositive}{TruePositive+FalsePositive}$. As we increase the threshold value to be stricter on the links we retain, PPV will go up, but this would imply that we falsely eliminate more links (i.e. False Negatives would increase), whereby TPR would fall. Researchers must use their judgement in deciding the trade off between these two meta-parameters.[12]

In order to do this, researchers must have an idea of how "aggressive" the RAs were overall in identifying non-matches when deciding on this as well.[13] If the RAs tend to eliminate links fairly aggressively, researchers might choose to put greater weight on TPR and choose a lower threshold value.

[12]This issue is typically solved by searching for the threshold value that maximizes a linear utility function of the form TPR + λPPV, where λ is the relative weight assigned by the researcher on PPV.

[13]This of course also depends on the instructions given to them.

Table 6: Confusion Matrix (Out of Sample)

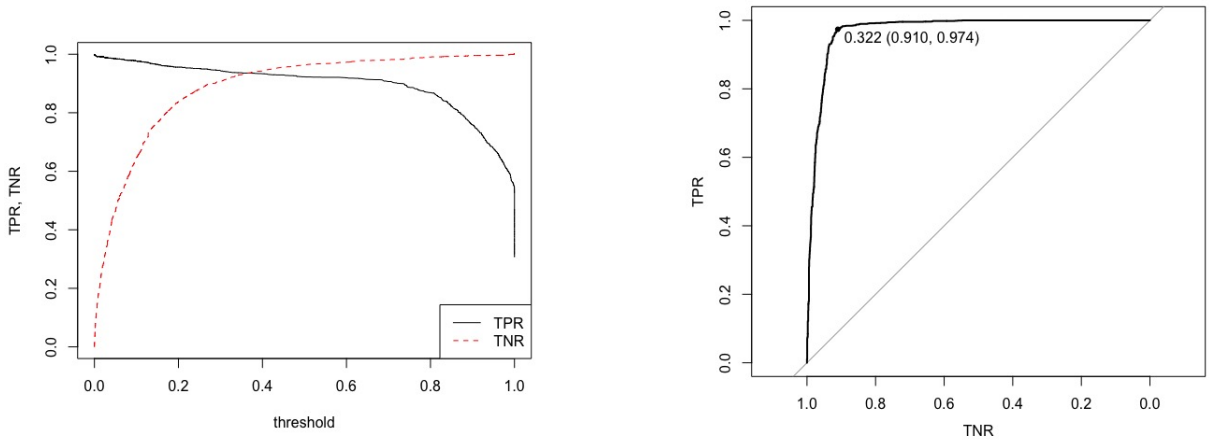
		Prediction	
		0	1
Reference	0	1611	129
	1	36	1610

		Prediction	
		0	1
Reference	0	1606	127
	1	41	1612

Notes: Prediction refers to probit or logit model prediction and Reference denotes how the RAs coded each link (non-match or not).

On the other hand if they are fairly restrained in their approach, a larger weight on PPV might be appropriate. Having multiple RAs work on the same training data set is therefore beneficial; in using the intersection of their work, the resulting output is less sensitive to idiosyncratic decisions. Our objective in using inputs from multiple RAs and using their intersection was to be agnostic about how aggressive each RA should be.

No specific instruction was given to any RA (except that they had to eliminate matches that were not likely), leaving them to use their own judgment. Because of this, we choose the threshold value that maximizes the sum of true predictions (both true positive and true negative) i.e. we maximize the area under the ROC (Receiver Operating Characteristic) curve. The threshold value that maximizes this value in our case is 0.322; in other words the algorithm denotes a link as a non-match if its score is less than 0.322.



Panel A: TPR, TNR Tradeoff

Panel B: Optimum Threshold

Figure 1: Selecting the Optimum Threshold Value

Figure 1 (Panels A and B) illustrates this. In Panel A, we plot TPR and TNR values for the entire range of probit scores. As we increase the threshold value and become more strict on the links that we retain, TPR (the rate at which links are correctly retained as a ratio of all actually retained links) falls, however that in turn implies that we would falsely retain fewer links (False Positives reduce), whereby TNR increases. Panel B shows the threshold value that maximizes the sum of these two measures. At the optimum value, we obtain a TNR of 0.910 and a TPR of 0.974.

5 Results

We report details of how we compute the bounds to help readers implement our method in their own context. Implementing the algorithm to find the upper bound is straightforward as long as one has access to a function that efficiently finds a maximum matching of a bipartite graph. In Matlab, in which we did all of our coding, we find that Ed Scheiner-

man's Matgrph toolbox to be fast.[14] For example, computation of the upper bound on the number of men who moved out of Arkansas (with 119,200 men) between 1850 and 1860, for example, took 382 seconds on a laptop equipped with CPU of 2.5GHz and installed RAM of 16GB.

To compute the lower bound, users of Matlab may use the function that we wrote that implements Algorithm 3 (available upon request). This function takes two following inputs:

1. a bipartite graph (I, J, E), where the set of edges E contains an edge between $i \in I$ and $j \in J$ if and only if j is a potential match for i; and

2. the state of residence for everyone in I and everyone in J

The output of the function is the lower bound computed in Algorithm 3. We chose to code in Matlab because of our familiarity with the language, but users may want to choose different languages, such as Python, Fortran, or C, for a speedier implementation of the algorithm.

To reduce the running time, we first divided up all men whose state of residence was Arkansas in 1850 Census by their birth state, and compute the lower bound for each of these sub-populations. The largest sub-population was those born in Arkansas with 42,700 men. Interestingly, we find that the largest bipartite graph created with the potential matches of Arkansas-native men is not the largest one. The largest bipartite graph is associated with the sub-population of men born in Tennessee. There were 29,172 Tennessee-born men living in Arkansas in 1850, and the size of the bipartite graph (i.e., the number of edges in the bipartite graph) associated with them is 402,082, which is more than twice as large as the size of the bipartite graph for Arkansas-born men

[14] Matgraph toolbox, created by Scheinerman, is downloadable at `https://www.mathworks.com/matlabcentral/fileexchange/19218-matgraph` after logging into mathworks.com.

(192,702). This implies that each Tennessee born men has on average twice as many potential matches than Arkansas born men. Computing lower bounds for the sub-population of Tennessee-born men took approximately 7.5 days (or 653,700 seconds) on a desktop computer equipped with 2.1 GHz CPU and 64GB RAM.

The final results for migration out of Arkansas are as follows. Without refinement i.e. without using any human input and retaining all candidate links after Step 1, the bounds on the number of people migrating out are [47490, 65937], which when divided by the baseline population of 119,200 individuals results in migration rates between [39.8%, 55.3%] out of Arkansas between 1850 and 1860. After using RAs to train a model that eliminates links that are unlikely to be true matches as described in section 4, we ran the algorithm again on the reduced set of potential candidate links to obtain the following bounds, [45591,58461] which translates to out-migration rates of [38.2%, 49%], which is substantially narrower. Note that in order to estimate migration rates given an upper and lower bound on the number of people migrating out of Arkansas, we divide each value by the baseline population in Arkansas in 1850. While this implies that the lower bound on the rate is truly the minimum (since we divide the lowest possible numerator with the highest possible denominator), the upper bound on the rate must be interpreted with caution.

Previous estimates of inter-state migration in the U.S. ranging from 19% to 28% are significantly lower than our estimates. There are two possible explanations for this. Firstly, in 1850 Arkansas was one of the more populous states in the U.S. and experienced significant westward migration, which could explain these high rates. Secondly, previous studies calculating migration rates generally focus on a subset of the population, typically individuals with uncommon names and families with children. To the extent that families with children are already established and less mobile, and individuals with uncommon names live in sparsely populated areas with low out migration, one would also

expect lower rates in previous works than our estimates.

We also compare the estimated bounds with migration rates in the Arkansas data, using matched samples obtained by replicating two state-of-the-art census linking approaches that are outlined in section 2. The results are presented in table 7 below.

Table 7: Comparison with matching algorithms

	Feigenbaum [2016]	Abramitzky et al. [2012a]	Migration Bounds
Baseline Population	119200	119200	119200
Matched Population	51728	33918	-
Match Rate	43.4%	28.5%	-
People Migrated	19,475	12166	[45591,58461]
Migration Rate	37.65%	35.87%	[38.2%, 49%]

Table 7 reports migration estimates from matched samples that we obtained by replicating census linking techniques outlined in Feigenbaum [2016] and Abramitzky et al. [2012a]. We obtain match rates of 43.4% and 28.5% respectively using these methods. In both cases the migration rate obtained is smaller than the lower bound we estimate. As discussed earlier, this difference is likely to be driven by the selected sample of matched people we obtain using the two linking methods. In eliminating individuals with multiple exact or very close matches, we are potentially matching people with more uncommon names at a higher rate. It is plausible that people with uncommon names generally reside in sparsely populated areas, with low out migration rates which attenuates estimates in the matched samples.

6 Simulations

To better understand the practical uses and limitations of the proposed algorithms for calculating upper and lower bounds, we present analysis on simulated data where we can explicitly compare true parameter values with their associated upper and lower bounds. Additionally, this allows us to develop some guidelines for applied work using this method. The key question this attempts to answer is, for what types of imperfectly matched data will these algorithms perform best, and where should we be most weary of its limitations. This may be especially useful in a context where a researcher is deciding whether additional time and money spent improving the quality of matches (for example by employing human research assistants) will provide sufficiently large gains in tightness of bounds.

6.1 Method

The simulations are generated by a simple method meant to parallel the process of matching individuals by name across datasets.

Step 1: Period 1 name and state.

Draw a 'name' value for each of the N individuals i in the simulated sample, where $name_{1i}$ is drawn from a lognormal distribution $name_{1i} = \exp(r_i)$, where $r_i \overset{i.i.d.}{\sim} N(0, \sigma)$. This 'name' value represents a one-dimensional summary of the identifiable characteristics of an individual that are time-invariant. This can include the person's actual name, but also their year of birth, state of birth, etc. Individuals with *name* values that are close to each other are thought of as being similar on those observable markers. So John A Smith, recorded to have been born in Utah in 1831 would have a 'name' value very close to that of a John M Smith, recorded to have been born in Utah in 1832. ('Recorded' because each observation of a given individual is measured with error, either from an imprecise

answer, a mistake in how that answer was recorded, or a later error in transcription.) Each individual is also assigned a potentially time-varying value S_{1i}, which could be for example their current state of residence.

Step 2: Period 2 name and state.

A second set of 'name' values is generated for each individual. To account for the fact that each instance of a person's time-invariant characteristics is recorded with error, we add a white-noise term to the name from period 1. That is,

$$name_{2i} = name_{1i} + \epsilon_i,$$

where ϵ_i is drawn from an i.i.d. $N(0, \gamma)$ distribution, and γ is a parameter that allows us to vary how noisy the recording process is (or rather, how much recorded names vary between rounds 1 and 2 of data collection).

As in period 1, each person is assigned a state. The probability of an individual moving states between periods 1 and 2 is denoted by ρ. Therefore, $S_{2i} = S_{1i}$ w.p. $(1 - \rho)$.

Step 3: Create matches between periods 1 and 2.

Given vectors $name_{1i}$ and $name_{2i}$, we can now create links based on name similarity, as in the census data. To do this we simply consider any pair of period 1 and period 2 observations to be a 'potential' link if $|name_{2i} - name_{1i}| \leq \delta^p$. We think of this set of links as being the analogue of the census procedure of starting with all pairs of people with the same birth state, less than three years difference in year of birth, and first and last name Jaro-Winkler similarity scores greater than 0.8.

A more restrictive set of 'likely' links is generated by selecting only links with $|name_{2i} - name_{1i}| \leq \delta^l$, where $\delta^l < \delta^p$. This in turn is the analogue of the links with a human RA (or a trained algorithm) deems to be a likely match.

Each link is assigned as either an observed mover or not depending on the state of

residence of the two individuals in period 1 and 2.[15]

These three steps create a dataset that has the key characteristics of census data for our analysis, and indeed any two datasets that are linked using imperfectly coded individual identifiers. These are:

1. Time-invariant individual identifiers, measured with error at time 1 and time 2

2. Time-varying characteristics, measured at time 1 and time 2

The purpose of the analysis in this case is to estimate the underlying parameter value of the time-varying characteristic, which would be trivial if the individual identifiers were measured without error. Here we treat the time-varying characteristic as being measured without error (since the state of residence is not subject to any of the types of measurement error described above), but adding error to this measure would be easy to incorporate into the simulation procedure.

The algorithm of finding upper and lower bounds on this parameter which is described above (section 3) can be applied to this simulated data, with the key advantage that we know its true value. In this case, the question is how the upper and lower bounds compare to the true value ρ.

6.2 Choice of parameters for simulations

Before trying to understand the effect of perturbations of the various parameters, we chose a set of benchmark parameter values that were chosen to roughly match the distribution of 'possible' links in the Arkansas 1850-1860 datasets. This is a reasonable starting point since the choice of 'possible' links is less subjective than the selection of 'likely' links, which involves subjective RA decisions.

[15]Note that if only true links were observed, the fraction of observed movers would match the true number of movers, ρ.

Column 1 of table 8 shows the set of parameter values for the benchmark simulation. The bottom rows report the mean and standard deviation of links from this simulation. For comparison, these values are reported for the Arkansas data in column 8.

The benchmark sets the number of individuals to be equal to 10,000. While a larger value would more closely match the scale of the census datasets, the reduced computational load from a smaller sample is useful in testing various alternatives parameter values.

The key parameters used to match the distribution of number of links are σ, γ and δ^p. Higher values of σ implies less overlap in the names of individuals in the dataset. Therefore populations with a smaller set of very popular names would have low values of σ. Populations with more diversity in names would instead have high values. Higher values of γ reduce the number of links, while increasing the proportion of incorrect matches. (This proportion is not benchmarked against since we cannot know its value in the Census data.) Finally, δ^p mechanically increases the number of links per person, since it simply widens the range of similar names that are considered a match.

In the benchmark case, $\delta^p = \gamma$, which implies that individuals with a ϵ_i noise draw which is within one standard deviation from the mean on either side will be included in the set of likely links. Recall that $name_{2i} = name_{1i} + \epsilon_i$, where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \gamma)$.

The other parameters in the benchmark, which have no effect on the number of possible links, are chosen as follows. The true number of movers ρ is set to 10%. The ratio of δ^p / δ^l is set to 2, so that the range of distance between names that is considered a match is half the size for 'likely' matches as it is for 'possible' matches.

Columns 2-7 each contain change in a single parameter relative to the benchmark. In column 2, the number of movers is decreased from 10% to 5%. This is to give us intuition on how far the upper and lower bounds of the share of movers are from the true value, as that value changes. In column 3, the sample size is doubled to 20,000.

Table 8: Simulation parameters and results

Label	(1) Benchmark	(2) Fewer movers	(3) Larger sample	(4) More unique names	(5) Stricter 'possible'	(6) Stricter 'likely'	(7) More noise	(8) Arkansas 1850-60
Data	*Simulated*	*Simulated*	*Simulated*	*Simulated*	*Simulated*	*Simulated*	*Simulated*	*Census*
N	10,000	10,000	20,000	10,000	10,000	10,000	10,000	119,200
σ	3.3	3.3	3.3	3.6	3.3	3.3	3.3	-
κ	0.2	0.2	0.2	0.2	0.2	0.2	0.2	-
γ	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0006	-
ρ (true migration rate)	0.1	0.05	0.1	0.1	0.1	0.1	0.1	-
δ^p	0.0003	0.0003	0.0003	0.0003	0.00015	0.0003	0.0003	-
δ^p / δ^o	2	2	2	2	2	4	2	-
Possible links								
Mean	8.86	9.26	17.27	12.64	6.55	8.87	11.08	8.68
Std Dev	15.53	17.20	34.86	26.78	11.13	16.51	18.51	15.52
Bounds using 'possible' links								
Migrants	[493, 2498]	[269, 2294]	[861, 5724]	[485, 2422]	[468, 2143]	[518, 2640]	[413, 2389]	[47490, 65937]
Percent	[0.049, 0.25]	[0.027, 0.229]	[0.043, 0.286]	[0.049, 0.242]	[0.047, 0.214]	[0.052, 0.264]	[0.041, 0.239]	[0.398, 0.553]
Spread (pct points)	0.20	0.20	0.24	0.19	0.17	0.21	0.20	0.15
Bounds using 'likely' links								
Migrants	[491, 2001]	[342, 1907]	[868, 4740]	[461, 2000]	[422, 1662]	[469, 1590]	[418, 1931]	[45591, 58461]
Percent	[0.049, 0.2]	[0.034, 0.191]	[0.043, 0.237]	[0.046, 0.2]	[0.042, 0.166]	[0.047, 0.159]	[0.042, 0.193]	[0.382, 0.49]
Spread (pct points)	0.15	0.16	0.19	0.15	0.12	0.11	0.15	0.11

Notice that the mean and standard deviation of links increases by doing so, as one would expect. In column 4, the distribution of names is more dispersed, which is equivalent to having a population with a smaller proportion of common names (more 'distance' between names). Column 5 halves the distance allowed for a 'possible' link, implying a stricter selection of initial matches. For example, this could represent limiting links to having a minimum Jaro-Winkler score of 0.9 instead of 0.8, or a maximum difference in year of birth of 2 instead of 3. Column 6 does a similar tightening of matches, but for 'likely' matches rather than 'possible'. Finally column 7 adds more measurement error to the name, reducing the probability that a given match is the true one. In each of these cases, the key question of interest is how far the bounds are from the true value of ρ.

6.3 Simulation results

The bottom two panels of Table 8 present the results of the seven simulations described above. The benchmark simulation has a true migration rate of 10%, and finds bounds of 5-25% when using only the looser criterion for 'possible' links. Changing the true rate of movers from 10% to 5% in column 2, the range of the bounds stays the same, but the bounds themselves shift to 3-23%. This suggests that, at least when the true rate is low, the lower bound is much closer to the true statistic than the upper bound. Column 4 considers the effect of wider distances between names, which we interpret as having fewer common names, or more unique names. Here, surprisingly, we find the bounds essentially unchanged. Column 5 halves the maximum distance between names that is considered a 'possible' links. This lowers the upper bound to 21% without moving the lower bound by much, thereby shrinking the total spread. Column 6 does not imply any change in the bounds calculated using 'possible' links.

These bounds are all substantially larger than the bounds we find using the US Census Arkansas data. This is likely because either the sample size is much larger in the Arkansas

case, or simply because the true rate in that data is far higher than in the simulation (though, of course, we do not know the true rate in the Census data).

When reducing possible matches to the stricter definition of 'likely' links, the bounds in the benchmark case (column 1) shrink to 5-20%. In all simulations we consider, the spread of the bounds shrinks by 4-5 percentage points, with the notable exception of column 6, where we only reduce the allowed distance for these 'likely' links. Here, the bounds shrink by 10 percentage points, dropping almost by half.

These results, while preliminary, suggest we can achieve something comparable to real-world bounds using simulated data. For each of the parameter values we consider, the true rate of movers is well within the bounds we calculate. The spread in bounds in each case is larger than we find using the Arkansas Census data, suggesting more fine-tuning of the simulation parameters to better match existing data may be useful. The bounds we find with the simulated data, while not particularly narrow, would not be so small as to be uninformative in cases with real data.

7 Conclusion

We propose a method to calculate upper and lower bounds on statistics of interest in datasets that lack unique identifiers. This method first requires dropping all obvious non-matches. Instead of then dropping records with more than one potential match, or choosing one amongst their matches as being correct, we explore the entire range of combinations of match combinations across the population. This is implemented by algorithms of which we require feasibility and non-wastefulness in the set of chosen matches. Feasibility innocuously limits actual links to the set of potential links, while non-wastefulness prevents the algorithm from trivially linking very few records.

We implement this method on data from the 1850 and 1860 US Census, to study mi-

gration out of Arkansas in this ten-year period. To do so, we first limit the 1850 Census to males currently living in Arkansas. For each of these men, we select potential links to males across the US in the 1860 Census. We do this by selecting all records that are sufficiently similar in name, with years of birth less than three years apart, and with the same state of birth. Importantly, we do not use co-residents in selecting matches. While using the names of household members may improve the correct match rate, it can also bias the data towards men who are less likely to move.

We also explore the added benefit from further refining the set of potential links. Rather than doing this simply by restricting the initial criteria for potential links, we train a machine learning model using the judgement of research assistants about which potential links can safely be dropped. We then re-estimate bounds on inter-state migration rates out of Arkansas using this subset of links.

Using both the looser and stricter set of potential links, we find bounds that are relatively tight, and surprisingly high. Using the first set of links the lower bound is 39.8% and the upper is 55.3%, a spread of 15 percentage points. Notably, these bounds are tight enough to exclude a wide range of low but plausible migration rates. Using the smaller set of links that are not excluded by the RA-trained machine learning algorithm, we find bounds of 38.2% to 49%. This almost one-third reduction in spread (11 points, down from 15) is achieved with what we think of as being modest and uncontroversial elimination of links that are highly unlikely to be true matches.

These bounds have the appealing features of being (a) tight enough to be informative, (b) highly reproducible and inexpensive to obtain, and (c) true under a very limited set of assumptions about which links between census rounds are correct. We believe this measure acts as an important complement to existing methods of selecting matches across census rounds, which generate point estimates of statistics of interest. Our method is useful both at the beginning and end of any research project where historical data need

to be linked. Initially, it can guide whether and how much to invest in manual selection of links. Later, it can point to the size and direction of bias from sample selection, and speaks to how sensitive point estimates are to decisions in the linking process.

A Proofs

Proof. Proposition 1.

Suppose, by way of contradiction, that there exists a feasible and non-wasteful matching $\mu' \in \mathbb{M}$ for which

$$\sum_{i \in I, \mu'(i) \in J} \mathbb{1}\{s_i \neq s_{\mu'(i)}\} > |E^{max}|$$

Then define a set of edges E' such that an edge between $i \in I$ and $j \in J$ is in E' iff $\mu'(i) = j$ and $s_i \neq s_j$. Then the matching associated with so constructed set of edges E', denoted by $M' = (I, J, E')$, is a matching in the bipartite graph $G = (I, J, E)$, since for any given edge (i,j) in E', $\mu'(i) = j$ by construction, hence $j \in J(i)$ by the feasibility of μ'. The size of the matching M' is equal to the number of people in I that is matched to a person in j whose state of residence is different from theirs by construction, i.e.,

$$|E'| = \sum_{i \in I, \mu'(i) \in J} \mathbb{1}\{s_i \neq s_{\mu'(i)}\}$$

which implies that the size of matching M' is strictly larger than that of M^{max}, which is a contradiction. □

References

R. Abramitzky, L. Platt Boustan, and K. Eriksson. Europe's tired, poor, and huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, 102(5), 2012a.

R. Abramitzky, L. Platt Boustan, and K. Eriksson. Have the poor always been less likely to migrate? evidence from inheritance practices during the age of mass migration. *Journal of Development Economics*, 2012b.

R. Abramitzky, L. Platt Boustan, and K. Eriksson. A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, 106(4), 2014.

Martha Bailey, Connor Cole, Morgan Henderson, and Catherine Massey. How well do automated methods perform in historical samples? evidence from new ground truth. Technical report, National Bureau of Economic Research, 2018a.

Martha Bailey, Connor Cole, and Catherine Massey. Simple strategies for improving inference with linked data:acase study of the 1850-1930ipumslinked representative historical samples. Technical report, Working Paper, 2018b.

J. J. Feigenbaum. Automated census record linking: A machine learning approach. *Unpublished Manuscript*, 2016.

J. P. Ferrie. A new sample of males linked from the 1850 public use micro sample of the federal census of population to the 1860 federal census manuscript schedules. *Historical Methods*, 29(4), 1996.

John Hopcroft and Richard Karp. An $n^{\frac{5}{2}}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2(4):224–231, 1973.

Joshua L Rosenbloom and William A Sundstrom. The decline and rise of interstate migration in the united states: Evidence from the ipums, 1850-1990. Working Paper 9857, National Bureau of Economic Research, July 2003.

D. Schaefer. A statistical profile of frontier and new south migration, 1850-1860. *Agricultural History*, 59, 1985.

R. Steckel. Census matching and migration: A research strategy. *Historical Methods*, 21, 1988.