Linking Individuals Across Historical Sources: a Fully Automated Approach*

Ran Abramitzky[1], Roy Mill[2], and Santiago Perez[3]

[1]Stanford University and NBER, email: ranabr@stanford.edu

[2]At-Bay Inc. email: milmilon@gmail.com

[3]UC Davis, email: seperez@ucdavis.edu

October 2018

* An early version of this paper, entitled "Linking Records across Historical Sources", was Chapter 3 of Roy Mill's dissertation completed at Stanford in June 2013. We have benefited from conversations with Jaime Arellano-Bover, Leah Boustan, Raj Chetty, Katherine Eriksson, James Feigenbaum, Randall Walsh, Tom Zohar and participants in the UC Berkeley complete count census workshop.

Abstract:

Linking individuals across historical datasets relies on information such as name and age that is both non-unique and prone to enumeration and transcription errors. These errors make it impossible to find the correct match with certainty. In the first part of the paper, we suggest a fully automated probabilistic method for linking historical datasets that enables researchers to create samples at the frontier of minimizing type I (false positives) and type II (false negatives) errors. The first step guides researchers in the choice of which variables to use for linking. The second step uses the Expectation-Maximization (EM) algorithm, a standard tool in statistics, to compute the probability that each two records correspond to the same individual. The third step suggests how to use these estimated probabilities to choose which records to use in the analysis. In the second part of the paper, we apply the method to link historical population censuses in the US and Norway, and use these samples to estimate measures of intergenerational occupational mobility. The estimates using our method are remarkably similar to the ones using IPUMS', which relies on hand linking to create a training sample. We created a Stata command that implements this method.

Keywords: census, record linkage, micro data

I Introduction

Linking individuals across datasets offers rich possibilities for economic history research.[1] However, because historical data often lack identifiers such as a Social Security Number, linking individuals relies on personal information such as names and reported ages that is prone to enumeration and transcription errors. These errors make it impossible to find the correct match with certainty. Furthermore, multiple individuals with identical names and reported ages introduce the problem of non-unique matches. Economic historians have developed useful ways to link individuals across historical datasets in the presence of such issues (for example, Atack et al. [1992], Ferrie [1996], Abramitzky et al. [2012, 2014, 2017] and Feigenbaum [2016a]; Massey [2017], Bailey et al. [2017], and Abramitzky et al. [2018] compare various matching algorithms).

A record matching method should aim to trade-off three goals. First, make as few false matches as possible (minimize type I errors). Second, make as many true matches as possible (minimize type II errors). Third, for given levels of type I and type II errors, create linked samples that resemble the population of interest as closely as possible. Different research projects may have different implications for compromising on each of these three goals.

In the first part of the paper, we suggest a fully automated method for linking historical datasets that enables researchers to create samples at the frontier of these three goals. The method has three steps. In the first step, we guide researchers in the choice of which variables to use for linking. In the second step, we combine distances in reported names and ages between each two potential records into a single score, roughly corresponding to the probability that both records belong to the same individual. We estimate these probabilities using the Expectation-Maximization (EM) algorithm, a standard technique in the statistical literature (Dempster et al., 1977, Winkler. 1989). In the third step, we suggest a number of decision rules that use these estimated probabilities to determine which records to use in the analysis.

[1] Recent examples include Abramitzky et al. [2012, 2014, 2017]; Aizer et al. [2016], Bleakley and Ferrie [2016, 2013], Collins and Wanamaker [2014, 2015, 2017], Eli et al. [2016], Eriksson [2015], Feigenbaum [2016b, 2017], Ferrie [1997], Fouka [2016], Long [2006], Long and Ferrie [2013], Hornbeck and Naidu [2014], Mill and Stein [2016], Kosack and Ward [2014], Modalsli [2017], Parman [2015], Perez [2017], Salisbury [2014].

Although there is a large literature that uses linked historical records, existing linking methods either do not use insights from statistics or are not easily replicable. Specifically, existing automated methods (e.g. Ferrie, [1996]; Abramitzky, Boustan and Eriksson, [2012, 2014, 2017]) are replicable but they rely on common sense rather than formal statistics. Existing hand-linking (e.g. Bailey et al, [2017], Costa et al, [2018]) and semi-automated methods (e.g. Feigenbaum, [2016a]; Ruggles et al., [2011]) rely on hand coders and may not be fully replicable. Unlike automated methods, it is unlikely that any two hand coders that started from the same raw data sources (for example, the 1920 and 1940 censuses) will make the exact same choices and generate the exact same training samples. Hence, relying on hand coders or on a training sample implies that two researchers starting from the same underlying data might end up building different linked samples. We suggest a method that is estimated in a fully automated way that is both grounded in statistics and is easily replicable.

In addition, while the EM algorithm is a standard method used in statistics for record linkage, this method has not been used for linking historical records (other than by the coauthors of this paper). Hence, our aim is to bring together practitioner experience of economic historians with insights from statistics for record linking.

It is an empirical question whether this method can generate meaningful samples when linking historical records, where there are unique challenges such as enumeration error, transcription error, mortality, return migration, and under-enumeration between Census years. In the second part of the paper, we test how our suggested method performs by using it to link fathers and sons across the US 1850-1880 and the Norwegian 1865-1900 censuses of population. We use these data to construct father-sons occupational transition matrices and to compute summary measures of intergenerational occupational mobility. We then compare our results to those obtained when using the widely-used linked samples constructed by IPUMS [Goeken et al., 2011]. These samples were constructed by first manually linking a subset of the records and then using this training sample to predict the linking status of the remaining records. For both countries, we document that the

patterns of intergenerational occupational mobility that we find using our linked samples are remarkably similar to those that we find when using the samples created by IPUMS.

This new method for historical record linking helps address concerns about false positives. Moreover, the method is flexible in that it can accommodate different researchers' preferences with respect to the trade-off between match quality and sample size. To facilitate the use of the method by practitioners, we have developed a Stata command that implements it. We provide the program and its corresponding documentation on our website:
https://people.stanford.edu/ranabr/matching-codes

II The matching problem

Imagine you are a researcher who wants to link people from the 1900 to the 1910 census. Imagine that one observation in 1900 is "Ran Abramitzky" who is reported being 10 years old. When you look up this record in 1910, you are looking for a "Ran Abramitzky" who is reported to be a 20 year old. However, when you search the 1910 census, you find three potential matches. One is a "Ran Abramitzky" who is reported to be a 21 year old. One is a "Ran Abramtziky" who is reported to be a 20 year old. And one is a "Ran Abramitzky" who is reported to be a 20 year old.

How would you know which one is the true match? It may be tempting to choose the exact match (third record). However, the other two may as well be the right one given that enumerators can easily make spelling errors and people may not report their exact age but rather round it up or down. An alternative is to declare this record as an impossible to match and drop it from the analysis, but this will result in a smaller sample size.

This problem of record linkage in the presence of errors in identifying information was already discussed almost 50 years ago in statistics [Fellegi and Sunter, 1969]. Much of this paper simply translates the insights from the statistics literature to the problem of historical record linking.

There are three goals that need to be taken into account when linking records:

1. Make as few false matches as possible: This corresponds to minimizing type I errors (minimizing false positives). In other words, we want the least number of cases where the potential match is a false match but we deem it as matched.

2. Make as many true matches as possible: This corresponds to minimizing type II errors (minimizing false negatives). In other words, we want the least number of cases where the potential match is a true match but we deem it as unmatched.

3. Create a sample that is as representative as possible: For given levels of type I and type II errors, we want the linked sample to resemble the population from which we draw matches as much as possible.

The first two goals describe a standard type I versus type II error trade-off, and are the ones emphasized in the Fellegi and Sunter [1969] framework. The third goal is an additional challenge that is faced by researchers in the social sciences who are interested in creating linked samples.

III Selecting identifying and blocking variables and measuring string distances

Before calculating probabilities that each two records are a true match (section IV) and choosing a match to be used in the analysis (section V), there are three main decisions that the researcher has to make. This section discusses these three decisions in turn.

Selecting identifying variables

The first decision is to choose which identifying variables to use in the matching procedure. The "Ran Abramitzky" example used name and age as identifying variables, but historical datasets often contain other potentially identifying information such as gender, occupation, race, place of birth and place of residence.

Here statistics theory does not provide a guidance, and instead the economics research question should guide the decision of which variables to use in the linking procedure. The selection of identifying variables will affect all three goals of the match. As we use more variables, we are better able to distinguish between otherwise equally-likely matches. For example, adding age to the list of identifying variables we are potentially able to distinguish between two different Ran Abramitzkys. If we use county of residence, we can distinguish between two Ran Abramitzkys who have the same age.

While adding variables to the list of identifying variables may increase the match rates and decrease false match rates, it may also introduce non-representativeness. For instance, a variable like county of residence appears in all censuses and can significantly increase match rates and even help us identify the true individual. However, using such a variable would result in excluding those who switched their county of residence from the analysis. This exclusion will be an issue in a study on geographical mobility, but will not be an issue in a study of fertility among residents who stay in Indiana. Similarly, using occupation for matching will bias any analysis of occupational mobility, but may not be an issue when studying outcomes unrelated to occupations.

The decision of whether to use a variable as an identifying variable thus depends on the research question at hand. In most economics applications, using outcome variables such as occupation or place of residence may be problematic. We suggest following standard practice in economic history and only use predetermined individual level characteristics in the matching procedure. Usually, this restriction reduces the matching variables to names, age and place of birth, which will be the focus of the rest of the paper.[2]

Blocking

The second decision has to do with reducing the computational requirements. In principle, we might want to compare every individual in dataset *A* to every individual in dataset *B*. In practice, this is currently not possible computationally unless the size of datasets *A* and *B* is very small. The reason is that we would need to perform $n_A X n_B$ comparisons, where n_A and n_B are the sizes of datasets *A* and *B*, respectively. For example, if you need to match 100 records in dataset *A* to 100 records in dataset *B*, you will need to make 100*100 = 10,000 comparisons and assign 10,000 probabilities. In a census of millions of people, this can be computationally impractical. The solution to this computational issue is to only compare individuals who agree on certain blocking variables. Ideal blocking variables are those for which mistakes are very unlikely. For instance, if individuals rarely misreport their state of birth, we would be unlikely to miss any true matches by not comparing individuals who declared different states of birth. Further reductions in computational time can be obtained by blocking on gender, or the first letter of the last name. Nevertheless, even though finer blocking results in a lower number of comparisons, blocking is

[2]Another variable that could potentially be used in linking is race. However, using this variable could be problematic if individuals selectively report a different race in different historical sources, a pattern documented in Mill and Stein [2016] and Nix and Qian [2015].

not an innocuous process because it rules out any potential matches across blocks. For instance, if we block by the first letter of the first name, we rule out the name Emmanuel from ever matching to the name Immanuel.

Similar to the choice of identifying variables, the decision on which variables to block on depends on the research question. For example, it will not make sense to block on race in a study of racial passing. Current applications of this method (Mill and Stein, [2016]; Perez, [2017]) restrict the set of comparisons to individuals who are: (1) born on the same state, (2) have the same first letter in first and last names and, (3) have an age difference no larger than five years in absolute value.

Measuring string distances

The third decision is how to map differences in name spellings into a numerical distance.[3] There is more than one way to compare two strings to each other. One straightforward option is to use an indicator of whether the names are exactly the same. In our example, 1910 "Ran Abramitzky" will have a distance of 0 and 1910 "Ran Abramtziky" will have a distance of 1 from 1900 "Ran Abramitzky". Another option is to use a phonetic algorithm such as NYSIIS instead of the exact name. When using a phonetic algorithm, words that have a similar pronunciation are assigned the same phonetic code. These phonetic codes are designed to overcome name spelling discrepancies that stem from the translation of a heard name to a written name.[4] A third option is to use a continuous string distance measure. When discrepancies in names stem mainly from hearing a name to writing it down, then using phonetic codes such as NYSIIS is a reasonable solution. When the discrepancies come from the exact spelling or digitization of the handwritten record, then string distances can produce better results. Phonetic code match can be used in addition to string distances.

[3] A related decision is how to map numerical distances (for instance, age differences) into a distance metric. We usually use the absolute difference in reported age, but we note that other distance metrics are also possible (for instance, an indicator that takes a value of one if both ages agree and is zero otherwise).

[4] Recent economic history papers use the NYSIIS algorithm. Other examples of phonetic algorithms include Soundex [Odell and Russell, 1918] and Metaphone [Philips, 1990]. Some phonetic algorithms are better suited for dealing with languages other than English. For example, the Spanish Metaphone algorithm is designed to match Spanish names [Mosquera et al., 2012].

There are many string distance measures available in the literature. We use the Jaro-Winkler string distance (Jaro [1989]; Winkler [2006]) since it is specifically designed for the comparison of names and was developed in the context of record linking. We note, however, that the method that we discuss in this paper (and the Stata command we provide) is more general and does not require using the Jaro-Winkler distance as its input. In principle, other string distance measures could be used as inputs in the estimation procedure. The Jaro-Winkler string distance calculates a function of the number of matching characters and required transpositions between the two compared strings (names). It gives a higher weight to discrepancies in the first part of the string, where errors are less likely to be made. The original measure is a measure of agreement spanning between 0 (no common characters) and 1 (exact string match). Since we want to treat all discrepancies in identifying variables as distances, we actually calculate 1 minus the Jaro-Winkler distance as originally defined, thereby having 0 as the distance between two exact names and 1 as the distance between two strings with no common characters.

In the Ran Abramitzky example, "Abramtziky" will be coded as a different name than "Abramitzky" using the NYIIS algorithm, but the Jaro-Winkler distance between these two names will be very low (0.02). At the same time, there are examples in which names have the same NYIIS code but a high Jaro-Winkler distance.[5]

IV Assigning a probability that each two records are a true match

After calculating name and other distances such as distances in reported age, we want to combine them into a single distance metric. A natural meaningful measure is the probability that a record pair is a true match. Several ways to estimate this probability have been suggested in non-historical settings (see Winkler [2006] for a rich survey of literature on the subject). In historical settings, Ruggles [2011] and Feigenbaum [2016a] estimate these probabilities using a training sample of manually classified records. We suggest an alternative method that does not rely on a training sample, which has the advantage of making the matching easier to replicate by other researchers. The method has been widely used for record linkage in non-historical contexts and is an application

[5] For instance, "James Tennes" and "James Thomas" have the same NYSIIS code, but the Jaro-Winkler distance between "Tennes" and "Thomas" is 0.4.

of the Expectation-Maximization (EM) algorithm.[6] This section describes how to apply the EM algorithm to the problem of matching historical records.

To gain intuition about the method, imagine that there are 10 Ran Abramitzkys in the 1900 census, and 10 Ran Abramitzkys in the 1910 census. Each Ran Abramitzky in 1900 is aged from 1 to 10 year old. Our goal is to link these two datasets using information on reported ages, but the challenge is that age is potentially misreported in the 1910 census. For example, somebody who is reported to be 11 in 1900 is reported to be 20 in 1910 instead of 21. This misreporting implies that the age distance will sometimes be greater than zero when comparing two records that belong to the same Ran Abramitzky. Each Ran Abramitzky in 1900 has 10 potential matches in 1910, so we would like to assign a probability that each of these 10 potential matches is the true one. There are 10 Ran Abramitzkys, so there are 100 such probabilities to assign.

To illustrate this example, we simulate 100 age distances. We assume that 10 of these distances correspond to a comparison of true matches, while 90 of them correspond to a comparison of true non-matches. The distances that correspond to true matches are drawn from a normal distribution with mean 0 and standard deviation of 1. The distances that correspond to true non-matches are drawn from a normal distribution with mean 5 and a standard deviation of 1. Panel (a) of figure 1 shows the distribution of observed age distances in this example, if we knew what are true matches and what are non-matches. There are 100 such distances drawn in this graph, each represented as a circle. These age distances come from two different populations: "matches" (that is, the observations belong to the same individual, corresponding to the 10 circles drawn in red) and "non-matches" (that is, the observations do not belong to the same individual, corresponding to the 90 circles drawn in blue).

However, the challenge is that in reality we do not know whether each distance belongs to a comparison of true matches (red) or to a comparison of non-matches (blue). Instead, our actual

[6] The general EM algorithm was described in Dempster et al. [1977]. The specific use of the EM algorithm for record linkage problems was developed by Winkler [1989]. For a Bayesian approach to record linkage problems see Larsen [2005].

data look like panel (b) in figure 1. The goal is to use these data to estimate the likelihood that each distance corresponds to a true match, even though we do now know for sure what records are a true match and what records are a non-match.

The EM algorithm starts from assuming that distances between records follow a particular type of distribution, and allowing two different distributions for matches and non-matches. For instance, one possible assumption is that, with probability p_M, distances are distributed normally with mean μ_M and standard deviation σ_M and, with probability $(1- p_M)$, distances are distributed normally with mean μ_U and standard deviation σ_U. The procedure then estimates p_M, μ_M, μ_U, σ_M and σ_U, and uses the parameter estimates to identify two separate clusters (one from which true matches are more likely to come and one from which non-matches are more likely to come).

Intuitively, we expect age distances to be on average smaller when comparing the same individual than when comparing different individuals. Panel (c) shows the estimated distributions under the assumption that distances are normally distributed. Given these estimated distributions, it is clear that observations that are closer to zero are going to be predicted to be more likely to belong to the population of true matches. In addition, it is clear given the size of each of the clusters that the fraction of true matches (p_M,) is smaller than the fraction of true non-matches (1-p_M,). At the same time, the degree of confidence on each of the links will depend on how informative the identifying information (in this case, reported ages) is. The further apart μ_M is from μ_U, the more confident we will be in distinguishing matches and non-matches. Similarly, when σ_M and σ_U are small (that is, if there is very little noise in the identifying information), then we will have more confidence in distinguishing matches and non-matches (there will be less overlap between the estimated distributions).

Imagine now that you try to link both Ran Abramitzky and Santiago Pérez. This will add to the problem the string distance dimension in addition to the difference in reported age. The intuition remains the same, but clustering will be two-dimensional in this case. Figure 2 shows an example in which records differ both with respect to their reported names (x-axis) and ages (y-axis). In panel (a), each data point is labelled as if we knew which records belong to true matches. Panel (b) is how our actual data look like: observations are not labelled as belonging to a comparison of

true matches or as a comparison of true non-matches. More generally, consider the set of ordered pairs of records $A \times B$ and partition this set to the set of true matches (M), if the records in A and B describe the same person, and the complementing set of true non-matches (U). Suppose that the distance, or the degree of non-agreement, in identifying variable k for pair $i \in A \times B$ is given by γ_{ik}, and the vector of such distance measures for pair i is γ_i. Our goal is to estimate for each pair how likely it is to be a true match given the vector of distances in the identifying variables. A pair with shorter distances should be more likely to be a match relative to a non-match.

The probability that a pair i in $A \times B$ is a true match conditional on the distances in the identifying variables γ_i (in our case, reported names and year of birth) can be inferred from Bayes Rule as:

$$\Pr(i \in M \,/\, \gamma_i) = \frac{\Pr(\gamma_i \cap i \in M)}{\Pr(\gamma_i)} \tag{1}$$

However, we obviously do not really know if pairs are true matches (in M) or non-matches (in U). In other words, pairs are not labeled as being in M or in U. In the data, we just observe a sample analogue of $\Pr(\gamma_i)$ (that is, we observe the empirical distribution of distances across pairs of records, which in our previous example corresponds to panel (b) of figure 1). At the same time, we know that:

$$\Pr(\gamma_i) = \Pr(\gamma_i \,/\, i \in M)\, p_M + \Pr(\gamma_i / i \in U)\,(1 - p_M) \tag{2}$$

where p_M is the unconditional probability that a pair is a match.

The method requires that we assume a statistical distribution for $\Pr(\gamma_i \,/\, i \in M)$ and $\Pr(\gamma_i / i \in U)$. We can then use maximum likelihood to find the parameters of the statistical distribution that maximize the likelihood of observing the observed distances. Once we find these parameters, we can compute an estimate of:

$$\Pr(i \in M \,/\, \gamma_i) = \frac{\Pr(\gamma_i \,/\, i \in M)\, p_M}{\Pr(\gamma_i \,/\, i \in M)\, p_M + \Pr(\gamma_i / i \in U)\,(1 - p_M)} \tag{3}$$

That is, the probability that a pair of observations is a match given the observed distances in identifying variables.

If we observed true match status, finding the parameters that maximize the likelihood function would be a straightforward exercise. The reason why we need the EM algorithm to estimate these parameters is because we do not observe true match status, which makes the direct maximization

of the likelihood function complicated computationally. The EM algorithm is just a numerical tool that enables us to estimate these parameters without information on true match status. Indeed, in the original Fellegi and Sunter [1969] introduction of the problem of record linkage, estimation was conducted through the method of moments. We use the EM algorithm as it has been shown to have better convergence properties than other numerical optimization tools (see, for example, Meilijson, [1989]; Xu and Jordan, [1996]).

In particular, the EM algorithm suggests an iterative process to estimate the parameters of the distributions above. It starts by calculating the probability of being a true match (left-hand-side of (1)) given a guess of the distributions of distances conditional on being a match or a non-match (right-hand-side of (1)). Then, based on these probabilities it makes a better guess of the same conditional distribution for another iteration. This process is repeated until the parameters converge. According to Dempster et al. [1977] (and specifically in this context according to Winkler [1989]) the algorithm reaches a local maximum of the likelihood function.

The EM algorithm

1. Define a distribution family for $\Pr(\gamma_i / i \in M)$ and $\Pr(\gamma_i / i \in U)$. The algorithm will estimate the parameters of the distributions. Denote the vectors of unknown distributional parameters as θ_S, where $s \in \{M,U\}$.
2. Guess initial values for parameters of the conditional distributions $\theta_s^{(0)}$ and the unconditional probability to be a true match $p_M^{(0)}$.
3. Loop over steps *E* and *M* until convergence:

   E-step: Given $\theta_s^{(t)}$ and $p_M^{(t)}$ infer $w_i^{(t)}$ according to Equation 1:

$$w_i^{(t)} = \Pr\left(i \in M / \gamma_i, \theta_M^{(t)}, p_M^{(t)}\right) = \frac{\Pr\left(\gamma_i / \theta_M^{(t)}\right) p_M^{(t)}}{\Pr\left(\gamma_i / \theta_M^{(t)}\right) p_M^{(t)} + \Pr\left(\gamma_i / \theta_U^{(t)}\right)(1 - p_M^{(t)})} \quad (4)$$

   M-step: Given $w_i^{(t)}$, infer $\theta_s^{(t+1)}$ and $p_M^{(t+1)}$ using Maximum Likelihood. The distribution of γ_i (an observable measure) is given by:

$$\Pr(\gamma_i) = \Pr(\gamma_i \,/\, i \in M)\, p_m + \Pr(\gamma_i / i \in U)\,(1 - p_m) \tag{5}$$

i. Hypothetically, if the classification of pairs to true matches and nonmatches was known and denoted by $z_i = I\{i \in M\}$, then we could have estimated θ_M and θ_U from the sample subsets of true matches and nonmatches:

$$log\ \ L(\gamma, z, \theta, p_M) = \sum_{i=1}^{N} [z_i\ log\ p_M \Pr(\gamma_i/\theta_M) + (1 - z_i)\ log\ (1 - p_M) \Pr(\gamma_i/\theta_U)] \tag{6}$$

ii. Since the classification z_i is unknown we replace it with $w_i^{(t)}$.The maximum likelihood

estimates are then:

$$p_M^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} w_i^{(t)}$$

$$\theta_M^{(t+1)} = arg \max_{\theta} \sum_{i=1}^{N} w_i^{(t)}\ log \Pr(\gamma_i/\theta) \tag{7}$$

$$\theta_U^{(t+1)} = arg \max_{\theta} \sum_{i=1}^{N} (1 - w_i^{(t)})\ log \Pr(\gamma_i/\theta)$$

After obtaining the maximum likelihood estimates, we can then compute, for any given pair i in *AXB* an estimate of $\Pr(i \in M/\gamma_i)$.

The procedure described above requires assuming a statistical distribution for the observed distances in identifying variables. In most applications of the EM algorithm, a conditional independence assumption is invoked: distances in each identifying variable are assumed to be independent of distances in the other variables, conditional on being a match/non-match. Thus, a distribution can be defined separately for each variable *k*: $\Pr(\gamma_{ik} \cap\ i \in M)$ and $\Pr(\gamma_{ik} \cap\ i \in U)$. Such an assumption greatly simplifies the estimation, as it reduces the number of parameters that need to estimated. Although this assumption is unlikely to exactly hold in real-world data, practical matching projects have shown that it is possible to still achieve high quality matches under this assumption (e.g. Herzog et al. [2007]; Christen, [2012]).

In the empirical applications discussed below, the distribution selected for the birth year distance was multinomial with six possible outcomes, each corresponding to an age difference ranging from 0 to 5 years in absolute value. Name distances, which are spanning the [0,1] range, were grouped in four ranges following Winkler [1988], roughly corresponding to agreement, partial agreement, partial disagreement, and disagreement: [0, 0.067], (0.067,0.12], (0.12, 0.25], and (0.25,1]. We then assumed a multinomial distribution of which range a name distance falls into.

Discretizing the string distance information into bins has a key practical advantage. With this approach, each pair of records can fall into a finite number of bins. For instance, if there are 6 possible age differences, 4 possible first name distances and 4 possible last name distances, there are just 6×4×4 = 96 possible combinations of distances. The properties of the multinomial distribution then imply that one just needs to store 96 numbers to run the EM algorithm, instead of storing a potentially very large matrix with all possible combinations. This simplification dramatically improves computational time. In other words, it is sufficient to know how many of the pairs fall into each of the bins to estimate the parameters of the multinomial model.[7] Recent applications of the EM algorithm for large-scale record linkage also adopt this simplification (see, for example, Enamorado et al, [2017]).

Intuition of the approach and further limitations

As described above, the goal of the method is to split the full set of pairs of records into two groups ("clusters"): matches and non-matches. The simplest way of thinking about this grouping problem would be to use *k-means* clustering. In this approach, the data are split into *k* clusters so as to (1) minimize the within-cluster differences across observations and (2) maximize the between-clusters differences. Intuitively, pairs of records that are closer to each other with respect

[7] This reasoning is analogous to the one that indicates that to estimate the probability of heads for a coin using N tosses, one just needs to have information on the number of tosses that resulted in heads. So, instead of storing N numbers, it is enough to know just one.

to their name and age distances should be grouped together in the cluster of "matches", and observations that are further away should be grouped together in the cluster of "non-matches". The EM algorithm instead computes probabilities of observations belonging to each of the clusters. The goal of the method is to maximize the overall probability or likelihood of the data, given the assigned clusters.

Ideally, we would like pairs of records that are close to each other in terms of identifying information to belong to the cluster of matches, while observations that are further apart to belong to the cluster of non-matches. However, a limitation of the approach is that there is no guarantee that the parameters that locally maximize the likelihood function will split the sample into matches and non-matches. Given this, one important sanity check is that the estimated match probabilities are indeed decreasing in the distance between observations. Formally, we want that:

$$\gamma_i \leq \gamma_j \Longrightarrow \Pr(i \in M/\gamma_i) \geq \Pr(i \in M/\gamma_j)$$

In the case of conditional independent distributions, this will be satisfied by a monotone likelihood ratio in each of the distances. That is, for each of the distances we want that:

$$\gamma_i \leq \gamma_j \Longrightarrow \frac{\Pr(\gamma_i/i \in M)}{\Pr(\gamma_i/i \in U)} \geq \frac{\Pr(\gamma_j/i \in M)}{\Pr(\gamma_j/i \in U)}$$

In addition, note that if there are no duplicates in either datasets A and B, the unconditional match probability p_M cannot be higher than $\frac{\min(n_a,n_b)}{n_a \times n_b}$. Hence, another restriction on the parameters that should be checked is whether the condition $p_m \leq \frac{\min(n_a,n_b)}{n_a \times n_b}$ holds.

When initializing the EM algorithm, we impose these two constraints into our initial parameter guesses. A more sophisticated version of the code could impose this sanity check as further restrictions on the probabilities (rather than just checking ex post that they are satisfied).

One case in which the algorithm typically fails is when the fraction of true matches (p_m) is very small. One fix to this issue is to use what Yancey [2002] calls a "match enriched sample": a sample in which we oversample observations that are ex-ante more likely to be a true match. One adjustment that works well in practice is to restrict the set of comparisons to individuals who match on place of birth, and first letter of the first and last names. This adjustment ("blocking") largely

excludes pairs of records who are very unlikely to belong to the same individual. This issue with the EM algorithm is an additional reason why blocking on some identifying variables is useful.

Finally, note that if the proportion of matches in the data (p_m) is relatively low, then it will be the case that:

$$\Pr(\gamma_i / i \in U) \approx \Pr(\gamma_i)$$

Hence, it is possible to obtain a close approximation of $\Pr(\gamma_i / i \in U)$ by using the observed frequencies of γ_i in the data. For instance, if we observe that 20% of the age distances are equal to 0, we can say that $\Pr(Age\ Distance_i = 0 / i \in U) \approx 20\%$. It is possible to directly incorporate this insight into the procedure.

V Choosing records to use in the analysis

Now that we have estimates of the probabilities that each two records are a true match, we can use these probabilities to choose which matches to use in the analysis. There are several ways to choose a match. One option, for example, is to just choose the match that yields the highest probability of being true. One issue with this approach, however, is that the highest probability can be low, for example 30% of being the true match. Even if the match with the highest probability is very likely (say 90% chance of being the true match), another issue is that there could be a second best match with very similar probability to be the true match (say 80%).[8] A better option is thus to only choose matches with high enough probability to be the true match (say 90%), for which the second best match is unlikely to be the true one (say below 15%). This option will also exclude observations that are non-unique, i.e. observations that have the exact same name and age combination.

Formally, this decision rule can be stated in the following way: To be considered a unique match for a record in dataset *A*, a record in dataset *B* has to satisfy three conditions. Specifically, the researcher should:

1. choose the match with highest probability of being a true match out of all potential matches for the record in A.

2. choose a match that is true with a sufficiently high probability, i.e. a match with a probability p_1 that satisfies $p_1>p$ for a given p in (0; 1] chosen by the researcher.

3. choose a match for which the second best match is unlikely, i.e. the match score of the next best match, denoted as p_2, satisfies $p_2<l$ for a given l in (0,p] chosen by the researcher.

Similarly, to be considered a unique match for a record in dataset B, a record in dataset A has to satisfy these three conditions.[9] Our linked sample is the set of pairs of records *(a, b)* in $A \times B$ for which: (1) *a* matches uniquely to *b*, and (2) *b* matches uniquely to *a*.

An additional assumption of the maximum likelihood procedure described above is that the observed distances are independently distributed from each other. In many economic history settings, this assumption will be violated because each observation in dataset *A* can be a match for at most one observation in dataset *B* (for instance, a one-to-one matching such as linking individuals across censuses). As a consequence, the algorithm does not require these probabilities to add up to 1. That is, for a given record in *A*, the sum of the probabilities across all potential matches in *B* will not in general add up to one. Hence, it is possible, for instance, to have a first best score of 0.8 and a second best score of 0.7. Indeed, in the empirical application below we will show the empirical distribution of estimated linking scores and show instances of second-best scores above 0.5 (which would be inconsistent with probabilities adding up to 1). This assumption can be relaxed, but doing so makes the estimation significantly more complex (Enamorado et al, 2017). Note that, even if the independence assumption does not hold in practice, the estimated linking scores are still useful as they guide researchers regarding how much weight each of the identifying variables (in our case, names and age distances) should have when classifying records as matches or non-matches. Moreover, as mentioned above, even if this assumption does not exactly hold in the data, high quality matches can still be achieved under this assumption (Herzog et al., 2007; Christen, 2012).

Which set of parameters should a researcher choose? Depending on the choice of values for the first- and second- best matches (*p* and *l)*, it is possible to generate samples with more or less

[9] We impose this symmetry condition because linking historical censuses is an example of one-to-one linking. Imposing this condition prevents situations in which a record *b* in *B* is the best candidate for a record *a* in *A*, but the best candidate for *b* in *B* is a different record *a'* in *A*.

confidence on the links. Intuitively, higher values of p and lower values of l will yield samples with fewer observations but higher average quality of the links. This possibility enables researchers to assess the robustness of their findings to the quality of the links.

When the main concern is to avoid false positives, we suggest two rules of thumb. First, we suggest researchers to choose l to be close to zero. This is a conservative choice because it implies that the second best match is very unlikely (probability close to zero). Second, because names are the most important source of identifying information, we suggest choosing p such that only records in which there is at least "partial agreement" (Jaro-Winkler distance below 0.12, as discussed above) in both first and last name will be linked. In any case, one useful tool (as shown by the empirical application below) to guide the choice of parameters is to plot a histogram of the estimated first and second-best probabilities. Such a histogram enables a visualization of where the mass of the distribution of first and second-best scores is located.

There are analogies between these decision rules and existing automated linking methods in economic history, such as Ferrie [1996] and Abramitzky et al. [2012, 2014, 2017]. When a method requires exact match of the names, it essentially requires that the first best match will have a high enough probability. Similarly, when a method requires uniqueness of the names within a five years window, it essentially requires that the second-best match will be unlikely. Requiring both exact match of names and uniqueness within a five years window is parallel to requiring both that the first best match has a high probability and that the second-best match is unlikely.

Indeed, the three steps described above allow researchers to make specific choices that will generate similar samples as other existing fully automated methods in economic history. Specifically, if: (1) in the first step of our method we block on a phonetic version (SOUNDEX or NYSIIS) of the first and last name, and (2) in the second step we use the EM algorithm to estimate probabilities based on age alone, and (3) in the third step we use a decision rule that picks the match with the highest probability (or, to mimic the more conservative approach, pick a match for which the second closest age is larger than two), then we are back to the traditional automated approaches.

One promising direction not discussed in this paper is how to adjust regression coefficients when dealing with imperfectly linked data. While there is a literature in statistics on this topic (see, for instance, Lahiri and Larsen [2005]), these methods are unfortunately still not directly applicable to the situations that typically arise in historical linkage problems. For instance, Lahiri and Larsen [2005] assume that all of the observations in one dataset have a potential link in the other, which does not hold when linking historical censuses due to mortality and underenumeration.

VI Application: Linking the US and Norwegian censuses using our method and IPUMS'

Next, we apply the method to create two linked samples: one linking the 1850 and 1880 US censuses of population, and one linking the 1865 and 1900 Norwegian censuses of population. We then use these data to construct father-son occupational transition matrices and to compute summary measures of intergenerational occupational mobility.

We chose to create these samples for three primary reasons. First, the most common datasets economic history papers attempt to link are historical censuses of population, making them especially attractive to test our methods. Second, IPUMS has constructed widely-used linked samples for both the US and Norway for these census years [Ruggles et al., 2011] using the exact same identifying information that we use to create our samples, hence enabling us to compare two algorithms that use the same information. Unlike our method, the method used by IPUMS is not fully automated. It starts by first manually classifying a subset of the records and then using these records to predict the classification status of the remaining ones. Finally, testing the method in two different countries enables us to assess how well our method does in two countries with different naming conventions, enumeration quality, outmigration rates, etc.

Creating linked samples

To create the US sample, we followed white males across the 1850 and 1880 US censuses of population.[10] To do so, we used the 1850 and 1880 full count US censuses available through the North Atlantic Population Project [Ruggles et al., 2011]. To construct the Norwegian sample, we followed males through the 1865 and 1900 Norwegian full count censuses. These two censuses are also available through the North Atlantic Population Project. In both cases, and as we discussed in section III, our linking was based only on predetermined characteristics: first and last names, place of birth and predicted year of birth.

As discussed above, to reduce the computational burden, we restricted our attention to pairs of individuals who: (1) reported the same place of birth[11], (2) had a predicted age difference of no more than five years in absolute value, and (3) had first and last names starting with the same letter. This blocking strategy attempted to avoid unnecessary comparisons between observations that were very unlikely to belong to the same individual.

Figure 3 shows the empirical distribution of the first best and second-best scores, both for the US (top panel) and Norway (bottom panel). There are three things worth noticing about these figures. First, the maximum estimated probability (which corresponds to two observations that agree exactly on all of their identifying information) is in both cases below one. Hence, a decision rule that imposes a very high level of p would essentially result in a sample with no observations. This is expected since, given that one observation can have multiple exact matches, no method can be sure about whether two observations are a match or not. Second, note that in a high fraction of the cases the first best probability is quite close to zero, indicating that all of the potential matches are quite unlikely. Third, note that the place of birth information is more detailed for Norway (municipalities) than for the US (states). This more detailed birthplace information makes multiple candidates for a match less likely in Norway, which explains why there is a sharper difference between the distribution of first-best matches and the distribution of second-best matches in Norway than in the US. More generally, this illustrates how having more detailed information on identifying variables facilitates uniquely linking individuals.

Figure 4 illustrate how our method can generate samples that are on different points of the type I vs. type II errors frontier. In this figure, we present the matching rate as a function of the cutoff

[10] The sample is restricted to whites because slaves, who constituted the majority of the US black population at the time, were not individually listed in the 1850 population census.

[11] Place of birth corresponded to states in the case of the US, municipalities in the case of Norway, country of birth for the foreign born in both countries.

for the first (x-axis) and second-best (y-axis) linking scores. The figure shows that our match rates can range from very low (less than 5%) to high (above 30%) depending on the choice of parameters.

With this trade off in mind, we created two linked samples, one using a more conservative choice of parameters and one using a less conservative choice. As our less conservative choice of parameters, we adopted the following decision rule: (1) we only kept observations for which the best match had a value of at least 0.6 and (2) the second-best match had a value of at most 0.3. As our more conservative choice, we only kept observations for which the best match is at least 0.7 and the second-best match is at most 0.1. These levels are indicated by the blue (less conservative) and red (more conservative) vertical bars in each panel of the figure. We note that there is little mass in the distribution of first-best scores between 0.3 and 0.6 and, similarly, there is little mass in the distribution of second-best scores between 0.3 and 0.6. Hence, the samples will be similar if we move the cutoffs up or down within that range.

When linking the Norwegian data, we had to deal with the fact that the patronymic naming scheme was still in place in Norway in the 19th century. Under this naming scheme, an individual received a last name based on the name of his or her father. For instance, the sons of William would receive the surname Williamson. We followed IPUMS in truncating the suffix in the patronymic surnames to minimize inconsistencies in the spelling of these suffixes.

Similarities and differences with IPUMS' linking method

The method used by IPUMS to generate their linked samples shares some similarities with the one proposed in this paper, but there are also some important differences.[12] Similar to us, the method starts by identifying a set of potential matches for each individual record[13], and then creates age and name similarity scores for each pair of potential links.

[12] This method is described in detail in Goeken et al. [2011].

[13] Unlike in our case, the method used by IPUMS does not block on first letter of first and last names, but rather just restricts the comparisons to individuals with a given race and birthplace. This coarser blocking dramatically increases

The key distinction with respect to our proposed method is that, after computing the similarity scores, IPUMS constructed a training sample of manually classified records.[14] In particular, data entry operators from the Minnesota Population Center classified a random subsample of potential links into matches and non-matches. Then, the remaining potential links were classified using a machine learning tool called Support Vector Machine, or SVM.[15] This tool uses information from the training sample to predict the classification status (matches or non-matches) of the remaining records. In this regard, the method used by IPUMS is close in spirit to the method discussed in Feigenbaum [2016a]. In contrast, our method does not rely on a training sample, and is thus cheaper and replicable.

The second distinction with respect to our samples is that, in the case of the US (but not in Norway), the IPUMS method started from a 1% sample rather than from the full 100% population data. That is, the IPUMS sample links a 1% sample of the 1850 census to the full count 1880 census, whereas we link a full count version of the 1850 census to a full count version of the 1880 census. Starting from a sample might be problematic. Assume there are two Ran Abramitzkys in 1850 US, but only one of them in the 1% sample. In 1860, one of the Ran Abramitzkys (the one who was originally in the 1% sample) decides to move outside of the US. By 1880, there will be just one Ran Abramitzky in the census, so a linking method that starts from a sample will likely link the unique (in the 1% sample) 1850 Ran Abramitzky to the unique

the number of calculations that need to be made. Nevertheless, in the IPUMS samples, about 98% of the individuals in the linked US data and about 92% in the Norwegian data agree on the first letter of both the first and last names. Hence, although the method does not explicitly block on these characteristics, in practice there are only few individuals in the resulting samples for which these characteristics do not agree. This is expected because the Jaro-Winkler similarity score, which is used as an input in the construction of the linked samples, has a larger penalty for mistakes that take place in the first letter of a word. Hence, names with such mistakes are unlikely to have a high estimated probability of being a true link.

[14] The procedure used to create the training sample is described in the following way: "For our project, we selected a random sample of potential links, and had a group of MPC data entry operators code each potential link as a "yes" or "no" based on a visual examination of names and ages of potential links (with yes indicating that it was in their opinion a true link). If a majority had the potential link as a "yes", then it was coded as a "yes" in the training data (with the remainder coded as "no")".

[15] As described in Goeken et al. (2011), "The SVM classifier analyzes the training data, plots them in a multidimensional space, and then constructs a boundary between the two classes of records that maximizes the distance from the hyperplane and the nearest data points in both of the classes (i.e., between the true and false links)."

1880 full count Ran Abramitzky, even though the two are different people. Note that immigration is not the only issue with a sample to population linking; any source of attrition from one census to the other (mortality, underenumeration) can generate a similar problem.

Comparison of match rates and representativeness

We next compare our linked samples to the IPUMS linked samples with respect to matching rates and representativeness. The IPUMS website does not explicitly report the matching rates for either the US or Norway, but we calculate these rates to be around 8% for the US and 15% for Norway.[16] As we discussed, the match rate in our case depends on the choice of parameters for the first- and second-best matches and can range from less than 5% to more than 30% (see Figure 4). In our more conservative sample (first best match at least 70% and second best at most 10%), our match rate is quite close to IPUMS' at 5% for the US and 12% for Norway. In our less conservative sample (first best match at least 60% and second best at most 30%), the match rate is 15% in the US and 24% in Norway.

To check representativeness, we compare our resulting linked samples to the population (using the non-linked cross sectional census data). In this exercise, we focused on our less conservative samples, but results are similar when focusing on the more conservative one. Specifically, we calculated the proportion in each occupational category of fathers in 1850 US and 1865 Norway. For example, in the entire US population 8.1% of fathers worked in white collar occupations and 59% were farmers. In our matched sample, these numbers are 9 and 66%, respectively, whereas in the IPUMS linked sample they are 9 and 64%. More generally, Figure 5 shows that while our linked sample is not completely representative of the population, it is very

[16] There are about 45,000 males aged 16 or less in the 1850 US census 1% sample, and about 3,500 in the 1850-1880 US linked sample. There are about 340,000 males aged 16 or less in the 1865 Norwegian census, and about 51,000 in the Norway 1865-1900 linked sample.

close. Moreover, the figure suggests that our linked sample is similarly representative as IPUMS'.[17]

Comparison of estimated intergenerational occupational mobility

We next use our data and IPUMS to compute rates of intergenerational occupational mobility. Specifically, we ask whether a researcher using linked samples constructed using these two different methodologies would have arrived to substantively different conclusion with respect to patterns of intergenerational mobility in this time period.

Tables 1 and 2 shows the father-son occupational transition matrices constructed using our linked samples and the IPUMS linked samples. Table 1 shows the data for the US 1850-1880 links, whereas Table 2 shows the corresponding Norway 1865-1900 links. As can be seen from the tables, both methods produce quite similar occupational transition matrices, both when linking US records and when linking Norwegian records. In most cases, the estimated percentage of sons who are in each occupational category is very similar across methods. As a result, both methods also generate a very similar occupational structure among sons in the later census year (last row of each matrix in each of the tables).

In table 3, we create summary measures of intergenerational occupational mobility using the linked samples. In panel (a), we report the simplest measure of occupational mobility: the fraction of sons working on a different occupational category than their father. In panel (b), we use instead the Altham statistic [Altham, 1970], which measures the distance of each occupational transition matrix with respect to a matrix representing independence (so that larger values imply higher departures from independence, that is, less mobility). This approach for measuring mobility is the one used in some recent economic history papers and is more appropriate when comparing

[17] The linked samples (both ours and the one built by IPUMS) might differ from the cross-section for reasons unrelated to the linking procedure. For instance, if there is differential mortality or outmigration by father's occupational category, the occupational distribution in the initial census year will differ from the cross section even if the method linked everyone who was still in the US by 1880/Norway in 1900.

countries with different occupational structures (see Long and Ferrie [2013] and Modalsli [2017] for further details).

In both the US and Norway, the fraction of sons working in a different occupational category than their father is similar when using the IPUMS linked samples than when using our linked samples (both in their more and less conservative versions). In the US, we estimate that about 45% of sons worked in a differential occupational category when using the less conservative sample, and 44% when using the more conservative sample (compared to 44% when using the IPUMS sample). In Norway, when using our linked samples we estimate that between 44 and 45% of sons in a different occupational category than their father, compared to 44% when using the IPUMS sample.

We next turn to analyze differences in estimated mobility across methods when using the measures based on the Altham statistic. The distance with respect to a matrix representing full independence is similar regardless of the linked samples that we use, both for the US and Norway. For the US, the estimated departure with respect to independence is 14.67 when using our linked sample, 15.18 when using our more conservative sample and of 17.37 when using the IPUMS sample. For Norway, the departure from independence is 25.94 when using our less conservative sample, 26.08 when using our more conservative sample, and of 25.01 when using the IPUMS sample.

Overall, while there are small differences in the magnitudes, the evidence indicates that researchers using any combination of these datasets would have arrived to the same conclusion: that the US had higher rates of intergenerational occupational mobility than Norway in the second half of the 19th century (as measured by the Altham statistic).

VII Conclusion

Fully-automated methods for linking historical records are transparent and easy to replicate. We suggest a fully automated method that adapts standard techniques from the statistical literature to the problem of historical record linkage. While this method is more computationally expensive

than automated methods based on simple name and age comparisons, it enables researchers to create samples at the frontier of minimizing type I and type II errors. A researcher can choose to create a sample with very low rates of false positives (at the cost of more false negatives), a sample with very low rates of false negatives (at the cost of more false positives), or anything in between. When applying our method to measure rates of intergenerational occupational mobility in historical US and Norway, we find that the estimates using our fully-automated method are remarkably similar to the ones using IPUMS' widely-used linked data.

References

1. Abramitzky, R., Boustan, L. P. & Eriksson, K. A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. *Journal of Political Economy* **122** (2014).

2. Abramitzky, R., Boustan, L. P. & Eriksson, K. Cultural assimilation during the age of mass migration. *National Bureau of Economic Research* (2016).

3. Abramitzky, R., Boustan, L. P. & Eriksson, K. Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. *American Economic Review* **102,** 1832–1856 (2012).

4. Abramitzky, R., Boustan, L. P. & Eriksson, K. Have the poor always been less likely to migrate? Evidence from inheritance practices during the Age of Mass Migration. *Journal of Development Economics* **102,** 2–14 (2013).

5. Abramitzky, R., Boustan, L. P., Eriksson, K. & Feigenbaum James J.and Pérez, S. Best Practices for Automated Linking Using Historical Data (2018).

6. Aizer, A., Eli, S., Ferrie, J. P. & Lleras-Muney, A. The long-run impact of cash transfers to poor families. *The American Economic Review* **106,** 935–971 (2016).

7. Altham, P. M. The Measurement of Association of Rows and Columns for an rXs Contingency Table. *Journal of the Royal Statistical Society. Series B (Methodological),* 63–73 (1970).

8. Atack, J., Bateman, F. & Gregson, M. E. "Matchmaker, Matchmaker, Make Me a Match" A General Personal Computer-Based Matching Program for Historical Research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* **25,** 53–65 (1992).

9. Bailey, M., Cole, C., Henderson, M. & Massey, C. *How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth* tech. rep. (National Bureau of Economic Research, 2017).

10. Bleakley, H. & Ferrie. Shocking behavior: Random wealth in antebellum Georgia and human capital across generations. *The Quarterly Journal of Economics* **131,** 1455–1495 (2016).

11. Bleakley, H. & Ferrie. Up from poverty? The 1832 Cherokee Land Lottery and the long-run distribution of wealth. *National Bureau of Economic Research* (2013).

12. Christen, P. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection* (Springer Science & Business Media, 2012).

13. Christen, P., Churches, T., *et al.* Febrl-freely extensible biomedical record linkage (2002).

14. Collins, W. J. & Wanamaker, M. H. Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data. *American Economic Journal: Applied Economics* **6,** 220–252 (2014).

15. Collins, W. J. & Wanamaker, M. H. The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants. *The Journal of Economic History* **75,** 947–992 (2015).

16. Collins, W. J. & Wanamaker, M. H. Up from Slavery? African American Intergenerational Economic Mobility Since 1880. *National Bureau of Economic Research* (2017).

17. Costa, D. L., Kahn, M. E., Roudiez, C. & Wilson, S. Data set from the Union Army samples to study locational choice and social networks. *Data in brief* **17,** 226–233 (2018).

18. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39,** 1–38 (1977).

19. Eli, S., Salisbury, L. & Shertzer, A. Migration responses to conflict: evidence from the border of the American Civil war (2016).

20. Enamorado, T., Fifield, B. & Imai, K. *Using a probabilistic model to assist merging of large-scale administrative records* tech. rep. (Technical Report. Department of Politics, Princeton University, 2017).

21. Eriksson, K. Access to Schooling and the Black-White Incarceration Gap in the Early 20th Century US South: Evidence from Rosenwald Schools. *National Bureau of Economic Research* (2015).

22. Feigenbaum, J. J. Automated Census Record Linking: A Machine Learning Approach. *mimeo* (2016).

23. Feigenbaum, J. J. Intergenerational Mobility during the Great Depression. *mimeo* (2016).

24. Feigenbaum, J. J. Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940. *Economic Journal* (2017).

25. Fellegi, I. P. & Sunter, A. B. A Theory for Record Linkage. *Journal of the American Statistical Association* **64,** 1183–1210 (Dec. 1969).

26. Ferrie. A New Sample of Males Linked from the Public Use Micro Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* **29,** 141–156 (1996).

27. Ferrie. The entry into the US labor market of antebellum European immigrants, 1840–1860. *Explorations in Economic History* **34,** 295–330 (1997).

28. Fouka, V. Backlash: The Unintended Effects of Language Prohibition in US Schools after World War I. *Stanford Center for International Development Working Paper* **591** (2016).

29. Herzog, T. N., Scheuren, F. J. & Winkler, W. E. *Data quality and record linkage techniques* (Springer Science & Business Media, 2007).

30. Hornbeck, R. & Naidu, S. When the levee breaks: black migration and economic development in the American South. *The American Economic Review* **104,** 963–990 (2014).

31. IPUMS. *IPUMS Linked Representative Samples, 1850-1930 Final Data Release* Minnesota Population Center, University of Minnesota.

32. Jaro, M. A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* **84,** 414–420 (June 1989).

33. Kosack, E. & Ward, Z. Who Crossed the Border? Self-Selection of Mexican Migrants in the Early Twentieth Century. *The Journal of Economic History* **74,** 1015–1044 (2014).

34. Lahiri, P. & Larsen, M. D. Regression Analysis with Linked Data. *Journal of the American Statistical Association* **100,** 222–230 (Mar. 2005).

35. Larsen, M. D. *Hierarchical Bayesian Record Linkage Theory* Aug. 2005.

36. Long, J. The Socioeconomic Return to Primary Schooling in Victorian England. *Journal of Economic History* **66,** 1026–1053 (Dec. 2006).

37. Long, J. & Ferrie. Intergenerational occupational mobility in Great Britain and the United States since 1850. *The American Economic Review* **103,** 1109–1137 (2013).

38. Massey, C. G. Playing with matches: An assessment of accuracy in linked historical data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* **50,** 129–143 (2017).

39. Meilijson, I. A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society. Series B (Methodological),* 127–138 (1989).

40. Mill, R. & Stein, L. C. Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America. *Working Paper, Stanford University* (2016).

41. Modalsli, J. Intergenerational Mobility in Norway, 1865–2011. *The Scandinavian Journal of Economics* **119,** 34–71 (2017).

42. Mosquera, A., Lloret, E. & Moreda, P. *Towards facilitating the accessibility of web 2.0 texts through text normalisation* in *Proceedings of the LREC workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)* (2012), 9–14.

43. Nix, E. & Qian, N. The Fluidity of Race:"Passing" in the United States, 1880-1940. *National Bureau of Economic Research* (2015).

44. Odell, M. & Russell, R. The soundex coding system. *US Patents* **1261167** (1918).

45. Parman, J. Childhood health and sibling outcomes: Nurture Reinforcing nature during the 1918 influenza pandemic. *Explorations in Economic History* **58,** 22–43 (2015).

46. Pérez, S. The (South) American Dream: Mobility and Economic Outcomes of First-and Second-Generation Immigrants in Nineteenth-Century Argentina. *The Journal of Economic History* **77,** 971–1006 (2017).

47. Philips, L. Hanging on the metaphone. *Computer Language* **7** (1990).

48. Ruggles, S. Intergenerational Coresidence and Family Transitions in the United States, 1850–1880. *Journal of Marriage and Family* **73,** 136–148 (2011).

49. Ruggles, S., Roberts, E., Sarkar, S. & Sobek, M. The North Atlantic population project: Progress and prospects. *Historical methods* **44,** 1–6 (2011).

50. Salisbury, L. Selective migration, wages, and occupational mobility in nineteenth century America. *Explorations in Economic History* **53,** 40–63 (2014).

51. Winkler, W. E. *Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage* in *Proceedings of the Fifth Annual Census Bureau Research Conference* (1989).

52. Winkler, W. E. Overview of Record Linkage and Current Research Directions. *U.S Bureau Statistical Research Division Research Report Series* **2** (2006).

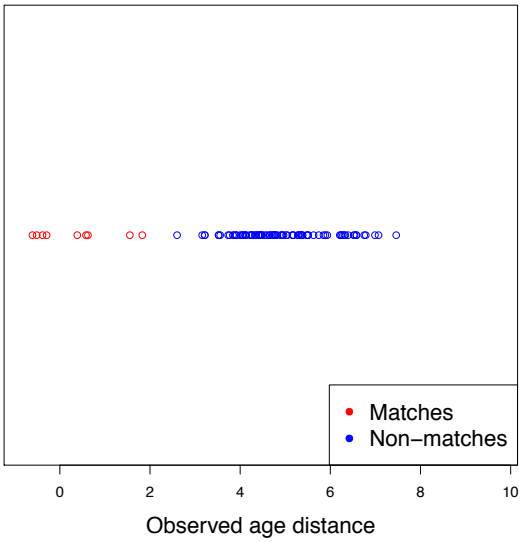
53. Winkler, W. E. *Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage* in *Proceedings of the Section on Survey Research Methods, American Statistical Association* **667** (1988), 671.

54. Xu, L. & Jordan, M. I. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural computation* **8,** 129–151 (1996).

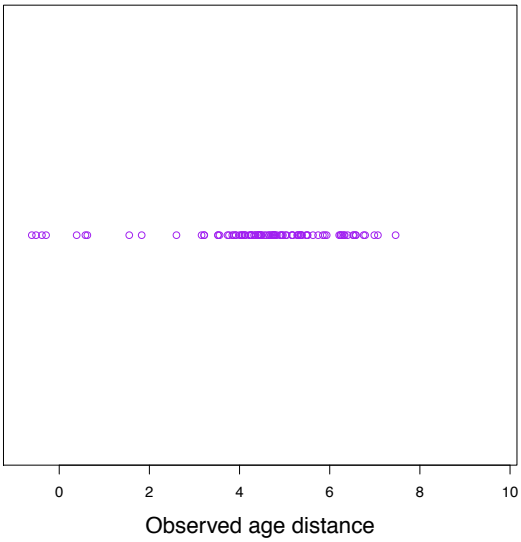
55. Yancey, W. E. *Improving EM Algorithm Estimates for Record Linkage Parameters* in *Proceedings of the Section on Survey Research Methods, American Statistical Association* (2002).

Figure 1: Illustration of the EM algorithm

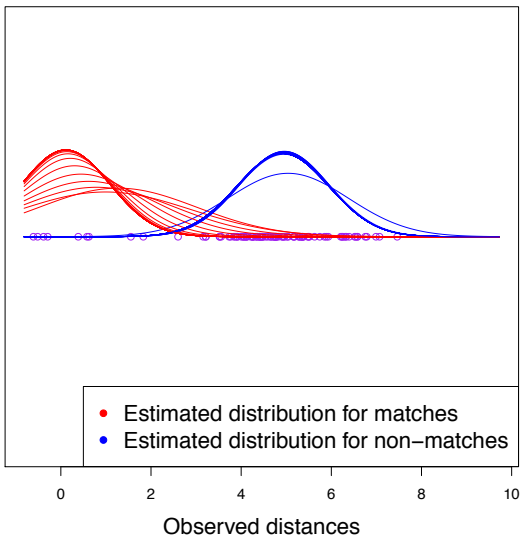
(a) If true matches were known



(b) Actual data (true matches are unknown)



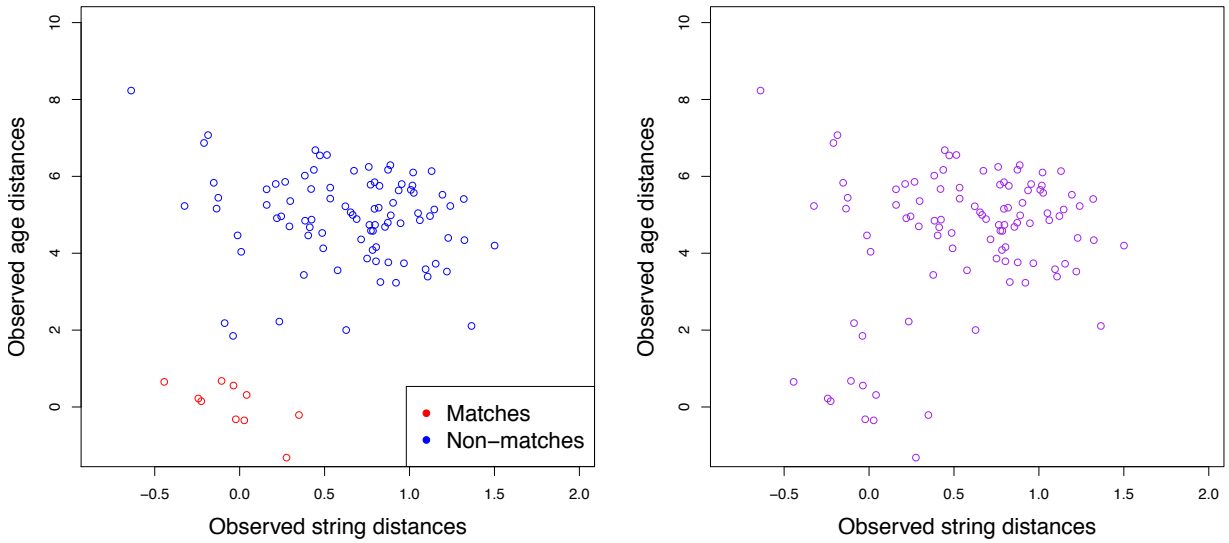
(c) Initial guess and convergence



Notes: This figure shows an hypothetical example that illustrates the EM algorithm. Panel (a) shows the situation in which the researchers knows whether the distances correspond to true matches or to true non-matches. Panel (b) shows the actual data, in which true matches are unknown. Panel (c) shows the estimated distributions under the assumption that the distances observed in panel (b) stem from two normal distributions, one corresponding to true matches and one corresponding to true non-matches.

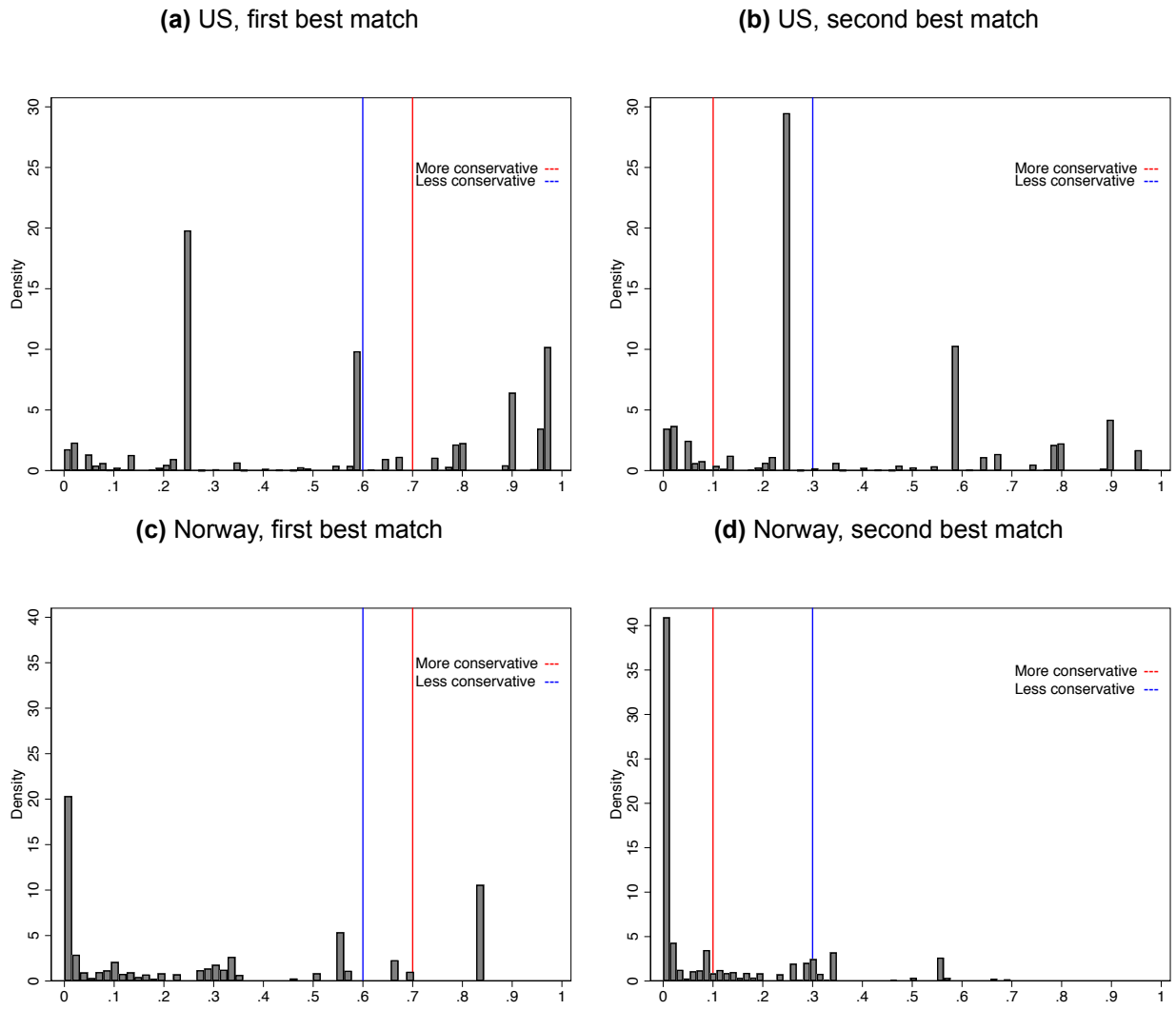
Figure 2: Illustration of the EM algorithm, two-dimensional case

(a) If true matches were known **(b)** Actual data (true matches are unknown)



Notes: This figure shows the case in which observations are compared to each other along two dimensions instead: reported ages and names. Panel (a) shows the situation in which the researchers knows whether the distances correspond to true matches or to true non-matches. Panel (b) shows the actual data, in which true matches are unknown.

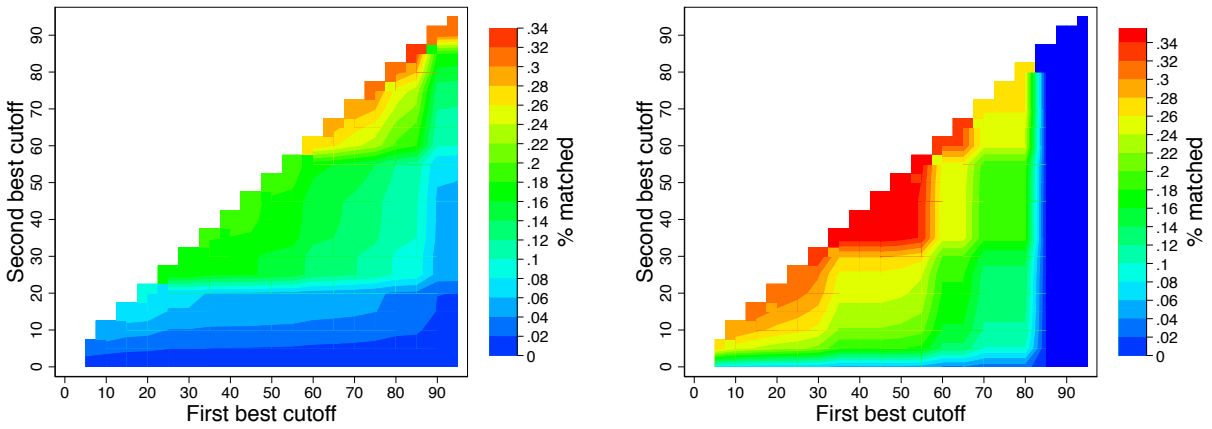
Figure 3: Empirical distribution of estimated linking scores



Panels (a) and (b) show the empirical distribution of the first best and second best linking scores when linking the 1850-1880 US censuses. Panels (c) and (d) correspond to the respective figures when linking the 1865-1900 Norwegian censuses. The vertical bars represent our parameter choices in the more (red) and less (blue) conservatively linked samples.

Figure 4: Matching rates as a function of first and second-best cutoffs

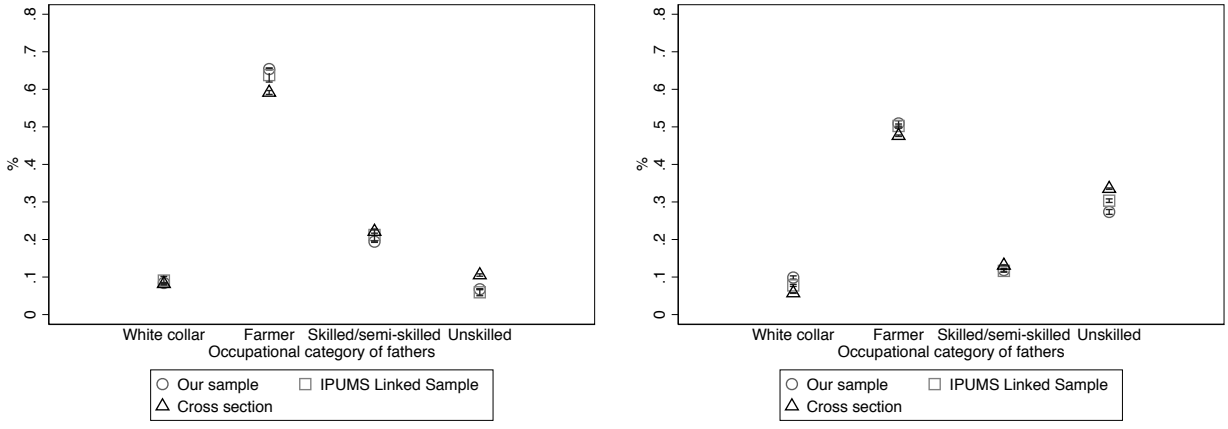
(a) US **(b)** Norway



Notes: This figure shows the matching rates in our linked samples as function of the cutoff first and second-best probability scores. A linking approach is more conservative the higher the first-best cutoff and the lower the second best cutoff.

Figure 5: Representativeness, comparison of our samples and IPUMS

(a) US **(b)** Norway



Notes: This figure shows the occupational structure among fathers in the initial census year (US 1850, Norway 1865) in our less conservative linked sample and in the linked samples compiled by IPUMS. Each proportion is reported around a 95% confidence interval.

Table 1: Comparison of occupational transition matrices, US 1850-1880

	Son's occupation				
Father's occupation	White collar	Farmer	Skilled/semi-skilled	Unskilled	Row total
Less conservative					
White-collar	0.53	0.21	0.19	0.08	1
	(7798)	(3094)	(2792)	(1129)	(14813)
Farmer	0.13	0.62	0.13	0.12	1
	(15282)	(72296)	(15470)	(13785)	(116833)
Skilled/semi-skilled	0.23	0.24	0.41	0.12	1
	(7881)	(8374)	(14084)	(4312)	(34651)
Unskilled	0.13	0.29	0.32	0.25	1
	(1601)	(3512)	(3903)	(3067)	(12083)
Column total	0.18	0.49	0.20	0.12	1
	(32562)	(87276)	(36249)	(22293)	(178380)
More conservative					
White-collar	0.56	0.21	0.16	0.07	1
	(2028)	(771)	(590)	(254)	(3643)
Farmer	0.13	0.64	0.12	0.11	1
	(3612)	(17198)	(3295)	(2941)	(27046)
Skilled/semi-skilled	0.24	0.25	0.40	0.11	1
	(1768)	(1835)	(2917)	(799)	(7319)
Unskilled	0.14	0.31	0.30	0.25	1
	(319)	(734)	(709)	(584)	(2346)
Column total	0.19	0.51	0.19	0.11	1
	(7727)	(20538)	(7511)	(4578)	(40354)
IPUMS					
White-collar	0.52	0.21	0.23	0.04	1
	(121)	(49)	(52)	(9)	(231)
Farmer	0.14	0.62	0.14	0.10	1
	(233)	(1035)	(232)	(166)	(1666)
Skilled/semi-skilled	0.23	0.26	0.40	0.11	1
	(127)	(140)	(219)	(60)	(546)
Unskilled	0.09	0.33	0.29	0.28	1
	(14)	(51)	(45)	(43)	(153)
Column total	0.19	0.49	0.21	0.11	1
	(495)	(1275)	(548)	(278)	(2596)

Notes: This table shows father-son occupational transitions constructed using our linked samples and the linked samples created by IPUMS.

Table 2: Comparison of occupational transition matrices, Norway 1865-1900

	Son's occupation				
Father's occupation	White collar	Farmer	Skilled/semi-skilled	Unskilled	Row total
Less conservative					
White-collar	0.80	0.05	0.11	0.05	1
	(1455)	(84)	(191)	(84)	(1814)
Farmer	0.09	0.62	0.14	0.15	1
	(813)	(5799)	(1325)	(1454)	(9391)
Skilled/semi-skilled	0.30	0.06	0.52	0.13	1
	(640)	(129)	(1116)	(277)	(2162)
Unskilled	0.10	0.24	0.30	0.36	1
	(481)	(1211)	(1473)	(1801)	(4966)
Column total	0.18	0.39	0.22	0.20	1
	(3389)	(7223)	(4105)	(3616)	(18333)
Conservative					
White-collar	0.82	0.04	0.09	0.04	1
	(1050)	(55)	(119)	(56)	(1280)
Farmer	0.09	0.61	0.14	0.15	1
	(491)	(3310)	(760)	(825)	(5386)
Skilled/semi-skilled	0.32	0.06	0.51	0.11	1
	(415)	(77)	(665)	(144)	(1301)
Unskilled	0.11	0.24	0.29	0.36	1
	(291)	(669)	(806)	(989)	(2755)
Column total	0.21	0.38	0.22	0.19	1
	(2247)	(4111)	(2350)	(2014)	(10722)
IPUMS					
White-collar	0.77	0.04	0.13	0.06	1
	(2192)	(126)	(358)	(173)	(2849)
Farmer	0.09	0.59	0.14	0.18	1
	(1645)	(11005)	(2595)	(3251)	(18496)
Skilled/semi-skilled	0.27	0.06	0.52	0.15	1
	(1133)	(267)	(2188)	(643)	(4231)
Unskilled	0.09	0.23	0.30	0.37	1
	(1028)	(2585)	(3309)	(4119)	(11041)
Column total	0.16	0.38	0.23	0.22	1
	(5998)	(13983)	(8450)	(8186)	(36617)

Notes: This table shows father-son occupational transitions constructed using our linked samples and the linked samples created by IPUMS.

Table 3: Comparison of summary measures of intergenerational occupational mobility

(a) Fraction working in different occupational category than father

US			Norway		
Less conservative	Conservative	IPUMS	Less conservative	Conservative	IPUMS
0.45	0.44	0.45	0.45	0.44	0.47

(b) Distance with respect to independence

US			Norway		
Less conservative	Conservative	IPUMS	Less conservative	Conservative	IPUMS
14.67 ***	15.18 ***	17.37 ***	25.94 ***	26.09 ***	25.01 ***

Notes: This table reports summary measures of mobility computed using the our linked samples and the linked samples created by IPUMS. Panel (a) reports the fraction of sons who worked in a different occupational category than their father (that is, the fraction of sons outside of the main diagonal in the transition matrix(. Panel (b) reports the mobility measures based on the Altham statistic. Higher distance with respect to independence indicates lower mobility. Significance levels are indicated by *** p<0.01, ** p<0.05, * p<0.1.