NBER WORKING PAPER SERIES

AUTOMATED LINKING OF HISTORICAL DATA

Ran Abramitzky
Leah Platt Boustan
Katherine Eriksson
James J. Feigenbaum
Santiago Pérez

Working Paper 25825
http://www.nber.org/papers/w25825

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2019

We are grateful to Jaime Arellano-Bover, Helen Kissel, and Tom Zohar for superb research assistance and useful comments and conversations, and to Horace Lee and Antigone Xenopoulos for help with data collection. We are grateful to Steven Durlauf (the editor) and six anonymous referees, as well as to Alvaro Calderón, Jacob Conway, John Parman, Laura Salisbury and Marianne Wanamaker for their most useful comments and suggestions. We are grateful to the Laura and John Arnold Foundation for financial support. We especially wish to thank Joe Price and Jacob Van Leeuwen from the BYU Record Linking Lab for comparing linkages made using our codes to the hand linkages made by users in the Family Tree Data on the FamilySearch.org website. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Automated Linking of Historical Data
Ran Abramitzky, Leah Platt Boustan, Katherine Eriksson, James J. Feigenbaum, and Santiago Pérez
NBER Working Paper No. 25825
May 2019, Revised June 2020
JEL No. C81,N0

ABSTRACT

The recent digitization of complete count census data is an extraordinary opportunity for social scientists to create large longitudinal datasets by linking individuals from one census to another or from other sources to the census. We evaluate different automated methods for record linkage, performing a series of comparisons across methods and against hand linking. We have three main findings that lead us to conclude that automated methods perform well. First, a number of automated methods generate very low (less than 5%) false positive rates. The automated methods trace out a frontier illustrating the tradeoff between the false positive rate and the (true) match rate. Relative to more conservative automated algorithms, humans tend to link more observations but at a cost of higher rates of false positives. Second, when human linkers and algorithms use the same linking variables, there is relatively little disagreement between them. Third, across a number of plausible analyses, coefficient estimates and parameters of interest are very similar when using linked samples based on each of the different automated methods. We provide code and Stata commands to implement the various automated methods.

Ran Abramitzky
Department of Economics
Stanford University
579 Jane Stanford Way
Stanford, CA 94305
and NBER
ranabr@stanford.edu

James J. Feigenbaum
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and NBER
jamesf@bu.edu

Leah Platt Boustan
Princeton University
Industrial Relations Section
Louis A. Simpson International Bldg.
Princeton, NJ 08544
and NBER
lboustan@princeton.edu

Santiago Pérez
Department of Economics
University of California at Davis
One Shields Avenue
Davis, CA 95616
and NBER
seperez@ucdavis.edu

Katherine Eriksson
Department of Economics
University of California, Davis
One Shields Avenue
Davis, CA 95616
and NBER
kaeriksson@ucdavis.edu

1. Introduction

The recent digitization of historical complete count population censuses and advances in computing power allow social scientists to create large historical panel datasets for the first time. These longitudinal datasets offer new evidence on questions such as the selection of immigrants (Abramitzky et al. 2012, 2013, Kosack and Ward 2014), how immigrants integrate into the economy and society (Abramitzky et al. 2014, 2019, Fouka 2019, Pérez 2017), how intergenerational mobility varied across groups and over time (Parman 2011, Ferrie and Long 2013, Feigenbaum 2015, 2018, Modalsli 2017, Collins and Wanamaker 2017, Abramitzky et al. 2019), the long-run effects of governmental programs and historical events (Aizer et al. 2016, Bleakley and Ferrie 2016, Salisbury 2014, Parman 2015, Eli and Salisbury 2016, Eriksson 2019), the geographic and economic mobility of African Americans (Collins and Wanamaker 2014, 2015, Hornbeck and Naidu 2014), and the return to human capital (Feigenbaum and Tan 2019).[1]

Linking historical data, however, introduces particular challenges. Because historical data lack unique identifiers such as Social Security Number, finding the same individual in two datasets requires using characteristics such as reported names and ages. Consider the following (real-world) linking problem: you want to match a man who is listed in the 1915 Iowa census as Paul Coulter, 3 years old, born in Kansas, to the 1940 Federal census. After much searching, you find two possible candidates in the 1940 census: Paul Coater, 28 years, born in Kansas, and Paul Courter, 29 years old, born in Kansas. Who, if anyone, would you choose as the correct match?

This example illustrates a recurrent challenge in historical record linkage: Common names, along with transcription and enumeration errors, age misreporting, mortality, under-enumeration and international migration between census years, often make it impossible to find the correct match with certainty (regardless of whether hand- or automated-methods are used).[2] In the face of this

[1] Historical Census data are released to the public 72 years after the survey was taken, including the personally identifying information like first and last name that is suppressed in modern public-use files. These historical censuses were fully digitized in a partnership between Ancestry.com and the Minnesota Population Center and have been made available for scholarly use.

[2] For instance, King and Magnuson (1995) report US census undercounts ranging from 6.5% in 1880 to 5% in 1940. Gould (1980) and Bandiera, Rasul, and Viarengo (2013) estimate rates of outmigration ranging from 20% to more than 50% among immigrants entering the US in the 1900-1920 period. Age misreporting is also pervasive in US historical censuses: Close to 30% of the individuals report an age ending with 0 or 5 in the 1860 census, although the true number should be closer to 20% (own calculation from IPUMS).

inherent challenge, a record matching method should therefore aim to satisfy four goals. First, a method should be *accurate*, making as few false matches as possible (minimize type I errors). Second, it should be *efficient*, creating as many of the true matches as possible (minimize type II errors). Third, it should be *representative*, generating linked samples that resemble the population of interest as closely as possible. Fourth, it should be *feasible* for most scholars to implement given current limitations of computing power and resources.

The main goals of this paper are to evaluate how various widely used automated methods of older and newer vintage perform on these metrics, suggest best practices for linking historical records, and provide user-friendly implementation codes. We evaluate three classes of the most widely used automated algorithms in economic history. First, we consider the algorithms developed by Abramitzky, Boustan, and Eriksson (2012, 2014, 2019; henceforth ABE), which are similar in spirit to Ferrie (1996) and suggest a simple fully automated linking approach that encompass a variety of different approaches to name comparisons (including using exact names, phonetic versions of names or "edit distance" measures between strings). Second, we test the machine-learning algorithm developed by Feigenbaum (2016), which uses a sample of manually classified records to train an algorithm to make matches like a human research assistant would at scale. Finally, we evaluate the fully automated probabilistic algorithm described in Abramitzky, Mill, and Pérez (2019), which uses the Expectation Maximization (EM) algorithm to combine age and name distances into a single score reflecting the probability that each potential pair of records is a true match.

A common feature of these approaches is that they link based on variables that are not expected to change over time, typically birthplace (in the US census: birth country for the foreign-born or birth state for the American-born), birth year, name and gender.[3] While adding other characteristics to the list of linking variables may increase the match rates and decrease false positive rates, it may also bias the economic analysis. For instance, using current county of residence or current

Mason and Cope (1987) show that, in the 1900 census (which was the first to ask individuals both their age and their year of birth), about 3% of the individuals reported inconsistent age and year of birth information.

[3] Current linking studies are often only able to systematically link men, because women tend to change their last name upon marriage (more recent studies that use marriage certificates that include information on maiden name also link women, see Craig, Eriksson, and Niemesh 2019). Olivetti and Paserman (2015) also proposed a group-based linking procedure to link daughters from childhood to adulthood.

occupations as identifying variables for linking can both significantly increase match rates and help us identify the true individual. However, using such a variable would result in excluding those who switched their county of residence and occupation from the analysis. This exclusion is an obvious issue in a study on geographical or occupational mobility, but could bias the results of many studies (see more discussion in section III of Abramitzky, Mill and Perez 2019).[4]

Another goal of this paper is to compare the performance of automated algorithms and manual linking. Linking by hand has the advantage that we instinctively trust other humans more than we trust computer algorithms, but hand linking is expensive, non-replicable (in the sense that any two scholars may link differently and even the same scholar might link differently at two different points in time), and impractical (consider linking the entire US population across two censuses by hand).[5] For example, the project of digitizing and linking the Union Army records that we use below took twenty years and several major grants to complete. In contrast, automated linking is rule based, cheap, and replicable. Moreover, a substantial body of evidence (going back to Meehl (1954)) argues that automated algorithms make more accurate and consistent decisions than humans along a wide variety of tasks.[6] Ultimately, when computers and humans use the same information, it is an empirical question whether they make similar links and which method performs better, relative to some benchmark.

One test of the accuracy of automated methods is to compare the links created by various algorithms with genealogical hand linkages. While we use high quality hand links, we hesitate to call these links "ground truth" because there is no way to know for sure what the true links are in this case. In the first exercise, we asked the Record Linking Lab at Brigham Young University (BYU) to check the quality of the links made by automated methods. The BYU lab linked the 1910 and 1920 US censuses using provided code and compared the links made by the automated

[4] Even linking based on race may be problematic when individuals selectively report their race (a pattern documented in Mill and Stein 2016 and Dahis, Nix and Qian 2019).

[5] In practice, there is no way to conduct hand linking without some use of a computer algorithm. Asking a human clerk to compare one record to all other records in the Census would be physically impossible. Instead, hand linking procedures first screen out "impossible" matches based on name and age similarity and then offer the human linkers a set of possible matches (usually fewer than 10).

[6] This literature is summarized in Kahneman (2011). Kahneman and Tversky (1973) coined the term "illusion of validity" to the refer to the cognitive bias through which people overestimate their ability to interpret and predict outcomes when analyzing a set of data.

methods to those made by users of the website FamilySearch.org. Specifically, the BYU team uses the Family Tree data, in which users of the platform connect individuals together on a wiki-style family tree (typically, their own family members and ancestors). To create these links, users attach historical documents (such as a US Census record) to that person as evidence of that person's existence or relationship. This genealogical linking has been considered the gold standard of hand linking (Bailey et al. 2017), although users of FamilySearch tree might not be representative of the entire US population. The Record Linkage Lab found that ABE links (EM links) agree with users of the Family Tree data in over 95% (97%) of cases. This implies a rate of "false positives" of less than 5% between the genealogical links made by humans and those made by the automated methods.

In the second exercise, we use a similar method to compare the links created by a wider set of algorithms to data linked from the Union Army Records to the 1900 US Census, which were carefully (and expensively) hand collected using trained research assistants who had access to extra information not typically available for linking (Costa et al. 2017). Treating these data as a benchmark, we compare the relative performance of automated and hand linking methods that use only typically-available variables (that is, names, year of birth, and state or country of birth). Unlike comparisons with the Record Linkage Lab (which links a full count census to a full count census), here the linking is from a sample (the Union Army records) to a full count census, which will tend to increase false positive rates.[7] We find that these widely-used automated methods lie along a frontier that illustrate the tradeoff between type I and type II errors. Researchers can choose to use algorithms that generate very low "discrepancy rates" from these high-quality benchmark links, which we refer to here as the "false positive rate" (as low as 5-10% in the context of the Union Army records). However, achieving a low false positive rate comes at a cost of accepting a

[7] The following example illustrates why this might be the case: Assume there are two John Smiths in 1880 US, but only one of them is in the Union Army records. In 1890, one of the John Smiths (the one who was originally in the Union Army records) moves out of the US. By 1900, there will be just one John Smith in the census, so a linking method that starts from a sample will likely link the unique (in the Union Army records) 1880 John Smith to the unique 1900 full count John Smith, even though the two are different people.

relatively low (true) match rates (10-30%). Alternatively, researchers can choose algorithms with higher (true) match rates (50-60%) at a cost of higher discrepancy rates (15-30%).[8]

Hand linking that relies on the typically-used linking variables (name, age and place of birth) is also on this frontier, producing "false positive" rates of around 25% and (true) match rates of around 63%. Not surprisingly, "false positive" and true match rates for hand links look very similar to links created by machine learning algorithms that are trained on hand-linked data. Relative to more conservative automated methods, humans using typical linking variables tend to match more observations than automated methods but at a cost of higher rates of false positives. When humans and machine use the *same* information for linking, automated methods have very low discrepancy rates relative to hand links.

We next use data from two different transcriptions of the 1940 Federal Census, one transcribed by FamilySearch and one by Ancestry.com. In this case, we can establish a real "ground truth" because records listed on the same census manuscript page and line number are known to refer to the same individual but the set of possible errors is more limited (transcription errors are still possible, but mortality and return migration are not). We find that differences in transcription for names (but not ages) are generally high, particularly for the foreign born from non-English speaking countries. Between 7-14% of first names (and 17-32% of last names) have at least one-character difference in the two transcribed versions.[9] Despite transcription differences, we find that automated methods produce links between these two versions of the 1940 Census that are almost 100% correct. We note that even when linking a census to itself, we can only link 43-67% of the observations: Non-unique records due to common names like John Smith can explain around 55 percent of non-matches, whereas transcription errors explain the remaining 45 percent. This suggests that there is an upper bound for match rates of any method, either with automated or hand linking.

[8] Depending on specific choices of name strings and age differences, the more conservative ABE algorithms generate samples with "false positive" rates of about 10% and (true) match rates of about 25-30%, and their less conservative algorithms generate samples with about 20-25% "false positive" rates and about 35% (true) match rates. The ML algorithms generate samples with "false positive" ranging from about 17-30%, and 50-60% (true) match rates, and the EM algorithms generate samples with about 5-15% "false positive" rates with a 5-30% (true) match rates range.

[9] We compare one version of the 1940 Census to the other by line and page number.

Ultimately, the goal of constructing linked samples is to conduct economic analyses. We study how automated linking methods affect inference using two samples. First, we examine the sensitivity of regression estimates to the choice of linking algorithm using linked data from the 1915 Iowa Census to the 1940 Federal census. These data allow us to study a set of typical regressions documenting intergenerational mobility between fathers and their sons. Here we do not have a proxy for ground truth, but we can compare the results we obtain in samples linked by hand to those we obtain in samples linked using automated methods. For multiple outcome and explanatory variables, we find that parameter estimates are stable across linking methods, with parameter estimates using automated and hand links similar in magnitude and well within each other's 95% confidence intervals. For example, in the case of intergenerational income elasticity, our weighted estimates mostly range between 0.15 to 0.20. This stability is not surprising, because we also find that human linkers and automated methods agree in over 90% of cases. In the few cases of discord, it is not clear from inspection whether computer algorithms or hand linkers is correct.

Second, to evaluate automated methods in another setting and across countries, we use data linked from the 1850 to the 1880 US censuses of population, and from the 1865 to the 1900 Norwegian population censuses, comparing our links to the widely-used linked samples constructed by IPUMS (Goeken et al. 2011). We measure intergenerational occupational mobility (another typical use of linked data) and find that in both the US and Norway the automated methods generate very similar measures of intergenerational mobility to the ones computed using the IPUMS linked samples.

Overall, we conclude that automated methods perform well. Automated methods involve a tradeoff between the number of matches made and the accuracy of the matches. It is possible to select an algorithm that will generate samples with very low rates of false positives, and estimates using different automated methods are in most cases stable. Throughout the paper we also provide general guidance for researchers and offer practical tips. Our overarching advice to researchers using linked historical data is to create alternative samples using various automated methods and test the robustness of results across samples. Whenever high quality hand linked data can be created, the researcher should consider using them as well (Bailey et al. 2019), although we note

that such data are often impractical and expensive to create, and that even well trained hand linkers can make errors.[10]

2. Linking algorithms

In this section, we briefly describe the automated record linkage algorithms evaluated in this paper, with an eye towards making them useful to practitioners. The goal of these algorithms is to link individuals from one set of records (dataset A) to another set of records (dataset B). Although the discussion below is focused on the case in which the linking is solely based on predetermined characteristics (names, place of birth and year of birth), we emphasize that all of these algorithms can be easily modified to use additional characteristics for linking (for instance, place of residence or information on other family members).

Implementation code and Stata commands for all of these algorithms can be found on our webpage at: https://people.stanford.edu/ranabr/matching-codes. We also provide a general coding structure to facilitate the adaptation of the codes to specific linking projects.

[10] Our paper was written in parallel to Bailey et al. (2017, 2019), which evaluates some of the same historical linking methods. Yet, our takeaway about the usefulness of automated methods is more positive that Bailey et al (2019) for a number of reasons. First, we show that links done by automated methods are very similar to the best available hand links created by users of FamilySearch.org website. Second, in real ground truth data based on comparing two independent transcriptions of the 1940 census, we find that while automated methods miss many links because of transcription differences, the links they create are almost 100% correct. Third, Bailey et al (2019) continues to report results that mix the Abramitzky, Boustan, and Eriksson approach with an outdated name standardizing algorithm (Soundex) that is not used in contemporary linking papers (see Koneru et al. (2016), which suggests that NYSIIS result in fewer false positives than Soundex). Not surprisingly, this method is reported to have the highest false positive rates (43%). Fourth, Bailey et al. (both 2017 and 2019) reports the highest error rates in an exercise that assumes that hand links performed by well-trained research assistants are ground truth (LIFE-M). In the LIFE-M data, hand linkers matched 45% of the records. Bailey et al. (2017, 2019) assume that any match made by an automated method in the remaining 55% of the data is necessarily wrong, but this assumption cannot be evaluated in the absence of real ground truth data. Finally, Bailey et al. (2017, 2019) also classifies any discordant match for which the automated method selects a different match than the hand linkers to be a "false positive" but it is also possible that the hand linkers selected the incorrect pair. We also note that our findings are more in line with the lower false positive rates reported in Bailey et al. (2019) than with the higher false positive rates reported Bailey et al. (2017). See our NBER Working Paper version (2019) for additional reasons why the false positive rates of automated methods reported in Bailey et al. (2017) are high.

One promising method that we do not evaluate is a probabilistic matching approach, whereby instead of making unique matches, the researcher decides to use several possible matches and adjust the analysis accordingly (see, for instance, Lahiri and Larsen 2005 and Poirier and Ziebarth 2019). One advantage of such a method is that no observations are dropped, so uniquely matched observations can be used together with observations that have multiple matches. This could, potentially, lead to efficiency gains. This method has not been used widely, perhaps because it is challenging to apply to historical linkage. For instance, Lahiri and Larsen (2005) assume that all observations in one dataset have a potential link in the other, which is usually not the case when linking historical censuses because of mortality and under enumeration (Abramitzky et al. 2019). More research is needed on the conditions required for such methods to work well in the historical context, although tools discussed in our paper could be used to identify the set of possible matches and their relative likelihood.

2.1. Abramitzky, Boustan and Eriksson (ABE) algorithms

The approach is described in more detail in Abramitzky, Boustan, and Eriksson (2012, 2014, 2019), and is similar in spirit to Ferrie (1996). Here are the basic steps:

1. Clean names in datasets A and B to remove any non-alphabetic characters and account for common misspellings and nicknames (e.g. so that Ben and Benjamin would be considered the same name).
2. Restrict the sample to people who are unique by first and last name, implied birth year calculated from calendar year and age, and place of birth (either state or country) in dataset A.[11]
3. For each record in dataset A, look for records in dataset B that match on first name, last name, place of birth, and exact birth year. At this point there are three possibilities:
    a. If there is a *unique* match, this pair of observations is considered a match.
    b. If there are multiple potential matches in dataset B with the same year of birth, the observation is discarded (it is impossible to tell which potential match is correct).

[11] We note that historical US censuses did not include a question on year of birth, but rather only asked people about their current age. To the extent that people misreport their age, we only observe an imperfect proxy of the actual year of birth.

c. If there are no matches by exact year of birth, the algorithm searches for matches within ± 1 year of reported birth year, and if this is unsuccessful, it looks for matches within ± 2 years. In each of these steps, only unique matches are accepted. If none of these attempts produces a unique match, the observation is discarded.

d. In an updated version of this method, this procedure is then done for each record in dataset B, after which the intersection of the two matched samples is taken.[12]

The steps described above represent the basic structure of the algorithm.[13] However, the papers that implement this approach typically also implement the following variations of the basic algorithm to check the robustness of the results.

1. Requiring matches on exact year of birth. In the standard process matched pairs are allowed to differ by up to 2 years in reported year of birth. Alternatively, matched pairs can be required to have the exact same reported year of birth to minimize the chance of false positives. However, this will result in a smaller matched sample, and will increase the number of Type II errors.

2. Requiring names to be unique within a 5-year band (within ± 2 years of the implied birth year). In historical records, reported age is an imperfect measure of true year of birth (due to misreporting, rounding, and timing of census enumeration). In the standard matching process, we attempt to link a record as long as there is no one else with the same name and same birth year. In this robustness check we only attempt to match a record if there is no one else with the same name who was born within +- 2 years of the implied birth year. Given this more restrictive uniqueness requirement, there will be fewer false positives but

[12] An example is given in the Appendix of a match that could have been made in the original ABE code but which has since been fixed. Because of this possibility, it is necessary to do the matching in both directions and take the intersection of the resulting matches. The updated code is not used in the papers cited, but is used for the **abematch** Stata command. We posted replication files of all of the relevant papers on https://ranabr.people.stanford.edu/matching-codes.

[13] In practice, this approach runs faster for large datasets when looping over blocks. For example, in the United States, it is more efficient to block first by birth state or country and then append the resulting datasets together. In addition, the Stata command can make multiple robustness samples, as described below, at once so the matching only needs to be done once if matching based on the same variables.

the number of (true) matches created will also be lower.[14] Two versions of this restriction are possible. The first version introduces a step 0 that flags observations in each dataset that are unique within ± 2 years of birth and then keeps only those observations after the match has occurred. The second version also flags matched cases in each dataset that do not have another potential match within the ± 2 years age band in the other dataset.[15]

3. Using NYSIIS (New York State Identification and Intelligence System) standardized names. Another concern with historical records is misspelling and mis-transcription of names. This risk can be exacerbated when focusing on immigrants with foreign names that census enumerators may not be familiar with. One way of accounting for this is to use the NYSIIS standardized names, rather than exact names, in the matching procedure. The NYSIIS phonetic algorithm standardizes names based on their pronunciation so that names can be matched even if there are minor spelling differences.

4. Jaro-Winkler adjustment. Jaro-Winkler string distance gives a measure of the similarity of two strings, placing more weight on characters at the beginning of the strings. The measure is based on the "edit distance" between two strings (that is, the number of changes that need to be made to one string to convert it into the other). An alternative to using NYSIIS standardized names is to compute the Jaro-Winkler string distance between the first and last names of all potential matches within each dataset. The matching process can then consider any two records with a string distance below a given cutoff to be a match. We describe this approach in more detail below.

[14] For instance, imagine that two men named James Alexander were born in 1892, but dataset A incorrectly reports that one of these men was born in 1890. Both men appear to be unique by name and exact year of birth in dataset A, and would be used in the standard matching process. Because in reality these men were born in the same year, they are non-unique and should not be considered in the matching process (it would be impossible to tell when James Alexander is the correct match).

[15] Generally, the first and second versions of this algorithm match the same set of people. However, here is an example of how the two versions could produce different matches. Imagine that there are three men named Alex Smith in dataset A with reported birth years of 1890, 1891, and 1894, and only one Alex Smith in dataset B with a birth year of 1892. In the first version, the men in dataset A with a birth year of 1890 and 1891 will be dropped, but Alex Smith born in 1894 will remain since he is unique within ± 2 years of his birth year. In this case we would match Alex Smith, born in 1894 to the Alex Smith in dataset B, born in 1892. However, in the second version, we would require that Alex Smith in dataset B have only one potential match within a ±2 year band in the other dataset. Given that there are three potential matches within ± 2 years, no successful match will be made.

5. Adding middle names/initials as a linking characteristic for those who have them.[16] In cases where no middle name is reported, we link on the full name. An alternative is to truncate all middle names to a single initial and then to match on middle initials only because the use of full middle names is very unstable across Census waves. Indeed, Census enumerators were instructed to only record middle initials, rather than full name.[17]

2.2. Abramitzky, Boustan, and Eriksson algorithm using string comparators (ABE-JW)

A limitation of the ABE algorithm described above is that it relies on matching observations based on either exact or standardized names. This reliance has two potential limitations. First, match rates could be lower, because cases in which names do not coincide exactly either in their raw or their standardized version might be very similar and could correspond to the same individual, yet would be dropped by the algorithm. For instance, "Santiago Perez" and "Santiago Perz," which are close to one another in string distance and may reflect a simple enumeration or transcription error, would count as a different name both when using raw and NYSIIS standardized names. Second, relying on standardized names might lead to higher rates of false positives because the name standardization may treat two very different names as the same string. For instance, the names "Thomas Calemon" and "Thomas Colenan" seem sufficiently different that they are unlikely to be the same person but would be assigned the same NYSIIS standardized name, "Tan Calanan." To address this limitation, we follow recent literature in economic history and incorporate a string comparator to the linking procedure—the Jaro-Winkler (JW) score.[18] More specifically, we implement the following procedure (Abramitzky, Boustan, and Eriksson 2019):

1. For each observation in dataset A, we identify a set of *potential* matches in dataset B. An observation in dataset B is a potential match for an observation in dataset A if:
    (a) It has the same place of birth.
    (b) The predicted year of birth is within minus/plus five years of the reported year of

[16] There is a wide variation in the use of middle names across populations. As many as 9.75% of enlistees in the Union Army had a middle name. In contrast, the foreign born in the US in 1900 were very unlikely to use middle names (0.47%). We recommend that researchers choose to use or not use middle names in their matching algorithms according to context.

[17] See instruction # 124 in 1900 here (https://usa.ipums.org/usa/voliii/inst1900.shtml).

[18] See, for instance, Mill and Stein (2016), Feigenbaum (2018), Nix and Qian (2015), Perez (2017, 2019).

birth in dataset A.[19]

(c) The first letter in the first and last names are the same first letters as the observation to be matched.[20]

2. For each pair of potential matches, we compute the Jaro-Winkler score for the first and last names (and middle name, if available). The JW score ranges from 0 to 1. We normalize the JW score such that 0 corresponds to two identical strings and 1 corresponds to two strings with no common characters.[21]
3. For each pair of potential matches, we define a "name match" as a pair of observations with a JW score in the first, (middle) and last names less than or equal to 0.1.[22] At this point, there are three possibilities:
   (a) We do not find any "name match" for a given observation in dataset A. In this case, we consider it "unmatched" and drop it from the analysis.
   (b) The observation in the source data (dataset A) has a unique "name match" in the destination data, and the observation in the destination data (dataset B) has a unique "name match" in the source data. In this case, we incorporate the observation to the analysis.
   (c) We find more than one "name match" for an observation in the source data. In this case, we use the following decision rule:

[19] We choose this rule of +/- five years for *potential* matches even though we only consider +/-2 years for *actual* matches to avoid cases like the following: Ran Abramitzky is 20 years old in dataset A and 22 years old in dataset B, and there is another Ran Abramitzky who is 23 years old in dataset B. In dataset B, the two Ran Abramitzky are only 1 year apart, so a more conservative choice is not to match either one.

[20] The goal of imposing this condition is to dramatically reduce computational time. It can be relaxed when the datasets are relatively small. Moreover, an alternative to blocking on the first letter of the exact first and last names is to block on the first letter of the *standardized* first and last names (so as to allow for matches between a name like "Katherine" and a name like "Catherine").

[21] Stata's "jarowinkler" command produces string similarity scores rather than distances so exact matches are actually scored at 1 and strings without common characters are scored at 0, but we follow the method used in the R package stringdist.

[22] The method does not require this cutoff to be exactly 0.1. However, in practice, 0.1 is a reasonably conservative cutoff. For instance, Feigenbaum (2016) shows that 95% of the links made by IPUMS have a Jaro-Winkler distance in first and last names below 0.2. To give a sense of the scale of Jaro Winkler scores, the strings "Smith" and "Smyth" have a Jaro-Winkler distance of 0.106, and the strings "Frederick" and "Francis" have a Jaro-Winkler distance of 0.329.

i. We compute the absolute value of the age difference with respect to the observation with the closest age (let's call this difference *d1*).
ii. We compute the absolute value of the age difference with respect to the observation with the *second* closest age (let's call this difference *d2*).
iii. We keep only individuals such that $d2-d1>x$, such that larger values of x represent more conservative linking choices. In other words, we choose the closest observation with respect to age, but only if the second closest observation is "far enough" away, where the value of x determines how far the second closest observation must be to create an acceptable match.

4. Repeat steps 1-3, only matching observations from dataset B to dataset A (instead of from A to B).
5. Our sample is comprised of the intersection between the observations that uniquely link from dataset A to dataset B and the observations that uniquely link from dataset B to dataset A.
6. Because the JW distance measure is not transitive, in some cases we will find a unique match for a record in dataset A, even though there is another record in dataset A with the same reported age and within 0.1 JW distance in name.[23] To account for this, we additionally require that for each successful match there is no other individual with $\leq$ 0.1 distance in name within $\pm x$ years of reported age.

We typically show results using $x = 0$ and $x = 2$, and allow matched observations to differ by up to 2 years in reported year of birth. The case when x = 0 is parallel to the standard ABE method, and in both cases we consider the potential match closest in age to be a successful match, so long as this match is unique. Using x = 2 is parallel to the ABE method with a 5-year uniqueness band. In this case we require that each record have only one potential match within ±2 year of reported age, and additionally that each individual is unique by name or JW string distance within ±2 years in his own dataset.

[23] As an example, imagine that dataset A contains the men "John Anders" and "John Aders" born in 1890 and there is a man named "John Andersson" in dataset B, also born in 1890. The last names "Anders" and "Aders" are within 0.1 JW distance, and the last names "Anders" and "Andersson" are also within 0.1 JW distance. However, the names "Aders" and "Andersson" are not within 0.1 JW distance. As a result, "John Anders" is the only potential match for "John Andersson."

One limitation of the class of Abramitzky, Boustan, and Eriksson algorithms is that it is not clear how to appropriately weight differences in name spelling versus differences in age when comparing two records. For instance, the basic approach requires names (or their standardized version) to match exactly, but enables age to differ by up to two years, whereas the JW considers string similarity first (that is, only considers as potential matches those within a 0.1 JW distance) before screening on age differences. This limitation prompted the development of two additional linking methods that estimate the optimal weight for each of these differences and other record match discrepancies. The goal of these methods is to "let the data speak" with respect to which features (differences in names, and differences in year of birth) should be weighted more heavily, and then combining these differences into a single linking score. We describe these two methods below.

2.3. Fully automated probabilistic approach (EM)

The EM method is the first method designed to systematically weigh name versus age differences in a fully automated probabilistic approach for record linkage. This approach is described in detail in Abramitzky, Mill, and Perez (2019) and has the following steps:

1. For each observation in dataset A, we identify a set of potential matches in dataset B. An observation in dataset B is a potential match for an observation in dataset A if:
    (a) It has the same place of birth.
    (b) The predicted year of birth is within minus/plus five years of the reported year of birth in dataset A.
    (c) The first letter in the first and last names are the same first letters as the observation to be matched.
2. For each pair of potential matches, we compute measures of similarity in their reported year of birth and name. To measure similarity in names, we compute the Jaro-Winkler score for the first and last names.[24] To measure similarity in reported ages, we compute the absolute value of the difference in reported years of birth.[25]

[24] In principle we could compute the JW score for middle names as well, which possibly could reduce false positives further.

[25] Note that steps 1 and 2 are shared with the ABE-JW approach.

3. We combine distances in reported names and ages between each record in dataset A and its potential match in dataset B 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 1988). Abramitzky, Mill and Perez (2019) provides the logic and intuition for this estimation.[26]
4. We use these scores to inform our decision rule on which records to use in the analysis. Specifically, to be considered a unique match for a record in dataset *A*, a record in dataset *B* must satisfy three conditions:
   (a) choose the match in dataset B with the highest probability (score) of being a true match out of all potential matches for the record in dataset A.
   (b) choose a match that is true with a sufficiently high probability (score), i.e. a match with a probability p_1 that satisfies $p_1>p$ for a given p in (0,1), with p chosen by the researcher. Intuitively, the higher the value of p that the researcher chooses, the lower the likelihood of false positives.
   (c) 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). with l also chosen by the researcher.

[26] 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 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. To compute these probabilities, 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.

Similarly, to be considered a unique match for a record in dataset B, a record in dataset A has to satisfy these three conditions.[27] 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*.

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 confidence in 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. When the main concern is to avoid false positives, we suggest two rules of thumb. First, we suggest to choose a low value of *l*. For instance, in the applications below we construct "conservative" samples with l=0.3*,* which in our data corresponds to pairs of observation with a Jaro-Winkler distance of at least 0.12 in both the first and last names and an age distance of 3. 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" (in our current implementation, a Jaro-Winkler distance below 0.12) in both first and last name will be linked.

2.4. Machine learning algorithm (ML)

The second method designed to systematically weight name versus age differences is a machine learning approach to record linkage, using hand linked data to train the algorithm on how much (or little) to penalize a potential match based on certain discrepancies in record features. This approach is explained in detail in Feigenbaum (2016).[28] The method has the following steps:

[27] 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 dataset *B* is the best candidate for a record *a* in dataset *A*, but the best candidate for *b* in dataset *B* is a different record *a'* in dataset *A*.

[28] Price et al. (2019) outline a new and promising tactic to create census linked samples. The method applies a supervised machine learning approach as in Feigenbaum (2016). But rather than rely on RA-linked data, Price et al. (2019) train their algorithm on data linked by the public on the wiki-style family tree hosted by FamilySearch.org. Using potentially high quality linked data – matched by descendants and possibly relying on information that goes well beyond the fields in any given census – the method performs quite well on our standard measures of recovering true matches (PPV) and minimizing false positives (TPR), as described in the next section. Price et al. (2019) link between the 1900, 1910, and 1920 Censuses. Although more study on the representativeness of the FamilySearch tree – or how much any non-representativeness in training data could matter – broader access to the FamilySearch tree could make this sort of strategy an attractive matching technique in the future. Of particular interest: the FamilySearch tree training data contains links of women over time (and across marital statuses), though such links could only be automated in the presence of extra data (marriage certificates, for example).

1. For each observation in dataset A, we identify a set of potential matches in dataset B using the rules articulated below. In principle, any record in B could be a match for a record in A, but to reduce computational complexity, the following screen is applied. An observation in dataset B is a potential match for an observation in dataset A if:
    a. It has the same place of birth.
    b. The year of birth distance between the two records is less than three years
    c. The Jaro-Winkler distance between the first names of the two records is less than 0.2.
    d. The Jaro-Winkler distance between the last names of the two records is less than 0.2.[29]
2. Build a training dataset on a small share of these possible links and use the training data to tune a matching algorithm. To do this, a trained human researcher views a record in dataset A and the set of possible matches in dataset B and selects a match, if the researcher is confident such a match is correct. If no hand match can be made, either because all possible candidate matches are too different or because there are multiple matches deemed to be equally likely, then no match is recorded.[30]
3. We fit a probit model to a reshaped version of the training dataset. Each observation in the reshaped data is a pair of records from dataset A and dataset B that could be a match, filtered as in step 1. The outcome is a simple indicator variable: 1 if the human researcher matched these records to one another, 0 if not. The right-hand-side variables are record linking features. These features include the main variables used by other methods such as Jaro-Winkler string distance in first name, the Jaro-Winkler string distance in last name,

[29] Specifically, the optimal filter values on Jaro-Winkler distance or year of birth distance may be larger for data with high rates of transcription error. For example, when working with historical datasets with many first names reported as initials (J.J. rather than James J., for example), including any pairs with either Jaro-Winkler distance less than .2 OR agreement in the first letter of first name string could prevent true matches from escaping the initial filter. Similarly, working with samples with more or less age heaping or age misreporting could push the researcher to change the filtering.

[30] We provide an example of the full instructions given to RAs creating the training sample at https://jamesfeigenbaum.github.io/linking/. The algorithm will "learn" from this researcher-produced training dataset. If the trainer is more conservative or more aggressive with linking, these traits will be reflected in the learned algorithm. Though most trainers tend to link similarly, such preferences could be shaped by how the trainers are trained or instructed in the linking process and can be set differently in different projects.

and the absolute difference in year of birth. Other variables are also included: indicators for phonetic matches in first or last name, agreement of first or last letter of first or last names, the number of possible matches for the record from A in B, and agreement on middle initial. The method is flexible to the extent that this list of variables could change depending on the underlying data and historical context. The coefficients on each feature in the model represent the weight or penalty the human implicitly put on the various record features when making or not making links.

4. Applying the fitted model to the full data, we generate a predicted probability of being a match for each pair of records in A and B.
5. For each record in A, we compare the scores (fitted values) of all possible matches in B. A record is considered a match if it meets three conditions, similar in spirit to the three conditions in the EM method described previously:
    a. It is the match with highest score (probability of being linked by the human RA) out of all potential matches for the record in A.
    b. The score is sufficiently large – that is, greater than some threshold *b1*.
    c. The score is sufficiently better than the second-best match – that is, the ratio of the top score to the second-best score is larger than *b2*. When there is no second-best match because only one candidate match in dataset B was found for a given record in dataset A, this condition is trivially satisfied.
6. To define the parameters *b1* and *b2* above, we use cross-validation within the training dataset, searching over the *b1* and *b2* space to find values that minimize a weighted average of both "false positives" and "false negatives" where the truth is taken as the matches made by the human in the training data. This step finds the parameters *b1* and *b2* that tune the machine matching to be most like the human training data when deciding whether a given candidate pair is likely or not likely to be a match. The researcher may choose these weights —essentially how much to care about type I versus type II errors – based on the context and the data. In this paper, we present ML linked samples based on even weights on type I and type II errors, as well as samples weighting 3 to 1 against type I errors and 3 to 1 against type II errors.

3. Comparing automated and hand linking methods

Having described the various automated methods, we turn to a comparison of their performance in a number of settings. In theory, comparing the performance of automated and hand linking methods requires a third set of "ground truth" data in which true links are known with certainty. Unfortunately, completely certain linked historical data do not exist. In the absence of such data, we consider two genealogical samples to proxy for ground truth, following by a matching exercise across two transcriptions of the same census where we *do* know whether two records are a true match due to their placement in the census manuscript.

3.1. 1910 to 1920 Censuses using FamilySearch data

We asked the Record Linking Lab at Brigham Young University (BYU) to construct automated links of the entire 1910 and 1920 US censuses using provided code, and to compare these matches to the links from the FamilySearch's Family Tree data. These data are a large wiki-style network of individuals connected together in a "Family Tree" by users of the FamilySearch platform.[31] In order to verify that a person on the Family Tree is the correct person, users of the platform attach records (such as a US Census record or a birth or death certificate) to that person as evidence of that person's existence or relationship. For the purposes of this exercise, we take advantage of those instances in which records from two different census years were attached to the same individual in the Family Tree (hence creating a census-to-census link).

These hand links are expected to be high quality, because users have personal motivation to find their ancestors, as well as private information not typically available to researchers conducting automatic linking. For instance, these users might have information on maiden names or on other family members, and from sources that go far beyond the census. Moreover, the website is structured so as to enable different individuals to collaborate (and potentially correct each other) when they have a family member in common (Price et al. 2019).

[31] Users of FamilySearch can add records to their ancestors' pages in two main ways. The first is through automatically generated hints. To generate these hints, FamilySearch, like other genealogical platforms (Ancestry.com, Geni.com, MyHeritage.com, etc), employs teams of data scientists and machine learning record linking experts to generate automated links. In contrast to the automated linking methods we discuss in this paper, FamilySearch uses information from sources other than the census, as well as information on other household members. The second approach is by directly searching their ancestor's names (and other identifying information including place of birth and death) in the platform.

Table 1 shows the results of this exercise. The standard ABE method linked approximately 11 million people (8 million with the more conservative ABE method), out of about 47 million males in the 1910 census, a link rate of about 23% (17% in the more conservative ABE method). Among these matches, there were 3.39 million (2.78 million in the more conservative ABE method) that overlapped with people for whom we also observed a match in the Family Tree data.[32] Of these matches, 95.23% (97.13% in the more conservative ABE method) of the time the match identified by the ABE method is the same as the match identified on the Family Tree. If we treat the Family Tree data from FamilySearch as the "ground truth" (again in quotation marks because we are hesitant to call any hand linked data truth), then the rate of false positive of the ABE method is 4.77% (and only 2.87% with the more conservative ABE method). The EM method linked approximately 3.6 million people. Among these matches, 1.358 million overlapped with people for whom we also observed a match in the Family Tree data. Of these matches, 97.64% of the time the match identified by the EM method is the same as the match identified on the Family Tree, a "false positive" rate of 2.36%.

3.2 Union Army: Oldest Old data linked to the 1900 census

The Family Tree has a number of advantages: it contains millions of records, is created by linking two complete-count censuses, and is considered the gold standard of hand linking. However, because the data is proprietary, we are not able to use it for comparison with a wide set of algorithms or for inference. We turn then to the Union Army data linked to the 1900 Census (Union Army-Oldest Old data). These data include individuals known to have survived past the age of 95, and are part of a larger data collection effort that started in 1991 with the *Early Indicators* project. The benefit of this sample is that the data were carefully and expensively hand-collected by genealogists who had access to data sources that go beyond the Union Army records (and that are typically unavailable in other record linking projects). For example, the military pension system collected detailed information on each individual's exact birth and death dates and location. Given the unusual wealth of data that was used to construct the links, they are more likely to be correct.[33]

[32] The BYU researchers emphasized that it is important to note that not every individual who appears in a US Census record is attached to a person in the Family Tree data due to the vagaries of genealogical interest. As a result, many automated links do not correspond to a record attached to a person in the Family Tree data.

[33] We are grateful to Dora Costa for sharing these data with us.

The downside of this data is that it includes substantially fewer observations than a link across two census years and that the Union Army records themselves contain fewer outcomes of economic interest.

The construction of the Union Army data illustrates just how expensive and impractical hand-linking might be. In particular, the project to link these records, called the Early Indicators project, was led by Robert Fogel beginning in 1992 and was funded through several grants from the NIH spanning from 1991 to 2019, with Dora Costa leading the project starting 2013. The project of digitizing and linking the records was only complete twenty years after beginning the work. In addition, the researchers who made the links were usually trained genealogists with many years of experience. Finally, these genealogists often spent *days* considering the links of just one individual. While the linkage rates are high, and some have considered the Union Army data to be a sort of "ground truth" (though that can be debated), the project is completely infeasible for most researchers because the funds and time spent on the project were enormous (see http://uadata.org/ for a description of the project, and Wimmer 2003 for an early discussion of the data collection). Automated linking methods, in contrast, have a dramatically lower entry cost for researchers.

We have also created our own new hand-linked samples matching the Union Army records to the 1900 census. These samples were linked by two independent hand coders (an undergraduate student at Boston University and an assistant professor, James Feigenbaum) who observed only the variables that are used in standard automated algorithms: first and last names, year of birth, and state of birth.[34] We call their links "limited-information hand linking" and test them against the automated methods that use the same information (for the ML method, we use these hand links to train the algorithm). Unlike in the case of the Union Army Data, we acknowledge that in this case our hand coders are not professional genealogists. However, they did link independently,

[34] Specifically, we use a Stata dataset sent to us by Dora Costa and her team called milinfo_final.dta. We pull names from the recname1 field (corresponding to the name taken from the Regimental Books at the beginning of the data collection process) which includes comma delimited first and last names. We extract years of birth from the rb_date1 field. As indicated by the variable names, neither are the only fields with names or birthdates for the veterans in the UA data. The UA data was compiled from a variety of sources and each record potentially has multiple name and birth date variants. If soldiers were lying about age to enlist (claiming to be older to enlist earlier) some years of birth, particularly those on enlistment papers, could contain substantial noise. But to make the comparison with automated methods as straight forward as possible, we select the "first" variables (those marked with 1s) and treat them as noisy truth.

ended up finding very similar links, and the results are unchanged when we use only links that were agreed by both. Comparing these hand-linked samples to those generated by automated methods enables us to distinguish between the role of having additional information for linking and whether human linking per se is more accurate than automated algorithms.

To compare how the various automated algorithms (ABE, ABE-JW, EM, and ML) perform relative to each other and relative to the manually linked data, we use two standard measures of performance capturing the tradeoff between Type I and Type II errors. These are:

1. *Positive Prediction Value (PPV) = #correct matches/#matches (= 1 – false positive rate)*

This measure captures the degree of accuracy of a linking algorithm. That is, out of all the records that are matched, how many of these are correct?

2. *True Positive Rate (TPR) = #correct matches/#observations (= 1 – false negative rate)*

This measure captures the degree of efficiency of a linking algorithm. That is, out of all the matches that were made by the original hand-linking team, how many correct matches are made? Note that this measure is different from the usually reported "match rate", because the number in the numerator is the "correct" number of matches rather than the total number of matches.

We deem a match to be correct if it links two records that were also matched by the genealogy-quality hand linkers who constructed the Union Army-Oldest Old sample. Because the purpose of this exercise is to compare the performance of different methods relative to the Union Army-Oldest Old links, we only attempt to match those records that were initially matched in this sample (this makes the maximum potential TPR value equal 1).[35] Even in this context, genealogists were not

[35] For the matches underlying Figures 1-6, we attempted to link all observations that were successfully matched by the relevant hand linkers. In Appendix Figures A.1-A.3, we instead attempt to link all records in each dataset. In this case, we consider a link created by an automated algorithm to be false if that observation was deemed "unlinked" by the hand linkers (although we emphasize that we cannot really know if the hand linkers failure to produce a link is due to a link not existing, or to the hand linkers being reluctant to make a choice between similar potential matches). This procedure will weakly increase the false positive rate for automated methods. In the Union Army dataset, only 386 people, out of 2,033, were unable to be assigned high quality links by hand coders, and as such including these observations does not make a significant difference. However, only 62 percent of the observations in the 1915 Iowa Census could be linked to the 1940 Census by the hand linkers (= 4,276 out of 6,881 men). Attempting to find matches for all observations in the Iowa data lowers the reported "accuracy" of the automated algorithms, but the mean PPV is still 85% and the lowest PPV is 70% (Appendix Figure A.3).

equally certain about all the links they made. We thus focus on those links that the research team was most certain about (quality code equal to 1 in the Union Army-Oldest Old data). Specifically, we only attempt to match the 1,619 records (out of 2,033) in the Union Army data that were matched in the Union Army-Oldest Old data and had a quality code equal to one.[36]

We summarize the performance of the different algorithms through a series of figures. In each of these figures, the vertical axis represents the PPV, while the horizontal axis represents the TPR. These figures enable a visual interpretation of the tradeoff between accuracy (minimizing incorrect links) and efficiency (maximizing correct links).

Figure 1 summarizes the performance of different algorithms when linking Union Army records to the 1900 census. Methods are represented multiple times in the graph because we use different versions of each, changing various parameters. Black circles represent different versions of the ABE algorithm using NYSIIS standardized names. Grey circles represent different versions of the ABE algorithm using exact names. Blue diamonds represent different versions of the ABE-JW. Green squares represent links constructed using the EM method. Pink triangles represent the links done using the ML method. Finally, the purple plus signs represent the hand matches that were done by independent hand linkers.[37]

We find that automated methods roughly trace a frontier, with different automated methods on different parts of it. The EM-linked samples tend to have fewer false positives (high PPV), whereas the ML samples make a large number of true matches (high TPR). EM samples can achieve false positive rates as low as 5-10%, but at the expense of rejecting a larger number of true matches. ML samples link up to 60% of true matches, but end up with a 30-35% false positive rate. ABE methods are in between the two sets of algorithms.

Hand linking is also on the frontier, in a region close to ML samples with relatively high (true) match rates (high TPR) but also relatively high false positive rates (low PPV). This outcome

[36] Some of these records are non-unique based on place of birth, name and age. Hence, regardless of the linking method we use, it is not possible to uniquely link these observations without access to additional data. Including these observations in the denominator when computing the TPR only affects the level of the TPR, but does not affect the ranking of methods.

[37] Appendix Table A.1 explores the performance of different variants of the ABE algorithms, experimenting with including or excluding matches where there is an age difference, using middle names, etc. The main takeaway from this table is that we recommend not using observations with age difference of more than one year, as a high fraction of matches tend to be false positives.

suggests that human linkers tend to be more willing to declare matches, and are thus less conservative than automated algorithms.[38] Consider also that our measure of "ground truth" is based on hand links (albeit hand links made with additional information), and so some of the reported success of our low-information hand links might simply reflect shared biases between hand linkers in the same social context (American universities).[39]

In Figure 2, we provide more details on the performance of the different versions of each of the algorithms. Panel (a) focuses on the ABE algorithms using NYSIIS standardized names and panel (b) instead uses exact names in the matching process. The variants are determined by whether the algorithm uses middle initials or not, imposes a 5-year uniqueness band or not, and whether we allow matches to differ in reported age by up to 2 years or require an age difference of ≤ 1 year.[40] Panel (c) focuses on the ABE-JW method. We report versions imposing the age difference between the first- and the second-best match to be of zero (our least conservative choice) or two (our most conservative choice); versions using middle initials or not; and versions allowing matches to have an age difference of ≤ 2 years or ≤ 1 year. Panel (d) focuses on the EM method. We report results corresponding to a "conservative" and to a "lenient" sample. Our "conservative" sample uses p=0.9 and l=0.3, whereas our "lenient" sample uses p=0.75 and l=0.6. To illustrate the trade-off between type 1 and type 2 errors, we also show how the PPV and TPR change as we progressively decrease the value of p (from 0.9 to 0.75) and increase the value of l (from 0.3 to 0.6). Finally, Panel (e) focuses on the ML method, where the variants are determined by: (1) which person linked

[38] In a different context, Chan, Gentzkow, and Yu (2018) also find that humans faced with a classification task put more weight on avoiding false negatives than on avoiding false positives. In their case, they consider doctors diagnosing patients as having pneumonia or not.

[39] The proximity of the ML approach and the hand-links on which it was trained is not surprising but a good check that that the computer does replicate choices made by the human trainer. It may also suggest that using higher quality linked data to train with---either based on the double- and triple-entry practices at LIFE-M described in Bailey et al. (2017) or the public genealogy trees used by Price et al (2019) or other sources---could be a straightforward way to increase both PPV and TPR for machine learning approaches to census linking.

[40] As a robustness check when linking individuals across different censuses, we typically recommend keeping only matches with the same implied year of birth. However, in the UA-1900 linked data, an abnormally high number of matches are off by one year in reported year of birth. This is likely because, unlike census data where all records are collected at the same point in time, the UA data include information collected at different points in the calendar year. Hence, we instead use all matches with ≤ 1-year difference in reported age as a robustness check. In addition, this problem might be particularly acute in the 1900 census which asked year and month of birth as well as current age but seemed not to check for consistency.

the training sample, (2) whether the algorithm is allowed to use middle initials in determining matches, and (3) the function of PPV and TPR that the algorithm tries to maximize. We consider three functions: "Even" refers to matches that arise when the algorithm maximizes PPV+TPR; "PPV" refers to maximizing 3*PPV + TPR; and "TPR" refers to maximizing PPV+3*TPR.[41] Appendix Figure A.4 shows the PPV and TPR of the ABE-JW method with alternative requirements for the age difference between the first- and the second-best match.

These within-method comparisons also clearly show the tradeoff between minimizing false positives (high PPV) and maximizing the (true) match rate (high TPR). More conservative versions of each method (for instance, imposing 5 years uniqueness in the ABE, choosing a higher value of *p* in the EM or giving higher weight to PPV in the ML) yield fewer false positives, but at the expense of overlooking true matches. Indeed, the most conservative version of each of the methods achieves a positive prediction value (PPV) of at least 0.8 (with some of the methods achieving PPVs close to 1), suggesting a low false positive rate, but at the expense of matching fewer true observations.[42]

One intuitive approach to further reduce false positives is to construct a sample using the *intersection* of the methods.[43] In Figure 3 we show the false positive and truth match rates (PPV and TPR) of all possible 2-way, 3-way, 4-way, and 5-way intersections of five automated methods. Taking the intersection of two or more methods indeed reduces false positives; the most accurate single method has a PPV of slightly above 90% but taking the intersection of all five of these methods increases the PPV to approximately 96%. However, this increase in PPV comes at the expense of a lower TPR, implying that the "intersection" method also lies along the type 1/type 2

[41] We note that PPV and TPR in this maximization correspond to the training sample of hand links we generate. The Union Army-Oldest Old sample links are *not* used in the training at any stage.

[42] Bailey et al. (2017) also evaluates the performance of different automated algorithms using the Oldest Old data (in addition to the Life-M data and a "synthetic ground truth" data). Our results using the Oldest Old data are similar to theirs when using the same linking algorithms. However, their analysis includes some methods with high rates of false positives that are rarely or never used in economic history (such as ABE with Soundex). In addition, we show that there a number of automated methods that generate low false positive rates. As a result, our takeaway on automated methods is more positive than theirs.

[43] Intuitively, if different methods were wrong for different reasons, then looking at their intersection will reduce false positives: it would be unlikely for several methods to make the same mistakes. Bailey et al (2019) also uses this approach.

frontier (in a region of high PPV but relatively low TPR). The PPV and TPR values, along with sample size, of each intersection are listed in Appendix Table A.2. Appendix Figure A.5 shows a Venn diagram of the overlap in matches found by each of these methods. One notable feature of this diagram is that the ML method yields the greatest number of matches that do not overlap with any other single method.

Another important aspect of the quality of a matching algorithm is its ability to generate a representative sample of the population. In particular, individuals who report their identifying information with high accuracy will be disproportionately represented under a more stringent matching rule, and these individuals could be more literate and numerate than the average population. To evaluate the performance of the different methods with respect to representativeness, in panel (a) of Table 2 we compare the observable characteristics of matched and unmatched individuals. Overall, as scholars using linked historical data often acknowledge, none of the methods generates fully representative samples. Specifically, in our application, all the methods are more likely to link literate and taller individuals. Note, however, that the differences are in all cases quite small. For instance, the largest coefficient on height is 0.373 inches (relative to a mean height of 67.5 inches) and the largest coefficient on literacy is 0.017 (relative to an average literacy of 0.938). In contrast, differences in occupational scores between matched and unmatched individuals are not statistically significant (and are sometimes positive and sometimes negative).

Given that differences in observable characteristics between matched and unmatched individuals are typical in studies using linked data, we suggest that papers that rely on linking should carefully discuss (lack of) representativeness and run versions of the analysis that reweight the sample to match the population on observable characteristics.[44] It is also important to acknowledge that external validity is limited to people with characteristics similar to the ones in the matched sample.

Automated methods versus limited-information hand links

Thus far, we have compared automated methods to a hand linked sample created from a wealth of personal information. Overall, the false positive rates implied by this comparison are relatively

[44] For reweighting using inverse propensity score, see also Pérez (2017), Zimran (2019), and Bailey et al (2019).

low, but they are not zero. Is this because hand linking is per se more accurate than automated methods or because the hand-linkers in question had access to additional information? In Figure 4, we directly compare our automatic links based on four linking variables (first name, last name, birth place and implied year of birth) to those made by hand linkers who had access to precisely the same set of limited variables.[45] To simplify presentation, we treat these limited-information hand links as the "truth" in Figure 4, and compute "PPV" and "TPR" for each automated method relative to these hand links (although, from Figure 1, we know that limited-information hand links do not always coincide with genealogy-quality matches). All of the automated algorithms much more closely replicate the limited-information hand links in Figure 4 compared to the full-information hand links in Figure 1. Indeed, the lowest PPV across all methods is quite high in this context (around 0.77), most of the PPVs are above 0.9 and many of them are above 0.95, implying a discrepancy or "false positive" rate between 5 and 15%.

Computational time. Comparing the effectiveness ("the benefits") of various methods does not take into account the fact that the cost and computational time requirements to implement each of them are rather different. For the UA-1900 exercise we measured the amount of time it took to match when using each of the fully automated methods (ABE, ABE-JW, and EM). The ABE is the least demanding method; it completed a full match using NYSIIS names in 10 minutes. The ABE-JW and EM methods take more time than the ABE, mostly due to the computation of Jaro-Winkler distances: Both the ABE-JW and EM methods took approximately 30 minutes to link the UA-1900 census data. All algorithms will be more time intensive when using larger data sets.[46] Finally, note that hand linking is similarly computationally intensive, since it requires computer time to filter which records to show to the human linker. For instance, manually linking a full

[45] Automated methods need not use only these four linking variables. Other approaches have included a wider set of fixed characteristics, such as race or parents' birthplace. Still other approaches, including those underway by the Minnesota Population Center, include many linking variables, even if they can change within person (e.g., name of spouse, occupation, location of residence). One advantage of linking based only on characteristics that should not change within person over time is to allow all records to have an equal chance of linking (for example, linking based on spouses' name will increase the match rates of individuals who are married, or who remain married to the same person over a decade).

[46] We also measured the time it took to link the 1915 Iowa data to the 1940 census. Linking in this case took approximately 30 minutes when using the ABE method with NYSIIS or exact names, and about two hours with the ABE-JW and EM methods. These are all rough estimates of the differences in computing time, given that the work was done on the NBER server whose speed varies dramatically day to day depending on other users' load.

census to a full census is practically impossible without first using a computer to filter which records to consider as potential candidates. Of course, in addition to the computational time, manual linking also requires the time of human clerks. In our case, the research assistants that we hired for performing the manual links made an average of four or five links per minute. Hence, creating a double-entry, hand-linked sample of approximately 2000 observations would take 14 to 16 hours of research assistant time. But while all automated methods scale reasonably, hand-linking does not: double-entry hand-linking all 40 million men in 1900 ahead to another census would take more than 30 years.

It is worth keeping in mind that most of the extra computational-time costs attached to the EM or ABE-JW methods are due to the calculation of JW distances across potential matches. When linking observations across censuses or other commonly used sources, this additional extra time is a kind of "fixed cost." Once we compute these distances, they can be stored and used for different research projects using the same datasets, or used within the same research project for different linking methods (for the EM and the ABE-JW). One natural option is to use the standard ABE match first before assessing robustness of the results to the methods that require additional computational time.

3.3. Matching two independent transcriptions of the 1940 census to each other

Thus far, we have conducted linking exercises without a "ground truth" because, even in the case of hand links made by highly trained genealogists, it is not possible to know with certainty whether selected matches are correct. In this section, we perform a matching exercise where we *do* know whether two records are a true match. Specifically, we link two different transcriptions of the 1940 Census: one version was transcribed by Ancestry.com and one version was transcribed by Family Search (Ancestry-FS). Both transcriptions were performed using the same underlying data — the 1940 Federal Census – but were transcribed by two different genealogy companies with their own methods. In each dataset, we observe the census page and line of each record. This information provides the "true nexus" between the two different transcriptions. In addition to comparing the performance of different automatic algorithms relative to actual ground truth, this exercise allows us to quantify the degree of human error in transcribing records. As there is no mortality or migration or name-changing or enumeration error between these two versions – they are based on the same exact underlying enumeration – only transcription will vary.

Table 3 shows that there are substantial differences between the names in the two transcriptions. In this table, we show the fraction of observations that differ in the first name (after different string cleaning), last name or middle name and initial. We also perform this exercise separately for persons born in Iowa, North Carolina, Ohio, England, and Italy to assess variation in transcription across population subgroups.[47] The table shows that transcription error is present and meaningful (e.g. between 7% and 14% of records disagree in the first name, and between 17% and 32% disagree in the last name). In addition, those who are born in Italy have higher rates of disagreement (likely reflecting that many transcribers are English-speaking). Similarly, in Appendix Figure A.8, we compute the Jaro-Winkler distance in first (x-axis) and last (y-axis) names between the two independent transcriptions. If both transcriptions were identical, all the observations would be on the origin, but this is far from being the case. In contrast to the differences in transcribed names, we find small (although non-zero) differences in transcribed ages. In this case, the two transcriptions disagree in at most 2% of the cases.[48]

In Table 4, we display the results from linking the two transcriptions of the 1940 Census. For simplicity, we just focus on the different versions of the ABE algorithm. Depending on whether we use a more or less conservative version of the ABE, we match between 43%-67% of all observations. However, among the ones that the ABE algorithm matches, practically all of them are correct (PPVs are between 0.98 and 1). This implies that, despite substantial transcription errors, the automated methods perform well in finding correct matches, but that they are conservative in the sense that they avoid matching observations that are not "unique" enough or where differences in transcription are very severe. Moreover, this exercise implies that it is likely not possible to obtain matching rates that are higher than the ones here, unless further information is used for disambiguation.

[47] Ohio and North Carolina are two populous states, one in the North and on in the South. We also included Iowa as it has similar characteristics to the average US states in this time period and because our previous analysis also uses data from Iowa. Finally, we chose two immigrant countries (one English speaking, one non-English speaking) to assess transcription discrepancies for foreign names.

[48] Note that, if the two transcriptions were identical to each other, all our algorithms would by construction have zero false positives: All the methods privilege "exact links" over links in which there is some disagreement in the identifying formation.

3.4. 1915 Iowa records linked to the 1940 census

We next turn to a sample that will allow us to conduct a set of typical intergenerational mobility regressions using samples based on hand and automated links. Specifically, we use a sample constructed in Feigenbaum (2018) that links the 1915 Iowa Census to the 1940 Federal Census. Unlike in the previous exercises, we do not have a genealogically-based sample to constitute "ground truth," and so we will focus on comparing automated links to hand links based on standard matching variables (names, place of birth, and year of birth). Feigenbaum began with the 6,071 boys in the Iowa State Census sample digitized in Goldin and Katz (2000, 2007). To locate these sons in 1940, he utilized the 100% 1940 census sample deposited by Ancestry.com with the NBER. In this context, it is not possible to establish whether the algorithm or the human is correct due to the absence of ground truth data.

As in the UA-1900 case, we find that there is little disagreement between automated and hand links when humans and algorithms use the same information for linking. Here, for presentational purposes, we treat limited-information hand links as the "truth" in our figures, and compute "PPV" and "TPR" assuming that the hand links are correct. Figure 5 shows the results of this exercise. The EM and ML methods consistently produce PPV values which are close to 1, and the ABE method produced PPV values that range from 0.88-0.99, implying a "false positive" rate between 2 and 12%. The intersection of the automated methods, shown in Figure 6, further increases PPV (which becomes nearly equal to one), but at the expense of substantially reducing TPR. In Appendix Table A.3, we list the PPV and TPR values, along with sample size, of each intersection. Appendix Figure A.6 shows a Venn diagram of the overlap in matches found by each of these methods.

In addition, as in the case of the Union Army linked data, the matched sample is quite similar to the population, but not fully representative (panel (b) of Table 1). Specifically, most methods are more likely to find younger individuals, with the largest age difference between matched and unmatched individuals being of -0.312 years (relative to an average age in 1915 of 9.636 years old). There is also a correlation between having a foreign-born parent and the likelihood of matching, although this correlation is sometimes positive and sometimes negative depending on the linking method. Individuals who were literate or had more years of schooling in 1915 are also

less likely to be linked in this case. However, given that the sample includes many children in 1915, these measures do not capture final educational attainment and this pattern instead likely reflects the fact that younger individuals are more likely to be linked.

3.5. Assessing discrepancies between hand links and automated methods

One of the general takeaways from this and the previous exercises is that, whenever humans and algorithms have the same amount of information, the level of disagreement in matches is actually quite low. However, there are still cases when humans and algorithms disagree. Once we look at disagreements, is it obvious that humans are "right"? To better understand this issue, we took the complete list of disagreements from a case where one of the algorithms had a very low number of disagreements with hand linkers. Specifically, we focus on the EM with hyperparameter choices 0.75-0.6 (see section 2). The EM 0.75-0.6 had TPR=0.25 and PPV=0.98, implying that out of all the links made by the algorithm, 98% were also made by the hand linkers. In Appendix Figure A.7, we provide the full list of the 16 instances in which the algorithm and the human chose a different link. The left column shows the records that we want to match, the middle column shows the records that human linkers chose, and the right column shows the choices that the automated method made.

In most cases, it is not straightforward to establish without further information whether the hand-link is "correct" or whether the algorithmic link is "correct." Going back to our example from the introduction, in line 14 (Paul Coulter, born in Kansas, predicted year of birth 1912), the algorithm chooses Paul Courter, born one year apart, whereas the hand linker chooses Paul Coater, born in the same year. In line 7 (John Obman), the algorithm chooses John M Orman, born one year apart, whereas the hand linker chooses John Ohmann, born in the same year. In line 16 (William Noel born in 1910), the algorithm chooses William F Noxsel born in 1911, but the hand linker chooses William G Noll born in 1909.

Overall, this exercise suggests that when humans and algorithms have the same amount of information, the extent of disagreement is usually low, and when they do disagree, it is unclear whether the human or the algorithm is more likely to be correct. Beyond this lack of a clear advantage in performance, there are some practical and conceptual difficulties in implementing hand linking. First, as discussed above, when matching by hand, even with unlimited budget it is infeasible to compare each record to each potential match (each potential match being for example

every one of 100s of millions of Census observations). Hence, it becomes necessary to choose who to show to the human linker as a potential match for each record. By contrast, a computer in principle *can* make these very large comparisons.

The fact that human linkers cannot compare all potential matches simultaneously creates additional conceptual difficulties. If the potential matches are shown "without replacement" (that is, each observation is shown as a potential match for just one observation), the order in which the human linker sees the potential matches might influence the final set of links. If, on the other hand, the potential matches are shown with replacement, the researcher might end up with duplicate links, which cannot obviously all be "true."

Finally, an additional important disadvantage of hand linking is the issue of replicability. While automated methods are transparent and easy replicate, hand links are by definition impossible to replicate without the intervention of the same hand linker (and even the same hand linker can make different choices at different points in time). While we do find that two independent hand linkers trained in similar ways made similar choices that need not always be the case.

4. Results: Automated methods and inference

Ultimately, the goal of producing linked samples is to conduct economic analyses. Thus, the key question for any automated linking method is to what extent it affects inference downstream. There are two main channels through which errors in linking might affect inference. First, incorrectly linked individuals could introduce measurement error and result in attenuation bias. Second, the inability to link every observation could affect inference if the people who can be linked are not representative of the population.

Before considering the extent to which automated linking affects regression coefficients, it is important to note that, even if automated linked samples produced different point estimates, doing so might not substantially affect a paper's conclusions. In particular, researchers must consider whether false positives will understate or overstate their findings. For instance, when the goal of a study is to test whether a certain variable has a positive or a negative effect on another, attenuation bias will bias the estimated effect toward zero but will not change the sign of the relationship. In contrast, studies of intergenerational mobility rely on estimating a persistence parameter (the

degree to which father outcomes predict son outcomes), which is inversely related to mobility. Thus, attenuation of the persistence parameter would overstate the degree of mobility. More generally, we encourage researchers to think carefully about the bias caused by false positives and unrepresentative sample. For example, in many regression settings, false links will create measurement error that will bias OLS estimates toward zero. However, family fixed effects estimates (FE) are actually biased towards OLS because picking two random people and calling them brothers (due to incorrect matches) will tend to recover the population-wide estimates.

We perform two exercises to assess how choice of linking method will affect substantive research conclusions. In the first exercise, we use the Iowa-1940 census to compute intergenerational correlations in earnings and education. In the second exercise, we use the US 1850-1880 and the Norway 1865-1900 linked samples to compute measures of intergenerational occupational mobility.

4.1. Inference using the Iowa 1915 Data linked to the 1940 census

One common application of linked data in economic history is to quantify the extent to which a son's economic outcomes can be predicted by his father's characteristics. We compare intergenerational correlations in earnings and education in the Iowa-1940 data using various automated and hand linking methods. First, we regress the log of son's earnings in the 1940 census on the log of his father's earnings in 1915 Iowa.[49] Second, we regress a son's completed years of schooling in 1940 on his father's completed years of schooling as measured in the 1915 data.

Panels (a) and (b) of Figure 7 present the results of these regressions. We find that hand and automated linked samples recover similar estimates of the inter-generational elasticity of income and education. In all cases, the coefficients are precisely estimated and the 95% confidence interval using hand links methods includes the point estimate using any of the automated methods. Estimates of the inter-generational elasticity of income range from 0.15 to 0.20, with the smallest coefficient 25% lower than the largest, producing a set of results that would not change the qualitative conclusion of any study in question.

[49] Bailey et al (2019) computes intergenerational elasticities between fathers' and sons' earnings using the LIFE-M data.

Figure 7 also reports the coefficients from these regressions after reweighting the data by the following observable characteristics in the Iowa census records: year of birth, literacy, years of schooling and foreign-born status of parents. By reweighting, we ensure that each linked sample matches the proportions in the 1915 Iowa census. The balance table in Appendix Table A.4. shows that the sample rebalances on observables after weighting. Weighted results look similar to their unweighted counterparts, with automated methods producing similar coefficients to the hand-linked data.

4.2. Inference using linked fathers and sons across US and Norwegian censuses of population

Abramitzky, Mill, and Perez (2019) use the EM method to create two sets of linked samples: one linking the 1850 and 1880 US censuses of population, and one linking the 1865 and 1900 Norwegian censuses of population. Here, we recreate these linked samples using the full set of other automated methods. We chose to create these linked samples for three primary reasons. First, economic history papers often attempt to link historical censuses of population (rather than some outside source like Union Army data to a census), making linked census samples 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). Third, testing the method in two different countries enables us to assess how well automated methods perform in two settings with different naming conventions, enumeration quality, outmigration rates, etc.

We compute transition matrices documenting rates of intergenerational occupational mobility using our various linked datasets as well as the published files from IPUMS. We show that researchers using our various linked samples would have arrived at substantively similar conclusions about the patterns of intergenerational mobility, as compared to the widely used IPUMS samples.

Appendix Tables A.5 and A.6 show the father-son occupational transition matrices constructed using our linked samples and the IPUMS linked samples. Appendix Table A.5 shows the data for the US 1850-1880 links, whereas Appendix Table A.6 shows the corresponding Norway 1865-1900 links. As can be seen from the tables, the automated 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, the 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).

Table 5 reports 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 in a different occupational category than their fathers. We also report a test of whether this fraction is the same as the one that would arise if son's occupations were independent of father's occupations. 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 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 the automatically linked samples. In the US, we estimate that about 45% of sons worked in a differential occupational category when using the IPUMS sample, and between 44 and 50 when using our own linked samples. In Norway, we estimate that 47% of sons worked on a different occupational category than their father when using the IPUMS sample, and between 44 and 48% when using our data.

The distance with respect to a matrix representing full independence is relatively similar regardless of the linked samples that we use, both for the US and Norway. For the US, the departure from independence displayed in the Altham statistic is 17.37 when using the IPUMS sample and ranges between 12.17 and 15.18 when using our own samples. For Norway, the departure from independence is 25.01 when using the IPUMS sample and between 24.19 and 26.09 when using our linked samples. In all cases, no matter which linking method we use, we reach the same conclusion: mobility was higher in the US than in Norway in the 19th century.

5. Conclusions

We evaluate the performance of widely used automated algorithms for historical record linkage, as well as hand linked samples that were created using the same set of standard linkage variables (that is, names, year of birth and state or country of birth). One benchmark for accuracy is the links made by genealogists and users of the website FamilySearch of the 1910 and 1920 US censuses (population to population). Another is the links of the Union Army Records to the 1900 US Census done by trained research assistants who had access to extra information not typically available for linking (sample to population).

In the population-to-population case, we find that automated methods agree with the links made by genealogists and users of FamilySearch.org over 95% of the time. In the sample-to-population case, we find that many automated methods lie along a frontier that traces out the tradeoff between type I and type II errors. Researchers can choose to use algorithms that generate very low discrepancy rates from these benchmark links (as low as 5-10% in the context of the Union Army records) at a cost of relatively low (true) match rates (10-30%). Alternatively, they can choose algorithms with higher (true) match rates (50-60%) at a cost of higher discrepancy rates (15-30%). Hand linking is not a perfect solution either: hand links that rely on the standard matching variables are also on this frontier, producing relatively high false positive rates (around 25%) and relatively high (true) match rates (around 63%). Compared to more conservative automated methods, humans tend to link more observations but at a cost of higher rates of false positives. The frontier provided here can help guide researchers in creating alternative samples using the various automated methods and testing the robustness of their results across samples.

Next, we compare automated methods to another hand linked sample, namely the 1915 Iowa Census linked to the 1940 Census. When humans and machine use the *same* set of linking variables, human linkers and automated algorithms agree in the vast majority of cases. When hand and automated links do not agree, it is not clear from inspection which links are correct.

We then use data from two different transcriptions of the 1940 Federal Census, one transcribed by FamilySearch and one by Ancestry.com. In this case, we can establish a real "ground truth" because records listed on the same census manuscript page and line number are known to refer to the same individual. We find that differences in transcription are generally high, especially for the foreign born from non-English speaking countries. Even in this case of linking a census to itself, we can only link 43-67% of the observations, suggesting an upper bound for match rates of any

method due to transcription quality and common names for whom we cannot find a unique match in the two datasets. Nevertheless, when linking these two versions of the 1940 Census, we find that automated methods produce links that are almost 100% correct.

Finally, we study how automated linking methods affect inference. Across a number of regression analyses, we find that coefficient estimates and parameters of interest are very similar when using linked samples based on each of the different automated methods. Point estimates generated from alternative linking algorithms are well within the confidence intervals of the benchmark estimate and exhibit a degree of measurement error no larger than other commonly-used variables (see, for example, the "years of schooling" variable studied in Ashenfelter and Krueger 1994).

We intentionally do not pick a favorite method – the choice of the method ultimately depends on the research question and historical setting. Instead, we have provided implementation codes to the methods discussed in this paper, so researchers are encouraged to test the robustness of their results to the linking method. It is natural to start with the more and less conservative versions of the ABE method, given they are the easiest and fastest to implement and given they span much of the tradeoff between type I and type II errors. When the potential sample size is very large, we encourage researchers to also use the EM method and pick conservative parameters that ensure very low rates of false positives. When sample size is low, we encourage researchers to also use the ML method that tends to generate larger samples.

Overall, we conclude that automated methods perform well; it is possible to use automated methods to create linked samples with few erroneous links. The ability to create large historical panel datasets by linking between census waves or other sources of data (such as the Union Army records or state censuses or other novel sources) has the potential to represent a major advance in our understanding of economic history, and of important questions of economic, social, and geographic mobility more broadly. Although linking across historical sources is challenging and inevitably imperfect, the rates of false links are low and can be minimized further by judicious choice of algorithm. Ultimately, we find that the effect of algorithm choice on conclusions from inference is small. Finally, we want to highlight our Census Linking Project (https://censuslinkingproject.org/), an ongoing effort to link and make publicly available every pair of historical complete-count Censuses from 1850 to 1940 (36 pairs in all) using the different

linking methods suggested in our paper.[50] Our goal is to reduce barriers and open up new research possibilities by providing customizable linked historical datasets to the broad research community. We will continue to post links created using a variety of matching algorithms, along with linking code and technical documentation for public dissemination.

[50] Beyond federal Census to federal Census linking, we want to emphasize the power and potential of using automated linking methods to connect other sources, either to one another or to the census. There are always new (old) sources of microdata for clever researchers to find and questions that can be answered once the data are linked. Automated methods for record linking makes that possible for scholars at all levels.

References

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, 102(5), 1832-56.

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2014). A nation of immigrants: Assimilation and economic outcomes in the age of mass migration. *Journal of Political Economy*, 122(3), 467-506.

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2019). To the new world and back again: Return migrants in the age of mass migration. *Industrial and Labor Relations Review,* 72(2), 300-322.

Abramitzky, R., Boustan, L. P., Jácome, E., & Pérez, S. (2019). *Intergenerational Mobility of Immigrants in the US over Two Centuries* (No. w26408). National Bureau of Economic Research.

Abramitzky, R., Mill, R., & Pérez, S. (2019). Linking Individuals Across Historical Sources: a Fully Automated Approach, *Historical Methods*.

Aizer, A., Eli, S., Ferrie, J., & Lleras-Muney, A. (2016). The long-run impact of cash transfers to poor families. *American Economic Review*, 106(4), 935-71.

Altham, P. M. (1970). The Measurement of Association of Rows and Columns for an r×s Contingency Table. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(1), 63-73.

Ashenfelter, O., & Krueger, A.B. (1994). Estimates of the Economic Return to Schooling from a New Sample of Twins. *American Economic Review,* 84(5), 1157-1173.

Bailey, M., Cole, C., Henderson, M., & Massey, C. (2017). How Well Do Automated Methods Perform in Historical Samples? Evidence from New Ground Truth. National Bureau of Economic Research Working Paper No. 24019.

Bailey, M., Cole, C., Henderson, M., & Massey, C. (2019). How Well Do Automated Methods Perform? Lessons from US Historical Data. *Journal of Economic Literature* (forthcoming)

Bandiera O., Rasul, I., & Viarengo, M. (2013). The Making of Modern America: Migratory Flows in the Age of Mass Migration. *Journal of Development Economics*, 102, 23-47.

Bleakley, H., & Ferrie, J. (2016). Shocking behavior: Random wealth in antebellum Georgia and human capital across generations. *The Quarterly Journal of Economics*, *131*(3), 1455-1495.

Chan, D.C., Gentzkow, M., & Yu, C. (2018). Selection with Skills: Evidence from Radiologists. Working Paper.

Collins, W. J., & Wanamaker, M. H. (2014). Selection and economic gains in the great migration of African Americans: new evidence from linked census data. *American Economic Journal: Applied Economics*, 6(1), 220-52.

Collins, W. J., & Wanamaker, M. H. (2015). The great migration in black and white: New evidence on the selection and sorting of southern migrants. *Journal of Economic History*, 75(4), 947-992.

Collins, W. J., & Wanamaker, M. H. (2017). Up from Slavery? African American Intergenerational Economic Mobility Since 1880. National Bureau of Economic Research Working Paper No. 23395.

Costa, D. L., DeSomer, H., Hanss, E., Roudiez, C., Wilson, S. E., & Yetter, N. (2017). Union Army veterans, all grown up. *Historical Methods*, 50(2), 79-95.

Craig, J., Eriksson, K., & Niemesh, G. T. *Marriage and the Intergenerational Mobility of Women: Evidence from Marriage Certificates 1850-1910*. Mimeo

Dahis, R., Nix, E., & Qian, N. (2019). *Choosing Racial Identity in the United States, 1880-1940* (No. w26465). National Bureau of Economic Research.

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

Eli, S., & Salisbury, L. (2016). Patronage Politics and the Development of the Welfare State: Confederate Pensions in the American South. *Journal of Economic History*, 76(4), 1078-1112.

Eriksson, K. (2019). Moving North and into jail? The great migration and black incarceration. *Journal of Economic Behavior & Organization*, 159, 526-538.

Eriksson, K. (Forthcoming) Education and Incarceration in the Jim Crow South: Evidence from Rosenwald Schools. *Journal of Human Resources.*

Feigenbaum, J. J. (2015). Intergenerational Mobility during the Great Depression. Working Paper. Available at https://open.bu.edu/handle/2144/27525

Feigenbaum, J. J. (2016). Automated census record linking: A machine learning approach. Working Paper. Available at https://open.bu.edu/handle/2144/27526.

Feigenbaum, J. J. (2018), Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940. *The Economic Journal*, 128(612),446-481

Feigenbaum, J.J. & Tan, H. (2019). The Return to Education in the Mid-20th Century: Evidence from Twins. National Bureau of Economic Research Working Paper No. 26407.

Ferrie, J. P. (1996). A new sample of males linked from the public use microdata sample of the 1850 US federal census of population to the 1860 US federal census manuscript schedules. *Historical Methods*, 29(4), 141-156.

Ferrie, J. P. (1997). The entry into the US labor market of antebellum European immigrants, 1840–1860. *Explorations in Economic History*, 34(3), 295-330.

Fouka, V. (2019). Backlash: The unintended effects of language prohibition in US schools after World War I. *The Review of Economic Studies*, *87*(1), 204-239.

Goeken, R., Huynh, L., Lynch, T. A., & Vick, R. (2011). New methods of census record linking. *Historical methods*, 44(1), 7-14.

Goldin, C., & Katz, L. F. (2000). Education and income in the early twentieth century: Evidence from the prairies. *Journal of Economic History*, 60(3), 782-818.

Goldin, C., & Katz, L. F. (2007). Long-Run Changes in the Wage Structure: Narrowing, Widening, Polarizing. *Brookings Papers on Economic Activity*, (2), 135.

Gould, J.D. (1980). European Inter-Continental Emigration: The Role of "Diffusion" and "Feedback." *Journal of European Economic History,* 9(2), 189-194.

Hornbeck, R., & Naidu, S. (2014). When the levee breaks: black migration and economic development in the American South. *American Economic Review*, *104*(3), 963-90.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kahneman, D. and A. Tversky. (1973) "On the psychology of prediction." *Psychological Review*, 80(4), 237-251.

Koneru, K., Pulla, V. S. V., & Varol, C. (2016, July). Performance Evaluation of Phonetic Matching Algorithms on English Words and Street Names. In *Proceedings of the 5th International Conference on Data Management Technologies and Applications* (pp. 57-64). SCITEPRESS-Science and Technology Publications, Lda.

King, M.L., & Magnuson, D.L. (1995) Perspectives on Historical U.S. Census Undercounts. *Social Science History*, 19(4), 455-466.

Kosack, E., & Ward, Z. (2014). Who crossed the border? Self-selection of Mexican migrants in the early twentieth century. *Journal of Economic History*, 74(4), 1015-1044.

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American statistical association*, *100*(469), 222-230.

Long, J., & Ferrie, J. (2013). Intergenerational occupational mobility in Great Britain and the United States since 1850. *American Economic Review*, 103(4), 1109-37.

Mason, K.O., & Cope, L.G. (1987). Sources of Age and Date-of-Birth Misreporting in the 1900 U.S. Census. *Demography*, 24(4), 563-573.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

Mill, R., & Stein, L. (2016). Race, skin color, and economic outcomes in early twentieth-century America. Working Paper.

Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2741797

Modalsli, J. (2017). Intergenerational mobility in Norway, 1865–2011. *The Scandinavian Journal of Economics*, 119(1), 34-71.

Nix, E., & Qian, N. (2015). The fluidity of race: "Passing" in the United States, 1880-1940. National Bureau of Economic Research Working Paper No. 20828.

Olivetti, C., & Paserman, M. D. (2015). In the name of the son (and the daughter): Intergenerational mobility in the united states, 1850-1940. *American Economic Review*, *105*(8), 2695-2724.

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

Pérez, S. (2017). The (South) American Dream: Mobility and Economic Outcomes of First-and Second-Generation Immigrants in Nineteenth-Century Argentina. *Journal of Economic History*, 77(4), 971-1006.

Pérez, S. (2019). Intergenerational Occupational Mobility across Three Continents. *The Journal of Economic History*, 1-34.

Poirier, A., & Ziebarth, N. L. (2019). Estimation of models with multiple-valued explanatory variables. *Journal of Business & Economic Statistics*, *37*(4), 586-597.

Price, J., Buckles, K., Riley, I., & Van Leeuwen, J. (2019). Combining Family History and Machine Learning to Link Historical Records. Working Paper.

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

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

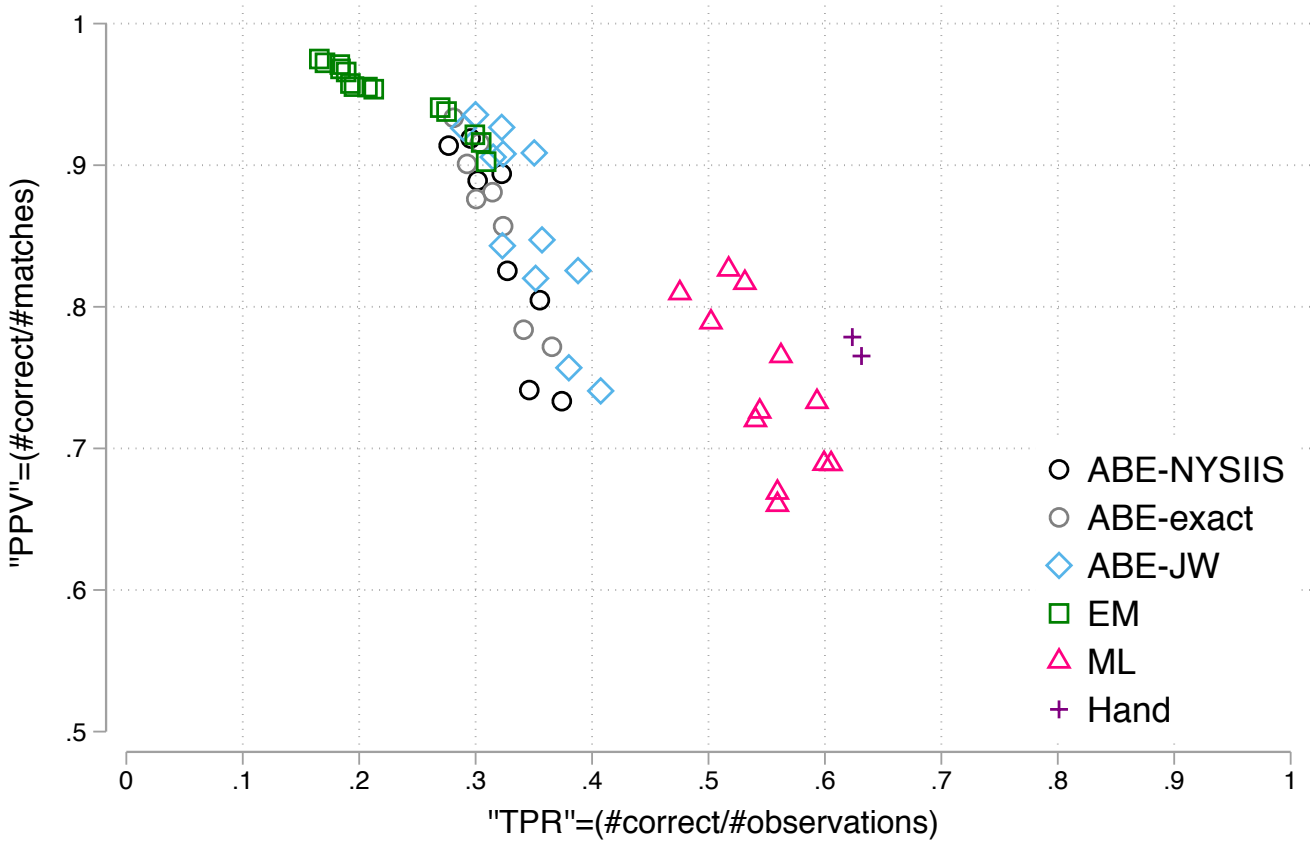
Wimmer, L. T. (2003). Reflections on the Early Indicators Project. A Partial History. In *Health and labor force participation over the life cycle: Evidence from the past* (pp. 1-10). University of Chicago Press.

Winkler, W.E. (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

Zimran, A. (2019). Sample-Selection Bias and Height Trends in the Nineteenth-Century United States. *The Journal of Economic History*, *79*(1), 99-138.

Figures and Tables

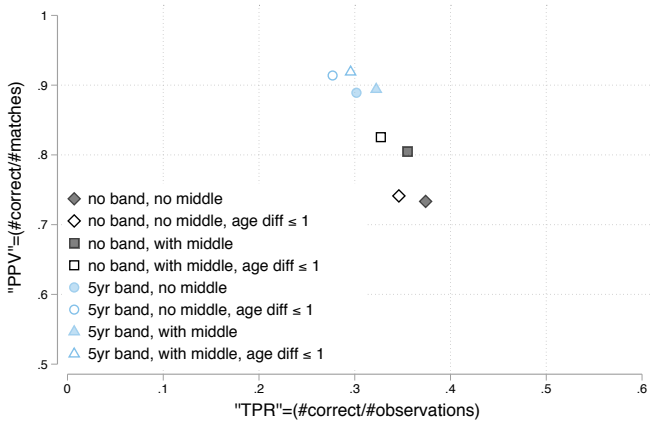
Figure 1: Accuracy vs. Efficiency - Comparing Linking Algorithms (UA-1900 census)



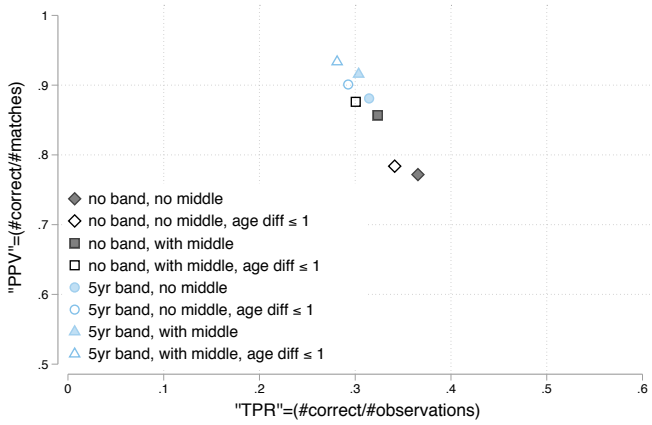
Notes: PPV ($\frac{\#truelinks}{\#matched}$) and TPR ($\frac{\#truelinks}{\#ofobservations}$) for the exercise that links the Union Army records to the 1900 census using different variations of each linking algorithms (ABE-NYSIIS, ABE-exact, ABE-JW, EM and ML). "Hand" compares the hand-linking carried out by two different people that only used the same information that the automated algorithms use (name, place of birth, and year of birth). A match is defined as "true" if it coincides with the links made in the Union Army-Oldest Old sample.

Figure 2: Accuracy vs. Efficiency-Comparing Versions of each Linking Algorithm (UA-1900 census)

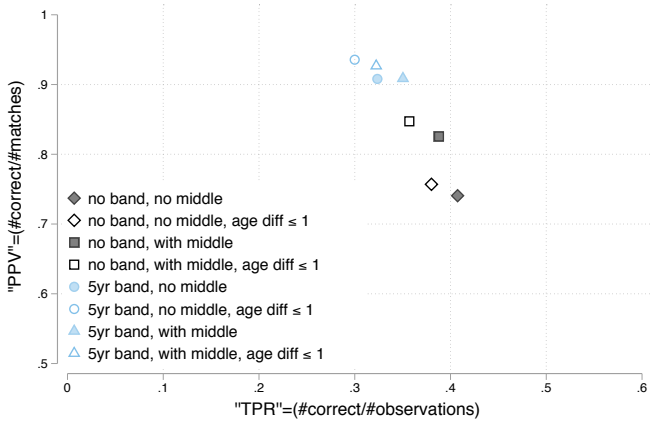
(a) ABE-NYSIIS



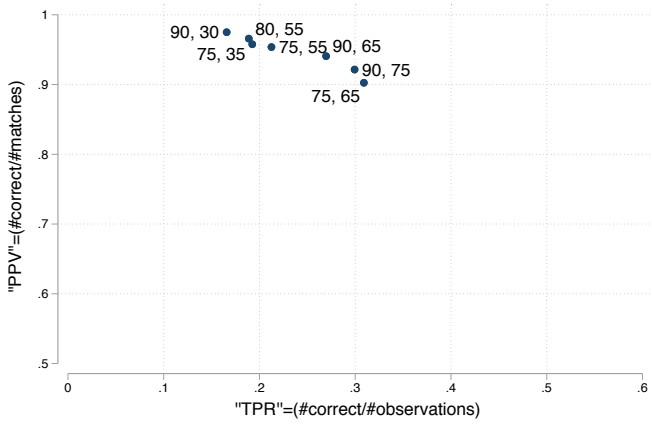
(b) ABE-Exact



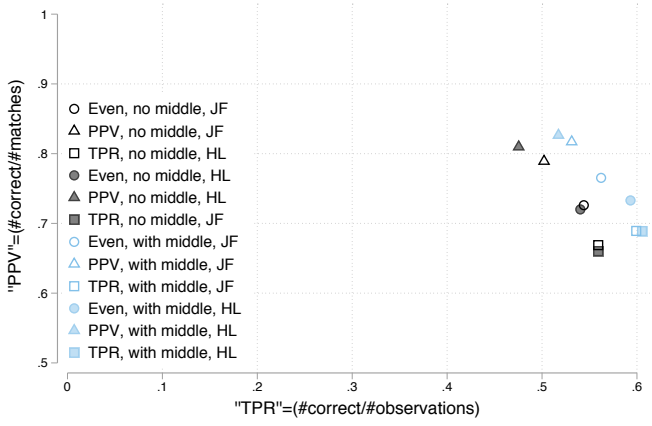
(c) ABE-JW



(d) EM

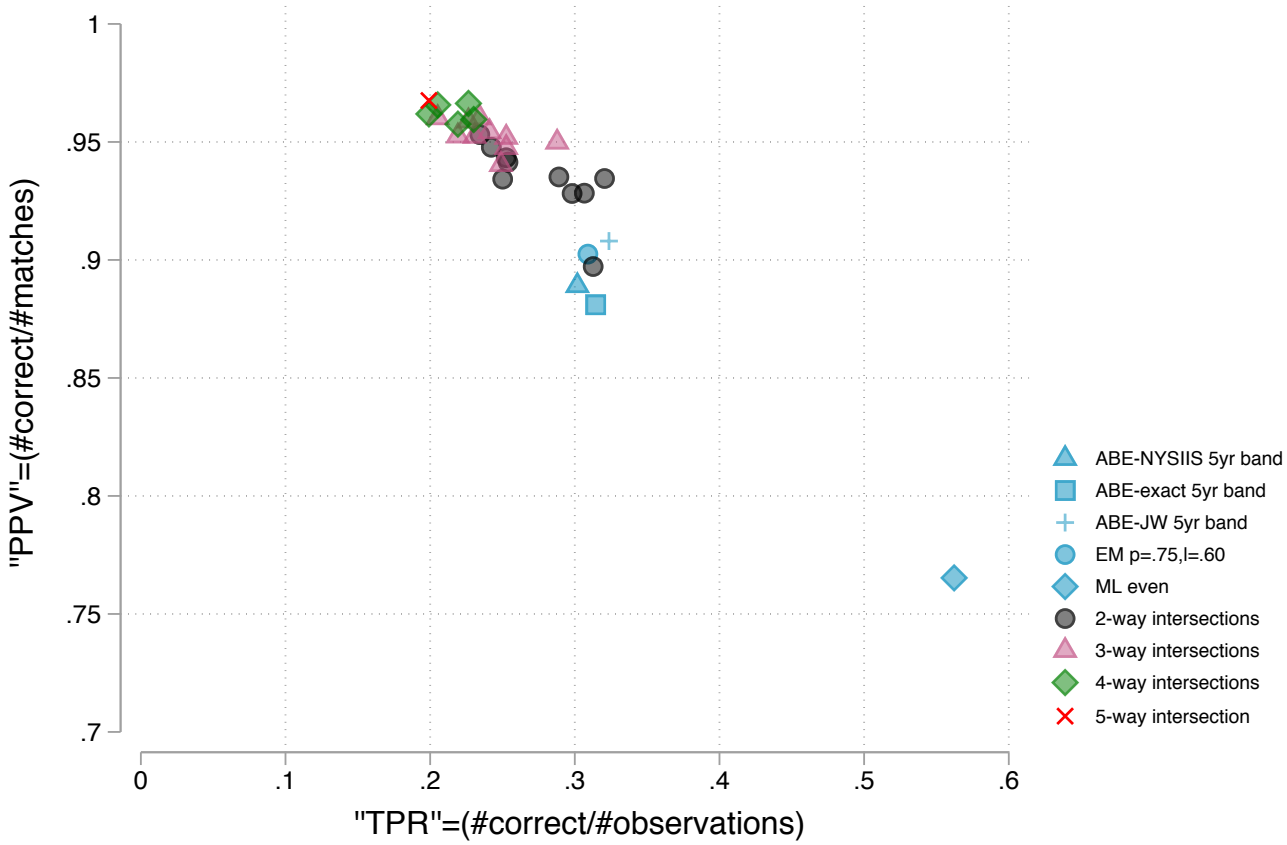


(e) ML



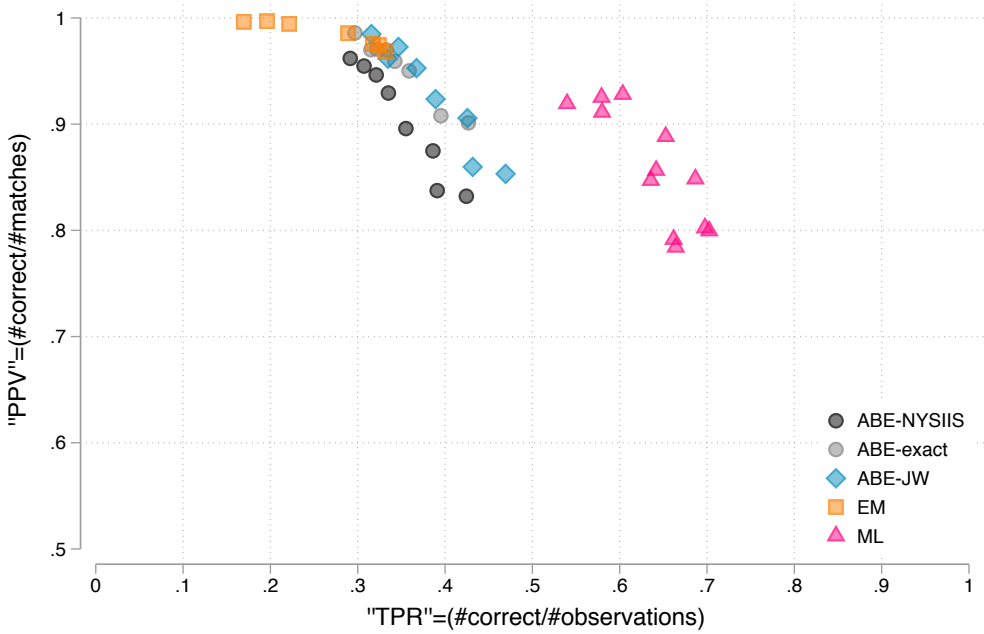
Notes: Panels (a-c) show PPV and TPR for the data linked from the Union Army Records to the 1900 Census using different variations of the ABE algorithm. Panel (a) uses NYSIIS-standardized names, Panel (b) uses exact names, and Panel (c) uses JW distance. In each panel we report versions using middle initials or not, different uniqueness requirements, and allowing matches to differ in reported age by ≤ 2 years or ≤ 1 year. Panel (d) shows versions of the EM algorithm that use different combinations of the hyperparameters p (the minimum score for the first-best match) and l (the maximum score for the second-best match). Panel (e) shows different variations of the ML algorithm. These variants change the person who carried out the training hand-linking (JF or HL), whether the algorithm is allowed to use middle initials or not, and the function of PPV and TPR that the algorithm aims to maximize. "Even" refers to the function PPV+TPR. "PPV" refers to the function 3PPV+TPR. "TPR" refers to the function PPV+3TPR.

Figure 3: Accuracy vs. efficiency using intersections of matching methods (UA-1900 census)



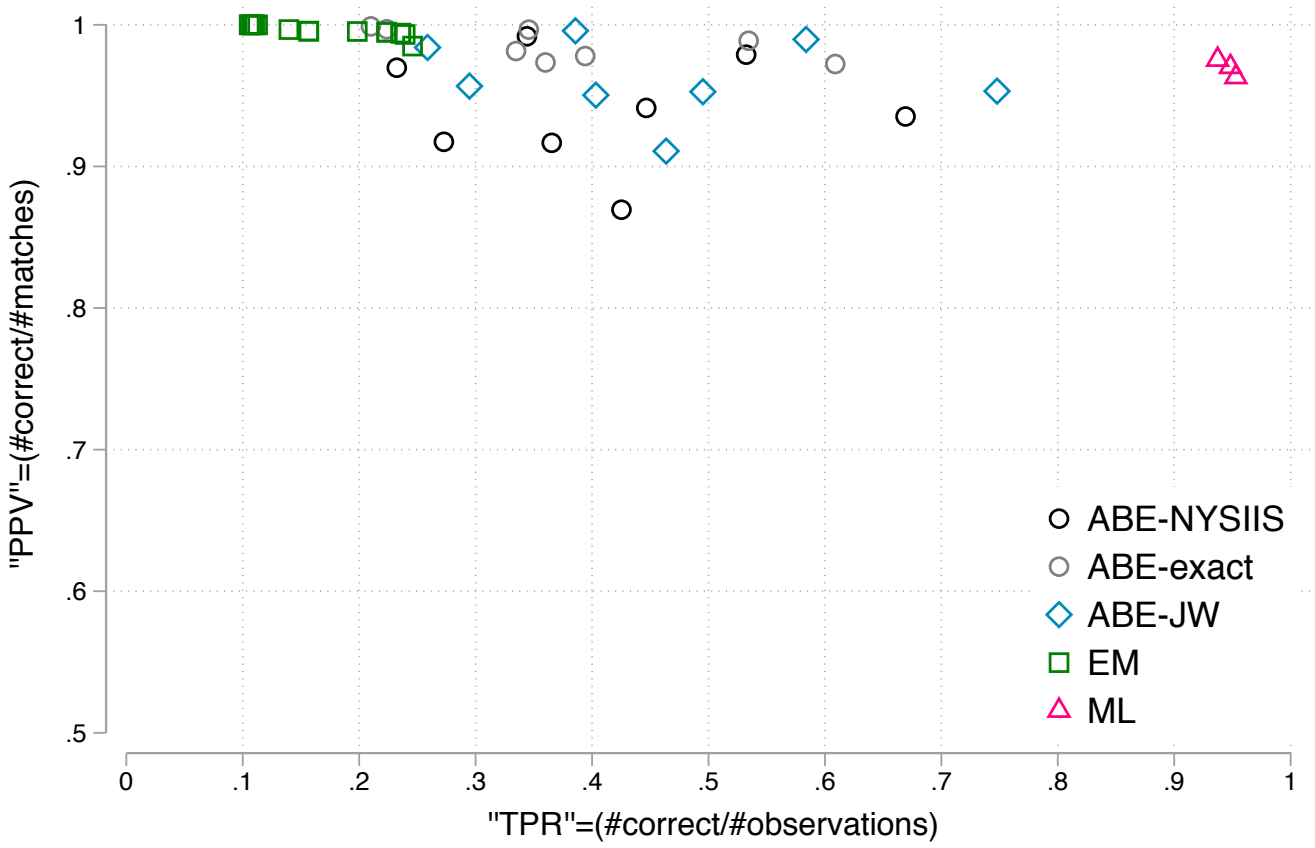
Notes: PPV ($\frac{\#truelinks}{\#matched}$) and TPR ($\frac{\#truelinks}{\#ofobservations}$) for the Union Army records linked to the 1900 census using five alternative matching algorithms and their intersections. Each intersection sample keeps only pairs that were found using all methods listed. For instance, the intersection between ABE-NYSIIS and ABE-exact is the sample of all pairs found both when using the ABE algorithm with NYSIIS names and with exact names. The PPV and TPR values of each intersection are listed in Appendix Table 2. All ABE methods require uniqueness within ± 2 years of age (5-year uniqueness band). None of the methods use middle names or middle initials in matching. The ML method uses equal weighting on PPV and TPR.

Figure 4: Differences Between Hand and Automated Methods when Using Same Information for Linking (UA-1900 census)



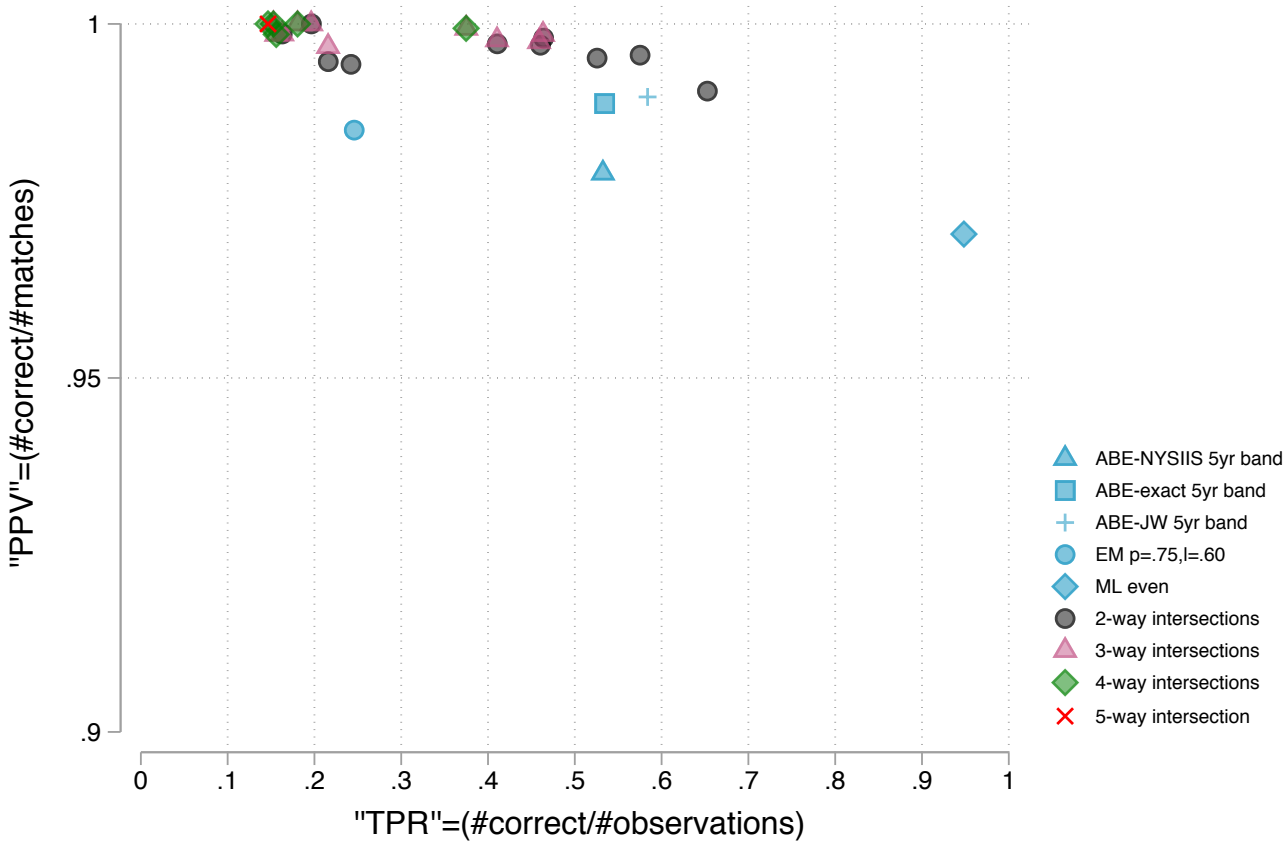
Notes: "PPV" ($\frac{\#truelinks}{\#matched}$) and "TPR" ($\frac{\#truelinks}{\#ofobservations}$) for the exercise comparing the automated matches to the hand-link matches when matching the Union Army records to the 1900 census. For presentational purposes, a match is defined as "true" if it coincides with the hand-link that a person made when using the same information that the automated algorithms use (name, place of birth, and year of birth).

Figure 5: Differences Between Hand and Automated Methods when Using Same Information for Linking (Iowa-1940 census)



Notes: PPV ($\frac{\#truelinks}{\#matched}$) and TPR ($\frac{\#truelinks}{\#ofobservations}$) for the 1915 Iowa records linked to the 1940 census using variations of each linking algorithm. A match is defined as "true" if it coincides with the match that human hand-linkers realized. The different estimates correspond to different variations of the ABE, EM and ML algorithms.

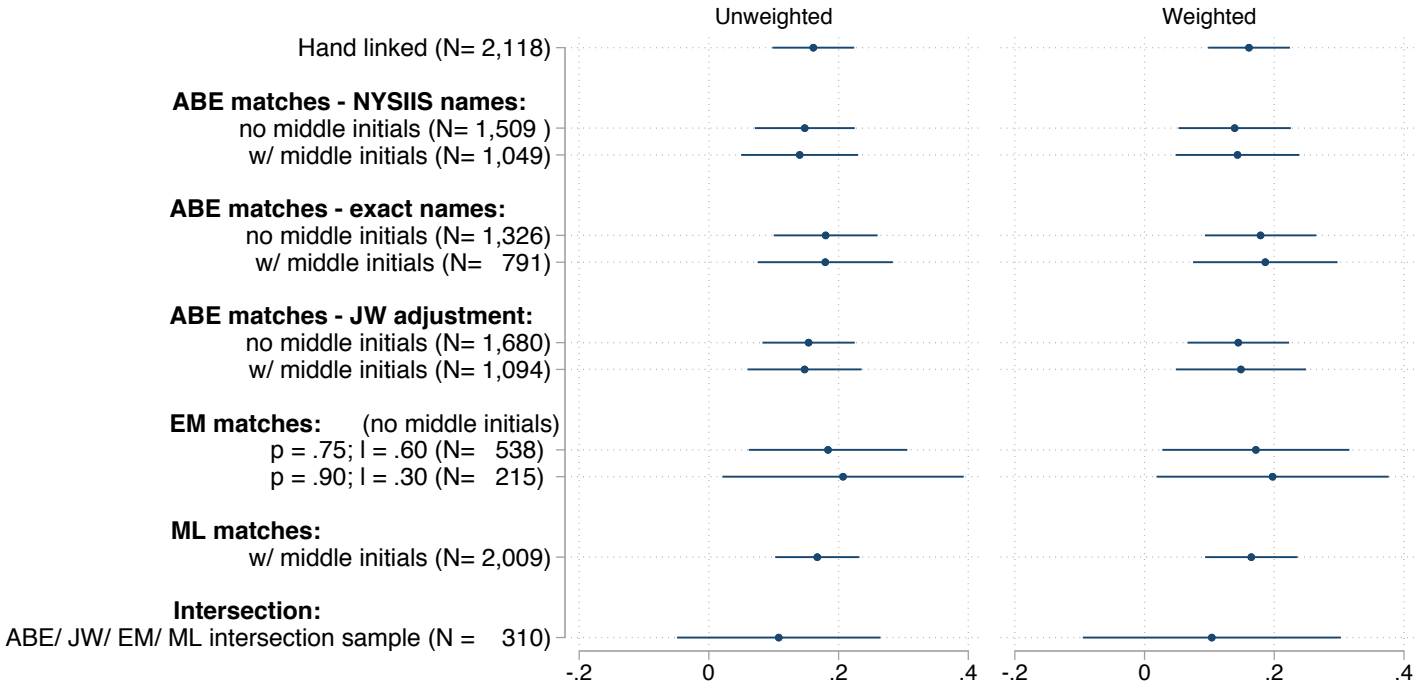
Figure 6: Accuracy vs. efficiency using intersections of matching methods (Iowa-1940 census)



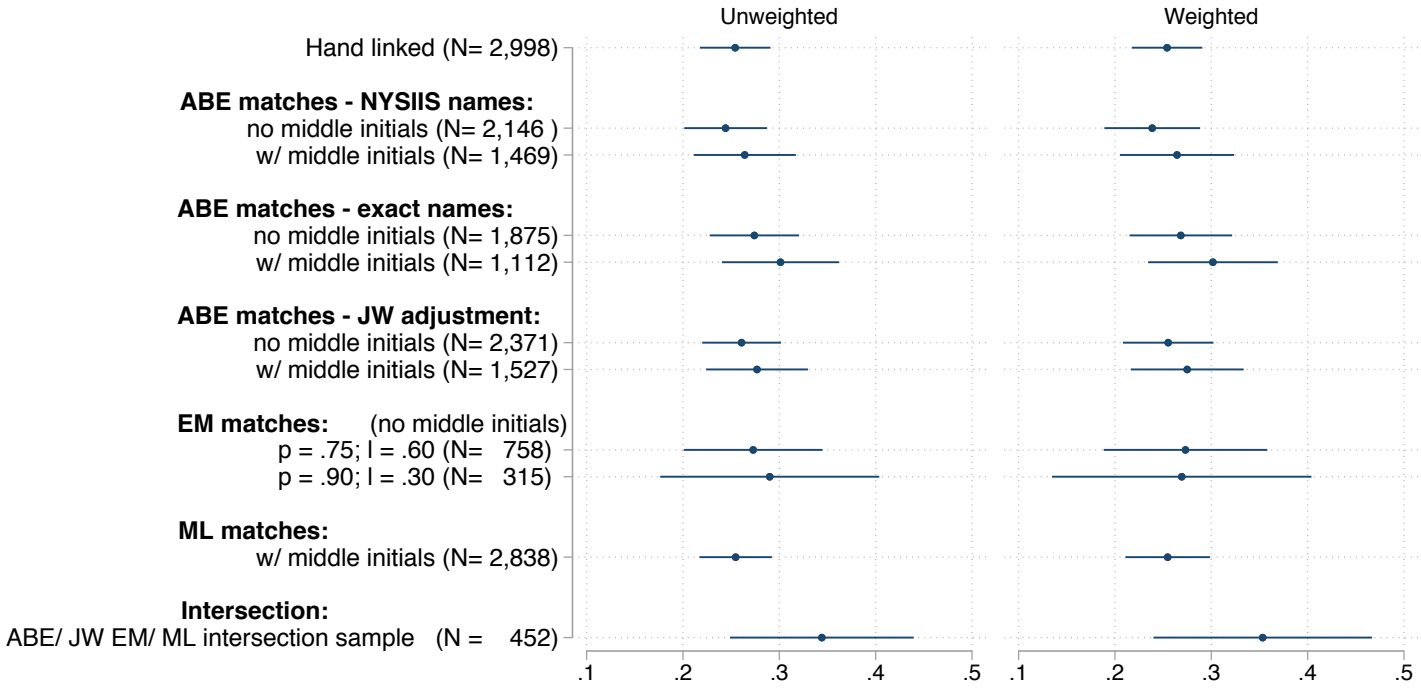
Notes: PPV ($\frac{\#truelinks}{\#matched}$) and TPR ($\frac{\#truelinks}{\#ofobservations}$) for the 1915 Iowa records linked to the 1940 census using five alternative matching algorithms and their intersections. Each intersection sample keeps only pairs that were found using all methods listed. For instance, the intersection between ABE-NYSIIS and ABE-exact is the sample of all pairs found both when using the ABE algorithm with NYSIIS names and with exact names. The PPV and TPR values of each intersection are listed in Appendix Table 3. All ABE methods require uniqueness within ± 2 years of age (5-year uniqueness band). None of the methods (with the exception of ML) use middle names or middle initials in matching. The ML method uses middle initials and equal weighting on PPV and TPR.

Figure 7: Regression analysis with the Iowa-1940 data

(a) Earnings intergenerational elasticity



(b) Relationship between fathers' and sons' education



Notes: Panel (a) plots the coefficient and 95% confidence interval from a regression of the ln(sons' earnings) on ln(fathers' earnings). Sons' earnings are reported in the 1940 Census, fathers' earnings are from the 1915 Iowa Census. Panel (b) plots the coefficient and 95% confidence interval from a regression of sons' years of schooling, reported in the 1940 Census, on their fathers' years of schooling reported in the 1915 Iowa census. We also show results weighted to account for differences between the linked sample and the population with respect to the following variables in the Iowa 1915 records: year of birth, literacy, years of school, and foreign-born status of parents. The samples are the same as in Figure 5; we only attempt to match those individuals that were successfully linked by the hand linkers. The ABE samples are linked without the 5-year uniqueness band requirement and allow matched pairs to differ by up to 2 years in reported age. The ML results use the "even" method, which aims to maximize PPV+TPR. The intersection sample is described in Figure 6.

Table 1. Link rates to FamilySearch of potential matches

	N	Percent
A. Traditional ABE method		
Both Censuses Attached to FamilySearch	3,387,588	100.00
Attached to Same Individual	3,225,847	95.23
Attached to Different Individuals	161,741	4.77
B. Conservative ABE method		
Both Censuses Attached to FamilySearch	2,779,618	100.00
Attached to Same Individual	2,699,833	97.13
Attached to Different Individuals	79,785	2.87
C. EM method		
Both Censuses Attached to FamilySearch	1,358,101	100.00
Attached to Same Individual	1,326,101	97.64
Attached to Different Individuals	32,000	2.36

Notes: This table compares the links made by automated methods to the links made by the Record Linking Lab at Brigham Young University (BYU) when linking the 1910 to the 1920 US censuses of population. Panel (a) shows the comparison using the standard version of the ABE algorithm, whereas panel (b) uses the conservative version that requires an individual to be unique within a 5-years window. Panel (c) uses the EM algorithm.

Table 2. Balance Test of Matched vs Unmatched Observations for ABE Algorithm

(a) UA-1900 census

		5 Years Band				No Band			
		Middle Initials		No Middle Initials		Middle Initials		No Middle Initials	
Variable	Mean	NYSIIS	Exact	NYSIIS	Exact	NYSIIS	Exact	NYSIIS	Exact
Year of Birth	1839.190	-0.379	0.275	-0.563	-0.180	-0.007	0.575	-0.335	0.193
	(6.007)	(0.317)	(0.315)	(0.323)	(0.315)	(0.302)	(0.305)	(0.298)	(0.299)
Literate	0.938	0.009	0.017	0.010	0.006	0.004	0.011	-0.002	-0.004
	(0.241)	(0.015)	(0.015)	(0.015)	(0.015)	(0.014)	(0.014)	(0.014)	(0.014)
Height (inches)	67.474	0.271	0.373	0.146	0.300	0.227	0.316	0.088	0.249
	(2.316)	(0.121)	(0.123)	(0.123)	(0.122)	(0.117)	(0.120)	(0.117)	(0.117)
Occupation Score	35.345	-0.212	-0.554	0.075	-0.029	-0.416	-0.707	0.443	-0.492
	(9.903)	(0.536)	(0.551)	(0.543)	(0.533)	(0.517)	(0.535)	(0.512)	(0.513)
Enlistment Age	22.746	0.396	-0.220	0.557	0.243	0.079	-0.414	0.278	-0.065
	(5.567)	(0.307)	(0.302)	(0.313)	(0.304)	(0.291)	(0.293)	(0.287)	(0.288)

(b) Iowa-1940 census

		5 Years Band				No Band			
		Middle Initials		No Middle Initials		Middle Initials		No Middle Initials	
Variable	Mean	NYSIIS	Exact	NYSIIS	Exact	NYSIIS	Exact	NYSIIS	Exact
Urban dummy	0.413	-0.025	0.009	-0.037	-0.002	-0.015	0.016	-0.026	0.021
	(0.492)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)	(0.017)	(0.016)
Age	9.636	-0.191	-0.100	-0.300	-0.243	-0.104	-0.102	-0.312	-0.183
	(4.359)	(0.136)	(0.142)	(0.134)	(0.134)	(0.133)	(0.139)	(0.148)	(0.138)
US-born father dummy	0.777	0.010	-0.016	0.026	0.022	-0.009	-0.011	0.058	0.051
	(0.416)	(0.013)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.015)	(0.013)
US-born mother dummy	0.812	0.010	-0.013	0.019	0.015	-0.002	-0.012	0.045	0.036
	(0.391)	(0.012)	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)	(0.014)	(0.013)
Years schooling	3.979	-0.161	-0.076	-0.168	-0.140	-0.092	-0.086	-0.203	-0.137
	(3.394)	(0.106)	(0.110)	(0.104)	(0.104)	(0.104)	(0.108)	(0.116)	(0.107)
Literacy	0.743	-0.016	-0.029	-0.016	-0.022	-0.013	-0.022	-0.000	-0.006
	(0.437)	(0.014)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.015)	(0.014)

Notes: Panel (a): The first column corresponds to the population mean and standard deviation of all observations in the hand-linked Union Army records. The table presents the balance test across the 8 different variations on the ABE algorithm corresponding columns 2 - 9 (e.g.: the second column is the balance test for the matched population under the ABE algorithm with 5 years band, including middle initials and using the NYSIIS standardized names). Estimates correspond to the difference in averages between the matched and unmatched observations for a given algorithm and are reported as the first line of each variable. The standard deviations are reported in the corresponding second line in parenthesis. Panel (b) repeats the same exercise for the Iowa 1915 to 1940 census linking exercise.

Table 3. Transcription Differences By Place of Birth: Ancestry vs. Family Search

	Iowa	North.Carolina	Ohio	England	Italy
First name differ after cleaning odd characters	0.116	0.160	0.109	0.096	0.152
First name differ after removing suffix	0.116	0.160	0.109	0.096	0.152
First name differ after uppercasing	0.116	0.159	0.109	0.096	0.152
First name differ after removing middle names	0.095	0.138	0.091	0.078	0.149
First name differ after removing nicknames	0.090	0.133	0.086	0.072	0.143
First name differ after NYSIIS standardization	0.063	0.093	0.060	0.050	0.085
First name differ less than 0.1 JW distance	0.944	0.916	0.950	0.961	0.935
Last name differ after cleaning odd characters	0.197	0.198	0.192	0.182	0.320
Last name differ after uppercasing	0.185	0.177	0.176	0.170	0.315
Last name differ after NYSIIS standardization	0.126	0.115	0.121	0.116	0.204
surname differ less than 0.1 JW distance	0.907	0.903	0.910	0.914	0.865
Middle name differ	0.031	0.028	0.025	0.022	0.006
Middle initial differ	0.027	0.020	0.021	0.020	0.006
Middle name differ after NYSIIS standardization	0.029	0.025	0.023	0.022	0.006
Middle name differ less than 0.1 JW distance	0.970	0.974	0.976	0.979	0.994
Ages differ	0.020	0.020	0.020	0.020	0.030

Notes: This table summarizes the transcription differences in names and ages between two versions of the 1940 US census, one transcribed by FamilySearch and one transcribed by Ancestry. We focus on comparing first, last and middle names under the different stages of our cleaning algorithm. For example, for those born in Iowa 11.6% of the transcriptions had some difference in first name transcription after cleaning odd characters, removing suffix and uppercasing. But after removing middle names this figure was reduced to 9.5%. Note that a small number of the transcription differences are due to errors in indexing which makes it impossible to exactly crosswalk between the Ancestry and Family, so these reported differences are an upper bound on transcription errors.

Table 4. Confusion Table - ABE Algorithm (Ancestry-FS)

	N (1)	Matched (2)	Matching Rate (3)=(2)/(1)	Number Correct (4)	PPV (5)=(4)/(2)	TPR (6)=(4)/(1)
I. No Middle Name						
a. Exact Names						
Unique 5-years band	14790293	7507944	0.51	7470350	0.99	0.51
Age difference=0		7458587	0.50	7435896	1.00	0.50
Age difference=1		27542	0.00	19602	0.71	0.00
Age difference=2		21815	0.00	14852	0.68	0.00
Non-Unique 5-years band	14790293	9104684	0.62	9047893	0.99	0.61
Age difference=0		9041350	0.61	9008543	1.00	0.61
Age difference=1		35605	0.00	22390	0.63	0.00
Age difference=2		27729	0.00	16960	0.61	0.00
b. Standardized names						
Unique 5-years band	14790293	6410781	0.43	6365447	0.99	0.43
Age difference=0		6357854	0.43	6335851	1.00	0.43
Age difference=1		29332	0.00	16846	0.57	0.00
Age difference=2		23595	0.00	12750	0.54	0.00
Non-Unique 5-years band	14790293	8908661	0.60	8821739	0.99	0.60
Age difference=0		8829030	0.60	8785296	1.00	0.59
Age difference=1		44881	0.00	20830	0.46	0.00
Age difference=2		34750	0.00	15613	0.45	0.00
II. Middle Initials						
a. Exact Names						
Unique 5-years band	14790293	8512580	0.58	8477630	1.00	0.57
Age difference=0		8461533	0.57	8437952	1.00	0.57
Age difference=1		28605	0.00	22503	0.79	0.00
Age difference=2		22442	0.00	17175	0.77	0.00
Non-Unique 5-years band	14790293	9645242	0.65	9597270	1.00	0.65
Age difference=0		9584373	0.65	9553970	1.00	0.65
Age difference=1		34189	0.00	24553	0.72	0.00
Age difference=2		26680	0.00	18747	0.70	0.00
b. Standardized names						
Unique 5-years band	14790293	7980056	0.54	7936506	0.99	0.54
Age difference=0		7922904	0.54	7898963	1.00	0.53
Age difference=1		31801	0.00	21370	0.67	0.00
Age difference=2		25351	0.00	16173	0.64	0.00
Non-Unique 5-years band	14790293	9925823	0.67	9851756	0.99	0.67
Age difference=0		9848221	0.67	9808523	1.00	0.66
Age difference=1		43623	0.00	24625	0.56	0.00
Age difference=2		33979	0.00	18608	0.55	0.00

Notes: This table reports the results from the matching exercise between the transcription done by Ancestry and the transcription done by FamilySearch. Instead of matching the entire 1940 census to itself we match people born in four different places of birth: England, Italy, Iowa, Ohio and North Carolina. We report the overall matching rate, PPV, and TPR for each of the methods. A correct link is defined as a link that shares the same Census page and line. Sum of matches in a given method is equal to the sum of =0 + =1 + =2 age difference in that method. For instance, first 4 rows: we were able to match 7,507,944 records, so Matched= 7,507,944, and out of these 7,507,944, age difference = 0 for 7,458,587 observations, age difference = 1 for 27,524 observations and age difference = 2 for 21,815 observations. Age difference=2, PPV=0.68 means that among all the matches that are two years apart in terms of year of birth (i.e. we match John Smith 1838 to John Smith 1840), 68% of them are right.

Table 5. Comparison of summary measures of intergenerational occupational mobility

(a) Fraction working in different occupational category than father

	US 1850 - 1880			Norway 1865 - 1900		
	Observed fraction	Expected fraction under father-son independence	Difference	Observed fraction	Expected fraction under father-son independence	Difference
IPUMS	0.45	0.62	0.16***	0.47	0.70	0.23***
ABE - less conservative	0.50	0.65	0.14***	0.48	0.70	0.22***
ABE - more conservative	0.47	0.63	0.16***	0.46	0.70	0.24***
JW - less conservative	0.50	0.64	0.15***	0.48	0.70	0.22***
JW - more conservative	0.47	0.63	0.16***	0.46	0.70	0.24***
EM - less conservative	0.45	0.62	0.16***	0.45	0.70	0.25***
EM - more conservative	0.44	0.60	0.16***	0.44	0.71	0.27***

(b) Distance with respect to independence

	US 1850 - 1880	**Norway 1865 - 1900**
IPUMS	17.37***	25.01***
ABE - less conservative	12.17***	24.19***
ABE - more conservative	14.07***	25.04***
JW - less conservative	12.47***	24.21***
JW - more conservative	14.45***	25.17***
EM - less conservative	14.67***	25.94***
EM - more conservative	15.18***	26.09***

Notes: This table reports summary measures using 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 occupational transition matrix), along with the expected fraction under independence and the difference between these measures. 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.

Appendix 1: ABE Matching Procedure Example

This appendix outlines an update to the standard ABE matching code after the publication of Abramitzky, Boustan, and Eriksson (2012, 2014, 2019). The original code matched from one direction to another. It also had a small logical mistake where it could potentially match the wrong two people. Correcting this logical mistake requires a small change to the code; in addition, so that the example outlined below could not happen, we now require matching in both directions, after which the code takes the intersection.

Example:

Dataset A:

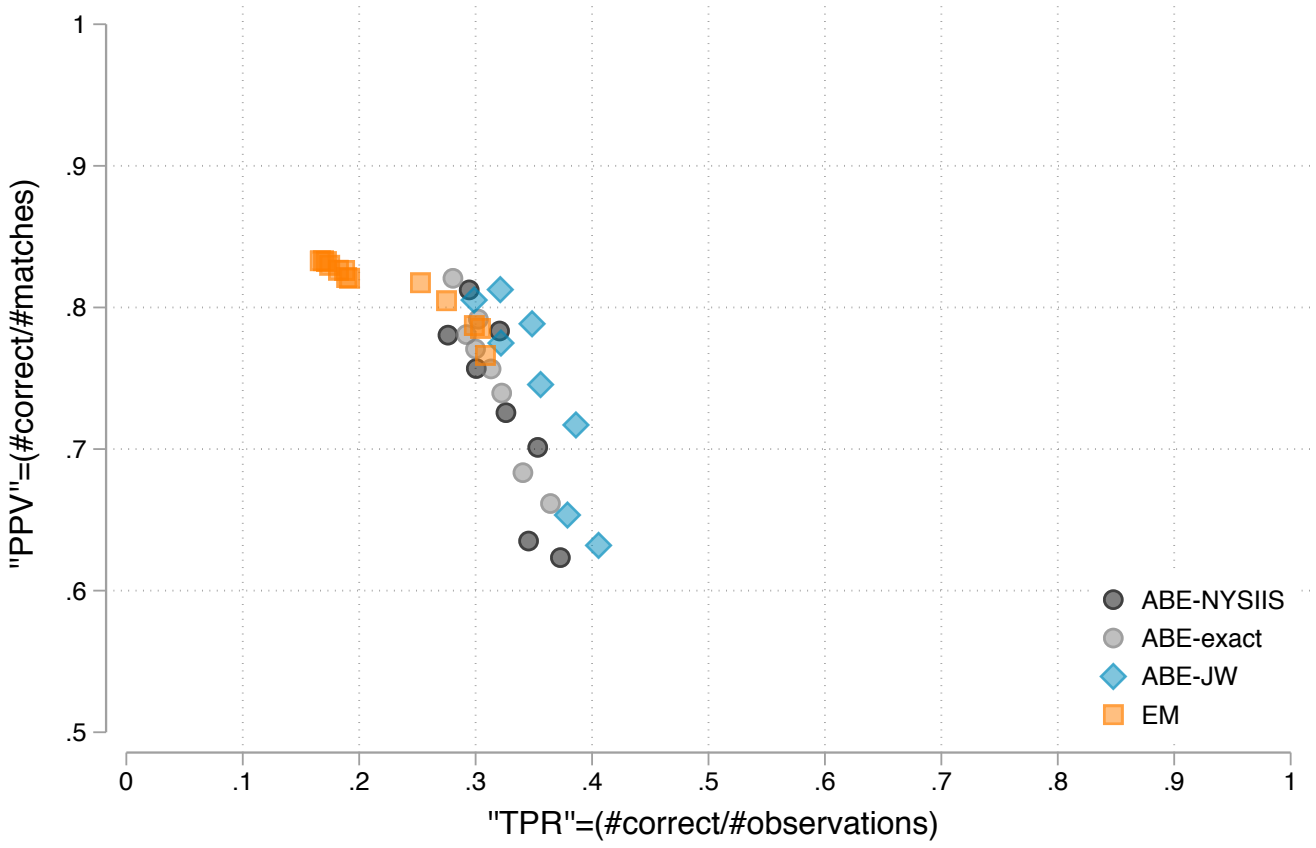
ID	Name	Birthyear
1A	John Smith	1900
2A	John Smith	1899

Dataset B:

ID	Name	Birthyear
1B	John Smith	1900
2B	John Smith	1900
3B	John Smith	1901

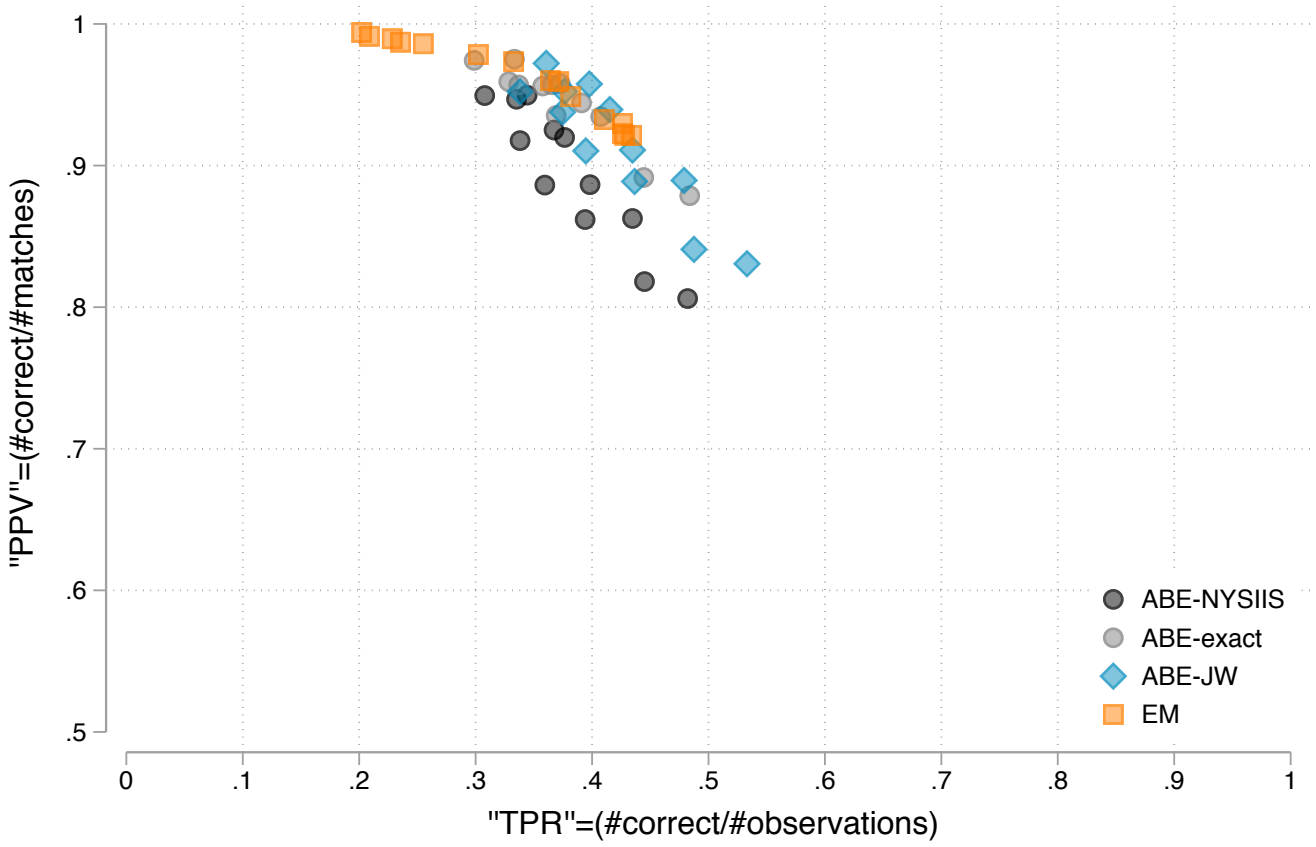
If we match from Dataset A to Dataset B using the original iterative ABE method, we will match 1A to 1B and 2B, and drop these individuals from the dataset because 1A has multiple matches. Then we will match 2A to 3B after iterating to plus or minus two years of age. The new abematch command does not make this second match because it flags 1B and 2B as possible matches for 2A. Because this pattern could happen in either direction, we now require that the code matches both directions and takes the resulting intersection of the two matched samples. This change affects approximately 1% of matches in most samples.

Figure A.1: Accuracy vs. Efficiency Using all Available Records (UA-1900 census)



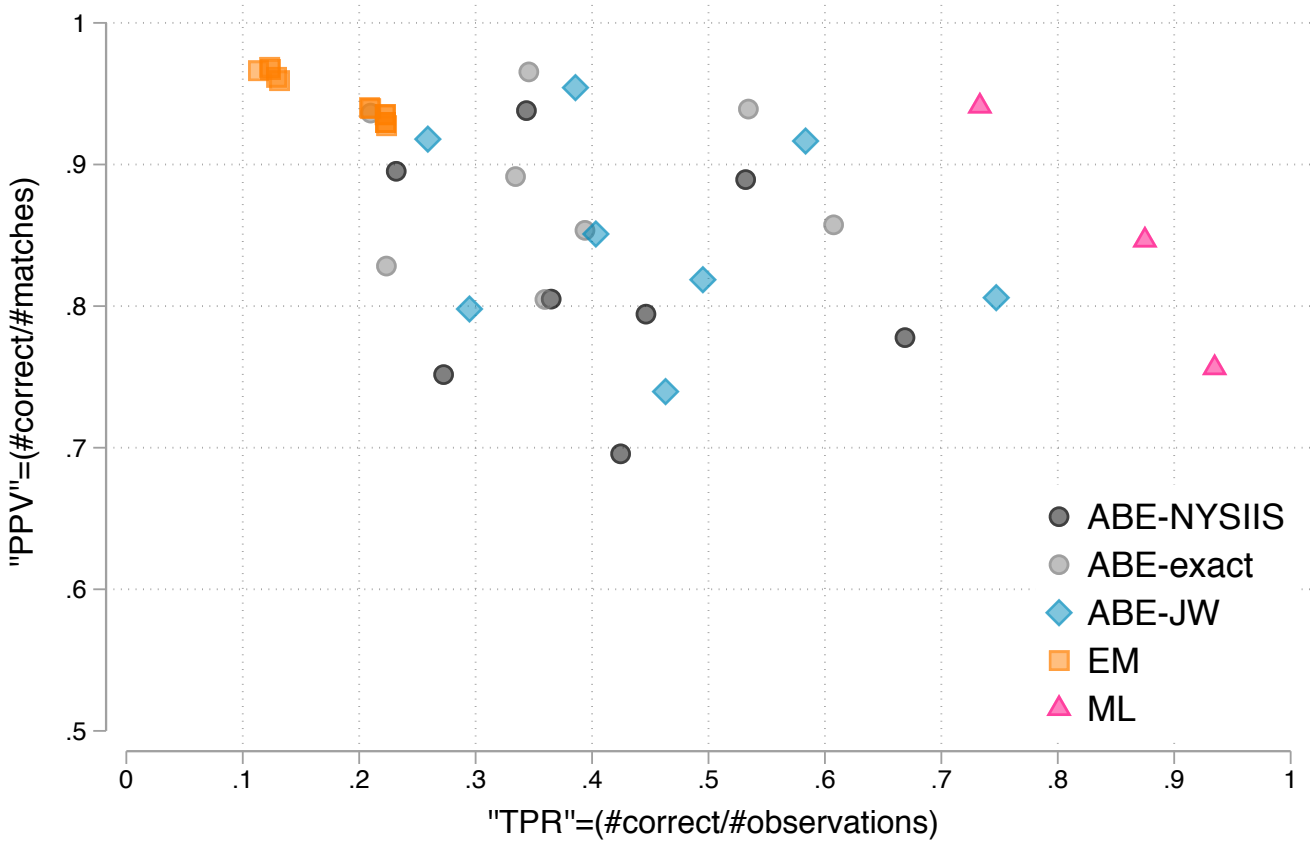
Notes: PPV ($\frac{\#truelinks}{\#matched}$) and TPR ($\frac{\#truelinks}{\#ofobservations}$) for the exercise the Union Army records to the 1900 census using different variations of the four linking algorithms (ABE, ABE-JW, EM and ML). In this version we attempt to find links for all available records in the full UA data set. "Hand" compares the hand-linking carried out by two different people that only used the same information that the automated algorithms use (name, place of birth, and year of birth). A match is defined as "true" if it coincides with the links in the Oldest Old sample. In this case, we also consider a match to be "false" if the automated algorithm identified a match but the hand linker did not.

Figure A.2: Differences Between Hand and Automated Methods when Using Same Information for Linking, Using All Available Records (UA-1900 census)



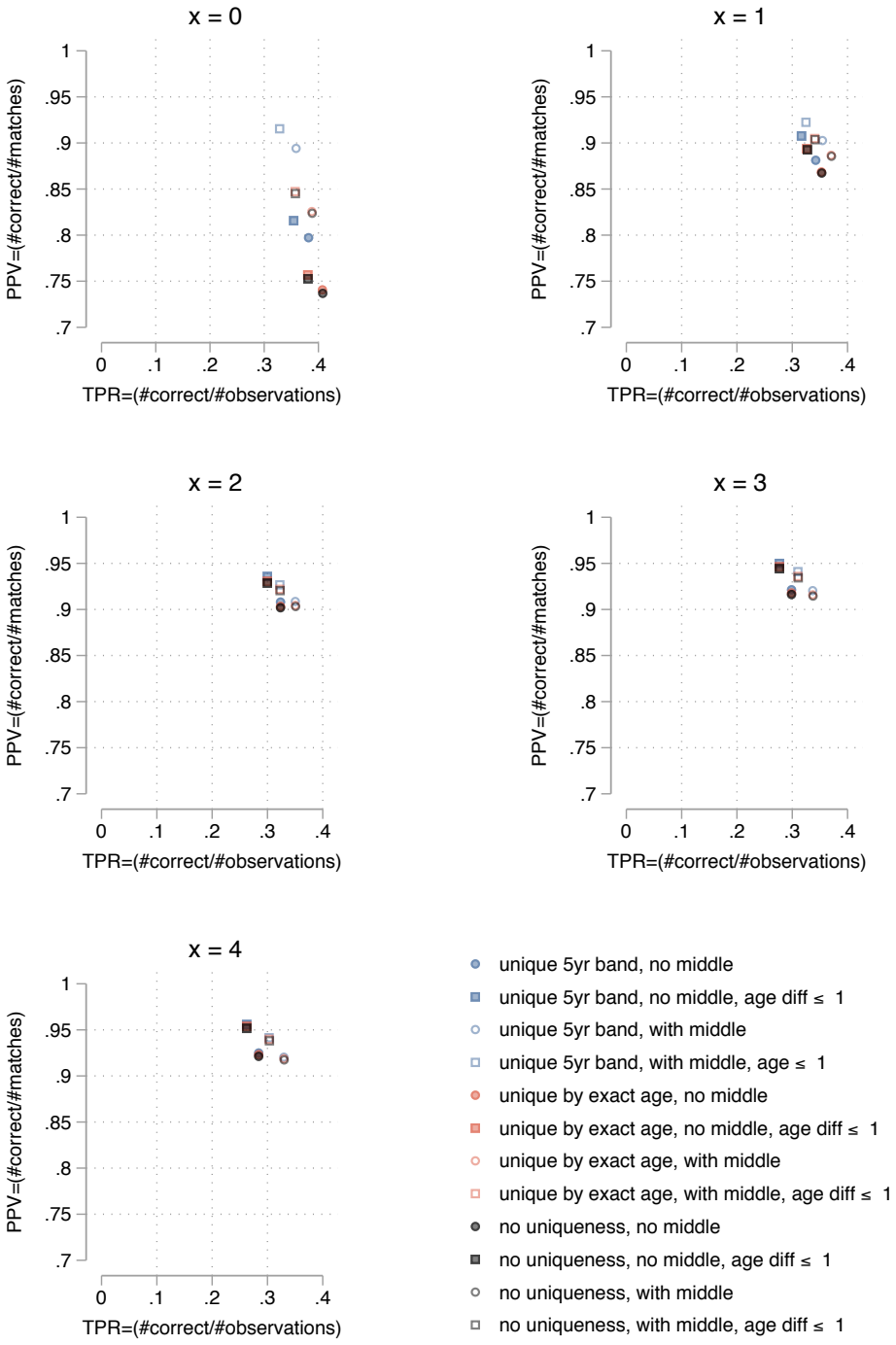
Notes: "PPV" ($\frac{\#truelinks}{\#matched}$) and "TPR" ($\frac{\#truelinks}{\#ofobservations}$) for the exercise comparing the automated matches to the hand-link matches when matching the Union Army records to the 1900 census. In this version we attempt to find links for all available records in the full UA data set. For presentational purposes, a match is defined as "true" if it coincides with the hand-link that a person did using only the same information that the automated algorithms use (name, place of birth, and year of birth). In this case, we also consider a match to be "false" if the automated algorithm identified a match but the hand linker did not.

Figure A.3: Differences Between Hand and Automated Methods when Using Same Information for Linking, Using All Available Records (Iowa-1940 census)



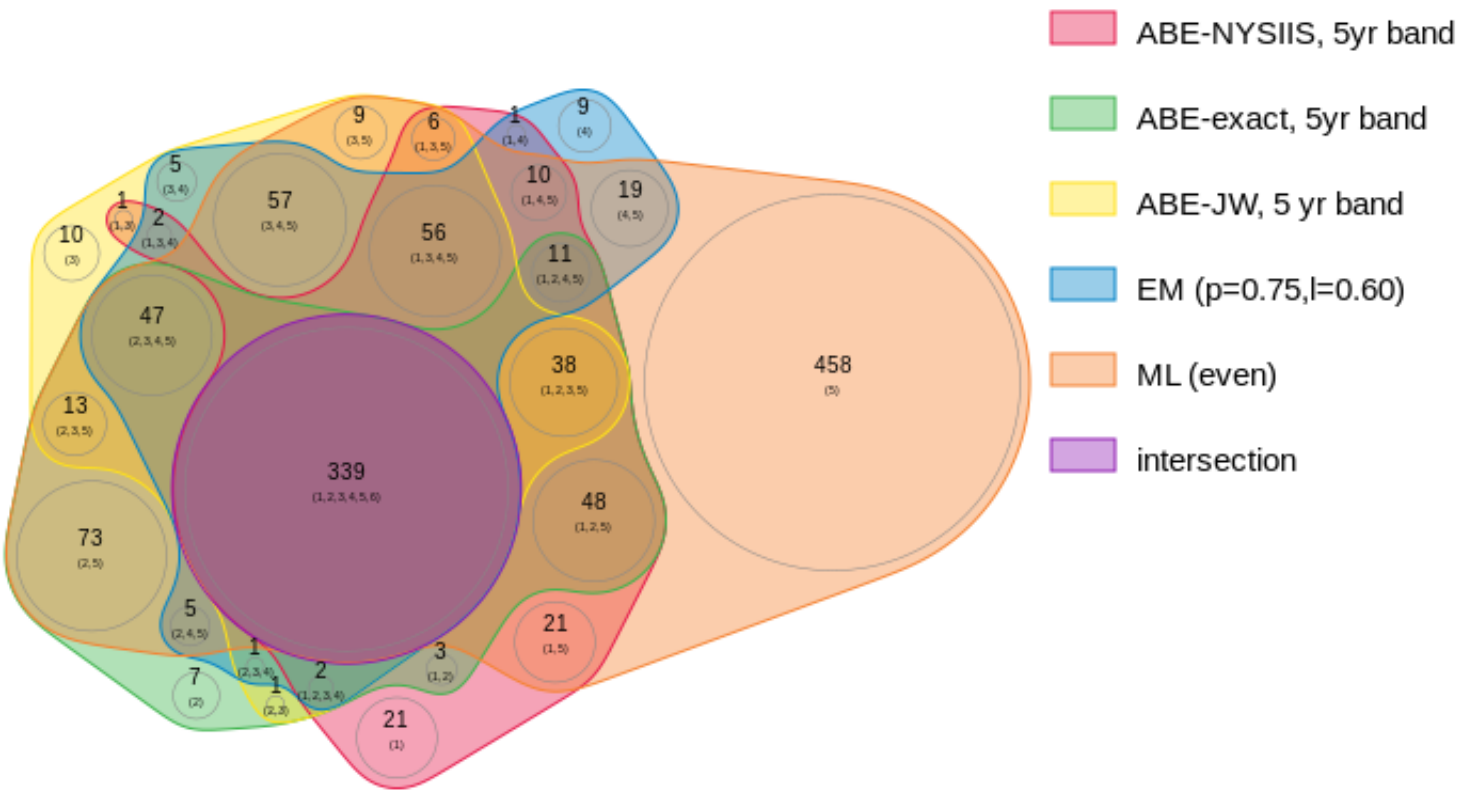
Notes: PPV ($\frac{\#truelinks}{\#matched}$) and TPR ($\frac{\#truelinks}{\#ofobservations}$) for the 1915 Iowa records linked to the 1940 census using variations of two linking algorithms (ABE, and EM). In this version we attempt to find links for all available records in the full Iowa data set. A match is defined as "true" if it coincides with the match that human hand-linkers identified. In this case, we also consider a match to be "false" if the automated algorithm identified a match but the hand linker did not.

Figure A.4: Comparing Versions of the ABE-JW Algorithm (UA-1900 census)



Notes: The figure shows the PPV and TPR for the data linked from the Union Army Records to the 1900 US Census using different versions of the ABE-JW algorithm. The parameter x determines how far in age the second closest match can be to the first closest match. For each value of x we also present versions requiring all observations to be unique by JW string distance within their own data set either by exact age or within $\pm$ 2 years.

Figure A.5: Overlap of matching methods (UA-1900 census)



Notes: This Venn-diagram shows the overlap in matched pairs found using five alternative matching algorithms to link the Union Army records to the 1900 census. The purple intersection area contains all the 339 matched pairs that were found by all 5 methods. All ABE methods require uniqueness within ± 2 years of age (5-year uniqueness band). None of methods use middle names or middle initials in matching. The ML method uses equal weighting on PPV and TPR.

Figure A.6: Overlap of matching methods (Iowa-1940 census)

Notes: This Venn-diagram shows the overlap in matched pairs found using five alternative matching algorithms to link the 1915 Iowa records to the 1940 census. The purple intersection area contains all the 633 matched pairs that were found by all 5 methods. All ABE methods require uniqueness within ± 2 years of age (5-year uniqueness band). None of the methods (with the exception of ML) use middle names or middle initials in matching. The ML method uses middle names and equal weighting on PPV and TPR.

Figure A.7: Disagreement Cases: Hand Links vs. EM Algorithm (Iowa-1940 census)

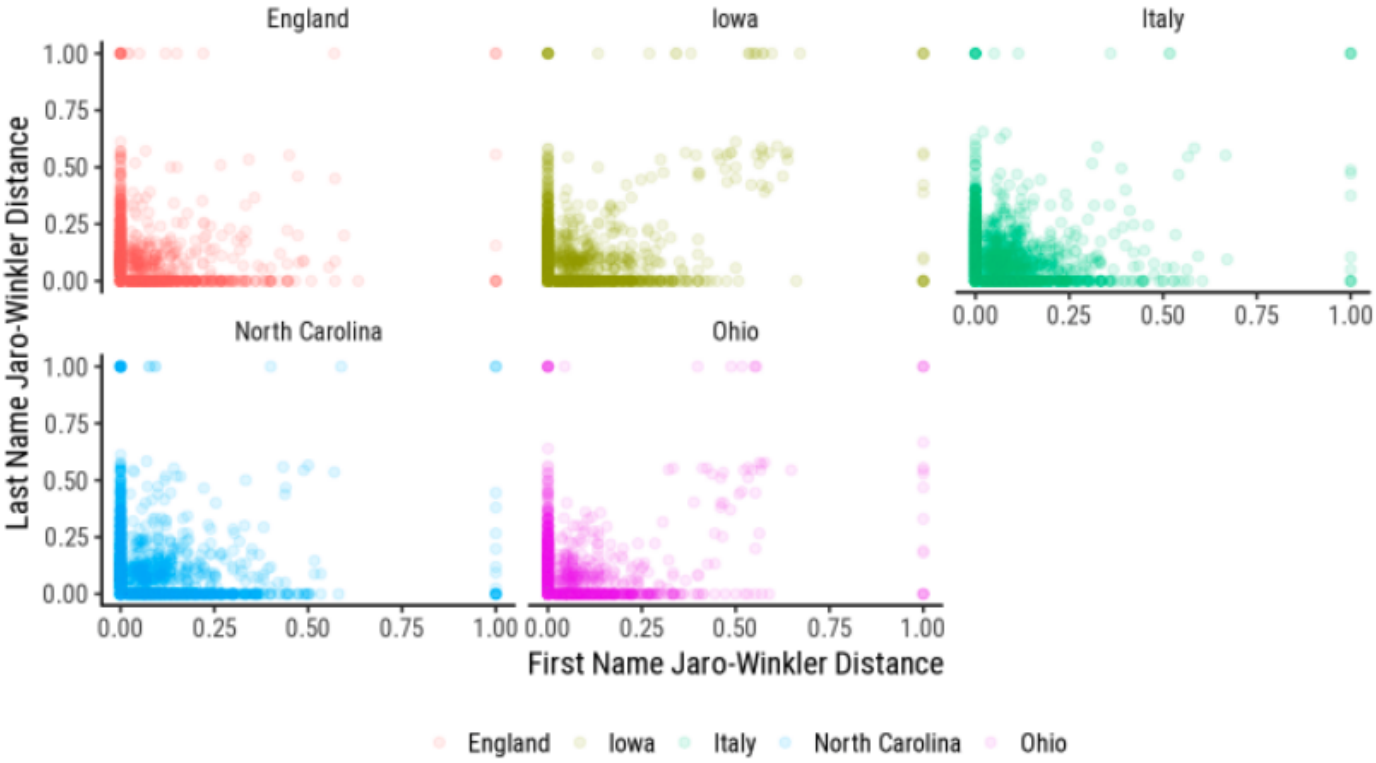
Origin Data to Match			
first name	last name	year of birth	state of birth
oren	miller	1899	Iowa
lloyd f	scheel	1911	Iowa
lee	jarrett	1901	Indiana
lawrence	frick	1905	Iowa
lelo	lunardi	1904	Italy
lester	groff	1912	Iowa
john	obman	1908	Iowa
otto	strasser	1904	Iowa
lem	carl	1902	Iowa
leonard	haas	1903	Iowa
leslie r	boling	1904	Iowa
leymour	morrison	1900	Missouri
raymond g	mack	1902	Iowa
paul	coulter	1912	Kansas
paul	briggs	1908	Iowa
william	noel	1910	Iowa

Hand Matches			
first name	last name	year of birth	state of birth
Orin	Miller	1900	IA
Lloyd F	Cheel	1912	IA
Leo	Jarrett	1902	IN
Sawrence	Frick	1904	IA
Leon	Lenardi	1905	ITA
Leslie	Groff	1911	IA
John	Ohmann	1908	IA
Otto W	Stuesser	1904	IA
Leo	Carl	1900	IA
Leonard	Hass	1902	IA
Lester	Boling	1905	IA
Seymour	Morrison	1900	MO
Raymond	Muck	1902	IA
Paul	Coater	1912	KS
Parl	Briggs	1906	IA
William G	Noll	1909	IA

Automated Matches			
first name	last name	year of birth	state of birth
Oren	Mills	1899	IA
Leroy	Schemmel	1911	IA
Lewis	Jarrett	1901	IN
Lawrence E	Fick	1907	IA
Lee G	Lunardi	1904	ITA
Lewis A	Groff	1912	IA
John M	Orman	1907	IA
Otto A	Sasse	1905	IA
L M	Carl	1902	IA
Leo J	Haas	1903	IA
Lessie	Bowlin	1904	IA
Leo J	Morrison	1900	MO
Raymond E D	Macy	1903	IA
Paul	Courter	1911	KS
Paul	Briggle	1908	IA
William F	Noxsel	1911	IA

Notes: This table shows all the matches that were found using the EM algorithm with parameters $p = 0.70$, $l = 0.65$ and that disagree with the matches identified by hand linkers, for the Iowa-1940 matching exercise. EM 70-65 had TPR = 0.25 (correct / observations) and PPV = 0.98 (correct / matches). The very high PPV means that, out of those observations that EM 75-60 matched, very few of them disagree with the hand matches. This is the complete list of those disagreements, a total of 16 cases.

Figure A.8: Jaro Winkler distances, Ancestry-Family Search 1940 census



Notes: This figure shows the Jaro-Winkler distance in first (x-axis) and last (y-axis) names between the Ancestry and Family Search transcriptions of the 1940 census, by place of birth. If both transcriptions were identical, all the observations would be on the origin.

Table A.1: Confusion Table - ABE Algorithm (UA-1900 census)

	N (1)	Matched (2)	Matching Rate (3)=(2)/(1)	Number Correct (4)	PPV (5)=(4)/(2)	TPR (6)=(4)/(1)
I. No Middle Name						
a. Exact Names						
Unique 5-years band	1647	588	0.36	518	0.88	0.31
Age difference=0		233	0.14	211	0.91	0.13
Age difference=1		302	0.18	271	0.90	0.16
Age difference=2		53	0.03	36	0.68	0.02
Non-Unique 5-years band	1647	780	0.47	602	0.77	0.37
Age difference=0		357	0.22	265	0.74	0.16
Age difference=1		360	0.22	297	0.82	0.18
Age difference=2		63	0.04	40	0.63	0.02
b. Standardized names						
Unique 5-years band	1647	559	0.34	497	0.89	0.30
Age difference=0		223	0.14	206	0.92	0.13
Age difference=1		276	0.17	250	0.91	0.15
Age difference=2		60	0.04	41	0.68	0.02
Non-Unique 5-years band	1647	840	0.51	616	0.73	0.37
Age difference=0		407	0.25	278	0.68	0.17
Age difference=1		362	0.22	292	0.81	0.18
Age difference=2		71	0.04	46	0.65	0.03
II. Middle Initials						
a. Exact Names						
Unique 5-years band	1647	546	0.33	500	0.92	0.30
Age difference=0		212	0.13	198	0.93	0.12
Age difference=1		284	0.17	265	0.93	0.16
Age difference=2		50	0.03	37	0.74	0.02
Non-Unique 5-years band	1647	622	0.38	533	0.86	0.32
Age difference=0		256	0.16	217	0.85	0.13
Age difference=1		309	0.19	278	0.90	0.17
Age difference=2		57	0.03	38	0.67	0.02
b. Standardized names						
Unique 5-years band	1647	594	0.36	531	0.89	0.32
Age difference=0		227	0.14	211	0.93	0.13
Age difference=1		303	0.18	276	0.91	0.17
Age difference=2		64	0.04	44	0.69	0.03
Non-Unique 5-years band	1647	727	0.44	585	0.80	0.36
Age difference=0		310	0.19	245	0.79	0.15
Age difference=1		343	0.21	294	0.86	0.18
Age difference=2		74	0.04	46	0.62	0.03

Notes: This table reports the results from the matching exercise between the Union Army and the 1900 census. We report the overall matching rate, PPV, and TPR for each of the ABE variants. A correct link is defined as a link that agreed with the links that were done by a team of genealogists as part of the Union Army Oldest-Old project. Sum of matches in a given method is equal to the sum of =0 + =1 + =2 age difference in that method. For instance, first 4 rows: we were able to match 588 records, so Matched=588 , and out of these 588, age difference = 0 for 233 observations, age difference = 1 for 302 observations and age difference = 2 for 53 observations. Age difference=2, PPV=0.68 means that among all the matches that are two years apart in terms of year of birth (i.e. we match John Smith 1838 to John Smith 1840), 68% of them are correct.

Table A.2: PPV and TPR of Intersection Methods (UA-1900 census)

Matching methods	N total matches (1)	N "correct" matches (2)	PPV (3)	TPR (4)
I. Single methods				
ABE-NYSIIS 5yr band	559	497	0.89	0.30
ABE-exact 5yr band	588	518	0.88	0.31
ABE-JW 5yr band	587	533	0.91	0.32
EM: $p = 0.75$, $l = 0.60$	564	509	0.90	0.31
ML: even	1234	896	0.73	0.54
II. Two-way intersections				
ABE-NYSIIS & ABE-exact	441	412	0.93	0.25
ABE-NYSIIS & JW 5-yr band	444	418	0.94	0.25
ABE-NYSIIS & EM	421	399	0.95	0.24
ABE-NYSIIS & ML	529	491	0.93	0.30
ABE-exact & ABE-JW	441	416	0.94	0.25
ABE-exact & EM	405	386	0.95	0.23
ABE-exact & ML	574	515	0.90	0.31
ABE-JW & EM	509	476	0.94	0.29
ABE-JW & ML	565	528	0.93	0.32
EM & ML	544	505	0.93	0.31
III. Three-way intersections				
ABE-NYSIIS & ABE-exact & ABE-JW	379	361	0.95	0.22
ABE-NYSIIS & ABE-exact & EM	352	338	0.96	0.21
ABE-NYSIIS & ABE-exact & ML	436	410	0.94	0.25
ABE-NYSIIS & ABE-JW & EM	399	380	0.95	0.23
ABE-NYSIIS & ABE-JW & ML	439	416	0.95	0.25
ABE-exact & EM & ML	416	397	0.95	0.24
ABE-exact & ABE-JW & EM	389	373	0.96	0.23
ABE-exact & ABE-JW & ML	437	416	0.95	0.25
ABE-exact & EM & ML	402	386	0.96	0.23
ABE-JW & EM & ML	499	474	0.95	0.29
IV. Four-way intersections				
ABE-NYSIIS & ABE-exact & ABE-JW & EM	341	328	0.96	0.20
ABE-NYSIIS & ABE-exact & ABE-JW & ML	377	361	0.96	0.22
ABE-NYSIIS & ABE-JW & EM & ML	395	379	0.96	0.23
ABE-NYSIIS & ABE-exact & EM & ML	350	338	0.97	0.21
ABE-exact & ABE-JW & EM & ML	386	373	0.97	0.23
V. Five-way intersection				
ABE-NYSIIS & ABE-exact & ABE-JW & EM & ML	339	328	0.97	0.20

Notes: This table shows the number of matches found between the Union Army records and the 1900 census using five alternative matching algorithms and their intersections. Each intersection sample keeps only pairs that were found using all methods listed, for instance the intersection between ABE-NYSIIS and ABE-exact is the sample of all pairs found both when using the ABE algorithm with NYSIIS names and with exact names. A match is considered "correct" if it coincides with the match that human hand-linkers found. PPV = # "correct" / # matches. TPR = #"correct"/#observations. These PPV and TPR values are plotted in Figure 3. All ABE methods require uniqueness within ± 2 years of age (5-year uniqueness band). None of the methods use middle names or middle initials in matching. The ML method uses equal weighting on PPV and TPR.

Table A.3: PPV and TPR of Intersection Methods (Iowa-1940 census)

Matching methods	N total matches (1)	N "correct" matches (2)	PPV (3)	TPR (4)
I. Single methods				
ABE-NYSIIS 5yr band	2325	2276	0.98	0.53
ABE-exact 5yr band	2311	2285	0.99	0.53
ABE-JW 5yr band	2522	2496	0.99	0.58
EM: $p = 0.75$, $l = 0.60$	1067	1051	0.99	0.25
ML: even	4178	4054	0.97	0.95
II. Two-way intersections				
ABE-NYSIIS & ABE-exact	1995	1989	1.00	0.46
ABE-NYSIIS & JW 5-yr band	2009	2005	1.00	0.46
ABE-NYSIIS & EM	849	849	1.00	0.20
ABE-NYSIIS & ML	2847	2820	0.99	0.65
ABE-exact & ABE-JW	1779	1774	1.00	0.41
ABE-exact & EM	705	704	1.00	0.16
ABE-exact & ML	2282	2271	1.00	0.53
ABE-JW & EM	938	933	0.99	0.22
ABE-JW & ML	2496	2485	1.00	0.58
EM & ML	1052	1046	0.99	0.24
III. Three-way intersections				
ABE-NYSIIS & ABE-exact & ABE-JW	1621	1620	1.00	0.38
ABE-NYSIIS & ABE-exact & EM	660	660	1.00	0.15
ABE-NYSIIS & ABE-exact & ML	1987	1982	1.00	0.46
ABE-NYSIIS & ABE-JW & EM	780	780	1.00	0.18
ABE-NYSIIS & ABE-JW & ML	2005	2002	1.00	0.46
ABE-exact & EM & ML	848	848	1.00	0.20
ABE-exact & ABE-JW & EM	675	674	1.00	0.16
ABE-exact & ABE-JW & ML	1777	1773	1.00	0.41
ABE-exact & EM & ML	705	704	1.00	0.16
ABE-JW & EM & ML	935	932	1.00	0.22
IV. Four-way intersections				
ABE-NYSIIS & ABE-exact & ABE-JW & EM	633	633	1.00	0.15
ABE-NYSIIS & ABE-exact & ABE-JW & ML	1620	1619	1.00	0.37
ABE-NYSIIS & ABE-JW & EM & ML	780	780	1.00	0.18
ABE-NYSIIS & ABE-exact & EM & ML	660	660	1.00	0.15
ABE-exact & ABE-JW & EM & ML	675	674	1.00	0.16
V. Five-way intersection				
ABE-NYSIIS & ABE-exact & ABE-JW & EM & ML	633	633	1.00	0.15

Notes: This table shows the number of matches found between the 1915 Iowa records linked to the 1940 census using five alternative matching algorithms and their intersections. Each intersection sample keeps only pairs that were found using all methods listed, for instance the intersection between ABE-NYSIIS and ABE-exact is the sample of all pairs found both when using the ABE algorithm with NYSIIS names and with exact names. A match is considered correct if it coincides with the match that human hand-linkers found. PPV = # "correct / # matches. TPR = #"correct/#observations. These PPV and TPR values are plotted in Figure 6. All ABE methods require uniqueness within ± 2 years of age (5-year uniqueness band). All methods(other than ML) do not use middle name or middle initial in matching. The ML method uses middle initials and equal weighting on PPV and TPR.

Table A.4: Balance Test of Matched vs Unmatched Observations for ABE Algorithm, using weights

(a) UA-1900 census - weighted

		5 Years Band				No Band			
		Middle Initials		No Middle Initials		Middle Initials		No Middle Initials	
Variable	Mean	NYSIIS	Exact	NYSIIS	Exact	NYSIIS	Exact	NYSIIS	Exact
Year of Birth	1839.190	-0.004	0.497	-0.380	-0.006	0.134	0.741	-0.213	0.325
	(6.007)	(0.290)	(0.295)	(0.308)	(0.300)	(0.292)	(0.291)	(0.290)	(0.290)
Literate	0.938	-0.002	0.011	0.003	-0.000	-0.000	0.006	-0.006	-0.009
	(0.241)	(0.017)	(0.016)	(0.016)	(0.016)	(0.015)	(0.016)	(0.015)	(0.015)
Height (inches)	67.474	-0.048	0.109	-0.033	0.131	0.090	0.157	-0.030	0.122
	(2.316)	(0.125)	(0.127)	(0.126)	(0.125)	(0.119)	(0.123)	(0.118)	(0.119)
Occupation Score	35.345	-0.091	-0.503	0.148	0.040	-0.361	-0.642	0.491	-0.440
	(9.903)	(0.543)	(0.558)	(0.547)	(0.537)	(0.519)	(0.538)	(0.514)	(0.515)
Enlistment Age	22.746	-0.043	-0.471	0.345	0.040	-0.084	-0.606	0.136	-0.218
	(5.567)	(0.276)	(0.280)	(0.295)	(0.287)	(0.280)	(0.276)	(0.279)	(0.278)

(b) Iowa-1940 census - weighted

		5 Years Band				No Band			
		Middle Initials		No Middle Initials		Middle Initials		No Middle Initials	
Variable	Mean	NYSIIS	Exact	NYSIIS	Exact	NYSIIS	Exact	NYSIIS	Exact
Urban dummy	0.413	-0.025	0.008	-0.038	-0.002	-0.015	0.016	-0.026	0.021
	(0.492)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)	(0.017)	(0.016)
Age	9.636	0.014	0.075	-0.216	-0.159	-0.011	0.021	-0.248	-0.110
	(4.359)	(0.136)	(0.141)	(0.133)	(0.133)	(0.133)	(0.138)	(0.148)	(0.137)
US-born father dummy	0.777	0.000	-0.023	0.022	0.018	-0.014	-0.016	0.055	0.048
	(0.416)	(0.013)	(0.014)	(0.013)	(0.013)	(0.013)	(0.013)	(0.015)	(0.013)
US-born mother dummy	0.812	0.000	-0.021	0.015	0.011	-0.007	-0.018	0.042	0.033
	(0.391)	(0.012)	(0.013)	(0.012)	(0.012)	(0.012)	(0.013)	(0.014)	(0.013)
Years schooling	3.979	-0.009	0.054	-0.107	-0.078	-0.024	0.004	-0.156	-0.083
	(3.394)	(0.106)	(0.111)	(0.104)	(0.104)	(0.104)	(0.108)	(0.116)	(0.107)
Literacy	0.743	-0.000	-0.015	-0.009	-0.015	-0.005	-0.013	0.005	-0.000
	(0.437)	(0.013)	(0.014)	(0.013)	(0.013)	(0.013)	(0.014)	(0.015)	(0.014)

Notes: Panel (a): The first column is the population mean and standard deviations of the linked observations in the Union Army records. The table presents the balance test across the 8 different variations on the ABE algorithm. Matched observations are weighted to account for the distribution of the following variables in the Union Army records: year of birth, enlistment age, height, literacy, and occupational category. Panel (b) repeats the same exercise for the Iowa 1915 to 1940 census linking exercise. Matched observations are weighted to account for the distribution of the following variables in the Iowa 1915 records: year of birth, literacy, years of school, and foreign-born status of parents.

Table A.5: Comparison of Occupational Transition Matrices, US 1850-1880

	Son's occupation				
Father's occupation	White collar	Farmer	Skilled/semi-skilled	Unskilled	Row total
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)
ABE - Less conservative					
White-collar	0.44	0.23	0.23	0.10	1
	(22583)	(11650)	(11492)	(5078)	(50803)
Farmer	0.13	0.58	0.16	0.13	1
	(50840)	(219560)	(59571)	(48923)	(378894)
Skilled/semi-skilled	0.22	0.25	0.39	0.14	1
	(29726)	(33842)	(52706)	(18693)	(134967)
Unskilled	0.14	0.27	0.34	0.24	1
	(7965)	(15157)	(18854)	(13616)	(55592)
Column total	0.18	0.45	0.23	0.14	1
	(111114)	(280209)	(142623)	(86310)	(620256)
ABE - More conservative					
White-collar	0.50	0.21	0.21	0.08	1
	(14877)	(6247)	(6160)	(2466)	(29750)
Farmer	0.13	0.61	0.14	0.12	1
	(29802)	(139300)	(32990)	(27983)	(230075)
Skilled/semi-skilled	0.22	0.24	0.41	0.13	1
	(17100)	(18246)	(30952)	(9838)	(76136)
Unskilled	0.13	0.28	0.33	0.26	1
	(3784)	(7992)	(9551)	(7347)	(28674)
Column total	0.18	0.47	0.22	0.13	1
	(65563)	(171785)	(79653)	(47634)	(364635)
ABE w/ JW adjustment - Less conservative					
White-collar	0.45	0.23	0.22	0.10	1
	(24016)	(11941)	(11818)	(5018)	(52793)
Farmer	0.13	0.58	0.15	0.13	1
	(52648)	(230872)	(61072)	(50496)	(395088)
Skilled/semi-skilled	0.22	0.25	0.39	0.14	1
	(30515)	(34938)	(54595)	(19050)	(139098)
Unskilled	0.14	0.28	0.34	0.24	1
	(8017)	(15616)	(19109)	(13737)	(56479)
Column total	0.18	0.46	0.23	0.14	1
	(115196)	(293367)	(146594)	(88301)	(643458)
ABE w/ JW adjustment - More conservative					
White-collar	0.51	0.21	0.20	0.08	1
	(14937)	(6004)	(5868)	(2259)	(29068)
Farmer	0.13	0.61	0.14	0.12	1
	(29231)	(138232)	(31497)	(26992)	(225952)
Skilled/semi-skilled	0.23	0.24	0.41	0.13	1
	(16663)	(17487)	(29925)	(9161)	(73236)
Unskilled	0.13	0.28	0.33	0.25	1
	(3499)	(7580)	(8824)	(6770)	(26673)
Column total	0.18	0.48	0.21	0.13	1
	(64330)	(169303)	(76114)	(45182)	(354929)

Continued on next page

Table A.5 — *Continued from previous page*

Father's occupation	White collar	Farmer	Skilled/semi-skilled	Unskilled	Row total
EM - 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)
EM - 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)

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

Table A.6: Comparison of occupational transition matrices, Norway 1865-1900

	Son's occupation				
Father's occupation	White collar	Farmer	Skilled/semi-skilled	Unskilled	Row total
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)
ABE - Less conservative					
White-collar	0.73	0.05	0.15	0.07	1
	(1931)	(141)	(396)	(189)	(2657)
Farmer	0.08	0.59	0.16	0.18	1
	(1369)	(10629)	(2825)	(3310)	(18133)
Skilled/semi-skilled	0.24	0.07	0.53	0.16	1
	(1062)	(309)	(2376)	(709)	(4456)
Unskilled	0.08	0.23	0.32	0.37	1
	(984)	(2654)	(3703)	(4254)	(11595)
Column total	0.15	0.37	0.25	0.23	1
	(5346)	(13733)	(9300)	(8462)	(36841)
ABE - More conservative					
White-collar	0.76	0.05	0.13	0.06	1
	(1763)	(117)	(300)	(137)	(2317)
Farmer	0.08	0.60	0.15	0.17	1
	(1046)	(7974)	(1976)	(2266)	(13262)
Skilled/semi-skilled	0.27	0.07	0.52	0.15	1
	(870)	(213)	(1694)	(485)	(3262)
Unskilled	0.09	0.23	0.30	0.37	1
	(702)	(1874)	(2417)	(2986)	(7979)
Column total	0.16	0.38	0.24	0.22	1
	(4381)	(10178)	(6387)	(5874)	(26820)
ABE w/ JW adjustment - Less conservative					
White-collar	0.73	0.05	0.15	0.07	1
	(1937)	(143)	(387)	(178)	(2645)
Farmer	0.08	0.59	0.15	0.18	1
	(1400)	(10542)	(2747)	(3270)	(17959)
Skilled/semi-skilled	0.25	0.07	0.53	0.16	1
	(1094)	(292)	(2286)	(679)	(4351)
Unskilled	0.09	0.23	0.32	0.36	1
	(1006)	(2621)	(3604)	(4143)	(11374)
Column total	0.15	0.37	0.25	0.23	1
	(5437)	(13598)	(9024)	(8270)	(36329)
ABE w/ JW adjustment - More conservative					
White-collar	0.77	0.05	0.13	0.06	1
	(1778)	(118)	(294)	(134)	(2324)
Farmer	0.08	0.60	0.15	0.17	1
	(1056)	(7844)	(1917)	(2199)	(13016)
Skilled/semi-skilled	0.28	0.06	0.52	0.14	1
	(876)	(187)	(1629)	(449)	(3141)
Unskilled	0.09	0.24	0.30	0.37	1
	(708)	(1821)	(2315)	(2822)	(7666)
Column total	0.17	0.38	0.24	0.21	1
	(4418)	(9970)	(6155)	(5604)	(26147)

Continued on next page

Table A.6 — *Continued from previous page*

Father's occupation	White collar	Farmer	Skilled/semi-skilled	Unskilled	Row total
EM - 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)
EM - More 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)

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