Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project[†]

Kasey Buckles[‡], University of Notre Dame, IZA, & NBER

Adrian Haws, Cornell University

Joseph Price, Brigham Young University, IZA, & NBER

Haley Wilbert, University of Notre Dame

PRELIMINARY DRAFT - PLEASE DO NOT CITE WITHOUT PERMISSION

Abstract

The Census Tree is the largest-ever database of record links among the historical U.S. censuses, with over 700 million links for people living in the United States between 1850 and 1940. These links allow researchers to construct a longitudinal dataset that is highly representative of the population, and that includes women, Black Americans, and other under-represented populations at unprecedented rates. In this paper, we describe our process for creating the Census Tree, beginning with a collection of over 317 million links contributed by the users of a free online genealogy platform. We then use these links as training data for a machine learning algorithm to make tens of millions of new matches. Finally, we incorporate other recent efforts to link the historical U.S. censuses and introduce a procedure for filtering the links and adjudicating disagreements. Our complete Census Tree achieves match rates between adjacent censuses that are between 69 and 86% for men, and between 58 and 79% for women, with over 41.5 million links for Black Americans.

[†] This work is supported by grants from the Russell Sage Foundation (G-1063) and the National Science Foundation (SES-2049762). We are grateful to the excellent research assistants at the University of Notre Dame and the BYU Record Linking Lab for their work on this project, including Brianna Felegi, Nate McGhie, and Tommy Morgan.

[‡] Corresponding author: Kasey Buckles, kbuckles@nd.edu

I. Introduction

Record linking, or the process of combining a subject's information from multiple datasets, is often a key component of empirical work in history, medicine, and the social sciences. These links allow the researcher to observe a person over time, to study relationships among variables that are not available in a single data source, and to identify connections between people in families and communities. Recent advances in record linking have been facilitated by growing access to restricted-use data that include stable and unique personal identifiers (e.g. social security numbers, registry numbers, or exact birth dates) that can be used to determine that two records correspond to the same person (Chetty et al. 2014; Chetty, Hendren, and Katz 2016; Mazumder 2005; Black, Devereux, and Salvanes 2005; Kleven, Landais, and Søgaard 2019) Unfortunately, many datasets that researchers would like to link—including many historical or publicly available sources—do not include these identifiers. In this situation, researchers must try to find unique matches using relatively stable characteristics like names, birth years, and birth places. These requirements have frequently resulted in unrepresentative samples; for example, women have been omitted entirely from several notable linking efforts because their surnames typically change when they marry (e.g. Abramitzky et al. 2022, Fogel and Wimmer 1992, Feigenbaum 2018).

In the Census Tree project, we use information provided by members of the largest genealogy research community in the world to create hundreds of millions of new links among the historical U.S. Censuses (1850-1940). The users of the platform link data sources—including decennial census records—to the profiles of deceased people as part of their own family history research. In doing so, they rely on private information like maiden names, family members' names, and geographic moves to make links that a researcher would never be able to make using the observable information. To date, users have created over 317 million census-to-census pairs, where nearly half of these are for women.

We describe our process for adding to these links using a machine learning model that employs the user-made links as training data. We also add pairs identified by other recent linking methods and develop a process to verify the quality of the matches and to adjudicate disagreements between methods. The result is the publicly-available Census Tree dataset, which contains over 700 million links among the 1850-1940 censuses. The data include an unprecedented number of links for women (314 million) and Black Americans (41.5 million).[1] We show that the Census Tree links are high quality and yield samples that are highly representative of the population, and discuss the potential for their use in research.

II. Genealogy Research on FamilySearch

A. The Platform

Founded in 1894, FamilySearch is "a nonprofit family history organization dedicated to connecting families across generations" (FamilySearch 2023a). Sponsored by the Church of Jesus Christ of Latter-day Saints, FamilySearch introduced a free website featuring family history tools and digitized records in 1999. It has since become one of the most widely-used genealogy websites in the world, with over 400,000 visitors per day in 2020 and visitors coming from 238 countries. The website also includes over two billion indexed historical records and over one billion unique individual profiles for deceased persons (FamilySearch 2023b).

FamilySearch.org has several features that contribute to its popularity among the genealogy community, including its sophisticated search tools, its enormous set of digitized and indexed historical records, and the fact that it does not charge any fees. But perhaps its most distinctive feature is that, rather than each user building their own tree, all users contribute to a single, interconnected family tree. The tree operates as a "wiki," in which users can edit and build on the contributions of others. As a result, the FamilySearch users have collaborated to produce an incredibly comprehensive and accurate population-wide family tree.

Critically for our purposes, users can attach digitized historical records to the profiles of people on the tree, including the decennial U.S. censuses from 1850 to 1940.[2] In cases where records in two different decennial censuses are linked to the same profile, this creates a user-made link that identifies the records as describing the same person. Thus, the process of record linking is "crowd-sourced" to millions of users with private information that helps them make links—including some

[1] All of the data described in this paper are available at censustree.org, along with the code and training data for the machine learning methods and the code for creating the full Census Tree.

[2] The Census Bureau releases the full-count censuses to the public after 72 years. The 1950 census was released in April of 2022 and is in the process of being digitized and indexed. The 1890 census is not included in the set of historical decennial censuses, as the majority of the records for that year were destroyed in a fire in 1921.

information that would be unavailable to trained research assistants or machine learning algorithms. For example, family members often know their female ancestors' maiden names, which allows them to link women between childhood and adulthood in a way that has not been possible using traditional linking methods that rely on a name match. Users may also know details that make it possible for them to solve the problem of common names—they may know the names of other family members within the same household that allow them to correctly identify which "John Smith" is the right one among many choices. This information can also help them to confirm that two records are a match, even if the digitized spelling of the name is different or if other information is not an exact match.[3]

B. User-Made Links: The Family Tree

The set of user-made links between censuses constitutes a dataset that we call the "Family Tree." The Family Tree dataset alone contains over 317 million unique links among the 1850-1940 censuses, with nearly half of the links being for women.[4] In Figure 1, we compare the match rates attained between adjacent censuses in four different sets of links, by gender. The match rate is calculated as the number of people for whom a link is made to the previous census, divided by the number of people who are old enough to have been alive at the time of the previous census, with an adjustment for rates of immigration and under-enumeration.[5]

In addition to the Family Tree, Figure 1 shows match rates for our complete Census Tree and for the two other largest sets of publicly-available links among the historical U.S. censuses—the Census Linking Project (CLP) (Abramitzky et al. 2022) and the IPUMS Multigenerational Longitudinal Panel (MLP) (Helgertz et al. 2023).[6] We describe these other datasets in more detail in the next sections, but here we note two facts about the match rates obtained in the Family Tree. First, the Family Tree contains between 24 and 48% of the possible matches between adjacent

[3] Appendix Figure 1 shows the sources linked to "Delilah A. 'Minnie' Jenkins," who appears in the digitized censuses as Delila Jenkins (1870), Deliah M Jinkins (1880), Minnie Sharone (1900), Minnie Shearom (1910), and Minnie Sherman (1920). The consistent presence of other family members across these records helps to confirm that they do indeed reference the same person.

[4] We apply a simple process to de-duplicate user-made links, where we remove any possible links which have a conflict with another possible link.

[5] See Price et al. (2021) for a detailed description of how these match rates are calculated.

[6] In Figure 1 we use the Exact-Conservative matches from the CLP. We choose this method when comparing match rates because its standards for a match are closest to those of the MLP and the Family Tree.

censuses for men; these match rates are comparable to those obtained by previous efforts to link records using unsupervised (CLP) and machine learning (MLP) methods[7]. Second, the match rates for women in the Family Tree are nearly as high as those for men. As a result, the Family Tree contains far more links for women than the MLP (and the CLP does not attempt to link women).

How reliable are the Family Tree links, given that they are crowdsourced and not directly validated? To investigate this, we conduct an exercise in which trained research assistants hand-check a random sample of 760 of the 1900-1910 links from the full Census Tree—440 of which appear on the Family Tree. We asked the assistants to use the full set of information available in each census record to classify the link as correct, incorrect, or unsure. Among the Family Tree links, 98% were determined to be correct—an exceptionally high number that is consistent with a similar check conducted on a different sample in Price et al. (2021).[8]

One potential limitation of the Family Tree data is that the users may be a selected group. Among other possible factors, they have a demonstrated interest in family history, and are able to access and use the internet. We explore this in Section IV, where we compare the observable characteristics of people who can be linked in the four datasets in Figure 1 to the full census population.

III. Creating the Census Tree Dataset

Figure 2 illustrates the process we use to create the Census Tree dataset. We first add links made using our machine learning process, where we use Family Tree links to inform decisions and as training data. We then include links obtained from other recent linking efforts and develop a process for filtering low-quality links and adjudicating disagreements. We elaborate on these steps in the following subsections.

A. Machine Learning Using Training Data from the Family Tree

[7] Unsupervised methods can be automated but do not require training data.

[8] There is also external evidence that the user-provided information is high quality. Using data from a similar genealogy platform, Kaplanis et al. (2018) compare DNA data to information provided by the site's users, and conclude "that millions of genealogists can collaborate in order to produce high quality population-scale family trees" (p. 172). Furthermore, the creators of other linked datasets have used the Family Tree as a benchmark for measuring the quality of their own matches (Bailey et al. 2020), referring to genealogy data as the "gold standard" (Abramitzky et al. 2021, 868).

1. *Pre-processing and blocking*

We begin by preparing the data to be linked by the machine learning process, drawing on information provided by the user-made links. We standardize the names of places (states and countries) to correct misspellings and abbreviations. For names, we convert nicknames to a standard set of formal names, using a list of the most common nickname-name pairs observed in the Family Tree.

The computational costs of our machine learning process also require that we limit the set of potential matches by grouping the data into blocks based on features like name, birthplace, and birth year. A challenge when choosing the features to create the blocks is that the most stable features, like race, sex, or birth state, are not very unique. Requiring that the potential matches also have, for example, the same birth year, might exclude many true matches. We are able to test several blocking strategies to see how they perform when trying to recreate the links in the Family Tree data (see Price et al. 2021). Table S2 in the Supplemental Materials identifies the variables that we use in our blocking strategy.

2. *Training Data*

We use millions of the user-made links from the Family Tree to train our machine learning models. We first remove any non-unique links across census years to avoid incorrect links. Then we use the "true" links to create a set of "false" links that satisfy the same blocking criteria but are not the same as the "true" link. For each of the 36 year-to-year pairs, we train the model using training data from those specific years; see Appendix Table S3 for the size of the training data for each pair of years. Because increasing the size of training data has been found to improve the accuracy and number of record links (Feigenbaum, 2016; Gross and Mueller-Smith 2021), we use a large set of the available "true" and "false" links. This also ensures that we have sufficient support in the data for training the algorithm to make matches for under-represented groups. We have over 2,000 observations for women in the training data in all but one year, and at least 800 observations for the Black sample for all pairs 50 years apart or less.

Each census record contains basic information about the person's name, birth year, residences, demographic characteristics, household relationships, and occupation. To prepare the training data, we convert these variables into "features" that capture the rich amount of information available.[9] For example, when comparing the birth year between two records, we create four features: a binary variable indicating that the absolute difference between them is less than or equal to 3, a variable that is equal to the absolute difference in birth years, an indicator that the sign of the birth year difference is positive, and a measure of the age in the earlier census. Table S2 in the supplemental materials shows the full list of 70 features created across the nine censuses, and the years that the feature is available. [10]

3. *Tuning the Model and Filtering Predictions*

The supervised machine learning algorithm, XGBoost, uses gradient-boosted decision trees to assign a score to each potential link.[11] This score, between zero and one, is similar to a predicted probability of a link being "true" that could be calculated using a logistic regression.[12] We use a cross-validation process with the training data to select the values of our model parameters—maximum tree depth and number of estimators—to optimize the model's performance. For each set of census years, we randomly select 2/3 of 500,000 training pairs to train a model and use the remaining 1/3 to test the out-of-sample performance. The model with the highest F1 score (balancing precision and recall) is then used with the full set of training data to produce the final model. We provide the trained models for all 36 year-to-year pairs at censustree.org.[13]

As a way of getting "under the hood" of the machine learning algorithm, Table 1 lists the fifteen most important features used in the process of linking the 1900 and 1910 censuses, after the core set of features used for blocking (see Table S6 in the appendix for feature importance for other adjacent year pairs). The importance measure is calculated as the average increase in accuracy across nodes of the decision tree which use the feature.[14] The most important individual feature is the distance in miles between the two towns. This illustrates the value of the machine learning approach—using a traditional blocking and matching procedure, one would not want to require that

[9] Names are not available in the publicly released versions of the IPUMS census files, but users can apply for the restricted-use versions of the data that include them. We obtain names from versions of the censuses provided to us by FamilySearch.

[10] This extensive set of features benefits from indexed census variables provided by Ruggles et al. (2021) as well as geographic coordinates from the Census Place Project (Berkes, Karger, and Nencka 2023).

[11] We use the XGBClassifier package within the xgboost library in Python.

[12] The highly flexible XGBoost algorithm out-performs logit (see Price et al. 2021).

[13] The website also includes the full training data set for 1900-1910.

[14] This is the "gain" method of feature importance calculated by the XGBoost algorithm.

two records be from the same (or nearby) towns, as people frequently move. However, if the person *is* in a similar place in the two censuses, that increases the probability that the records are a match. [15] Most of the other important features are variations on the characteristics most commonly used in blocking—birth year, name, and birth place. In Appendix Table S6, we rank the importance of the feature categories for all 8 adjacent census pairs. Features that relate to the person's name are most important, followed by residence and birth year. The importance of names is not as apparent in the individual feature importance ranking because names are used in blocking and we use a total of 33 name-based features.

The machine learning algorithm generates a match score for each combination of potential matches within the blocking cell. We identify a pair of records as a match if it satisfies three conditions. First, it should have the highest match score among possible links. Consider a record "A" in 1900 which has potential links to both "B" and "C" in 1910. We retain only the link to "B" if this has a higher match score than "C". In practice, we also keep the link to "C" if its match score with "A" is the highest among all of its potential links to 1900. Second, a possible link should have the highest sheet count, where the sheet count is the total number of individual links between the census pages which contain the records.[16] If record "A" and four additional records are linked to the sheet containing record "B" then the A-to-B has a sheet count of five. In this step, we require potential links to have the highest sheet count for potential links in 1910 with A and the highest sheet count for potential links in 1900 with B. Third, there must be no remaining conflicts between the two years. We tested this method using a "truth set" from the Family Tree and determined that over 98% of true links satisfy these conditions. We additionally remove a small set of links for women with consistent surnames but who transitioned from single to married between the census years we attempt to link.[17] This represents only 0.9% for 1900-1910 women links because these cases are penalized by the machine learning model.

[15] See Folkman, Furner, and Pearson Pearson (2018) and Price et al. (2021) for a more in-depth discussion of this issue, and for a demonstration of the effects of excluding geographic information from the set of features.

[16] We calculate sheet counts using the set of potential links with a match score above 0.1. While many of these potential links are later removed from the sample, this match score criterion removes 92.5% of the blocked pairs between 1900 and 1910. The occurrence of multiple links between a set of sheets could almost never occur by random chance, as there are 40 million potential links remaining and about 2.8 trillion possible combinations of census sheets between 1900 and 1910.

[17] Because marital status is not available for the 1850 through 1870 censuses, we remove links for women who are married to the household head in the later census but have a different household

B. Additional Sources for Links

1. Census Linking Project (CLP)

The CLP was the first effort to fully link the 1850-1940 decennial U.S. censuses and to make the links publicly available to the research community. These links are based on traditional, unsupervised blocking and matching strategies that rely on names, birth dates, and birth places; see Abramitzky et al. (2021) for a detailed description of their process. The CLP data contain multiple sets of links, which use slightly different features and more or less conservative rules to identify matches. We use the NYSIIS Standard links, which use the New York State Identification and Intelligence System Phonetic Code to standardize names based on their pronunciation and require that the names be unique within the birth year. We choose this set because it has a high match rate, allowing us to include more links; we discuss this choice further below.

2. Multigenerational Longitudinal Panel (MLP)

Helgertz et al. (2021) introduce an innovative two-step approach, in which they first use a machine learning approach to obtain high-quality matches for men, and then link together other individuals in the same households of the two men that were linked.[18] This approach allows them to match women as well as men. The MLP data are only available for adjacent censuses.

3. FamilySearch Hints

FamilySearch has a proprietary machine-learning algorithm for identifying possible record links. They have provided us with two sets of these "hints" for U.S. census records. The first type of hints, which we call "profile hints", suggest to users that a census record might belong to a profile in their family tree. When census records from two different years are both "hinted" to the same profile, this creates a possible census link. The second type, "direct hints", identifies a possible link directly between two census records. We have developed several tools that allow volunteers to validate both types of hints by attaching records to profiles on the Family Tree. In this way, these hints help to expand the set of user-made links on the Family Tree. FamilySearch hints include many links for

relationship in the earlier census. This alternative strategy removes 4.6% of 1870-1880 links for women.

[18] The MLP household-based strategy is similar to the dyad and household matching methods that were part of the process described in Price et al. (2021). Because the MLP data contain nearly all of the additional links generated by these methods, we do not implement them here.

women, which is made possible by the large corpus of digitized records on the website (including marriage records) and by personal information available on person profiles (including dates of marriage and spouse's surnames). While we do not have access to FamilySearch's machine learning models, the methods employed by genealogy companies can be quite rich (Folkman, Furner, and Pearson 2018). We use match scores provided by FamilySearch to apply the same three-step filtering process described for our machine learning model. As Appendix Table S4 shows, there are 26.5 million FamilySearch "direct hints" that are part of the 1900-1910 Census Tree links, of which 0.5 million are not also found by one of the other methods in the full linking process. A similar number of "profile hints" are used in our links.

C. Preparing the Data

1. *Filtering and Adjudication*

We combine unique links from the Family Tree, our machine learning process informed by the Family Tree, the Census Linking Project, the Multigenerational Longitudinal Panel, FamilySearch profile hints, and FamilySearch direct hints. Because these various links may disagree, we filter them using the same sheet checking procedure described in step two of filtering the machine learning links. In this case, we calculate sheet counts using links from all six methods (without double-counting the same link from multiple methods), keep potential links which have the highest sheet count for each year, and drop any links with remaining conflicts.

2. *Creating Implied Links*

This step takes advantage of the fact that if records from two different censuses are linked to a record in a third census, the original two should also be a match. For example, if a link has been established between a person's 1900 and 1910 census records, and the 1910 record is linked to a 1920 census record, we can also link the 1900 record to 1920. This step is especially helpful in expanding the set of links made by the MLP, which uses an innovative household-based matching strategy but only includes links from adjacent censuses. Identical to the adjudication process after combining the previous links, implied links are filtered by keeping potential links which have the highest sheet count for each year and dropping any links with remaining conflicts. We also remove links with an absolute birth year difference greater than three years. Even if birth years are similar for two links, these differences could become greater when creating an implied link. For example, a 3-year difference between 1900 and 1910 and between 1910 and 1920 can result in a 6-year difference for the implied 1900-1920 link. In these cases, it is likely that one of the underlying links is incorrect.

3. Creating the Crosswalks

After creating the implied links, we conduct one final round of sheet-checking and drop remaining conflicts. We also add flags to identify the linking method(s) used to create each link; as we discuss below, the link source flags should be helpful in the event that a researcher wishes to exclude links made by a particular method. Appendix Table S4 shows the number of total and unique links provided by each of the different linking methods. In the next section, we compare these datasets along three key dimensions: their size, their quality, and their representativeness.

IV. Results

A. Match Rates (Recall)

We return to Figure 1, which compares the match rates for the CLP, MLP, Family Tree, and Census Tree for adjacent censuses. Starting with the rates for men in Figure 1A, we see that the Census Tree obtains match rates between 69% and 76% for the 19th century censuses, and between 82% and 86% for the 20th century. These exceptionally high rates represent a large increase over existing linking methods. The Census Tree has five to six times as many links for men as the CLP (Exact-Conservative, or EC).[19] Comparing to the MLP, the Census Tree has between 41 and 80% more matches. Crucially, the MLP does not attempt to link non-adjacent censuses directly, so the Census Tree is an even more significant advance for those pairs. Finally, Figure 1A shows the gain that is made by our procedures for adding to the Family Tree. The Census Tree dataset is 1.7 to 3 times larger than the Family Tree for these adjacent census pairs.[20]

Match rates for women are in Figure 1B. The CLP has match rates of 0% for all years, as they do not attempt to link women. The MLP does, with rates between 32% and 46% for their adjacent-census pairs. The Census Tree's match rates are 1.6 to 1.9 times higher, and range from 58% to

[19] Match rates are higher for other sets of the CLP links (reaching 30-40%); we use the EC matches here because they are closest to the other datasets in the figure in terms of their quality.

[20] The Family Tree has the highest match rates for 1900-1910 and 1910-1920 because the Record Linking Lab at BYU has focused their initial efforts to expand the Family Tree on the 1910 census. The Lab's goal has been to ensure that every person in the 1910 census has a profile on the Family Tree, and as of July 2023, the coverage rate had reached 80%.

79%, with all four 20th century pairs obtaining match rates above 70%. As with men, the Census Tree process adds millions of observations to those in the Family Tree, increasing the match rates by 50 to 300%. We note that the gain in going from the Family Tree to the Census Tree is slightly smaller for women than it is for men. This is because users link their female and male ancestors at very similar rates, but our XGBoost algorithm is not able to "learn" to make matches for women in cases where the surname changes due to marriage.

We include match rates for all 36 census-to-census pairs in Table 2. Here, we make an adjustment to how we calculate the match rates, as our method of adjusting for immigration does not perform well for censuses that are further apart in time.[21] Even with this adjustment, the match rates for men are above 56% for all census-to-census pairs. As expected, the match rates are generally higher for more recent censuses. It is the case that the match rates are actually *above* 100% for pairs that are 80 or 90 years apart; this appears to be due to likely errors in the denominator (e.g. unreliable ages for those who are very old). The match rates for women show similar patterns, with rates of 44% or above for all pairs, and again reaching 70% or above in the 20th century.

Appendix Table S7 translates these match rates into the number of links between each of the 36 census-to-census pairs. These numbers show the unprecedented size of the Census Tree dataset, with over 391 million links for men and 314 million links for women. While the calculation of the match rates is sensitive to choices about how the denominator is constructed, the absolute number of links is not. Accordingly, the table also shows that the size of the crosswalks predictably declines as the length of time between the two censuses grows.

B. Quality (Precision)

While it is clear that the Census Tree is an advance in terms of the number of links made, what can we say about whether the links are likely to be "true" matches? As we described in Section II.B., we randomly selected 760 of the 1900-1910 Census Tree links and asked research assistants to use the full set of information available in each census record to classify the link as correct, incorrect, or unsure. Appendix Table S1 shows the fraction of each links that were determined to be correct, for

[21] As described in Price et al. (2021), our main match rate calculation adds the total number of legal immigrants in the U.S. between the two years, and subtracts this total from the denominator for our main match rate calculations. When the censuses are further apart, this will cause the denominator to be much too small, as many of those who immigrated between the two endpoints will not have survived to the latter year. Ideally we would use information on the year of immigration from the latter census to adjust the denominators, but this information is only available from 1900-1930.

the full Census Tree and for the links identified by each link source. This fraction—known as precision—depends on the treatment of the "unsures," and so we present results with different treatments that constitute upper and lower bounds.

Overall, between 89% and 94% of the links in the full Census Tree were determined to be correct, depending on whether the unsure links are treated as incorrect or correct or dropped altogether. When we look at the source of the links, we see that the implied links and the Family Tree links are least and most precise, respectively. The supervised methods (XGBoost, MLP, FS Hints) have very similar precision, and perform better than the unsupervised method (CLP). Note that each individual method has a higher rate of precision than the full Census Tree, because the calculations include links that are only identified by that method *and* those that are also identified by others.[22]

In Appendix Table S1 we also compare precision for links that are identified by one or more sources. When a link is only identified by one source, it is determined to be correct between 68% and 81% of the time. However, links that have two sources are much more precise (86% to 94%). Links that have at least four sources have precision rates of 94% or above—reaching 98% for those with six or seven sources.

The results in Appendix Table S1 highlight the well-known tradeoff between recall and precision when linking records (Abramitzky et al. 2021). The Census Tree constitutes a major advance in what is possible in terms of match rates, while maintaining high rates of precision. However, in some cases, researchers may prefer to have higher confidence in the matches, even if it means reducing their sample size. For this reason, the Census Tree crosswalks include flags that indicate the sources of the match. With these flags, the researcher could omit links from methods that they believe to be lower quality (e.g. the implied links), or that come from sources that use a less transparent linking process (e.g. the FamilySearch hints). Alternatively, one could restrict the sample to those links that are identified by at least two sources. For the 1900-1910 sample, this choice would increase precision significantly while decreasing the sample size by about 17% (but still leaving 39.4 million links).

[22] To see this, suppose that there are two methods (A and B) and six links. Two of the links are identified by A only, one of which is correct. Two are identified by B only, and again one is correct. There are two links identified by both A and B, and both are correct. Precision would be 0.75 (3/4) for each method, while precision for the entire set would be 0.67 (4/6).

Thus, the publicly available Census Tree crosswalks allow the researcher to choose their desired point along the recall/precision frontier.

C. Representativeness

Another desirable property of any dataset is that it be representative of the population it is meant to describe. This has been a challenge for those attempting to create linked datasets, as some people may be easier to link, leading to selected samples. The most serious issue has been the difficulty in linking women due to surname changes, which has led to their complete omission from some scholarship on the historical U.S. (Collins and Wanamaker 2022; Feigenbaum 2018). Other populations that have been difficult to link include those with common names, those whose names are less stable (e.g. immigrants), or those who are more likely to have been left out by the enumeration process (e.g. the enslaved or formerly enslaved).

To assess the representativeness of the Census Tree and its alternatives, we compare the observable characteristics of those linked between 1900 and 1910 by each method to the full population of those who are observed in the 1910 census. From the latter, we omit those who are under age 11, as those children would not have been born in 1900. The results are in Table 3 (see Appendix Table S8 for comparisons between the Census Tree and the population for other adjacent year pairs). As expected, the Census Tree has nearly the same fraction of women as the population (0.47 vs. 0.48), compared to 0.43 for the MLP and zero for the CLP. As with previous efforts, Black Americans are under-represented in the Census Tree, but our additional steps help improve upon the under-representation of this population on the Family Tree. Furthermore, the large sample size means in the Census Tree means that there are still 3.39 million links for Black Americans between 1900 and 1910.

Those linked in the Census Tree are very similar to the full population in terms of their marital status and family structure. There is some evidence that those on the Census Tree are positively selected by socioeconomic status—they are slightly more literate and more likely to speak English. They are also more likely to live in their birth state. On all of these dimensions, the Census Tree does at least as well at matching the population as the CLP, the MLP, or the Family Tree alone.

Critically, the summary statistics in Table 3 and Appendix Table S8 are unweighted. Bailey, Cole, and Massey (2020) propose a method for weighting linked data to match population characteristics and obtain representative samples. Buckles et al. (2023) apply their method and show that, once weighted, estimates of the intergenerational transmission of socioeconomic status are nearly identical when using links from either the CLP or the Census Tree, despite the fact the two datasets have different sample sizes and observable characteristics. Moreover, the Census Tree has such large samples that the reweighting procedure is likely to have sufficient support in the data for reweighting in cases where the study population is smaller (e.g. a single state or immigrant group).

To summarize, there is little evidence that the Census Tree dataset is a highly selected sample—as we would expect, given that each year-to-year pair has at least 60% of the linkable population. Where some non-representativeness remains, the dataset is large and complete enough to support re-weighting to produce results that match the population characteristics. The Census Tree also includes millions of observations for groups that have been omitted or under-represented in prior research, including women and the formerly enslaved and their descendants.

V. Discussion

The Census Tree is a resource that will allow researchers to link people across the historical United States censuses at an unprecedented scale. Scholars will be able to create longitudinal datasets that follow individuals over time, and to connect people to their families and communities. In this paper, we have described our process for creating this resource, beginning with links provided by the users of an online genealogy platform, and adding to them using machine learning and the contributions of previous linking efforts. The finished dataset contains over 700 million links, including 314 million links for women and 41 million links for Black Americans. The Census Tree is flexible enough to accommodate different preferences regarding the tradeoff between recall and precision, and it is large enough to support reweighting and work on small populations.

The Census Tree project also demonstrates the tremendous potential for using crowd-sourced genealogical research in academic work. There is a growing interest in genealogy throughout the world and several companies provide vast record collections and sophisticated search tools that allow people to do high quality research. While the focus of this paper has been on US census records, our approach could be used to link records within or across other countries, or other vital or administrative records (death certificates, enlistment records, marriage licenses). These links could be used to create rich datasets that facilitate work on topics including family formation, migration and immigration, and the determinants of health and well-being.

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Pérez. 2021. "Automated Linking of Historical Data." *Journal of Economic Literature* 59 (3): 865–918. https://doi.org/10.1257/jel.20201599.

Abramitzky, Ran; Boustan, Leah; Eriksson, Katherine; Rashid, Myera; Pérez, Santiago. 2022. "Census Linking Project: 1850-1940." Harvard Dataverse. https://doi.org/10.7910/DVN/XUXYSR.

Bailey, Martha, Connor Cole, and Catherine Massey. 2020. "Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850–1930 IPUMS Linked Representative Historical Samples." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53 (2): 80–93. https://doi.org/10.1080/01615440.2019.1630343.

Bailey, Martha J., Connor Cole, Morgan Henderson, and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from US Historical Data." *Journal of Economic Literature* 58 (4): 997–1044. https://doi.org/10.1257/jel.20191526.

Berkes, Enrico, Ezra Karger, and Peter Nencka. 2023-03-28. The Census Place Project: A Method for Geolocating Unstructured Place Names. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E179401V2.

Black, Sandra E, Paul J Devereux, and Kjell G Salvanes. 2005. "Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital." *American Economic Review* 95 (1): 437–49. https://doi.org/10.1257/0002828053828635.

Buckles, Kasey, Joseph Price, Zach Ward, and Haley Wilbert. 2023. "Family Trees and Falling Apples: Intergenerational Mobility Estimates from U.S. Genealogy Data." Working paper.\

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review* 106 (4): 855–902. https://doi.org/10.1257/aer.20150572.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. "Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States*." *The Quarterly Journal of Economics* 129 (4): 1553–1623. https://doi.org/10.1093/qje/qju022.

Collins, William J., and Marianne H. Wanamaker. 2022. "African American Intergenerational Economic Mobility since 1880." *American Economic Journal: Applied Economics* 14 (3): 84–117. https://doi.org/10.1257/app.20170656.

Feigenbaum, James J. 2018. "Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940." *The Economic Journal* 128 (612): F446–81. https://doi.org/10.1111/ecoj.12525.

Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard. 2019. "Children and Gender Inequality: Evidence from Denmark." *American Economic Journal: Applied Economics* 11 (4): 181–209. https://doi.org/10.1257/app.20180010.

Mazumder, Bhashkar. 2005. "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data." *The Review of Economics and Statistics* 87 (2): 235–55. https://doi.org/10.1162/0034653053970249.

FamilySearch. 2023. "The Largest Free Family History Resource." FamilySearch. 2023. https://www.familysearch.org/en/about/.

———. 2023. "Who Are Your Ancestors?" 2023. https://ancestors.familysearch.org/en/.

Feigenbaum, James J. 2016. "A Machine Learning Approach to Census Record Linking," 34.

Feigenbaum, James J. "Multiple measures of historical intergenerational mobility: Iowa 1915 to 1940." *The Economic Journal* 128, no. 612 (2018): F446-F481.

Fogel, Robert W., and Larry T. Wimmer. "Early indicators of later work levels, disease, and death." NBER Working Paper #h0038. (1992).

Folkman, Tyler, Rey Furner, and Drew Pearson. 2018. "GenERes: A Genealogical Entity Resolution System." In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 495–501. https://doi.org/10.1109/ICDMW.2018.00079.

Gross, Matthew, and Michael Mueller-Smith. 2021. "Modernizing Person-Level Entity Resolution with Biometrically Linked Records."

Helgertz, Jonas, Steven Ruggles, John Robert Warren, Catherine A. Fitch, David Hacker, Matt Nelson, Joseph Price, Evan Roberts, and Matthew Sobek. 2023. "IPUMS Multigenerational Longitudianal Panel: Version 1.1 [Dataset]." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D016.V1.1.

Kaplanis, Joanna, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, et al. 2018. "Quantitative Analysis of Population-Scale Family Trees with Millions of Relatives." *Science* 360 (6385): 171–75. https://doi.org/10.1126/science.aam9309.

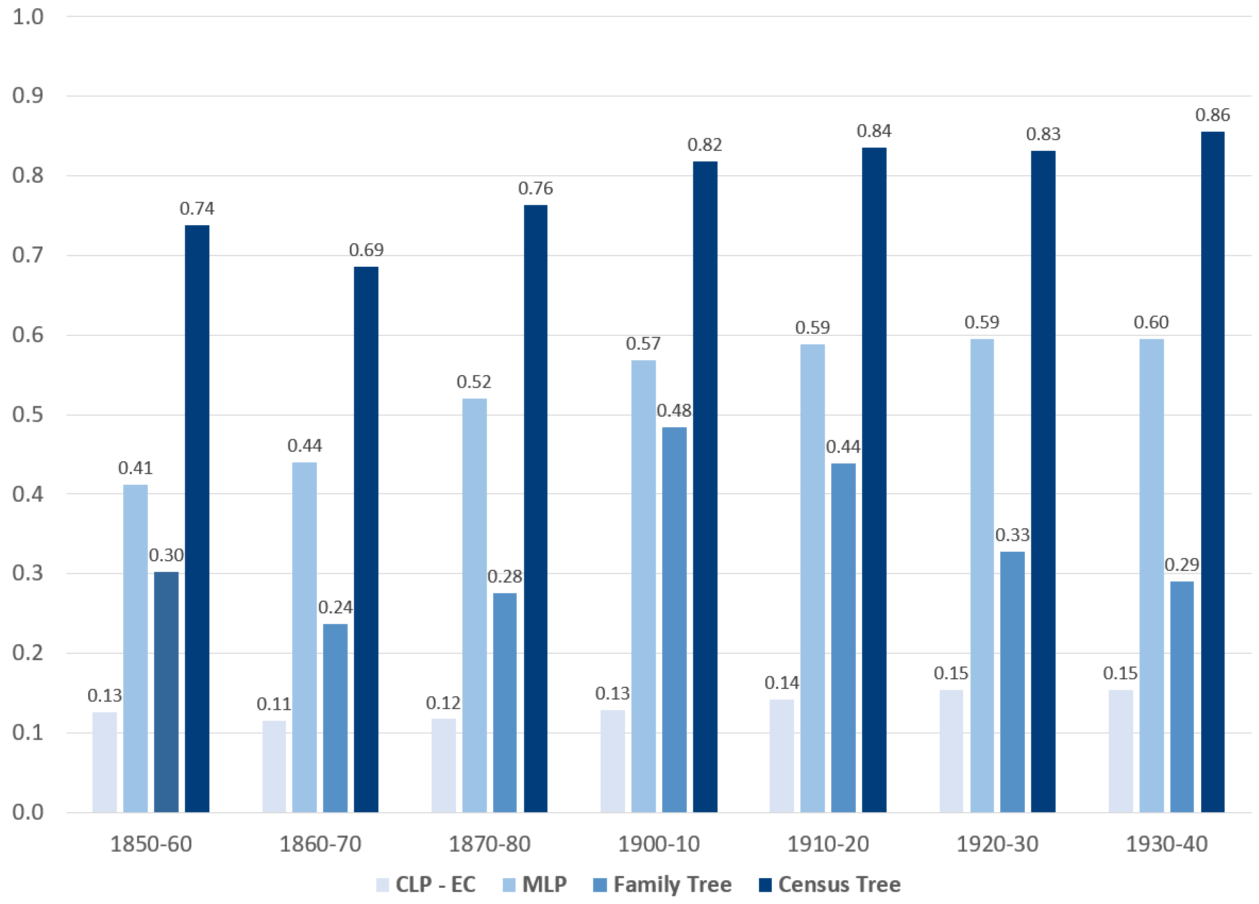
Kleven, Henrik, Camille Landais, and Jakob Egholt Søgaard. 2019. "Children and Gender Inequality: Evidence from Denmark." *American Economic Journal: Applied Economics* 11 (4): 181–209. https://doi.org/10.1257/app.20180010.

Mazumder, Bhashkar. 2005. “Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data.” *The Review of Economics and Statistics* 87 (2): 235–55. https://doi.org/10.1162/0034653053970249.

Price, Joseph, Kasey Buckles, Jacob Van Leeuwen, and Isaac Riley. 2021. “Combining Family History and Machine Learning to Link Historical Records: The Census Tree Data Set.” *Explorations in Economic History* 80 (April): 101391. https://doi.org/10.1016/j.eeh.2021.101391.

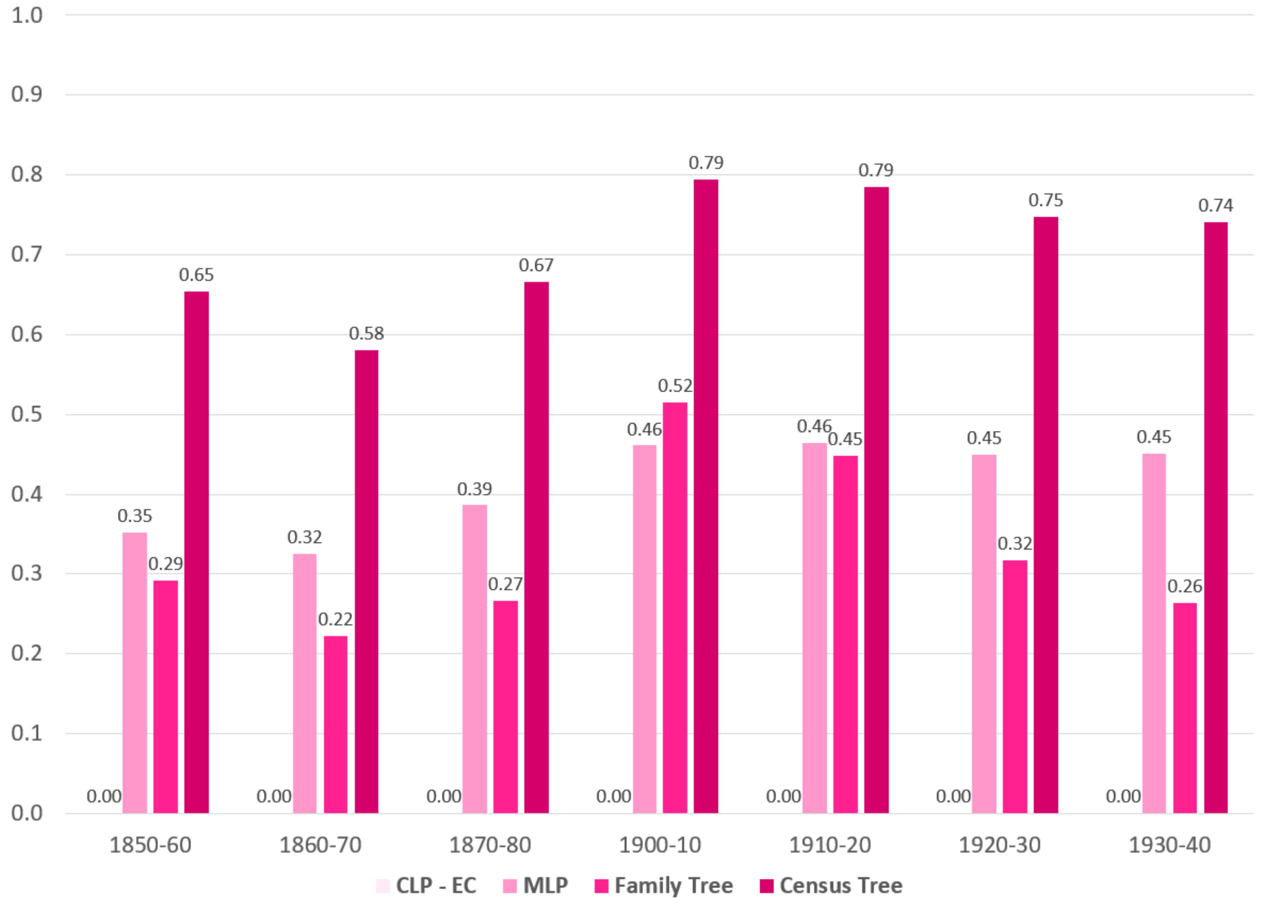
Steven Ruggles, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Evan Roberts, Megan Schouweiler, and Matthew Sobek. IPUMS Ancestry Full Count Data: Version 3.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D014.V3.0

Figure 1A: Match Rates for Men Using Various Linking Methods, for Censuses Ten Years Apart



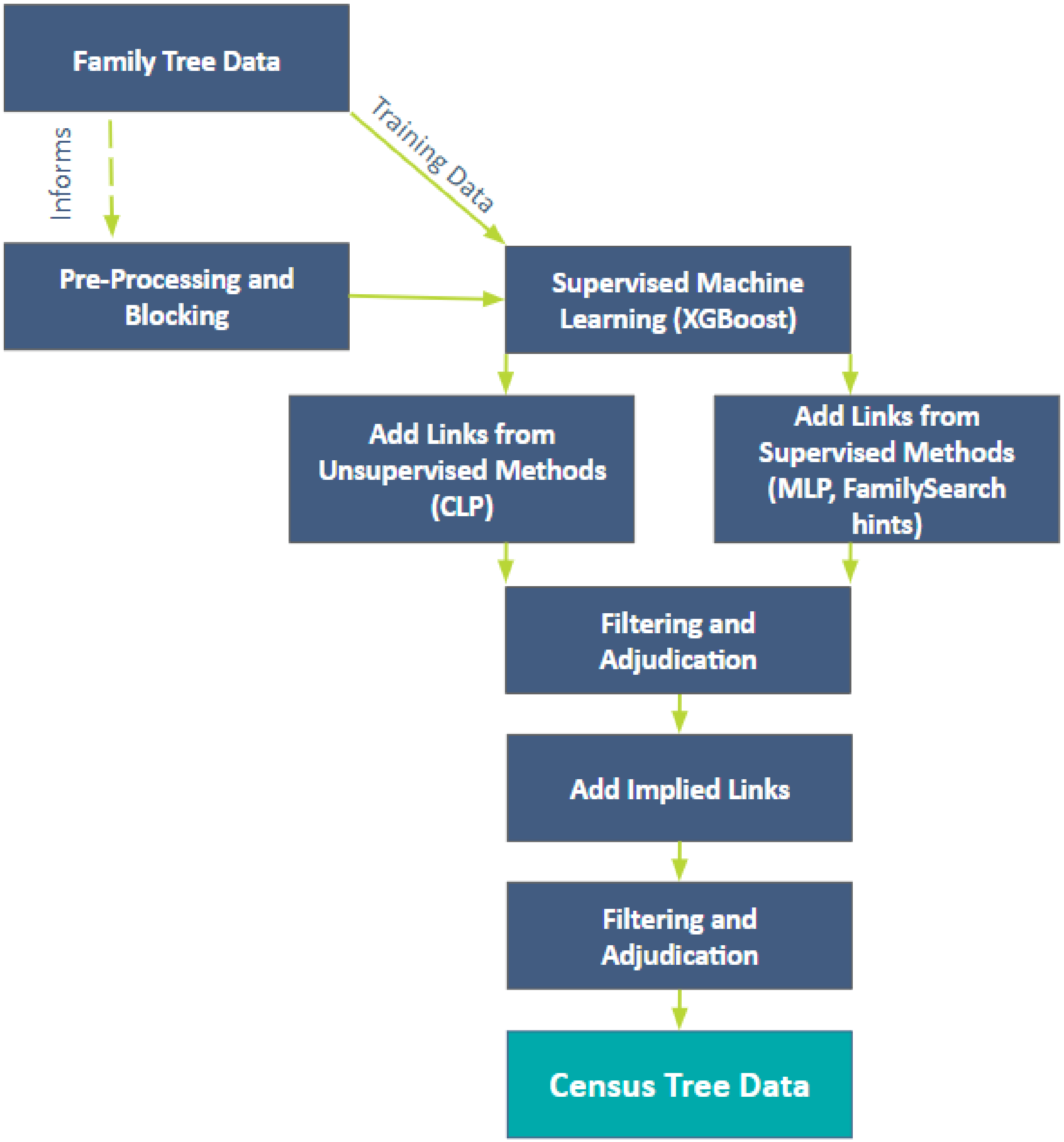
Notes: Match rates are constructed as the number of links between the two years, divided by the number of people age 11 and older in the latter year, with adjustment for rates of under-enumeration in the earlier census and for immigration. CLP – EC links are from the Census Linking Project, using the exact conservative approach. MLP links are from the IPUMS Multigenerational Longitudinal Panel. The Family Tree links are the links made by users on FamilySearch.org, and the Census Tree links are from the final Census Tree dataset.

Figure 1B: Match Rates for Women Using Various Linking Methods, for Censuses Ten Years Apart



Notes: Match rates are constructed as the number of links between the two years, divided by the number of people age 11 and older in the latter year, with adjustment for rates of under-enumeration in the earlier census and for immigration. CLP – EC links are from the Census Linking Project, using the exact conservative approach; the CLP does not include women so the match rate is 0% for all pairs. MLP links are from the IPUMS Multigenerational Longitudinal Panel. The Family Tree links are the links made by users on FamilySearch.org, and the Census Tree links are from the final Census Tree dataset.

Figure 2: The Process for Creating the Census Tree


Family Tree Data
Informs
Training Data
Pre-Processing and Blocking
Supervised Machine Learning (XGBoost)
Add Links from Unsupervised Methods (CLP)
Add Links from Supervised Methods (MLP, FamilySearch hints)
Filtering and Adjudication
Add Implied Links
Filtering and Adjudication
Census Tree Data

Notes: CLP links are from the Census Linking Project, using the NYSIIS standard links. MLP links are from the IPUMS Multigenerational Longitudinal Panel, and FamilySearch hints are created by FamilySearch using their proprietary algorithm. See the text for a description of implied links and of the filtering and adjudication process.

Table 1: Fifteen Most Important Features Used by XGBoost Algorithm, 1900-1910

Feature	Description	Importance
Township distance	Geographic distance between townships	0.1890
Birth year difference	Absolute difference in birth years	0.1047
Middle initial exact	Indicator for middle name exact match	0.0961
Last name uniqueness * last name Levenshtein	Levenshtein string distance in last name, weighted higher for more unique names	0.0722
Last name uniqueness * last name exact	Indicator for last name exact match, weighted higher for more unique names	0.0606
Sign of birth year difference	Sign of difference in birth years	0.0452
Mother's birthplace exact	Indicator for mother's birthplace exact match	0.0383
First name uniqueness * first name Jaro-Winkler	Jaro-Winkler string distance in first name, weighted higher for more unique names	0.0367
State exact * not living in birth state	Indicator for residence state exact match and living outside birth state	0.0304
Immigrant in starting year	Indicator for immigrant in 1900	0.0277
Standardized first name uniqueness * Standardized first name Levenshtein	Levenshtein string distance in standardized first name, weighted higher for more unique names	0.0264
Last name Jaro-Winkler	Jaro-Winkler string distance in last name	0.0246
Relationship exact	Indicator for relationship to head exact match	0.0220
First name uniqueness * first name Levenshtein	Levenshtein string distance in first name, weighted higher for more unique names	0.0214
Father’s birthplace exact	Indicator for father’s birthplace exact match	0.0212

Notes: The importance measure is calculated as the average increase in accuracy across nodes of the decision tree which use the feature. This is the “gain” measure of feature importance calculated by the XGBoost algorithm. The model used 70 features in total.

Table 2: Match Rates for Each Census Pair in the Census Tree

Panel A: Men

	1850	1860	1870	1880	1900	1910	1920	1930
1860	0.6686							
1870	0.5656	0.6455						
1880	0.5865	0.6166	0.7217					
1900	0.6403	0.6333	0.6481	0.6808				
1910	0.7030	0.6688	0.6585	0.6616	0.7406			
1920	0.8371	0.7471	0.7052	0.6872	0.7057	0.7904		
1930	1.1291	0.8825	0.7755	0.7302	0.7033	0.7468	0.8042	
1940	1.7628	1.1424	0.8915	0.7751	0.7189	0.7483	0.7768	0.8527

Panel B: Women

	1850	1860	1870	1880	1900	1910	1920	1930
1860	0.5922							
1870	0.4441	0.5450						
1880	0.4569	0.4975	0.6277					
1900	0.4596	0.4743	0.5075	0.5861				
1910	0.4820	0.4794	0.4951	0.5523	0.7189			
1920	0.5447	0.4966	0.4948	0.5411	0.6200	0.7427		
1930	0.7041	0.5491	0.4920	0.5255	0.5690	0.6166	0.7223	
1940	1.0800	0.6846	0.5340	0.5213	0.5519	0.5693	0.6008	0.7381

Notes: Match rates in the table are constructed as the number of links between the two years, divided by the number of people age 11 and older in the latter year, with adjustment for rates of under-enumeration in the earlier census.

Table 3: Representativeness for Various Linking Methods, 1900-1910

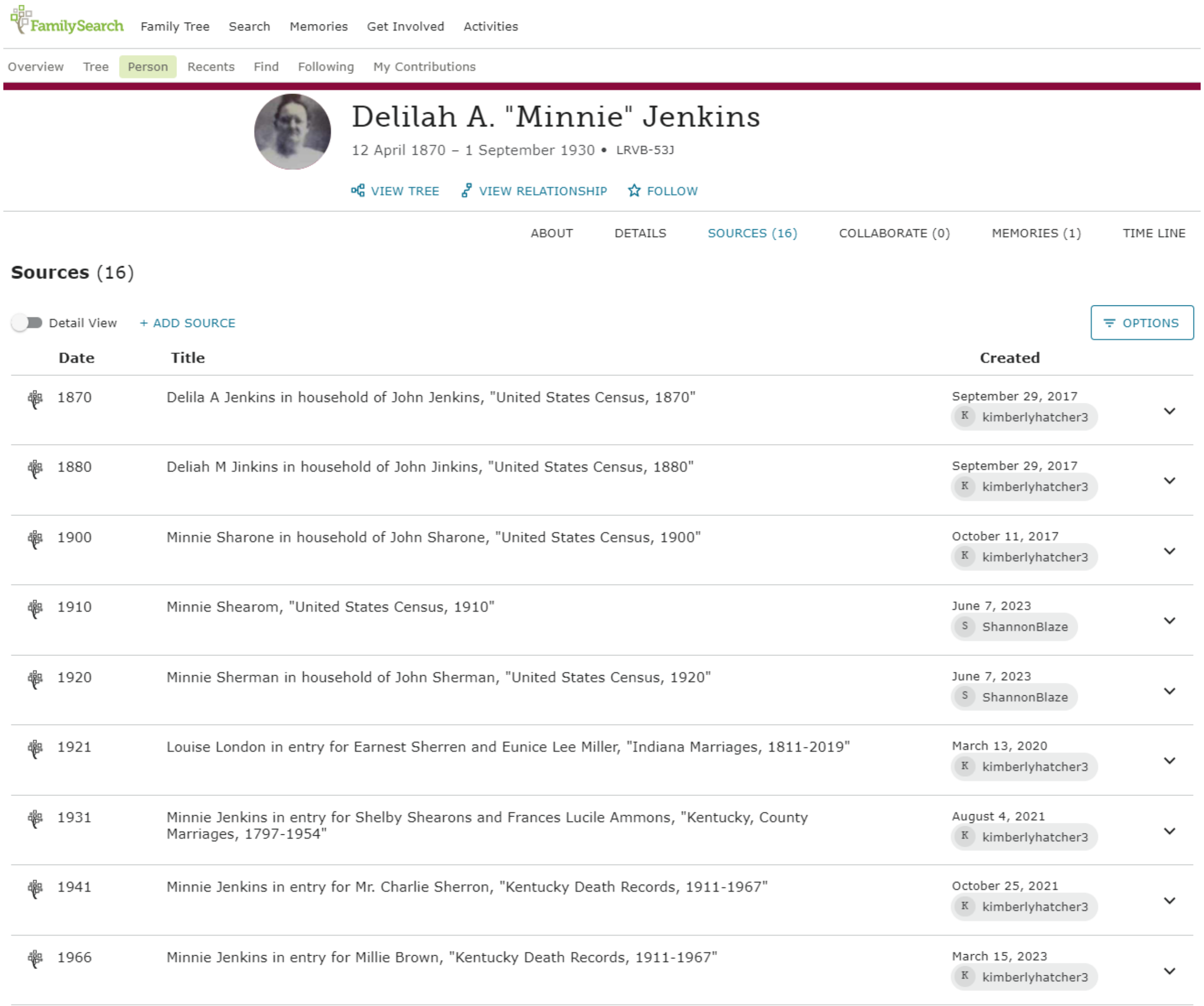
	CLP	MLP	Family Tree	Census Tree	Full Census (Age 11+)
Female	-	0.4273	0.4947	0.4714	0.4824
Age	33.58	33.62	33.02	34.22	33.59
White	0.9239	0.9377	0.9451	0.9248	0.8925
Black	0.0742	0.0619	0.0544	0.0740	0.1030
Married	0.4912	0.4874	0.5317	0.5198	0.5133
HH Head	0.4901	0.2928	0.2844	0.3069	0.2876
HH Size	5.71	6.05	5.93	5.72	5.79
Lives in Birth State	0.6650	0.6934	0.7062	0.6671	0.5905
Speaks English	0.9859	0.9860	0.9901	0.9844	0.9501
Literate	0.9463	0.9501	0.9529	0.9425	0.9150
N	9,806,617	29,238,890	28,267,717	45,772,617	69,725,595

Notes: Unweighted summary statistics for individuals observed in 1910, for which each data set has a link for 1900, compared to the population of individuals age 11 or older in 1910. CLP links are the NYSIIS-Standard links from the Census Linking Project; the CLP does not include women. MLP links are from the IPUMS Multigenerational Longitudinal Panel. The Family Tree links are the links made by users on FamilySearch.org, and the Census Tree links are from the final Census Tree dataset.

Appendix

A. Supplemental Figures

Figure S1: Sources on a FamilySearch Profile



Delilah A. "Minnie" Jenkins

12 April 1870 – 1 September 1930 • LRVB-53J

Sources (16)

Date	Title	Created
1870	Delila A Jenkins in household of John Jenkins, "United States Census, 1870"	September 29, 2017 K kimberlyhatcher3
1880	Deliah M Jinkins in household of John Jinkins, "United States Census, 1880"	September 29, 2017 K kimberlyhatcher3
1900	Minnie Sharone in household of John Sharone, "United States Census, 1900"	October 11, 2017 K kimberlyhatcher3
1910	Minnie Shearom, "United States Census, 1910"	June 7, 2023 S ShannonBlaze
1920	Minnie Sherman in household of John Sherman, "United States Census, 1920"	June 7, 2023 S ShannonBlaze
1921	Louise London in entry for Earnest Sherren and Eunice Lee Miller, "Indiana Marriages, 1811-2019"	March 13, 2020 K kimberlyhatcher3
1931	Minnie Jenkins in entry for Shelby Shearons and Frances Lucile Ammons, "Kentucky, County Marriages, 1797-1954"	August 4, 2021 K kimberlyhatcher3
1941	Minnie Jenkins in entry for Mr. Charlie Sherron, "Kentucky Death Records, 1911-1967"	October 25, 2021 K kimberlyhatcher3
1966	Minnie Jenkins in entry for Millie Brown, "Kentucky Death Records, 1911-1967"	March 15, 2023 K kimberlyhatcher3

Notes: Figure shows sources attached to the profile of Delilah A. "Minnie" Jenkins, including the name of the person who attached the record. Note that the name is different in each of the five attached census records.

B. Supplemental Tables

Table S1: Precision Estimates for the 1900-1910 Census Tree

	Treat Unsure as Incorrect (N = 760)	Drop Unsure (N = 715)	Treat Unsure as Correct (N = 760)
Record Source:			
CLP	0.875	0.949	0.953
MLP	0.933	0.962	0.963
XGBoost	0.912	0.968	0.970
Family Tree	0.961	0.970	0.971
FS Direct Hint	0.952	0.969	0.970
FS Profile Hint	0.950	0.964	0.965
Implied Link	0.892	0.937	0.940
Number of Sources:			
1	0.683	0.785	0.813
2	0.857	0.938	0.943
3	0.893	0.948	0.951
4	0.939	0.964	0.965
5	0.929	0.968	0.970
6	0.982	0.982	0.982
7	0.981	0.981	0.981
Full Census Tree	0.885	0.935	0.938

Notes: Table shows the results of an exercise in which research assistants hand-checked a random sample of the 1900-1910 links from the full Census Tree and classified each as correct, incorrect, or unsure. The top panel shows results by record source, where a record can have multiple sources. The bottom panel shows the results by the number of sources that identified the link. In the first column the unsure links are treated as incorrect, in the middle they are dropped, and in the last they are treated as correct.

Table S2: Features Used by XGBoost Algorithm

Category	Starting Year	Ending Year
Name		
First name JW, LV, LVN, EM	All	All
First name uniqueness interacted with JW, LV, LVN, EM	All	All
First nickname JW, LV, LVN, EM, **NYSIIS EM**	All	All
First nickname uniqueness interacted with JW, LV, LVN, EM	All	All
Middle initial EM (0 if missing)	All	All
The above feature interacted with first name EM and indicator for first initial only	All	All
Indicator for middle name longer than one letter in both years, interacted with middle name JW, LV, LVN, and EM	All	All
Last name JW, LV, LVN, EM, **NYSIIS EM**	All	All
Last name uniqueness interacted with JW, LV, LVN, EM	All	All
Birthplace		
Standardized birthplace EM	All	All
Standardized mother's and father's birthplaces EM	1880-1940	1880-1940
Standardized birthplace uniqueness	All	All
Birth year		
Absolute birth year difference <= 3	All	All
Absolute birth year difference	All	All
Sign of birth year difference	All	All
Age in starting census	All	None
Sex and marital status		
Sex EM	All	All
Female in starting census	All	All
Marital status EM	1880-1940	1880-1940
Married in starting census	1880-1940	None
Single-to-married across censuses	1880-1940	1880-1940

Table S2: Features Used by XGBoost Algorithm (continued)

Category	Starting Year	Ending Year
Household relationships		
Relationship to head EM (0 if missing)	All	All
Indicators for head, wife, son, daughter in starting census	All	All
Wife in ending census but not in starting census	All	All
Race		
Race EM	All	All
Black in starting census, and interacted with absolute birth year difference	All	All
Immigration		
Absolute immigration year difference	1900-1930	1900-1930
Indicator for immigrant in starting census	1900-1930	None
Residence		
State EM	All	All
Interaction of state EM and an indicator for not residing in birth state (0 if missing)	All	All
Township coordinate distance	All	All
Living in same place as in 1935	None	1940
Occupation		
Occupation category EM	All	All
Standardized occupation EM	All	All
Occupation string JW and EM	All	All
Raw occscore difference, and interacted with age	All	All

Notes: This table includes 66 features, of which 6 (bolded) are used for blocking. JW is Jaro-Winkler string distance. LV is Levenshtein string distance, with LVN being normalized by maximum string length. EM is exact match.

Table S3: Size of Training Data for Each Pair

Years	Total	Women	Black	Other Race
1850 to:				
1860	2,216,705	926,244	3,950	4
1870	1,493,049	492,216	2,488	0
1880	1,360,663	336,566	2,229	1
1900	675,397	84,119	854	0
1910	410,234	30,588	446	9
1920	172,391	9,196	171	12
1930	38,861	2,060	33	21
1940	2,757	206	4	4
1860 to:				
1870	2,833,051	1,201,792	5,188	51
1880	2,241,409	736,055	3,649	58
1900	1,183,510	219,844	1,392	35
1910	816,960	101,452	877	37
1920	444,347	34,165	406	36
1930	185,722	9,573	121	102
1940	43,853	2,671	34	75
1870 to:				
1880	4,768,988	2,028,633	51,288	236
1900	2,240,388	562,227	15,567	85
1910	1,688,831	312,420	10,148	123
1920	928,378	109,699	4,167	102

	1930	572,401	43,155	2,260	262
	1940	248,248	15,976	1,062	379
1880 to:					
	1900	5,372,938	1,867,517	64,425	479
	1910	4,033,175	1,030,487	36,618	559
	1920	2,744,518	527,366	20,359	488
	1930	1,678,316	209,626	9,654	981
	1940	950,792	82,914	6,077	1,357
1900 to:					
	1910	7,868,650	3,507,492	91,663	5,712
	1920	5,707,543	1,973,705	38,706	3,177
	1930	3,889,033	938,428	18,679	4,108
	1940	2,543,469	432,008	12,328	4,239
1910 to:					
	1920	11,687,579	5,226,222	98,074	12,633
	1930	6,940,992	2,341,283	38,351	12,054
	1940	4,682,932	1,171,371	22,874	9,821
1920 to:					
	1930	11,728,770	5,120,313	82,500	24,928
	1940	7,003,699	2,451,368	40,517	17,731
1930 to:					
	1940	12,860,670	5,597,041	91,956	43,715

Table S4: Number of Links in Census Tree from Each Source, 1900-1910

	Links Before F&A	% Dropped in F&A	In Census Tree	Unique Links
Family Tree	29,314,798	1.5%	28,874,030	672,841
XGBoost	27,407,692	7.6%	25,317,190	1,470,857
CLP	10,140,318	17.3%	8,388,152	406,770
MLP	30,313,883	1.9%	29,730,141	2,069,840
FS Direct Hint	26,963,154	1.6%	26,534,259	485,118
FS Profile Hint	26,455,508	3.3%	25,589,016	502,274
Implied Links	35,461,926	5.6%	33,468,423	2,314,368

Notes: F&A refers to the filtering and adjudication process described in the text.

Table S5: Number of Sources that Identify Each Link, 1900-1910

# Sources	Links
1	7,922,068
2	6,486,142
3	7,369,745
4	7,039,613
5	7,698,195
6	7,727,108
7	3,126,507
Total	47,369,378

Table S6: XGBoost Feature Importance for Adjacent Censuses

Category	Mean	1850-1860	1860-1870	1870-1880	1880-1900	1900-1910	1910-1920	1920-1930	1930-1940
Name	0.444	0.561	0.471	0.490	0.472	0.394	0.376	0.373	0.412
Residence	0.247	0.201	0.288	0.287	0.204	0.232	0.257	0.240	0.265
Birth year	0.145	0.106	0.114	0.124	0.144	0.156	0.138	0.189	0.187
Household relationships	0.065	0.083	0.087	0.063	0.078	0.058	0.061	0.055	0.030
Birthplace	0.038	0.007	0.010	0.011	0.044	0.068	0.063	0.055	0.042
Occupation	0.026	0.031	0.022	0.020	0.016	0.017	0.029	0.024	0.048
Sex and marital status	0.020	0.003	0.003	0.003	0.038	0.032	0.036	0.030	0.012
Immigration	0.014	0.000	0.000	0.000	0.000	0.042	0.038	0.032	0.000
Race	0.003	0.007	0.003	0.002	0.004	0.002	0.003	0.002	0.003

Notes: There are 66 features in the model, and here we have grouped them into categories. Blocking variables have zero feature importance; these include first name NYSIIS exact match, last name NYSIIS exact match, standardized birthplace exact match, absolute birth year difference within 3, sex exact match, and race exact match.

Table S7: Number of Observations for Each Census Pair in the Census Tree

Panel A: Men

	1850	1860	1870	1880	1900	1910	1920	1930
1860	5,953,064							
1870	4,807,141	8,042,153						
1880	4,333,204	7,125,423	11,783,697					
1900	2,695,500	4,931,251	8,201,564	13,335,312				
1910	1,747,475	3,679,203	6,594,868	11,171,363	24,976,021			
1920	871,805	2,370,506	4,846,225	8,897,191	20,404,710	30,164,025		
1930	260,567	1,207,552	3,179,977	6,656,254	17,397,956	25,541,131	35,888,058	
1940	35,159	366,292	1,607,257	4,318,696	13,941,716	21,358,133	29,849,380	42,665,479

Panel B: Women

	1850	1860	1870	1880	1900	1910	1920	1930
1860	4,957,971							
1870	3,645,222	6,733,140						
1880	3,181,939	5,499,005	9,947,076					
1900	1,826,668	3,426,055	5,898,085	10,712,058				
1910	1,157,056	2,427,618	4,469,454	8,329,403	22,384,230			
1920	600,127	1,533,752	3,150,035	6,333,400	16,832,433	26,957,112		
1930	199,280	791,811	1,993,076	4,533,003	13,312,424	20,314,679	31,584,817	
1940	31,704	276,390	1,061,427	2,988,185	10,411,260	15,925,055	23,141,851	37,516,261

Table S8: Representativeness of Census Tree, for Adjacent Censuses

	1850-1860		1860-1870		1870-1880		1880-1900	
	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)
Female	0.455	0.485	0.456	0.495	0.458	0.491	0.446	0.489
Age	31.33	30.81	32.18	31.33	32.78	32.09	40.56	33.16
White	0.988	0.978	0.982	0.878	0.925	0.877	0.929	0.887
Black	0.011	0.018	0.018	0.119	0.074	0.120	0.070	0.109
Married	0.476	0.466	0.485	0.464	0.489	0.475	0.633	0.470
HH Head	0.294	0.279	0.302	0.283	0.305	0.285	0.417	0.287
HH Size	6.72	6.52	6.42	6.33	6.21	6.13	5.43	5.85
Lives in Birth State	0.610	0.512	0.611	0.550	0.632	0.573	0.629	0.602
Literate	0.923	0.912	0.866	0.785	0.868	0.830	0.908	0.883
High School Grad	-	-	-	-	-	-	-	-
N	10,600,283	18,975,196	14,406,659	27,098,171	21,155,063	35,469,052	24,477,539	56,082,690

Table S8: Representativeness of Census Tree, for Adjacent Censuses (Continued)

	1900-1910		1910-1920		1920-1930		1930-1940	
	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)	Census Tree	Full Census (Age 11+)
Female	0.471	0.482	0.471	0.489	0.468	0.495	0.467	0.500
Age	34.22	33.59	34.95	34.60	35.83	35.37	37.12	36.69
White	0.925	0.892	0.932	0.899	0.936	0.902	0.934	0.905
Black	0.074	0.103	0.067	0.097	0.061	0.094	0.063	0.091
Married	0.502	0.482	0.521	0.509	0.528	0.517	0.527	0.523
HH Head	0.307	0.288	0.318	0.303	0.327	0.310	0.341	0.326
HH Size	5.72	5.79	5.38	5.36	5.11	5.03	4.70	4.58
Lives in Birth State	0.667	0.591	0.662	0.600	0.664	0.610	0.683	0.656
Literate	0.943	0.915	0.959	0.937	0.968	0.954	-	-
High School Grad	-	-	-	-	-	-	0.253	0.257
N	46,283,690	69,725,595	55,677,558	80,497,032	66,727,574	96,202,610	79,111,247	108,342,194

Notes: Unweighted summary statistics for people linked between the two years in the Census Tree, compared to the linkable population (those age 11 and older) in the latter census.