"Evaluating the Accuracy of Linked U. S. Census Data: A Household Linking Approach"

The Systematic Linking of Historical Records, University of Guelph, May 10-13, 2017

Ronald Goeken
University of Minnesota

Yu Na Lee
University of Minnesota

Tom Lynch
University of Minnesota

Diana Magnuson† Bethel University

December 2017
Working Paper No. 2017-1
https://doi.org/10.18128/MPC2017-1

†Correspondence should be directed to:
Diana Magnuson
University of Minnesota, 50 Willey Hall, 225 19th Ave S., Minneapolis, MN 55455
e-mail: ruggles@umn.edu, phone: 612-624-5818, fax:612-626-8375

Introduction

Despite the proliferation of published studies using linked decennial census records there has been little empirical work on the accuracy of the linked data. The primary reason, of course, is that you can never definitively state that two records taken from two distinct censuses represent the same person. Given the absence of unique identifiers (e.g., social security numbers) matching historical census records depends on high similarity between primary linkage variables; e.g., names, age, sex, and place of birth. Potential links are then classified as true or false according to rules or machine learning procedures. Estimating linkage rates is a straightforward exercise, but error rates can only be measured indirectly.

The goal of most historical census linkage projects is to create linked data that does not include corroborative evidence derived from co-resident kin and migration status because of bias issues. This is a valid concern, but it is also possible that relying on linkage methods that ignore a fair amount of corroborative evidence comes at a cost. The obvious effect would be to lower linkage rates. A potentially more significant concern would be the effect on error rates. The main issue is if the true link is unidentifiable (because of under-enumeration or a mismatch or low similarity on key linkage variables), then any link to this record will be false.

Most record linkage projects more or less assume that the inability to find true links due to mismatches or low similarity for key linkage variables is a relatively minor issue. Our strategy for investigating this topic is to use a maximum amount of information to establish a set of verified links. Primarily, we plan on using the presence of common kin and residential stability (i.e., living in the same place) in successive decennial censuses to supplement similarity at the individual level. Although many true links will not have corroborative household or residential information, we find that many can be verified. These verified links will then be used to optimize blocking strategies and to test procedures used to classify potential links generated by individual level classifiers, primarily by constructing linkage and error rates.

This is still our basic mission statement. However, the nineteenth century linking--which is part of a five-year project examining demographic change in the aftermath of the American Civil

War--is still in progress. We provide a status report in the last half of the paper, but in the first half we discuss the development of the household linking process.[1]

The 1880 Complete-Count Linkage Project (2003-2009)

In 2003 the Minnesota Population Center began work on a project that would eventually link the complete-count database of the 1880 U. S. population census to samples of other 19th and early 20th century U. S. decennial censuses. The original grant asserted that we would establish links at the individual level and only use a set of variables that would minimize linking bias; i.e., names, age, sex, race, and place of birth. We did not use place of residence or information gleaned from co-resident kin because of bias concerns (i.e., that non-migrants and those living with the same kin in both censuses would be overrepresented in the linked population).[2]

The decision to ignore corroborative evidence (because of bias concerns) ultimately resulted in the choice of a conservative linking strategy. The final linkage rates were modest, but we felt this was necessary in order to achieve (relatively) low false positive rates. Although we did not possess a "truth" sample for verification, indirect evidence indicated we had relatively low false positive rates. For example, if we independently linked two brothers who were co-resident in the 1880 census, rarely were they also not co-resident in 1870 (i.e., sets of brother came from the same households in both census years) . Another example would be consistency in our male-only and couple-only linked samples; if a male from the 1880 census was linked in both of these samples, we rarely had this individual linked to two different records in the 1870 census.[3]

Both of these diagnostics offer evidence of consistency and indirectly imply precision. They also cherry-pick a bit, in that the selected universe was native-born whites in 1880; it is likely that error rates were higher for African Americans and the foreign-born (specifically the Irish). It is

[1] Hacker, J. David. Principal Investigator. "Models of Demographic and Health Changes Following Military Conflict" 1R01HD082120-01. National Institute of Child Health/Human Development.
[2] Ruggles, Steven. Principal Investigator. "Population Database for the United States in 1880." R01 HD39327, NICHD-DBSB.
[3] Ronald Goeken, Lap Huynh, T.A. Lynch and Rebecca Vick, "New Methods of Census Record Linking," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, volume 44, issue 1, 2011. Steven Ruggles, "Linking Historical Censuses: A New Approach," *History and Computing*, volume 14, March 2002, pp. 213-244.

also probable that some demographic sub-groups might be more likely to have consistent information in successive censuses and thus be more likely to be accurately linked (and this would probably apply to married men and children). We were also more confident in our 1870-1880 linked sample compared to linked samples with inter-censal gaps exceeding ten years (i.e., we expect false positive rates to increase as the years between linked censuses to increase).

Another reason we thought we had relatively low false positive rates (at least for married men and sons) was because we spent some time visually evaluating the linked households. Although we linked on the individual basis (primary links), the resulting linked data consists of the primary links along with their co-resident household members from the two specific censuses. If any of the non-primary records appeared to be the same person in the respective censuses, then we established the link based on a set of rules (with these linked records identified as secondary links).[4]

Many of the 1870-1880 primary links do not have co-resident secondary links for obvious reasons; an example would be a 24-year-old son living with his parents in 1870 linked to a 34-year-old household head living with his wife and three-year-old daughter in 1880. But many of the primary links have co-resident secondary links; in the male 1870-1880 linked sample 28 percent of the primary links have no secondary links, 19 percent have one and 53 percent have two or more secondary links. Although we did not do a systematic analysis, it is possible to pick out low-quality primary links, and the top panel in Figure 1 gives an example. Here the primary link is Henry McHugh, age 14 in 1870 and age 25 in 1880. However, no other record in either household appears to be the same person (with the only real possibility being James E, age 3 in 1870 and Edward J., age 12 in 1880). But an example like this is relatively rare in the 1870-1880 male linked sample. Much more common would be the linked records in the second panel. Here the primary link is James Felkins, age 61 in 1870 (linked to James H. Felkin, age 71 in 1880). And three of James' kin are secondary links and they appear to be correctly linked despite the differences in expected age. In fact, there is also a high probability that Martha, age 53 in 1870 is the correct link to Matilda, age 67 in 1880.

[4] See https://usa.ipums.org/usa/linked_data_samples.shtml.

Whether Martha is actually Matilda illustrates a basic dilemma with establishing some of the secondary links; they could be the same person, but maybe or probably not (it is definitely possible that James re-married to Matilda at some point between 1870 and 1880). But despite a somewhat conservative standard for establishing secondary links in cases of ambiguity, the secondary links had lower levels of similarity compared to our primary links. For example, in the 1870-1880 male file, less than 1 percent of the primary links have an expected age difference exceeding one year of age. For secondary links, over 20 percent have an expect age difference of two years of age or more.

The higher precision for our primary links resulted from our conservative linkage strategy. To use a simplified example, if we had two potential links for a given record, with one potential link being an exact match on all linkage variables and the other being an exact match on all variables with the exception of an expected age difference of four years, we would reject both potential links because of ambiguity (yes, the potential link with an exact age match would have a higher probability of being the true link, but we took a conservative linking approach). In addition, if our only potential link was an exact match except for an expected age difference of four years, we would reject because of low precision. In other words, we had a two-threshold approach, with the higher threshold determining eligibility to be a primary link, and the lower threshold identifying the area of ambiguity; a link was defined as one-and-only-one potential link above the higher threshold, and no other potential links above the lower threshold. This resulted in fairly accurate results, but also meant that our primary links were not representative of all true links. This finding, along with the understanding that many primary links can be verified through the presence of consistent co-resident kin in both census years, were important insights, but we really did not appreciate this until we were finished with the 1880 complete-count linkage project.

Linking Slave-Owners to the 1850 Complete-Count Population Database

Our next linkage project was the 1850 complete-count database of the 1850 U. S. Census,

which was a collaboration with the Church of Jesus Christ of Latter-Day Saints (LDS).[5] In addition to the population records, LDS had also entered the 1850 slave schedules. The slave census has the slave owner names and we wanted to link the slave owners (and their slaves) to the slave owner's population record. The population and slave enumerations were done simultaneously, so slave owners in the slave schedules and the population schedules should be (roughly) in the same order in their respective databases. However, it became apparent that some slave schedule pages were microfilmed out of their original order (and there are no page numbers or enumerator sequence numbers to verify the sort; the pages have an enumeration date field, but this information was often missing and was not considered to be incredibly reliable). The forms do not have information for slave owner age, birthplace or sex (and about 20 percent of slave owners in 1850 were female). The only owner-related information on the slave schedules is slave owner name, but the forms have legibility (and transcription) issues and given name often consists of a single initial.

Here we were not concerned about bias in linkage methods; the goal was to accurately link all of the slave owners to their respective population records. The basic rule was that slave owners and their population record would usually follow approximately the same sequence in both schedules (with some exceptions due to absentee slave owners). But we had to identify mini-sequences within counties when the slave schedule pages were out of order. To do this we blocked by county of residence and restricted potential links to records age 17 and older in the population data, and wrote out potential links that exceeded a preset threshold for given and surname similarity.

The creation of slave owner sequences relied on identifying clusters of potential links (i.e., a high proportion of slave owners from a slave page or range of slave pages that had potential links to a given page or range of pages in the population data). Figure 2 shows an 1850 slave schedule page; a slave page has 84 lines for individual slaves, and this page has 14 slave holdings (i.e., 14 slave owners). The population schedules have 42 lines per page in 1850 and Washington County, Missouri had a free population of 7,736 in 1850; thus Washington County,

[5] Alexander, Joseph Trent. Principal Investigator. "Baseline Microdata for Analysis of U.S. Demographic Change. PRF601864. National Institute of Child Health/Human Development.

Missouri had approximately 190 pages of population data in 1850. Again, the only information used to establish the link is name (i.e., we do not have age, birthplace or sex for the slave owners). The basic concept was that the 14 slave owners would have random potential links dispersed over the entire county (pretty much anywhere on pages 1 through 190 in the population data for this example). But typically we could identify a cluster of potential links on a given page or range of pages in the population data; probably not all 14, but we would see clusters, which would indicate that these potential links were the true link (even if another potential link elsewhere in the county had greater name similarity; in other words, sequence order was often a better predictor of the true link than name similarity).

We eventually began to understand that we could apply the slave owner sequencing logic to linking the population records taken from two distinct censuses on the household basis. Basically, a household is a subset of a page of population data. And household members are similar to a group of slave owners on a given page of slave data. The analogy breaks down a bit when dealing with individuals enumerated ten years apart. However, as mentioned above, under certain circumstances we would expect some co-residential stability. Basically, if we find certain combinations of nuclear kin age ten and older co-residing in a given census year, there is a very high probability they were also co-residing ten years earlier. For example, the expectation is that a household head, spouse and two teen-aged sons in the 1880 census will also have been enumerated together in the same household in the 1870 census. At the individual level each of the four records could have multiple potential links to the 1870 census, but the true link would be identifiable because it would be the household combination that also had potential links for other members of the household. Again, the correct household might not have potential links to all four, but three out of four probably would be enough to establish and confirm the link.

Household Linking the Two Enumerations of St. Louis in 1880

One issue with this approach is the large number of individual potential links that need to be generated in order to establish the household links. Most of our potential links will be the only

link between specific households in two different censuses (and will not be a true link), but we have no way of knowing this until we generate and process all of the potential links. And working with the complete-count tabulations would require improvements in our processing speed.

We also had to develop an actual process, which evolved during work we did linking the two enumerations of St. Louis in 1880. The first enumeration occurred in June and, because of allegations of an undercount, the Census Office authorized a second enumeration in November of the same year. This was not the first time that an American city would be re-enumerated, nor would it be the last.[6] But St. Louis in 1880 appears to be unique in that the second enumeration was an attempt at a complete re-enactment; the same enumeration sheets were used in both enumerations and enumerators were expected to complete all of the census questions.[7] Both enumerations also used the same June 1 reference date. The enumerator instructions for the November recount state that "enumerators will not ask the people of their district whether they have changed their residence since June 1, 1880, but they must ask, "Were you residents of St. Louis on the 1st of June?" or, "Was St. Louis your home on the 1st of June, 1880? … enumerators will make no inquiries as to removals from one family to another, and from one district to another since June 1 (as suggested in my circular); but they must be very particular to ask, "Has any member of this family or household left the city since June 1, 1880?" and "Has any person or family moved from the city from this neighborhood since June 1, 1880?"[8]

The use of the June 1 reference date for the November enumeration raises a number of issues regarding the accuracy of the results. The enumeration of individuals who were present on June 1 but had subsequently left the city would depend on relatives, neighbors or landlords reporting this information to enumerators as well as giving them information on the migrants' individual characteristics. Enumerators were always dealing with these issues, and getting fairly

[6] Francis A. Walker, *A Compendium of the Ninth Census* (Washington, D.C.: GPO, 1870), pp. xx-xxi.
[7] For example, New York City and Philadelphia had recounts in 1870. In both cases enumerators were only expected to fill in a subset of the questions on the original enumerator sheets.
[8] "The Census: Revised Instructions Issued to the Enumerators--One District Already Finished," *St. Louis Post Dispatch*, November 9, 1880.

accurate information on absentee residents would not be an insurmountable difficulty if the respondent had some familiarity with the absentees. But the five-month gap between the reference date and the actual enumeration would make it difficult to get an exact count and precise information on relatively transient population sub-groups: extended kin and unrelated individuals in general, and those residing in hotels and larger rooming and lodging establishments more specifically. But this should not affect our ability to link the data. In contrast to records taken from two separate decennial censuses, the two enumerations of St. Louis constitute a relatively closed universe; we expect to find the same individuals living with each other. In addition, we have street addresses for both enumerations. Although some individuals would relocate (within the city) between the two enumerations, the addresses would prove useful in the linking process. The use of corroborative evidence in the form of co-resident kin and street address undoubtedly produces biased linkage results. But this issue is not important here because our goal is to link, to the extent possible, all of the records.

Our linkage approach consists of initially establishing potential links at the individual level. Names are cleaned (i.e., non-alpha characters are removed) and parsed (i.e., the given name 'Mary E' becomes name1 = 'Mary' and name2 = 'E'). Records are blocked by sex and similarity scores based on the Jaro-Winkler algorithm are calculated for given name and surname.[9] Record pairs having a surname similarity score of at least 0.9, a given name similarity score of at least 0.7, and an absolute age difference of less than five years are selected as potential links. We did not standardize given names, nor did we use birthplace or race as a blocking factor. Some name standards are fairly obvious, but we decided to empirically determine the appropriate standards based on our initial links rather than impose standards based on assumptions. We hoped to use street address to facilitate the linking, but our initial attempts to link on the basis of matching street and house number information produced relatively few quality matches. In addition, we have enumeration district information, but there were 168 enumeration districts in the first enumeration compared to 450 in the second. For that reason we initially did not use district information to link records.

[9] Peter Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer, 2012. http://link.springer.com/book/10.1007%2F978-3-642-31164-2.

Although we find far fewer exact or near duplicates in St. Louis than we would if we were trying to link the entire country, we nonetheless encounter a fair amount of ambiguity when looking at potential links on the individual level. Much of this ambiguity is eliminated if we take into account characteristics of co-resident family members. For example, in the first enumeration we have a 'John O'Donnell' who was 43 years old. Restricting the potential links to exact name matches and a maximum age difference of four years, we have three men named John O'Donnell in the second enumeration with ages of 45, 45 and 46 (see Figure 3). We know that the 43-year-old in the first enumeration is actually the 46-year-old in the second enumeration after we take into account information from other household members.

Rather than creating variables for each individual pertaining to information gleaned from co-resident kin (e.g., father's name, father's age, mother's name, mother's age, etc.) we create potential links for each individual using the simple method outlined above. Then we sum the number of potential links between specific households in the two enumerations. Using the O'Donnell example, each household member in the first enumeration has numerous links to individual records in the second enumeration. For most of these potential links, however, only one of the household members has a link to a specific household in the second enumeration. Despite the inconsistent age for John O'Donnell in the two enumerations (age 43 and 46), we know that this is the correct link after determining that his spouse and children also have potential links between these two households.

This process also allows us to establish links even if some of the household members do not have potential links in our initial linking pass (see Figure 4). In this household the first and third members of the two households were not in our initial potential links file because of low given name similarity (Autonia-Anton has a Jaro-Winkler similarity score of 0.699, which is below the 0.7 threshold) and excessive enumerated age difference (Annie was 19 years old in the first enumeration and 14 years old in the second enumeration). However, after determining that there are four other links between these households, we can also establish links for the records that were not initially linked on the individual basis.

We established links between households based on the following rules. First, if we have four or more potential links between specific households in the two enumerations, and each of the households with four or more potential links did not have two or more links to any other household (in the other enumeration), then we flagged it as a linked household. Second, we also accepted households with three potential links, if neither of these households had two links to any other household. Finally, we reviewed our work by visually inspecting the households with the lowest composite similarity for names and age or linked households with a majority of household members unlinked at the individual level. Using this approach we were able to link about one third of the first enumeration households; 21,214 out of 63,325 households and 99,147 out of 276,683 related individuals.

This method only works on related individuals and will not link smaller households. However, after establishing high quality linked households, we set them aside and made additional passes through the data (discussed below). We also used the visual review process to assess why many households remained unlinked. A primary reason was lower levels of surname similarity for unlinked households, while some smaller households were unlinked because of the use of diminutives or abbreviations for given names in one enumeration or the other. We also began to explore ways to use place of residence to either verify or link households with relatively low similarity. For example, some linked households had street agreement, but their house number was off slightly (e.g., 2402 Market St. in one enumeration versus 2404 Market St. in the other). In addition, some of our initial set of household links had house number agreement, but the street name disagreed. An examination of the linked households identified many street name corrections and we were also able to construct an enumeration district translation table between the two enumerations. Although many linked households had street address disagreement, almost all of the linked households that had identical address information resided in one of a set of contiguously numbered districts in the second enumeration corresponding to a single district in the first enumeration. The correction to addresses and the use of the enumeration district equivalents allowed us to link households that had been difficult to link because of their small size or because of low surname similarity.

A second group of potential links was generated using the same thresholds used in the initial pass, except we lowered the surname threshold to a Jaro-Winkler score of 0.7 and applied some empirically-derived name standards to the given names. We then generated a second batch of household links using rules based on the number of potential links between specific households in the two enumerations. After identifying higher quality household links, the household linking rules allowed less precision if there was some evidence of residential persistence; either identical address information or similar address and residing in the same enumeration district equivalent.

Figure 5 shows two examples of linked households with surname similarity below our initial threshold of 0.9. The surname combination of Burgherdt-Burkhart generates a Jaro-Winkler score of 0.86, a level generally sufficient to establish a link if other linking variables also had relatively high similarity. And, after looking at the entire household, it is obvious that these households were correctly linked. The second household in Figure 5 is also linked, but has a surname similarity of 0.67. Here we suspect that an individual link with the surname combination of Fitzgerald-Vetzgura would be rejected by most classifiers. After looking at the household composition, however, we conclude that these are the same people. Any doubts are alleviated by looking at the household head's occupation ("stone mason" in both enumerations) and street address (the household was enumerated at 2405 Division Street in both enumerations).

Occupational information was never explicitly used to establish links. But we began to use street address and enumeration district information to link households, and this was useful in establishing links between smaller households (especially one- and two-person households). The bottom two linked households in Figure 5 give a couple of examples. The first household has a surname similarity of 0.63, and linking is further complicated by the head's given name (Frank vs. F.H. in the two enumerations). However, these households were enumerated at the same address, and are very likely the same people (with additional corroboration provided by the head having the occupation of 'Retail Grocer' in both enumerations). The second linked household in Figure 5 has higher surname similarity (0.85) but linking is complicated by the

head's given name (Caroline vs. Catherine in the two enumerations). Although they do not have identical street information (4th Street vs. 5th Street) they do have identical house number information and were enumerated in the same enumeration district equivalents, which was enough of a tell to establish the link (we also have occupational similarity for the head's occupation: Caroline's listed occupation was "Keep Millinery Store" and Catherine was a "Milliner").

The rules-based system, with its shifting thresholds and manual intervention, undoubtedly introduces bias. However, we are primarily interested in maximizing the number of links and making sure that they are correct links. Although we have not finished our work linking the related individuals, Table 1 shows that we have established links for 78 percent of the households in the first enumeration and 74 percent of the households in the second enumeration, which corresponds to 80 percent of the related individuals in the first enumeration and 76 percent of the related individuals in the second. As expected, given our household linking approach, we have more success linking households that contain more related individuals. Some of the currently unlinked households cannot be linked because the household is missing from one enumeration or the other. We anticipate, however, increasing our linkage rate through trial and error and the process of elimination. Some of the unlinked households have surname similarity below the thresholds used thus far, and we continue to modify our rules to link the smaller households. In addition, in the future we will attempt to link the unrelated population, although we suspect that many of the boarders and lodgers will be unlinkable due to the absence of corroborative information supplied by co-resident kin.

Table 2a gives the linked population's distribution by surname similarity measures. At higher levels of similarity we would typically assume a potential link with that combination of surnames would be a true link given sufficient similarity for other linkage variables (e.g., given name, age, birthplace, and sex). This assumption begins to break down as we see less similarity in the surname combinations. Figure 6 gives examples of surname combinations from the St. Louis linked records along with the Jaro-Winkler score, phonetic codes and matched letter metrics. There is no absolute rule for deciding at what point the similarity between sets of

linked surnames transitions from “plausible” to “maybe” to “doubtful.” Based on Figure 6, the transition from “maybe” to “doubtful” probably begins around 0.8 Jaro-Winkler similarity. And this means over 11 percent of our links would be treated with a fairly high level of scepticism without the corroborative information from other co-resident household members (or consistent place of residence information).

In addition to Jaro-Winkler score, Table 2a gives surname matching rates for two phonetic code algorithms, NYSIIS and doublemetaphone. We also construct measures indicating whether the first letter, the first two letters, and the first three letters of a surname match for our linked records. Almost half of the linked records are perfect matches, and all of these would also be considered matches using the phonetic codes and matching letters techniques. However, for linked records with non-exact matches for surname but a Jaro-Winkler score greater than 0.95, 66 percent would be a match using NYSIIS and 74 percent would be a match using doublemetaphone.[10] Overall, 69 percent of the surname combinations of the linked population have a NYSIIS match compared to 73 percent for doublemetaphone. Over 93 percent of the linked surname combinations match on the first letter, with 80 and 71 percent matching on the first two and first three letters.

The second panel in Table 2b shows the distribution by Jaro-Winkler score for given names. With given names, we are more concerned with standardizing abbreviations and diminutives than with whether we can match dissimilar combinations with phonetic codes. The distribution of linked records that have perfect given name similarity is 53 percent, with another 2.7 percent having a single initial matching the first letter of a full given name. This leaves over 44 percent of the links with less than perfect similarity. However, we constructed name standards after examining combinations of non-identical given name combinations in our linked data. In addition to the 53.8% of links with an exact name score, another 25% receive an exact score after standardization.

[10] https://en.wikipedia.org/wiki/New_York_State_Identification_and_Intelligence_System; https://en.wikipedia.org/wiki/Metaphone; http://www.b-eye-network.com/view/1596; https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

The overall imprecision in given names is also driven by the fact that some of our linked records have distinctly different given names in the two enumerations. Figure 7 gives a few example of these. The first set shows linked records that would have a given name match if we compared first names to middle names. The second set consists of examples where the given name matches a middle initial for the linked record (e.g., the "N" in "Bayard N" probably stands for "Nelson"). Nonetheless, the third set of linked records have little or no similarity in their given names, nor do they have middle initials that match a given name. Possible explanations for given name inconsistency would include changing personal preferences, respondent bias, enumerator error, and transcription error.

Table 3 gives the distribution of age precision for our linked records. If enumerators were giving a respondent's age as of the November enumeration (rather than age on June 1st) then being a year older in the second enumeration would be considered a good or perfect match. Being a year off in the other direction would also be considered a good match if we were linking across different decennial censuses. But that would still leave over 16 percent of our linked records with an age difference of two or more years. Some respondents may not have known their true age, and their response to enumerators may have been somewhat random. Some of the imprecision is caused by respondent bias, that co-resident kin or even neighbors might have been supplying information to a given enumerator. Transcription error would also contribute here. Regardless of the source of the error, we suspect that age differences in true links found in two different 19th century U. S. censuses would have similar (or possibly higher) rates of imprecision.[11]

The table also gives the somewhat surprisingly high levels of sex errors in our linked data, where almost one percent of the linked records have a sex mismatch. Although we did minimal blocking in linking the two enumerations, we did block by sex. After establishing links between households, we often have remaining unlinked related household members in the household in

[11] Peter R. Knights, "Accuracy of Age Reporting in the Manuscript Federal Census of 1850 and 1860," *Historical Methods Newsletter*, Vol. 4, Issue 3, 1971. Ronald Goeken, Lap Huynh, T.A. Lynch and Rebecca Vick, "New Methods of Census Record Linking, *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, Vol. 44, Issue 1, January 2011.

both enumerations. We automate a forcing procedure to link these records (if possible). We evaluated the results through clerical review, and in the process found many households with a single unlinked record in both enumerations that was very similar with the exception of a sex conflict. These records tended to be younger individuals, and often had given names that were gendered equivalents (e.g., Josephine to Joseph, Augusta to August, and Julia to Julius). It is possible that in the absence of a declaration of gender on the part of the respondent, infants and small children would not have been easily identified by the enumerator as male or female. This also reflects the oral nature of the census; enumerators recorded what they thought they had heard.

Table 3 gives place of birth and race consistency for the linked records. The reporting of the race variable was relatively consistent, especially after taking into account inconsistency in the black and mulatto categories. Only 0.2 percent of the linked records go from white to black/mulatto (or vice versa). In contrast, over 8 percent of our linked records have mismatched birthplaces and over 18 percent have mismatches on parental birthplaces. The disagreement rate goes down quite a bit if we combine all U.S. birthplaces into a single category and do the same for the foreign born. But even using this conservative measure, 1.3 percent of our linked records have a U.S. birthplace in the first enumeration and a foreign birthplace in the second enumeration, and 1.2 percent have a foreign birthplace in the first enumeration and a U.S. birthplace in the second enumeration.

Our evaluation of linkage variable precision for the St. Louis data is preliminary, since we have not finished linking the two enumerations. The overall impression at this point is that a significant number of the linked records would not be linkable at the individual level because of low similarity. The only way we were able to link some of the households was by using address information along with the assumption that the two enumerations were a relatively closed universe.

Household Linking the Complete-Count 1870 and 1880 U. S. Censuses

We suspended the St. Louis linkage project in late 2016 (although we anticipate finalizing the linking at some point). We initially hoped to use the St. Louis linked data to train individual-level classifiers that we would use to link the various 19th century U. S. censuses. One reason why this might not be a great idea is that the high levels of imprecision found in the St. Louis linked data might not be representative of what we would find in the population of all true links found in the decennial censuses. This is basically an issue of whether or not the two enumerations of St. Louis were of atypical poor quality. We have no way of directly answering this question; we suspect that overall the accuracy (or consistency) found in the 19th century U. S. censuses was less than ideal. The relative lack of precision in the linked St. Louis data could be a worse case example, but it could also be what we would typically expect in enumerations of large American cities in the 19th century.

Given concerns about using the St. Louis linked data as training data, we decided to apply the household linking process to the complete-count decennial censuses. It was unclear how many households we would be able to link, but we were confident that it would be a sufficient number to train and test individual-level classifiers. We would also be able to construct false positive estimates based on verified links (at least for the proportion of the population that we would link and confirm via the household linking process).

The only real impediment to applying the household linking process to the complete-count tabulations is the relative size of the databases; e.g., the United States had a population of 38 million in 1870 and 50 million in 1880. When we began work on linking the 1870 and 1880 complete count databases last fall it was taking at least a week of processing time to generate a basic potential links file. Earlier this year, however, we made some improvements and are currently able to generate a potential links file comparing 1870 to 1880 in about a day.

We block by sex and place of birth. We write out potential links if expected age difference is less or equal to five and both given and surname similarity is greater or equal to 0.8 (Jaro-Winkler). If the given name is an initial (in either year) and it matches the first letter of the given name for a record in the compare year (regardless of whether it is an initial or full name), then given name similarity is set at 0.8 (and is thus eligible to be included in the potential links file).

We also apply a relatively short list of given name standards (based on our St. Louis household linked data). [12] The outfile consists of 2.4 billion potential links.[13]

At this point we are only interested in records that constitute a cluster; basically we want to examine sets of two or more potential links between specific households in 1870 and 1880 (i.e., potential household links). Thus we filter out any potential link that is the sole link between specific 1870 and 1880 households. This reduces the file to 79 million individual potential links and 38 million 1870 and 1880 household combinations. Although the potential links file used a 0.8 surname threshold, we initially only process records that have surname similarity of at least 0.9. This further reduces the file to 48 million individual potential links and 21 million 1870 and 1880 household combinations. Most of the 1880 households that are included in the potential links file have multiple households in 1870 as potential links (e.g., only 10 percent have a potential link to a single household in 1870). At this stage we have ambiguity given that we are using relatively low age and name similarity thresholds, and some birthplace blocks contain a disproportionately large number of records (e.g., New York State, Ireland, Germany). We could attempt to disambiguate conflicting links based on composite household age or given name similarity, but we were fairly confident that applying rules similar to those used linking the St. Louis enumerations provide a good first approximation.

We work from the perspective of 1880 and calculate the number of individual level links between a specific 1880 household and 1870 households (the minimum will be a potential link

[12] And we do a four-way given name comparison on take the maximum value (i.e. 1. 70raw/80raw; 2. 70raw/80std; 3. 70std/80raw; 4. 70std/80std.

[13] We use custom software written in Python to compare records between complete-count datasets. Development of the software considers the performance effects of four main parameters: I/O time (including network communication), compute time, memory consumption, and disk space. Our software keeps data on disk as long as possible, only pulling in data when needed and immediately writing it back out to disk at the conclusion of processing. This strategy requires many more disk reads/writes than an alternative approach that keeps data in memory, but is relatively fault-tolerant, since the data are immediately persisted to long-term storage. With extra preprocessing, use of appropriate system calls, and proper balancing between data chunk size and number of tasks, I/O time is reduced relative to compute time. Random access to the data is enabled by generating an index on the data prior to running comparisons and processing is amortized across many small tasks, several of which can run concurrently. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. URL: http://www.msi.umn.edu"

to one household in 1870 consisting of 2 individual level links). An 1880 household is considered linked if it has at least 4 individual links to a specific 1870 HH and no more than 2 individual links to any other 1870 household. In addition, an 1880 household with 3 individual links to a specific 1870 household and no more than 1 link to any other 1870 household is linked. This initial rule establishes 1,553,420 household links consisting of 6,473,809 individual links.

We have no way of measuring our false positive rate. However, we can look for indirect evidence in the form of inconsistency. Since we do not use place of residence information to establish links, we can use the crude migration rate (defined as not living in the same state and county in both censuses) as a proxy for the false positive rate. In other words, we expect to see fairly consistent rates of living in the same state and county in our linked households regardless of non-demographic characteristics. For example, age and gender are likely to have an effect on migration behavior. But overall similarity or name commonness in our linked records should not have a large effect on migration behavior. All things being equal, if a linked household resides in the same state and county in both enumerations, our confidence that this is a true link increases; a linked record that is also a non-migrant is rarely an error. However, migrants are typically a mix of true links and false positives.[14]

Table 4 gives migration status for the first batch of linked households by various linkage metrics. The top panel gives migration rates based on surname similarity. This is a household measure (and we select the first potential link with a nuclear relationship to represent the household). There appears to be a relationship between surname similarity and being a non-migrant, although the range is relatively small. It is possible that migrants are less likely to have their surnames recorded accurately or consistently, but it is also possible that we are more likely to have false positives as surname similarity decreases (and thus higher levels of migration for linked records with lower surname similarity indicate a higher probability of false positives at lower levels of surname similarity).

[14] This a relative rather than an absolute rule. Some American counties have populations greater than the totals for the least populated states.

The second panel in Table 4 gives migration rates for overall record uniqueness. We construct a uniqueness score based on the number of potential links generated by the given potential link (which is dictated by whether a record has a relatively common combination of given and surname, but also by the overall size of their birthplace block). We take the inverse on the individual level, and calculate the average for the household. For example, if a given record in 1880 has only one potential link to 1870, the individual score = 1/1 (1.0). If a given record has 100 potential links in the 1870 data, then the individual score = 1/100 (0.01). Thus high household scores indicate relative uniqueness. There appears to be a clear relationship between lower household uniqueness scores and migration, although the range is again relatively small. We would not expect different levels of household uniqueness to affect the decision to move between censuses; thus the differential is indicative of higher false positive rates as household uniqueness decreases.

The bottom panel gives the migration rates based on how many records constitute the linked household. Here it is possible that the differential does not indicate false positives, but rather indicates that smaller households (and especially if they were younger couples) were in fact more likely to move between censuses. Nonetheless, we anticipate that there are false positives in our initial set of household links, and that Table 4 provides clues about where we would most likely find them; household links based on the minimum number of individual links, and those comprised of relatively common records and lower overall similarity (either overall age or given and surname similarity).[15]

Table 5 gives the household linkage rate after the first round of rules-based household linking. We only link 15 percent of all 1880 households, but most 1880 households (52 percent) are not at risk of being linked because they contain fewer than three linkable 1880 records (with linkable defined as a having a nuclear relationship to head and being at least 10 years old in 1880). However, we link 32 percent of the eligible households and over 40 percent of the households containing 5 or more linkable records. The table also gives household linkage rates by race and nativity (based on the household head's race and place of birth), with native-born

[15] One possibility explaining the differentials in Table 4 is that migrants are more likely to live in places where the overall enumeration quality is lower; i.e., urban and frontier areas.

whites the most likely to be linked under the rules-based approach. We suspect that non-white groups have lower overall precision (and possibly less stable households). The foreign-born might have lower linkage rates because of lower overall precision (especially in the recording of surname information), but the lower linkage rate could also be caused by the fact that some of them were not present in the United States in 1870 (and we do not have year of immigration information in the 19th century censuses).

Overall, the 32 percent household linkage rate is promising. And, based on our experience linking the St. Louis data, many true household links will be found if we lower the surname similarity threshold (for the initial pass we set the threshold at 0.9). But we also felt that many true household links were in our current potential links universe (i.e., at the 0.9 surname level) but remained unlinked because of ambiguity (multiple conflicting potential household links) or because of low numbers of linkable 1880 members (three or fewer potential links in a potential household link). And it is preferable to establish these links before we try to link households with lower surname similarity.

Linking Households Based on Evidence of Common Neighbors

Eventually we will develop measures to identify the most similar household in cases of ambiguity, but a quick and dirty approach would be to take the non-migrant household if there are multiple potential household links. However, while crude non-migration works well as a diagnostic tool, it is not always a precise linking variable. Large American cities are typically located in a single county. In addition, for some small states a high proportion of all individuals born in the state will reside in the largest city in that state (e.g., Boston Massachusetts; Providence Rhode Island; Baltimore, Maryland). Also some ethnicities tend to cluster in large cities. For example, linking an Irish household living in Boston in both 1870 and 1880 does not provide definitive evidence that this is the true link.

Although we plan on continuing to experiment with the following approach, we currently construct a measure of potential household neighbors. We have 38 million potential household

combinations in our initial potential links file and over 99 percent are the only potential household link for the given combination of 1870 census page and 1880 census page (there are 40 lines per page in 1870 and 50 lines per page in 1880). All things being equal, the presence of two or more potential household links on the same page combinations would increase our confidence that these potential household links are the true links. But many neighbors will not show up on exactly the same census page combination in the two enumerations. Typically households enumerated ten years apart would not be enumerated in the exact sequence even if they had not physically relocated; direct evidence of neighbors depends somewhat on whether or not the enumerator took the same route in two different enumerations. But many non-movers should have common neighbors in the enumerations regardless of whether or not they show up in the same exact sequence.

Currently we calculate the number of potential household links for specific grids consisting of ranges of images in the 1870 and 1880 data. The grid is calculated from the perspective of specific potential household links (thus each combination of 1870 page and 1880 page will have its own unique grid). For example, a potential household link is located on page *x* in 1870 and page *y* in 1880. The grid (for this potential household) is defined as *x* plus/minus 10 (pages) in 1870 and *y* plus/minus 8 (pages) in 1880 (there are 40 lines per page in 1870 and 50 lines per page in 1880; thus the grid, based on this definition, consists of a maximum of 840 records in 1870 and 850 records in 1880). And we want to know how many other potential household links are present in a grid.

Table 6 gives the distribution of the potential household links by the number of potential household neighbors (PHHN) in their respective grid. Approximately 59 percent of the time the specific potential household link will be the only potential household link in the grid (i.e., PHHN = 1). Some of these could be true links (if the household physically moved between censuses and thus does not have any common neighbors), but we suspect that most are false links. The right side of the table gives the PHHN distribution for the rules-based links. The table also gives the relationship between PHHN and migration status for our first batch of linked households; over 70 percent of our initial household links are migrants if they are the only potential link in

their grid. As grid count increases, the household links are increasingly non-migrants.[16]

Figure 8 shows the potential household links contained in a single grid. The reference household is highlighted (the "Turks"), and this is the only potential household link on the specific combination of 1870 page and 1880 page. Their grid is defined as 1870 page +/- 10 pages and 1880 page +/- 8 pages, and there are 12 other potential household links in this grid; thus the PHHN for the reference potential household link (the Turks) is 13 (and the PHHN for other potential households in the figure will have different values for PHHN because the grid moves as we calculate PHHN for other combinations of pages). The figure does not contain page information, but it does contain household serial number information. The serials for both years are set to zero for the reference household in the example (the Turks), with the values for other potential households equal to the difference between their actual household serial and the actual household serial for the reference household. For example, the Kime household has a serial 80 diff = 2, meaning there was one household located between the Turk household and the Kime household in 1880. For 1870 the value is -10, meaning there were nine households between the Turk household and the Kime household in 1870.

A high value for PHHN typically indicates the true household link, but we initially expected some potential household links to have relatively high values but still be a false link. Thus we combine the PHHN with the household uniqueness score discussed earlier. The average uniqueness score for a household ranges from 0 to 1.0, which we convert to an integer (i.e., 1 to 100). Combo score is the product of PHHN and the household uniqueness score. Using Figure 8 as an example, the range of PHHN is 10 to 27, the range of uniqueness score is 2 to 40, and the range for combo score is 26 to 1040.

Without much experimentation we decided to create another batch of linked households based on the combo score. We also decided to include smaller households (i.e., potential households with only two potential links) in the eligible universe. Thus any 1880 household not linked in the first pass (rule-based) that has at least two or more potential links is eligible. If the potential

[16] And the small percentage of potential household links that have high PHHN and are also a migrant are apparently residents of counties that experienced boundary changes between 1870 and 1880.

household link has the maximum number of individual potential links for that household, and the potential household has a combo score of at least 100, we consider it linked. Figure 8 shows how this rule affects the households in this grid. Five of nine households that were initially unlinked are now linked. In addition, it seems that the current combo score threshold is too conservative; all of the remaining unlinked households appear to be true household links.

Again, this first pass only used potential links above 0.9 surname J-W, and our original potential links file contains potential links down to the 0.8 surname level. After flagging linked households from the 0.9 level (both the rules based linked household and the household links based on combo score), we set them aside and include all records from currently unlinked households and repeat the process. Table 7 gives the number of households linked at the end of the 0.8 surname level pass (8 categories); 2.4 million linked households consisting of over 9 million individual links.

Table 7 also gives the non-migration rates for the 8 categories of household links. However, since we used the presence of common neighbors to establish 6 of the 8 categories of linked households, the non-migration rate is not an indication of consistency (at least not as a comparison to the categories of household links (i.e., rules based) where we did not use the presence of common neighbors to establish the link). A comparison of the 1st category (rules-based household links using a 0.9 threshold for surname) to the 5th category (rules-based household links using a 0.8 threshold for surname) shows that the latter category does have a lower rate of non-migration, which could be indicative of a higher rate of false positives. Table 8 replicates the diagnostics shown earlier in Table 4 (which used the 0.9 surname threshold, rules based household links). In general the 2nd batch of rules-based household links have lower rates of non-migration compared to the same categories in Table 4, but overall the range for the 0.8 threshold (rules-based) household links is similar to what we found for the 0.9 threshold (rules-based) household links.

The top panel in Table 9 shows the household linkage rate for all 1880 households by the number of 1880 linkable records. In contrast to Table 5, where we only included the first batch of rules-based household links (using the 0.9 surname threshold), this version includes all of our

current household links. Our overall linkage rate is now over 24 percent, although the linkage rate remains quite a bit lower for the smaller households. The bottom panel of the table restricts the universe to 1880 households at risk of being linked and gives the household linkage rate by race and nativity. Since we eventually were willing to link 1880 households with two linkable records, the only 1880 households not in the linkable universe are the 1880 households that only contain one linkable record. The linkage rate for 1880 linkable households is 26.3 percent, which is lower than the comparable figure in Table 5 (which was 32.4 percent). But the linkable household universe here is inflated by the inclusion of 1880 households containing only two linkable records (which make up almost half of the 1880 households, but are only linked 7 percent of the time). And we suspect that many of the households containing only two linkable records did not exist in 1880 (i.e., younger married couples).

Table 9 gave the number of individual potential links contained in our current batch of linked households. However, this underestimates the number of true links in the linked households; similar to what we found in our St. Louis linked households, we have many currently unlinked records in our linked households that appear to be the true link. Figure 9 shows a few examples of linked households. In the first example we establish the linked household based on the household head and spouse in 1880 (W. N. and Sarah Ann) and one of their children (Ida). However, there are other children in the 1880 household who were also present in the household in 1870. But we were unable to establish these links at the individual level because of birthplace inconsistency (John and Walter had missing birthplace information in 1870, while Howard was born in Iowa in 1870 and Illinois in 1880) and low given name similarity (Cora vs. Carrie for the daughter). And we can assume that eight-year-old Willie in the 1880 household was not yet born in 1870.

The second example shows a household with four explicit links. The three unlinked members in the 1880 household also appear to be in the 1870 household but were unlinked because of excessive differences in expected age (the head was age 28 in 1870 and age 46 in 1880, while the spouse was age 25 in 1870 and age 43 ten years later) and given name (Ann E. vs Analiscia). And we assume the two other members of the 1880 household were not present in 1870

(Minnie I. was 9-years-old in 1880 and was probably unborn in 1870 and John Peterman was a 21-year-old unrelated individual in 1880).

The households in the third example contain five individuals explicitly linked. We were unable to link Elwood C. at the individual level because he/she was enumerated as a male in 1870 and as a female in 1880. Despite the name difference, we are fairly confident that 0-year-old Rosetta J. in 1870 is actually 10-year-old Josephine R. in 1880. It is also possible that 21-year-old Minerva in 1870 is 29-year-old Louiza J. in 1880. But in contrast to Rosetta J.-Josephine R., where transposing first and middle names results in similarity, there is no obvious commonness between the names Minerva and Louiza J.

These examples are not strictly representative, but demonstrate that many of our linked households in 1880 contain unlinked records that also have their true link in the 1870 household. In general, if we establish a linked household, then we expect unlinked records with a nuclear relationship (i.e., head, spouse or child) and age greater or equal to 10 to also be present in the 1870 household. There are categories where this assumption is less likely to be true. For example, an older child in 1880 might have already left home at the time of the 1870 census despite being present for the 1880 enumeration. The youngest linkable children in 1880—ten- or even eleven-year-olds for example—might actually have not been born at the time of the 1870 census (and some of the nine- or even eight-year-olds in 1880 were actually alive in 1870). Spouses with low age or name similarity could be indicative of second marriages. Given these exceptions to our general assumptions about co-residential persistence, we initially adopted a fairly conservative approach to forcing linkages between records with low similarity for key linkage variables.

We will eventually develop a more nuanced approach to deal with this complex problem, but for this paper we adopted a simple procedure based on our household linking rules. First we drop all thresholds, and compare all unlinked household members from the 1870 household to all linkable members of the 1880 household (i.e., we block by household and exclude 1880 records younger than ten and those with a non-nuclear relationship to head). We award one point for each of the following: same sex, same birthplace, age within 4 years of expected age,

and given name similarity greater than 0.9. Using an example from Figure 9, Elwood C in 1880 would get three points for the comparison to Elwood C in 1870 (one point each for given name, age, and birthplace—but not for sex—for a total of three points). The maximum number of points for the forcing procedure is three points (because all of these records failed to link initially because of low similarity or mismatch in at least one of the key linkage variables). If a comparison gets three points, and no other comparison gets at least three points, then we force the link.

Figure 10 shows the forced linking procedure applied to the households from Figure 9. Despite failures to initially link at the individual level, all of the forced links look highly probable with the exception of Louiza J. to Minerva in the Miller household, but even here we would assume that there is a possibility that Louiza J is actually Minerva. The forcing procedure establishes links for 1,183,892 records, or about 71 percent of the unlinked but linkable 1880 records. Some of the current forced links are errors, but we anticipate refining the approach to address the issue of false positives. But it also appears that many of the linkable but still unlinked 1880 records do have their true link residing in the 1870 household. In the first example in Figure 11 we have one unlinked record in 1880 household, 21-year-old John W., who is probably 11-year-old Walker in the 1870 household. In addition to the low similarity between the given names, the two records have mismatched birthplaces. The second household in Figure 8 shows an extreme example of ambiguity in the forcing process. The 1870 household contains two 13-year-old males with given names of Abda F and Abba F. Despite the presence of two males in the 1880 household who were 23 years old, the forcing procedure cannot determine the correct link (i.e., because either could be Felix or Festus in the 1880 household).

A review of our forced links discloses that low given name similarity was the primary reason records were not linked as part of the initial household linking process. We anticipate improving our given name standardization process, which would increase the given name similarity for some of these records (and thus increasing the probability that these records will be compared to their true link at the individual level). But as seen in previous examples, many true links with low given name similarity were enumerated with distinctly different given names in the two

enumerations. We have 41,472 males with the given name of Henry in 1880 in the group of forced links. Approximately 45 percent also had a given name of Henry in 1870, with a much smaller percentage having names or variants that could be standardized as Henry (like Harry or Harvey). But most have given names that are definitely not Henry. For example, we have 1,714 Henry-William combinations and almost 40 percent of the Williams have a middle initial of 'H' in 1870. Many of the forced links that have low given name similarity also have a middle initial that increases confidence in the link, but a majority do not have middle name or initial information.

Although we have not finished developing a comprehensive approach to the household linking process, we have begun to assess the range of precision for our key linkage variables. Tables 10 and 11 give the range of imprecision for our current linked data, which includes both explicit and forced links. In general, precision levels are higher for our complete-count household links compared to the St. Louis household links (see Tables 2 and 3). However, the ability to make strict comparisons is limited by a number of factors. For example, approximately 11 percent of our complete-count household links have surname similarity below the 0.9 level. The comparable figure for the St. Louis household links was 28 percent. We expect the proportion of complete-count links where this is true to increase as we lower our surname similarity threshold in the potential link selection process; i.e., some of the currently unlinked households are unlinked precisely because all household members have low surname similarity to their true links.[17] The relatively closed universe of the two enumerations of St. Louis, along with the availability of street and house number information, allowed us to link smaller households or households with low levels of similarity; in other words, we were able to get closer to the bottom of the barrel than we will ever be able to do with households enumerated 10 years apart.[18]

[17] And this same logic would apply higher levels of imprecision for place of birth in the linked St. Louis data; we blocked by place of birth in constructing the complete-count individual level links; we suspect that some of the currently unlinked households are unlinked because most or all household members have mismatched birthplace information.

[18] It is also possible that the first enumeration of St. Louis was an example of a shoddily taken census, while the second enumeration—which used a reference date five months prior to the date of the recount—introduced imprecision in recording information for individuals who had left the city. A more

Going Forward

Our current linkage project will eventually include links covering the 1850, 1860, 1870, and 1880 complete-count census databases. Based on our initial results, we are fairly confident that we will link a fairly sizable proportion of 1880 records to all three of the previous decennial censuses using the household linking approach (year of birth permitting). Going forward, some of our work will focus on better methods of identifying and eliminating false positives. The use of additional evidence derived from common neighbors and co-resident kin implies that we have a higher standard; our (unachievable) goal is to never make an incorrect household link.

Quality control can be tedious (and demoralizing when it uncovers a logical flaw or two) but it is a necessary part of the process. And we will continue to evaluate quality issues as we proceed to create additional household links in the 1870-1880 data. Some households will never be linked, but we hope to ultimately double our current household linkage rate. Some of our optimism is based on our experience with St. Louis; although we already saw diminishing returns in our second pass using a lower surname threshold, we anticipate finding additional household links below a 0.8 surname similarity threshold. We also suspect a significant number of households remain unlinked because of birthplace inconsistency. Analysis of our forced links—often forced due to low given name similarity—will result in improved given name standardizations (or aliases). We will also refine our measurement of household uniqueness and neighbor calculations. The PHHN (i.e., common neighbors) approach needs some calibration, but promises to link many additional households.

Although we anticipate continuing to find households based on the process of elimination, some households will remain unlinked because they did not exist in the previous census. A common example would be older sons in the 1870 census who leave home and get married;

---

charitable interpretation would be that imprecision found in St. Louis in 1880 would be representative of enumerations in large American cities in the nineteenth century, and that we would expect greater precision for individuals enumerated in small towns and rural areas. Whether or not the imprecision in the linked St. Louis data is an outlier is an interesting issue, but we nonetheless also find relatively high imprecision in the complete-count linked data.

thus they will be living with a spouse and children under the age of 10 in the 1880 census. However, if we can link their household of origin in 1870 to an 1880 household and verify that they were absent from that 1880 linked household, then we are more confident in creating a household link absent the presence of any corroborative kin. Figure 12 gives an example based on the grid example (i.e., Figure 8). Figure 12 gives the entire 1870 and 1880 households for the Mathis household, and we can see that the four oldest sons in the 1870 household were not present when the household was enumerated in 1880. Although this household was not the reference point for this specific grid, we can identify what appears to be two of the absent sons (with their wives and children) in this grid, and they are located in close proximity to the 1880 household that contains their parents and younger siblings. We do not know how many of these types of households we will be able to link, but we believe the use of common neighbor information greatly expands our ability to confidently verify linkage decisions.

The household links will be useful for some types of analysis (e.g., where the relevant unit of study consists of married couples or related groups) but they will definitely be biased. But we also anticipate continuing to construct individual (minimal bias) level links. Here the household links can be used in two primary ways. They can be used as a verification set for links established at the individual level. And the household links are an important part of this process because of the presence of the forced links (i.e., links not initially present in our potential links file, typically because of low similarity or mismatch in at least one linkage variable). For the most part, these records will rarely be linked by individual-level classifiers. An accurate estimation of the false positive rate requires establishing all true links (despite low similarity or mismatched linkage variables) and the only way to do this is to use a maximum amount of information (i.e., the household linkage process).

One issue with the household links as a verification set is that they will not cover the entire population of individual-level links. This is true (i.e., some individual level links will not be verified because we did not link their household), but we suspect that our individual-level links will contain a disproportionately high number of links established at the household level. This is because the inability to be linked at the household level implies a number of conditions or

characteristics at the individual level.

We would (theoretically) expect similar levels of linkage variable precision for some groups of individuals not linked at the household level (compared to those linked at the household level). This would include the sons who transition to marriage and the establishment of their own households between censuses. This would also include households with relatively common names (combined with large birthplace blocks) that remain unlinked because of ambiguity (especially if they lack common neighbors in the two censuses). These two groups should have overall precision comparable to the household linked set (although ambiguous records at the household level will also be ambiguous at the individual level).

But many members of the household linking resistance are harder cases. Under-enumeration in the 19th century was fairly high (possibly as high as five percent). We also have some 1880 households that were not in the country in 1870 (and from a record linkage perspective they are similar to under-enumerated records). We are still in the speculative stage, but in addition to households missing from one enumeration or the other, it seems plausible that at least that many are underwater (i.e., we will never be able to link them even at the household level because of low similarity or mismatch for one or more linkage variables). Less extreme , but still problematic, is the sizable 19th century unrelated population. Rarely will they be co-resident with the same people in both censuses. And the accuracy of their names, age and birthplace will undoubtedly vary, but we suspect that the quality of information for unrelated individuals enumerated in the 19th century is relatively poor.

So the part of the 1880 population that is not part of the household linked universe will consist of a higher proportion records that either do not have a true link or have a relatively low similarity true link. It is possible that an individual-level classifier trained and tested on the household links (and calibrated to get an optimal combination of linkage and false positive rates) will not perform nearly as well on the set of records that were not linked at the household level (primarily because this universe contains many records without true links, and some of these records get linked randomly at lower levels of classifier-approved thresholds).

But maybe this really does not matter. It is possible that a well-designed individual level

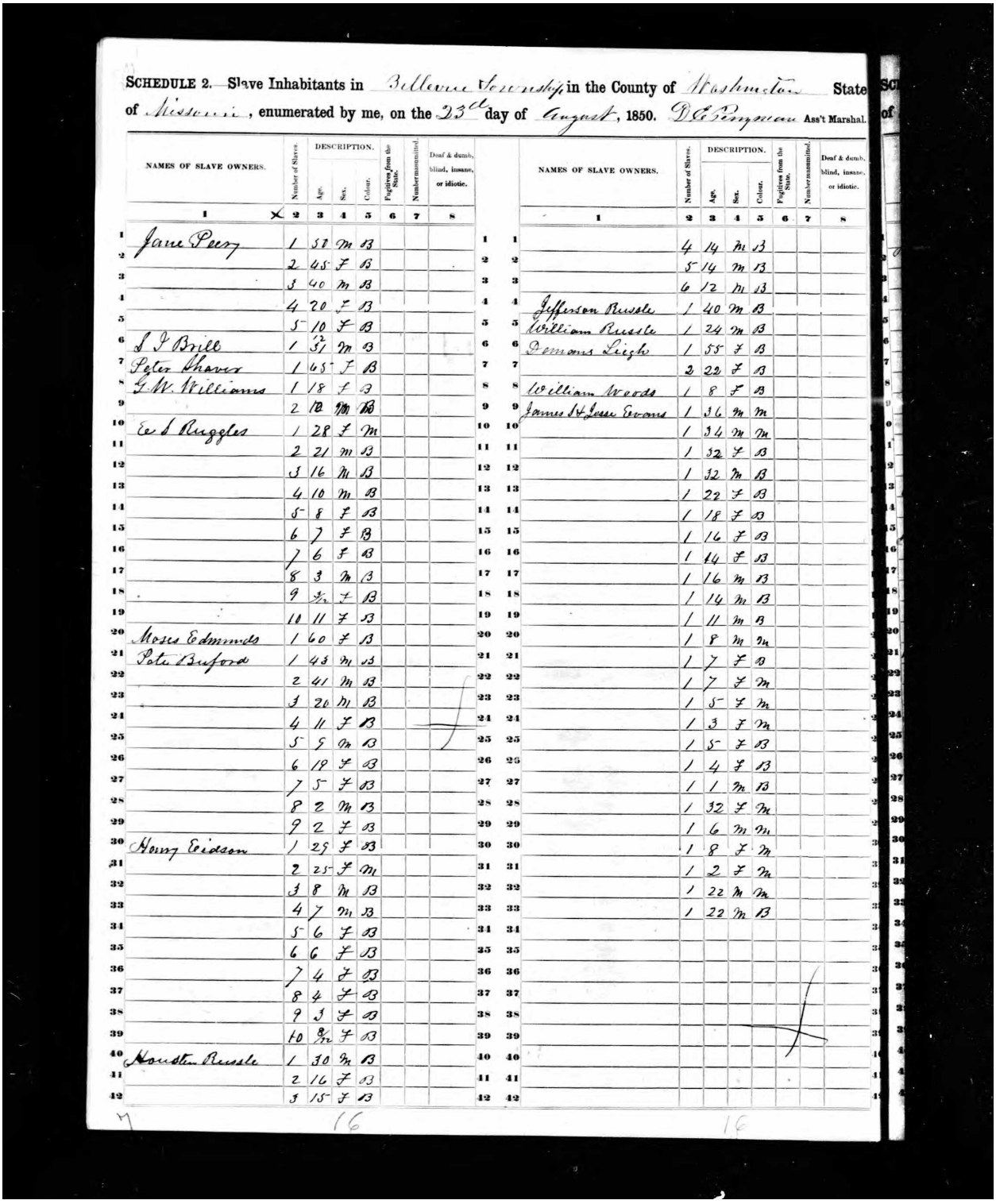
classifier achieves “acceptably” low false positive rates, in that the presence of some incorrectly linked records does not significantly affect research results. This has been the standard default position for previous linkage projects, but it has mostly been based on speculative optimism (i.e., faith-based record linkage). Ultimately we hope to produce a fairly comprehensive set of verified household links for 1850 through 1880. We will also produce linked data at the individual level. Thus we will have three different linked sets: household links; individual-level links; and individual-level links with the false positives removed (i.e., false positives identified by comparing the individual-level links to the household links). We plan on experimenting with different types of analysis (e.g. female-labor force participation, social-economic mobility, etc.) to see if we get different results based on which linked set we use.

Figure 1. Primary and Secondary Links, 1870-1880 Male-Only Sample

link type	fname 70	lname 70	age 70	relate 70	fname 80	lname 80	age 80	relate 80
unlinked	JOHN	MCHUGH	50	head				
unlinked	REBECCA	MCHUGH	37	spouse				
primary	**HENRY**	**MCHUGH**	**14**	**child**	**HENRY**	**MCHUGH**	**25**	**child**
unlinked	JAMES E	MCHUGH	3	child				
unlinked	JANE R	MCHUGH	0	child				
unlinked					CATHARINE	MCHUGH	64	head
unlinked					ELLEN	MCHUGH	38	child
unlinked					EDWARD	MCHUGH	35	child
unlinked					MARY F.	MCHUGH	27	child
unlinked					MARY E.	MCHUGH	16	grandchild
unlinked					EDWARD J.	MCHUGH	12	grandchild
link type	**fname 70**	**lname 70**	**age 70**	**relate 70**	**fname 80**	**lname 80**	**age 80**	**relate 80**
primary	**JAMES**	**FELKINS**	**61**	**head**	**JAMES H.**	**FELKIN**	**71**	**head**
unlinked	MARTHA	FELKINS	53	spouse				
secondary	NANCY	FELKINS	35	child	NANCY	FELKIN	42	child
unlinked	BUNELL	FELKINS	28	child				
secondary	ELISABETH	FELKINS	16	child	ELISIBETH	FELKIN	23	child
secondary	PAIKNEY	FELKINS	14	child	PINKNY	FELKIN	22	child
unlinked					MATILDA	FELKIN	67	spouse

Notes:
fname70 = first name in 1870
lname70 = last name in 1870
age70 = age in 1870
relate70 = imputed relationship to head in 1870
fname80 = first name in 1880
lname80 = last name in 1880
age80 = age in 1880
relate80 = imputed relationship to head in 1880

Figure 2. 1850 Slave Schedule



SCHEDULE 2. Slave Inhabitants in Belleview Township in the County of Washington State of Missouri, enumerated by me, on the 23d day of August, 1850. D. C. Pingman Ass't Marshal.

	NAMES OF SLAVE OWNERS.	Number of Slaves.	DESCRIPTION. Age.	Sex.	Colour.	Fugitives from the State.	Number manumitted.	Deaf & dumb, blind, insane, or idiotic.
	1	2	3	4	5	6	7	8
1	Jane Peery	1	58	M	B			
2		2	45	F	B			
3		3	40	M	B			
4		4	20	F	B			
5		5	10	F	B			
6	S J Brill	1	51	M	B			
7	Peter Shaver	1	65	F	B			
8	G W Williams	1	18	F	B			
9		2	12	M	B			
10	Ed S Ruggles	1	28	F	M			
11		2	21	M	B			
12		3	16	M	B			
13		4	10	M	B			
14		5	8	F	B			
15		6	7	F	B			
16		7	6	F	B			
17		8	3	M	B			
18		9	1/2	F	B			
19		10	11	F	B			
20	Moses Edmonds	1	60	F	B			
21	Pete Buford	1	43	M	B			
22		2	41	M	B			
23		3	20	M	B			
24		4	11	F	B			
25		5	5	M	B			
26		6	18	F	B			
27		7	5	F	B			
28		8	2	M	B			
29		9	2	F	B			
30	Henry Eidson	1	25	F	B			
31		2	25	F	M			
32		3	8	M	B			
33		4	7	M	B			
34		5	6	F	B			
35		6	6	F	B			
36		7	4	F	B			
37		8	4	F	B			
38		9	3	F	B			
39		10	8/12	F	B			
40	Houston Russle	1	30	M	B			
41		2	16	F	B			
42		3	15	F	B			

	NAMES OF SLAVE OWNERS.	Number of Slaves.	DESCRIPTION. Age.	Sex.	Colour.	Fugitives from the State.	Number manumitted.	Deaf & dumb, blind, insane, or idiotic.
	1	2	3	4	5	6	7	8
1		4	14	M	B			
2		5	14	M	B			
3		6	12	M	B			
4	Jefferson Russle	1	40	M	B			
5	William Russle	1	24	M	B			
6	Demons Leigh	1	55	F	B			
7		2	22	F	B			
8	William Woods	1	8	F	B			
9	James & Jesse Evans	1	36	M	M			
10		1	34	M	M			
11		1	32	F	B			
12		1	32	M	B			
13		1	22	F	B			
14		1	18	F	B			
15		1	16	F	B			
16		1	14	F	B			
17		1	16	M	B			
18		1	14	M	B			
19		1	11	M	B			
20		1	8	M	M			
21		1	7	F	B			
22		1	7	F	M			
23		1	5	F	M			
24		1	3	F	M			
25		1	5	F	B			
26		1	4	F	B			
27		1	1	M	B			
28		1	32	F	M			
29		1	6	M	M			
30		1	8	F	M			
31		1	2	F	M			
32		1	22	M	M			
33		1	22	M	B			
34								
35								
36								
37								
38								
39								
40								
41								
42								

16 16

Figure 3a. Potential matches for John O'Donnell, St. Louis 1880

fname1	lname1	age1	fname2	lname2	age2
JOHN	O'DONNELL	43	JOHN	O'DONNELL	45
			JOHN	O'DONNELL	45
			JOHN	O'DONNELL	46

Figure 3b. Households containing potential links for John O'Donnell, St. Louis 1880

fname1	lname1	age1	fname2	lname2	age2	p_link	sum_p_link
JOHN	O'DONNELL	43	JOHN	O'DONNELL	46	1	5
MARY	O'DONNELL	43	MARY	O'DONNELL	44	1	5
MICHAEL	O'DONNELL	15	MICHAEL	O'DONNELL	16	1	5
PATRICK	O'DONNELL	9	PATRICK	O'DONNELL	9	1	5
BRIDGET	O'DONNELL	6	BRIDGET	O'DONNELL	5	1	5
JOHN	O'DONNELL	43	JOHN	O'DONNELL	45	1	1
MARY	O'DONNELL	43	ELLEN	O'DONNELL	40	0	1
MICHAEL	O'DONNELL	15	JULIA	O'DONNELL	12	0	1
PATRICK	O'DONNELL	9				0	1
BRIDGET	O'DONNELL	6				0	1
JOHN	O'DONNELL	43	JOHN	O'DONNELL	45	1	1
MARY	O'DONNELL	43	MARGRET	O'DONNELL	39	0	1
MICHAEL	O'DONNELL	15	JOHN	O'DONNELL	19	0	1
PATRICK	O'DONNELL	9	ELIZEBETH	O'DONNELL	14	0	1
BRIDGET	O'DONNELL	6	FRANCIS	O'DONNELL	12	0	1
			WILLIAM	O'DONNELL	4	0	1

Notes:
fname1 = first name in first enumeration
lname1 = last name in first enumeration
age1 = age in first enumeration
fname2 = first name in second enumeration
lname2 = last name in second enumeration
age2 = age in second enumeration
p_link = indicates a potential link between individuals listed
sum_p_link = the sum of potential links between specific households.

Figure 4. A linked household, St. Louis 1880

fname1	lname1	age1	fname2	lname2	age2	p_link	sum_p_link	J-W fname	J-W lname
AUTONIA	STROUBEL	52	ANTON	STRUBE	53	0	4	0.69	0.94
ELIZABETH	STROUBEL	42	ELIZA	STRUBE	42	1	4	0.91	0.94
ANNIE	STROUBEL	19	ANNIE	STRUBE	14	0	4	1.00	0.94
MINNIE	STROUBEL	12	MINNIE	STRUBE	13	1	4	1.00	0.94
LOUISA	STROUBEL	10	LOUISE	STRUBE	11	1	4	0.93	0.94
DORETTA	STROUBEL	4	DORA	STRUBE	5	1	4	0.90	0.94

Notes:
fname1 = first name in first enumeration
lname1 = last name in first enumeration
age1 = age in first enumeration
fname2 = first name in second enumeration
lname2 = last name in second enumeration
age2 = age in second enumeration
p_link = indicates a potential link between individuals listed
sum_p_link = the sum of potential links between specific households.
J-W fname = Jaro-Winkler similarity score for first name strings
J-W lname = Jaro-Winkler similarity score for last name strings

Figure 5. Linked households, St. Louis 1880

fname1	lname1	age1	fname2	lname2	age2	fname J-W	lname J-W
MATHEW	BURGHERDT	40	MATHEW	BURKHART	47	1.00	0.86
ELIZABETH	BURGHERDT	40	ELIZABETH	BURKHART	40	1.00	0.86
CATHERINE	BURGHERDT	12	KATE	BURKHART	11	0.69	0.86
ELIZABETH	BURGHERDT	9	ELIZABETH	BURKHART	9	1.00	0.86
WILLIAM	BURGHERDT	4	WILLIAM	BURKHART	4	1.00	0.86

fname1	lname1	age1	fname2	lname2	age2	fname J-W	lname J-W
DAVID	FITZGERALD	48	DAVE	VETZGURA	45	0.85	0.67
MARY	FITZGERALD	34	MARY	VETZGURA	36	1.00	0.67
ANNIE	FITZGERALD	12	ANNA	VETZGURA	12	0.85	0.67
KATE	FITZGERALD	10	KATE	VETZGURA	11	1.00	0.67
ANDREW	FITZGERALD	5	ANDREW	VETZGURA	6	1.00	0.67
NORA	FITZGERALD	2	MONORA	VETZGURA	3	0.81	0.67
RICHARD	FITZGERALD	0	RICHARD	VETZGURA	0	1.00	0.67

fname1	lname1	age1	fname2	lname2	age2	fname J-W	lname J-W
FRANK	KLAESER	60	F. H.	CLASSEN	60	0.76	0.63
BRIDGET	KLAESER	56	BRIDGET	CLASSEN	58	1.00	0.63

fname1	lname1	age1	fname2	lname2	age2	fname J-W	lname J-W
CAROLINE	SCHWARTZ	60	CATHERINE	SCHMARG	60	0.76	0.85
AUGUSTA	SCHWARTZ	26	AUGUSTE	SCHMARG	25	0.94	0.85

Notes:
fname1 = first name in first enumeration
lname1 = last name in first enumeration
age1 = age in first enumeration
fname2 = first name in second enumeration
lname2 = last name in second enumeration
age2 = age in second enumeration
fname J-W = Jaro-Winkler similarity score for first name strings
lname J-W = Jaro-Winkler similarity score for last name strings

Figure 6. Selected surname combinations in the linked data, St. Louis 1880

lname1	lname2	J-W	NYSIIS	Double meta	match1	match2	match3
COBB	COBBS	0.96	1	1	1	1	1
MAIER	MIER	0.94	1	1	1	0	0
BLOCH	BLOCK	0.92	0	1	1	1	1
SCHLEGEL	SCHLAEGD	0.90	0	0	1	1	1
KAMPF	KEMPF	0.88	1	1	1	0	0
LAMPE	LAMPKING	0.86	0	0	1	1	1
NOOTEN	NEWTON	0.84	1	1	1	0	0
BORGERS	BORSGUS	0.82	0	0	1	1	1
GERRAN	GUERIN	0.80	1	1	1	0	0
THORNALLY	TOMALLI	0.78	0	0	1	0	0
BOETTE	BOOTH	0.76	0	0	1	1	0
BROCHRIGT	BROOKLINE	0.74	0	0	1	1	1
HEFFNER	HOFFMANN	0.72	0	0	1	0	0
RUBIN	LUBIER	0.70	0	0	0	0	0
GOTTMAYER	KOLMEYER	0.66	0	0	0	0	0
THOMA	TGNAZ	0.64	0	0	1	0	0
BOICE	NOYES	0.60	0	0	0	0	0
KOOKENBERG	GUEGGESBERY	0.55	0	0	0	0	0
KEEVIL	DRISCOLL	0.53	0	0	0	0	0

Notes:
0/1 indicates that the name combination would not match/match for phonetic/matching codes
J-W = Jaro-Winkler similarity score for last name combination
NYSIIS = whether the name combination has a NYSIIS match
Double meta = whether the name combination has a doublemetaphone match
Match1 = whether the name combination matches on first letter
Match2 = whether the name combination matches on first 2 letters
Match3 = whether the name combination matches on first 3 letters

Figure 7. Examples of first name mismatches, St. Louis 1880

fname1	lname1	age1	fname2	lname2	age2
C. ALBERT	RAHNER	24	ALBERT	RAHNER	24
BERNARD	HILL	23	C. BERNARD	HILL	22
BRIDGET	CARTEN	33	M. BRIDGET	CARTEN	34
C. AMELIA	SHEERER	32	AMALIA C.	SHERER	35

fname1	lname1	age1	fname2	lname2	age2
BAYARD N.	ABBOTT	4	NELSON	ABBOTT	3
BELLE	HILTON	5	IDA B.	HILTON	5
DAVID	SUTTMUELLER	47	JOHN D.	SULTMULLER	48
ELLEN	ROBINS	2	MARY E.	ROBBINS	3

fname1	lname1	age1	fname2	lname2	age2
THECKLA	NIEHAUS	57	MARY	NIEHAUS	57
TIMOTHY	LYNCH	17	BUD	LYNCH	18
LILLY	WALSER	0	GRACE	WALSER	0
WILLIAM	PERRIN	0	EUGENE	PERRIN	0

Notes:
fname1 = first name in first enumeration
lname1 = last name in first enumeration
age1 = age in first enumeration
fname2 = first name in second enumeration
lname2 = last name in second enumeration
age2 = age in second enumeration

Figure 8. Sample Neighbor Grid, Livingston County, Illinois, 1870-1880 Complete Count

rules only	rules plus neighbors	serial 70 diff	Serial 80 diff	fname70	lname70	age 70	fname80	lname80	age 80	neighbor count (PHHN)	unique score	combo score
		-12	-67	MARY	WOODRUFF	34	MARY	WOODRUFF	43	10	5	50
		-12	-67	ALPHONSO	WOODRUFF	15	ALPHONSO	WOODRUFF	25	10	5	50
linked	***	-19	-66	JOHN	ARNOLD	52	JOHN	ARNOLD	63	10	6	60
linked	***	-19	-66	LOUISA	ARNOLD	50	LOUISA	ARNOLD	61	10	6	60
linked	***	-19	-66	WILLIAM	ARNOLD	26	WILLIAM	ARNOLD	35	10	6	60
linked	***	-19	-66	FRANKLIN	ARNOLD	17	FRANKLIN	ARNOLD	27	10	6	60
	linked	-13	-64	MARY	BUSSARD	50	MARY	BUZZARD	61	10	12	120
	linked	-13	-64	OZILLA	BUSSARD	25	ROZILLA	BUZZARD	36	10	12	120
	linked	-13	-64	WILLIAM	BUSSARD	19	WILLIAM	BUZZARD	28	10	12	120
		-25	-17	GEORGE	CHRITTEN	16	GEO	CRITTEN	26	17	3	51
		-25	-17	WILLIAM	CHRITTEN	43	WILLIAM	CRITTEN	55	17	3	51
		0	**0**	**SARAH**	**TURK**	**39**	**SARAH**	**TURK**	**48**	**13**	**2**	**26**
		0	**0**	**EVALINE**	**TURK**	**4**	**EVALIENE**	**TURK**	**13**	**13**	**2**	**26**
		-10	2	JOSEPH	KIME	35	JOSEPH	KIME	48	17	2	34
		-10	2	SUSAN	KIME	31	SUSAN	KIME	39	17	2	34
linked	***	2	14	D	DEFENBAUGH	37	DAVID	DEFFENBAUGH	46	19	36	684
linked	***	2	14	ISABELL	DEFENBAUGH	37	ISABELLA	DEFFENBAUGH	48	19	36	684
linked	***	2	14	GEORGIANNA	DEFENBAUGH	9	GEORGANNA	DEFFENBAUGH	19	19	36	684
linked	***	-7	28	SAMUEL	THOMSON	52	SML	THOMPSON	62	17	4	68
linked	***	-7	28	HARIET	THOMSON	47	HARRIET	THOMPSON	58	17	4	68
linked	***	-7	28	EDGAR	THOMSON	5	EDGAR	THOMPSON	15	17	4	68
linked	***	-29	43	CALEB	MATHIS	46	CALEB	MATHIS	56	22	18	396
linked	***	-29	43	SOFLENA	MATHIS	43	SOPLENA	MATHIS	53	22	18	396
linked	***	-29	43	SOFLENA	MATHIS	9	SOPLENA	MATHIS	19	22	18	396
linked	***	-29	43	WILLIAM	MATHIS	7	WILLIAM	MATHIS	16	22	18	396
linked	***	-29	43	HELLAND	MATHIS	2	HOLLAND	MATHIS	12	22	18	396
	linked	-47	44	SAERTIS	SMITH	51	LAERTES	SMITH	61	27	8	216
	linked	-47	44	LOUISA	SMITH	48	LOUISA	SMITH	59	27	8	216
	linked	-38	47	ANDREW	WRIGHT	55	ANDREW	WRIGHT	65	26	14	364
	linked	-38	47	EMELINE	WRIGHT	44	EMMELINE	WRIGHT	54	26	14	364
	linked	-41	53	WILLIAM	BOATMAN	52	WILLIAM	BOATMAN	62	26	7	182
	linked	-41	53	ELENOR	BOATMAN	50	ELEANOR	BOATMAN	60	26	7	182
	linked	-40	75	CHARLES	THRASHER	55	CHARLES	THRASHER	65	26	40	1040
	linked	-40	75	MARY	THRASHER	43	MARY	THRASHER	52	26	40	1040
	linked	-40	75	THANKFUL	THRASHER	6	THANKFUL	THRASHER	16	26	40	1040

Figure 9. Linked Household Examples, 1870-1880 Complete-Count

linked80	name1_70	name2_70	name1_80	name2_80	relate80	age70	age80	sex70	sex80	bpl70	bpl80
linked	W N	AYERS	W. N.	AYERS	head	45	54	male	male	Ohio	Ohio
linked	SARAH	AYERS	SARAH ANN	AYERS	spouse	41	51	female	female	Vermont	Vermont
unlinked			JOHN	AYERS	child		24		male		Washington
unlinked			WALTER	AYERS	child		22		male		Washington
unlinked			HOWARD	AYERS	child		19		male		Illinois
linked	IDA	AYERS	IDA	AYERS	child	6	16	female	female	Illinois	Illinois
unlinked			CARRIE	AYERS	child		14		female		Iowa
unlinked			WILLIE	AYERS	child		8		male		Arkansas
	JOHN	AYERS				14		male		missing	
	WALTER	AYERS				12		male		missing	
	HOWARD	AYERS				9		male		Iowa	
	CORA	AYERS				4		female		Iowa	
linked80	**name1_70**	**name2_70**	**name1_80**	**name2_80**	**relate80**	**age70**	**age80**	**sex70**	**sex80**	**bpl70**	**bpl80**
unlinked			HENRY C.	CUTTING	head		46		male		Ohio
unlinked			CORDELIA	CUTTING	spouse		43		female		Vermont
linked	LUCY A	CUTTING	LUCY	CUTTING	child	10	20	female	female	Ohio	Ohio
linked	WILLIAM H	CUTTING	WILLIAM K.	CUTTING	child	7	19	male	male	Ohio	Ohio
unlinked			ANALISCIA	CUTTING	child		17		female		Ohio
linked	SAMUEL J	CUTTING	SAMUEL J.	CUTTING	child	4	14	male	male	Ohio	Ohio
linked	CORA A	CUTTING	CORA A.	CUTTING	child	1	11	female	female	Ohio	Ohio
unlinked			MINNIE I.	CUTTING	child		9		female		Ohio
unlinked			JOHN	PETERMAN	unrelated		21		male		Ohio
	HENRY	CUTTING				28		male		Ohio	
	CORDELIA	CUTTING				25		female		Vermont	
	ANN E	CUTTING				5		female		Ohio	
linked80	**name1_70**	**name2_70**	**name1_80**	**name2_80**	**relate80**	**age70**	**age80**	**sex70**	**sex80**	**bpl70**	**bpl80**
linked	NATHAN	MILLER	NATHAN	MILLER	head	54	63	male	male	Ohio	Ohio
linked	MARGARET D	MILLER	MARGARET D.	MILLER	spouse	53	63	female	female	Ohio	Ohio
linked	CHARLES H	MILLER	CHARLE N.	MILLER	child	10	20	male	male	Ohio	Ohio
linked	SARAH J	MILLER	SARAH M.	MILLER	child	13	23	female	female	Ohio	Ohio
unlinked			ELWOOD C.	MILLER	child		17		female		Ohio
unlinked			LOUIZA J.	MILLER	child		29		female		Ohio
linked	JOHN W	MILLER	JOHN W.	MILLER	child	2	12	male	male	Ohio	Ohio
unlinked			JOSEPHINE R.	MILLER	child		10		female		Ohio
	ELWOOD C	MILLER				7		male		Ohio	
	MINERVA	MILLER				21		female		Ohio	
	ROSETTA J	MILLER				0		female		Ohio	

Figure 10. Linked Household Examples After Forced Linking Process, 1870-1880 Complete-Count

linked80	name1_70	name2_70	name1_80	name2_80	relate80	age70	age80	sex70	sex80	bpl70	bpl80
explicit	W N	AYERS	W. N.	AYERS	head	45	54	male	male	Ohio	Ohio
explicit	SARAH	AYERS	SARAH ANN	AYERS	spouse	41	51	female	female	Vermont	Vermont
explicit	IDA	AYERS	IDA	AYERS	child	6	16	female	female	Illinois	Illinois
forced	CORA	AYERS	CARRIE	AYERS	child	4	14	female	female	Iowa	Iowa
forced	JOHN	AYERS	JOHN	AYERS	child	14	24	male	male	missing	Washington
forced	WALTER	AYERS	WALTER	AYERS	child	12	22	male	male	missing	Washington
forced	HOWARD	AYERS	HOWARD	AYERS	child	9	19	male	male	Illinois	Washington
unlinked			WILLIE	AYERS	child		8		male		Arkansas

linked80	name1_70	name2_70	name1_80	name2_80	relate80	age70	age80	sex70	sex80	bpl70	bpl80
forced	HENRY	CUTTING	HENRY C.	CUTTING	head	28	46	male	male	Ohio	Ohio
forced	CORDELIA	CUTTING	CORDELIA	CUTTING	spouse	25	43	female	female	Vermont	Vermont
explicit	LUCY A	CUTTING	LUCY	CUTTING	child	10	20	female	female	Ohio	Ohio
explicit	WILLIAM H	CUTTING	WILLIAM K.	CUTTING	child	7	19	male	male	Ohio	Ohio
forced	ANN E	CUTTING	ANALISCIA	CUTTING	child	5	17	female	female	Ohio	Ohio
explicit	SAMUEL J	CUTTING	SAMUEL J.	CUTTING	child	4	14	male	male	Ohio	Ohio
explicit	CORA A	CUTTING	CORA A.	CUTTING	child	1	11	female	female	Ohio	Ohio
unlinked			MINNIE I.	CUTTING	child		9		female		Ohio
unlinked			JOHN	PETERMAN	unrelated		21		male		Ohio

linked80	name1_70	name2_70	name1_80	name2_80	relate80	age70	age80	sex70	sex80	bpl70	bpl80
explicit	NATHAN	MILLER	NATHAN	MILLER	head	54	63	male	male	Ohio	Ohio
explicit	MARGARET D	MILLER	MARGARET D.	MILLER	spouse	53	63	female	female	Ohio	Ohio
explicit	SARAH J	MILLER	SARAH M.	MILLER	child	13	23	female	female	Ohio	Ohio
explicit	CHARLES H	MILLER	CHARLE N.	MILLER	child	10	20	male	male	Ohio	Ohio
forced	ELWOOD C	MILLER	ELWOOD C.	MILLER	child	7	17	male	female	Ohio	Ohio
forced	MINERVA	MILLER	LOUIZA J.	MILLER	child	21	29	female	female	Ohio	Ohio
explicit	JOHN W	MILLER	JOHN W.	MILLER	child	2	12	male	male	Ohio	Ohio
forced	ROSETTA J	MILLER	JOSEPHINE R.	MILLER	child	0	10	female	female	Ohio	Ohio

Figure 11. Linked Household Examples, 1870-1880 Complete-Count

name1_70	name2_70	name2_80	name1_80	relate80	age70	age80	sex70	sex80	bpl70	bpl80
WILLIAM	FENTON	WM. H.	FENTON	head	35	46	male	male	New Jersey	New Jersey
CORDELIA	FENTON	CORDELIA	FENTON	spouse	33	44	female	female	DC	DC
		JOHN W.	FENTON	child		21		male		DC
SAMUEL	FENTON	SAMUEL	FENTON	child	9	19	male	male	DC	DC
EMMA	FENTON	EMMA	FENTON	child	7	17	female	female	DC	DC
WILLIAM	FENTON	WILLIAM	FENTON	child	5	15	male	male	DC	DC
MARY	FENTON	MAY	FENTON	child	3	13	female	female	DC	DC
BESSIE	FENTON	BESSIE	FENTON	child	1	10	female	female	DC	DC
WALKER	FENTON				11		male		Virginia	
IDA	WALKER				16		female		DC	
LOUISA	BROWN				27		female		Maryland	

name1_70	name2_70	name2_80	name1_80	relate80	age70	age80	sex70	sex80	bpl70	bpl80
WILLIAM J	CANTRELL	W. J.	CANTRELL	head	56	67	male	male	Georgia	Georgia
AMANDA	CANTRELL	AMANDA	CANTRELL	spouse	43	54	female	female	Georgia	Georgia
		FELIX	CANTRELL	child		23		male		Georgia
		FESTUS	CANTRELL	child		23		male		Georgia
MARGARET A	CANTRELL	MAGGIE	CANTRELL	child	11	20	female	female	Georgia	Georgia
JOHN	CANTRELL	JOHN	CANTRELL	child	5	13	male	male	Georgia	Georgia
		EVA	CANTRELL	child		8		female		Georgia
JAMES R	CANTRELL				17		male		Georgia	
MARGARET F	CANTRELL				15		female		Georgia	
ABDA F	CANTRELL				13		male		Georgia	
ABBA F	CANTRELL				13		male		Georgia	
SUSAN	CANTRELL				38		female		Virginia	
CHARLES	CANTRELL				7		male		Georgia	
ARMSTEAD	CANTRELL				1		male		Georgia	

Figure 12. Linking Older Sons, Livingston County, Illinois, 1870-1880 Complete Count

fname70	lname70	age70		fname80	lname80	age80	serial80	serial80 diff
CALEB	MATHIS	46		CALEB	MATHIS	56	*799	0
SOPLENA	MATHIS	43		SOFLENA	MATHIS	53		
SOPLENA	MATHIS	9		SOFLENA	MATHIS	19		
WILLIAM	MATHIS	7		WILLIAM	MATHIS	16		
HOLLAND	MATHIS	2		HELLAND	MATHIS	12		
GEORGE	MATHIS	19						
JAMES	MATHIS	17						
ELBERT	MATHIS	13						
EUGENE	MATHIS	12						
				fname80	lname80	age80	serial80	serial80 diff
				GEORGE	MATHIS	29	*805	6
				SARAH	MATHIS	27		
				MAY	MATHIS	4		
				LENA	MATHIS	2		
				CARL	MATHIS	1		
				fname80	lname80	age80	serial80	serial80 diff
				JAMES	MATHIS	27	*819	20
				ANNA	MATHIS	25		
				NELIE	MATHIS	2		

Table 1. Linked households (top) and individuals (bottom), St. Louis 1880

A. **By households (HH)**

Number of Related in HH	1st Enumeration			2nd Enumeration		
	N HHs	N Linked HHs	Linked %	N HHs	N Linked HHs	Linked %
1	3,524	505	14.3	3,855	481	12.5
2	10,650	6,578	61.8	11,599	6,221	53.6
3	11,043	8,482	76.8	11,502	8,354	72.6
4	10,721	9,113	85.0	11,039	9,002	81.5
5	9,546	8,453	88.6	9,729	8,371	86.0
6	7,193	6,521	90.7	7,500	6,668	88.9
7	4,835	4,491	92.9	5,046	4,593	91.0
8	2,988	2,804	93.8	3,194	2,968	92.9
9	1,620	1,530	94.4	1,673	1,557	93.1
10+	1,205	1,155	95.9	1,361	1,274	93.6
All	**63,325**	**49,632**	**78.4**	**66,498**	**49,489**	**74.4**

B. **By individuals**

Number of Related in HH	1st Enumeration			2nd Enumeration		
	N Individuals	N Linked	Linked %	N Individuals	N Linked	Linked %
1	3,519	505	14.4	3,814	481	12.6
2	21,300	12,588	59.1	23,198	11,897	51.3
3	33,129	23,801	71.8	34,506	23,108	67.0
4	42,884	34,146	79.6	44,156	33,247	75.3
5	47,730	39,854	83.5	48,645	38,908	80.0
6	43,134	36,763	85.2	45,000	36,967	82.1
7	33,831	29,599	87.5	35,322	29,790	84.3
8	23,888	20,981	87.8	25,552	21,785	85.3
9	14,580	12,799	87.8	15,057	12,867	85.5
10+	12,688	11,147	87.9	14,487	11,896	82.1
All	**276,683**	**222,183**	**80.3**	**289,737**	**220,946**	**76.3**

Note: Related refers to household members related to the household head, either biologically or through marriage

Table 2a. Linked population's distribution by surname similarity measures, St. Louis 1880

	N	Dist. (%)	NYSIIS	Double Meta	Match1	Match2	Match3
Less than 0.6	2,751	1.2	0.6	3.3	13.9	0.1	0.0
0.60 to 0.649	2,604	1.2	3.1	7.1	44.0	4.1	0.0
0.65 to 0.699	3,573	1.6	7.4	16.4	59.1	8.4	0.0
0.70 to 0.749	6,910	3.1	10.6	18.1	68.3	20.1	1.4
0.75 to 0.799	9,506	4.3	19.5	29.5	79.7	33.9	7.1
0.80 to 0.849	15,918	7.2	32.2	40.7	86.1	41.6	18.4
0.85 to 0.899	20,644	9.3	39.1	47.0	90.2	64.9	34.0
0.90 to 0.949	27,128	12.2	49.1	57.5	96.4	83.2	68.0
0.95 to 0.999	25,348	11.4	66.3	74.7	99.5	93.9	86.5
1.00 (Exact match)	108,048	48.6	100.0	100.0	100.0	100.0	100.0
All	**222,430**	**100.0**	**69.4**	**73.6**	**93.4**	**80.7**	**71.5**

Table 2b. Distribution by Jaro-Winkler score for given names, St. Louis 1880

	N	Dist. (%)	N Name Std.	% Name Std. (by row)
Less than 0.6	13,092	5.9	3,080	23.5
0.60 to 0.649	3,607	1.6	971	26.9
0.65 to 0.699	4,538	2.0	1,695	37.3
0.70 to 0.749	6,491	2.9	2,136	32.9
0.75 to 0.799	8,407	3.8	3,545	42.2
0.80 to 0.849	13,813	6.2	9,415	68.2
0.85 to 0.899	14,407	6.5	9,983	69.3
0.90 to 0.949	19,464	8.8	14,418	74.1
0.95 to 0.999	13,063	5.9	10,461	80.0
1.00 (Exact match)	119,595	53.8	0	0.0
Initial Match	5,953	2.7	0	0.0
All	**222,430**	**100.0**	**55,704**	**25.0**

Table 3. Distribution of age, sex, race, birthplace precision, St. Louis 1880

A. **Age difference**

	N	Dist. (%)
−2 (and greater)years	13283	6.0
−1 year	17552	7.9
Same age	106,275	47.8
+1 year	61,686	27.7
+2 (and greater) years	23,634	10.6
Total	**222,430**	**100.0**

B. **Sex**

	N	Dist. (%)
Agrees	220,323	99.1
Disagrees	2,107	0.9
Total	**222,430**	**100.0**

C. **Race**

	N	Dist. (%)
Agrees	221,904	99.8
Disagrees	526	0.2
Total	**222,430**	**100.0**

D. **Own birthplace**

	N	Dist. (%)
Agrees	203,785	91.6
Disagrees	18,645	8.4
Total	**222,430**	**100.0**

E. **Father's birthplace**

	N	Dist. (%)
Agrees	182,620	82.1
Disagrees	39,810	17.9
Total	**222,430**	**100.0**

F. **Mother's birthplace**

	N	Dist. (%)
Agrees	180,917	81.3
Disagrees	41,513	18.7
Total	**222,430**	**100.0**

Table 4. Migration Status for Rules-Based Household Links, 1870-1880 Complete-Count

Surname Similarity	N Linked HHs	Non Migrant (1)	Same State Different County (2)	Different State (3)	Migrant (2+3)
.90 to .909	44,568	75.8	14.3	9.9	24.2
.91 to .919	25,779	75.2	14.7	10.1	24.8
.92 to .929	38,949	76.1	14.1	9.9	23.9
.93 to .939	44,984	76.7	13.8	9.5	23.3
.94 to .949	40,668	77.0	13.5	9.5	23.0
.95 to .959	38,060	77.1	13.3	9.6	22.9
.96 to .969	69,252	76.9	13.5	9.6	23.1
.97 to .979	70,961	77.5	13.0	9.5	22.5
.98 to .999	7,016	77.6	13.2	9.3	22.4
exact match	1,173,183	79.0	12.0	9.0	21.0
All	**1,553,420**	**78.5**	**12.4**	**9.1**	**21.5**

Household Uniqueness Score	N Linked HHs	Non Migrant (1)	Same State Different County (2)	Different State (3)	Migrant (2+3)
< 10	821,342	77.5	12.9	9.5	22.5
10 - 19	340,343	78.8	12.2	9.0	21.2
20 - 29	170,603	79.5	11.7	8.8	20.5
30 - 39	100,776	79.9	11.6	8.5	20.1
40 - 49	57,573	80.7	11.0	8.3	19.3
50 - 59	31,557	81.2	11.0	7.8	18.8
60+	31,226	81.7	10.5	7.8	18.3
All	**1,553,420**	**78.5**	**12.4**	**9.1**	**21.5**

N potential Links in Household	N Linked HHs	Non Migrant (1)	Same State Different County (2)	Different State (3)	Migrant (2+3)
3	561,326	76.8	13.0	10.2	23.2
4	541,916	78.3	12.6	9.1	21.7
5	272,968	80.0	11.7	8.3	20.0
6+	177,210	81.9	10.8	7.3	18.1
All	**1,553,420**	**78.5**	**12.4**	**9.1**	**21.5**

Table 5a. Household Linkage Rate, 1870-1880 Complete-Count (all 1880 households)

Number rules based linked households	1,553,420
Number of explicitly linked Individuals	6,473,809

N Linkable in 1880 Household	N 1880 Households	N 1880 Households Linked	% Linked
1	934,251	0	0.0
2	4,354,712	0	0.0
3	1,788,843	375,655	20.9
4	1,280,185	428,408	33.4
5	832,111	354,198	42.5
6+	889,856	395,159	44.3
All	**10,079,958**	**1,553,420**	**15.4**

Table 5b. Household Linkage Rate, 1870-1880 Complete-Count (1880 households with 3 or more linkable records only)

Race and Nativity (Household Head)	N 1880 Households	N 1880 Households Linked	% Linked
Native-born white	2,918,696	1,133,828	38.7
Foreign-born white	1,339,201	337,725	25.2
Black	456,746	67,722	14.8
Mulatto	71,053	13,872	19.5
Other	5,299	273	5.2
All	**4,790,995**	**1,553,420**	**32.4**

Table 6. Distribution of Neighbor Count (PHHN) For All Potential Household Links (0.9 Surname Threshold) and For Rules-Based Household Links, 1870-1880 Complete-Count

All Potential HH Links (0.9 Surname Threshold)			Rules-Based Household Links Only			
N Neighbors in Grid (PHHN)	N Potential HH Links	Distribution	N Neighbors in Grid (PHHN)	N Linked Households (Rules-based)	Distribution	% Non Migrant (linked households)
1	12,727,140	59.3	1	317,330	20.4	28.5
2	4,155,353	19.4	2	124,729	8.0	56.6
3	1,459,194	6.8	3	67,350	4.3	71.8
4	546,663	2.5	4	43,488	2.8	80.6
5	238,992	1.1	5	32,522	2.1	86.6
6	129,656	0.6	6	27,108	1.7	90.6
7	88,112	0.4	7	24,699	1.6	92.9
8	71,170	0.3	8	23,441	1.5	94.4
9	65,238	0.3	9	23,917	1.5	95.2
10	62,179	0.3	10	24,166	1.6	96.0
11	61,185	0.3	11	24,453	1.6	96.4
12	61,229	0.3	12	25,431	1.6	96.5
13	61,355	0.3	13	25,556	1.6	96.5
14	61,341	0.3	14	25,788	1.7	96.9
15	61,751	0.3	15	25,992	1.7	97.0
16	62,226	0.3	16	26,479	1.7	97.2
17	62,844	0.3	17	26,998	1.7	97.5
18	62,307	0.3	18	26,757	1.7	97.7
19	62,998	0.3	19	27,326	1.8	97.5
20+	1,345,209	6.3	20+	609,890	39.3	98.8
All	**21,446,142**	**100.0**	**All**	**1,553,420**	**100.0**	**78.5**

Table 7. Number of Linked Households and Individuals, Rules-Only and Rules Plus PHHN Grids

Link Type	N Potential Links in Household	N Households Linked	N Linked Individuals	% Non Migrant (By Households)
rules only (0.9 surname)	**3+**	**1,553,420**	**6,473,809**	**78.5**
rules plus (0.9 surname)	2	485,800	982,388	97.5
rules plus (0.9 surname)	3	87,326	266,460	97.6
rules plus (0.9 surname)	4+	36,400	211,712	96.5
rules only (0.8 surname)	**3+**	**144,469**	**879,008**	**76.9**
rules plus (0.8 surname)	2	20,418	62,193	96.9
rules plus (0.8 surname)	3	65,009	130,914	97.4
rules plus (0.8 surname)	4+	8,902	50,911	94.9
All		**2,401,744**	**9,057,395**	

Table 8. Migration Status for Rules-Based Household Links, 1870-1880 Complete-Count (Surname 0.8 only)

Surname Similarity	N Linked HHs	Non Migrant (1)	Same State Different County (2)	Different State (3)	Migrant (2+3)
.80 to .809	10,870	75.6	14.7	9.7	24.4
.81 to .819	6,908	76.9	13.7	9.5	23.1
.82 to .829	16,513	75.9	14.1	10.0	24.1
.83 to .839	9,119	76.9	13.6	9.5	23.1
.84 to .849	14,697	76.2	14.1	9.7	23.8
.85 to .859	15,230	76.7	13.7	9.6	23.3
.86 to .869	20,393	77.9	12.7	9.3	22.1
.87 to .879	10,765	76.8	13.5	9.7	23.2
.88 to .889	19,506	77.5	13.2	9.3	22.5
.89 to .899	20,468	77.8	12.8	9.5	22.1
All	**144,469**	**76.9**	**13.5**	**9.6**	**23.1**

Household Uniqueness Score	N Linked HHs	Non Migrant (1)	Same State Different County (2)	Different State (3)	Migrant (2+3)
< 10	65,579	76.4	14.1	9.5	23.6
10 - 19	28,190	77.5	13.1	9.4	22.5
20 - 29	21,886	77.2	12.9	10.0	22.8
30 - 39	12,690	76.8	13.5	9.7	23.2
40 - 49	7,500	77.5	12.5	10.1	22.5
50 - 59	4,126	78.3	12.4	9.3	21.7
60+	4,498	81.7	10.5	7.8	22.1
All	**144,469**	**76.9**	**13.5**	**9.6**	**23.1**

N potential Links in Household	N Linked HHs	Non Migrant (1)	Same State Different County (2)	Different State (3)	Migrant (2+3)
3	20,871	76.1	13.3	10.6	23.9
4	71,348	75.5	14.5	10.0	24.5
5	32,911	78.3	12.6	9.1	21.7
6+	19,339	80.1	11.8	7.3	19.1
All	144,469	76.9	13.5	9.6	23.1

Table 9a. Household Linkage Rate, 1870-1880 Complete-Count, Rules and Rules Plus PHHN Grids

Number rules based linked households	2,401,744
Number of explicitly linked Individuals	9,057,395

N Linkable in 1880 Household	N 1880 Households	N 1880 Households Linked	% Linked
1	934,251	0	0.0
2	4,354,712	305,461	7.2
3	1,788,843	585,676	33.6
4	1,280,185	572,697	45.8
5	832,111	442,524	54.5
6+	889,856	495,387	57.0
All	**10,079,958**	**2,401,744**	**23.9**

Table 9b. Household Linkage Rate, 1870-1880 Complete-Count, Rules and Rules Plus PHHN Grids (1880 households with 2 or more linkable records only)

Race and Nativity (Household Head)	N 1880 Households	N 1880 Households Linked	% Linked
Native-born white	5,729,709	1,743,796	30.8
Foreign-born white	2,307,905	510,277	22.3
Black	943,820	123,500	13.2
Mulatto	151,489	23,622	15.8
Other	12,784	549	4.3
All	**9,145,707**	**2,401,744**	**26.3**

Table 10a. Linked population's distribution by surname similarity measures

	N	Dist. (%)	NYSIIS	Double Meta	Match1	Match2	Match3
0.80 to 0.849	467,651	4.6	29.7	35.9	82.4	46.8	21.3
0.85 to 0.899	648,388	6.4	43.8	50.1	90.6	65.7	32.7
0.90 to 0.949	1,127,925	11.1	54.4	61.7	96.8	84.0	68.9
0.95 to 0.999	1,076,914	10.6	71.6	74.0	99.7	94.8	86.7
1.00 (Exact match)	6,920,409	67.4	100.0	100.0	100.0	100.0	100.0
Total	**10,241,287**	**100.0**	**85.1**	**86.9**	**98.2**	**93.1**	**87.3**

Table 10b. Distribution by Jaro-Winkler score for given names

	N	Dist. (%)	N Name Std.	% Name Std. (by row)
Less than 0.6	554168	5.4	151,288	27.3
0.60 to 0.649	64808	0.6	6,092	9.4
0.65 to 0.699	115222	1.1	37,332	32.4
0.70 to 0.749	162728	1.6	25,386	15.6
0.75 to 0.799	281654	2.8	92,664	32.9
0.80 to 0.849	274735	2.7	81,871	29.8
0.85 to 0.899	419735	4.1	218,682	52.1
0.90 to 0.949	783995	7.7	61,936	7.9
0.95 to 0.999	500938	4.9	40,576	8.1
1.00 (Exact match)	7083299	69.0	0	0.0
Total	**10,241,287**	**100**	**715,827**	**7.0**

Table 11. Distribution of age, birthplace, sex and race precision

A. **Age difference**

	N	Dist. (%)
−5 (and greater) years	191,952	1.9
−4 years	134,355	1.3
−3 years	240,630	2.3
−2 years	564,267	5.5
−1 year	1,957,802	19.1
Same age	4,940,063	48.2
+1 year	1,337,226	13.1
+2 years	402,158	3.9
+3 years	178,233	1.7
+4 years	105,844	1.0
+5 (and greater) years	188,757	1.8
Total	**10,241,287**	**100.0**

B. **Birthplace agreement**

	N	Dist. (%)
Agrees	9,983,344	97.5
Disagrees	257,943	2.5
Total	**10,241,287**	**100.0**

C. **Sex agreement**

	N	Dist. (%)
Agrees	10,200,689	99.6
Disagrees	40,598	0.4
Total	**10,241,287**	**100.0**

D. **Race agreement**

	N	Dist. (%)
Agrees	10,179,031	99.4
Disagrees	62,256	0.6
Total	**10,241,287**	**100.0**