US migration from 1850 to 1920:

A comparison of family trees with linked census data

Caglar Koylu[a], Alice B. Kasakoff[b] and Maryam Torkashvand[a]

[a]School of Earth, Environment, and Sustainability, University of Iowa; [b]Department of Geography, University of South Carolina

ABSTRACT

The quality and representativeness of longitudinal datasets play a central role in historical migration research. In this study, we apply the child-ladder (CL) method to a population-scale family tree dataset to analyze U.S. interstate family migration from 1850 to 1920. The CL method infers moves from changes in birthplaces between successive children, allowing for more precise dating of migration events. However, it is limited to families with at least two children. To evaluate the representativeness and utility of family trees for migration research, we compare the CL data to the IPUMS Multigenerational Longitudinal Panel (MLP), which tracks household moves across census decades and serves as a proxy for broader population migration. The CL data reveal higher migration rates, suggesting a likely closer approximation to migration levels in the overall population. Also, by capturing intercensal and return migrations, the CL method provide a detailed view of migration patterns across space and time. Despite differences in migration rates, both datasets reveal similar regional migration structures, especially in the earlier periods. These findings show that population-scale family trees when analyzed using the CL method, offer a valuable complement to linked census data by enhancing our understanding of long-term U.S. migration patterns and regional divisions.

Introduction

The population of the United States was highly mobile during the nineteenth century. People moved long distances due to the westward movement of the European settlers. Much of the migration was of families since hired labor was expensive on the frontier. Families relied on their older sons to clear the land. The most comprehensive source of information about internal migration for the U.S. has been the tables that show the state of birth and the states of residence of all persons born within the U.S. published as part of the census starting in 1850. They show that the percentage of the population living away from their states of birth declined steeply starting in 1850 (Lee and Lee 1960; Price, 1953; Hall and Ruggles 2004; Otterstrom and Bunker 2013).

Despite their breadth, the birth-to-residence tables have important limitations. Since it is a cumulative measure of moves made over a lifetime it misses the exact sequence of moves made by individuals as they came to their current states of residence. The period it covers is very long. Comparison between the decades is difficult because the measure changes depending upon the age structure of the population, the timing of moves over the life cycle, and the death rate by age. Although the census birth-to-residence tables are the most comprehensive source, the data exclude foreign-born residents, and they still cannot reveal when or how often people moved. Complementing the general picture from the census, studies conducted at local or regional scales, such as Mathews (1962) on the expansion of New Englanders into the northern tier of the states, offer valuable insights into patterns of settlement. However, in the absence of population registers, the U.S. still lacks detailed, long-term, individual-level migration data that span the entire population. This persistent gap limits our ability to reconstruct patterns of movement over time and fully understand the dynamics shaping historical internal migration.

Family trees derived from crowdsourced genealogy platforms offer an opportunity to address the lack of detailed, long-term migration data in the U.S. When

individual trees are connected, cleaned, and deduplicated, they can be used to generate a population-scale kinship network that encompass kinship ties among a considerable number of individuals within a specific population (Koylu and Kasakoff 2024). Family tree records include birth and death dates and places, as well as kin ties that allow us to track migration and settlement patterns over time. The short intervals between children during the nineteenth century make it possible to identify turning points in migration rates and patterns using the child-ladder (CL) method, which infers migration events from changes in the birthplaces of consecutive siblings (Lathrop 1948). However, this method has limitations because it can only be used for families with at least two children and therefore applies only to a subset of the population. As a result, the extent to which migration rates and spatial patterns derived from the CL method are representative of the broader population remains uncertain.

In this article, we apply the child-ladder method on a population-scale family tree dataset (Koylu et al. 2021) to analyze interstate migration flows in the U.S. between 1850 and 1920. To evaluate the representativeness of the child-ladder method in capturing family migration patterns, we compare it with a more comprehensive dataset, the Multigenerational Longitudinal Panel (MLP) (Helgertz et al. 2022). The MLP tracks household migration by linking individuals and their households across two consecutive censuses and identifying changes in household residences. Linkage begins with males and subsequently includes other household members, both male and female. While the MLP attempted to link everyone from one census to the next, it proved more difficult to link certain segments of the population, such as single individuals and people living in newly settled regions. Nevertheless, the MLP dataset is much larger and more comprehensive than the family tree dataset. Both the CL and MLP datasets provide a more temporally precise view of U.S. settlement and interstate migration than the commonly used birth-to-residence tables in the census. Because both sources capture migration over shorter intervals, they allow for a dynamic view of change across time.

Building on this comparison, our central research question is: How well does the child-ladder method, applied to population-scale family tree data, capture the family migration patterns relative to the more comprehensive MLP data, which serves as a proxy for migration in the entire population? To address this, we construct and analyze the interstate migration networks from both sources for the census periods 1850–1860, 1860–1870, 1870–1880, 1900–1910, and 1910–1920. We compare migration rates, and flow patterns, and the structural cohesiveness of regions derived from interstate migration flows to evaluate structural cohesiveness of migration regions to evaluate the degree of alignment between the two datasets.

Background

Rosenbloom and Sundstrom (2004) found that migration rates in the U.S. showed a U-shaped pattern, declining from 1850 to 1900 and then rising to 1970, a rise which they attributed to increased education. They were able to study smaller time intervals by extrapolating from a census question which asked whether a person had moved within the last 5 years. This question was first asked in the census in 1940 and continued to be included in later years. They extrapolated the results by age to the censuses before 1940. If age specific rates changed over time, as we think they did, this would not produce an accurate measure of migration. Still, their estimates confirmed the decline from 1850 at least until 1900 which others have found.

Koylu and Kasakoff (2022) used population-scale family trees and the child-ladder method to study migration. They were able to date the changes in migration rates more precisely than previous work using 10-year census intervals. They divided the period from 1789 to 1924 into seven periods. There were two peaks in migration rates, 1837 and 1853, and then rates declined to a low point in 1897. Migration rates bounced back a bit in the twentieth century but remained far below the earlier peaks. Most interstate family migration in the U.S. was from East to West. As the settled area grew larger, there was less family migration between states over long distances. Frontiers were fed by the adjacent areas. By the early twentieth century, there was much less interstate migration in the longest settled regions in the East and the Middle West than had existed earlier.

Early migration patterns resulted in dialectical and cultural regions that persist to this day. Those in the East were along the major rivers run from East to West (with the exception of the Mississippi) and were the major means of transport during the Colonial period and reflected the importance of crops with different climate needs. Steckel (1983) pointed out the importance of specific crops and animals that could thrive only in small range of latitudes. For example, corn, a staple crop, could not grow above a certain latitude due to the number of frost-free days required.

This changed later when new varieties with a greater tolerance of cold were created. Otterstrom and Bunker (2013) employed ancestry data to investigate the origins of populations in various cities across the U.S., supporting Fischer's (1989) study of migration patterns that contributed to the emergence of cultural regions within the country, usually bands along different latitudes, i.e., from North to South.

Population-scale family trees

Family tree records provide information about both family relationships and individuals, which is useful in many different domains including medical research (Daelemans et al. 2013; Williams et al. 2001), local history (Hey 2010), population change, and migration (Adams and Kasakoff 1984; Otterstrom and Bunker 2013; Wrigley and Schofield 1983) and social mobility (Blanc 2024; Clark and Cummins 2024; Shiue 2019). Koylu et al. (2021) cleaned, connected, and deduplicated crowd-sourced family trees with 250 million individuals from one of the major crowdsourced genealogy websites, Rootsweb. They applied a fuzzy matching approach to identify and link individuals and spousal couples across different family trees, using the personal characteristics and family information available, including spouse and parent-child relationships. By identifying candidate spousal pairs, they connected the family trees into tree clusters and removed duplicate records after merging the trees (Koylu et al. 2021). Given the largest connected component of nearly 40 million individuals, and a total of 80 million individuals, Koylu et al. (2021) generated, to date, the largest population-scale and longitudinal kinship network extending over centuries. Other scholars also utilized population-scale family tree data to study migration (Charpentier and Gallic 2020; Han et al. 2017; Kaplanis et al. 2018; Otterstrom and Bunker 2013). Kandt, Cheshire, and Longley (2016) combined genetic data with census records to benefit from the ability to link individuals with detailed demographic information and gain deeper insights into population dynamics and historical patterns. Also using family tree data, Kaplanis et al. (2018) revealed that females exhibited a higher propensity for migration compared to males, albeit their relocations generally spanned shorter distances.

There are several methods for extracting migration from family trees. The most common have been birth to death, life-course, generational, and child-ladder migration methods. The birth to death migration measure considers birth location as origin and death location as destination. However, because death records rarely exist in family tree data from the U.S., the birth-death method does not produce a robust estimation of migration and disregards the series of moves that happen in the life course. To address this issue, Adams, Kasakoff, and Kok (2002) introduced life-course migration, dividing the life course into three stages: birth to birth of first child, birth of first to birth of last child, and birth of last child to death, and they relied most on the second stage. The generational migration method tracks migration by comparing the birthplace of each child with that of their parent (Otterstrom and Bunker 2013). It can also include comparisons with the birthplace of multiple generations, such as grandparents to parents, parents to children, and grandparents to grandchildren. In cases where migration is inferred from the birthplaces of children and their parents, families with more children are likely to generate more migration events. Including parents with multiple children further increases the number of potential moves, amplifying the influence of family size on migration estimates. Koylu and Kasakoff (2020) used the parent-child measure but minimized the effect of family size on migration estimates by counting children of the same sex and birthplace only once per family.

Crowd-sourced family trees have several limitations in studying historical populations and migration. The representativeness of family tree data varies widely by geography, age, sex, race, and other socioeconomic and demographic characteristics. Individuals of white, European descent were linked at notably higher rates, whereas Black, Native American, and many foreign-born groups, particularly those from eastern and southern Europe and Ireland, were significantly underrepresented. For a more in-depth discussion of these limitations and the ethical considerations surrounding the use of family tree data in historical research, we refer readers to Koylu and Kasakoff (2025).

Child-ladder method

Lathrop (1948) coined the term "child-ladder" to describe the method he employed to study the origins of European settlers who pioneered counties in East Texas. Lathrop was able to track families in the census through the birth states of their children. Although he intended to work on later periods, his study focused on 1850 to 1860. The effectiveness of the child-ladder method depended on the age of the county, as migrant families were prominent in newly settled areas. In these areas, migrant families captured by the child-ladder method accounted for about half of the population counted in the census. However,

the method does not include all migrants; it misses families with children that elude detection, childless couples, and single persons. This omission is significant because the migration patterns of childless individuals, particularly unmarried men, may differ from those of families. Despite this limitation, Lathrop (1948) found that the bulk of the migration into East Texas involved farmer families, and he suggested that even if childless persons behaved differently, family migration patterns still provided a nearly accurate representation of overall movement. Regardless of its potential effectiveness to study migration, the child-ladder method has rarely been used since Lathrop's study.

In this article, we use the child-ladder method (CL) to extract migration events and date them based on the mid-point or average of birth years between two consecutive siblings with different birthplaces, providing a more precise estimate of migration timing. On average, children in the family tree data were born two years apart between 1789 and 1924. Similarly, Lathrop (1948) reported that the interval between consecutive births in 1850 and 1860 ranged between two and three years. However, the method inherently favors larger families, which may have been more likely to move, and does not capture the moves of single individuals or those with only one or no children. Since the child-ladder method relies on the birth of children, it lacks information about life stages before and after childbearing years.

Given the rapidly growing U.S. population during the nineteenth century, childless individuals comprised <10% of the population, which is much lower than the proportion of single individuals in Europe (Hacker 2016). In such populations, the child-ladder method may not underestimate migration as significantly as it would in populations with more single individuals. However, because the proportion of single persons increased over time, the method is likely more representative of migration patterns in the earlier periods of U.S. history than in later ones.

Multigenerational longitudinal panel (MLP)

Historical census records are rich resources of population information as they contain names and demographic information as well as ethnicity and nativity not only for each person as an individual but also as a member of a household. The theoretically immutable characteristics in census records, such as individuals' names, birthdates, birthplaces, and genders, allows linking of individual and household records in consecutive censuses. This creates a linked dataset that enables tracking of changes in mutable characteristics, such as residence information, in the decennial census records.

Recently, the linked historical census dataset has been a resource for different historical demographic and long-term social and economic change studies (Antonie et al. 2022; Hacker et al. 2021; Roberts, Rahn, and Lazovich 2022). Historical census records in the 1800s and early 1900s include inaccurate and incomplete information, such as misspelled names, inaccurately reported birth years, and the absence of unique identifier variables, such as social security numbers, which makes it challenging to link the same individual record between multiple censuses. Machine learning algorithms have recently been used to automate a part of the linkage process and find matches in a large amount of data efficiently and accurately (Feigenbaum 2015; Fu, Christen, and Zhou 2014; Goeken et al. 2011; Ruggles 2002). One of the most recent attempts to link population records between censuses is the IPUMS Multigenerational Longitudinal Panel project (Helgertz et al. 2022; Ruggles, Fitch, and Roberts 2018). Using Full Count IPUMS Ancestry data, Helgertz et al. (2022) introduced an automated linkage procedure that uses a two-step probabilistic algorithm. In the first step, the algorithm searches for highly reliable links between men by using all available criteria in the censuses, including variables of individuals and their family ties, such as parents, spouses, and siblings. The second step links the individual records of the household members who did not link in the first stage. This two-step approach made it possible to engage extensive variables in exploiting linkages while searching among a large number of potential links. The resulting longitudinal dataset represented a large population from the census records with a high linkage accuracy among available datasets (Helgertz et al. 2022).

Women are usually underrepresented in longitudinal panels as their surnames normally change after marriage. Hacker (2013) evaluated the undercount of the native-born population in census records from 1850 to 1930 using demographic analysis. His findings indicate that children under the age of 5 and women over the age of 30 were the most underrepresented groups in the census, particularly in earlier years, such as 1880. Additionally, while women over the age of 50 were underestimated in earlier enumerations, the elderly population of men was over-counted. According to Hacker (2013), the white population born in the southern region of the U.S., especially women, was

undercounted in the 1870 census. Overall, it can be concluded that although the enumeration errors were the highest in 1870, the coverage of the historical census slightly improved from 1850 to 1930.

In the machine learning approach, training data quality plays a significant role in generating the final dataset (Bailey et al. 2020). Price et al. (2021) developed a "*census tree*" dataset using the family tree dataset developed by the public and available data on the FamilySearch genealogy website for training their algorithm for link records in the 1900–1920 censuses. They showed that the linkages made voluntarily by family members are the most reliable, especially for linking women. The advantage of a family tree dataset is that it is developed by family members themselves, and they usually have more specific knowledge of their family and ancestors, such as the women's maiden names. However, the across-marriage links for women were limited to the training dataset as the machine can only find matches based on the information available in the censuses.

Unlike the child-ladder method applied to the population-scale family tree dataset, the MLP allows extraction of migration events for also smaller households, such as single individuals or families with one or no children. However, the representativeness of the MLP data is also shaped by linkage constraints. Helgertz et al. (2022) found that white young men aged 7–20 who lived with their parents and came from larger households are overrepresented in the MLP dataset compared to a sample of the 1910 census. Similarly, although women are generally underrepresented in the MLP dataset, certain groups, such as white women living with family and those from larger households, are overrepresented relative to the census samples. In contrast, single individuals and those who have left home are difficult to link and are therefore underrepresented.

Methodology

This study aims to evaluate how well the child-ladder (CL) method, applied to population-scale family tree data, captures family migration patterns compared to those derived from the Multigenerational Longitudinal Panel (MLP), which serves as a broader proxy for historical U.S. population migration. Specifically, we compare migration rates and the structure of state-to-state migration networks across both data sources for the census periods between 1850 and 1920.

To ensure comparability, we limit the MLP data to a subset of households with at least two children who were successfully linked across two consecutive censuses. We refer to this subset as Households with at least Two Linked Children (H2LC). Each H2LC household is counted only once per period, consistent with the CL dataset, where each family is also represented once. This restriction allows for a more balanced comparison of migration patterns between the two sources.

We extract and analyze interstate migration networks from both CL and H2LC data for the periods 1850–1860, 1860–1870, 1870–1880, 1900–1910, and 1910–1920. The 1890 census was excluded due to data loss, and the 1880–1900 period was omitted to preserve consistent 10-year intervals. For the remainder of this manuscript, CL refers to migration data derived from the child-ladder method using family trees, and H2LC refers to migration data from households with at least two linked children in the MLP dataset.

Extracting migration data

Due to the difference between the two sources, it is not possible to make measures that are exactly comparable. However, we can narrow the difference. This section outlines our approach to extracting migration data using two primary methods: the child-ladder method for family trees and state-level changes of residences for households with at least two linked children (H2LC) from the MLP data.

The child-ladder method identifies migration events within nuclear families, defined as one or two parents with at least two children. Migration is recorded when there is a change in the birthplaces of consecutive children, indicating that the family moved between the births of these siblings. This approach captures family-level migration patterns tied to childbearing intervals.

The MLP data link individuals and households between consecutive censuses, such as 1850 and 1860 through unique household identifiers (serial IDs) and other variables, including age, sex, birthplace, and state of residence. One approach to analyzing household migration from the MLP data uses the first census serial ID (e.g., serialid1850) to track, but this can miss new members or those who change households by the second census. Alternatively, using the second census serial ID (e.g., serialid1860) captures newly formed households but loses track of members from the original household in the first census.

To extract household migration for the households with at least two linked children (H2LC), we use household identifiers from both the first census (e.g.,

serialid1850) and the second census (e.g., serialid1860). This allows us to consider household dynamics, including splits, where members of an original household end up in multiple households in the second census, and merges, where individuals from different households in the first census combine into a single household in the second census. However, we should note that the percentage of merges and splits ranges between 0.02 and 0.5% from 1850 to 1920, with a noticeable increase in later periods due to the inherent bias of the MLP method toward linking individuals who remained in the same household. A household is classified as "Migrated" if its state of residence changes between the two censuses, and "Stayed" if it remains in the same state.

Definition of networks

In our analysis, the migration networks for both the Child-Ladder (CL) and Household with at least Two Linked Children (H2LC) data are defined as weighted directed graphs that evolve over time. For each period t, we represent the migration network as $G_t = (V_t, E_t)$, where V_t is the set of nodes representing the states and territories existing at the end of period t; E_t is the set of directed edges representing migration flows between states during period t; and each edge $e_{ij} \in E_t$ is a directed link from node i (origin state) to node j (destination state).

The weight (volume) associated with each edge e_{ij} is denoted as w_{ij}, representing the volume of families or households moving from state i to state j during period t. Thus, the weighted adjacency matrix W_t for period t contains the migration volumes between all pairs of states:

$$W_t = \left[w_{ij}^t \right], \quad \text{where} \quad w_{ij}^t \geq 0, \quad \forall_{i,j} \in V_t$$

where V_t is the set of states/territories in period t. The units of analysis are the states and territories that existed at the end of each period under study. Specifically, the dates are determined based on the MLP data periods that link households across two censuses, such as 1850 and 1860, for the comparative evaluation of CL and H2LC networks. We use the boundaries of the ending period (e.g., 1860) to aggregate the migration flows between states to ensure consistency with the geopolitical context of that time. This approach accounts for any changes in state and territorial delineations over time.

For each period t, we construct the migration networks G_t^{CL} and G_t^{H2LC} for CL and H2LC, respectively. These networks capture the flows between the states and territories in existence at the latest date of the period.

Similarity of migration flows

To quantify the similarity between the flows of CL and H2LC networks for each period t, we represent each network's migration flows as vectors A^t and B^t, respectively. Each dimension of these vectors corresponds to a specific origin-destination pair (i, j), and the value is the migration volume w_{ij}^t:

$$A^t = \left[w_{ij}^{t,CL} \right], B^t = \left[w_{ij}^{t,H2LC} \right]$$

The cosine similarity between the two migration networks for period t is then calculated using the formula:

$$\begin{aligned} &Cosine\,Similarity_t \\ &= \frac{\sum_{i=1}^{V_t} \sum_{j=1}^{V_t} w_{ij}^{t,\mathrm{CL}} \cdot w_{ij}^{t,\mathrm{H2LC}}}{\sqrt{\sum_{i=1}^{V_t} \sum_{j=1}^{V_t} \left(w_{ij}^{t,\mathrm{CL}} \right)^2} \sqrt{\sum_{i=1}^{V_t} \sum_{j=1}^{V_t} \left(w_{ij}^{t,\mathrm{H2LC}} \right)^2}} \\ &= \frac{A^t \cdot B^t}{\| A^t \| \| B^t \|} \end{aligned}$$

where $w_{ij}^{t,CL}$ is the migration volume from state i to state j in the CL network during period t; $w_{ij}^{t,H2LC}$ is the migration volume from state i to state j in the H2LC network during period t; and $|V_t|$ is the number of states and territories at the end of period t.

A cosine similarity of 1 indicates that the two migration networks are identical in terms of the origin-destination pairs and their volumes. A value of 0 means that the networks are completely unrelated, with no overlap in their migration patterns, and a value of −1 indicates that the networks are diametrically opposed. By computing the cosine similarity for each period, we can assess how closely the migration flows between states captured by the CL network align with those captured by the H2LC network.

It is important to note that cosine similarity primarily considers the edges (origin-destination pairs) that exist in both networks. Therefore, the interpretation of the similarity measure is based on the overlap of migration flows between the two networks. However, since our units of analysis are states and territories, and the edges between the nodes significantly overlap in both CL and H2LC networks, this potential limitation is mitigated.

We acknowledge that using raw migration counts could conflate true changes in flow patterns with simple population growth, and thus, making it difficult to tell whether an increase in migration flows reflects greater migration propensity or merely more people overall. Proportional, row and/or column-scaled, and log-ratio normalizations could further clarify

period-to-period change; exploring these with scale-sensitive checks is a logical next step, especially for eras where rapid population growth compresses inter-period cosine distances.

Similarity of migration regions

In addition to assessing flow similarity, we assess the structural similarity of the CL and H2LC migration networks. Structural similarity refers to the similarity in the community structures of the networks—that is, how the nodes (states) are grouped into communities (regions) based on their migration connections.

To analyze the community structures within each network, we identify communities (also referred to as modules or regions) using the Leiden community detection algorithm (Traag, Waltman, and Van Eck 2019), which is an improvement over the commonly used Louvain method (Blondel et al. 2008). The Leiden algorithm addresses some limitations of the Louvain method by guaranteeing well-connected communities and providing better quality partitions. In this context, we use the terms communities, modules, and regions interchangeably to refer to groups of states that are densely connected through migration flows.

The Leiden algorithm is an iterative method that identifies communities with high modularity in a network. Modularity is a measure of the strength of a community structure in a network, with higher modularity values indicating that states within a region have more migration connections to each other than to states outside the region. In the first step of the Leiden algorithm, each state (node) forms an initial community in the network. The algorithm then moves individual states to neighboring states if such a move increases the modularity to account for the direction of edges.

In addition, the Leiden algorithm effectively handles directed graphs, which is crucial for our study of the nineteenth century migration flows. Migration during this period was predominantly westward, toward the ever-changing frontier in the Western United States. By considering the directionality of migration flows, the Leiden algorithm captures the asymmetric nature of these movements and thus provides a more accurate representation of structural patterns. For directed, weighted networks, the modularity formula is as follows:

$$Q = \frac{1}{M}\sum_{ij}\left(w_{ij} - \frac{k_i^{out}k_j^{in}}{m}\right)\delta\left(c_i c_j\right)$$

where w_{ij} is the weight of the directed edge from node i to node j; k_i^{out} is the total weight of the outgoing edges from node i (out-strength); k_j^{in} is the total weight of the incoming edges to node j (in-strength); m is the total weight of all edges in the network; $\delta\left(c_i c_j\right)$ is 1 if nodes i and j are in the same community and 0 otherwise. Unlike similarity measures, which quantify how closely two networks resemble each other (e.g., cosine similarity), modularity assesses how well the network is divided into communities. Although modularity is not strictly bounded between 0 and 1, higher values indicate stronger community structures. High modularity means there are dense connections between states within regions but sparse connections between states in different regions.

The optimal number of communities is determined by maximizing the modularity score, which reflects the strength of the community structure within the network. In other words, we do not predefine the number of regions; instead, the Leiden algorithm identifies the number of regions that maximize modularity.

Similar to the Louvain method, the Leiden algorithm does not enforce spatial contiguity, which allows regions to be geographically disjointed and distant from each other. We employ this type of community detection algorithm because this reflects historical migration patterns in the U.S., which often involved long-distance moves from eastern to western states.

After we detect the regions and visually assess the agreement between the two sources of CL and H2LC, we compare the similarity of regions using the z-score of the Rand coefficient, also known as the z-Rand score (Red et al. 2011). The z-Rand score computes the number of node (state) pairs that belong to the same community (region) in two different regionalization or clustering results. The z-Rand score is a normalized measurement that analytically compares the pair count measure to its expected value under a null model with the same size communities. The null hypothesis for the z-Rand score posits that the observed similarity between the partitions of the two networks is no greater than what would be expected by random chance. If the z-Rand is found to be significant, it indicates that the similarity between the partitions is statistically meaningful, which demonstrates that the structures of the two networks are aligned in a way that cannot be attributed to random variation.

Results

As can be seen in Figure 1, the H2LC dataset (green) is larger than the CL dataset (purple) and it increases over time, mirroring the growth of the population of

the U.S. In contrast, the CL dataset peaks in 1870 to 1880. This is partly due to the bias in the family tree data toward the earlier waves of settlement. However, in the first decade, the two datasets are much closer in size suggesting that the CL is more representative of the general population in the late nineteenth century than in the twentieth century.

Migration rates

We calculate migration rates for the CL network by including both migrant and stayer families. To determine whether a family moved or stayed, we require at least two children born within a given period. In the MLP data, a move is identified when the same household is censused in two different states across consecutive censuses, regardless of the number of children. To enable a meaningful comparison with the family tree dataset, which defines migration based on the birthplace changes of multiple children within the same family, we limit the MLP sample to households with at least two children who were successfully linked across both censuses. We refer to this subset as H2LC. We stipulate that those children had to be age 18 or younger in the first census. In the CL network, the moves are dated at the mean between the birth years of successive children. This places them within a particular decade. But in H2LC they were largely born in the prior decade because they had to exist in the first census to have been linked to the second census. Thus, the families in the CL network are younger on average than the families in the H2LC network.

As shown in Figure 2, the family migration rate in the CL network declines over time, dropping from 13.2% in 1850–1860 to 8.3% in 1910–1920. In contrast, the H2LC network shows relatively stable household migration rates, with 8.8% (83,518 households) migrating in 1850–1860 and 9.4% (550,732 households) in 1910–1920. Meanwhile, the total number of households increases significantly, from 951,591 in 1850–1860 to 5,834,362 in 1910–1920, reflecting the population growth in the U.S.

The H2LC data captures only the census-to-census moves, missing within-period migrations. The percentage of families that moved more than once among all migrating families in the CL data consistently ranges between 21 and 24% of all migration events, as shown in Table 1. The inability to capture within decade moves results in lower migration rates for the H2LC data, particularly in earlier periods when a larger percentage of families likely moved more than once. One possible way to estimate these missing moves in future work is to extrapolate the proportion of multi-move families from the CL data and adjust the H2LC migration rates accordingly for each period.

In addition to missing migrations in intercensal periods, individuals who left the state but later returned would be classified as stayers in H2LC. Return migration statistics in Table 1 show that about 22% of migrant families that moved more than once returned to their earlier states across all census periods (Table 1). CL detects return migration when a family leaves and later returns to a previous state. If the return occurs within the 10-year intercensal period, the family is classified as a return migrant. The ability to detect return migration may also explain

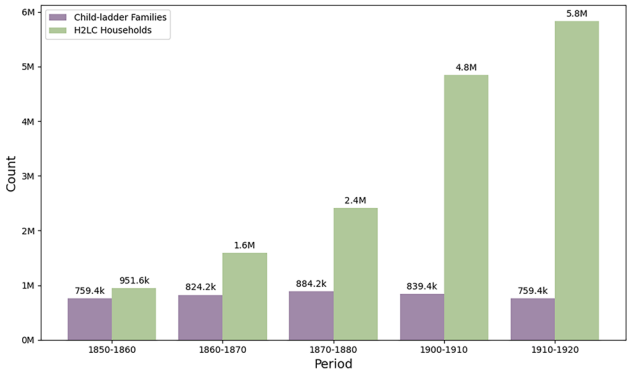


Figure 1. The total number of CL families and H2LC households per period.

Table 1. CL families with multiple migrations and return migration.

Period	>1 migration/all migration (%)	Return migration/>1 migration (%)
1850–1860	19.4	23.5
1860–1870	19.6	22
1870–1880	19.9	23
1900–1910	18.8	21.2
1910–1920	17.9	18.9

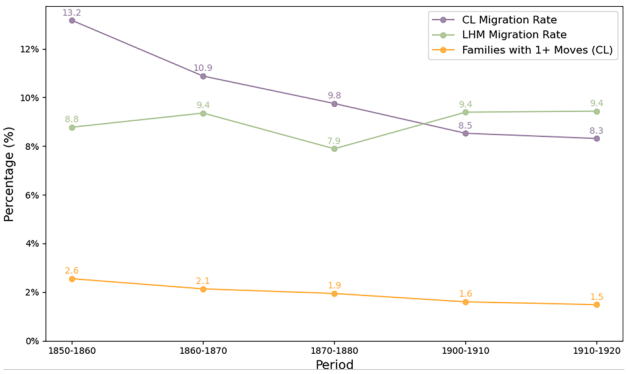


Figure 2. Migration rates for families (CL) and H2LC households per period.

part of the discrepancy between migration rates in the two data sets. However, in the first period adding return migrants to H2LC estimates would raise the migration rate but still fall short of the CL rate, while in the twentieth century, such an adjustment would further widen the gap between the two sources.

Migration regions

To gain a deeper understanding of the underlying mechanisms and the structure of migration patterns, we group states into regions based on how strongly they are connected to one another by migration flows. We do this using the Leiden community detection method, which identifies clusters of states that have more migration between them than with other states. The number of regions is chosen by the algorithm to maximize modularity, which is a measure of how well the network is divided into tightly connected regions. In simple terms, high modularity means that states within a region are strongly connected through migration, while connections between regions are weaker. Table 2 shows the number of regions and modularity values for both the CL and H2LC networks in each period. These values are similar across all periods, ranging from 0.35 to 0.41, suggesting that both networks reveal comparable levels of internal cohesion within migration regions.

In the earlier periods (1850–1860 and 1860–1870), both the CL and H2LC networks have identical modularity values (0.37 and 0.36, respectively) and the same number of regions. The H2LC network produces slightly higher modularity values than the CL network, likely due to five regions in H2LC *versus* four in CL during the 1870–1880 and 1900–1910 periods. Overall, the modularity scores remain very similar, indicating comparable levels of regional cohesiveness in both networks.

Table 3 presents the results of the z-Rand and cosine similarity measures used to compare the CL and H2LC networks across multiple periods. The z-Rand score measures the similarity between the partitions of nodes (states) into regions identified by the Leiden algorithm in the two networks. Comparing z-Rand scores across periods is only meaningful when the complexity of the networks remains similar. The complexity could be determined by factors, such as the number of nodes, edges, communities, degree distribution, the distribution of edge weights, and network density.

The z-Rand scores show statistically significant similarity between the CL and H2LC networks in all periods. This significant similarity indicates that the migration regions derived from both networks align well, which means that the partitions of states into regions are highly consistent between the two data sources. In other words, in each period, the z-Rand scores confirm that the regions formed by the CL and H2LC networks are more similar than would be expected by chance, which reflects meaningful structural correspondence between the networks over time.

In addition to comparing the regions, we use cosine similarity to assess the similarity of the flow volumes between common origin-destination pairs. This metric focuses on the magnitude of flows rather than the structural configuration of the networks. Across the periods, cosine similarity values remain relatively high, peaking in 1870–1880 (CS = 0.89), which reflects a strong alignment in migration flow patterns during this period. However, there is some fluctuation, with a slight decline in the later periods, such as 1910–1920 (CS = 0.82), indicating that while the flow volumes are somewhat aligned, they diverge slightly over time. This suggests that although the structural organization of the CL and H2LC networks may persist, the magnitude and patterns of movement along the common edges evolve over time.

To visualize how migration regions evolve over time and compare them between datasets, it is important to assign consistent colors to similar regions across maps. To track and compare these regions over time and between datasets using a color coordination of similar and distinct regions, we assign global meta-community IDs using a similarity-based approach. For each time period and dataset, the migration networks are first partitioned into communities (regions) using the Leiden algorithm. Communities from the two datasets and periods are then compared using the

Table 2. The Leiden community detection results for CL and H2LC regions.

	Child-ladder		H2LC	
Period	Modularity	# Regions	Modularity	# Regions
1850–1860	0.37	4	0.37	4
1860–1870	0.36	5	0.36	5
1870–1880	0.35	4	0.38	5
1900–1910	0.36	4	0.40	5
1910–1920	0.39	5	0.41	5

Table 3. Network comparison of child-ladder and H2LC regions.

Period	# Nodes	z-Rand	CS
1850–1860	40	27.91	0.79
1860–1870	47	18.78	0.74
1870–1880	47	24.63	0.89
1900–1910	49	26.36	0.83
1910–1920	49	19.14	0.82

z-Rand scores represent the degree of similarity between regions identified, while CS values represent the cosine similarity between CL and H2LC networks for each period.

Jaccard similarity index, which measures the overlap in membership between communities. If two communities have a similarity score above a predefined threshold (40%), they are assigned the same global meta-community ID, and thus the same color, which remains consistent across all maps.

Figure 3 illustrates the migration regions derived from the CL and H2LC networks across all periods. The migration regions identified from each data source display remarkable similarities to each other, which is supported by the z-Rand scores. Although the H2LC households are older and thus farther along in their childbearing than the CL families, the scores in any given period are very similar. In each period, the networks maintain a consistent structure in terms of origins and destinations, even though the number of the migrating families and households differs between the two sources.

The North-South divide appears in all the maps in Figure 3, although there were a few minor changes over time: for example, Kentucky is grouped with the South briefly from 1860 to 1880 period in both datasets, while Virginia is classified as part of the Southeast only in the H2LC regions in the twentieth century. Kansas and Missouri are consistently placed in the Southwest across both datasets in the twentieth century, with Colorado joining them in the final period (1910–1920).

The East became increasingly separated from the West over time in both sets of maps because people from the East were less likely to move to the West. The West came to include states like Wisconsin and Iowa in the twentieth century, as the Mississippi River became a dividing line rather than a way of connecting the states on either side.

East-West regional groupings appear in all periods but are especially prominent in the early periods of 1850–1860 and 1860–1870, as shown in blue color. Over time, these longitudinal groups of states or regions became more entrenched and self-contained in the East. In the earlier periods, a single region often spanned a broad swath of states from the East to the West, but in later periods, these groupings increasingly split into distinct Eastern and Western regions. For example, in both datasets, the upper Midwest was part of the East from 1850 to 1860 which included New England, New York State, and Pennsylvania. This is the Yankee ethnic group well known to have been the first to settle these regions, and to have voted for Lincoln.

In both datasets, the West was initially grouped with the Midwest during the 1850–1860 period, reflecting the region's early settler origins who were not foreign born. Between 1860 and 1880, however, the Far West emerged as a distinct and autonomous region in both sources, no longer integrated with the Midwest, and continued to expand in size over subsequent decades.

In the twentieth century, the West divides into northern and southern regions, each consisting of states west of the Mississippi, marking a return to the longitudinal bands previously seen in the East. These regions were historically shaped by rivers flowing into the Atlantic and Pacific, which served as early transportation routes, especially East of the Mississippi, later followed by the railroads running from East to West.

Despite the high degree of similarity, there are also differences between the maps from the two datasets. Interestingly, the maps do not show a middle band in the East, between the North and the South. Instead, a Middle West region, which is shown in pink, appears only in the H2LC network during the 1870–1880 period. By 1900–1920, this region shifts westward, now in an olive-green shade, encompassing Kentucky, Indiana, and Illinois, while excluding states east of Ohio and West Virginia. This is one of the few differences between the two data sets. The CL network lacks these regions. Instead, there is a larger region, in a teal shade, which includes these states starting in 1870 stretching all the way from New England to Illinois. Also, in the final map of the CL, there are two regions in the area between the Mississippi and the Far West. By contrast, the H2LC network combines this area into a single Northern region that merges with the Far West. This divergence may reflect differences in population composition: the CL data, drawn from family trees, overrepresents native-born populations who were more likely to remain settled in interior states. By contrast, the H2LC data includes a broader and more representative sample of the population, including more foreign-born households, who may have been more likely to continue migrating westward, contributing to the emergence of a single northern region.

In summary, the division of the U.S. into regions based on migration networks is highly consistent across both datasets. Both the CL and H2LC regions show the North-South division persistently over time and longitudinal bands of regions from East to West. In the earliest maps, the West is grouped with the Eastern regions from which it was originally settled. Over time, it gradually separated and expanded into a distinct and larger region. In the twentieth century, the Southwest separated from the Southeast. These regional divisions reflect historical migration patterns shaped by economic, geographic, and cultural factors.

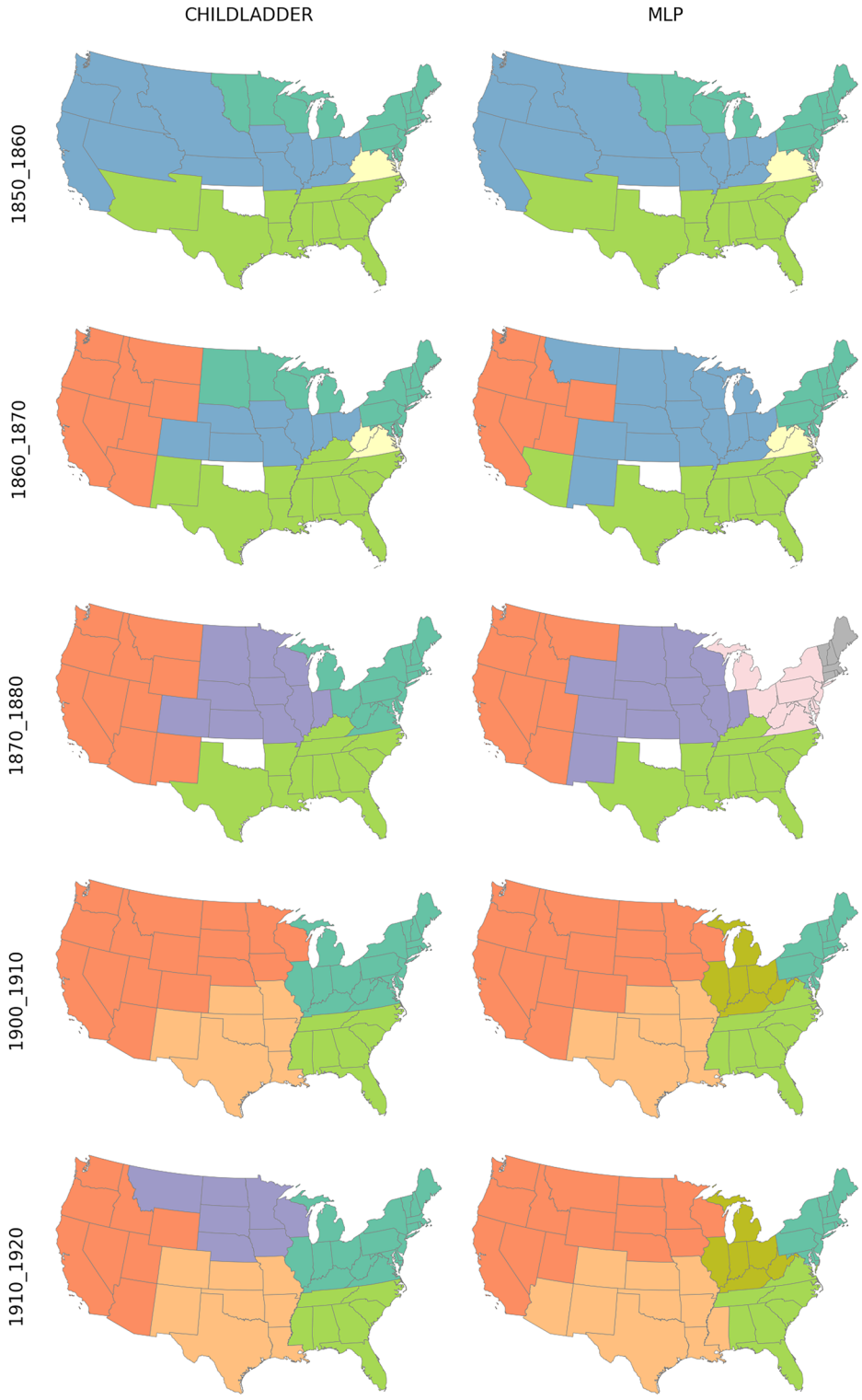


Figure 3. Migration regions by period for the child-ladder (CL, left) and households with at least two linked children (H2LC, right) networks.

Discussion

Kinship networks have long played a central role in migration patterns, particularly in the 19th and 20th centuries. Even as demographic transitions reduced family sizes and technological advances allowed people to live farther apart, kinship ties continued to influence migration decisions, especially immigrant migration streams (Massey, Goldring, and Durand 1994; Palloni et al. 2001; White, White, and Johansen 2005). As the U.S. was settled from abroad, kin networks

helped shape settlement patterns, not simply as a function of population density but by evolving as they aged in place and sent new migrants westward (Koylu et al. 2014). With declining fertility rates, kin networks shifted from initiating new settlements to strengthening established communities over time. This led to shorter distance moves within regions rather than long-distance migrations. Regions in Figure 3 show a change from more expansive family networks sending people to newly settled areas to moves at shorter distances within established regions. These networks adapted over time, continuing to shape migration flows and regions across generations.

Both datasets used in this paper are only approximations. But the steeper decline in the CL migration rate during the nineteenth century is probably closer to the true rate of migration than the H2LC rate derived from the MLP data. It is difficult to imagine a scenario in which the proportion of people living away from their state of birth (as shown in the state of birth state of residence tables published in the census) would decrease when most people were moving West without a decline in the rate of migration over time. Smaller families due to the fertility decline and the distance to the frontier would have led to a slowdown in migration in the East where much of the population lived. Why did the H2LC data not show this? The linking process in MLP which used addresses to "confirm" matches is biased toward stayers. The MLP has fewer individuals not living in their state of birth in the 1900–1910 sample (19 *vs.* 26% of men in the census population). Also, MLP has fewer people in regions that had been newly settled in 1910, such as the West South Central, and Mountain and Pacific, where 18% of the men in the census population lived there but only 14% were linked in the MLP sample (Helgertz et al. 2022). Women are similarly biased with 24% in MLP living away from their state of birth in the census but 20% in the MLP; 16% of women lived in the three newly settled regions in the census but only 12% in MLP. On the other hand, the CL data includes only large families and is biased toward them, and these large families were also more likely to move. Although we do not analyze the effects of family size in this paper, ongoing work suggests that the observed decline in migration rates may partly reflect demographic change, as large families made up a shrinking share of the population over time. In earlier periods, frontier settlement often depended on children for labor, especially where hired help was scarce. As family sizes declined, this migration dynamic likely weakened.

Yet, the findings from both sets of data provide evidence for an important change from a pioneering pattern, which is characterized by migration among larger families, to the modern pattern, where migration typically peaks in early adulthood among smaller families and declines with age, aside from a modest increase at retirement. The H2LC data allow us to date this change to the end of the nineteenth century. However, ongoing work shows that the sub population in the CL data continued to exhibit the earlier pattern albeit the rate of migration declined over time even in the largest families.

Both datasets underrepresent smaller households. MLP (Table 6 in Helgertz et al. 2022) shows that linking other household members to the men they started with (Step I) led to a bias toward larger households, not surprising since the larger the household the greater number of individuals they could have linked. In 1910, the linked sample has just over half the men in households from 2 to 3 individuals that existed in the census, and while on the census 34% of all men were in households with seven or more members, in MLP 46% of the men lived in those very large households. The H2LC sample is more skewed to larger households for women.

The CL method has an even stronger bias toward large families, as detecting a move requires that a child be born in a different location. This means larger families have a higher chance of showing migration. Some methods can address this bias; for example, a survival approach using the last child as a cutoff may help control for this effect and provide a clearer view of the connection between family size and migration. Evidence of this pattern appears in the H2LC data of the 1850–1860 period, indicating that larger households were indeed more mobile. However, the importance of this pattern declined over time due to the decrease in family farms, where children's labor was valuable, and the general decline in fertility rates.

As part of a separate, ongoing effort to evaluate biases in family tree data, we have conducted a linkage experiment between individuals likely alive in 1880 from the family tree data and records in the 1880 U.S. Census. We have successfully linked individuals in the family trees to ~3% of the total census population in 1880. The representativeness of the family tree data varies substantially across geography, age groups, sex, race, and other socioeconomic and demographic factors. White, European-descendant individuals were linked at much higher rates, while Black, Mexican, Native American, and many foreign-born populations, especially those from eastern and southern Europe and Ireland, were

significantly underrepresented. For a more detailed discussion of the limitations and ethical challenges involved in using family tree and census data in historical research, we refer readers to Koylu and Kasakoff (2025).

This paper introduces methods for grouping states into regions and measures of regional cohesion and similarity that have not been previously applied to the settlement process anywhere in the world. When we compare migration rates between the two data sources, the differences in data sources mean this comparison is only approximate and may be off by a decade. However, when the subject is a network of state-to-state migration, the network structures, as shown in our maps derived from the two distinct data sources directly comparable. It is no surprise that migration regions derived from the two datasets are remarkably similar to an extent that cannot be due to chance. These methods could also be applied to other sources, such as the census cross-tabulations of state of birth by state of residence. Because those tables report complete counts for each census, the same analyses could be tailored to specific population segments (e.g., by nativity, race, or sex).

In this article, we focus on the methods of extracting migration, particularly from family trees. However, the child-ladder method could also be applied to census data using the complete household records for the individuals linked by MLP. This approach would capture household members present in the first census but absent in the second due to events, such as death, migration, or adult children forming their own households. It would also include individuals who appear in the second census but not in the first, such as children born after the first census, new household members through marriage, or individuals who moved into the household between censuses. Drawing on the strengths of MLP and the child-ladder method, this approach would allow the study of changes in household composition and the actual size of the household including individuals not present in both censuses.

Conclusion

The child-ladder (CL) approach, applied to the population-scale family tree data, provides a more detailed view of migration than the data derived from the Households with at least Two Linked Children (H2LC) in the Multigenerational Longitudinal Panel (MLP). This advantage stems from the CL method's ability to capture and date moves that are not visible when only two points in time, a decade apart, are available. In contrast, migration measured in MLP, based on changes in residence between censuses, underestimates migration by at least 20%, likely more, due to unobserved moves occurring between census intervals and the exclusion of return migration, which is not counted as a move.

However, the more detailed picture of the child-ladder method is only available for a subset of the population, namely families in their childbearing years. As fertility declined over time, these families represented a diminishing share of the population. While both CL and H2LC data have representational biases, the CL method depends on the presence of multiple children to detect migration, whereas H2LC captures household moves more broadly across different household types. This makes certain comparisons challenging. Still, the CL method's precise dating of moves enables analysis of short-term migration shifts and responses to historical events, such as wars, climate change, and economic cycles.

Although migration rates differ between the two datasets, the migration regions they produce are strikingly similar, especially in the earliest period (1850–1860), when the CL dataset was likely most representative of the broader population. Still, the changes over time in the spatial patterns and regions are very similar. These regional patterns likely represent constraints from the past and geographic, transportation, and economic factors that left their imprint on migration regardless of the subpopulation. Used together, the two sources document a major change in migration patterns in the U.S., from an earlier pattern of long-distance, East-to-West movement to one characterized by shorter-distance migration and increasingly cohesive regional structures.

This study demonstrates that integrating population-scale genealogical data with linked census panel data significantly improves our ability to reconstruct long-term U.S. migration trends. By comparing these complementary sources, we gain a deeper understanding of the temporal and spatial structure of internal migration, while also recognizing the representational limits of each dataset. Together, they offer a more comprehensive view of how migration patterns and regional divisions in the U.S. evolved between 1850 and 1920.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Science Foundation (NSF) Grant No. 2215568 titled "Population-scale kinship networks and migration."

References

Adams, J. W., A. Kasakoff, and J. Kok. 2002. Migration over the life course in XIXth century Netherlands and the American North: A comparative analysis based on genealogies and population registers. Paper presented at the Annales de démographie historique. doi: 10.3917/adh.104.0005.

Adams, J. W., and A. B. Kasakoff. 1984. Migration and the family in colonial New England: The view from genealogies. *Journal of Family History* 9 (1):24–43. doi: 10.1177/036319908400900102.

Antonie, L., K. Inwood, C. Minns, and F. Summerfield. 2022. Intergenerational mobility in a mid-Atlantic economy: Canada, 1871–1901. *The Journal of Economic History* 82 (4):1003–29. doi: 10.1017/S0022050722000353.

Bailey, M. J., C. Cole, M. Henderson, and C. Massey. 2020. How well do automated linking methods perform? Lessons from US historical data. *Journal of Economic Literature* 58 (4):997–1044. doi: 10.1257/jel.20191526.

Blanc, G. 2024. Demographic transitions, rural flight, and intergenerational persistence: Evidence from crowdsourced genealogies.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10):P10008. doi: 10.1088/1742-5468/2008/10/P10008.

Charpentier, A., and E. Gallic. 2020. Using collaborative genealogy data to study migration: A research note. *The History of the Family* 25 (1):1–21. doi: 10.1080/1081602X.2019.1641130.

Clark, G., and N. Cummins. 2024. *Matriline versus patriline: Social mobility in England, 1754–2023*.

Daelemans, S., J. Vandevoorde, J. Vansintejan, L. Borgermans, and D. Devroey. 2013. The use of family history in primary health care: A qualitative study. *Advances in Preventive Medicine* 2013:695763. doi: 10.1155/2013/695763.

Feigenbaum, J. J. 2015. Automated census record linking: A machine learning approach.

Fischer, D. H. (1989). Albion's seed: Four British folkways in America (America: A Cultural History) (Vol. 1). New York: Oxford University Press.

Fu, Z., P. Christen, and J. Zhou. 2014. A graph matching method for historical census household linkage. Paper presented at the Advances in Knowledge Discovery and Data Mining, Cham, Switzerland.

Goeken, R., L. Huynh, T. Lenius, and R. Vick. 2011. New methods of census record linking. *Historical Methods* 44 (1):7–14. doi: 10.1080/01615440.2010.517152.

Hacker, J. D. 2013. New estimates of census coverage in the United States, 1850–1930. *Social Science History* 37 (1):71–101. doi: 10.1017/S0145553200010579.

Hacker, J. D. 2016. United States. *Demography* 53 (6):1657–92. doi: 10.1007/s13524-016-0513-7.

Hacker, J. D., J. Helgertz, M. A. Nelson, and E. Roberts. 2021. The influence of kin proximity on the reproductive success of American couples, 1900–1910. *Demography* 58 (6):2337–64. doi: 10.1215/00703370-9518532.

Hall, P. K., and S. Ruggles. 2004. "Restless in the midst of their prosperity": New evidence on the internal migration of Americans, 1850–2000. *The Journal of American History* 91 (3):829–46. doi: 10.2307/3662857.

Han, E., P. Carbonetto, R. E. Curtis, Y. Wang, J. M. Granka, J. Byrnes, K. Noto, A. R. Kermany, N. M. Myres, M. J. Barber, et al. 2017. Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nature Communications* 8 (1):14238. doi: 10.1038/ncomms14238.

Helgertz, J., J. Price, J. Wellington, K. J. Thompson, S. Ruggles, and C. A. Fitch. 2022. A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel. *Historical Methods* 55 (1):12–29. doi: 10.1080/01615440.2021.1985027.

Hey, D. 2010. *The Oxford companion to family and local history*. Oxford: OUP.

Kandt, J., Cheshire, J. A., and Longley, P. A. 2016. Regional surnames and genetic structure in Great Britain. Transactions of the Institute of British Geographers, 41 (4): 554-569.

Kaplanis, J., A. Gordon, T. Shor, O. Weissbrod, D. Geiger, M. Wahl, M. Gershovits, B. Markus, M. Sheikh, M. Gymrek, et al. 2018. Quantitative analysis of population-scale family trees with millions of relatives. *Science* 360 (6385):171–5. doi: 10.1126/science.aam9309.

Koylu, C., and A. B. Kasakoff. 2024. Population-scale kinship networks. In *International encyclopedia of geography*, ed. D. Richardson, N. Castree, M. F. Goodchild, A. Kobayashi, W. Liu, and R. A. Marston, 1–12. Malden, MA: John Wiley & Sons Ltd.

Koylu, C., and A. B. Kasakoff. 2025. Ethical challenges in analyzing and mapping historical demographic changes and migration using population-scale family trees. *Cartographic Perspectives* (105):55–62. doi: 10.14714/CP105.1945.

Koylu, C., and A. Kasakoff. 2020. Mapping temporal trends of parent-child migration from population-scale family trees. Paper presented at the AutoCarto International Research Symposium, World Wide Web.

Koylu, C., and A. Kasakoff. 2022. Measuring and mapping long-term changes in migration flows using population-scale family tree data. *Cartography and Geographic Information Science* 49 (2):154–70. doi: 10.1080/15230406.2021.2011419.

Koylu, C., D. Guo, A. Kasakoff, and J. W. Adams. 2014. Mapping family connectedness across space and time. *Cartography and Geographic Information Science* 41 (1):14–26. doi: 10.1080/15230406.2013.865303.

Koylu, C., D. Guo, Y. Huang, A. Kasakoff, and J. Grieve. 2021. Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S. *International Journal of Geographical Information Science* 35 (12):2380–423. doi: 10.1080/13658816.2020.1821885.

Lathrop, B. F. 1948. Migration into East Texas 1835–1860. *The Southwestern Historical Quarterly* 52 (1):1–31.

Lee, E. S., & and Lee, A. S. (1960). Internal Migration Statistics for the United States. Journal of the American Statistical Association, 55(292), 664-697. doi:10.1080/01621459.1960.10483367

Massey, D. S., L. Goldring, and J. Durand. 1994. Continuities in transnational migration: An analysis of nineteen Mexican communities. *American Journal of Sociology* 99 (6):1492–533. doi: 10.1086/230452.

Mathews, L. K. 1962. *The expansion of New England: The spread of New England settlement and institutions to the Mississippi River, 1620–1865*. New York, NY: Russell and Russell.

Otterstrom, S. M., and B. E. Bunker. 2013. Genealogy, migration, and the intertwined geographies of personal pasts. *Annals of the Association of American Geographers* 103 (3):544–69. doi: 10.1080/00045608.2012.700607.

Palloni, A., D. S. Massey, M. Ceballos, K. Espinosa, and M. Spittel. 2001. Social capital and international migration: A test using information on family networks. *American Journal of Sociology* 106 (5):1262–98. doi: 10.1086/320817.

Price, D. O. (1953). Estimates of Net Migration in the United States, 1870-1940. American Sociological Review, 18(1), 35-39. doi:10.2307/2087846

Price, J., K. Buckles, J. Van Leeuwen, and I. Riley. 2021. Combining family history and machine learning to link historical records: The Census Tree data set. *Explorations in Economic History* 80:101391. doi: 10.1016/j.eeh.2021.101391.

Red, V., D. E. Kelsic, J. P. Mucha, and A. M. Porter. 2011. Comparing community structure to characteristics in online collegiate social networks. *SIAM Review* 53 (3):526–43. doi: 10.1137/080734315.

Roberts, E., W. Rahn, and D. Lazovich. 2022. Life-course transitions in rural residence and old-age mortality in Iowa, 1930–2014. *The Russell Sage Foundation Journal of the Social Sciences* 8 (4):106–24. doi: 10.7758/rsf.2022.8.4.05.

Rosenbloom, J. L., and W. A. Sundstrom. 2004. The decline and rise of interstate migration in the United States: Evidence from the IPUMS, 1850–1990. In *Research in economic history*. Vol. 22, 289–325. Leeds: Emerald Group Publishing Limited.

Ruggles, S. 2002. Linking historical censuses: A new approach. *History and Computing* 14 (1–2):213–24. doi: 10.3366/hac.2002.14.1-2.213.

Ruggles, S., C. A. Fitch, and E. Roberts. 2018. Historical census record linkage. *Annual Review of Sociology* 44 (1):19–37. doi: 10.1146/annurev-soc-073117-041447.

Shiue, C. H. 2019. Social mobility in the long run: A temporal analysis of China from 1300 to 1900. CEPR, Discussion Paper No. DP13589.

Steckel, R. H. 1983. The economic foundations of east-west migration during the nineteenth century. *Explorations in Economic History* 20 (1):14–36. doi: 10.1016/0014-4983(83)90040-2.

Traag, V. A., L. Waltman, and N. J. Van Eck. 2019. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports* 9 (1):5233. doi: 10.1038/s41598-019-41695-z.

White, D., D. R. White, and U. Johansen. 2005. *Network analysis and ethnographic problems: Process models of a Turkish nomad clan*. Lanham, MD: Lexington Books.

Williams, R. R., S. C. Hunt, G. Heiss, M. A. Province, J. T. Bensen, M. Higgins, R. M. Chamberlain, J. Ware, and P. N. Hopkins. 2001. Usefulness of cardiovascular family history data for population-based preventive medicine and medical research (the Health Family Tree Study and the NHLBI Family Heart Study). *The American Journal of Cardiology* 87 (2):129–35. doi: 10.1016/s0002-9149(00)01303-5.

Wrigley, E. A., and R. S. Schofield. 1983. English population history from family reconstitution: Summary results 1600–1799. *Population Studies* 37 (2):157–84. doi: 10.1080/00324728.1983.10408745.