Mapping migration regions and their evolution from population-scale family trees: What can they tell us about cultural identities and regions today?

Caglar Koylu†[1], Maryam Torkashvand[1], Hoeyun Kwon[1], Alice Bee Kasakoff [2]

[1] Geographical and Sustainability Sciences, University of Iowa, Iowa City, Iowa, USA
[2] Department of Geography, University of South Carolina, Columbia, South Carolina, USA
caglar-koylu@uiowa.edu

ABSTRACT

Using a population-scale family tree dataset, this paper proposes a study of migration regions and their evolution in the U.S. between 1789 and 1924. To extract migration events, we use the child ladder approach, which traces family moves based on changes in birthplaces of consecutive children in each individual family. We calculate a time series measure of migration rate and partition the time into optimal periods so that each period has a distinct migration network. We apply community detection to derive migration regions from each network of different periods. We map these regions and use a pair-counting measure to statistically compare the similarity of regions in consecutive time periods. Migration regions reveal the extent to which the strong regional identities we see today, and, in the past, which were rooted in migration. The North/South divide was pervasive not only in the early periods but throughout U.S. history. Migration regions are important for understanding the development of regional and national cultural forms such as music, literature, foodways, and dialects, as well as political divisions and events.

CCS CONCEPTS

• Information systems • Information systems applications • Spatial-temporal systems

KEYWORDS

Family trees, Migration, Migration and Cultural Regions, Spatio-temporal Network Analysis

ACM Reference Format:
Caglar Koylu, Maryam Torkashvand, Hoeyun Kwon, Alice Bee Kasakoff. 2022. Mapping migration regions and their evolution from population-scale family trees: What can they tell us about cultural identities and regions today? In *6th ACM SIGSPATIAL Workshop on Geospatial Humanities (GeoHumanities'22), November 1, 2022, Seattle, WA, USA*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3557919.3565814

1 INTRODUCTION

Publicly available historical sources and genealogy websites have made it possible for people to compile and share their family trees. A family tree is a network of family members stretching over many generations. When there are places associated with vital events, as there are in a large proportion of the entries, a family tree becomes a network in both time and space. Through marriage and birth of children trees overlap. In our previous work, we cleaned, geocoded, and connected publicly shared family trees on rootsweb.com to develop a population-scale kinship network [1]. To date, this network is the largest population-scale kinship network available. Its largest connected component contains nearly 40 million individuals and spans over centuries. Our study was among the first of its kind to formally evaluate the representativeness of user-contributed big data from family trees with ground truth data (1880 Census). Our results, which were derived based on aggregate statistics of demographics from 1880 Census, showed that the trees represent the native-born white population of the U.S. quiet well, which accounts for 72% of the total population in 1880. But there are probably very few non-Whites in the trees and fewer foreign-born Whites than in the general population. Using this dataset, we introduced a methodology to measure and map long-term changes in interstate migration flows in the U.S. between 1789 and 1924 [2]. Despite a few scholars who used information from family trees to study migration and population dynamics [3-6], our study has been one of the first studies to uncover dynamic migration patterns on a large spatial and temporal extent. To extract migration from family tree records, we used the child-ladder approach which traces family migration using changes in birthplaces of consecutive children in each individual family. To identify the long-term changes that span across three centuries, we first developed a time-series measure of the family migration rate by dividing the number of family migration events by the number of birth events. We then partitioned the study period into discrete sections to create a small number of aggregated networks that summarize a complex process of change over the 135 years. Our evaluation of changes in migration rate and similarity of flows between time periods provide an overview of the long-term changes in migration flows. However, this tells us little about the structure of the migration network and how it changed over time. In this paper, we aim to reveal structural patterns, specifically, migration regions and how these regions

†Caglar Koylu is the corresponding author for this paper

GeoHumanities'22, November 1, 2022, Seattle, Washington USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9533-5/22/11. ...$15.00
https://doi.org/10.1145/3557919.3565814

changed over time building upon our work on measuring and mapping migration using family trees [2].

Migration flows naturally form regions, which are meaningful groupings of areal units based on the structure of migration [7]. Regions are often a product of economic, social, and cultural factors that produces a certain level of homogeneity within each region. These regions evolve over time both changing their social and cultural characteristics and borders. Clark [8] studied the temporal stability of migration regions in the U.S. and found that the regions were quite volatile over even short periods of time due to the adjustments to macroeconomic fluctuations and its geographical structure. Here we apply a similar approach, but for much longer periods. Although distance has been considered as a major influence on migration and the formation of migration regions, migrants may often be attracted to distant locations. For example, during the California Gold Rush, settlers from the eastern U.S. moved to the West in search for gold. Family ties also have been shown to play a significant role in chain migration and attracting individuals to far away locations. Therefore, migration regions may have been formed by groups of areas that are geographically distant from each other. Despite this fact, migration studies have employed regionalization methods that constrained regions to be geographically adjacent to each other [9]. In this preliminary study, we aim to derive regions which may include geographically distant places because of their strong migration ties among each other.

2 PRELIMINARY RESULTS

In our previous work, we identified the seven optimal time periods that captured the largest changes in migration rate between 1789 and 1924 [2]. We use the state-to-state migration flow matrix from each of these periods to obtain the migration regions. We perform the Louvain community detection method [10] to extract the regions that may be formed by geographically disjoint states. The Louvain method groups nodes (states) in a network into communities (partitions or regions) by optimizing modularity iteratively.

Figure 1 illustrates the migration regions and their corresponding modularity values for each period. Although the local maximum modularity values are not comparable between the networks and regionalization for different periods, they consistently partition the network into 3 (for the earlier periods with a smaller number of states), 4 and 5 communities. Although the Louvain method does not enforce spatial contiguity constraint, the resulting regions are mostly formed by spatially contiguous states. These maps show the extent to which the strong regional identities we see today, and, in the past, which were rooted in migration. The North/South divide was pervasive not only in the early periods but throughout U.S. history. It should be noted that our study is at the level of states, but the North/South divide ran through many states such as Indiana, Ohio, and Illinois, which were divided culturally. Some of those states moved between regions in the later periods. The North/South divide persisted during the settlement process so that the Southern part of the West was part of the Southern bloc, and the Northern part was part of the Northern bloc up until the Civil War in 1862. These regions were formed by complex human-environmental interactions, and cultural and social processes. Fischer [11] traced such latitudinal regions back to the colonies’ different origins within England, seeing it reflected in local government, housing types and diet [12]. Steckel [13] noted that the varieties of corn at that time had specific growing seasons and could not be grown outside specific latitudes. The West separated off after the Civil War. At first it included only those states that were in the Northern bloc but later those originally part of the Southern region gradually joined the West. “Middle” region, often found by dialectologists, emerge in the last two periods. In 1887-1901 period, western states of California, Oregon and Washington are in the same region with the states in the middle region that include mainly Great Plains and Midwest. The western and eastern parts of this region (orange) is geographically separated from each other by Idaho, Wyoming, Utah, and Nevada. In the last period of 1901-1924 the middle region is split into its northern (yellow) and southern (blue) parts.

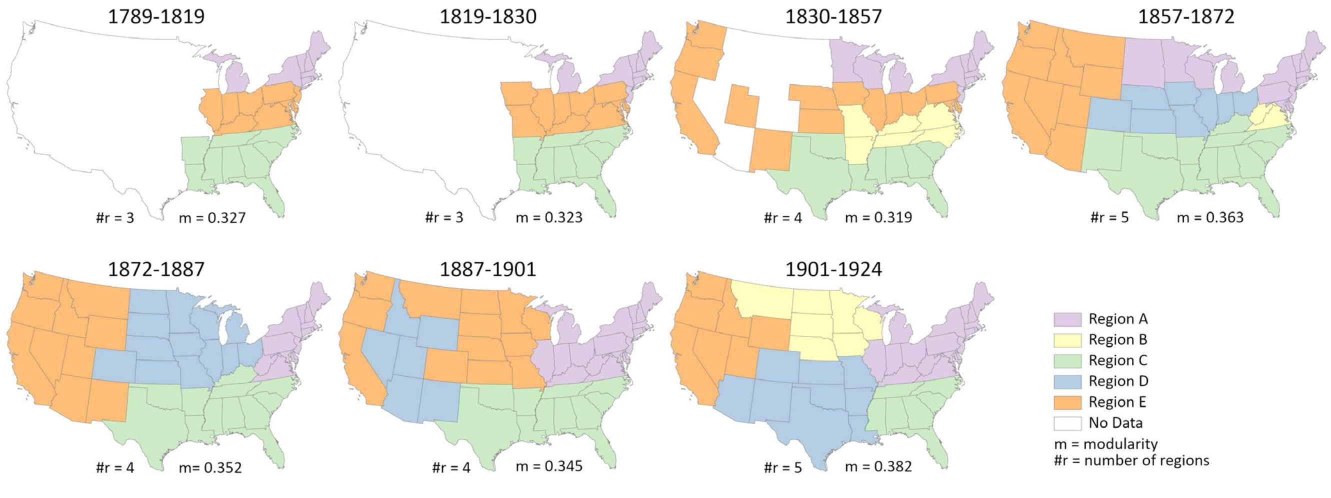

Figure 1: Temporal evolution of migration regions from the family tree data (1789-1924).

After visually examining the regions and their change over time, we used the z-Rand measure [14] to quantify the degree of similarity between regions over time (Table 1). The higher the z-Rand score the more similar the regions are between two consecutive periods. In Table 1, we only compare the regions with the same number of nodes (states) such as the regions of the first two periods and the regions of the last four periods which have about the same number of states. Overall, the z-Rand values confirm our interpretation of migration regions in Figure 1. Among all periods, structural similarity of the migration regions in 1857-1872 and 1872-1887 are the highest followed by 1887-1901 and 1901-1924, and 1872-1887 and 1887-1901 periods. In addition, regions of the first two periods, 1789-1819 and 1819-1830 are also very similar to each other. The reason for lower z-Rand value can be explained by smaller number of states in earlier periods.

Table 1: Z-Rand values

Periods for comparison	zRand
1789-1819 vs. 1819-1830	13.14
1857-1872 vs. 1872-1887	24.67
1872-1887 vs. 1887-1901	22.22
1887-1901 vs. 1901-1924	24.11

Figure 2 illustrates the migration of families between regions. Nodes are placed at the geometric centroid of regions, and the curved flow lines with half arrowheads illustrate the total volume of flows between regions. The choropleth base map depicts the migration efficiency, which is derived by dividing the netflow (i.e., subtraction of inflow from outflow) by the gross flow volume (i.e., the sum of inflow and outflow) for each region. Red hues illustrate regions with negative migration efficiency, which send more migrants than they receive. On the other hand, blue hues illustrate regions with positive migration efficiency, which receive more migrants than they send. The western region is depicted with dark blue in 1857-1872 period and attracts flows from all other regions although the volume of flows is relatively lower. Light blue region with the east-west band accounts for most of the migration that happened in longer distances in this period. Emergence of this region could be attributed to the growing expansion of railways to the West and the completion of the first transcontinental railroad of the U.S. in 1869. The West and the South become the destinations for migration in the 1872-1887 period. During the 1887-1901 period, Pacific states were regrouped with northern states of the Midwest and the Great Plains, and this region became a region that send large volume of migrant families to the Northeast and the South. However, apart from the population redistribution that happened in the Great Plains and Midwest states of this larger disjoint region, Pacific states continued to be the destination of migrant families during this period [2]. In the last period of 1901-1924 the West and Southwest become the major attraction for family migration. We also observe more of a polycentric structure of the regional migration network in which there are bidirectional flows between all regions. These flows also highlight the pattern of

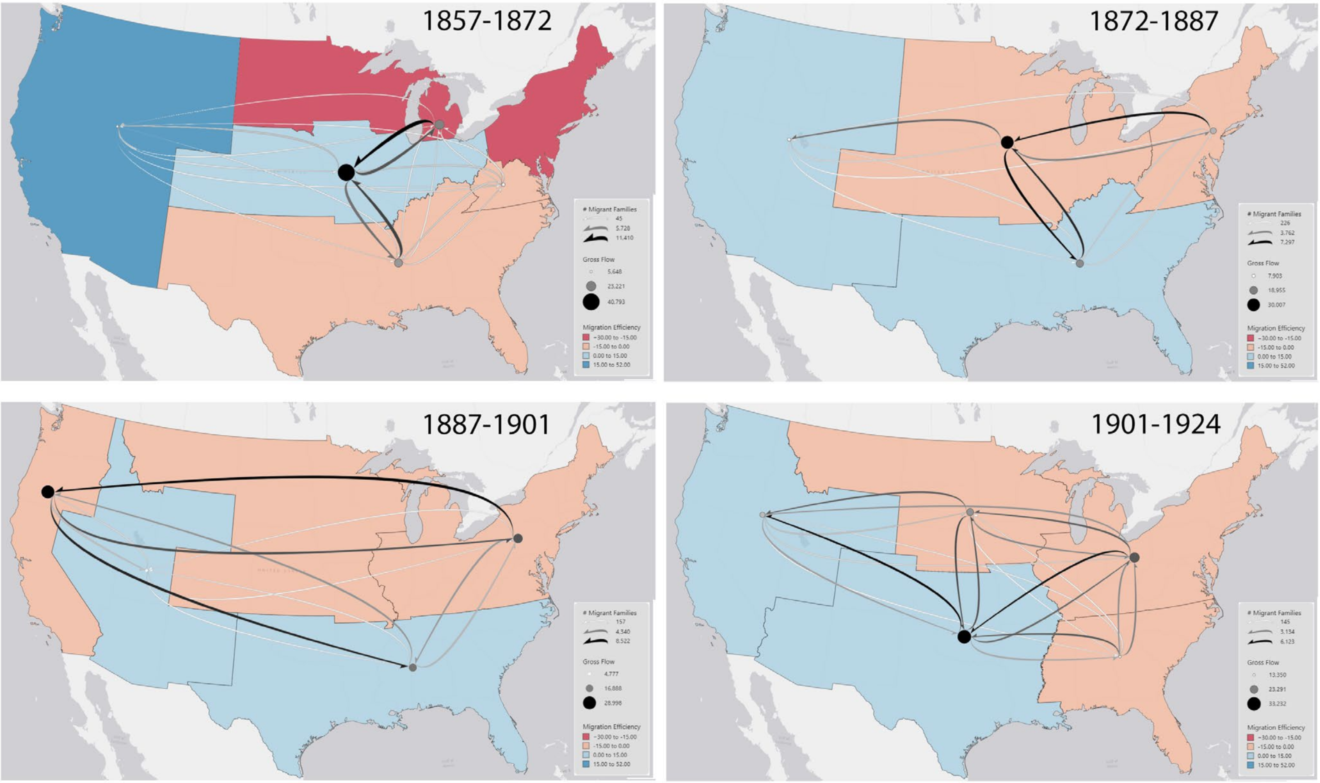

Figure 2: Family migration between regions over the four time periods

population redistribution after the Frontier ended. Across all periods, migration flows appear to be in all directions east to west, north to south and south to north. However, because these regions are very large, flows appear to be longer could in fact be a short distance flows between adjacent states that belong to two regions that cover very large geographic areas. To distinguish between short and long-distance flows, one can visualize state-to-state flows between regions. It would also be meaningful to show flows within these regions, which would illustrate the network structure that forms the regions.

3 LIMITATIONS AND FUTURE WORK

One of the limitations of our study is the state-level analysis, which disregard moves within states and between places among contiguous state borders. There are many cities and settlements in bordering states that are strongly connected to each other than the rest of settlements in their origin states. We plan to geocode birthplace locations at finer spatial resolution such as county or city. This will allow to distinguish regions that do not necessarily correspond with the state borders.

Modularity is a commonly used quality function, which includes all key components and issues to express the "strength" of communities, however, there are drawbacks associated to modularity such as the resolution limit [15]. The modularity optimization may lead to the clustering of smaller communities into bigger communities, especially for the modules with small number of internal links. On the other hand, the iteration process may cause production of arbitrary communities with poor connections [16]. During the algorithm iteration, some nodes that are known as bridges may move and so the badly connected communities may be created. To address any biases in the modularity optimization, we plan to perform the Constant Potts Model [17]. We also plan to use the Leiden algorithm to address another artefact of the Louvain method, which sometimes generates arbitrarily badly connected communities [16].

Migration regions are important for understanding the development of regional and national cultural forms such as music, literature, foodways, and dialects, as well as political divisions and events. The family tree data allow researchers to study the emergence of these forms against the backdrop of migration. Our data combined with longitudinal voting and other cultural data can help uncover the processes by which regions develop within countries and how they changed with changes in transport and media communication. For example, since the West originally did not include the southwestern states, were the cultural attributes we now associate with the West originally from the North or did the Southwest bring new attributes and did they spread to the rest of the region? In an earlier study Grieve et al. [18] and Huang et al. [19] used Twitter to map regional dialects and the spread of newly coined words. Similarly, Koylu [20] derived regions of interpersonal communication on Twitter, which corresponded with dialectal and cultural regions. The regions based on migration will allow us to see how the more recent dialect regions were influenced by the earlier family migration into and out of regions. As archives of longitudinal and georeferenced humanities data such as movies, carnivals, and radio broadcasts become available on a large scale, we can compare the geography of cultural material to migration regions to see to what extent the circulation of cultural forms was limited within the regions created by migration or whether they perhaps played a role in leading migrants to particular locations.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Koylu, C., Guo, D., Huang, Y., Kasakoff, A. and Grieve, J. 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 (2021/12/02 2021), 2380-2423.

[2] Koylu, C. and Kasakoff, A. Measuring and mapping long-term changes in migration flows using population-scale family tree data. *Cartography and Geographic Information Science*, 49, 2 (2022), 154-170.

[3] Han, E., Carbonetto, P., Curtis, R. E., Wang, Y., Granka, J. M., Byrnes, J., Noto, K., Kermany, A. R., Myres, N. M., Barber, M. J., Rand, K. A., Song, S., Roman, T., Battat, E., Elyashiv, E., Guturu, H., Hong, E. L., Chahine, K. G. and Ball, C. A. Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nature Communications*, 8, 1 (2017/02/07 2017), 14238.

[4] Otterstrom, S. M. and Bunker, B. E. Genealogy, migration, and the intertwined geographies of personal pasts. *Annals of the Association of American Geographers*, 103, 3 (2013), 544-569.

[5] Kandt, J., van Dijk, J. and Longley, P. A. Family Name Origins and Intergenerational Demographic Change in Great Britain. *Annals of the American Association of Geographers*, 110, 6 (2020), 1726-1742.

[6] Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M. and Gymrek, M. Quantitative analysis of population-scale family trees with millions of relatives. *Science*, 360, 6385 (2018), 171-175.

[7] Slater, P., B. Hierarchical regionalization of RSFSR administrative units using 1966-69 migration data. *Soviet Geography Review and Translation*, 16, 7 (1975), 453-465.

[8] Clark, G. L. Volatility in the geographical structure of short-run US interstate migration. *Environment and Planning A*, 14, 2 (1982), 145-167.

[9] Guo, D. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22, 7 (2008), 801-823.

[10] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008, 10 (2008), P10008.

[11] Fischer, D. H. *Albion's seed: Four British folkways in America*. Oxford University Press, 1991.

[12] Bailyn, B. *Voyagers to the West: A Passage in the Peopling of America on the Eve of the Revolution*. Vintage, 1988.

[13] Steckel, R. H. *The economic foundations of east-west migration during the nineteenth century*. 0898-2937, National Bureau of Economic Research, 1982.

[14] Traud, A. L., Kelsic, E. D., Mucha, P. J. and Porter, M. A. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53, 3 (2011), 526-543.

[15] Fortunato, S. Community detection in graphs. *Physics reports*, 486, 3 (2010), 75-174.

[16] Traag, V. A., Waltman, L. and Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, 9, 1 (2019), 1-12.

[17] Ronhovde, P. and Nussinov, Z. Local resolution-limit-free Potts model for community detection. *Physical Review E*, 81, 4 (2010), 046114.

[18] Grieve, J., Nini, A. and Guo, D. Mapping lexical innovation on American social media. *Journal of English Linguistics*, 46, 4 (2018), 293-319.

[19] Huang, Y., Guo, D., Kasakoff, A. and Grieve, J. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59 (2016), 244-255.

[20] Koylu, C. *Discovering Multi-Scale Community Structures from the Interpersonal Communication Network on Twitter*. In: Perez, L., Kim, EK., Sengupta, R. (eds) Agent-Based Models and Complexity Science in the Age of Geospatial Big Data. Advances in Geographic Information Science. Springer, Cham.