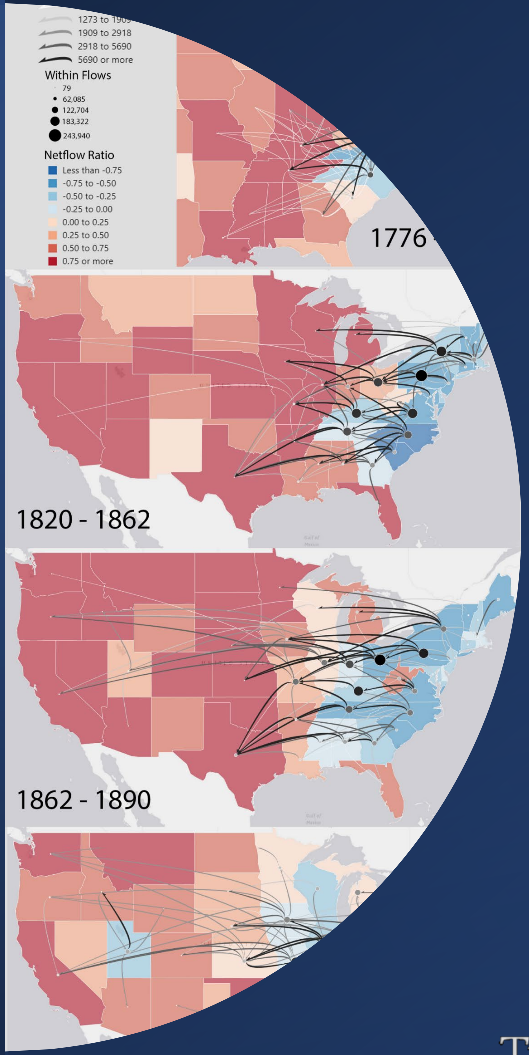
@caglarkoylu

Mapping Temporal Trends of Parent-Child Migration from Population-Scale Family Trees



Caglar Koylu* & Alice Kasakoff

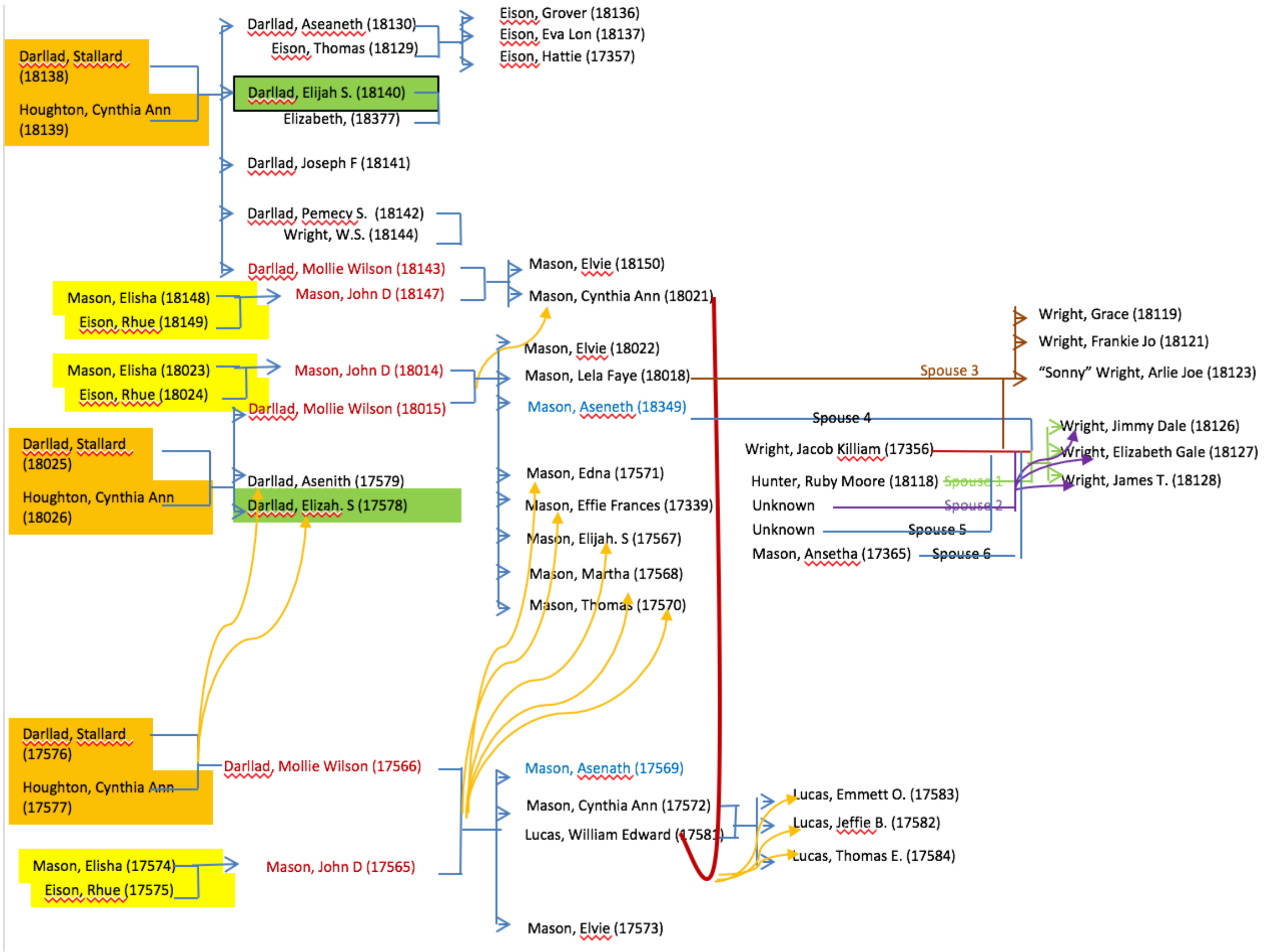
caglar-koylu@uiowa.edu

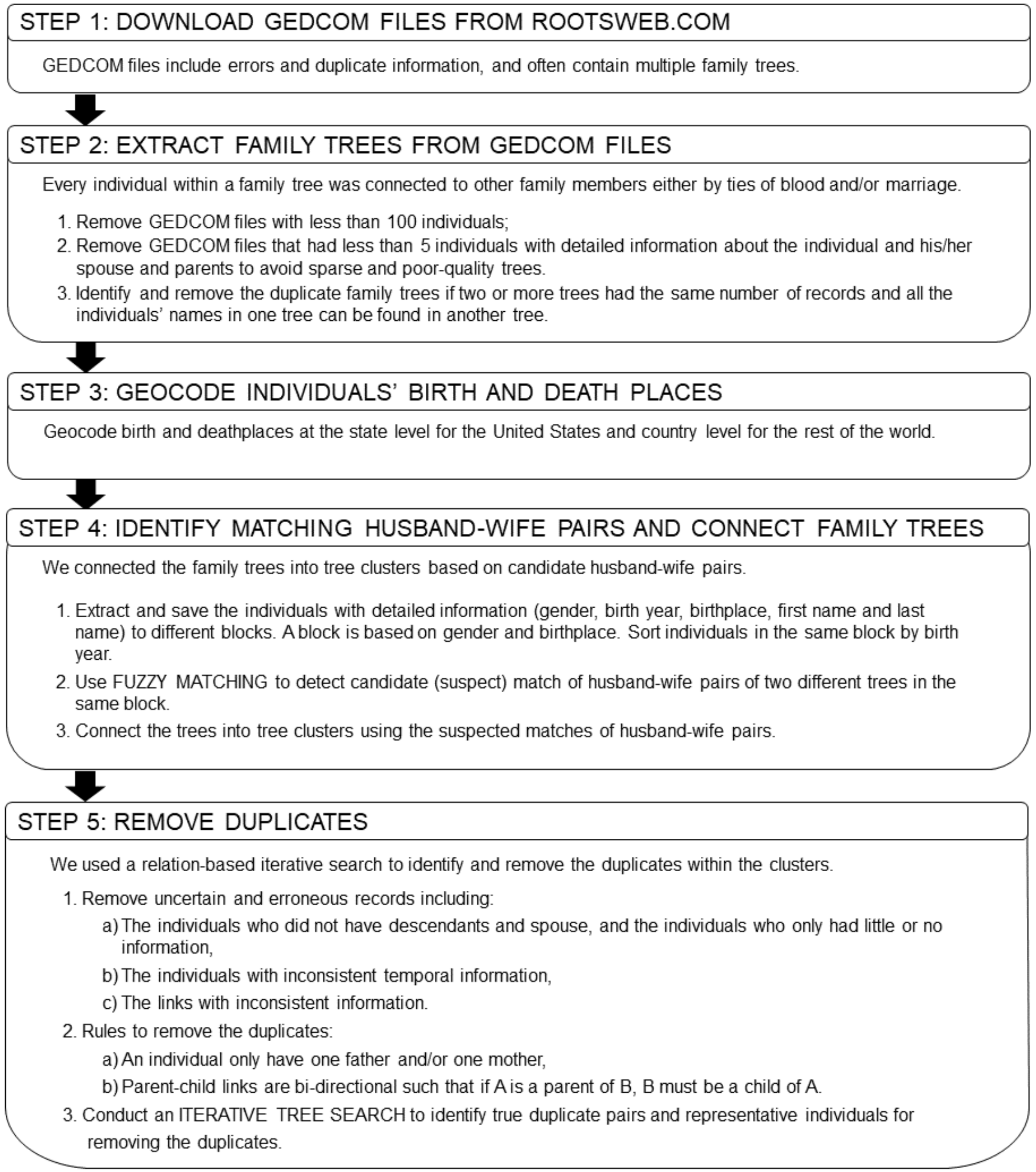
*Geographical and Sustainability Sciences

The University of Iowa

Koylu, C., & Kasakoff, A. (2020). Mapping Temporal Trends of Parent-Child Migration from Population-Scale Family Trees, *AutoCarto International Research Symposium,* November 17 -20, 2020 Redlands, California.

Uncertainty, inaccurate records, duplicates, representativeness,....





INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2020.1821885

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S.

Caglar Koylu[a], Diansheng Guo[b], Yuan Huang[b], Alice Kasakoff[b] and Jack Grieve[c]

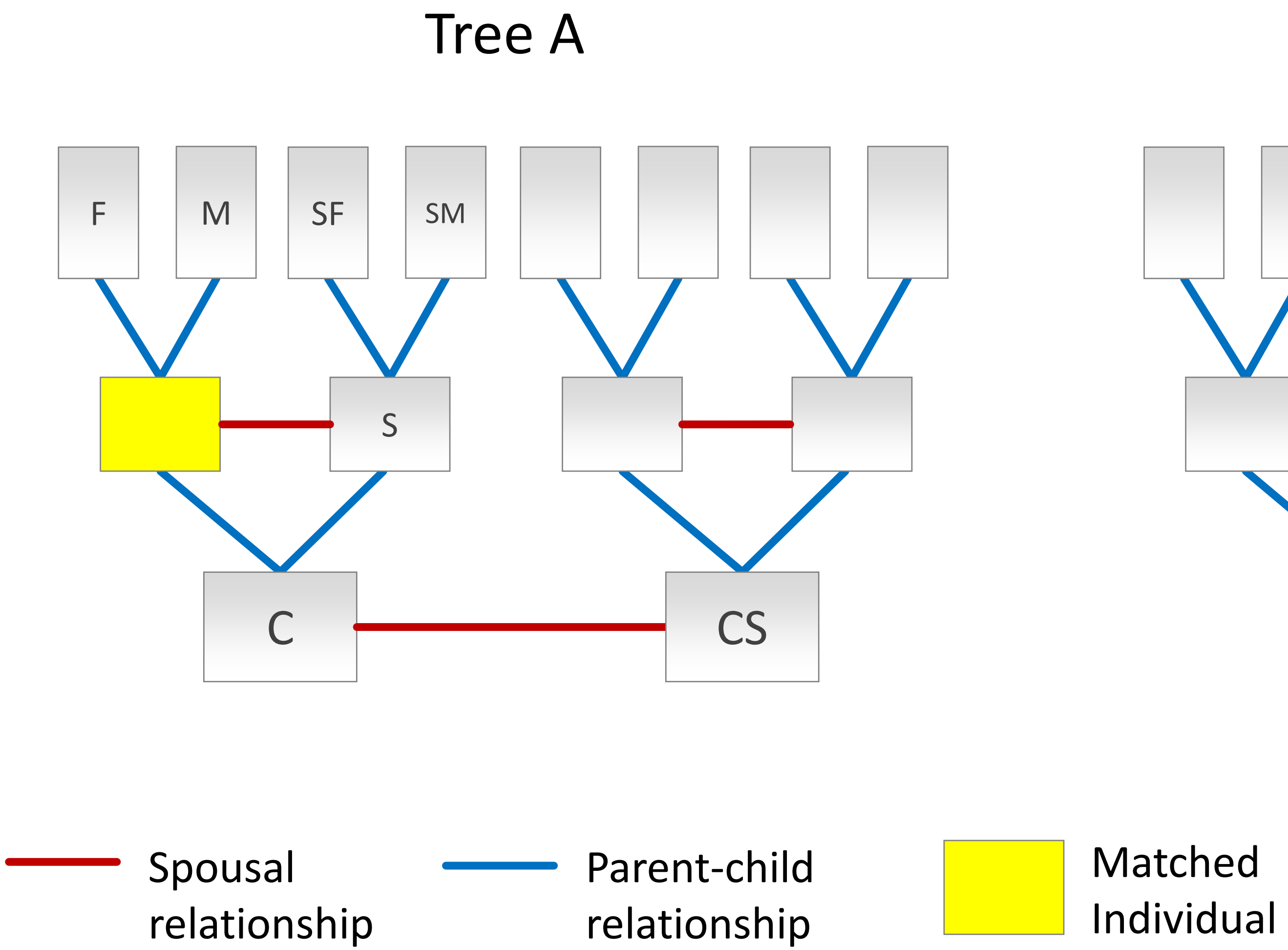
Fuzzy Record Linkage & Deduplication & Evaluation of Representativeness

Family Trees

- 92,832 GEDCOM files from Rootsweb
- One third of all data on Rootsweb because not all users want their data publicly shared
- ~248 million “individuals” and ~93 million families
- 1880 US Population = 50,000,000

Koylu et al. (2020) Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S., *International Journal of Geographical Information Science.* DOI: https://doi.org/10.1080/13658816.2020.1821885

Identifying matching spousal pairs



Deduplication and Cleaning of family trees

We clustered 87,308 trees into the following clusters:

1. 9,033 trees were stand-alone trees, which were not connected to any other tree
2. 2,866 trees formed 1,077 tree clusters
3. There were 75,409 trees in the largest connected cluster. After further cleaning **the largest connected component includes nearly 40 million individuals.**

Table 2. Deduplication and cleaning of the family tree clusters.

Family Tree Clusters	# trees/# clusters	# records	# duplicates	# inconsistent information	# cleaned records
Stand-alone trees	9,033	5,928,853	549,087	1,576,016	3,803,750
Small clusters	1,077	2,044,372	765,937	607,199	671,236
The largest connected cluster	75,409	231,700,575	67,918,977	88,083,932	75,697,666

Koylu et al. (2020) Connecting family trees to construct a population-scale and longitudinal geo-social network for the U.S., *International Journal of Geographical Information Science.* DOI: https://doi.org/10.1080/13658816.2020.1821885

Extracting migration from family trees

- Parents’ birth state or territory as the origin and the child’s birth state or territory as the destination.

- To reduce the bias of large families
  - We counted the four gender categories of parent-child relations once for those instances in which a parent had multiple children with the same birth state and gender.
  - If the same sex children were born in the same state, mother-child and father-child relations were counted only once.

Normalizing flows and geographic proximity

To account for the effect of geographic proximity and flow volumes in migration flows, we transformed the raw flows into modularity flows (Newman, 2006) using a double-constrained a gravity model (Roy & Thill, 2004).

$$\text{Modularity (i, j)} = \text{Observed Flows } (F_{ij}) - \text{Expected Flows } (E_{ij})$$

Modularity & Gravity Model

Modularity (i, j) = Observed Flows (F_{ij}) – Expected Flows (E_{ij})

Expected Flows: Double-constrained gravity model

The model constrains both origins and destinations and forces:

1. the sum of expected flows from an origin is equal to the observed
2. the sum of expected flows to a destination is equal to the observed volume of flows to that destination.

$$E_{ij} = \mathrm{A}_i * \mathrm{O}_i * \mathrm{B}_j * D_j * \mathrm{D}_{ij}{}^{-beta}$$

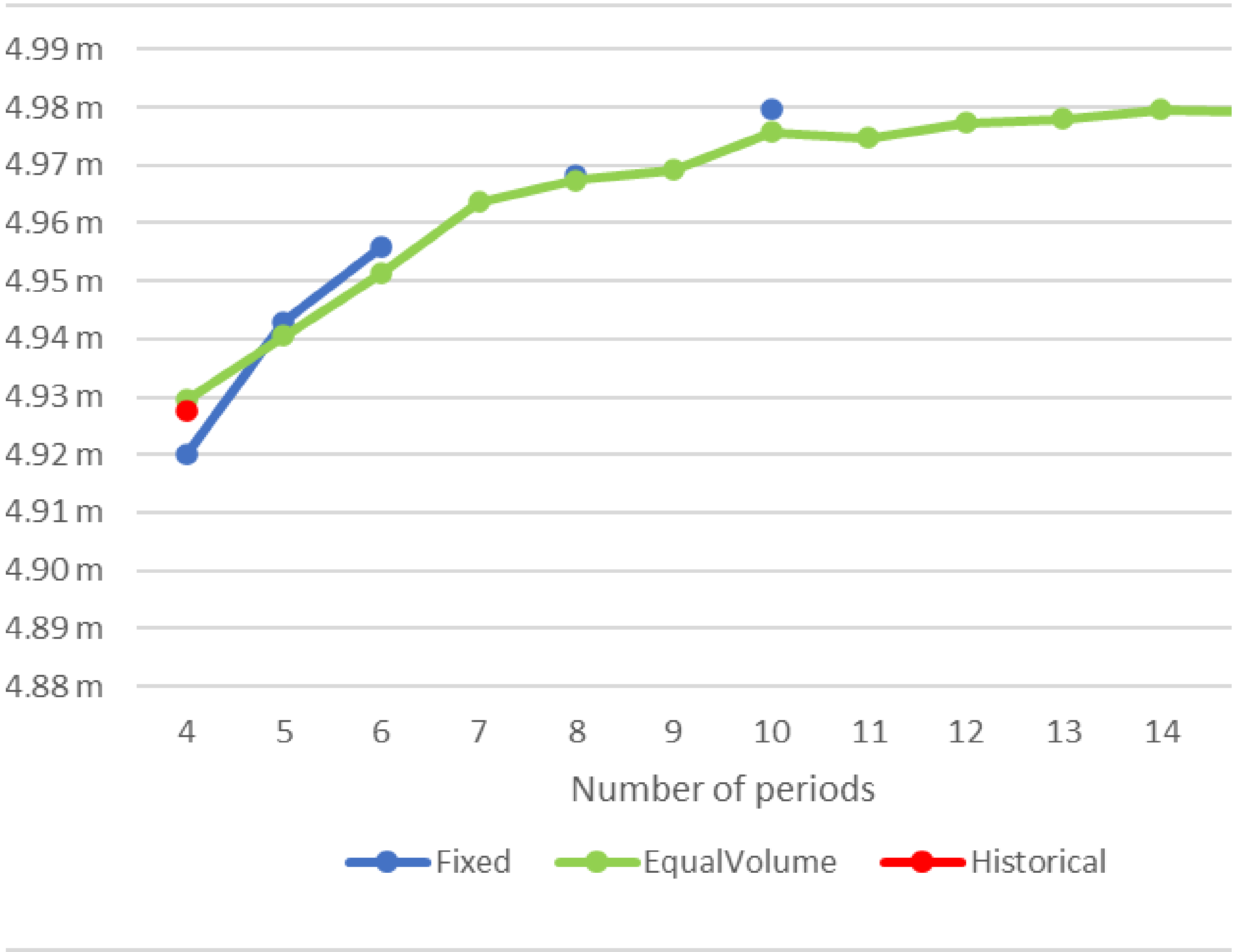
$$A_i = 1/\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\mathrm{B}_j\mathrm{D}_j * \mathrm{D}_{ij}{}^{beta}\right)$$

$$B_i = 1/\sum_{i=0}^{n}\sum_{j=0}^{n}\left(\mathrm{A}_j\mathrm{O}_j * \mathrm{D}_{ij}{}^{beta}\right)$$

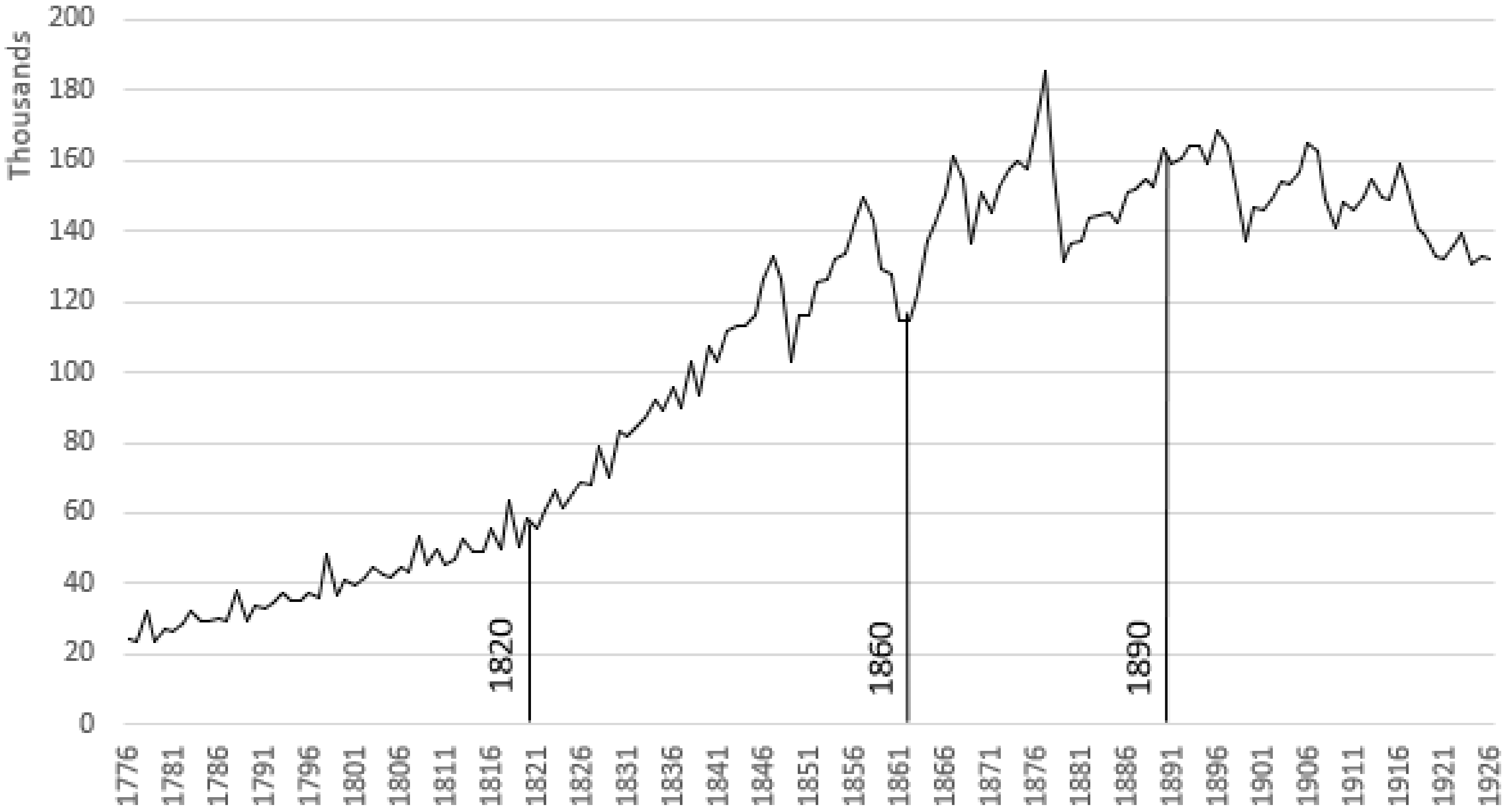
Temporal Partitioning

Total modularity of the three partitioning strategies using equal number of partitions:

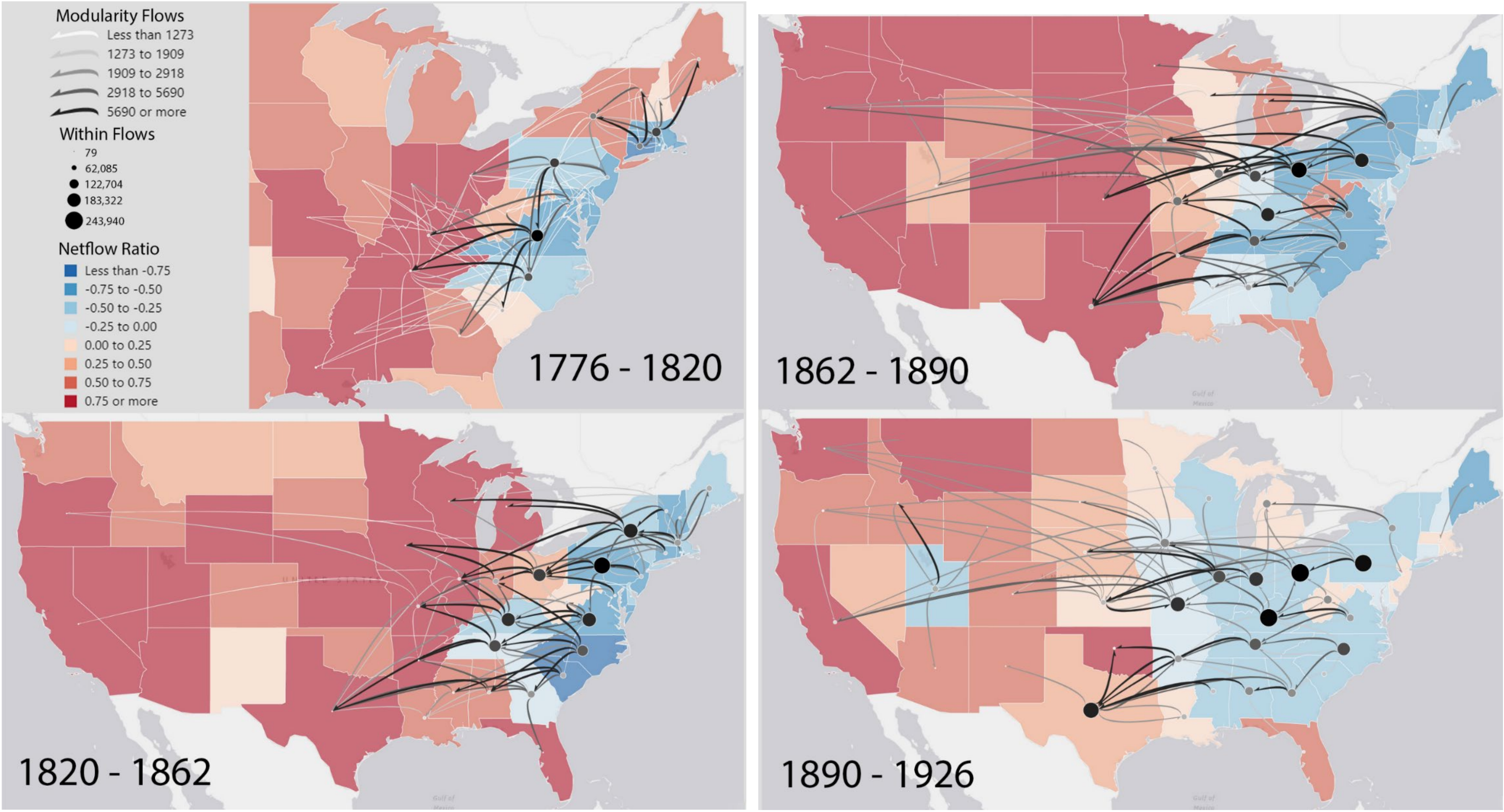
- historical periods
- fixed-length periods
- equal volume periods.



Migration Histogram & Historical Periods



Parent-Child Migration
in the U.S.
1776 - 1926



Modularity Flows
Less than 1273
1273 to 1909
1909 to 2918
2918 to 5690
5690 or more
Within Flows
79
62,085
122,704
183,322
243,940
Netflow Ratio
Less than -0.75
-0.75 to -0.50
-0.50 to -0.25
-0.25 to 0.00
0.00 to 0.25
0.25 to 0.50
0.50 to 0.75
0.75 or more
1776 - 1820
1862 - 1890
1820 - 1862
1890 - 1926

Conclusions

- Half the segment of the population whose parents had been born in the U.S. lived in a different state from where their parents had been born.
- The broad historical periods used by historians performed comparable to other partitioning methods.
- In a way, the importance of key events such as the Civil War and the closing of the frontier, has been validated through our comparison with other ways of partitioning time.

caglar-koylu@uiowa.edu

@caglarkoylu

Future Directions

- Use **the child-ladder approach** (Lathrop, 1948) to extract migration using changes in birthplaces of consecutive siblings in a family.
- Systematically evaluate the **changes in flow volumes and structures** using **temporal natural breaks**, **persistence measures** (Pamfil et al., 2019), and the goodness of absolute deviations from the median.
- Study **gender effects on migration** over time by disaggregating flows by gender into mother-daughter, mother-son, father-daughter and father-son relations and

Koylu, C., & Kasakoff, A. (2020). Mapping Temporal Trends of Parent-Child Migration from Population-Scale Family Trees, *AutoCarto International Research Symposium,* November 17 -20, 2020 Redlands, California.