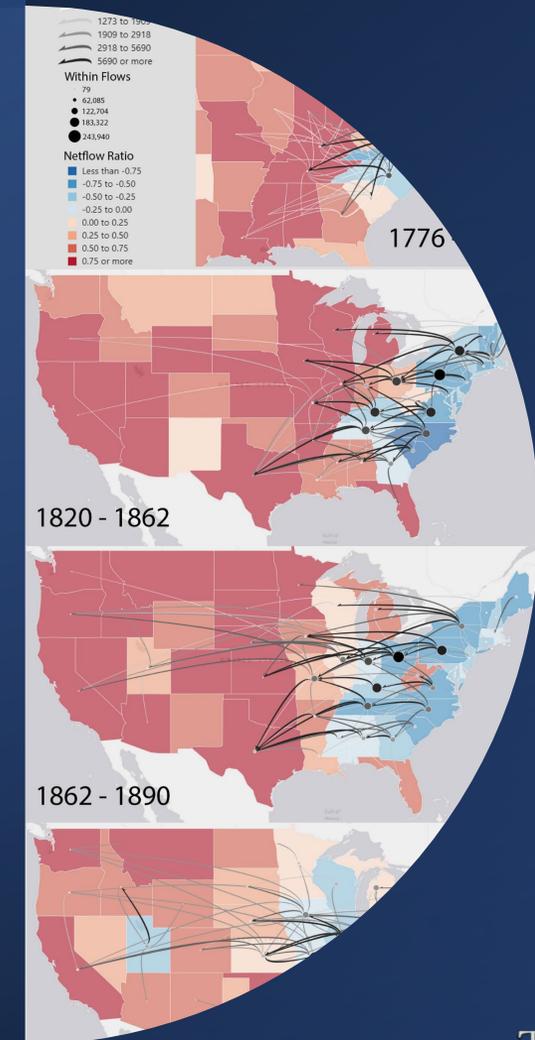


Mapping Temporal Trends of Parent-Child Migration from Population-Scale Family Trees

Caglar Koylu* & Alice Kasakoff

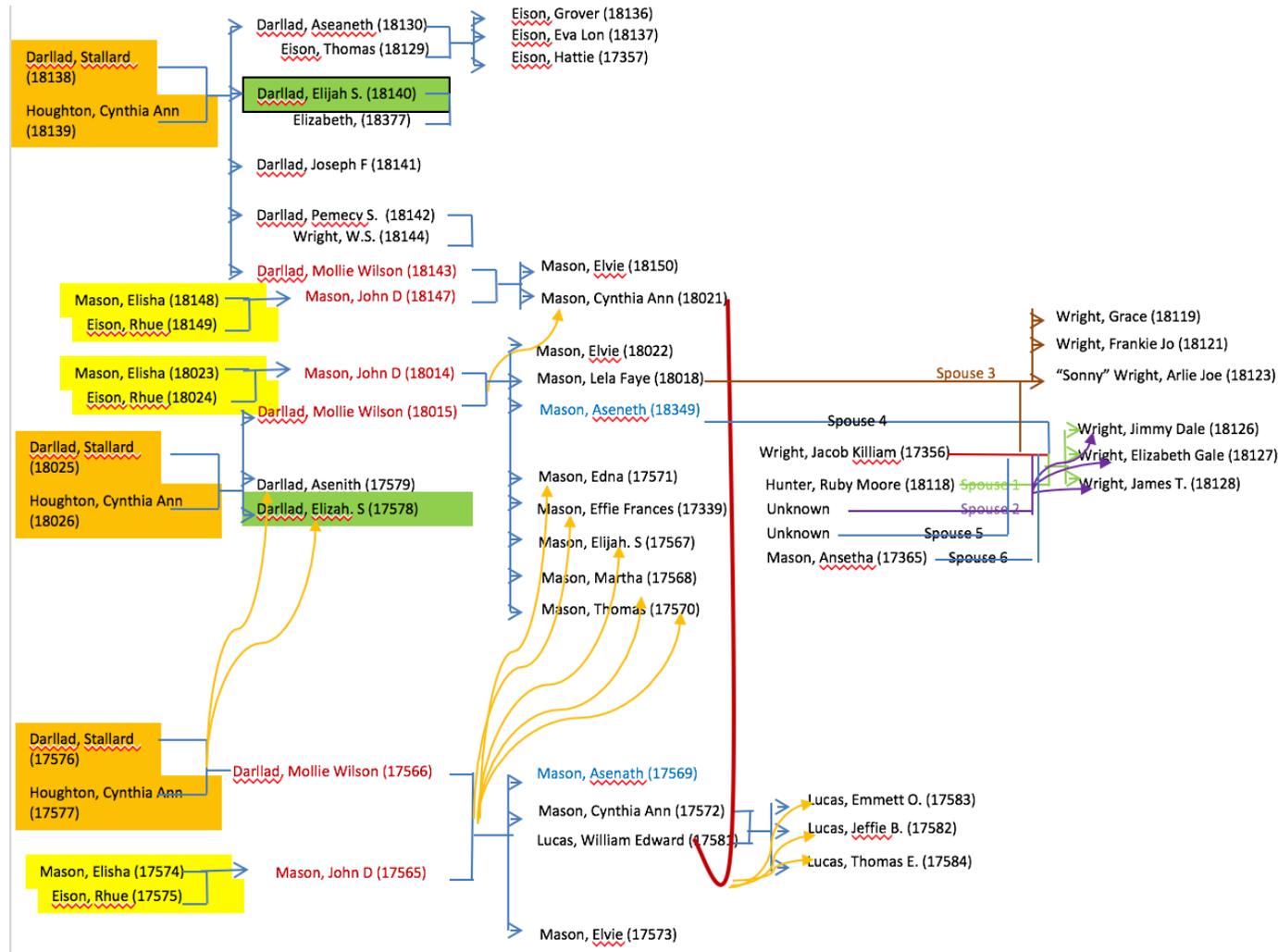
caglar-koylu@uiowa.edu

*Geographical and Sustainability Sciences



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,....



STEP 1: DOWNLOAD GEDCOM FILES FROM ROOTSWEB.COM

GEDCOM files include errors and duplicate information, and often contain multiple family trees.

STEP 2: EXTRACT FAMILY TREES FROM GEDCOM FILES

Every individual within a family tree was connected to other family members either by ties of blood and/or marriage.

1. Remove GEDCOM files with less than 100 individuals;
2. Remove GEDCOM files that had less than 5 individuals with detailed information about the individual and his/her spouse and parents to avoid sparse and poor-quality trees.
3. Identify and remove the duplicate family trees if two or more trees had the same number of records and all the individuals' names in one tree can be found in another tree.

STEP 3: GEOCODE INDIVIDUALS' BIRTH AND DEATH PLACES

Geocode birth and deathplaces at the state level for the United States and country level for the rest of the world.

STEP 4: IDENTIFY MATCHING HUSBAND-WIFE PAIRS AND CONNECT FAMILY TREES

We connected the family trees into tree clusters based on candidate husband-wife pairs.

1. Extract and save the individuals with detailed information (gender, birth year, birthplace, first name and last name) to different blocks. A block is based on gender and birthplace. Sort individuals in the same block by birth year.
2. Use FUZZY MATCHING to detect candidate (suspect) match of husband-wife pairs of two different trees in the same block.
3. Connect the trees into tree clusters using the suspected matches of husband-wife pairs.

STEP 5: REMOVE DUPLICATES

We used a relation-based iterative search to identify and remove the duplicates within the clusters.

1. Remove uncertain and erroneous records including:
 - a) The individuals who did not have descendants and spouse, and the individuals who only had little or no information,
 - b) The individuals with inconsistent temporal information,
 - c) The links with inconsistent information.
2. Rules to remove the duplicates:
 - a) An individual only have one father and/or one mother,
 - b) Parent-child links are bi-directional such that if A is a parent of B, B must be a child of A.
3. Conduct an ITERATIVE TREE SEARCH to identify true duplicate pairs and representative individuals for removing the duplicates.



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

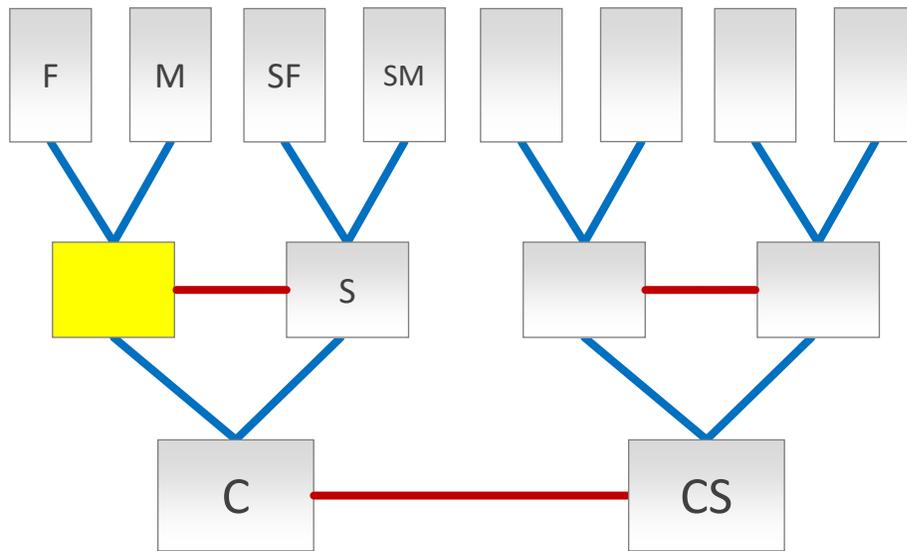
Fuzzy Record Linkage & Deduplication & Evaluation of Representativeness

Family Trees

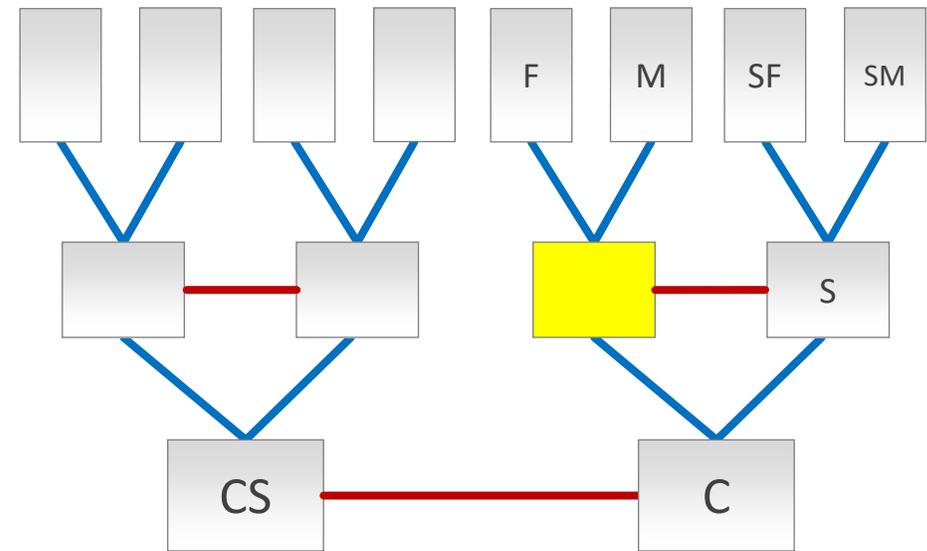
- 92,832 GEDCOM files from Rootsweb
- One third of all data on Rootsweb because not all users want their data publicly shared
- ~248 million “individuals” and ~93 million families
- 1880 US Population = 50,000,000

Identifying matching spousal pairs

Tree A



Tree B



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

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.

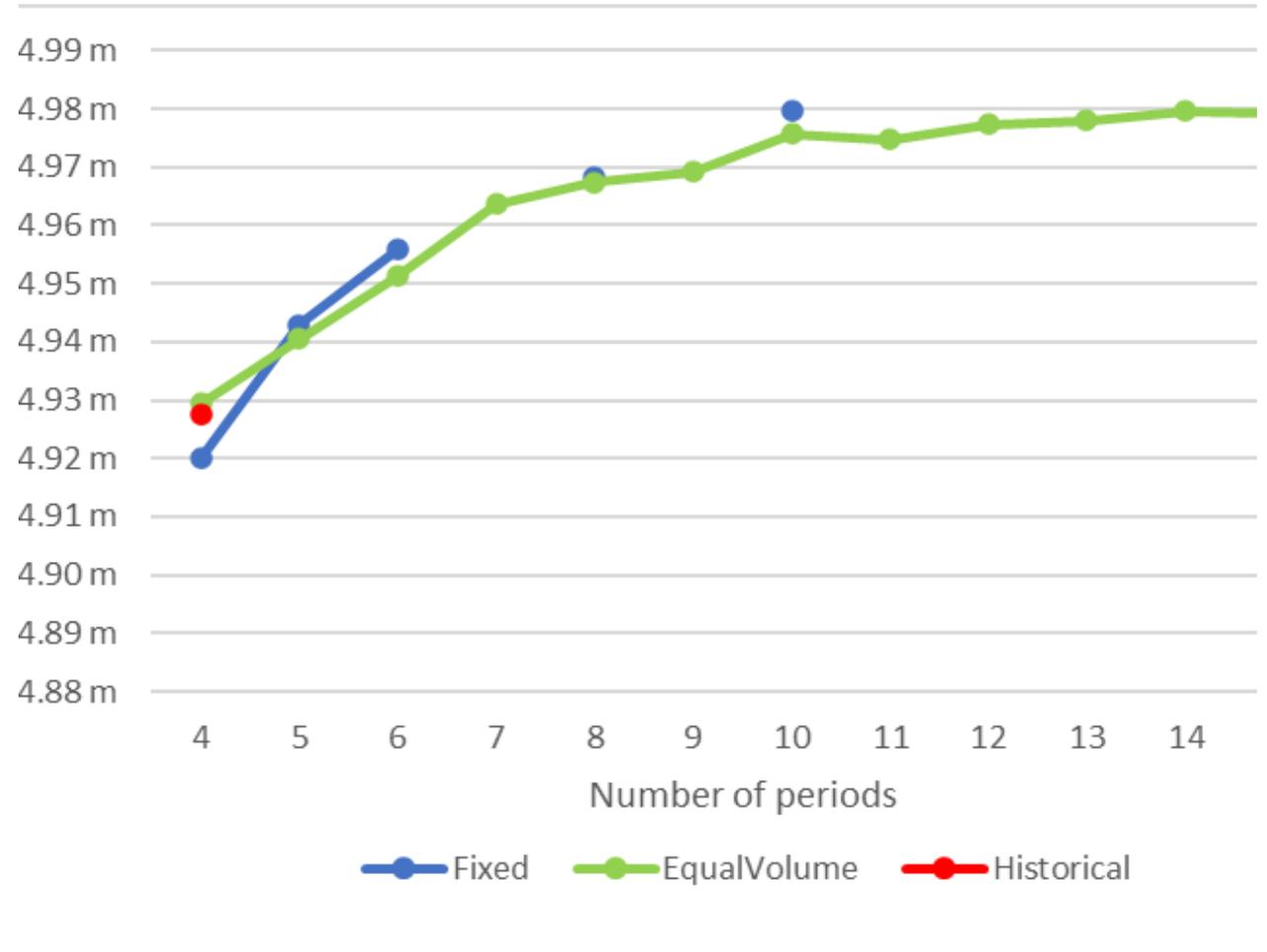
$$E_{ij} = A_i * O_i * B_j * D_j * D_{ij}^{-beta}$$

$$A_i = 1 / \sum_{i=0}^n \sum_{j=0}^n (B_j D_j * D_{ij}^{beta})$$
$$B_i = 1 / \sum_{i=0}^n \sum_{j=0}^n (A_j O_j * D_{ij}^{beta})$$

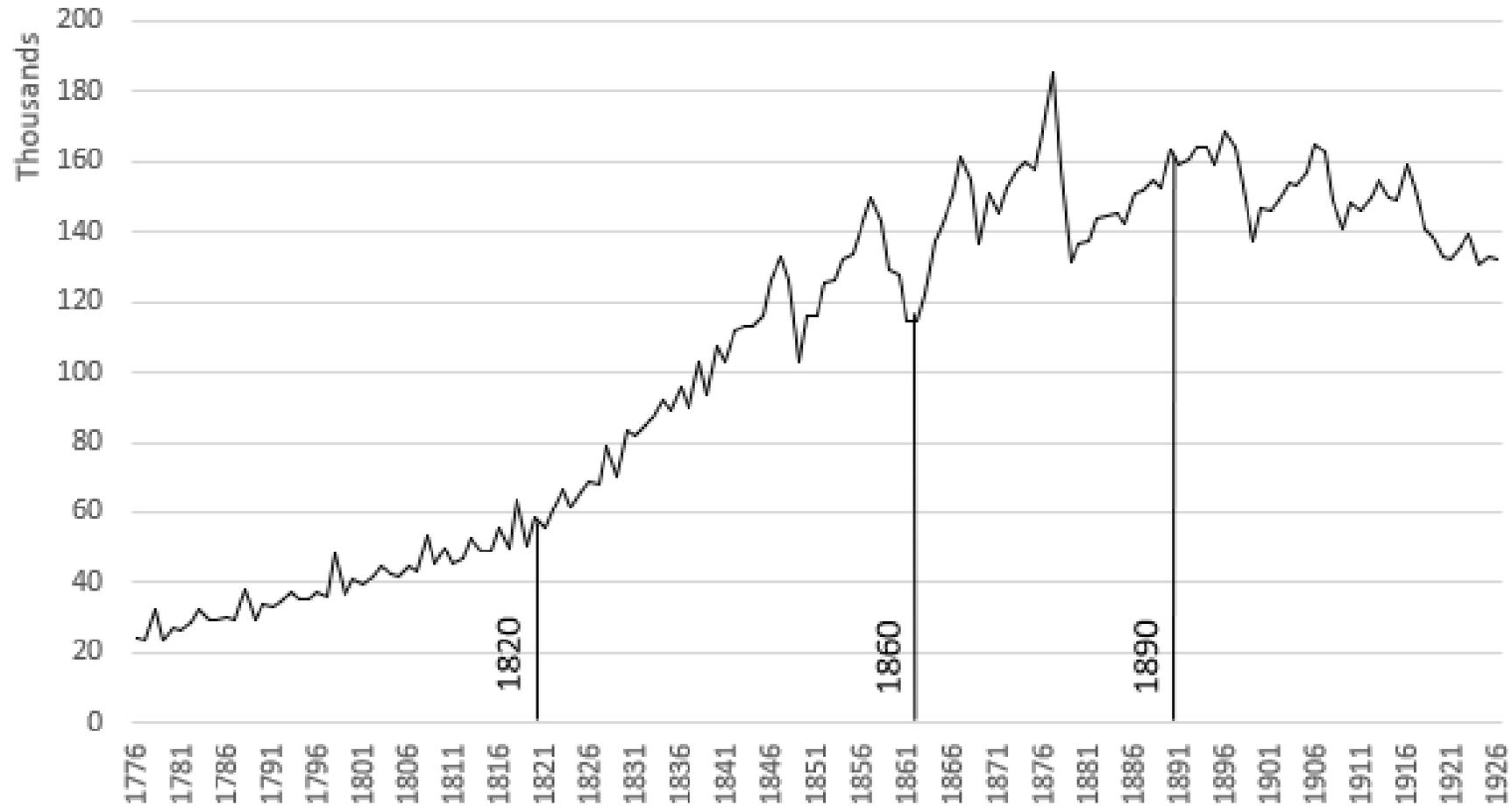
Temporal Partitioning

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

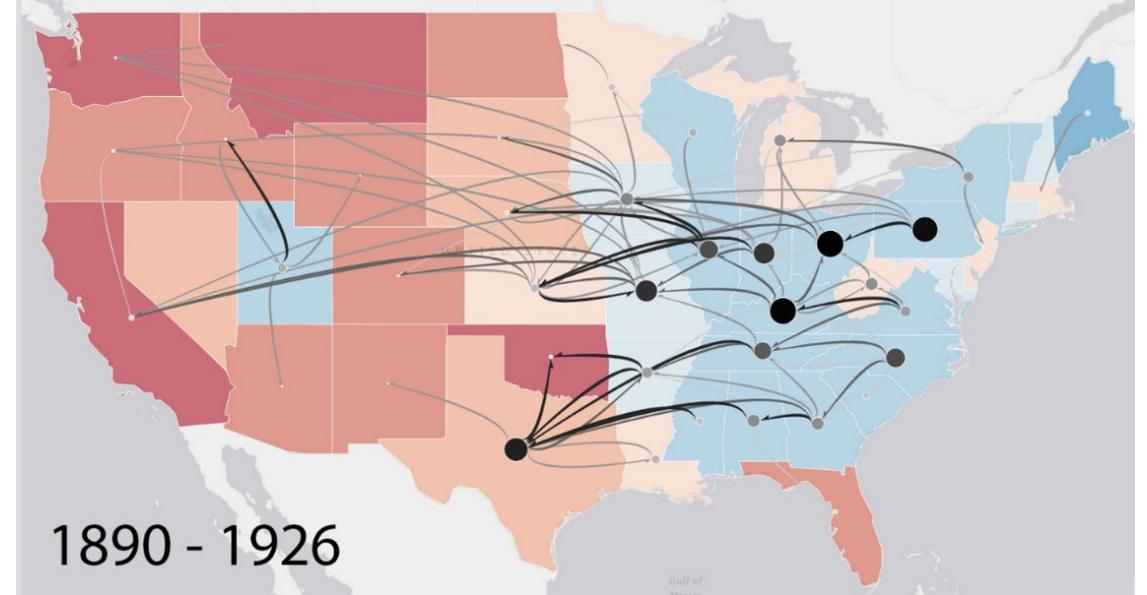
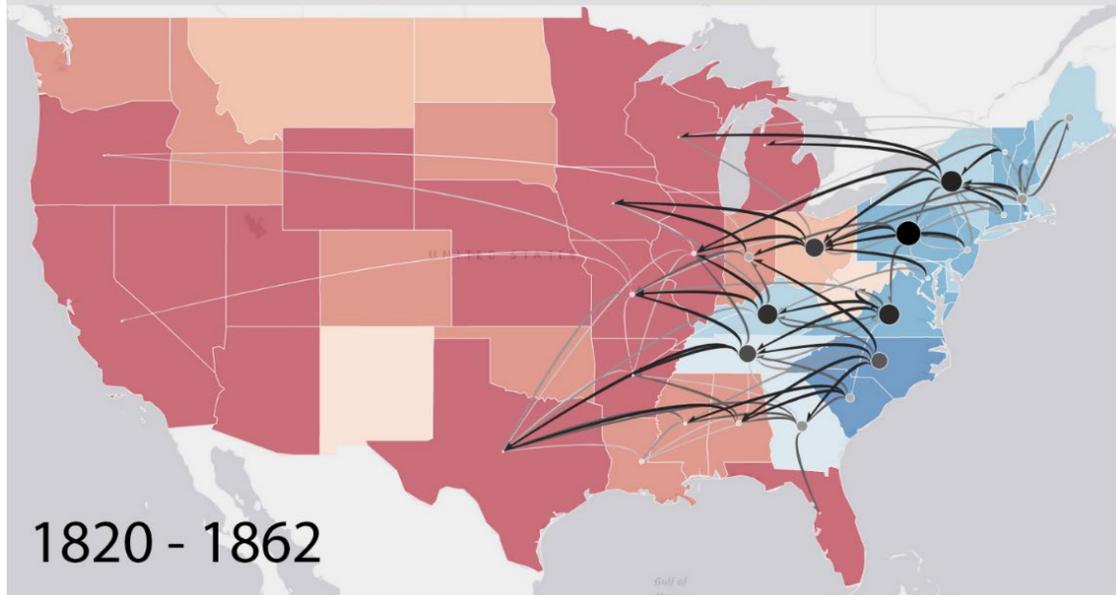
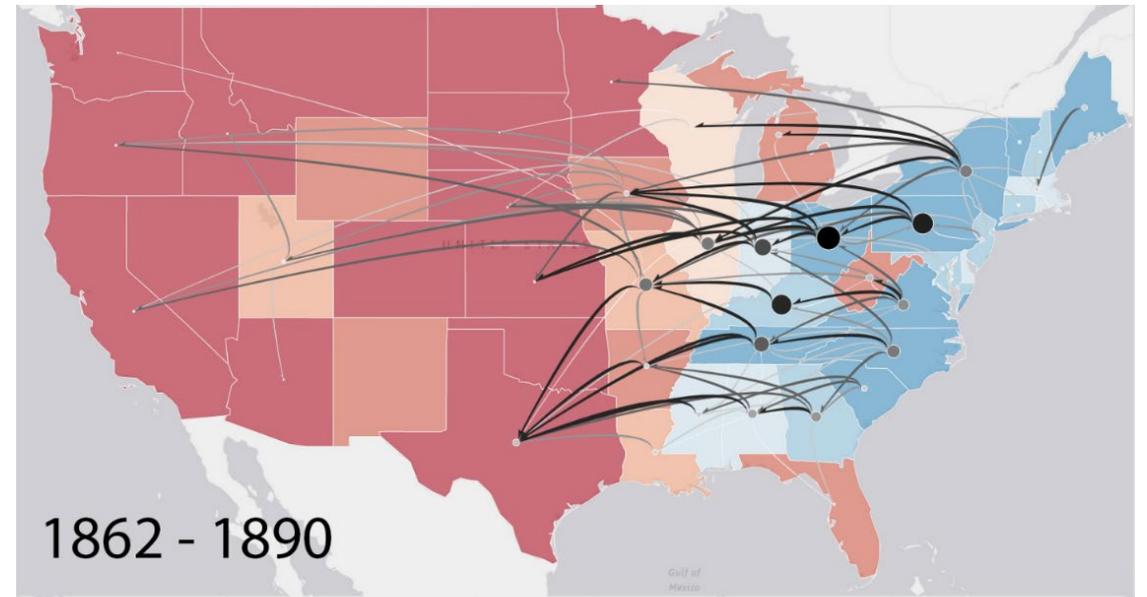
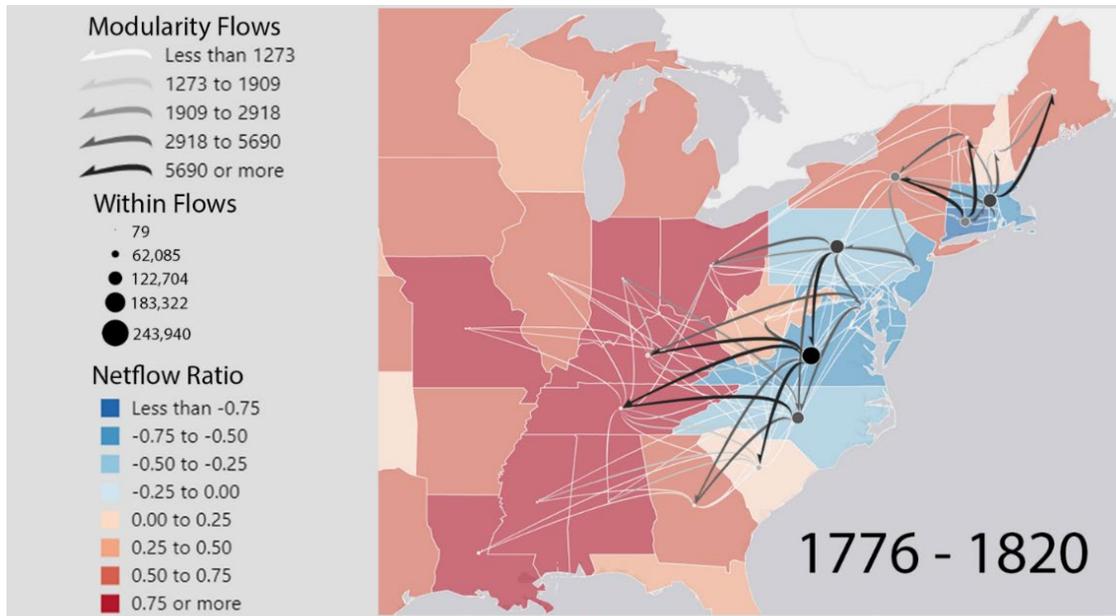
- historical periods
- fixed-length periods
- equal volume periods.



Migration Histogram & Historical Periods



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



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.

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