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

Caglar Koylu* & Alice Kasakoff

caglar-koylu@uiowa.edu

*Geographical and Sustainability Sciences

Uncertainty, inaccurate records, duplicates, representativeness,…
Fuzzy Record Linkage & Deduplication & Evaluation of Representativeness

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

Caglar Koylu, Diansheng Guo, Yuan Huang, Alice Kasakoff and Jack Grieve
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

Spousal relationship
Parent-child relationship
Matched Individual
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.

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).

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

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

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.

\[
E_{ij} = A_i \times O_i \times B_j \times D_j \times D_{ij}^{-\beta}
\]

\[
A_i = \frac{1}{\sum_{j=0}^{n} (B_jD_j \times D_{ij}^{\beta})}
\]

\[
B_i = \frac{1}{\sum_{j=0}^{n} (A_jO_j \times D_{ij}^{\beta})}
\]
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