# COVERAGE RATE OF THE FAMILY TREE

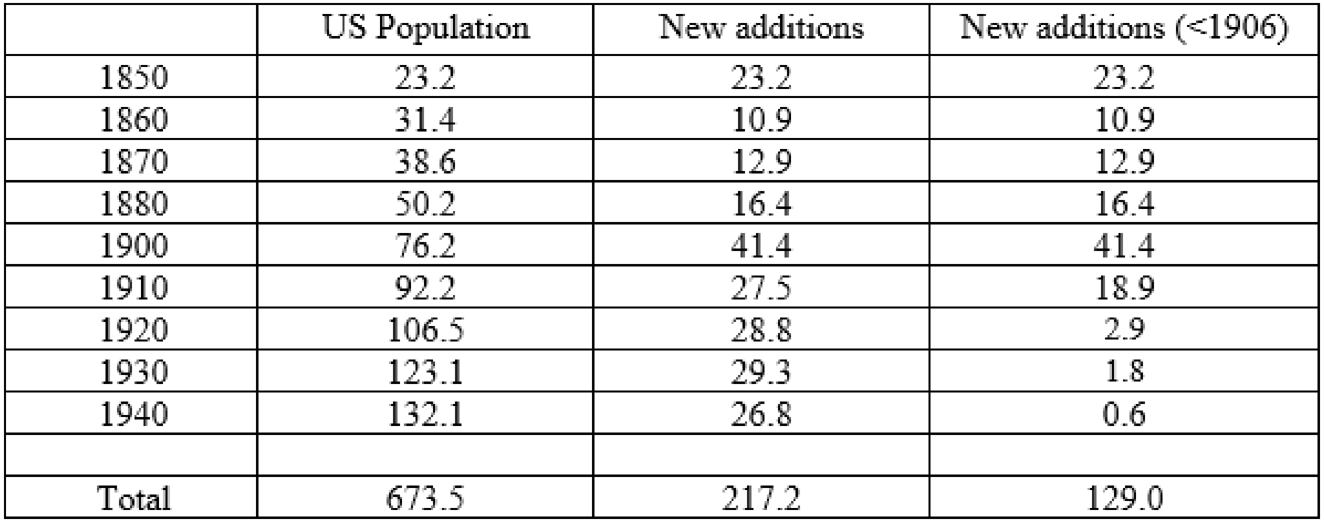
Joe Price
joe_price@byu.edu
rll.byu.edu

# Economic Research + Family History

- Economists are interested in: (1) economic and geographic mobility and (2) long-run determinants of health and wealth. Requires linking people/families across long periods of time.
- Large federal grants for these projects: NIH, NIA, NSF.
- Longitudinal, Intergenerational Family Electronic Micro-data (LIFEM). U. Michigan. $2.1M, NSF. Will link together the entire population of 5 states (including Ohio).
- Others in progress: American Longitudinal Infrastructure for Research on Aging (ALIRA); Census Longitudinal Infrastructure Project (CLIP).
- These projects could create natural partnerships with FamilySearch as a way to improve the Family Tree.

# BYU Record Linking Lab

- Our goal is to help complete the Family Tree for everyone who lived in the US between 1850 and 1940.
  - All deceased individuals found in a census record between 1850 and 1940 attached to the tree
  - All said individuals linked to parents, siblings, spouse, & children
  - Vital event and census records attached to FamilySearch profiles



| | US Population | New additions | New additions (<1906) |
|---|---|---|---|
| 1850 | 23.2 | 23.2 | 23.2 |
| 1860 | 31.4 | 10.9 | 10.9 |
| 1870 | 38.6 | 12.9 | 12.9 |
| 1880 | 50.2 | 16.4 | 16.4 |
| 1900 | 76.2 | 41.4 | 41.4 |
| 1910 | 92.2 | 27.5 | 18.9 |
| 1920 | 106.5 | 28.8 | 2.9 |
| 1930 | 123.1 | 29.3 | 1.8 |
| 1940 | 132.1 | 26.8 | 0.6 |
| | | | |
| Total | 673.5 | 217.2 | 129.0 |

# Pilot: Knox County, Ohio

- We started with Knox County, Ohio. We chose Ohio because of our collaboration on the LIFEM project and Knox County because it was the median size county in the state (population of 28,000 in 1900).
- First, we wrote a program to use the search feature on FamilySearch to check what fraction of these people were already on the Family Tree (found 30%).
- Second, we added the remaining families to the Family Tree. Volunteers then attached sources and expanded the family links until we connected them with someone already on the tree. Along the way we fixed misspelled names, incorrect dates, and merged any duplicates.

# Saturation Approach

- We found that working on the whole county at once allowed us to break through typical barriers by approaching the barrier from both directions.
- This also enabled us to split the work into micro-tasks (e.g. death dates, maiden names, parents, knots) that permitted the inclusion of volunteers possessing a full span of ability levels.
- This approach could even be used to involve a large group of volunteers from a single county to work on their own county. The google docs that distribute the work make it easy to even focus on a population with a specific last name.
- Benefits: allow everyone from that county to find family names, broaden the base of participants, and attract community members to the Family Tree

# Three Metrics of Success

- **Coverage**
  - What fraction of the target population is on the tree?
  - What changes have accelerated the work?
- **Quality**
  - How reliable is the information about the person?
  - Does quality improve over time?
- **Duplication**
  - What fraction of individuals show up multiple times on the tree?
  - Can we use what we learn from a saturated county to improve our possible duplicates algorithm?
- We will be quantifying each of these for well-defined populations and then keeping track of our progress.

# Coverage Rate

- Prior to starting this project, we asked people to predict what fraction of the Knox population in 1900 was on the Family Tree (the coverage rate).
- The predictions we got varied widely but sometimes were as low as 5%. I mentioned earlier that our initial search found 30%.
- We looked through the history of each person's profile to see when they were first added to the tree.
- What fraction do you think turned out to have been on the Family Tree prior to October 2016?

# Coverage Rate

- Prior to starting this project, we asked people to predict what fraction of the Knox population in 1900 was on the Family Tree (the coverage rate).
- The predictions we got varied widely but sometimes were as low as 5%. I mentioned earlier that our initial search found 30%.
- We looked through the history of each person's profile to see when they were first added to the tree.
- What fraction do you think turned out to have been on the Family Tree prior to October 2016?
- We found that 84% of the individuals living in Knox County in 1900 were already on the Family Tree

# Conclusion

- We will be doing this pilot for other counties and would love to partner with anyone that would like to help.
- The Family Tree is likely to be more complete than any of us imagines and it is getting better every day (a shared tree and open edit were key innovations).
- We can combine automated approaches with human volunteers to dramatically hasten the work. With a concerted effort we could complete the US part of the tree (217M) by 2020, in time for the 1950 census.
- This same approach could be used to hasten the work in other countries.

# COVERAGE RATE OF THE FAMILY TREE

Joe Price

joe_price@byu.edu

rll.byu.edu

Record Linking LAB