IPUMS

Multigenerational Longitudinal Panel

Life histories for the U.S. population, 1850-1940

- Censuses
- Social Security
- Military records (draft, enlistment)
- Vital records (birth, death, marriage, divorce)

IPUMS

Multigenerational Longitudinal Panel

Life histories for the U.S. population, 1850-1940

- Impact of early life conditions on later health and well-being
- Social, Economic, Geographic Mobility
- Life course transitions

IPUMS

Multigenerational Longitudinal Panel

Life histories for the U.S. population, 1850-1940

Link across 5+ generations

- Impact of forebears on health and well-being
- Socioeconomic mobility across generations: Do we have dynasties?

IPUMS

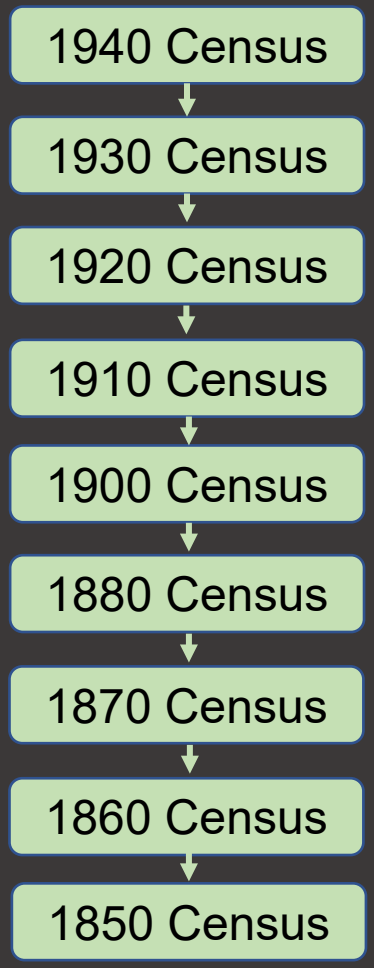
Multigenerational Longitudinal Panel

Life histories for the U.S. population, 1850-1940

Understanding the great transformations:

demographic transition, family transition,
urbanization, immigration, industrialization

IPUMS Multigenerational Longitudinal Panel

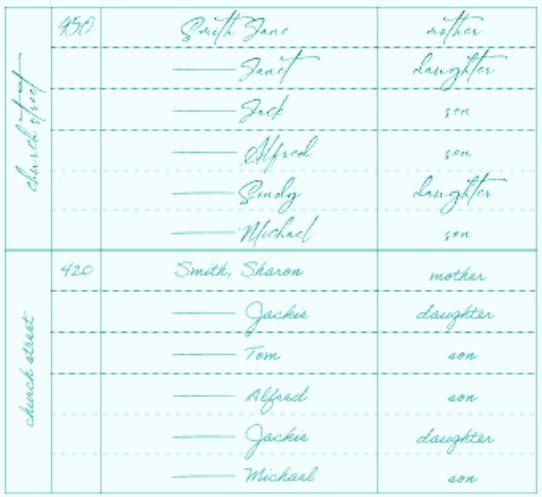


IPUMS MLP

- Availability of complete-count data makes it feasible to link most people with extremely low Type I errors, provided that we use all information available, including characteristics of others in the household, neighbors, and location
- This can introduce selection bias, but this bias turns out to no more severe than other linkage methods.

Census Linking Project

The Census Linking Project offers researchers the ability to create longitudinal datasets using historical US Census data (1850-1940). We provide links between each pair of complete-count Censuses using a wide variety of linking algorithms.



Get the Data

Download the crosswalk files →

Learn more

Watch our video series →

IPUMS USA
ABOUT
REGISTER
DONATE TO IPUMS

DATA
BROWSE AND SELECT DATA
ANALYZE DATA ONLINE
IPUMS ABACUS
DOWNLOAD OR REVISE MY DATA

SUPPLEMENTAL DATA
GEOGRAPHY & GIS
LINKED CENSUS DATA: 1850-1940
SLAVE SAMPLES: 1850-1860
FULL COUNT DATA: 1790-1840
FEDERAL STATISTICAL RDC

DOCUMENTATION
USER GUIDE
SAMPLE DESCRIPTIONS
QUESTIONNAIRES
PUBLISHED CENSUS VOLUMES
COMPLETE COUNT DATA

IPUMS LINKED DATA

IPUMS disseminates full count census enumerations for nine census years from 1850 to 1940. These full count data, covering almost 700 million individual records, are the fruit of collaboration between IPUMS and the world's two largest genealogical organizations—Ancestry.com and FamilySearch—to leverage genealogical data for scientific purposes. Full count data have opened the possibility of automated record linkages across census years to construct millions of individual life histories and trace millions of families over multiple generations.

MULTIGENERATIONAL LONGITUDINAL PANEL

The IPUMS Multigenerational Longitudinal Panel project links individuals' records between censuses. Our first IPUMS MLP data release consists of a set of crosswalks between pairs of adjacent censuses from 1850-1940. We plan to build on this work in future releases. We expect that IPUMS MLP will eventually serve as a general framework that can incorporate records from a wide range of sources.

- Data description
- Linking method
- Downloads

MLP vs. CLP

- As expected, MLP has far lower Type I error rate (false positives) than CLP
- MLP also has lower Type II error rates than CLP (higher linkage rate)
- More surprisingly, MLP and CLP have similar overall selection bias (representativeness)
- MLP will soon be much easier to use than CLP as it is integrated into the IPUMS data access system

Links and legibility: Making sense of historical U.S. Census automated linking methods*

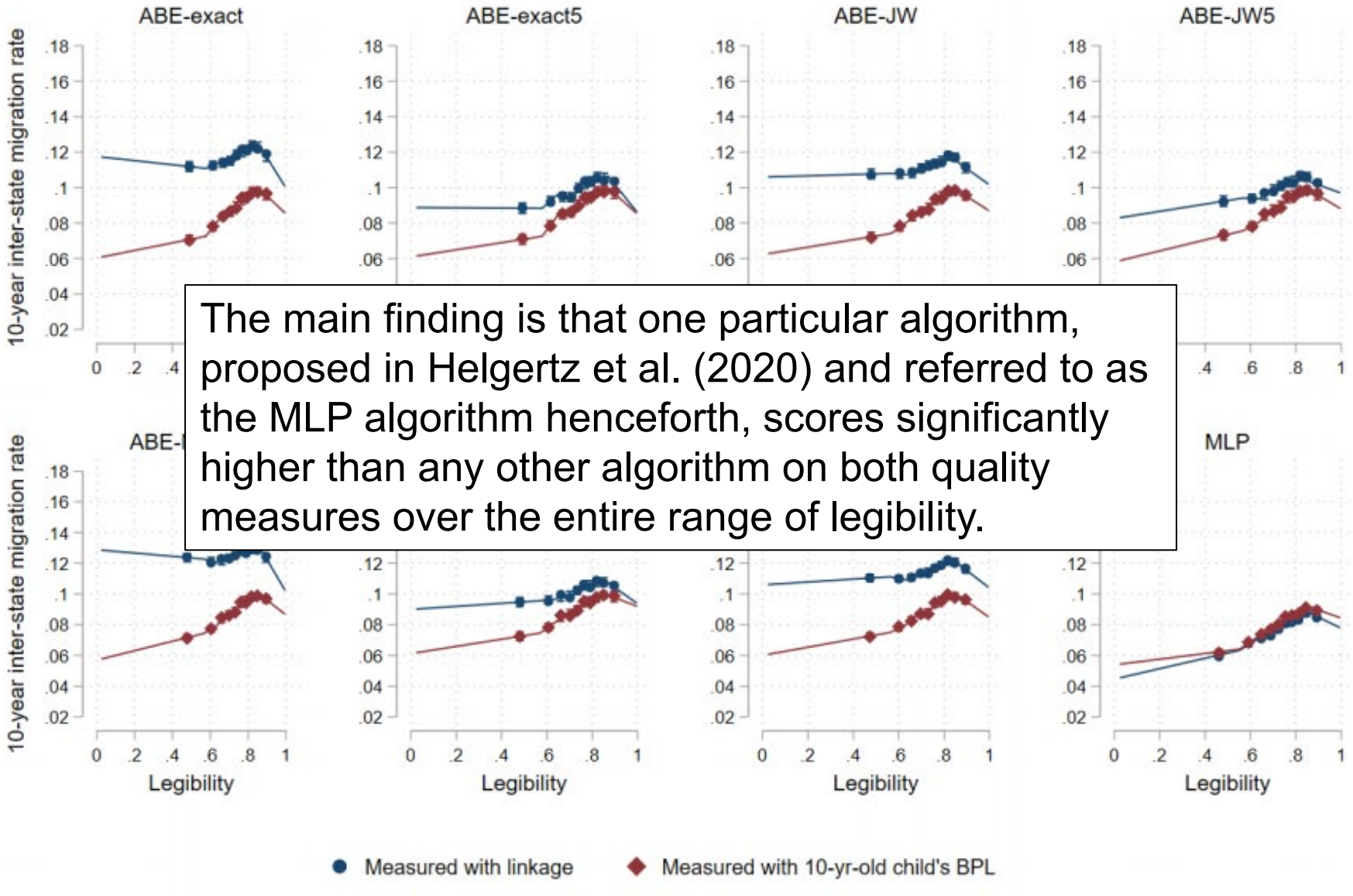
Arkadev Ghosh, Sam Il Myoung Hwang, Munir Squires

University of British Columbia

July 1, 2021

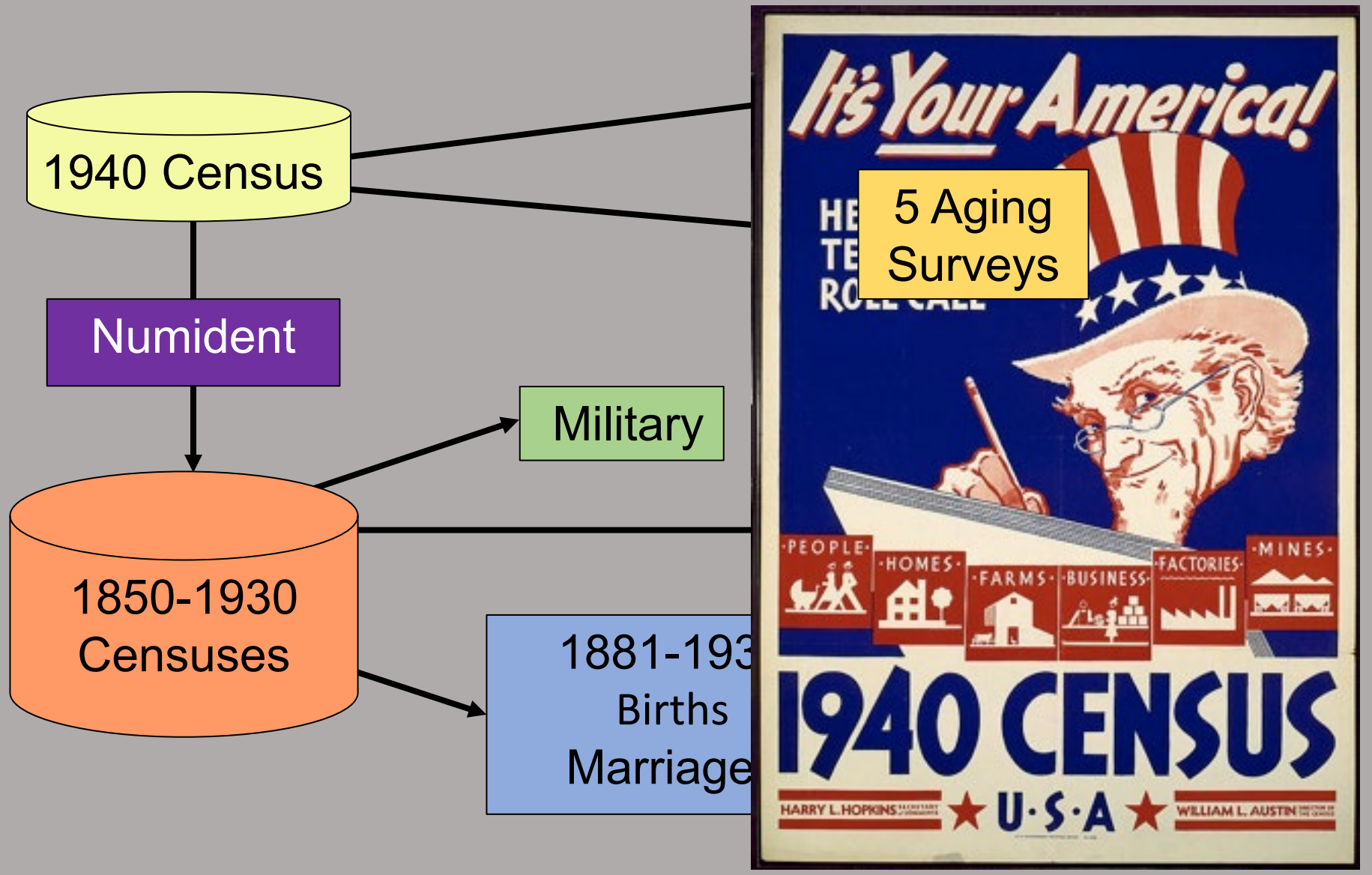
Abstract

This paper explores the effect of handwriting legibility on the performance of algorithms that link individuals across census rounds. We propose a measure of legibility which we implement at scale for the 1940 US Census, and find strikingly wide variation in enumerator-level legibility. Using boundary discontinuities in enumeration districts, we estimate the causal effect of low legibility on the performance of a set of popular automated linking algorithms. We show that one algorithm out-performs the rest across the spectrum of high to low legibility, and find that it provides a better measure of 10-year interstate migration.



The main finding is that one particular algorithm, proposed in Helgertz et al. (2020) and referred to as the MLP algorithm henceforth, scores significantly higher than any other algorithm on both quality measures over the entire range of legibility.

Figure 7: Comparison of 10-year inter-stage migration rates: the BPL10 vs. the linkage migration rates



MLP Linking Strategy

IPUMS-MLP Goals

- Use all available information to minimize errors and maximize linkage rates
- Weights to adjust for selection bias
- Links of individuals and families across nine censuses and other sources
- Large
- Easy-to-use
- Maintainable
- Expandable
- Interoperable with CLIP

NUMIDENT (Social Security Claims Database)

- Includes persons who had Social Security and had died or reached the age of 110 (not clear by when)
- Includes: name, up to three previous names used with date of use, date of birth, date of death, place of birth, sex, race, citizenship, mother’s maiden name, father’s name.
- Released by NARA in October 2018

HLink

- Hlink provides an end-to-end linking solution, replacing our earlier process using multiple programs (FEBRL, LIBSVM, C, Stats Packages) and data formats (ASCII, MySQL, binary files).
- Written in Python and Spark SQL, some Scala
- Leverages Apache Spark
- Enables parallel processing throughout
- Uses Parquet column-store data structure
- Two orders of magnitude faster than the old system

Thank You.