FILE LINEAGE AT CENSUS AND THE 2018 IRS SAFEGUARD A-1 FINDING

U.S. Census Bureau

Carla Medalia – Assistance Division Chief for Business Development

Keith Finlay – Research Economist

CIVIC DIGITAL FELLOWSHIP

United States™ Census Bureau

NOAM KANTOR

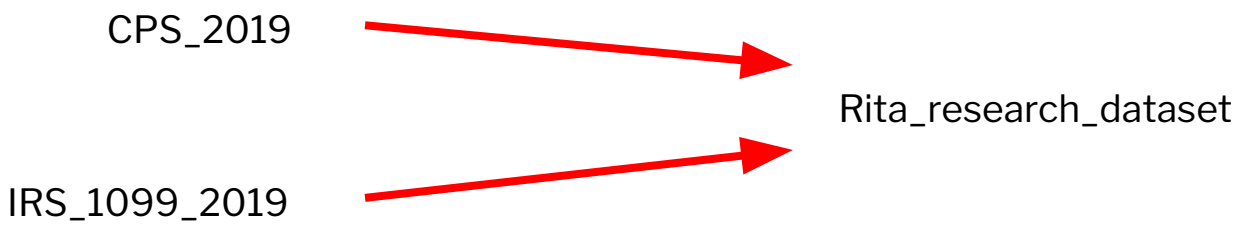
Oxford University

DPhil, Mathematics

MOTIVATION

The Census Bureau must do a better job of tracking all products of Federal Tax Information. – Paraphrased IRS Safeguard Review

- **3 Phases:** Turn on file system access logs, create dashboard that links logs to permissions, and…
- **Problem:** Predict the origin/parentage of commingled datasets at Census



CIVIC DIGITAL FELLOWSHIP

APPROACH #1: VALUE MATCHING

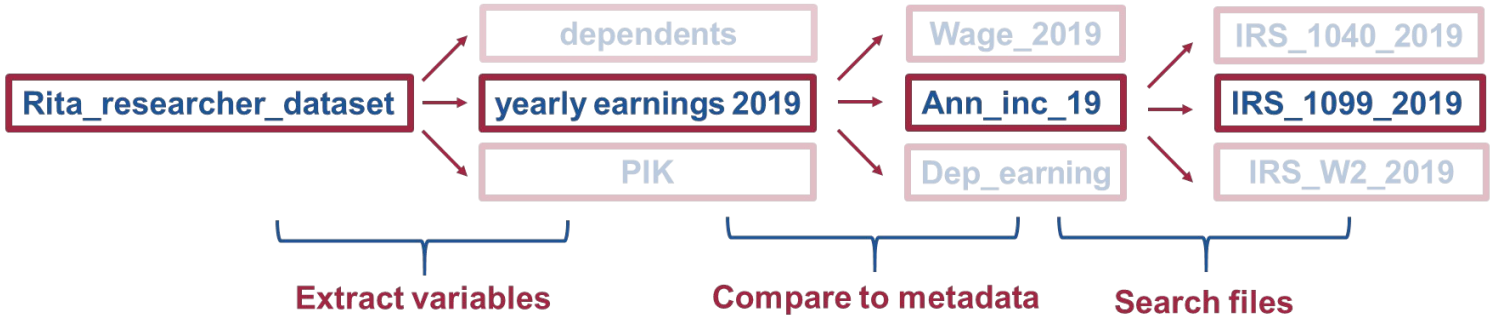
- **Naive approach:** Find “most identifying values” in a dataset D using Census metadata project
- For example, D might be the only dataset with the number 132,458 in it
- If another dataset has that number, we know they are probably related

CIVIC DIGITAL FELLOWSHIP

APPROACH #2: VARIABLE SIMILARITY MATCHING

- Often infeasible to search datasets for data matches

- **Instead:** use Metadata Project to predict when a user-created variable comes from data warehouse (FB FastText)

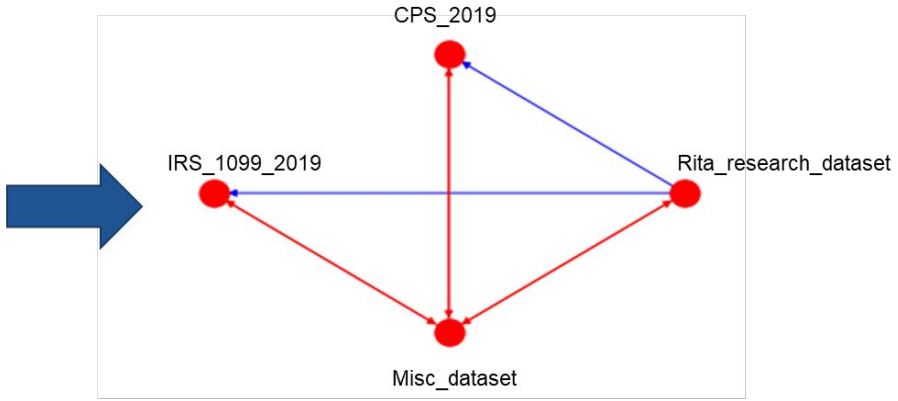


CIVIC DIGITAL FELLOWSHIP

APPROACH #3: LOG ANALYSIS

- In Phase 1: richer file system access log on Census servers
- The impact for lineage: two files can only be commingled if they are opened nearby in the log
- Result is a commingling network:

	A	B	C	D
1	Time	jbid	name	key
2	4/28/2017 13:51	rita007	database/IRS_1099_2019.csv	access
3	4/29/2017 13:51	rita007	database/CPS_2019.csv	access
4	4/29/2017 13:52	rita007	database/rita_research_dataset.csv	access



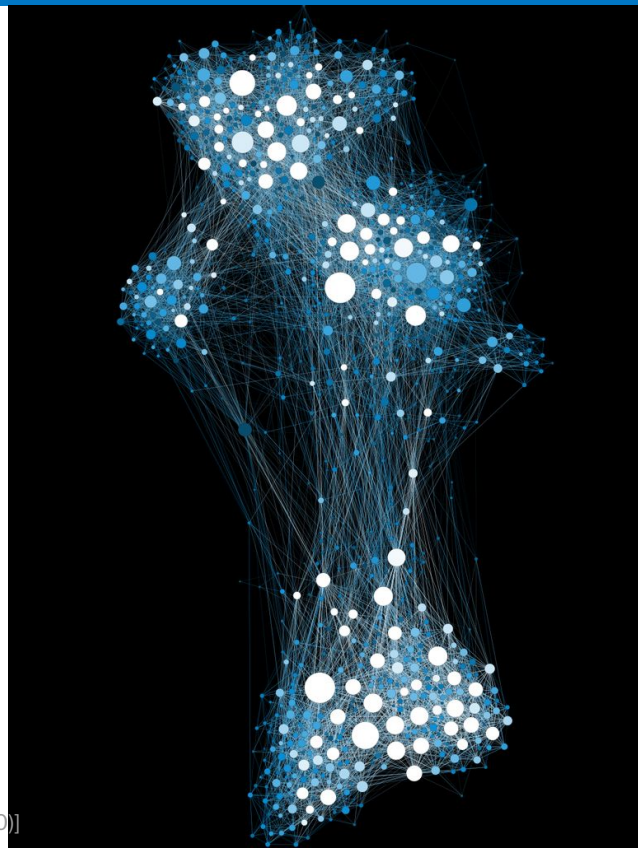
CIVIC DIGITAL FELLOWSHIP

ENTERPRISE-WIDE IMPLEMENTATION

- Variable matching: Seems to be successful on identifying provenance of LEHD datasets, but slow

- Log analysis and value matching: Seem feasible to implement on a large scale and are fast and computationally cheap

- Potential for analytics regarding file usage at Census: Which users use which kinds of files? Data recommendation system? Data prioritization?



CIVIC DIGITAL FELLOWSHIP