Arlington DHS Probabilistic Record Linkage Phase II

Fangfang Lee (NYU), Keren Chen (VT) with Ian Crandell (SDAL) & Aaron Schroeder (SDAL)
Anita Friedman, Martha Coelho, Michael-Dharma Irwin, Legesse Alemu, Jennifer Shell & Meheret Asfaw (Arlington DHS)

Introduction

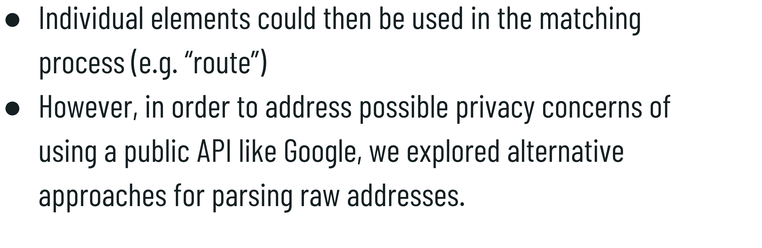
- The Arlington Department of Human Services (DHS) has a no wrong door philosophy meaning customers can sign-up for any DHS service from any DHS department, therefore, data is collected from multiple locations into different systems.
- These different systems include their own; those provided by the **Commonwealth of Virginia**; and, those provided by the **Federal Government.** These systems do not communicate directly with each other.
- In Phase I, a method to probabilistically deduplicate and link individuals across these systems, using demographic information (e.g. SSN, address, name, gender), was created.
- In Phase II, the DSPG team worked on providing three requested enhancements: **enhancing the performance & accuracy of the deduplication processes**, **addressing possible geocoding data privacy issues**, and **updating the linkage method used between deduplicated data sets**.

Project Overview

- Data Source Overview

	Anasazi	ETO	WVS	HCV (New)
Columns	44	37	33	37
Rows	56533	25775	64927	4347

- Project Workflow



Linkage Methodology

- Deterministic Linkage
  - Two entries are compared. If they do not match exactly, they are not a match and vice versa.
- Probabilistic Linkage
  - The probabilistic linkage method will compute likelihood for whether two entries (pairwise comparison) are the same even if they do not match exactly

$$\gamma\left[\alpha(a), \beta(b)\right] = \left\{\gamma^1\left[\alpha(a), \beta(b)\right], ..., \gamma^K\left[\alpha(a), \beta(b)\right]\right\}$$

$$m(\gamma) = P\left\{\gamma\left[\alpha(a), \beta(b)\right] | (a, b) \in M\right\}$$

  - Each columns is given two weights:
    - M probability--determines the reliability of a record
    - U Probability--determines the uniqueness of a record
  - Composite weight scores:
    - Weights are aggregated using probability of linkage formula

Problem I : Performance & Inconsistent Links

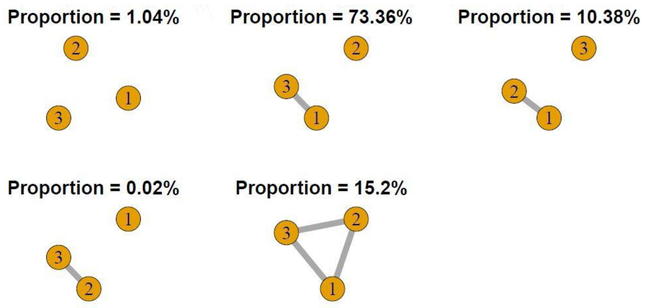
- Performance Issue: Unnecessarily Comparing Fields with Partial Data
  - Every demographic field is used in the probabilistic match
  - When fields have only partial data, they can generate unbalanced probabilistic weights that reduce accuracy while adding extra processing time
- Inconsistent Links: Weight based linkage can create inconsistent links.
  Record 1 **is linked** with Record 2
  Record 2 **is linked** with Record 3
  Record 1 **is not linked** with Record 3

Results I: Added Deterministic Deduplication

- Enhance Performance by adding a Deterministic Linkage Step to remove partial data fields from probabilistic matching
  - We introduced a preprocessing step, linking deterministically on SSN, which will reduce the processing time when data file is large
  - The incomplete field of SSN is removed from probabilistic linkage, resulting in better accuracy of probabilistic linkage
  - This reduced number of records modestly, as well as helping the weight problem

System	ETO
Before Det-dedup	25775
After Det-dedup	25339

- For inconsistent linkage, we introduced a more robust probabilistic method that estimates linkage *configurations*, rather than just pairs.
  - Estimation is performed using a Metropolis algorithm, which estimates probabilities for all possible configurations



Above is an example of the algorithm output. We reconciled 634 linkage groups, reaching near unanimous consensus on 98% of them. Fully integrating this algorithm is future work.

Problem II: Address Parsing with Public API

- Previously, we relied on Google Geocode API for parsing raw addresses
  - Example of a manually entered address: **900 north 8th Str, Philadelphia**
  - Example of geocoded address by Google:

street_number	route	locality	postal_code
900	N 8TH ST	Philadelphia	19123

- Individual elements could then be used in the matching process (e.g. "route")
- However, in order to address possible privacy concerns of using a public API like Google, we explored alternative approaches for parsing raw addresses.

Result II : New Address Parsing Engine

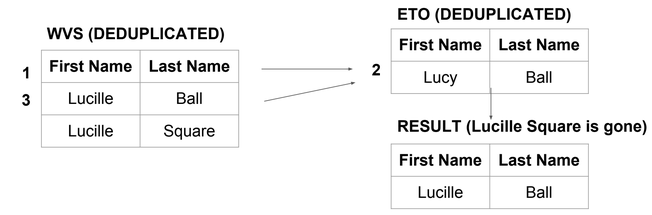
- We used an open source address parsing library written in Python called the Parserator (instead of Google Geocoding) (https://parserator.datamade.us/usaddress)
  - Example of raw full address: 90 N 9th street
  - Example of parsed address:

Address part	Tag
90	AddressNumber
N	StreetNamePreDirectional
9th	StreetName
street	StreetNamePostType
Arlington	PlaceName
VA	StateName
22203	ZipCode

- Comparison to Original Approach:
  - Parserator returned 16508 out of 20280 non missing addresses correctly (81.3%).

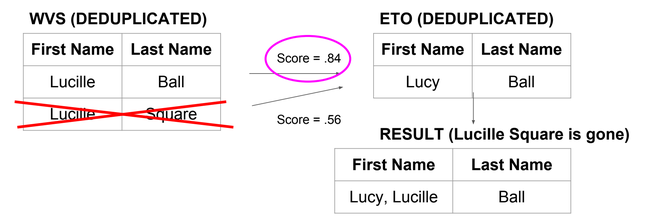
Problem III : Disappearing Unique Individuals

- Occasionally, unique individuals identified through the deduplication process for a single data source could disappear after combing that deduplicated data set with the deduplication data set from another source.
- As one of the examples involved an individual with a first name of Lucy and/or Lucille, this became known as the **Lucy Problem**



Result III: Keep Source Individual Attributes

- A new algorithm was designed for linking between two already deduplicated data sets
- As each deduplicated data set has only unique individuals, it is consistent to assume that there can be, at most, only one match between two deduplicated data sets
- The two records with the highest probability score are taken as the unique match between those two systems
- If any attributes are different, all variations of that attribute are returned



References/Acknowledgements

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.

bi.vt.edu/sdal
Stay Connected: facebook.com/biocomplexity
twitter.com/biocomplexityvt
youtube.com/c/biocomplexity
google.com/+biocomplexity