SDAL SOCIAL & DECISION ANALYTICS LABORATORY

DATA SCIENCE FOR THE PUBLIC GOOD

HOW MANY CLIENTS DO WE SERVE?
ARLINGTON DEPARTMENT OF HUMAN SERVICES (DHS)

Zarni Htet (NYU) and Sayali Phadke (PSU) with Aaron Schroeder (SDAL),
Anita Friedman, Martha Coelho, Michael-Dharma Irwin, Biljana Culibrk, and Meheret Asfaw (Arlington DHS)

Motivation

Name (Sys2)	SSN	DOB	Gender	Zipcode
Lata	NA	10/19/1990	M	22203
Htet Zarni	NA	05/27/1991	M	22202
Claire	789013445	--/07/1975	F	22203

If records do not have unique IDs and exact matching fails, how do we link across multiple systems to find records for the same person?

Name (Sys1)	SSN	DOB	Gender
Sayali Phadke	123-45-6789	10/19/1990	F
Zarni Htet	456-78-8901	05/25/1991	F
Aaron Schroedner	789-01-2345	03/07/1975	M
Zarni	456-78-8901	05/25/1991	M

Is this the same "Zarni"? If so, how would we determine it?

- In a single system, clerical errors and misspellings result in duplicates
- Across systems, field requirements and conventions are different; shared IDs do not exist

Problem Statement

- Each day, multiple clients utilize services offered by DHS
- A client may use more than one service on a given day, creating multiple records
- Arlington DHS client data are:
  - Stored across nine different systems
  - Operated by various service providers, from in-house to State and Federal contractors
- An individual appearing twice in a single system, or in multiple systems, may or may not have the same ID and demographics, due to clerical error
- To answer important questions below, we must identify unique clients:
  - On a given day, how many citizens does the DHS serve?
  - What are the demographics of the clients they are serving?
  - Could DHS be missing eligible individuals they can serve?
- Three systems for testing the method:
  - Web Vision : Behavioral Health for Children and Aging Population
  - ETO : Economic Independence Data
  - Anasazi : Community Service

Methodology

- Probabilistic Linkage
  - A method to determine whether two items are the same, even if they do not match exactly due to different specification, or spelling error etc.
- Each column is given two weights:
  - M probability (quality/reliability)
    - Determines how well a record has been documented
  - U probability (commonness)
    - Determines the commonness of a column
- Composite Weight Scores
  - Agreement on a given column calculated using Jaro Winkler distance between strings
  - Weights are aggregated using probability of linkage formula below (Fellegi and Sunter, 1969)

j = record pair in question
k = identifier (linking variable) in question
n = number of identifiers per record
m_k = estimated identifier agreement weight among true links
n_k = estimated identifier agreement weight among false links
γ_k^j = observed agreement or disagreement (0 or 1) of identifier k in record pair j

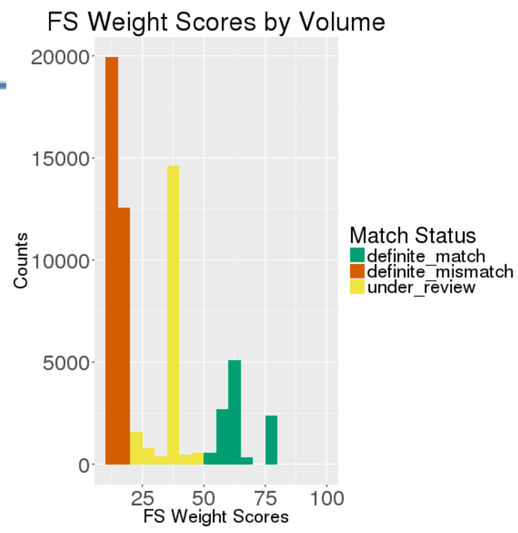
$$\text{probability of linkage}^1 = \sum_{k=1}^{n} \log\left(\frac{m_k}{u_k}\right)^{\gamma_k^j} \log\left(\frac{1-m_k}{1-u_k}\right)^{1-\gamma_k^j}$$

Results

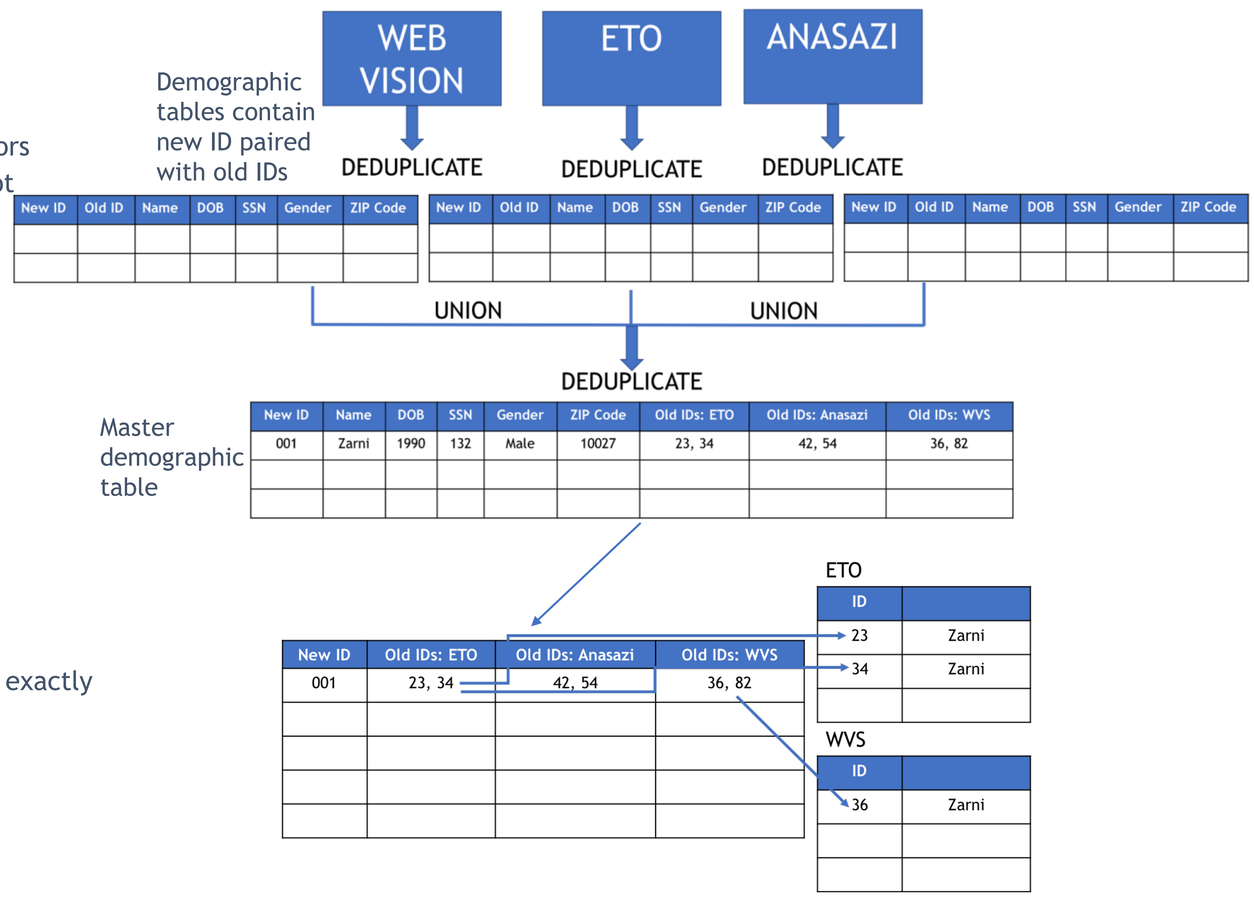
Results from an initial run on Web Vision system with our current weights and thresholds:

Out of ~64,000 administrative records:

- 25,745 have repeated system IDs
  - reduced to 18,222, removing 7,523 duplicated rows, and merged demographic variables
- 105 instances of different system IDs pairs
  - reduced to 70, removing 35 duplicated system IDs
- Clerical review on those below threshold score



Schematic for Merging the Three Systems



DEDUPLICATION PROCEDURE
- De-duplicate and merge demographic variables
- De-duplicate records with mismatched system IDs
- Combine the above 2 data outputs with the remaining unmatched set

Next Steps

- Running deduplication process on Anasazi and ETO systems
- Finalizing weights that work for all three systems
- Calculating weights for additional social service systems
- Testing automated weight calculation
  - Epiweights
  - EM algorithm
- Testing automated thresholding
  - Epiweights
- Comparing automated results with manual inspection

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