SUPERSEDING NOISE THROUGH HASHING AND DATA LINKAGES

National Institute on Aging: Division of Behavioral and Social Research

Advised by Dr. Partha Bhattacharyya

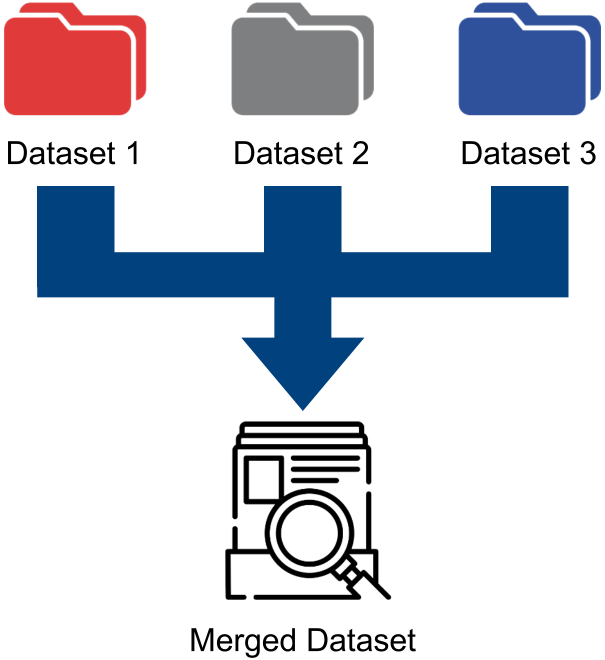
CIVIC DIGITAL FELLOWSHIP

NIH

VISHAL DUBEY
Duke University
CS + Statistics

CURRENT PROBLEM

- Datasets are linked across across different entities and sources [1]
- Errors in identifying the same subjects occur
- For example, Paul C. Smith and Paul D. Smith could be linked due to errors in name, among others such as in DOB (Date of Birth)
  - Lack of fully identifiable information
  - inherent noise and randomness



SOLUTION

Created hash value to mask the identifiable data in reproducible way

Created a synthetic and realistic dataset to conduct linkage comparisons

1. Merged CRISP Legacy Data of NIH awards from the online available datasets for the past 39 years
   - Approximately 2.1 million data points and 300,000 unique persons
2. Imputed values
   - MOB, YOB, DOB, Gender [3], Location-specific [4]
3. Adding noise
   - Key-stroke, gender, zip-code based, MOB/DOB/YOB-errors

TOKEN ANALYSIS STEPS

Matched hashed tokens from a Private Contractor and an internal SHA-512 Hashing Algorithm with known unique identifiers [5]

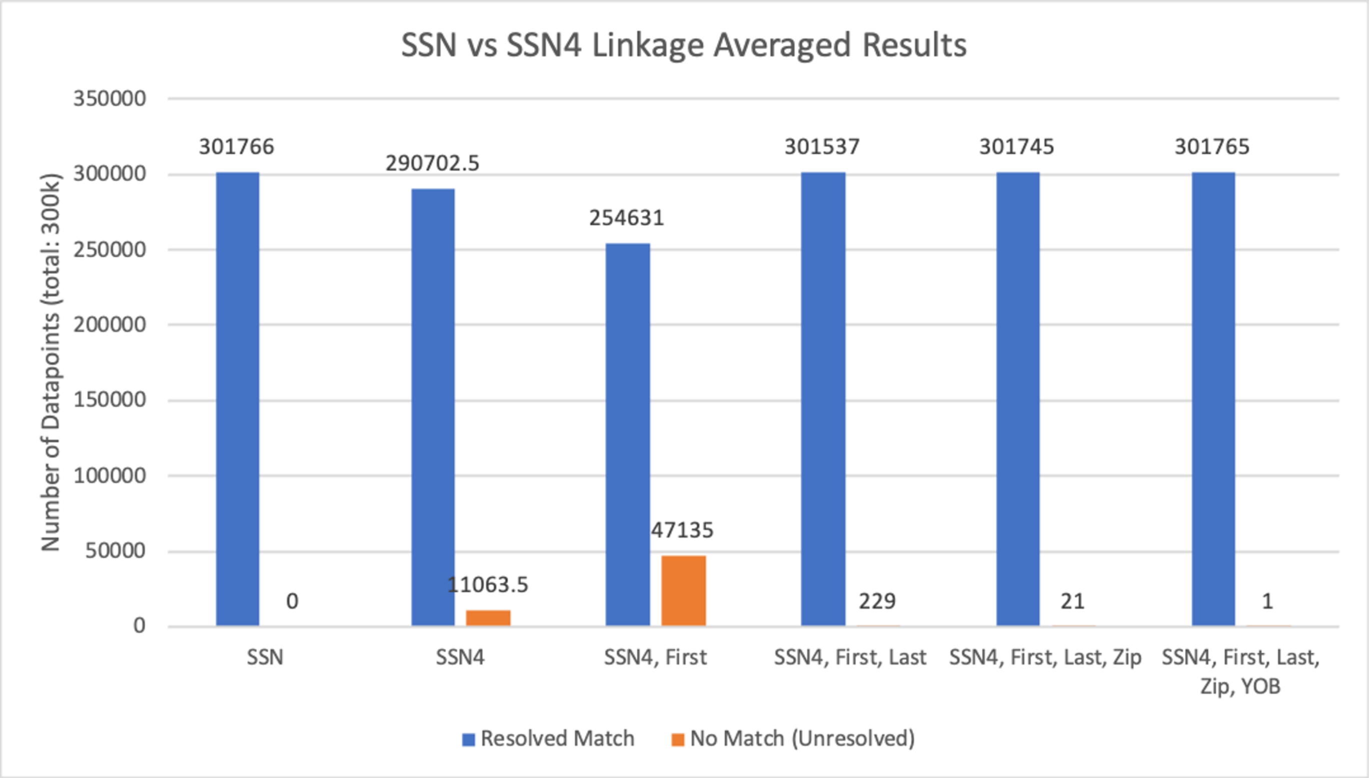
1. Grouped results by the hashed token groups and unique identifiers (SSN)
2. Matched within dataset containing noise encodings (i.e. where exactly noise was imputed)
3. Counted misses where the SSNs did not result in a match
4. Repeated Steps 1-3 for results for many permutations of token groups for both SSN and SSN4

COMPARISON OF RESULTS

Token	Token Grouping	Private Contractor		SHA512 Hash	
		Misses	% Incorrect	Misses	% Incorrect
1	SSN	0	0	0	0
2	SSN4	11075	3.67	11052	3.66
3	SSN + MOB + DOB + YOB	0	0	0	0
4	SSN4 + MOB + DOB + YOB	6738	2.23	6738	2.23
5	SSN4 + Gender + MOB + DOB + YOB	4739	1.57	4739	1.57
6	SSN4 + First Name + Last Name	229	0.08	229	0.08
7	Last Name + First Name + Gender	72184	23.92	72260	23.94
8	Last Name + First Name + Gender + DOB	7054	2.34	7079	2.35
9	Last Name + First Name + MOB + DOB + YOB	46	0.02	46	0.02
10	Last Name + First Name + Gender + COB	18584	6.16	18828	6.24
11	Last Name + Gender + MOB + DOB + YOB + COB	18	0.01	18	0.01

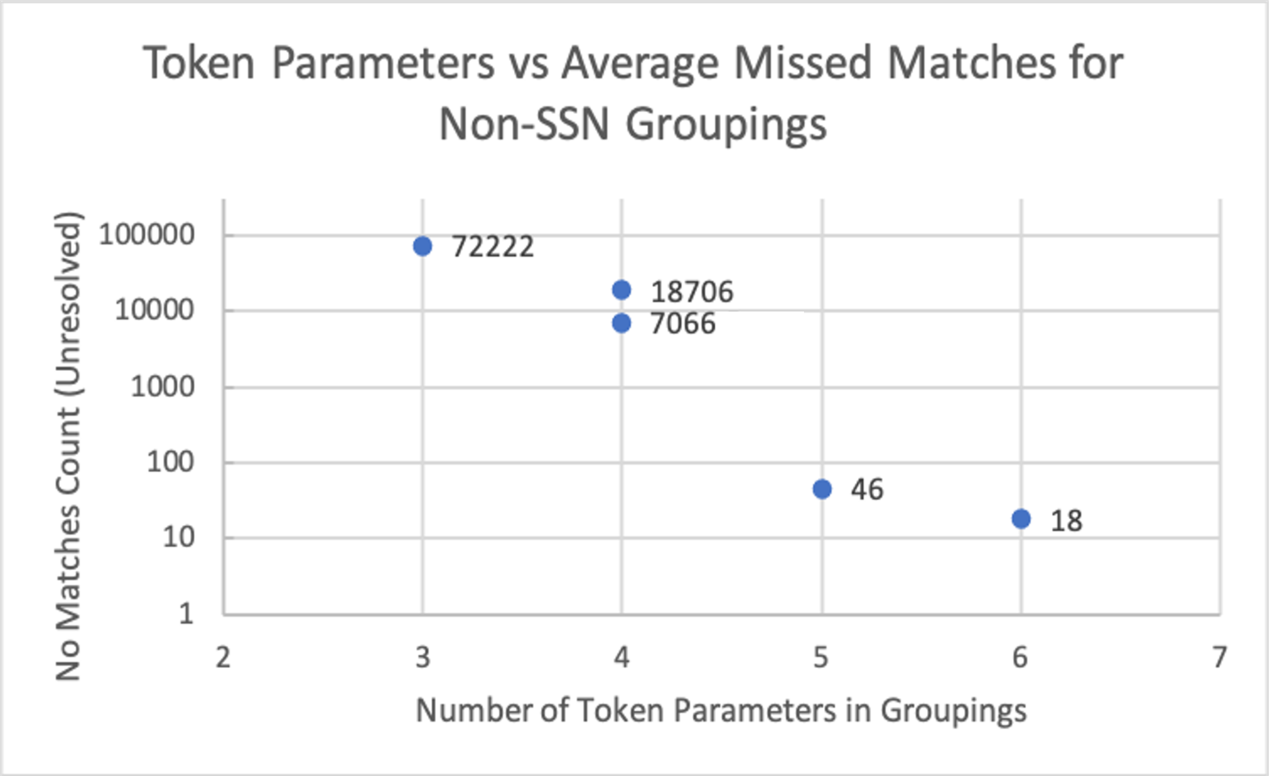
Note: MOB denotes Month of Birth, SSN4 denotes last 4 digits of SSN, and COB denotes City

VISUALIZING SSN VS SSN4 HASHING



CIVIC DIGITAL FELLOWSHIP

INCREASING PARAMETERS YIELDS BETTER MATCHES



CIVIC DIGITAL FELLOWSHIP

DELIVERABLES AND FUTURE WORK

- Conclusion
  - Based on the results from the SHA512 Hash and the Private Contractor, the unresolved miss rates are roughly similar
  - The SSN4 was a worse indicator of linking on it’s own, but combined with other columns (First Name, Last Name, YOB) it is still formidable
- Next Steps
  - Obtain final hashing results from NIH GUID ID’s and Datavant (January 2021)
  - Provide final recommendation and comparison on most appropriate hashing approach

SPECIAL THANKS AND CONTACT

Special Thanks to:

- NIA: Division of Behavioral and Social Research
  - Dana Plude
  - Partha Bhattacharyya
  - Jonathan King
- National Institutes of Health
  - Jessica Mazerik
- Coding it Forward
  - Rachel Dodell
  - Chris Kuang
  - Ariana Soto

Contact Info:

Vishal Dubey
Duke University
Computer Science and Statistics
E: dubeyv2@nih.gov | vishal.dubey@duke.edu
T: (407) 687 - 8998

CITATIONS

[1] An Introduction to Probabilistic Record Linkage (http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/problinkage.pdf)

[2] Probabilistic record linkage (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5005943/)

[3] Gender predictor model (https://pypi.org/project/gender-guesser/)

[4] Zipcode and city database (https://pypi.org/project/uszipcode/)

[5] SHA-512 Cryptographic Hash Algorithm (https://www.movable-type.co.uk/scripts/sha512.html)

APPENDIX

SOLUTION

A hash is a value that masks the original data in an identifiable way so that it may be compressed.

- Generated synthetic dataset with imputed noise
  - Ensures that noise is controlled and known for each variable/column in the dataset
- Groupings based on variables where noise was imputed
  - Analyze differences in hashed values for insight into unresolved versus resolved matches
- An analysis of tokens hashed in different pairings to observe how the noise affected the error rates for linkages within a dataset

TOKEN GROUPS

Token	Private Contractor and SHA512 Hashing Token Elements
1	SSN
2	SSN4
3	SSN + MOB + DOB + YOB
4	SSN4 + MOB + DOB + YOB
5	SSN4 + Gender + MOB + DOB + YOB
6	SSN4 + First Name + Last Name
7	Last Name + First Name + Gender
8	Last Name + First Name + Gender + DOB
9	Last Name + First Name + MOB + DOB + YOB
10	Last Name + First Name + Gender + COB
11	Last Name + Gender + MOB + DOB + YOB + COB