Private Set Intersection on Noisy Identifiers.

René Peralta Angela Robinson

2026-10-27

Presented at WPEC 2026

THE PROBLEM

- Two entities each hold a list of medical records.
- They want to find out which identifiers occur in both lists.
- No SSN or National Medical Record Identifier.
- Names can be noisy : e.g. John and Jon can be the same individual.
- The two entities are not allowed to share identifiers that are not common to both lists.

Noisy Identifiers

- There are many reasons why the same person may appear under a different name in different lists.
  - Robert vs Rob or Bobbie
  - Liz vs Elizabeth
  - Juan Ramirez Santos vs Juan Ramirez
  - Robert vs Robrt

Noisy Identifiers

- DOB may be different, as in 2/10/1970 vs 10/2/1970.
- Middle initial may be missing, as in John D Carter vs John Carter.

Noisy Identifiers

- DOB may be different, as in 2/10/1970 vs 10/2/1970.
- Middle initial may be missing, as in John D Carter vs John Carter.
- For this preliminary work i) two name strings match if they are within a low edit distance from each other; ii) we encode MM/DD/YYYY as **min{MM,DD}/max{MM,DD}/YYYY.**
- For edit distance we allow deletions, insertion, substitutions, transpositions.

A Simple Heuristic

- Want Robert, Roberto, Robrt to map to the same name.
- Process a list of names as follows:
  - Start with an empty list of unique identifiers (UIDs);
  - For each new name S, if S is within edit **distance 2** of a UID, assign that UID to S and go to the next name;
  - Otherwise make S a new UID.

A Simple Heuristic

- Want Robert, Roberto, Robrt to map to the same name.
- Process a list of names as follows:
  - Start with an empty list of unique identifiers (UIDs);
  - For each new name S, if S is within edit **distance 2** of a UID, assign that UID to S and go to the next name;
  - Otherwise make S a new UID.

So Robert, Roberto, Robrt, Robbt all have UID **Robert**.

Order matters

- But if the order of processing is Roberto, Robbt, Robert, Robrt then
  - Robbt has UID Robbt
  - Roberto, Robert, Robrt have UID Roberto

Order matters

- But if the order of processing is Roberto, Robbt, Robert, Robrt then
  - Robbt has UID Robbt
  - Roberto, Robert, Robrt have UID Roberto

Also note that Robert and Albert may have the same UID.

Distribution of Names Matter

- Consider the three names Alberto, Robert, Roberto.
- In this order, the UIDs are Alberto, Robert, Alberto respectively.

Distribution of Names Matter

- Consider the three names Alberto, Robert, Roberto.
- In this order, the UIDs are Alberto, Robert, Alberto respectively.

This works better when most common names are processed first.

Lists of Names and Surnames

The Census Bureau publishes lists of names and surnames, ordered by frequency of occurrence in the US Census.

- MICHAEL JOHN JAMES DAVID ROBERT
  WILLIAM MARY MARIA DANIEL JOSEPH

Lists of Names and Surnames

The Census Bureau publishes lists of names and surnames, ordered by frequency of occurrence in the US Census.

- SMITH JOHNSON WILLIAMS BROWN JONES
  GARCIA MILLER RODRIGUEZ DAVIS MARTINEZ

Lists of Names and Surnames

The Census Bureau publishes lists of names and surnames, ordered by frequency of occurrence in the US Census.

- SMITH JOHNSON WILLIAMS BROWN JONES
  GARCIA MILLER RODRIGUEZ DAVIS MARTINEZ

156621 surnames and 1000 first names

Search Structure

- Process a list of names as follows:
  - Start with an empty list of unique identifiers (UIDs);
  - For each new name S, if S is within edit **distance 2** of a UID, assign that UID to S and go to the next name;
  - Otherwise make S a new UID.

Search Cost

- We used the Census lists to create UIDs for names and surnames.
- "justin": "justin", "stephen": "steven", "stephanie": "stephanie", "gary": "mary", "scott": "scott", "juan": "john"

Search Cost

- We used the Census lists to create UIDs for names and surnames.
- This is stored in a data structure that allows for constant-time retrieval for known names.
- Noisy names (e.g. Robrt, Welthx) require sequential computation of edit distance to known UIDs.

Pseudopeople

- Pseudopeople is a project that generates data records that mimic the US population.
- We use that resource for generating test data.

Experimental Results

- Tested algorithm with two lists containing 5000 pseudopeople each and with randomly generated noisy matches.
- About 90 percent of matches are retrieved in about 90 seconds on a laptop.
- No false positives, which means we can turn some knobs to increase the detection rate.

Back to Fuzzy PSI

- We reduced Fuzzy PSI to exact PSI for a target population (the US) and random noise.
- Two system models
  - Use a semi-honest "combinator" third party
  - No combinator

Back to Fuzzy PSI

- We reduced Fuzzy PSI to exact PSI for a target population (the US) and random noise.
- Two system models
  - Use a semi-honest "combinator" third party
  - No combinator
    - Quantum vulnerable
    - Quantum secure

Conclusions

- For medical record identifiers, this seems very feasible. Even for large data bases.
- Need to have some knowledge of a distribution of names.
- We haven't thought much about non-random noise in the identifiers.
- Not clear that the US legal framework for PSI of medical records fully accounts for cryptographic techniques.