# A Private Set Intersection Use Case

UNECE Input Privacy Preserving Workstream

A. Dasylva and Z. Zanussi, Nov. 2021

Statistics Canada

Delivering insight through data for a better Canada

# Outline

1. Objectives
2. Methodology
3. Conclusion and next steps

Disclaimer: The content of this presentation represents the authors' opinions and not necessarily those of Statistics Canada. It describes theoretical methods that may not reflect those implemented by the Agency.

# Objectives

- A National Statistical Office (NSO) measures the coverage of a file from a 3rd party, with limited trust in the NSO.
  - To provide a value-added service to the 3rd party.
  - To guide the NSO data acquisition strategy.
- The NSO is trusted to limit data access on a need-to-know basis and to provide the related evidence.

# Methodology (cont'd)

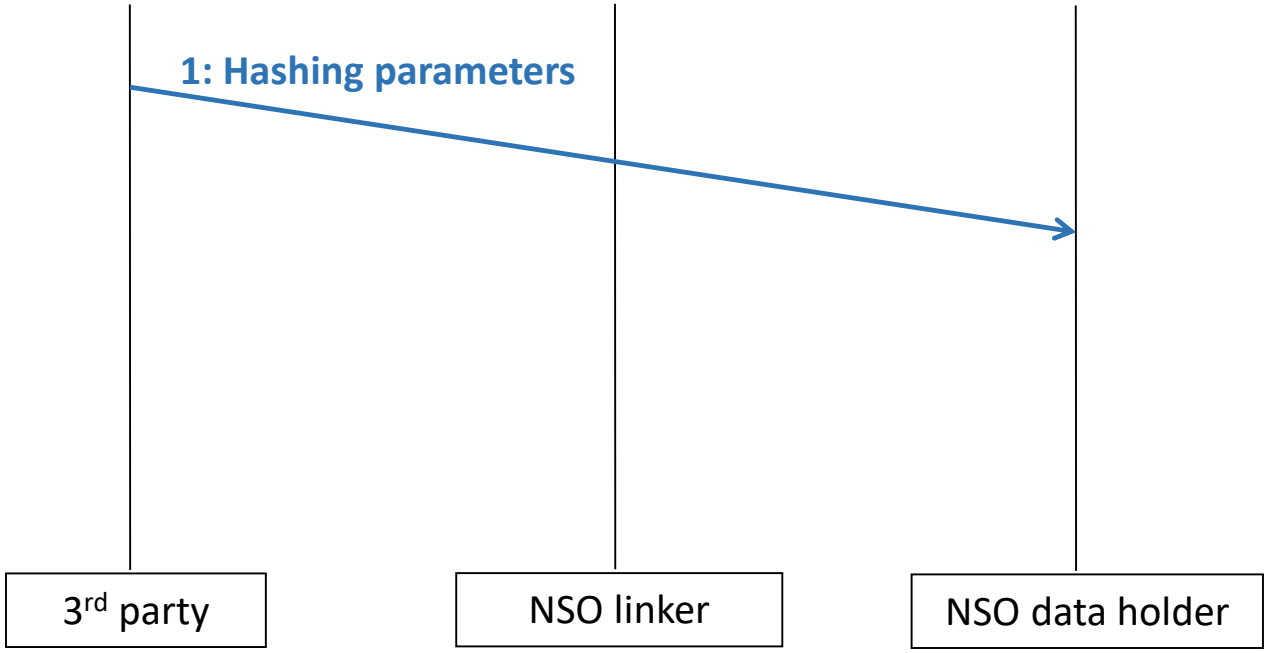
- Estimate the coverage of the 3rd party file by comparing it to a reference file from the NSO, with no data transfer and such that
  - The 3rd party does not discover the identity of the units on the NSO file.
  - No one at the NSO is in a position to discover the identity of the units on the 3rd party file.

# Methodology (cont'd)

- A simple software-based solution combining capture-recapture estimation, hashing and the three-party protocol from Christen (2012, chap. 8) including
  - The 3rd party.
  - Two noncolluding NSO parties including a linker and a data holder, where the NSO reference file is kept with the identifiers.

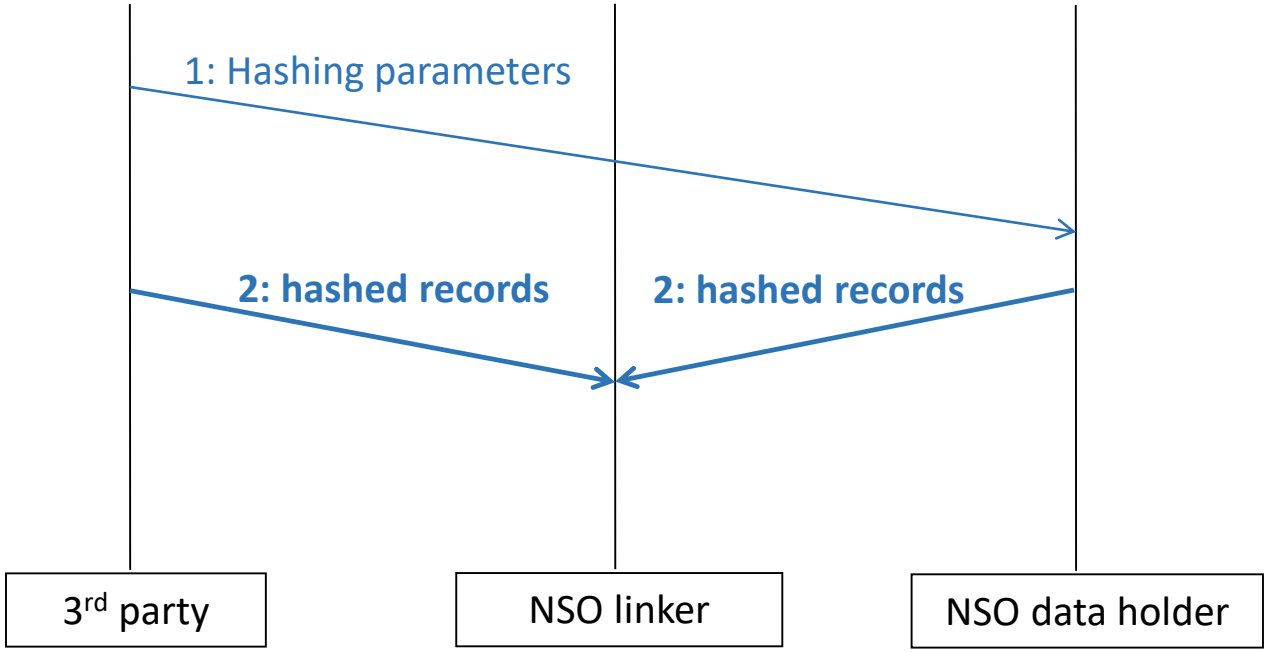
# Methodology (cont'd)

- The steps
  1. The 3rd party sends the hashing parameters to the data holder.



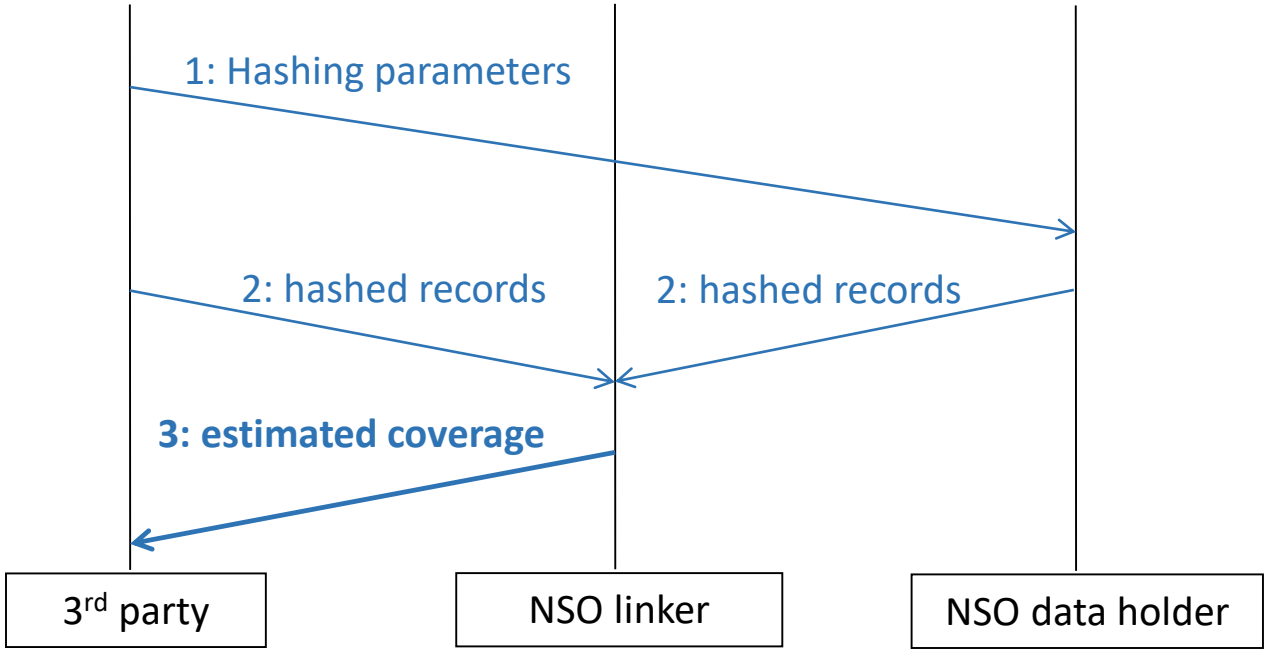
# Methodology (cont'd)

- The steps
  2. The hashed records are sent to the linker, where they are linked and the coverage is estimated.



# Methodology (cont'd)

- The steps
  3. The estimated coverage is communicated to the 3rd party.



# Methodology (cont'd)

- The hashed records are linked based on exact comparisons.
- Obtain bounds on the coverage while accounting for linkage errors by adapting the statistical model described by Blakely and Salmond (2002) and recently extended by Dasylva and Goussanou (2020).
- A proof of concept was implemented in Python with synthetic census data.

# Conclusion and next steps

- This work demonstrates the usefulness of private set intersection even when there is no data transfer.
- A simple software-based solution may be implemented in Python using open source code.
- In future work
  - How to perturb the estimated coverage for a safe output?
  - Conduct an experiment where different parties play the role of the 3rd party and the NSO.

# Thank you!

abel.dasylva@canada.ca

# References

Blakely, T., and Salmond, C. (2002). “Probabilistic record linkage and a method to calculate the positive predicted value“, International Journal of Epidemiology, 31, 1246-1252.

Christen, P. (2012). Data Matching, Berlin:Springer.

Dasylva, A. and Goussanou, A. (2020). “Estimating linkage errors under regularity conditions”, Proceedings of the Survey Methods Section, American Statistical Association.