Private Set Intersection (PSI)

A. Dasylva (StatCan)

UNECE-IPP Workshop

Nov. 24

Contributors

CBS	ISTAT	StatCan
L. Franssen	M. De Cubellis	N. Boushey
D. Ramondt	F. De Fausti	C. Maloney
R. Schreijen		J. Nightingale

Aknowledgement

- UN PET Lab
- Openmined

Outline

- International trade use case
- Methodology
- Experiment
- Lessons learnt
- Potential next steps

Disclaimer: The content of this presentation represents the authors' opinions and not necessarily those of Statistics Canada.

The Use Case

International trade use case

- Two agencies (e.g., Statistics Canada and Statistics Netherlands (CBS)) wish to match their international trade transactions.
  - To resolve bilateral trade asymmetries.
  - To study how exporters take advantage of preferential tariffs in the context of the Canada EU Trade Agreement (CETA).
- For confidentiality reasons, the micro-data cannot be freely shared across the two agencies.

International trade use case (cont'd)

- The transaction microdata
  - Available to both agencies: exporter id, product code, date and value
  - Only available at the importing side (e.g., StatCan): preferential code (nature of the tariff), importer size
  - Only available at the exporting side (e.g., CBS): exporter size

Exporter id	Product code	date	value	Importer size	Preferential code	Exporter size

International trade use case (cont'd)

- Disaggregate the exported value by tariff regime and by exporter size subject to the following confidentiality constraints.
  - StatCan must not be able to infer the size of any exporter.
  - CBS must not be able to infer the size of any importer or the tariff regime of any specific transaction.

  The micro-data is sensitive statistical information (SSI).

Exporter id	Product code	date	value	Importer size	Preferential code	Exporter size

The Methodology

Private set intersection background

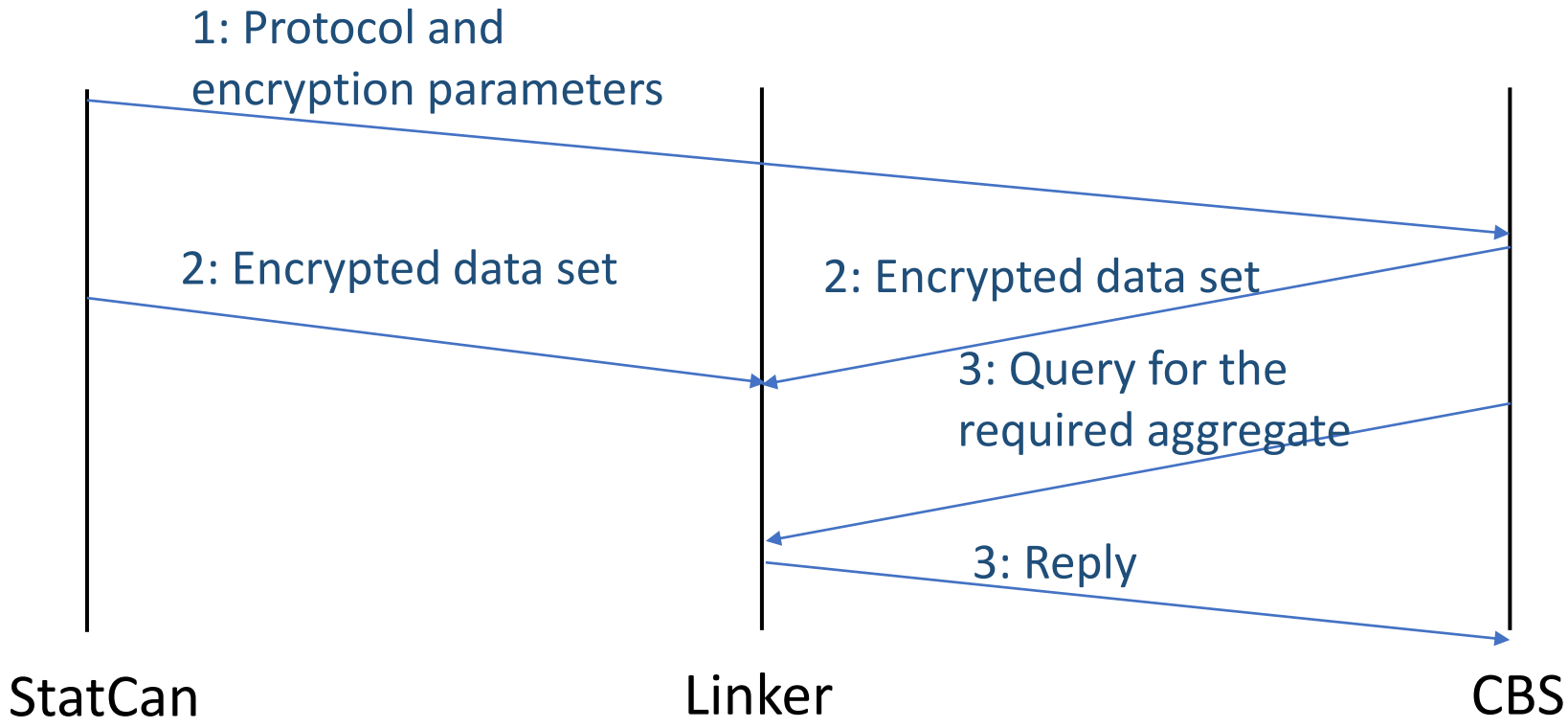
- Two parties wish to privately determine the intersection of their data sets, i.e. to find the subset of units that are in both data sets.
  - To compute some total over this intersection, e.g., the intersection size or the coverage (Dasylva and Zanussi, 2021).
  - To learn some additional information about the units in the intersection.
- A simple idea: if there is a shared unique identifier
  - Encrypt the unique ids.
  - Determine the intersection by comparing the two subsets of encrypted ids.

Private set intersection background (cont'd)

- In general the parties are assumed to be honest but curious.
  - They follow the protocol but seek additional information about the other party.
- Numerous solutions: see Kamara et al. (2014)
  - Three-party solutions: with a trusted third party that does the linkage.
  - Two-party solutions: for dishonest parties. These solutions are more complex than three-party solutions.
  - With data transfer: a three-party solution where the third party sends some micro-data to one of the parties. It requires a perfect linkage.

Implemented PSI protocol

- See De Cristofaro and Tsudik (2010) and Bruno et al (2021) for details.



Implemented PSI protocol (cont'd)

- A unique identifier is to be used with an exact comparison on its hashed/encrypted value.
  - A quasi-identifier (i.e. a nonunique) may be used but this has implications.
- The linker computes an aggregate on the intersection.
- The linker may perturb the outputs (e.g., via differential privacy) to make them safe (not implemented currently).
- Code available on Github.

PSI with a quasi-identifier

- PSI is more challenging when using a quasi-identifier, e.g., the date or value of an international trade transaction.
  - Nonunique
  - Possibly recorded with typos or spelling variations
- Then linkage errors may occur.
  - False negative: not linking records from the same unit
  - False positive: linking records from different units

PSI with a quasi-identifier (cont'd)

- These linkage errors are a potential source of bias.
- A comprehensive solution
  1. Perform approximate comparisons to allow for typos.
  2. Estimate the linkage accuracy: Clerical reviews are impossible. Use an error model that is based on the number of links from a given record, when one data set is contained in the other (Dasylva and Goussanou, 2022).
  3. Account for the linkage accuracy when computing an aggregate, e.g., based on Lahiri and Larsen (2005), Chambers and Kim (2016) or Dasylva (2018).

PSI with a quasi-identifier (cont'd)

- Measures of linkage accuracy
  - Call a record pair matched if the records are from the same unit.
  - A true positive (TP) is a matched pair that is linked.
  - A true negative (TN) is an unmatched pair that is not linked
  - A false negative (FN) is a matched pair that is not linked. It is an error.
  - A false positive (FP) is an unmatched pair that is linked. It is an error.

	Linked	Not linked
Matched	TP	FN
Unmatched	FP	TN

Recall = $TP/(TP + FN)$

False positive rate (FPR) = $FP/(FP + TN)$

PSI with a quasi-identifier (cont'd)

- To account for the linkage errors, estimate the total by

$$\frac{\overbrace{\begin{pmatrix} total\ over\ the \\ linked\ pairs \end{pmatrix}}^{naive\ estimator} - FPR \times \begin{pmatrix} total\ over\ the \\ Cartesian\ product \end{pmatrix}}{recall - FPR}$$

  - It equals the naïve estimator for a perfect linkage, i.e. recall=1.0 and FPR=0.0.
  - Use blocking to estimate the total over the Cartesian product.
  - See Lahiri and Larsen (2005), Chambers and Kim (2016) or Dasylva (2018) for other error adjustment methods when fitting a statistical model.

The Experiment

Overview

- Produce synthetic international trade micro-data.
- Perform the PSI when the linkage variables have no typos and compute an aggregate.
- Perform the PSI when the linkage variables have typos and estimate the aggregate while accounting for the linkage errors.

Synthetic micro-data

- 100K transactions with the following variables

Variable	StatCan	CBS	Used for linkage	Description
exp_id	✓	✓	✓	Exporter id
hs6	✓	✓	✓	Product code
date	✓	✓	✓	Transaction date
value	✓	✓	✓	Transaction value
exp_size	x	✓	x	Exporter size, small or large
imp_size	✓	x	x	Importer size, small or large
pref_kode	✓	x	x	Tariff regime

Analytical need

- To study how small exporters take advantage of the preferential tariffs between EU and Canada based on CETA.
- For small exporters from the Netherlands
    a. Estimate the number of transactions, which benefited from the preferential tariffs. It requires no change to the PSI protocol.
    b. Estimate the total exported value, which benefited from the preferential tariffs. It requires homomorphic additions at the linker.

PSI with no typos

- Test the Python implementation of the PSI protocol.
- The variables have no typos. Thus each linkage variable has the same value on the CBS and StatCan data sets.
- There is no unique identifier but the combination of the exporter id, product code, date and value is essentially unique.
- Link two transactions if they agree perfectly on all these variables.
- Compute the number of transactions of interest.

Results with no typos

- A successfully test on a single machine
  - The implemented solution behaves as expected and can handle the volume of the synthetic data set.
  - The number of transactions of interest is obtained.
- Some challenges in performing a test across the Internet for cybersecurity reasons.

PSI with typos

- The date and value have errors on the exporter data set. However the exporter id and the product code have no errors.
  - With probability 0.8 the date and value are recorded with no errors.
  - Otherwise, the values are perturbed. The date is moved by a number of days, according to a normal distribution with mean 0 and standard deviation 20. The value is deflated at random with a mean of 15% and a standard deviation of 5%.
- Compare two strategies for approximate comparisons in R.

PSI with typos (cont'd)

- Use two strategies to compare and link the records

    1. Exact comparison: a simple solution requiring no change to the PSI protocol. However the record comparisons are coarse.
    2. Bloom filters: a solution for finer record comparisons (Schnell et al., 2009). However it is more complex and requires changes to the PSI protocol.

    Can we get away with the first solution?

PSI with typos (cont'd)

- Bloom filters
  - A Bloom filter is an encoding of a string into a long sequence of bits (e.g., 1,000 bits). Initially set all the bits to 0. Break the string into bigrams and for each bigram flip a number of bit positions to 1 based on a set of hash functions (e.g., 20 such functions). A flipped position is never set back to 0.
  - The Dice similarity of two filters is based on the number of common bit positions. It is a number between 0 (no common position) and 1 (same bit positions).

PSI with typos (cont'd)

- Bloom filters (cont'd)
  - Block based on the exporter id and product code.
  - Encode all the variables into a single Bloom filter (i.e., a record-level Bloom filter) and link two records if the Dice similarity of their filters is equal to or greater than a threshold.
  - It requires a major modification of the PSI protocol.

PSI with typos (cont'd)

- Break the data set into independent chunks (each with 100 transactions) and perform 100 Monte Carlo repetitions. In each repetition
  - Link the data sets.
  - Estimate the linkage errors.
  - Estimate the total of interest.
- Use the error model because the two data sets overlap perfectly.

Results with exact comparisons

- Linkage accuracy: It is accurately estimated with the model.

Measure	Estimator	Mean	Standard error
Recall	Actual	0.8	0.038
	Estimate	0.8	0.038
FPR	Actual	0.0	0.0
	Estimate	0.0	0.0

Results with exact comparisons (cont'd)

- Estimated total value for small exporters and preferential tariffs
  - The adjusted estimator has a small bias.

Estimator	Mean (M)	Standard error (M)
Actual	6.260	1.697
Naïve	5.214	1.585
Adjusted	6.514	1.952

Results with exact comparisons (cont'd)

- Number of transactions with a value equal to or greater than 500K by small exporters with preferential tariffs
  - The adjusted estimator has a small bias.

Estimator	Mean (M)	Standard error (M)
Actual	6.370	2.116
Naïve	5.320	2.000
Adjusted	6.680	2.530

Results with Bloom filters

- Linkage accuracy: It is accurately estimated with the model.

Dice threshold	Measure	Estimator	Mean	Standard error
0.9	Recall	Actual	0.796	0.042
		Model	0.796	0.042
	FPR	Actual	0.0	0.0
		Model	0.0	0.0
0.8	Recall	Actual	0.949	0.020
		Model	0.949	0.020
	FPR	Actual	0.0	0.0
		Model	0.0	0.0

Results with Bloom filters (cont'd)

- Total value by small exporters with preferential tariffs
  - The adjusted estimator has a small bias.
  - The naïve estimator has a smaller bias when the recall is higher.

Dice threshold	Estimator	Mean (M)	Standard error (M)
0.9	Actual	6.231	1.705
	Naïve	5.023	1.665
	Adjusted	6.310	2.066
0.8	Naïve	5.977	1.722
	Adjusted	6.293	1.785

Results with Bloom filters (cont'd)

- Number of transactions with a value equal to or greater than 500K by small exporters with preferential tariffs
  - The adjusted estimator has a small bias.
  - The naïve estimator has a smaller bias when the recall is higher.

Dice threshold	Estimator	Mean (M)	Standard error (M)
0.9	Actual	6.310	2.187
	Naïve	5.190	2.083
	Adjusted	6.481	2.634
0.8	Naïve	5.960	2.183
	Adjusted	6.266	2.303

Conclusion

- The implemented PSI protocol allows the computation of totals on the intersection when
  - the analytical variables are categorical (e.g., whether the transaction value exceeds 500K) and
  - the linkage variables have no typos and can be combined into a key that is essentially unique.

Conclusion (cont'd)

- With typos, a simple solution is to still perform exact comparisons but to adjust the estimated total for the linkage errors.
  - When one data set is contained in the other (i.e., assuming that each transaction is recorded on both sides, possibly with different information), model the linkage errors by the number of links from a given record.
  - When the recall is high, one may use the naïve estimator based on the links without any adjustment.

Conclusion (cont'd)

- Major changes are required to the implemented PSI protocol for
  - sophisticated approximate comparisons, e.g., using Bloom filters or homomorphic encryption,
  - totals over quantitative analytical variables (e.g. the actual transaction value), e.g., using homomorphic encryption,
  - fitting a statistical model, e.g., using homorphic encryption.

Lessons

Lesson 1

- Private set intersection technologies are promising but further work is required.
  - The described methodology may be used to study certain bilateral trade asymmetries, under the assumption that each transaction is recorded by each trading partner and when considering categorical analytical variables.
  - More work is required when the data sets overlap partially, some analytical variables are quantitative or when fitting a statistical model.

Lesson 2

- There are many technological components but no complete solution that integrates the following required features.
  - Flow governance, i.e. policies and procedures to control who can use the Privacy Enhancing Technology (PET) infrastructure.
  - Verification of the inputs, i.e. only accepting legitimate inputs.
  - Verification of the outputs, i.e. only permitting legitimate computations.
  - Private inputs, e.g. through encryption.
  - Safe outputs, e.g. through differential privacy.

Lesson 2 (cont'd)

- In general the governance is lacking.
- This means that it is hard to provide evidence that the environment is safe and the privacy guaranteed, if requested.

Lesson 3

- The main obstacles are not technological.
  - Cybersecurity concerns currently preventing a joint test of the PSI protocol.
  - Legal concerns preventing the transmission of encrypted micro-data from the statistical agency to an external party. From a legal stand point, encryption is not sufficient for de-identification.
- There is a need for a dedicated IT infrastructure to further test PETs.

Lesson 4

- Collaboration and team work are key.
- Build multi-disciplinary teams including subject matter specialists, computer scientists, methodologists and legal experts.
- Pool the resources, e.g., with a solution for multi-party computation as a service.

Potential next steps

- Perform a joint test of the implemented PSI protocol
- Evaluate PSI options with only two parties and no trust.
- Leverage the UN PET Lab for exploring more robust PSI options, e.g. multiparty computation.

Thank You! / Merci!

abel.dasylva@statcan.gc.ca

Disclaimer: The content of this presentation represents the authors' opinions and not necessarily those of Statistics Canada.

Avertissement: Le contenu de cette presentation représente le point de vue de son auteur et pas nécessairement celui de Statistique Canada.

References

Bruno, M., De Cubellis, M., De Fausti, F., Scannapieco, M. and Vaccari, C. (2021). "Privacy set intersection with analytics - an experimental protocol (PSI De Cristofaro)", UNECE-IPP presentation.

Chambers, R. and Kim, G. (2016), "Secondary analysis of linked data" in *Methodological Developments in Data Linkage*, Wiley, pp. 83-108.

Dasylva, A. (2018). "Pairwise estimating equations with linked data", International Methodology Symposium.

References (cont'd)

Dasylva, A. and Goussanou, A. (2022). "On the consistent estimation of linkage errors without training data", Japanese Journal of Statistics and Data Science.

Dasylva, A. and Zanussi, Z. (2021). "A private set intersection use case", presentation at the UNECE-IPP workshop.

De Cristofaro, E. and Tsudik, G. (2010). "Practical private set intersection protocols with linear computational and bandwidth complexity", in Proceedings of Financial Cryptography and Data Security, 2010.

References (cont'd)

Kamara, S., Mohassel, P., Raykova, M. and Sadeghian, S. (2014). "Scaling private set intersection to billion-element sets", in Proceedings of the International Conference on Financial Cryptography and Data Security.

Lahiri, P. and Larsen, M. (2005). "Regression analysis with linked data", Journal of the American Statistical Association, vol. 100, pp. 222-230.

Schnell, R., Bachteler, T., Reiher, J. (2009): "Privacy-preserving record linkage using Bloom filters", BMC Medical Informatics and Decision Making, vol. 9.