A three-party private set intersection protocol for data sets with typographical errors

A. Dasylva

Feb. 2, 2023

Abstract: Private set intersection methods may be used when two parties must privately link their data sets to compute a total over the units that are found in both data sets. These methods typically rely on the presence of a unique identifier and are more difficult to implement when using a quasi-identifier due to the occurrence of linkage errors. To address the issue, a three-party protocol is proposed when there is a blocking key that is recorded without error in each data set and each data set is a Bernoulli sample from the underlying finite population. In this protocol, the linkage accuracy is estimated with a model by each data holding party and the estimated total is adjusted accordingly to remove the bias caused by the linkage errors.

Definitions and assumptions

For simplicity, suppose that the two data sets are independent Bernoulli samples from the same finite population. Call two records matched if they are related to the same unit. Also call a pair matched if its records are from the same unit. Otherwise call it unmatched. When linking records with quasi-identifiers, two types of linkage error may occur, including false negatives and false positives, where a false negative is not linking two matched records and a false positive is linking unmatched records. To facilitate the linkage of large files, blocking keys may be used, which are derived from the available quasi-identifiers. These keys are used to select a small subset of the Cartesian product where one expects to find most if not all matched pairs. Thus a pair is not linked if its records agree on no blocking keys. In general, the blocking keys are non-unique and such that two matched records have a high probability of agreeing on at least one such key. One can also use a linkage key that is also derived from the quasi-identifiers, such that two records are linked if they agree on at least one blocking key and on the linkage key. For simplicity, it is assumed that there is an error-free blocking key that is available on each data set. Thus, all matched pairs agree on this key. Also suppose that there is a linkage key such that two records agree on this key only if they also agree on the blocking key.

The modified protocol

The new protocol is obtained by modifying the first and the third steps of the current protocol as follows. In the first step, the blocking key is used to build the intersection file. Additionally, the linkage accuracy is estimated by each data-holding party, when the linkage is based on the blocking key and when it is based on the linkage key. In the third step, the linker computes two totals, including the total over the pairs that agree on the blocking key, and the total over those that agree on the linkage key. The two totals are returned to the requesting party, who computes the adjusted total according to the estimated linkage accuracy. The following paragraphs provide more details.

First step: To build the intersection file, each data-holding party first exchanges the hashed and encrypted values of the blocking key with its peer (the other data-holding party), where a record is included in the intersection if its value of the blocking key is found in the other data set. For a such a record, the associated unit is deemed to be in the intersection. In the process, the same party estimates the coverage of the other data set and the false positive rate when the linkage is based on having the same blocking key, where the estimates are obtained by adapting the solution from Dasylva and Goussanou (2022) as described in the next section. Next the party exchanges the hashed and encrypted values of the linkage key with its peer, to estimate the linkage accuracy when linking based on having the same linkage key, this time without creating a new intersection file. Again the estimates are obtained by adapting the solution from Dasylva and Goussanou (2022) as described in the next section. At the end of this step, the intersection file contains the record of each unit, which is truly in the intersection because the blocking key is error-free. However, the intersection also contains records of units that are outside the intersection. In the source code, the first step is implemented by modifying the files psiClient.py and psiServer.py.

Second step: As before each data-holding party sends its encrypted intersection file to the linker, including the record of each unit that is deemed to be in the intersection.

Third step: The linker receives a query to compute a total by one party. In response, it computes the total over the links (i.e., the pair that agree on the linkage key), as before. However, in the modified protocol, it also computes the corresponding total over the pairs that agree on the blocking key. Then it replies to the requesting party with both totals. When it receives the two totals, the requesting party computes the adjusted estimate according to the following formula.

$$\frac{\overbrace{\begin{pmatrix} \mathit{total\ over\ the} \\ \mathit{linked\ pairs} \end{pmatrix}}^{\mathit{naive\ estimator}} - \frac{FPR}{(Blocking\ FPR)} \times \begin{pmatrix} \mathit{total\ over\ the} \\ \mathit{blocked\ pairs} \end{pmatrix}}{recall - \frac{FPR}{(Blocking\ FPR)}}.$$

In the source code, the third step is implemented by modifying the files py and py.

<u>Estimating the linkage accuracy</u>

The methodology described by Dasylva and Goussanou (2022) applies when one data set has a complete coverage of the finite population. It requires an adaptation when both data sets have some under-coverage, e.g., when they are Bernoulli samples. In this case, the straightforward application of the solution may be applied to estimate the false positive rate and the product of the coverage by the recall. Estimating the recall is harder because the absence of a link may be due to the unit not being included in the data set or the occurrence of a false negative. However, the coverage may be estimated if the linkage is known to produce no false negatives, i.e., if the recall is known to be equal to 1.0, as is the case when linking based on having the same error-free blocking key value. Then the linkage accuracy may be estimated with a double application of the error estimation procedure by Dasylva and Goussanou (2022), in the first step. When the data-holding party exchanges the hashed and encrypted values of the blocking key, it also applies the model from Dasylva and Goussanou (2022) to estimate the coverage of the other data set and the blocking false positive rate, i.e., the probability that two records agree on the blocking key given that they are unmatched. The input of the error estimation procedure is the distribution of the number of links from a given record, when linking based on having the same value of the blocking key. Then the same party applies the error estimation procedure a second time, when it exchanges the hashed and encrypted values of the linkage key. This yields estimates of the false positive rate and the product of the recall by the coverage, when linking based on having the same value of the linkage key. The estimated recall is obtained by dividing the estimate of the product of the coverage and recall by the estimated coverage in the first application of the error estimation procedure. For the second application of the error estimation procedure, the input is the distribution of the number of links from a given record, when linking based on having the same value of the linkage key.

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.

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