**Re-identification risk-based anonymization component - Proof of Concepts Documentation**

*Hebert Silva, Tania Basso, Regina Moraes - University of Campinas – UNICAMP*

**PRIVaaS re-identification risk-based anonymization component**

Even when data is anonymized - in the ETL process (Anonymization 1) or during the data analytics process (Anonymization 2) - these data can be re-identified, especially through privacy attacks (e.g., background knowledge attack). The PRIVaaS re-identification risk-based component evaluates the re-identification risk of the data resulting from data mining and, based on predefined risk thresholds, increases the anonymity level in order to reduce this risk.

The component implements functionalities from ARX tool [1] and is adequate to be used before exporting the statistical results of big data analytics. For its correct operation, a risk threshold must be defined in the anonymization policy. Then, the component identifies the quasi-identifiers in the sample and calculates the re-identification risk. If this risk is higher than the risk threshold established in the policy, the k-anonymity algorithm is applied in order to increase the anonymity level and, consequently, reduce the risk. The value of k is increased successively, until the risk is equal or lower the threshold (the higher the k, the lower the risk).

As the re-identification risk calculation is based on ARX tool, we considered the *prosecutor model* profile, which is the most threatening profile provided by ARX. We also considered the *highest prosecutor risk*, which represents the highest risk (in percentage) that the data set can present among all records. For example, if a data set has at least one record with 90% of risk to be re-identified, the data set re-identification risk is classified with this mark (90%).

**1. Attributes classification**

The attributes (i.e., the fields in the database) can be classified according to the risk that they can present for the information disclosure of the individuals. This classification is based on the literature and made as follows: (i) *identifiers*, which identify individuals uniquely (e.g., Name, CPF, GR); (ii) *quasi-identifiers*, which when combined with external information expose individuals or increase certainty about their identities (e.g., date of birth, ZIP code, position, blood type); and (iii) *sensitive attributes*, which relate to specific information about the individuals that must be private (e.g., salary, medical examinations). To use the PRIVaaS re-identification risk-based component the fields must be classified as *sensitive, quasi-identifiers, identifiers* or *non-sensitive*.

**2. Input Data**

The Input Data represents the data set that must be anonymized. This data set must be loaded to the component as CSV file. This CSV file must use semicolon to separate the columns and numerical data need to use point (.) to separate double data type. Figure 1 shows an example of the format of the input data. This example refers to a small sample of a result of data mining, with some aggregated results (MIN, MAX, COUNT, SUM), a non-sensitive field (DATETIME) and a quasi-identifier attribute (ZIPCODE).



Figure1. Sample of the input data.

To calculate the re-identification risk, it is necessary that the data set has at least one quasi-identifier. Otherwise, the component will show a message informing that it is not possible to calculate the risk and no anonymization is applied, finishing the process.

**3. Anonymization Policy**

The purpose of the anonymization policy file is to define the parameters necessary to the operation of the component, letting the user free from system interactions during the anonymization process.

The policy must define: (i) the fields (i.e., the names that identify the columns of the table); (ii) the type of attribute (identifier, quasi-identifier, sensitive or non-sensitive), which implies the technique that will be applied; (iii) the re-identification risk threshold. Figure 2 shows an example of anonymization policy. We explain its structure below.



Figure 2. Anonymization policy.

The first column in Figure 2 (before ";") represents the fields of the database. In the first line, using the abbreviation "SR", the ZIP Code will be categorized as *quasi-identifier* and suppressed from right to left. In the remaining lines (except the last one), the number "1" categorizes the fields as *non-sensitive*, and they will not be anonymized.

Besides "SR" and "1", used in the example in Figure 2, other classifications can be used when defining a policy: the suppression can be done from left to right using the abbreviation "SL"; date generalization can be done through “DT”; the number "2" categorizes the fields as *sensitive*, and they will be automatically suppressed; the number "3" categorizes the fields as *quasi-identifier*, and they will be automatically micro-aggregated; the number "4" categorizes the fields as *identifier*, and they will be automatically suppressed.

The last line of the policy must define the re-identification risk threshold, through the abbreviation "*Rmax*" followed by the value for the desired risk threshold, expressed in decimal format. Some works (e.g., [2][3]) suggest thresholds from 1% (0.01) to 5% (0.05) as acceptable to be used for research data.

It is important to mention that, when defining the threshold, the data utility must be considered (the lower the threshold, the lower the data utility).

**4. Using the Jar Application**

1. Download Branch ARX\Risk\Integrated using the command terminal:

git clone -b ARX\Risk\Integrated http://gitlab.dei.uc.pt/nuno/priva.git

2. Compile the project and generate the runable jar.

3. Run the java executable:

java -jar path... priva\developmentarx-poc\dist\arx-poc.jar "path to csv dataset " "path to csv policy"

(the location of .jar file can be changed depending of compilation software used).

4. the results will be showed at the screen, and the anonymized file will be created in the home path of PRIVA project.

Figure 3 shows an example of the component execution. In this example, the input data was the one whose sample was showed in Figure 1. The risk threshold was defined as 0.01 (see the policy in Figure 2). To reach a risk value lower than the threshold, the value of k (from k-anonymity algorithm) was increased to 24.



Figure 3. Example of the component execution.

Figure 4 shows a small sample of the anonymized file. The Zip Code was anonymized from right to left, with three digits suppressed. The number of digits to be suppressed is determined by the k-anonymity algorithm. It suppresses the digits until it reaches the number of similar data equal to k (for this example, three digits were enough to have a result with at least 24 similar zip codes). It is possible to observe in the sample that, after anonymization, some zip codes have the same value.



Figure 4. Sample of the anonymized dataset.

**References**

[1] ARX> (2017) ARX data anonymization tool. [Online]. Available: http://arx.deidentifier.org.

[2] K. El Emam, D. Paton, F. Dankar, and G. Koru, “De-identifying a public use microdata file from the canadian national discharge abstract database,” BMC medical informatics and decision making, vol. 11, no. 1, p. 53, 2011.

[3] H. L. Howe, A. J. Lake, and T. Shen, “Method to assess identifiability in electronic data files,” American Journal of Epidemiology, vol. 165, no. 5, pp. 597–601, 2006.