Data Anonymization Tool

Mario Münzer
coordination@supercloud-project.eu
TECHNIKON Forschungs- und Planungsgesellschaft mbH
May, 2017

Contents

Chapter 1 Anonymization

1.1 Introduction

Anonymization techniques open the possibility of releasing personal and sensitive data, while preserving individual's privacy. Therefore, data anonymization guarantees that revealed data cannot be assigned to a natural person nor inferences to user's identity can be made. There are several techniques known, which are applicable for data anonymization, as described in the SUPERCLOUD deliverable D3.2 [3]. The characterized data anonymization tool within this chapter is among others based on k-anonymity, whereby the focus is put on the irreversibility of released data.
The aim of the data anonymization tool is to meet the goal of k-anonymity regarding irreversibility. As a result of this, the disclosure of sensitive data (e.g. health data) is impossible and the opening of data is enabled, whereby each data record is at least $k - 1$ from other records indistinguishable with respect to the quasi-identifier. However, the tool is not simply performing calculations on medical data and anonymizing them. Moreover, the tool aims to calculate the best solution for the given data in terms of cost-efficiency. This is done by means of so-called cost metric calculation as well as the Optimal Lattice Anonymization (OLA) algorithm, as described in D3.2 in detail.
Privacy-enabling mechanisms for untrusted cloud(s) represent an explorative subdomain of the SUPERCLOUD architecture. Therefore, data anonymization techniques, such as k-anonymity, were from the beginning of the project under consideration in the overall architecture design (depicted in SUPERCLOUD's deliverable D3.1 [2]). In the architecture proposal of WP3 of SUPERCLOUD, the data anonymization tool is used to enable the release of medical-related sensitive data. In order to guarantee a secure storage as well as a trusted health data exchange, the tool will be integrated within the SUPERCLOUD data management architecture, and in particular with the Janus storage service.

1.2 Component Description

As already mentioned, the data anonymization tool is among others based on k-anonymity. Therefore, generalization and suppression is used, as described in detail in the previous deliverable D3.2 [3]. While suppression deletes uniquely identifying attributes, generalization is necessary in order to obtain k-anonymity, respectively to obtain k-identical values by generalizing the pre-selected attributes (a set of these attributes is called quasi-identifier in the following). One further element of the data anonymization tool is the precision cost metric algorithm by Sweeney [4]. Besides the achievement of the anonymity level (k) for given medical data, the precision of each applicable node has to be calculated. As a result, the height and depth of the generalization hierarchy will be considered and the ratio between applicable and total generalization steps determined. As mentioned previously, the main goal of the data anonymization tool is to (besides the irreversible anonymization and opening of data) determine the best solution in terms of cost-efficiency with most minimal information loss for the given data. Hence, a potential solution is given by its k and precision. However, there could be several potential solutions available with same characteristics based on anonymity level and precision only. Therefore, the data anonymization tool includes one more important element in order to determine the optimal solution for the provided medical data. The OLA algorithm [1] is the last element of the anonymization tool and is responsible for determining efficiently the optimal solution, respectively the optimal node in the so-called lattice, by means of divide-and-conquer technique. The detailed steps of

the Optimal Lattice Anonymization as well as a pseudo code of all including functions of the algorithm can be found in deliverable D3.2 [3].
Since the OLA algorithm has to traverse through all possible solutions, a list of nodes, also known as lattice, is required as input. The lattice represents a stepped generalization of the given data in form of a node list and contains the current generalization step, the k and the information loss, respectively the cost metric precision. The lattice itself is built automatically within the anonymization tool based on the quasi-identifier and its total possible generalization steps, respectively the total generalization step of each attribute. The OLA algorithm traverses through the lattice according to the divide-and-conquer principle and marks all non-applicable as well as already traversed nodes with tags for best efficiency. The detailed procedure of the lattice traverse can be found in the previous SUPERCLOUD's deliverable D3.2 [3] as well.

1.2.1 Software Development

The software development of the data anonymization tool was completely done in Microsoft's object-oriented programming language C#. Therefore, a graphical user interface (GUI) was made as well in order to support the user's input and control. Since the software tool was built by means of Microsoft's Visual Studio default libraries (included in .NET 4.x framework), such as *System.IO*; *System.Windows.Forms*; *System.Threading*; *System.Data*; etc, there is no further external or third-party library necessary to run the program. The final software tool is resulting in an executable program (*.exe), which is supported on computers with Windows operating system only. Further details about the tool access can be found at the end of this chapter (section **??**).

1.2.2 Procedure

The data anonymization tool is composed of three main components: k-anonymity calculation; cost metric computation; OLA algorithm. Therefore, the procedure from the input of plain health data to the output of irreversible anonymized data records is straightforward. The tool accepts as input plain data records in comma-separated values (CSV) file format. Given a valid source file, the user then has to select the unique-identifying attribute(s) as well as the quasi-identifier, as it is depicted in Figure 1.1 (selected quasi-identifier attributes highlighted in green).

Figure 1.1: Selection of unique- and quasi-identifier attributes

By means of look-up tables (LUTs), the tool automatically checks the total generalization level of each selected quasi-identifier attribute[1]. Based on the quasi-identifier and its total generalization level, the node list (lattice) can be built, which is depicted in Figure 1.2.

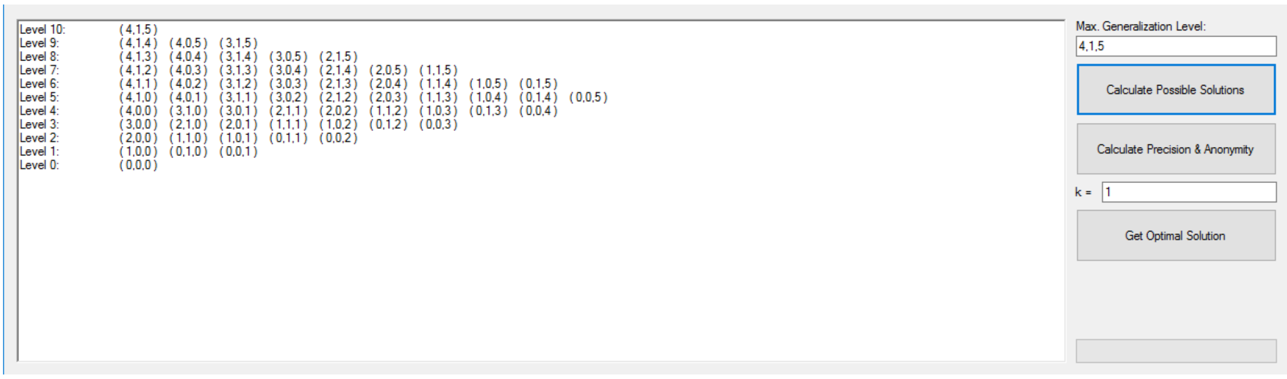

Figure 1.2: Generation of lattice based on maximum generalization level [4,1,5]

Since the OLA algorithm requires for the calculation of the best/optimal solution the k (anonymity level) and the precision (information loss), a further computation has to be done. The resulting output can be found in Figure 1.3

Figure 1.3: Calculation of anonymity level and precision based on medical data records and its maximum generalization level [4,1,5]

Besides the plain health data and identifier selection, the user has to select the lower bound for the anonymity level (k). In the end of the procedure, the OLA algorithm based on the valid inputs (source file with data records; quasi-identifier; total generalization level; lattice; anonymity level boundary) can be applied. At this stage, the OLA algorithm computes the best-fitting node of the lattice based on the provided characteristics and the optimal solution will be displayed. However, the calculation is performed on the lattice only, so no anonymization on the provided data will take place at this time. Therefore, the user has now the possibility to apply the optimal solution on all loaded data records or decline the result, as seen in Figure 1.4.

[1]If there is no pre-defined LUT for the selected quasi-identifier attribute, a default value will be assumed

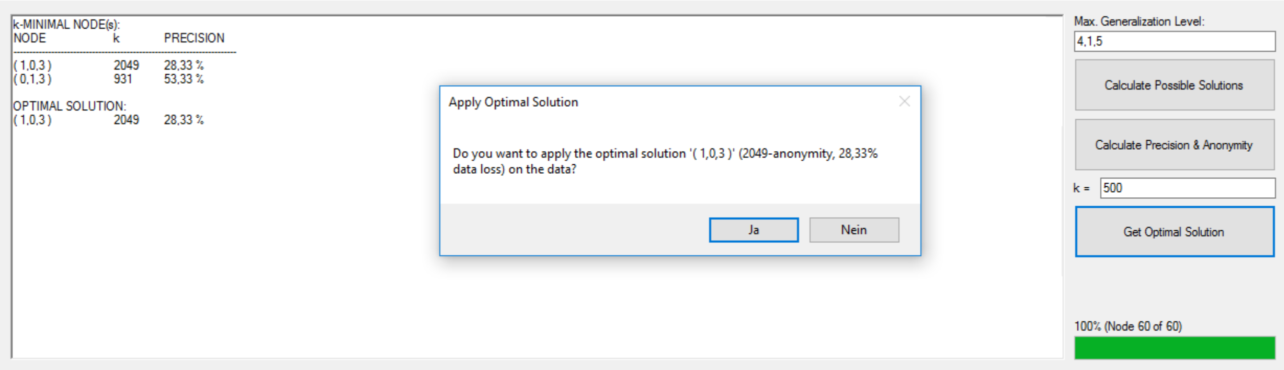

Figure 1.4: Calculation of optimal anonymization node by means of OLA algorithm based on anonymity level boundary of $k = 500$

The final outcome of the data anonymization tool is illustrated in Figure 1.5. Within the stated example, the OLA algorithm resulted for anonymization level boundary of $k = 500$, the node [1,0,3] is resulting. Since the quasi-identifier is composed of *Age*, *Gender* and *ZIP Code*, consequential the age is generalized once, the ZIP code three times and the gender not once at all. Therefore, the information loss is about 28% and the anonymization level is at $k = 2049$. Thus, there exist (at least) 2049 (out of 100,000) data records, which are not distinguishable from each other (considering the selected quasi-identifier attributes only).

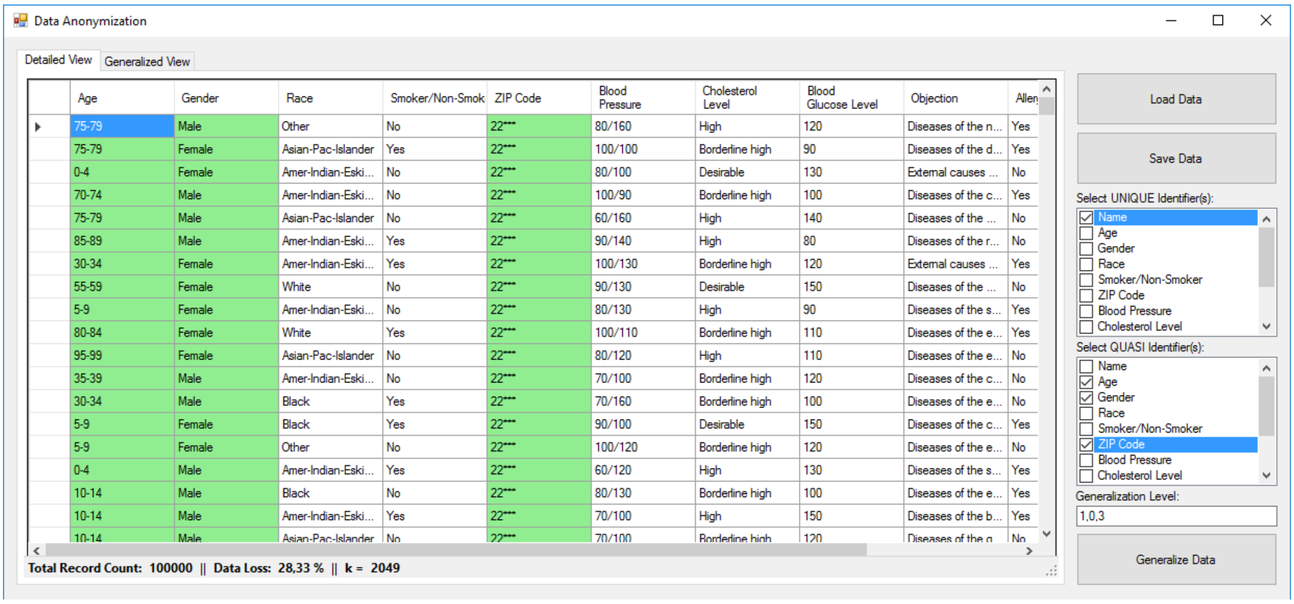

Figure 1.5: Representation of the final outcome of the data anonymization tool by applying the node [1,0,3]

The subsequent listing sums up the necessary steps of the data anonymization tool, as described in this chapter.

1. Input of valid data records
2. Selection of unique- and quasi-identifier attributes
3. Generation of node list (lattice)
4. Calculation of anonymity level (k) and information loss (precision)

5. Set of lower anonymity level boundary
6. Calculation of optimal anonymization by means of OLA algorithm

1.3 Documentation

As already mentioned, a more detailed description of all included elements of the data anonymization tool can be found in Chapter 13 of SUPERCLOUD's deliverable D3.2 [3].

Bibliography

[1] K. El Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J. P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley. A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

[2] Mario Münzer, Sébastian Canard, Marie Paindavoine, Alysson Bessani, Caroline Fontaine, Krysztof Oborzyński, Meilof Veeningen, and Paulo Sousa. D3.1 - Architecture for Data Management. *SUPERCLOUD*, 2015.

[3] Mario Münzer, Sébastian Canard, Marie Paindavoine, Andre Nogueira, Antonio Casimiro, João Sousa, Joel Alcântara, Tiago Oliveira, Ricardo Mendes, Alysson Bessani, Christian Cachin, Simon Schubert, Caroline Fontaine, Daniel Pletea, Meilof Veeningen, and Jialin Huang. D3.2 - Specification of Security Enablers for Data Management. *SUPERCLOUD*, 2016.

[4] Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002.