A Two-Levels Data Anonymization Approach

Sarah Zouinina, Younès Bennani, Nicoleta Rogovschi, Abdelouahid Lyhyaoui

A two-levels data anonymization approach *

Sarah Zouinina[1,2,4], Younès Bennani[1,2], Nicoleta Rogovschi[1,3] and Abdelouahid Lyhyaoui[4]

[1] LIPN UMR 7030 CNRS, Université Sorbonne Paris Nord, France
[2] LaMSN, La Maison des Sciences Numériques, Université Sorbonne Paris Nord
[3] LIPADE, Université de Paris, France
[4] LTI-ENSA, Université Abdelmalek Essaadi, Tanger, Morocco
Firstname.Lastname@lipn.univ-paris13.fr
rogovschi@parisdescartes.fr
lyhyaoui@gmail.com

Abstract. The amount of devices gathering and using personal data without the person's approval is exponentially growing. The European General Data Protection Regulation (GDPR) came following the requests of individuals who felt at risk of personal privacy breaches. Consequently, privacy preservation through machine learning algorithms were designed based on cryptography, statistics, databases modeling and data mining. In this paper, we present two-levels data anonymization methods. The first level consists of anonymizing data using an unsupervised learning protocol, and the second level is anonymization by incorporating the discriminative information to test the effect of labels on the quality of the anonymized data. The results show that the proposed approaches give good results in terms of utility what preserves the trade-off between data privacy and its usefulness.

Keywords: Data Anonymization · Learning Vector Quantization · Data Anonymization · Privacy Utility Tradeoff · Microaggregation

1 Introduction

Due to the saturation of cities with smartphones and sensors, the amount of information gathered about each individual is frightening. Humans are becoming walking data factories and third-parties are tempted to use personal data for malicious purposes. To protect individuals from the misuse of their precious information and to enable researchers to learn from data effectively, data anonymization is introduced with the purpose of finding balance between the level of anonymity and the amount of information loss. Data anonymization is therefore defined as *it is the process of protecting individuals' sensitive information while preserving its type and format* [17] [12].

Hiding one or multiple values or even adding noise to data as an attempt to anonymize data is considered inefficient because the reconstruction of the initial

* Supported by the ANR Pro-Text project. N° ANR-18-CE23-0024-01

information is very probable [15]. Machine learning for data anonymization is still an underexplored area [3], although it provides some good assets to the field of data security. Inspired from the k-anonymization technique proposed by Sweeney [16], we aim to create micro clusters of similar objects that we code using the micro cluster's representative. In this way, the distortion of the data is minimal and the usefulness of the data is maximal. This can be achieved using supervised or unsupervised methods. For the unsupervised methods [8], the most used approach is the clustering that allows to open a new research direction in the field of anonymization i.e. create clusters of k elements and replace the data by the prototypes of the clusters (centroids) in order to obtain a good trade-off between the information loss and the potential data identification risk. However, usually, these approaches are based on the use of the k-means algorithm which is prone to local optima and may give biased results.

In this paper we answer the question of *how can the introduction of discriminative information affect the quality of the anonymized datasets*. To this purpose, we revisited all the previously proposed approaches, and we added a second level of anonymization by incorporating the discriminative information and using Adaptive Weighting of Features to improve the quality of the anonymized data. This aims to improve the anonymized data quality without compromising its level of privacy. The paper is organised into four sections: the first dresses the different approaches of privacy preserving using machine learning, the second sums up the previously proposed approaches, the third discusses the introduction of the discriminative information and the fourth validates the method experimentally on six different datasets.

2 Privacy Preservation using Machine Learning

Anonymization methods for microdata rely on many mechanisms and *data perturbation* is the common technique binding them all. Those mechanisms modify the original data to improve data privacy but inevitably at cost of some loss in data utility. Strong privacy protection requires masking the original data and thus reducing its utility. Microaggregation is a technique for disclosure limitation aimed at protecting the privacy of data subjects in microdata releases. It has been used as an alternative to generalization and suppression to generate k-anonymous data sets, where the identity of each subject is hidden within a group of k subjects. Unlike generalization, microaggregation perturbs the data and this additional masking freedom allows improving data utility in several ways, such as increasing data granularity, reducing the impact of outliers and avoiding discretization of numerical data [4] microaggregation. Rather than publishing an original variable V_i for a given record, the average of the values of the group over which the record belongs is published. In order to minimize information loss, the groups should be as homogeneous as possible. The impact of microaggregation on the utility of anonymized data is quantified as the resulting accuracy of a machine learning model trained on a portion of microaggregated

data and tested on the original data [13]. Microaggregation is measured in terms of *syntactic distortion*.

Achieving microaggregation might be done using machine learning models, like *clustering* and/or *classification*. LeFevre et al. [7] propose several algorithms for generating an anonymous data set that can be used effectively over predefined workloads. Workload characteristics taken into account by those algorithms include selection, projection, classification and regression. Additionally, LeFevre et al. consider cases in which the anonymized data recipient wants to build models over multiple different attributes. Nearest neighbor classification with generalization has been investigated by [11]. The main purpose of generalizing exemplars (by merging them into hyper-rectangles) is to improve speed and accuracy as well as inducing classification rules, but not to handle anonymized data. Martin proposes building non-overlapping, non-nested generalized exemplars in order to induce high accuracy. Zhang et al. discuss methods for building naive Bayes and decision tree classifiers over partially specified data [6] [18]. Partially specified records are defined as those that exhibit nonleaf values in the taxonomy trees of one or more attributes. Therefore generalized records of anonymous data can be modeled as partially specified data. In their approach classifiers are built on a mixture of partially and fully specified data. Inan et al. [5] address the problem of classification over anonymized data. They proposed an approach that models generalized attributes of anonymized data as uncertain information, where each generalized value of an anonymized record is accompanied by statistics collected from records in the same equivalence class. They do not assume any probability distribution over the data. Instead, they propose collecting all necessary statistics during anonymization and releasing these together with the anonymized data. They show that releasing such statistics does not violate anonymity.

3 Clustering for Data Anonymization

3.1 *k*-TCA and Constrained TCA

In previous articles we introduced an approach of k-anonymity using Collaborative Multi-view Clustering [22] and a k-anonymity through Constrained Clustering [21]. The two models propose an algorithm that relies on the classical Self Organizing Maps (SOMs) [10] and collaborative Multiview clustering in purpose to provide useful anonymous datasets [9]. They achieve anonymization in two-levels, the pre-anonymization step and the anonymization step. The pre-anonymization step is similar for both algorithms and it consists of horizontally splitting data so each observation is described in different spaces and then using the collaborative paradigm to exchange topological information between collaborators. The Davies Bouldin index (DB) [2] is used in this case as a clustering validity measure and a stopping criterion of the collaboration. When DB decreases, the collaboration is said to be positive, but if it increases, the collaboration is clearly negative, since it is degrading the clustering quality and

therefore the utility of the provided anonymous data. The topological collaborative multiview clustering outputs homogeneous clusters after the clustering, the individuals contained in each view are coded using the Best Matching Units of each neuron in the case of k-TCA and using the linear mixture of models in the case of Constrained TCA. The pre-anonymized views are then gathered to be reconstructed in the same manner as the original dataset.

The anonymization step of the algorithms is totally different between the two. In the k-TCA, the pre-anonymized dataset will be fine-tuned using a SOM model with a map size determined automatically using the Kohonen heuristic. Each individual of the dataset is then coded using the BMU of the cluster and the level of k-anonymity is evaluated. In those model we have the advantage of determining the k-anonymity level automatically. In the second algorithm, Constrained TCA (C-TCA), the k level of anonymity is fixed ahead, before starting the experiments. A SOM is created using the pre-anonymized dataset as an input. Each node is examined to determine if it respects the constraint of k element in each cluster. Respectively the elements captured by the neurons that don't respect the predefined constraint are redistributed on the closest units. By using this technique, we design clusters of at least k elements and we code the objects using the BMUs in order to have k-anonymized dataset. We then evaluate the best k level that gives a good tradeoff between anonymity and utility.

3.2 Attribute Oriented Kernel Density Estimation for Data Anonymization

Another method that we proposed to anonymize a dataset was the Attribute Oriented Kernel Density Estimation [20]. The choice of 1 dimensional KDE was motivated by the ability of the model to determine where data is grouped together and where it is sparse relying on its density. KDE is a non parametric model that uses probability density to investigate the inner properties of a given dataset. The algorithm that we propose clusters the data by determining the points where density is the highest (local maximas) and the points with the smallest density (local minimas): those local minimas refer to the clusters' boarders and the local maximas are the clusters' prototypes. KDE is a non-parametric approach to approximate the distribution of a dataset and overcome the inability of the histograms to achieve this estimation because of the discontinuity of the bins. Each object that falls between two minimas is recoded using the corresponding local maxima. Doing this at a one dimensional level helps preserving the characteristics of each feature in the dataset and thus doesn't compromise its utility.

4 Incorporating discriminative power during anonymization process

After evaluating the different results of data anonymization using the methods in the previous works, we asked the question *What if data was labelled?* and

How the supervision can influence the obtained utility results? To answer to those questions we used the Learning Vector Quantization approach (LVQ). We applied it to enhance the clustering results of each of our proposed methods. LVQ is a pattern recognition model that takes advantage of the labels to improve the accuracy of the classification. The algorithms learns from a subset of patterns that best represent the training set.

The choice of the Learning Vector Quantization (LVQ) method was motivated by the simplicity and rapidity of convergence of the technique, since it is based on the hebbian learning. This is a prototype-based method that prepares a set of codebook vectors in the domain of the observed input data samples and uses them to classify unseen examples. Kohonen presented the self organizing maps as an unsupervised learning paradigm that he improves using a supervised learning technique, called the learning vector quantization. It is a method used for optimizing the performances of a trained map in a reward-punishment scheme.

Learning Vector Quantization was designed for classification problems that have existing data sets that can be used to supervise the learning by the system. LVQ is non-parametric, meaning that it does not rely on assumptions about that structure of the function that it is approximating. Euclidean distance is commonly used to measure the distance between real-valued vectors, although other distance measures may be used (such as dot product), and data specific distance measures may be required for non-scalar attributes. There should be sufficient training iterations to expose all the training data to the model multiple times. The learning rate is typically linearly decayed over the training period from an initial value until it is close to zero. Multiple passes of the LVQ training algorithm are suggested for more robust usage, where the first pass has a large learning rate to prepare the codebook vectors and the second pass has a low learning rate and runs for a long time (perhaps 10-times more iterations).

In the Learning Vector Quatization model, each class contains a set of fixed prototypes with the same dimension of the data to be classified. LVQ adaptively modifies the prototypes. In the learning algorithm, data is first clustered using a clustering method and the clusters' prototypes are moved using LVQ to perform classification. We chose to supervise the results of the clustering by moving the center clusters' using the wLVQ2 proposed in algorithm 1 for each of the approaches. We use the wLVQ2 [1] since this upgraded version of the LVQ respects the characteristics of each features and adapts the weighting of each feature according to its participation to the discrimination. The system learns using two layers: the first layer calculates the weights of the features and then it is presented to the LVQ2 algorithm.

The cost function of this algorithm can be written as follows:

$$R_{wLVQ2}(x, m, W) = \begin{cases} \|Wx - m_j\|^2 - \|Wx - m_i\|^2, \ If \ C_k = C_j \\ 0, \quad otherwise \end{cases}$$

Where $x \in C_k$ and W is the weighting coefficient matrix; m_i is the nearest codeword vector to Wx and m_j is the second nearest codeword vector to

Algorithm 1 Adaptive Weighting of Pattern Features During Learning

Initialization :
Initialize the matrix of weights W according to :

$$w_j^i = \begin{cases} 0, \ when \ \text{i}\neq\text{j} \\ 1, \ when \ \text{i}=\text{j} \end{cases}$$

The codewords **m** are chosen for each class using the k-means algorithm.
Learning Phase:

1. Present a learning example x.
2. Let $m_i \in C_i$ be the nearest codeword vector to x.
   - **if** $x \in C_i$, then go to 1
   - **else then**
     - let $m_j \in C_j$ be the second nearest codeword vector
     - **if** $x \in C_j$ then
       * a symmetrical window *win* is set around the mid-point of m_i and m_j.
       * **if** x falls within *win*, then
         Codewords Adaptation:
       * m_i is moved away from x according to the formula
         $$m_i(t+1) = m_i(t) + \alpha(t)[Wx(t) - m_j(t)]$$
       * m_j is moved closer x according to the formula
         $$m_j(t+1) = m_j(t) - \alpha(t)[Wx(t) - m_j(t)]$$
       * for the rest of the codewords
         $$m_k(t+1) = m_k(t)$$
         Weighting Patterns features:
       * adapt w_k^k according to the formula:
         $$w_k^k(t+1) = w_k^k(t) - \beta(t)x^k(t)(m_i^k(t) - m_j^k(t))$$
       * go to 1.

Where $\alpha(t)$ and $\beta(t)$ are the learning rates

Wx. The wLVQ2 with the Collaborative Paradigm enhances the utility of the anonymized data by the k-TCA and the Constrained TCA (C-TCA) models, the use of wLVQ2 is done after the collaboration between cluster centers' to improve the results of the Collaboration at the pre-anonymization and the anonymization steps.

The experimental protocol of using wLVQ2 with Attribute-oriented data anonymization and Kernel Density Estimation, takes in account the labels of

the dataset and improves the found prototypes and then represents the micro-clusters using them.

5 Experimental Validation

5.1 Datasets

Six datasets from the UCI machine learning repository are used in the experiment. The table below presents the main characteristics of these databases.

Table 1. Some Characteristics of Datasets

Datasets	**#Instances**	**#Attributes**	**#Class**
Ecoli	336	8	8
Electrical	10000	14	2
Glass	214	10	7
Page blocks	5473	10	5
Waveform	5000	21	3
Yeast	1484	8	10

5.2 Quality Validity Indices

Cluster validity consists of techniques for finding a set of clusters that best fits natural partitions without any a priori class information. The outcome of the clustering process is validated by a cluster validity index. Internal validation measures reflect often the compactness, the connectivity and the separation of the cluster partitions. We choose to validate the results of the proposed methods using Silhouette Index and Davies Bouldin Index. The results are given in the tables 2 and 3.

Table 2. Silhouette Index

	Ecoli	Electrical	Glass	Page Blocks	Yeast	Waveform
k-TCA	0.26	-0.05	0.42	-0.40	0.13	0.18
k-TCA^{++}	0.89	0.08	0.59	-0.49	0.84	0.24
C-TCA	0.24	-0.05	0.43	-0.34	0.07	0.13
C-TCA^{++}	0.84	0.08	0.45	-0.43	0.81	0.25
KDE	0.26	0.069	-0.19	-0.54	0.28	-0.27
KDE^{++}	0.99	0.99	0.57	0.98	0.99	1

As illustrated, the Attribute oriented microaggregation using wLVQ2 (++: Discriminative version of each approach, KDE^{++}, k-TCA^{++},C$-TCA^{++}$) outperforms by far the Attribute Oriented microaggregation in both Silhouette and Davies Bouldin indices.

Table 3. Davies Bouldin Index

	Ecoli	Electrical	Glass	Page Blocks	Yeast	Waveform
k-TCA	2.68	2.28	0.40	3.23	2.31	1.51
k-TCA^{++}	0.59	3.38	0.40	3.11	0.24	1.37
C-TCA	1.61	2.58	0.55	3.04	2.95	1.92
C-TCA^{++}	0.14	3.38	0.51	3.10	0.26	1.35
KDE	0.57	3.96	4.99	3.83	2.43	6.96
KDE^{++}	9.91E-08	0.02	1.32	0.52	4.20E-08	3.58E-06

5.3 Combined Utility Measure

Separability Utility To measure the utility of the anonymized datasets we propose a test on the original and the anonymized data. The test consists of comparing the accuracy of a decision tree model with 10 folds cross validation before and after microaggregation to evaluate the practicality of the proposed anonymization. We call it separability utility since it measures the separability of the clusters. We give the results of this measure in table 4, we also provide a comparison between the separability utility measures of the original and the anonymized datasets.

The separability measure was improved after LVQ for 83% of the tests done on the datasets, this can be explained by the tendency of microaggregation to remove non decisive attributes from the dataset in order to gather together elements that are similar. The $^{++}$ in the name of the methods refers to discriminant version.

Table 4. Separability Utility: Glass, Ecoli, Electrical, Page Blocks, Waveform & Yeast

Datasets	***Glass***	***Ecoli***	***Electrical***	***PageBlocks***	***Waveform***	***Yeast***
Before Anonymization	0.692	0.821	0.995	0.966	0.748	0.812
k**-TCA**	0.943	0.845	0.999	0.905	0.830	0.862
k**-TCA** $^{++}$	0.944	0.988	0.735	0.919	0.884	1
C**-TCA**	0.747	0.848	0.999	0.915	0.816	0.876
C**-TCA** $^{++}$	0.859	0.863	0.745	0.918	0.884	0.887
KDE	0.701	0.801	0.988	0.955	0.755	0.834
KDE $^{++}$	0.743	0.806	0.982	0.962	0.758	0.845

Structural Utility using the Earth Mover's Distance We believe that measuring the distance between two distributions is the way to evaluate the difference between the datasets. The amount of utility lost in the process of anonymization can be see as the distance between the anonymized dataset and the original one.

The Earth Mover's distance (EMD) also known as the Wasserstein distance [14], extends the notion of distance between two single elements to that of a

distance between sets or distributions of elements. It compares the probability distributions P and Q on a measurable space (Ω, Ψ) and is defined as follows (We are using the distance of order 1):

$$W_1(P,Q) = \inf_{\mu} \left\{ \int_{\Omega\times\Omega} |x-y| \mathrm{d}\mu(x,y) \,\middle|\, \begin{array}{l} \mu \text{ :prob. measure on } (\Omega\times\Omega, \Psi\otimes\Psi) \\ with\ marginals : P, Q \end{array} \right\} \quad (1)$$

where $\Omega \times \Omega$ is the product probability space. Notice that we may extend the definition so that P is a measure on a space (Ω, Ψ) and Q is a measure on a space (Ω', Ψ').

Let us examine how the above is applied in the case of discrete sample spaces. For generality, we assume that P is a measure on (Ω, Ψ) where $\Omega = \{x_i\}_{i=1}^{n}$ and Q is a measure on (Ω', Ψ') where $\Omega' = \{y_i\}_{j=1}^{n'}$ - the two spaces are not required to have the same cardinality.

Then, the distance between P and Q becomes:

$$W_1(P,Q) = \inf_{\{\lambda_{i,j}\},i,j} \left\{ \sum_{i=1}^{n}\sum_{j=1}^{n'} \lambda_{i,j}|x_i - y_j| : \sum_{i=1}^{n} \lambda_{i,j} = q_j, \sum_{j=1}^{n'} \lambda_{i,j} = p_i, \lambda_{i,j} \geq 0 \right\}$$

EMD is the minimum amount of work needed to transform a distribution to another. In our case we measure the EMD between the anonymized and the original datasets, attribute by attribute, to get an idea about the distortion of the anonymized datasets. We then normalize all distances between 0 and 1, then we define the utility by $1 - W_1(P,Q)$. The smaller the distance W_1 is, the more the data utility is preserved.

Preserving combined utility To choose the anonymization method which best addresses the separability-Structural utility Trade-off, we propose to combine the two types of utility structural and separability in a combined form while $\alpha = \frac{1}{2}$:

$$Comb\_Utility = \alpha.Separability + (1-\alpha).Structural$$

Table 5 summarize the clustering results of the proposed approaches in terms of combined utility ($Comb\_Utility$). As it can be seen, our approach Attribute-oriented generally performs best on all the datasets. To further evaluate the performance, we compute a measurement score by following [19]:

$$Score(A_i) = \sum_j \frac{Comb\_Utility(A_i, D_j)}{\max_i Comb\_Utility(A_i, D_j)}$$

where $Comb\_Utility(A_i, D_j)$ refers to the combined Utility value of A_i method on the D_j dataset. This score gives an overall evaluation on all the datasets, which shows our approach Attribute-oriented outperforms the other methods substantially in most cases.

As shown in the table 5, the introduction of the discriminant information improves the utility of the anonymized datasets for all of the methods proposed.

Table 5. Combined separability and structural utility Comb_Utility

	Ecoli	Electrical	Glass	Page Blocks	Waveform	Yeast	**Score**
k-TCA	0.63	0.74	0.71	0.60	0.66	0.51	**4.96**
k-TCA^{++}	0.78	0.62	0.74	0.82	0.69	0.92	**5.18**
C-TCA	0.74	0.75	0.54	0.70	0.65	0.51	**4.92**
C-TCA $^{++}$	0.62	0.62	0.76	0.71	0.70	0.87	**5.20**
KDE	0.60	0.54	0.44	0.71	0.63	0.73	**4.98**
KDE $^{++}$	0.83	0.95	0.91	0.77	0.75	0.79	**5.17**

6 Conclusion

In this paper we studied the impact of incorporating the discriminative information to improve data anonymization level and to preserve its usefulness. The anonymization is achieved in two levels process. The first, uses one of these three methods: k-TCA or Constrained TCA (C-TCA) or Attribute Oriented KDE, that we introduced for data anonymization through microaggregation approach. And the second, through the use of labels and the learning of the vectors weights adaptively using the weighted LVQ. The experimental investigation shown above prove the efficiency of the methods and illustrate its importance. The main contributions of the article are the addition of the supervised learning layer to improve utility of the model without compromising its anonymity. The separability utility reflects the usefulness of the data and the structural utility shows its level of anonymity. The combined utility is a weighted measure that combines both measures, we can change the weight of the utility tradeoff depending on wich side we want to emphasise on.

References

1. Bennani, Y.: Adaptive weighting of pattern features during learning. In: IJCNN'99. International Joint Conference on Neural Networks. Proceedings. vol. 5, pp. 3008–13. IEEE Service Center, Piscataway, NJ (1999)
2. Davies, D., Bouldin, D.: A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-1**(2), 224–227 (1979)
3. Domingo-Ferrer, J., Soria-Comas, J., Mulero-Vellido, R.: Steered microaggregation as a unified primitive to anonymize data sets and data streams. IEEE Transactions on Information Forensics and Security **14**(12), 3298–3311 (2019)
4. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., De Wolf, P.P.: Statistical disclosure control. John Wiley & Sons (2012)
5. Inan, A., K.M.B.E.: Using anonymized data for classification. In: Data Engineering. pp. 429–440. IEEE 25th International Conference (2009)
6. J. Zhang, D.-K. Kang, A.S., Honavar, V.: Learning accurate and concise naive bayes classifiers from attribute value taxonomies and data. Knowl. Inf. Syst. **9**, 157–179 (2006)

7. K. LeFevre, D.J.D., Ramakrishnan, R.: Workload-aware anonymization. in KDD'06 pp. 277–286 (2006)
8. Khan, S., Iqbal, K., Faizullah, S., Fahad, M., Ali, J., Ahmed, W.: Clustering based privacy preserving of big data using fuzzification and anonymization operation. ArXiv **abs/2001.01491** (2019)
9. Kim, S., Chung, Y.D.: An anonymization protocol for continuous and dynamic privacy-preserving data collection. Future Generation Computer Systems **93**, 1065 – 1073 (2019)
10. Kohonen, T.: Self-organizing Maps. Springer-Verlag Berlin, Berlin (1995)
11. Martin, B.: Instance-based learning: Nearest neighbour with generalisation. Master's thesis, Computer Science Department, Hamilton, New Zealand, University of Waikato (1995)
12. Raghunathan, B.: The Complete Book of Data Anonymization: From Planning to Implementation. CRC Press (2013),
13. Rodríguez-Hoyos, A., Estrada-Jiménez, J., Rebollo-Monedero, D., Parra-Arnau, J., Forné, J.: Does k-anonymous microaggregation affect machine-learned macrotrends? IEEE Access **6**, 28258–28277 (2018)
14. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)
15. Sharma, A., Singh, G., Rehman, S.: A review of big data challenges and preserving privacy in big data. In: Kolhe, M.L., Tiwari, S., Trivedi, M.C., Mishra, K.K. (eds.) Advances in Data and Information Sciences. pp. 57–65. Springer Singapore, Singapore (2020)
16. Sweeney, L.: K-anonymity: A model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. **10**(5), 557–570 (Oct 2002)
17. Venkataramanan, N., Shriram, A.: Data Privacy: Principles and Practice. Chapman & Hall/CRC (2016)
18. Zhang, J., Honavar, V.: Learning from attribute value taxonomies and partially specified instances. pp. 880–887. Washington, DC, USA (2003)
19. Zhao, H., Fu, Y.: Dual-regularized multi-view outlier detection. In: IJCAI (2015)
20. Zouinina, S., Bennani, Y., Ben-Fares, M., Lyhyaoui, A., Rogovschi, N.: Preserving utility during attribute-oriented data anonymization process. Australian Journal of Intelligent Information Processing Systems (2019)
21. Zouinina, S., Grozavu, N., Bennani, Y., Lyhyaoui, A., Rogovschi, N.: Efficient k-anonymization through constrained collaborative clustering. In: IEEE Symposium Series on Computational Intelligence, SSCI 2018 (2018)
22. Zouinina, S., Grozavu, N., Bennani, Y., Lyhyaoui, A., Rogovschi, N.: A topological k-anonymity model based on collaborative multi-view clustering. In: Artificial Neural Networks and Machine Learning - ICANN 2018 (2018)