A FRAMEWORK FOR ENHANCED MISSING VALUE RECORD SETS ANONYMIZATION

[1]A. RANGAMMA, [2] K. APARNA, [3] KS. ARCHANA

[1,2, 3] Assistant Professor, Department of Computer Science and Engineering, Sri Indu College of Engineering and Technology, Hyderabad, Telangana-501510

ABSTRACT

Increase in use of digital platforms generate a very huge amount of user data which on processing reveals more valuable information about users while mining or it may also reveal some future events. Privacy preserving data mining (PPDM) is the current developing area of research that precise to ensure a certain level of privacy by increasing the utility of information. Data anonymization is PPDM technique that protects sensitive information in the dataset with high confidence. Anonymizing missing value record sets suffers from serious information loss due to the missing value pollution that arises because of missing values present in the original micro data. In this paper, we propose an enhanced framework to anonymize the missing value record sets with decreased information loss and increased utility. We study the properties of missing value pollution on generalization. Guided by these properties, we develop (k,R_m) anonymity model which is to preserve data utility on missing value records sets. The results obtained by executing our framework over real-world dataset proved that it is optimized in providing better trade-off between utility and privacy for missing value record sets than the existing procedures.

Keywords: *Missing Value, k –anonymization, generalization and suppression*

INTRODUCTION

Nowadays society has faced massive growth in the acquisition of data's. So, protecting the sensitive information of the individual data is important. We should also protect privacy against by joining multiple public data sources to re-identify the person is very easy so we are proposing k-anonymity.E.g. In online social media it may release de-identified relationship dataset so that the analyzer can study the characteristics of online social media .However, dataset has the Potentially Identifiable Information (PII) of personal disclosure. According to Sweeny article expel the personal identities from microdata is insufficient due to the linking attacks. The competitor can link some attributes named Quasi Identifier (Qid) with the external dataset to re- identify individuals. To preserve privacy during publishing the data many approaches [1] has been proposed E.G. k-anonymity, l-diversity-closeness.In conventional anonymization approaches [4-6] mainly assumes that microdata does not contain any null/missing values. They consider records which have missing values as outliers and they will be removed in pre- processing step. Some of the works [7, 8] will allow missing values to take part in anonymization and to be published with normal values.E.g as we prove both methodologies are not appropriate for generalization based algorithms, even the missing values are a minority. The first scheme suffers from record suppression; while second suffers from extensive information loss due to the missing value pollution.

Now let's consider a scenario where the hospital wants to release patient's medical records [5]. The dataset contains three attributes (i.e.,) age, zip code, disease.

Table 1 :

Null/missing values microdata and directly anonymization with Mondrian

Tuple Id	Age	Zip Code	Disease
1)ALEX	14	632002	Hypertension
2)BOB	18	632006	Cancer
3)LUCY	25	*	Birth defects
4)JANET	26	*	Bird flu
5)ALICE	*	632007	Bird flu
6)SIMON	37	632009	Heart disease
7)JOHN	*	632012	Hypertension
8)TOM	48	632014	Bird flu
BY MONDRIAN			
	*	*	Hypertension
	*	*	Cancer
	*	*	Birth defects
	*	*	Bird flu
	*	*	Bird flu
	*	*	Heart disease
	*	*	Hypertension
	*	*	Bird flu

The first two attributes are considered as a Qid. Disease attribute is considered as a sensitive attribute (Sa).While collecting data of some patients they do not want their personal information to be published or to be public. So we get the microdata as shown in table 1A [2].

Here we denote * as a missing value. In table 1A, we get 25%of values as missing values, up to 50%of records containing missing values. Removing the records with the missing values will lose 50%of records that are nearly we lost 50%of information. The automatically the solution will lose more information if there is the higher uncompleted rate [3].

If we allow the missing values to participate in anonymization there it will be suffered from another problem that is missing value pollution. While we are performing generalization based anonymization, the normal values will be polluted because of the missing value and it will cause vast information a loss E.g., if we apply the widely used Mondrian [4] algorithm to the table 1A, then the entire Qid nearly 66.67%information will be lost or polluted because of only 4 missing values which has been showed in table 1A and Fig 1.They have developed top-down anonymization technique named Enhanced Mondrian, based on Mondrian. Enhanced Mondrian can reduce missing value pollution. But they cannot further split partitions .when some sub- partitions do not satisfy anonymity. This is a major drawback which increases both missing value pollution and information loss. They have developed semi-partition to increase data utility .By balancing records in the sub-partitions, semi-partitions cannot be further split by the Enhanced Mondrian and Mondrian .Though Enhanced Mondrian preserve more information than Mondrian but top down anonymization has been implemented to protect both missing value records and normal value records which is not that much capable enough in producing the fine k-groups.

To address this problem, we analyze the influence of missing value pollution on generalization based anonymization algorithms. Then we develop bottom-up anonymization (k,Rm) anonymization that partitions the missing values records from the normal values records recursively and anonymize the missing value record set and normal value records set separately and publish both of them as single anonymized record set. The rest of the paper will be followed by Section 2 inspect again on related works. Section 3 introduces the concepts and analyzes missing value pollution. In section 4 we present bottom-up anonymization algorithm for uncompleted microdata, which achieves k-anonymity. Section 5 describes our experiments and demonstrates our algorithms. Section 6 concludes the paper with the future works [9].

RELATED WORKS

Preserving privacy in the publication of data is an important problem in the privacy literature. In 2002, Sweeny et al [3, 9] initially discovered that removing personal identities is deficient to guard privacy throughout data publication due to the existence of Qid. To handle this issue they planned k-anonymity which needs every record is indistinguishable with a minimum of k-1 alternative records on Qid.Machanavajjhala et al [10] discovered that k-anonymity is vulnerable to homogeneity attack and background knowledge attack. They framework L-diversity model with reserves diversity constraint to boost privacy protection. Li et al [11] discovered skewness attack and similarity attack on l-diversity and more projected t-closeness model with distribution constraint to preserve privacy.Cao et al. [12] found that t-closeness cannot sufficiently shield occasional Sa values, and framework β-likeness with sturdy (strong) constraints on relative confidence gain to attain anonymity.Terrovitis et al. [13] argued that competitor have partial knowledge regarding high dimensional transactions. They proposed k anonymous to stop opponent has knowledge of m items in particular transaction. Gong et al. [14] found out that multiple records of particular individual might result privacy breaches and large information loss. They proposed (k, l)-diversity to preserve privacy and utility with restricted assumptions. Intuitively, the initial microdata typically cannot satisfy the aforesaid privacy models, unless they are properly anonymized. According to [11, 15, 16], carry through optimal anonymization achieving best anonymization i.e. minimum data loss beneath existing models, is NP-hard. Therefore all existing anonymization approaches attempt to come through nearly-optimal obscurity with the approximation algorithms. In 2006, LeFevre et al. [4] planned a methodical top-down greedy approximation k-anonymization algorithm has been used and it is called Mondrian which was based upon local recoding. This uncomplicated formula has been widely used in different literatures [17, 18]. Xu et al. [5] planned two clustering-based algorithms that outperformed Mondrian on data loss by sacrificing potency.Ghinita et al. [6] mapped multi- dimensional microdata to one-dimension, and proposed two systematic microdata algorithms named Hilb and iDist to achieve k-anonymity. Ni et al. [19] framework a clustering-oriented methodology to stay nearest neighborhood structures of data points throughout anonymization.Guo et al. [20] developed a clustering-based anonymization approach to preserve the characteristics of knowledge streams. Bhuyans et al. [21] framework a privacy protective sub-feature choice approach based mostly on fuzzy possibilities. Recently, Wong et al.

[22] found that k-anonymity will be achieved by non-homogeneous generalization, and proposed a method named ring generalization to realize higher utility whereas providing identical privacy guarantee. Xue et al. [23] adapted ring

generalization for anonymizing high-dimensional data, and proposed a non-reciprocal recording anonymization theme for such information/data. Doka et al. [24] developed the optimal-utility k-anonymization drawback as a network flow, and proposed freeform generalization for higher utility. The problem of incomplete values or missing information arises frequently [25].Its standard that this sort of data encompasses a negative result on data processing [26].To deal with this issue, researchers have frame worked a serial of approaches to handle missing information. One of the foremost used approaches is not to include missing data or null data for instance. Another popular methodology in the place of missing values based on the assumption e.g. mean substitution regression imputation has been applied. But unwanted imputations can create us a major biases between real and assumption data. To the most effective of our data no other previous works uses missing values throughout anonymization. That is as a result anonymization is usually separated from usage of data. By using missing values throughout anonymization which could change the essence of raw dataset and misdirect date recipient. In the all existing works [4-6] and [14, 29] in this works they will remove the records which has missing values in the preprocessing step. As we shown in the section 1, due to this approach make more information loss. To ignore this problem.Nergiz et al. [7] for incomplete dataset they framework cell-based suppression. But this approach causes serious missing value pollution. Gong et al. [8] to avoid the missing value pollution they frame worked anatomize approach for incomplete datasets. The existing anonymization which is based on generalization this solution is not applicable. For this issue and for preserving more data utility in generalization we are going to analyze missing value pollution.

DEFINITIONS

Definition 1: K-Anonymity

If every Ec in S has at least k records then the microdata S is known as k-anonymity. Now, data anonymization is especially done by generalization, e.g. Mondrian [4] and Hilb [6]. Typically anonymization of data preferably done by generalization (i.e.)means replacing original values with generic values that was obtainable under their domain context where many number of records seems to be indistinguishable from each other in terms of Qid values.

Definition 2: Generalization

In a record group G={s 1, s2, s3, s4…sj} and generalization function is denoted as KGen. The result of generalization/ KGen [KG] is a record s* which covers all the record in G, such that $\forall s_j \in G$, $s_j[i] \subseteq s*[i]$.Generalization hierarchies conduct generalization. In the example of generalization hierarchies in Fig 3.In table 3[A],Alex age value is 14 and Bob age is 18 while generalizing to interval [11,30],it covers both Alex and Bob ages. So the competitor cannot re- identify the Alex record or even if they knew the age.

Proposed Approach (K,) Anonymity model

Precondition: Input record set with missing valueR_m.
Post condition: Anonymized record setR_m.

Partition records with missing value If Recordset cannot be divided then Include record set to global group else.

Pivot Attribute ← Choose_ Attribute (Recordset);

MVG ← {r ← recordset: t [pivot Attribute] = *, #,}; Return MVG.

NVG ← {r ← recordset: t [pivot Attribute] ≠ *, #,-}; Return NVG.

For each (NVG)

Bottom-Up anonymize (NVG). return NVG*.

For each MVG.

Bottom-Up anonymize (MVG). Return (MVG*)

Publish {MVG*} U {MVG*} End.

We consider input record set with missing value R_m as the precondition. In the post condition

we are considering anonymizing records with R_m . Now let us partition records with the missing value separately from normal value. If we cannot further split the Recordset then we can include the Recordset to global group. We can find missing value group either by choosing the particular pivot attribute or we can find the missing value group for the whole Recordset Now let us choose the any one pivot attribute in the record set and check how many missing value group has been found. While analyzing if the Recordset contains any null/missing values [*.#, /] in the Recordset we anonymize that record separately. After that we return to the Missing Value Group[MVG].We are using Bottom-Up anonymization for both Normal Value Group[NVG] and Missing Value Group and returning it has NVG*and MVG*.

EXPERIMENTAL APPROACH

As shown in fig 1 we are loading the dataset without performing any anonymization or missing values. But in fig 2 we are loading the dataset with missing values and we are performing anonymization by choosing which column should be anonymized after anonymizing process we are getting the results that all the records which has the missing value for the column which we have selected will be opened separately/viewed to us. We are not only performing anonymization by selecting the particular column we are also performing anonymization for the normal dataset also which has the missing values.

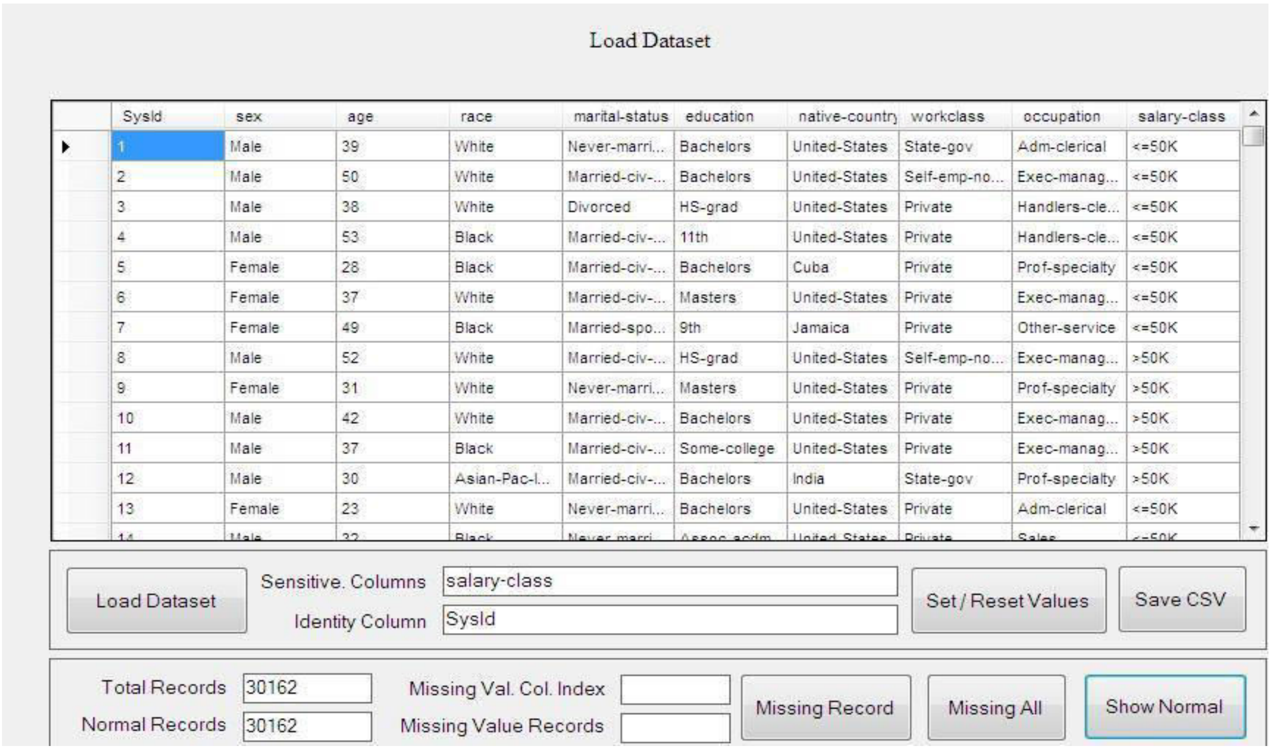

Fig 1. Loading the dataset without anonymization

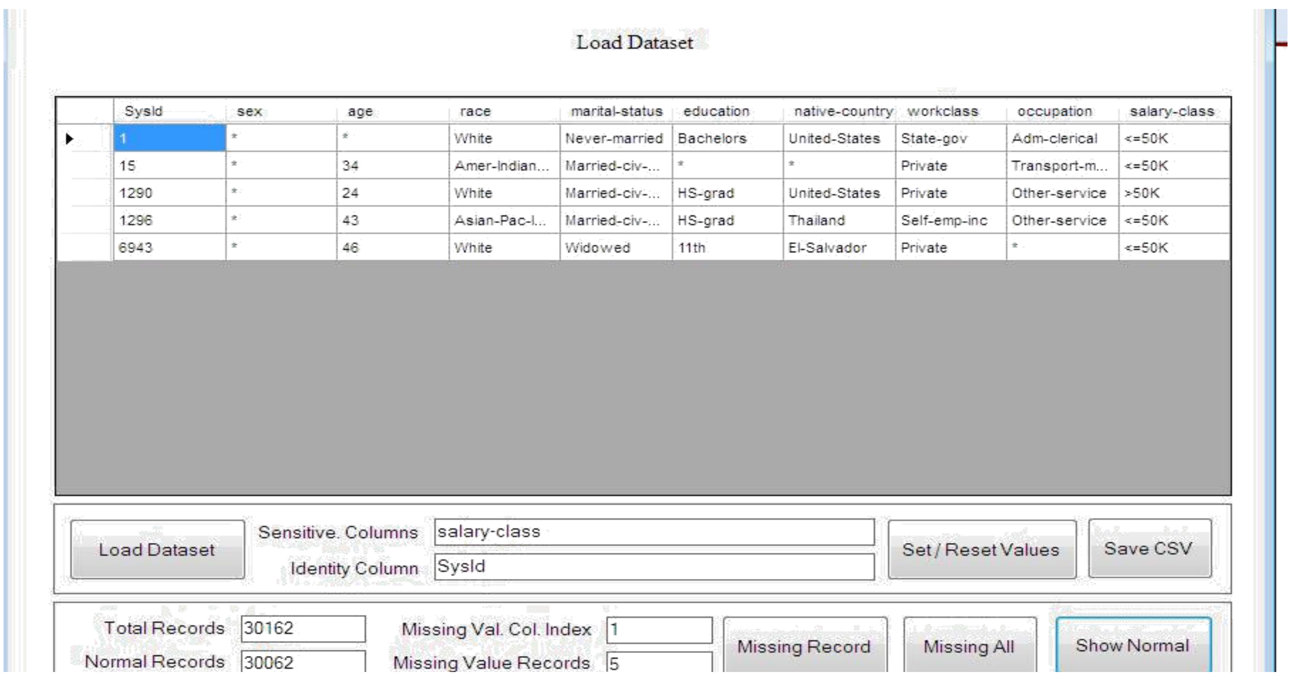

Fig 2 Dataset with missing values column wise

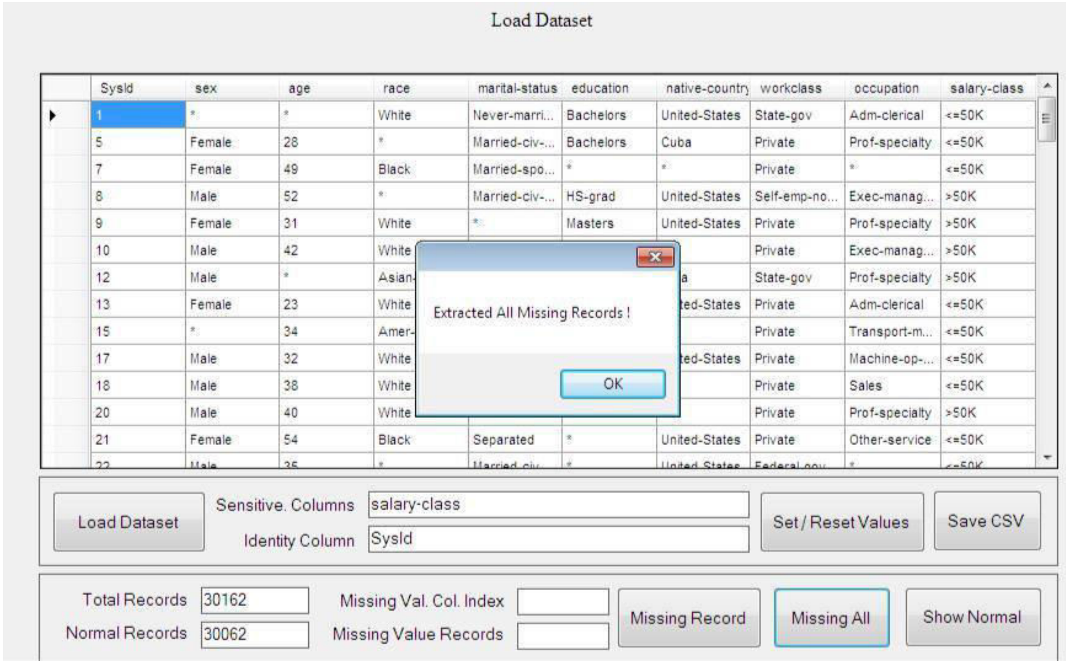

Fig 3 Finding dataset with missing value for normal dataset

Conclusion and future study

This paper presents a study on privacy preserving data for incomplete recordsets.To address this issue we are using bottom-up anonymization. We will be analyzing the properties of missing value pollution. We have done our experiments over real-world datasets The results obtained by executing our framework over real-world dataset proved that it is optimized in providing better trade-off between utility and privacy for missing value record sets than the existing procedures. This work also initiates several directions for future work.

REFERENCES

[1] Bhuyan, H.K., Kamila, N.K.: Privacy preserving sub-feature selection based on fuzzy probabilities. Cluster Comput.17 (4), 1383–1399 (2014)

[2] Cao, J., Karras, P.: Publishing microdata with a robust privacy guarantee. Proc. VLDB Endow.5 (11), 1388–1399 (2012).

[3] Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J.Uncertainty Fuzziness Knowl. Based Syst.10 (5), 557–570 (2002).

[4] Li, N., Li, T., Venkatasubramanian, S.: t-closeness: privacy beyond k-anonymity and l- diversity. In: IEEE 23rd International Conference on Data Engineering (ICDE), IEEE, pp. 106– 115(2007)

[5] He, Y., Naughton, J.F.: Anonymization of set-valued data via top-down, local generalization. Proc. VLDB Endow.2 (1), 934–945(2009)

[6] LeFevre, K., DeWitt, D.J., Ramakrishnan, and R.: Incognito: efficient full-domain k- anonymity. In SIGMOD'05: Proceedings of the2005 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM, New York (2005)

[7] Zakerzadeh, H., Aggarwal, C.C., Barker, K.: Privacy-preserving big data publishing. In: Proceedings of the

27th International Conference on Scientific and Statistical Database Management, ser.SSDBM'15, pp. 26:1–26:11. ACM, New York (2015)

[8] Bayardo, R.J., Agrawal, R.: Data Privacy Through Optimal k-Anonymization. IEEE Computer Society, Los Alamitos (2005)

[9] K. Bhargavi. An Effective Study on Data Science Approach to Cybercrime Underground Economy Data. Journal of Engineering, Computing and Architecture.2020;p.148.

[10]M. Kiran Kumar , S. Jessica Saritha. AN EFFICIENT APPROACH TO QUERY REFORMULATION IN WEB SEARCH, International Journal of Research in Engineering and Technology. 2015;p.172

[11] K BALAKRISHNA,M NAGA SESHUDU,A SANDEEP. Providing Privacy for Numeric Range SQL Queries Using Two-Cloud Architecture. International Journal of Scientific Research and Review. 2018;p.39

[12] K BALA KRISHNA, M NAGASESHUDU. An Effective Way of Processing Big Data by Using Hierarchically Distributed Data Matrix. International Journal of Research.2019;p.1628

[13] P.Padma, Vadapalli Gopi,. Detection of Cyber anomaly Using Fuzzy Neural networks. Journal of Engineering Sciences.2020;p.48.

[14] Kiran Kumar, M., Kranthi Kumar, S., Kalpana, E., Srikanth, D., & Saikumar, K. (2022). A Novel Implementation of Linux Based Android Platform for Client and Server. In A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems (pp. 151-170). Springer, Cham.

[15] Kumar, M. Kiran, and Pankaj Kawad Kar. "A Study on Privacy Preserving in Big Data Mining Using Fuzzy Logic Approach." Turkish Journal of Computer and Mathematics Education (TURCOMAT) 11.3 (2020): 2108-2116.

[16] M. Kiran Kumar and Dr. Pankaj Kawad Kar. "Implementation of Novel Association Rule Hiding Algorithm Using FLA with Privacy Preserving in Big Data Mining". Design Engineering (2023): 15852-15862

[17] K. APARNA, G. MURALI. ANNOTATING SEARCH RESULTS FROM WEB DATABASE USING IN-TEXT PREFIX/SUFFIX ANNOTATOR, International Journal of Research in Engineering and Technology. 2015;p.16.

[18] Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertainty Fuzziness Knowl.Based Syst.10 (5), 571–588 (2002).

[19] LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multi-dimensional k-anonymity. In: ICDE'06: Proceedings of the 22nd International Conference on Data Engineering, p. 25.IEEE Computer Society, Washington, DC (2006).

[20] Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst.10 (5), 557–570 (2002)

[21] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, and A.W.-C.: Utility-based anonymization using local recoding. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD'06, pp. 785–790. ACM, New York (2006)

[22] LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional k-anonymity. In: ICDE'06: Proceedings of the 22nd International Conference on Data Engineering, p. 25.IEEE Computer Society, Washington, DC (2006)

[23] Xue, M., Karras, P., Raïssi, C., Vaidya, J., Tan, K.-L.: Anonymizing set-valued data by nonreciprocal recoding. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD'12, pp. 1050–1058. ACM,New York (2012)

[24] Brown, M.L., Kros, J.F.: Data mining and the impact of missing data. Ind. Manag. Data Syst.103 (8), 611–621 (2003)

[25] Doka, K., Xue, M., Tsoumakos, D., Karras, P.: k-anonymization by freeform generalization. In: Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security,ser. ASIA CCS'15, pp. 519–530. ACM, New York (2015)

[26] A. A. S. Kumar, K. Ovsthus, and L. M. Kristensen, "An industrial perspective on wireless sensor networks - A survey of requirements, protocols, and challenges," IEEE Commun. Surveys Tuts., vol. 16, no. 3, pp. 1391–1412, 3rd Quart., 2014.