ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

# A Hybrid K-anonymity Data Relocation Technique for Privacy Preserved Data Mining in Cloud Computing☆

Yousra Abdul Alsahib S.Aldeen[1]　　Mazleena Salleh[2*]

ABSTRACT

The unprecedented power of cloud computing (CC) that enables free sharing of confidential data records for further analysis and mining has prompted various security threats. Thus, supreme cyberspace security and mitigation against adversaries attack during data mining became inevitable. So, privacy preserving data mining is emerged as a precise and efficient solution, where various algorithms are developed to anonymize the data to be mined. Despite the wide use of generalized K-anonymizing approach its protection and truthfulness potency remains limited to tiny output space with unacceptable utility loss. By combining L-diversity and (α,k)-anonymity, we proposed a hybrid K-anonymity data relocation algorithm to surmount such limitation. The data relocation being a tradeoff between trustfulness and utility acted as a control input parameter. The performance of each K-anonymity`s iteration is measured for data relocation. Data rows are changed into small groups of indistinguishable tuples to create anonymizations of finer granularity with assured privacy standard. Experimental results demonstrated considerable utility enhancement for relatively small number of group relocations.

☞ keyword : K-anonymity, privacy, L-diversity, data relocation, generalization

## 1. Introduction

In a global client foundation, the cloud service providers (CSP) offer efficient data storage and computing facilities [1]. Certainly, this is a creative blending of innovative internet services and notions that are useful for the future economic solutions [2], [3]. Stakeholders engaged in computing business chain can save their financial investment substantially in terms of hardware, data storage utilization, and computational power by availing the benefits of this novel electronic-trade model [4]. Lately, customers from every field are gaining the advantages of CC [5]. This widespread popularity of CC prompted majority of the well-known organizations to develop their IT systems on cloud. Moreover, the easy access of CC posed new privacy and security challenges. Users' data are often exposed to several threats, malicious attacks, and security breaches upon full access to cloud services [6].

The privacy protection of the sensitive financial and health records against fraud hands and phishing incursion remains challenging issue [7]. For instance, the medical data containing the sensitive information are highly useful to research community for disease analyses, drug development, and subsequent cure. Thus, anonymization mediated privacy preserving data mining emerged as a new solution because it not only assures the privacy requirements but also protects the data utility. Furthermore, it is mandatory to remove the link between the sensitive data and individual prior to publishing. Research revealed that a direct re-identification of dataset information is possible using partial identification and available information (quasi-identifiers) such as individual age, gender, zip-code, and data records [8]. Earlier, privacy metrics including adversary models are developed to prevent such identification [9], [10], [11].

Over the years, many algorithms are introduced to achieve the fundamental privacy standards by generalized manipulation, where the data values are replaced through general values. These general values characterize the original one by suggesting other atomic values such as rose changed flower, which is common

[1] Department of Computer Science, College of Education _Ibn Rushd, Baghdad University, Baghdad, Iraq.
2 Department of Computer Science, Universiti Teknologi Malaysia (UTM), Malaysia
[Received 9 May 2016, Reviewed 24 May 2016, Accepted 5 June 2016]
[*] Corresponding author (mazleena@fc.utm.my)
☆ A preliminary version of this paper was presented at ICONI 2015 and was selected as an outstanding paper.

http://dx.doi.org/10.7472/jksii.2016.17.5.51

in numerous proposed algorithms. Unlike perturbation technique, this generalization applies the noise data cells individually before publishing and thus preserves the data truthfulness. Moreover, the occurrence of information loss and over-generalization limits their privacy requirements. To overcome such shortcoming, various heuristics approaches are designed. Issues such as over-generalization of outliers in private datasets and increase of the number of groups are the other concerns [12].

This paper proposes new hybrid *K*-anonymity data relocation algorithm by combining the generalization technique with data relocation approach. The overall utility is enhanced at the cost of truthfulness. Through data relocation the number of groups is reduced and their tuples are populated that belonged to a small equality groups. This target is achieved by bounding the number of relocations and controlling the tradeoff between utility and truthfulness.

## 2. RELATED WORK

The phrase value of utility preservation in dataset is significant in anonymization-based privacy protection. The novel heuristic algorithms comprising of nearby tuples can produce the equality groups to attain efficient utilized generalizations. All conceivable one dimensional mappings over the subset operation with binary search over the lattice is used to find an optimal K-anonymous generalization for minimizing a utility cost metric [8]. Two algorithms regarding the top-K high utility pattern mining are analyzed [13], where several experiments are performed for the algorithms on real data sets. This approach is extended [14] using bottom-up pruning method, which searched for all optimal K-anonymous generalizations. A more flexible approach is proposed [15] by relaxing the constraint, where every value in the generalization domain is considered identical. Numerous methods including t-closeness, L-diversity, and δ-presence are introduced [16], [17], [11] to achieve the privacy.

Clustering techniques for heterogeneous generalizations are developed to provide *K*-anonymity [18], [19], [20]. Multidimensional space is partitioned to form L-diverse and *K*-anonymous groups of tuples [21], [22], [23], where the usage of space filling curves with decreased dimensionality of the database is proposed. A new approach for optimal sub-tree anonymization over big data is introduced. It combined the Bottom Up Generalization (BUG) with Top Down Specialization (TDS) [24] to develop a Map Reduce algorithm for achieving the scalability. Hybrid approach has improved the efficiency and scalability of sub-tree anonymization over the conventional methods. Another hybrid generalization is presented based on data relocation mechanism [25], [12]. A simple anonymization technique using sub-clustering is also introduced [26] to achieve maximum privacy with minimum execution time. An intelligent approach based on association mining [27] called Adaptive Utility-based Anonymization (AUA) is proposed. For performance evaluation, AUA model is tested on National Family Health Survey (NFHS-3) dataset. It is established that the data anonymization can be performed without compromising the quality of data mining results.

Most of the existing approaches are based on pure and orthogonal generalizations. Conversely, our proposed relocation approach depends on the most generalizations irrespective of the underlying algorithm. Being independent of standalone anonymization algorithms the present approach can be made efficient amid equality groups based on truthfulness cost. In addition, it can improve the utility by releasing more information from the equality groups.

## 3. THE PROPOSED ALGORITHM

The presence of relatively small output space and strict privacy requirements of the conventional techniques are responsible for huge data loss. Thus, utility preservation using generalization-based technique remains challenging. Factors including the over-generalization of outliers in private datasets and the presence of large number of groups continued to be the primary concerns. In this view, a hybrid *K*-anonymity data relocation technique is introduced to solve the negative impacts of outliers and overgeneralization. This section highlights the problem analysis, describes the datasets, and explains the benefits of hybrid K-anonymity data relocation technique in terms of performance.

### 3.1 PROBLEM ANALYSIS

The most common feature of all anonymization algorithms is the data handling through generalizations. The resultant dataset is comprised of original and other atomic values, in which ‘Rose’ is replaced by ‘Flower’. Besides, more tuples can present analogous meanings. Typically, the generalizations protect the data truthfulness using the noise and collect the data cells autonomously before publishing. However, generalizations often lead to the information loss unless inhibited. Thus, over-generalization must be avoided upon fulfilling the privacy requirements. The outliers in private datasets originate from over-generalization. Efficient and accurate techniques are thus needed to prevent the destruction of overall datasets. The negative impacts of outliers and over-generalization is resolved by introducing the hybrid *K*-anonymity data relocation. Being sequential, the hybrid K-anonymity can easily overcome the limitations of generalization and data relocation method. Furthermore, it can improve the generalization method without being stuck in local optimal solution and permits the relocation of groups to enhance the overall utility at the cost of truthfulness. The main idea behind data relocation is to merge the complementary groups with lower K values. This reduces the number of groups to further populate small equality groups of tuples. It is customary to provide the following definitions to clarify hybrid K-anonymity.

1. ***QI* (Quasi-identifier):** Given a table *U* denoted as $T\ (A_1 \cdots A_n)$, $f_c{:}U \rightarrow T$, $f_g{:}T \rightarrow U'$, where $U \subseteq U'$, a quasi-identifier of $T\ (Q_T)$ is a set of attributes $\{A_i \cdots A_j\} \subseteq \{\ A_1 \cdots A_n\}$ where $\exists p_i \in U$ *such that* $f_g\ (f_c\ (p_i)[\ QT]) = p_i$ [28].
2. **Generalization:** A *generalization*, written $\{c\} \rightarrow p$ replaces all child values $\{c\}$ with the parent value $\{p\}$. This is *valid* if all values that are lower than c are generalized to c. *A* vid is *generalized* by $\{c\} \rightarrow p$ if the vid contains some value in $\{c\}$ [27].
3. ***K*-anonymity:** A table *T* satisfies the *K*-anonymity if for every tuple $t \in T$ there exists (*k-1*) other tuples $^{t}{}_{i1}\ ^{t}{}_{i2} \cdots\ ^{t}{}_{k-1} \in T$ such that $^{t}{}_{i1}[C] = {}^{t}{}_{i2}[C] = \cdots\ ^{t}{}_{ik-1}[C]$ for all $C \in QT$[16].
4. Equality group: The equality group of tuple *t* in dataset $T^{\ast}$is the set of its all tuples with identical quasi-identifiers to t [29].

## 3.2 DATASETS DESCRIPTION

Preset study considers direct marketing (bank) dataset collected from the UCI (University of California at Irvine) Machine Learning Repository [30] [31]. It is associated with different marketing campaigns through phone calls of a Portuguese banking institution. Regularly, more than one contact is required with one client to complete and analyze the product (e.g. term deposit). This dataset is categorized into two types including Bank-full.csv and Bank.csv. The first dataset used all examples and date-wise (May 2008 to November 2010) arranged. Conversely, the second dataset used only 10% from examples with random selection from bank-full.csv. The bank direct marketing dataset includes three hundred samples with seventeen attributes without any missing values [32]. The numeral attribute includes age, day, balance, duration, campaigning, present, and previous days. The categorical type encloses job, marital, education, contact, month, and outcome. The binary type comprises of yes or no and their classes such as loan, housing, default and output.

## 3.3 HYBRID *K*-ANONYMITY DATA RELOCATION TECHNIQUE

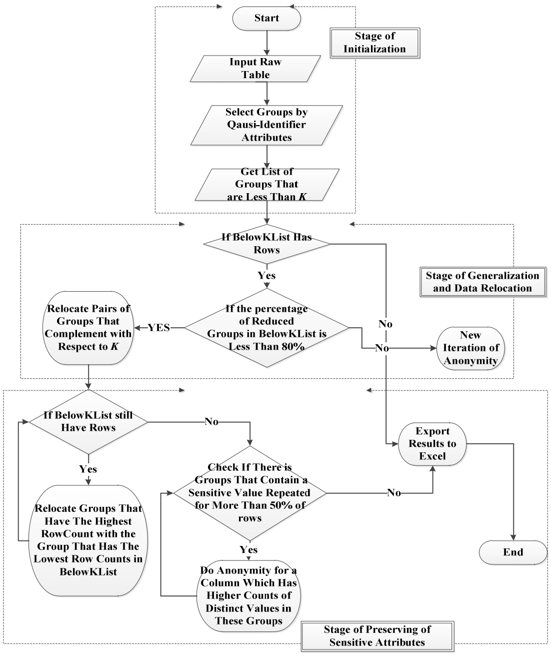
The hybrid *K*-anonymity data relocation technique measures the performance of each *K*-anonymity‘s iteration and decides whether data relocation has to be conducted or not. This technique is accomplished in three stages including initialization, generalization, and preservation as follows.

**Initialization:** After inputting the dataset (row table), the quasi-identifier groups are selected. Then, these groups are checked to get the list lower than *K*.

**Generalization and Data Relocation:** A *K*-anonymity operation is executed and the reduction percentage of BelowKList is logged. BelowKList means the groups of quasi-identifier are selected to get the list lower than *K*. This percentage being a good indicator determined the capacity of one step generalization or suppression to performing K-anonymity. If the reduction is less than 80% in BelowKList then the data relocation is required. Figure 1 depicts the performance flowchart when the data relocation is called. First, a list of groups is created by complementing each other with respect *K*. The data relocation is performed for uniting each pair of that group under one K group. Next, the remaining groups are united by data relocation according to their RowCount. In short, the groups with lowest row count are united with the groups with the highest one. This operation is executed iteratively until the BelowKList is empty.

**Stage of Preserving of Sensitive Attributes:** The BelowKList is checked and the groups with lowest row count are united with the highest one if the BelowKList is non-empty. Otherwise, a data processing step is performed to maintain *L*-diversity, where the sensitive attribute is checked with respect to their frequency values in one group. If this frequency is more than 50% then the new K-anonymity iteration is conducted with (K+1) values until the sensitive value above this frequency is exhausted.

The data relocation (not shown) is performed after checking the percentage (more than 10%) of data modification (Table 1). This choice of 10% is to avoid the over-relocation that harms truthfulness. This also limits the number of relocations that the algorithm can apply, thus dominates the trade-off between truthfulness and utility. A roll back called new *K*-anonymity iteration is performed.


Start
Stage of Initialization
Input Raw Table
Select Groups by Qausi-Identifier Attributes
Get List of Groups That are Less Than K
If BelowKList Has Rows
Yes
Stage of Generalization and Data Relocation
Relocate Pairs of Groups That Complement with Respect to K
YES
If the percentage of Reduced Groups in BelowKList is Less Than 80%
No
No
New Iteration of Anonymity
If BelowKList still Have Rows
No
Export Results to Excel
Yes
Check If There is Groups That Contain a Sensitive Value Repeated for More Than 50% of rows
No
Relocate Groups That Have The Highest RowCount with the Group That Has The Lowest Row Counts in BelowKList
Yes
End
Do Anonymity for a Column Which Has Higher Counts of Distinct Values in These Groups
Stage of Preserving of Sensitive Attributes

(Figure 1) Flowchart of hybrid K-anonymity data relocation

# 4. RESULTS

Table 1 summarizes the BelowKList groups. The algorithm is obtained by creating a list of groups that complement each other respecting *K* value. The complement groups are labeled with A, B, and C as enlisted in Table 2. The data relocation is performed to combine each pair of that group under one k group as enlisted in Tables 3 and 4. Next, the remaining groups are united by data relocation according to their RowCount as summarized in Tables 5 to 7. The groups with lowest row count are re-integrated with that of the highest one. This operation is executed iteratively until the BelowKList is empty. Table 8 depicts the BelowKList new groups upon completing the relocation.

(Table 1) BelowKList

| Job | Martial | Education | Group Count |
|---|---|---|---|
| Unemployed | Divorced | Unknown | 1 |
| Student | Divorced | Primary | 1 |
| Student | Divorced | Secondary | 1 |
| Self-Employed | Divorced | Unknown | 1 |
| Student | Divorced | Unknown | 1 |
| Student | Married | Primary | 2 |
| Student | Divorced | Tertiary | 3 |
| Unknown | Divorced | Tertiary | 3 |
| Unknown | Divorced | Secondary | 3 |
| Unknown | Divorced | Primary | 4 |
| Blue-Collar | Divorced | Tertiary | 4 |
| Retired | Single | Unknown | 5 |

(Table 2) Complementary groups are labeled as A, B, and C

| Job | Martial | Education | Group Count |
|---|---|---|---|
| Unemployed | Divorced | Unknown | 1A |
| Student | Divorced | Primary | 1 |
| Student | Divorced | Secondary | 1 |
| Self-Employed | Divorced | Unknown | 1 |
| Student | Divorced | Unknown | 1 |
| Student | Married | Primary | 2B |
| Student | Divorced | Tertiary | 3C |
| Unknown | Divorced | Tertiary | 3C |
| Unknown | Divorced | Secondary | 3 |
| Unknown | Divorced | Primary | 4B |
| Blue-Collar | Divorced | Tertiary | 4 |
| Retired | Single | Unknown | 5A |

(Table 3) BelowKList after first step of data relocation

| Job | Martial | Education | Group Count |
|---|---|---|---|
| Retired | Single | Unknown | 1A |
| Student | Divorced | Primary | 1 |
| Student | Divorced | Secondary | 1 |
| Self-Employed | Divorced | Unknown | 1 |
| Student | Divorced | Unknown | 1 |
| Unknown | Divorced | Primary | 2B |
| Unknown | Divorced | Tertiary | 3C |
| Unknown | Divorced | Tertiary | 3C |
| Unknown | Divorced | Secondary | 3 |
| Unknown | Divorced | Primary | 4B |
| Blue-Collar | Divorced | Tertiary | 4 |
| Retired | Single | Unknown | 5A |

(Table 4) BelowKList new groups after first step of data relocation

| Job | Martial | Education | Group Count |
|---|---|---|---|
| Student | Divorced | Primary | 1A |
| Student | Divorced | Secondary | 1A |
| Self-Employed | Divorced | Unknown | 1B |
| Student | Divorced | Unknown | 1B |
| Unknown | Divorced | Secondary | 3B |
| Blue-Collar | Divorced | Tertiary | 4A |
| Unknown | Divorced | Primary | 6 |
| Retired | Single | Unknown | 6 |

(Table 5) BelowKList after second step of data relocation

| Job | Martial | Education | Group Count |
|---|---|---|---|
| Blue-Collar | Divorced | Tertiary | 1A |
| Blue-Collar | Divorced | Tertiary | 1A |
| Self-Unknown | Divorced | Secondary | 1B |
| Unknown | Divorced | Secondary | 1B |
| Unknown | Divorced | Secondary | 3B |
| Blue-Collar | Divorced | Tertiary | 4A |
| Unknown | Divorced | Primary | 6 |
| Retired | Single | Unknown | 6 |

(Table 6) BelowKList new groups after second step of data relocation

| Job | Martial | Education | Group Count |
|---|---|---|---|
| Unknown | Divorced | Secondary | 5A |
| Blue-Collar | Divorced | Tertiary | 6A |
| Unknown | Divorced | Primary | 6 |
| Retired | Single | Unknown | 6 |

(Table 7) BelowKList after third step of data relocation

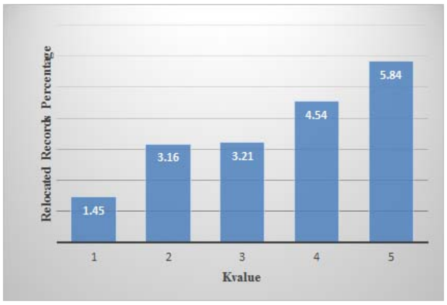
| Job | Martial | Education | Group Count |
|---|---|---|---|
| Blue-Collar | Divorced | Tertiary | 5A |
| Blue-Collar | Divorced | Tertiary | 6A |
| Unknown | Divorced | Primary | 6 |
| Retired | Single | Unknown | 6 |

(Table 8) BelowKList new groups after third step of data relocation

| Job | Martial | Education | Group Count |
|---|---|---|---|
| Unknown | Divorced | Primary | 6 |
| Retired | Single | Unknown | 6 |
| Blue-Collar | Divorced | Tertiary | 11 |

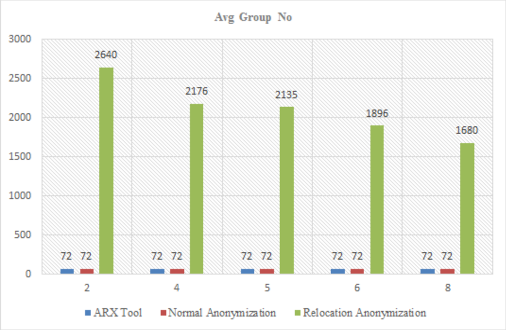
## 5. TRUTHFULNESS AND UTILITY EVALUATION

Figure 2 displays the *K* value dependent relocation percentage, which is a measure of the truthfulness of the proposed hybrid *K*-anonymity. The percentage of data relocation is found to saturate to maximum of 6% regardless of the values of K, indicating the data truthfulness of over 94%.



(Figure 2) Relocation percentage as measure of truthfulness

The average number of groups that share the same values of quasi-identifier is calculated as a measure of utility (Figure 3). The average group size is found to decrease significantly with the implementation of hybrid *K*-anonymity. This is equivalent to an increase of the utility of the resulted anonymized data.



(Figure 3) Average group size as an indicator to utility

## 6. CONCLUSION

This paper emphasized the significance of anonymization based privacy protection approaches for ensuring the non-linkage of published data back to an individual. The *K*-anonymity method is demonstrated to be most appropriate for applying generalizations in private data to maintain the privacy standard. The limitations associated to existing generalization-based approaches such as improper utility loss for stringent privacy requirements for relatively small output space is overcome. A hybrid *K*-anonymity data relocation algorithm is developed and the performance of each K-anonymity's iteration is determined to decide the conductivity of data relocation. The approach of data relocation is discerned to modify certain data rows into small groups of indistinguishable tuples. Furthermore, the proposed approach allowed the anonymizations with fine granularity and thereby validated the privacy standards. This data relocation scheme has established a trade-off between truthfulness and utility. An input parameter to control this tradeoff together with privacy metrics such as L-diversity, and (α-k)-anonymity is provided. Experimental results revealed that the developed approach achieved relatively small number of relocations of groups' with enhanced utility.

## Reference

[1] X. Dong, J. Yu, Y. Luo, Y. Chen, G. Xue, and M. Li, "Achieving an effective, scalable and privacy-preserving data sharing service in cloud computing," *computers & security*, pp. 151-164, 2014. http://doi.org/10.1016/j.cose.2013.12.002

[2] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and L. Brandic, "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," *Future Generation computer systems*, vol. 25, pp. 599-616, 2009. http://doi.org/10.1016/j.future.2008.12.001

[3] W. Cohen and D. Levinthal, "Absorptive capacity: a new perspective on learning and innovation," *Administrative science quarterly*, pp. 128–152, 1990. http://doi.org/10.2307/2393553

[4] L. Wang, J. Zhan, W. Shi, and Y. Liang, "In cloud, can scientific communities benefit from the economies of scale?" *Parallel and Distributed Systems, IEEE Transactions on.* 23, no. 2, pp. 296-303, 2012. http://doi.org/10.1109/TPDS.2011.144

[5] X. Yang, L. Wang, and G. Laszewski, "Recent Research Advances in e-Science," *Cluster Computing,* 2009, vol. 12, no. 4, pp. 353–356. http://doi.org/10.1007/s10586-009-0104-0

[6] G. Ateniese, R. Di Pietro, L. V. Mancini, and G. Tsudik, "Scalable and efficient provable data possession," *Proceedings of the 4th international conference on Security and privacy in communication netowrks.* ACM, 2008. http://doi.org/10.1145/1460877.1460889

[7] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Generation computer systems*, vol. 28, no. 3, pp. 583–592, 2012. http://doi.org/10.1016/j.future.2010.12.006

[8] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001. http://doi.org/10.1109/69.971193

[9] R. C. Wong, J. Li, A. W. Fu, and K. Wang, "(α,k)-Anonymity : An Enhanced k -Anonymity Model for Privacy-Preserving Data Publishing," *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2006. http://doi.org/10.1145/1150402.1150499

[10] S. Kumara, S. Singhb, A. Singhc, and J. Alid,

"Virtualization, The Great Thing and Issues in Cloud Computing," *International journal of Current Engineering and Technology*, pp. 338–341, 2013. http://inpressco.com/wp-content/uploads/2013/03/Paper18338-341.pdf

[11] M. E. Nergiz and C. Clifton, "δ-presence without complete world knowledge," *IEEE Transactions on Knowledge and Data Engineering*, 2010, vol. 22, no. 6, pp. 868–883. http://doi.org/10.1109/TKDE.2009.125

[12] M. E. Nergiz, M. Z. Gök, and U. Özkanli, "Preservation of utility through hybrid k-anonymization," *Trust, Privacy, and Security in Digital Business.* Springer Berlin Heidelberg, pp. 97–111, 2013. http://doi.org/10.1007/978-3-642-40343-9_9

[13] C. Kim, "Performance Analysis of Top-K High Utility Pattern Mining Methods," *JICS*, vol. 16, no. 15, pp. 89–95, 2015. http://dx.doi.org/10.7472/jksii.2015.16.6.89

[14] K. Lefevre, "Incognito : Efficient Full-Domain K-Anonymity," *Proceedings of the 2005 ACM SIGMOD international conference on Management of data.* ACM, 2005. http://doi.acm.org/10.1145/1066157.1066164

[15] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on.* IEEE, 2005. http://doi.org/10.1109/ICDE.2005.42

[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity," *ACM Transactions on Knowledge Discovery from Data*, vol. 1, no. 1, p. 3–es, 2007. http://doi.org/10.1145/1217299.1217302

[17] M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals from shared databases," *Proceedings of the 2007 ACM SIGMOD international conference on Management of data.* ACM, 2007. http://doi.org/10.1145/1247480.1247554

[18] M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," *Data & Knowledge Engineering, 2007,* vol. 63, no. 3, pp. 622–645. http://doi.org/10.1016/j.datak.2007.03.009

[19] G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas, K. Kenthapadi, S. Khuller, and A. Zhu, "Achieving anonymity via clustering," *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.* ACM, 2006. http://doi.org/10.1145/1798596.1798602

[20] J. L. Lin, M. C. Wei, C. W. Li, and K. C. Hsieh, "A hybrid method for k-anonymization," *Asia-Pacific Services Computing Conference, 2008. APSCC'08. IEEE.* IEEE, 2008. http://doi.org/10.1109/APSCC.2008.65

[21] K. Lefevre and D. J. Dewitt, "Mondrian Multidimensional K-Anonymity," *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on.* IEEE, 2006. http://doi.ieeecomputersociety.org/10.1109/ICDE.2006.101

[22] B. Hore, R. C. Jammalamadaka, and S. Mehrotra, "Flexible Anonymization For Privacy Preserving Data Publishing : A Systematic Search Based Approach," SDM, 2007. http://dx.doi.org/10.1137/1.9781611972771.51

[23] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," *Proceedings of the 33rd international conference on Very large data bases.* VLDB Endowment, 2007. Retrieved from http://dl.acm.org/citation.cfm?id=1325938\nhttp://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.138.3217

[24] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, and J. Chen, "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud," *Journal of Computer and System Sciences*, vol. 80, no. 5, pp. 1008–1020, 2014. http://doi.org/10.1016/j.jcss.2014.02.007

[25] M. E. Nergiz and M. Z. Gök, "Hybrid k-Anonymity," *Computers & Security*, vol. 44, pp. 51–63, 2014. http://doi.org/10.1016/j.cose.2014.03.006

[26] J. J. Panackal and A. S. Pillai, "Adaptive Utility-based Anonymization Model: Performance Evaluation on Big Data Sets," *Procedia Computer Science*, vol. 50, pp. 347–352, 2015. http://doi.org/10.1016/j.procs.2015.04.037

[27] E. T. Wang and G. Lee, "An efficient sanitization algorithm for balancing information privacy and

knowledge discovery in association patterns mining," *Data & Knowledge Engineering,* Jun., vol. 65, no. 3, pp. 463 – 484, 2008.. http://doi.org/10.1016/j.datak.2007.12.005

[28] Y. Pan, X. L. Zhu, and T. G. Chen, "Research on privacy preserving on K-anonymity," *Journal of Software*, vol. 7, no. 7, pp. 1649 – 1656, 2012. http://doi.org/10.4304/jsw.7.7.1649-1656

[29] M. E. Nergiz, M. Z. Gök, and U. Özkanli, "Preservation of utility through hybrid k-anonymization," *in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8058 LNCS, pp. 97 – 111. http://doi.org/10.1007/978-3-642-40343-9_9

[30] S. Moro and R. M. S. Laureano, "Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology," *European Simulation and Modelling Conference*, 2011. Retrieved from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[31] H. A. Elsalamony, "Bank Direct Marketing Analysis of Data Mining Techniques," *International Journal of Computer Applications,* 2014, pp. 12 – 22. http://www.ijcaonline.org/archives/volume85/number7/14852-3218

[32] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, 2014, vol. 62, pp. 22 – 31. http://doi.org/10.1016/j.dss.2014.03.001

## ◐ Authors ◑

### Yousra Abdul Alsahib S.Aldeen

1995- Present (20 years 8 months) lecture Department of Computer Science, College of Education _Ibn Rushd, Baghdad University, Baghdad, Iraq.

1994 She has got BSc in Computer Science (Al-Mustansiriyah University). 2006 She has got MSc in Computer Network (Iraqi commission for computers and informatics). 2013- Present PhD student at Faculty of Computing, University Technology Malaysia, UTM, 81310 UTM Skudai, Johor, Malaysia.

### Mazleena Salleh

is an associate professor at Universiti Teknologi Malaysia (UTM), lecturing under the Department of Computer Science, Faculty of Computing. She received her PhD in Computer Science at UTM in the field of computer networking while her Master's degree from Virginia Polytechnic State University (US) in the field of electrical engineering. She has published several journal and conference papers related to her research works that include watermarking, steganography, chaos image encryption, network analysis, e-learning and knowledge management. Her current research is on computer security related issues namely data survivability and availability in cloud, privacy preserving in cloud environment, elliptic curve cryptography, and detection of misuse in computer forensic.