A New Heuristic Anonymization Technique for Privacy Preserved Datasets Publication on Cloud Computing

S. Aldeen Yousra¹,², Salleh Mazleena²

¹Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq

²Faculty of Computing, University Technology Malaysia, UTM, Skudai, Johor, Malaysia

yohrmz_8@yahoo.com¹,²; mazleena@fc.utm.my²

Abstract. Recent advancement in Information and Communication Technologies (ICT) demanded much of cloud services to sharing users' private data. Data from various organizations are the vital information source for analysis and research. Generally, this sensitive or private data information involves medical, census, voter registration, social network, and customer services. Primary concern of cloud service providers in data publishing is to hide the sensitive information of individuals. One of the cloud services that fulfill the confidentiality concerns is Privacy Preserving Data Mining (PPDM). The PPDM service in Cloud Computing (CC) enables data publishing with minimized distortion and absolute privacy. In this method, datasets are anonymized via generalization to accomplish the privacy requirements. However, the well-known privacy preserving data mining technique called K-anonymity suffers from several limitations. To surmount those shortcomings, I propose a new heuristic anonymization framework for preserving the privacy of sensitive datasets when publishing on cloud. The advantages of K-anonymity, L-diversity and (α, k)-anonymity methods for efficient information utilization and privacy protection are emphasized. Experimental results revealed the superiority and outperformance of the developed technique than K-anonymity, L-diversity, and (α, k)-anonymity measure.

1. Introduction

The innovative notion so called cloud computing (CC) with prolific computing resources allows the users to store the data remotely into the cloud for diverse applications [1]. Numerous third-party cloud computing services also offer data support, computing management and internet services including Microsoft Azure, Amazon’s EC2, etc. They are not only economic but have great bounciness to rendering progressively the service providers in outsourcing their data to the cloud platform. Organizations recognize opportunities as well as critical values of sharing prosperity of data information through numerous dispersed databases. The sensitive and private data collected from various governmental and non-governmental organizations are rapidly increasing and stored in the electronic repository [2]. In recent times, varieties of data mining techniques on CC are implemented to support the decision-making process. These techniques are used to quotation the hidden information from enormous datasets in form of new models, trends, and diverse
patterns. However, due to privacy reason the personal data of any individual need to be protected during data mining. Privacy in CC implies the individual’s personal information (called sensitive data) protection while publishing [2];[3];[4]. Investigates exposed that attackers often detect and goal the information efficiently from third party clouds [5]. Individuals sensitive data information requires supreme privacy protection before being outsourced to the cloud [6]. This issue is widely addressed via anonymization in the form of K-anonymity [7];[8], (a, k)-anonymity [9], L-diversity [10], t-closeness [11], m-invariance [12], etc. The original information that is released to the cloud by data providers necessitates different privacy requirements. Despite many dedicated efforts, an efficient and accurate privacy preserving anonymization technique is far from being achieved [1]. For the first time, I report the design and development of a heuristic anonymization technique for privacy preserving dataset publishing on cloud. The methods for L-diversity, K-anonymity and (a, k)-anonymity are meticulously combined to avoid the adversary from aggressive the dataset privacy. The proposed method overcomes the limitations of heuristic K-anonymity algorithm to efficiently protecting the sensitive datasets privacy from linking attack. This is achieved by integrating the heuristic L-diversity with heuristic (a, k)-anonymity method. A heuristic anonymization technique provides robust privacy to datasets publishing on CC. In achieving so, two major contributions are made:

1. A systematic experimentation on anonymization is introduced. The benefits of K-anonymity, L-diversity and (a, k)-anonymity -anonymity are amalgamated to ensure strong privacy by balancing simultaneously the privacy protection and information utilization.
2. The proposed privacy framework and anonymization technique outperformed the distinct L-diversity measure, and K-anonymity with (a, k)-anonymity measure.

This article is composed of seven sections. Section 2 overviews the literature on related work especially the anonymization method. The discussion about the relationship between cloud computing and preserving of privacy of published data (Section 3). Then, the technique and flowchart of this work are explained in Section 4. Section 5 is explained the experimental results and evaluation Section 6 concludes the paper. 'Figure(1)’ illustrates the conceptual framework of this article.

![Conceptualization of privacy preserving data publication.](image)

Figure 1. Conceptualization of privacy preserving data publication.

2. Related Work
Currently, the privacy preserved data publishing are conducted broadly in integrated databases [13]. It is important to develop privacy values, such as K-anonymity, (a, k)-anonymity, L-diversity, t-closeness and m-invariance for greatly secured data sharing and mining on cloud. All data supplies undertook this matter by suggesting a third party trusted
approach. They upload the data to confidential third party for anonymization and data integration in which the clients inquire on centralized database [14]. Jiang et al. (2006) suggested a two-party application based model for generating K-anonymous data, which vertically partitioned the sources without revealing the data from site to site [15]. Zhong et al. (2005) presented a solution for K-anonymous generalization for distributed setup by maintaining the end-to-end privacy of original data together with final K-anonymous results [16]. Jurczyk et al. (2009) developed a distributed anonymization algorithm for achieving the anonymity of data providers and subjects [14]. Ciriani et al. (2007) clear the likely threats to K-anonymity that arise from executing the mining on a group of data, where two main methods are considered to syndicate K-anonymity in data mining. Numerous approaches are developed to distinguish K-anonymity violations and to remove them in suggestion rule and cataloging mining. Despite extensive use of K-anonymous privacy preservation it suffers from various challenges that requires special attention and further improvements [17]. Patil et al., 2013 analyzed the current K-anonymity model and its applications. They addressed some of the multidimensional K-anonymous approach that is accustomed attain the elementary knowledge. Nowadays, majority of the K-anonymity algorithms hang on multidimensional datasets, where nearest neighborhood strategy is incorporated to recover the superiority of anonymity and to decrease the information loss [18]. Generally, the addition of all local anonymized datasets is performed through the secure algorithm. The problem of secured outsourcing of frequent item-set mining on the multi-cloud environments is scrutinized [19]. To provide protection against a knowledgeable attacker with exact support information a K-support anonymity is proposed [20], [1] suggested a distributed anonymization protocol for privacy-preserving data publishing from numerous data suppliers on cloud. Recently, the scalability issue of sub-tree anonymization over big data on cloud is reported by Zhang et al. (2013) and Zhang et al. (2014). They suggested a hybrid method which combined the Top-Down Specialization (TDS) and Bottom-Up Generalization (BUG). Both TDS and BUG are talent in an extremely scalable manner via a sequence of deliberately designed Map Reduce jobs [21],[22]. Zhang et al. (2014) also proposed a highly scalable two-phase TDS approach using Map Reduce on cloud, where datasets are first divided and anonymized in parallel to create intermediate results [23]. Literature clearly hinted that the existing privacy preserving techniques cannot efficiently protect the privacy from combined attacks including homogeneity, similarity and background knowledge. However, my proposed anonymization technique ensures strong data privacy during publishing in CC. Conversely, it satisfies the privacy need of data providers and individual within the budget. It is significant for anonymous data that already exists in servers as individual databases, where all local anonymized datasets are inserted through the secure algorithm.

3. Privacy Preservation of Published Data

In recent times, numerous data mining techniques are suggested and rummage-sale to assist the decision-making process. These methods mined the concealed information from vast data in the form of new models, trends and different patterns. Privacy in data mining during publication indicates the defense of private information [4]. Repeated studies indicated that attackers often detect and aim the information from third party clouds [5]. Figure 2 demonstrates the three domains of privacy preservation of the stored data, which contain the user domain, cloud domain, and recipient domain. In public practice, any user share the stored data records to the cloud service provider. Service providers further publish these
datasets to any research centre (viz. medical research). For example, hospital assists as the data owner and the medical research centre acts as data recipient. The cloud domain enables the hardware and software infrastructures to the service provider for providing the shared medical records as outsourced storage [24].

![Figure 2. Various domains of the stored data privacy preservation.](image)

Before outsourcing the data to cloud for publication, the data privacy (sensitive information of individuals) must be protected or well-preserved. This is determined using data anonymization methods such as K-anonymity, (α, k)-anonymity and L-diversity [25], [26]. The data providers implement various privacy requirements while releasing their original data to the cloud, which often pose a challenge. Numerous privacy preservation methods are accessible for data mining such as K-anonymity, distributed privacy preservation, L-diversity, randomization, taxonomy tree, condensation and cryptographic techniques [27],[28],[29]. To resolution the cloud data mining related privacy preservation problem, this paper proposed a practical solution to the data provider.

4. Heuristic anonymization Technique

The aim heuristic anonymizing techniques for privacy preserving data mining in cloud is to ensure strong confidentiality for data to be published in cloud. The privacy preserving techniques for data publishing are categorized into theoretical and heuristic types. This research proposed a heuristic technique [30]. Although numerous heuristic techniques are introduced to guarantee strong privacy preservation but a unified algorithm is far from being developed. Thus, heuristic K-anonymity, heuristic L-diversity and (α, k) - anonymity techniques are combined to achieve this target. Figure 3 depicts an overview of privacy preserving techniques for published data in cloud environment. It composed of two parts. The first part explains the heuristic K-anonymity technique and its result. The second part explains heuristic L-diversity and (α, k) - anonymity technique to overcome the limitations of the heuristic K-anonymity technique. It is customary to describe datasets and the attack model before introducing the concept of heuristic K-anonymity technique.

![Figure 3. Overview of privacy preserving techniques for published data](image)
I use the bank direct marketing dataset, which is collected from different web sources such as the UCI (University of California at Irvine) Machine Learning Repository. This dataset is implemented and the privacy related performances of proposed heuristic anonymization approach is evaluated. [31],[32], [33], [34] collected and arranged this dataset, which is utilized by [35].

4.1 Heuristic K-anonymity Technique

Generally, the procedure starts from promoting an expert in the data to determine the QI, where QI, diversion method, and K values are determined by the user. The user is able to determine QI, K value and diversion method. K value denotes the length of the K group in the resulting K-anonymized data as shown in Figure 4. In this research, K values are changed to examine its effect on other metric. For the first experiment, the value of K is taken as 20. Every group possessed many tuples and the range is increased to 25. This produced a strong privacy at the cost of enhanced information loss. Therefore, it is essential to reduce the K value. In the second experiment, the K value is chosen to be 10. The information loss is still increased. Then, the value of K is changed to 5 and this time the information loss is reduced with lower range. Thus, the trade-off between the privacy and utility of data is achieved. Next, the grouping is performed. The criterion of stopping the technique is based on confirming the condition of reaching a minimum group of set of unique QI with a size equals or more than the K value. Figure 4 demonstrates three diversion methods of performing the K-anonymity including Incremental Range, Incremental Suppression, and Unification.

![Figure 4. Flowchart of heuristic K-anonymity technique](image-url)
In the Incremental Range method, the user must select the suitable mode with corresponding attributes. For example, when the attribute is a range of numbers (set of numbers) like Age, Date, Time… etc., the user might select incremental range. Consequently, the user provides the value of increment to combine records together with a sub-range in the Age attribute. For instance, if a user provided 5 as range incremental, then the technique start to generalizing by combining the sub-range 20-24 into one group, then 25-29 into another group and so on. The Incremental Suppression method is performed by incrementally replacing the characters. Thus, different values are included in one group once they are matched each other. This diversion can be used with ZipCode. The third method is the unification, which simply combines more than one of the distinct values of the attributes. This is similar to combine different cities in the attribute address under the name of the state that includes them. Variable X in the flowchart indicates the index of the number of diversion that is performed on the attribute. In Figure 4, the QI, X (the index of the number of diversion method) and K value are determined. If X is incremental range, then incremental range method is chosen. In this method the attribute is a range of numbers (set of numbers) such as age, date, time, etc. The user might select incremental range, and provide the value of increment to combining the records together with respect to a sub-range in the Age attribute. For X to be incremental suppression, then this method is applied by performing it incrementally to replacing the characters of the attribute to "*". Consequently, once two values or more become matching each other they are included in one group and the diversion is used with ZipCode. Otherwise, the unification method is applied. This operation is iterative until the desired K value is reached. 'Figure( 5)' represents the pseudo code of the heuristic K-anonymity steps.

1. **INPUT**: Quasi_Identifiers, Division _Method, K
2. **OUTPUT**: K_Anonymized_Table
3. **Main**
4. **WHILE**(MinGroup<K)
5. X=Find_Index_of_Min (Division _Level[])
6. **IF**(Division _Method(x)==INCREMENTAL_RANGE)
7. Incremental_Range(x);
8. **ELSEIF**(Division _Method(x)==INCREMENTAL_SUPPRESSION)
9. Incremental_Suppression(x);
10. **ELSEIF**(Division _Method(x)==UNIFICATION)
11. Unification(x);
12. **End While**
13. Division _Level[x]++;
14. **END**

**Figure 5.** Algorithm representing the pseudo code for Heuristic K-anonymity technique.

4.2. *Heuristic (L-diversity) and (α, k)-anonymity Technique*

K-anonymity is extended using an L-diversity and (α, k)-anonymity methods to maintain a limit of maximum frequency of distance (sensitive) values in the critical attributes. Figure 6 shows a new value of K that is enabled once the K- groups are generated. To overcome the weakness of K-anonymity from homogeneity attack such as the QI group having several tuples with same sensitive attributes value, one needs to achieve L-diversity with (α, k)-anonymity methods. Furthermore, K is increased to enhance the diversity in the data until the achieved maximum frequency is less than 20%. This percentage is selected based on the natural of data as tuneable parameter. In other word, to decrease the frequency of sensitive attributes, this percentage is selected heuristically.
**Figure 6.** Flowchart of heuristic (L-diversity) and (α, k)-anonymity technique. Figure(7) represents the steps of pseudo code of heuristic (L-diversity) and (α, k)-anonymity. Meanwhile, the frequencies are calculated for distinct values of sensitive attributes. If they meet the condition (less 20%), the loop is ended else it is continued to increase K until the condition is met.

1. **INPUT**: Quasi_Identifiers, Diversion_Method, K
2. **OUTPUT**: K_Anonymized_Table
3. **Main**
4. **K_Anonymity(K);**
5. **While(Max_Freq()<0.2)**
6. **K++;**
7. **K_Anonymity(K);**
8. **Export_Table**
9. **End**

**Figure 7.** Pseudo-code for Heuristic (L-diversity) and (α, K)-anonymity.

The frequencies of distinct values of sensitive attributes are measured with respect to L-diversity with (α, k)-anonymity techniques by increasing K to maintain a limit of maximum frequency of distance values in the sensitive attributes. Figure 8 demonstrates the frequency of distinct values with respect to increase K. This demonstrates the spreading of the critical values among the whole equivalent group. It is observed that there is a less possibility of attacking record when the frequencies of critical values are decreased in every equivalent group of tuples. For example, if a record has critical value of high frequency the background or homogeneity attack can affect the individual. There is diversity on critical values to prevent these attacks from breaching the privacy. This histogram illustrates the reduction in frequency value of tuples and increase of diversity in critical values.
The experimental results demonstrate that the present heuristic anonymization technique can prevent the privacy breach in distributed datasets on cloud environment. By taking the advantage of K-anonymity, L-diversity and (α, k)-anonymity the unified technique overcomes the weakness of each method. Table 3 enlists the result after applying the heuristic (L-diversity) and (α, k)-anonymity technique. It is clear that the equivalent group contains at least L acceptable values of sensitivity attributes, which has 6053, 643, 1411, 1483, 397, 499, 2235, 0, 1381, 2140, and 5944 balances. This signifies the presence of diversity in sensitive attributes with diminished frequency values. For example, in the first sensitive value (6053) the frequency is reduced with increasing K value, similarly for other critical attributes. Furthermore, the group size is found to increase with the decrease in frequencies values of sensitive attributes. Small group size revealed the higher frequency. These frequencies are reduced by increasing the group size to avoid background and homogeneity attack.

5. Experimental Results and Evaluation
The experimental results demonstrate that the present heuristic anonymization technique can prevent the privacy breach in published datasets on cloud environment. By taking the advantage of K-anonymity, L-diversity and (α, k)-anonymity the unified technique overcomes the weakness of each model. For the first experiment, the value of K is taken as 20. Every group possessed many tuples and the range is increased to 25. Therefore, it is essential to reduce the K value. In the second experiment, the K value is chosen to be 10. Then, the value of K is changed to 5 and this time the information loss is reduced with lower range as shown in Table 1. Table 1 enlists the result after applying the heuristic K-anonymity technique. As shown in Table 1, two groups are resulting from using the proposed technique where K value is 5. The QI attributes that generalized are (age and job). The first group has nine-tuples. Which mean that every tuple in this group has matched with other eight tuples in the same group so; it is hard to distinguish every tuple. The second group has eleven tuples that are indistinguishable. The reason of difference in a number of anonymized tuples in every group, which is based on condition Mingroup ≥ K as shown in Figure(4)
Table 1. Two groups resulting from applying the method (K=5, QI: age and job)

| Age | Job                           | Marital | Education | Balance |
|-----|-------------------------------|---------|-----------|---------|
| 1   | 90-105 Retired-Services-Admin.-Technician | Divorced | Secondary | 1       |
| 2   | 90-105 Retired-Services-Admin.-Technician | Divorced | Primary   | 712     |
| 3   | 90-105 Retired-Services-Admin.-Technician | Married  | Unknown   | 775     |
| 4   | 90-105 Retired-Services-Admin.-Technician | Married  | Unknown   | 775     |
| 5   | 90-105 Retired-Services-Admin.-Technician | Married  | Unknown   | 775     |
| 6   | 90-105 Retired-Services-Admin.-Technician | Married  | Unknown   | 775     |
| 7   | 90-105 Retired-Services-Admin.-Technician | Divorced | Secondary | 1234    |
| 8   | 90-105 Retired-Services-Admin.-Technician | Divorced | Primary   | 2282    |
| 9   | 90-105 Retired-Services-Admin.-Technician | Married  | Secondary | 0       |
| 1   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Unknown   | 4984    |
| 2   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Unknown   | 1780    |
| 3   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Unknown   | 4984    |
| 4   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Unknown   | 4984    |
| 5   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Unknown   | 1780    |
| 6   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Tertiary  | 5619    |
| 7   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Unknown   | 1780    |
| 8   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Unknown   | 1780    |
| 9   | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Primary   | 6483    |
| 10  | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Tertiary  | 0       |
| 11  | 75-90 Unemployed-Entrepreneur-Self-Employed-Management | Married  | Secondary | 0       |
Table (2) enlists the result after applying the heuristic (L-diversity) and (α, k)-anonymity technique. It is clear that the equivalent group contains at least L acceptable values of sensitivity attributes, which has 6053, 643, 1411, 1483, 397, 499, 2235, 0, 1381, 2140, and 5944 balances. This signifies the presence of diversity in sensitive attributes with diminished frequency values. For example, in the first sensitive value (6053) the frequency is reduced with increasing K value, similarly for other sensitive attributes. Furthermore, the group size is found to increase with the decrease in frequencies values of sensitive attributes. Small group size revealed the higher frequency. These frequencies are reduced by increasing the group size to avoid threats.

**Table 2. Results of heuristic (L-diversity) and (α, k)-anonymity technique**

| Age  | Job                        | Marital       | Education                  | Balance |
|------|---------------------------|---------------|---------------------------|---------|
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 6053    |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 6053    |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 643     |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 1411    |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 1411    |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 1483    |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 397     |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 397     |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 499     |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 2235    |
| 75-100 | Unknown-Student-Housemaid-Unemployed-Entrepreneur-Blue-Collar | Divorced-Single-Married | Unknown-Primary-Tertiary-Secondary | 0       |
The efficiency and effectiveness of the heuristic anonymization technique is compared with the recent art-of-the technique (Zhang et al., 2014). The performance evaluation of the heuristic anonymization technique is carried out using UTD Anonymization ToolBox software, where entropy and execution time are calculated as privacy measure.

**Entropy Evaluation:** UTD Anonymization Tool Box is used to compare the results of heuristic anonymization technique with K-anonymity and original data as shown in Figure 9. Figure 9 displays that the K-anonymity reduced the entropy level of the data by 13.5% from 0.3823 to 0.3308. However, by implementing the heuristic anonymization technique on the same dataset, the entropy revealed a reduction of 0.75% only, which is much lower compared to K-anonymity technique (13%). This clearly indicates the outperformance of heuristic anonymization technique in preserving the entropy of the data with small percentage difference from the original data. It is important to note that the entropy measure depends on the data. Moreover, the entropy measure of heuristic anonymization technique is observed to be too close to entropy measure of original data. This signifies that the entropy measure for heuristic anonymization technique is superior compared to the entropy measure.
of K-anonymity. Besides, the entropy is discerned to be insensitive to the value of K. This can be interpreted with the large amount of data with respect to small ranges of values of K.

![Figure 9. Comparison of entropy measure for K-anonymity and heuristic anonymization technique](image)

**Privacy Measure:** Privacy being complementary of entropy, the privacy measure of the heuristic anonymization technique is evaluated using UTD Anonymization Tool Box. The privacy is observed to decrease with the increase of information contents. Privacy measure for K-anonymity and heuristic anonymization technique are compared in Figure 10. The privacy level is found to enhance in the presence of K-anonymity. Thus, it is established that the heuristic anonymization possesses more utility than the normal K-anonymity. Owing to this attribute the heuristic anonymization technique protects the privacy without increasing information loss, where the preserved privacy level is over 65%. Moreover, the percentage remains the same even with the increase of information and the preserving entropy of the original data does not exceed 10%. Indeed, it is a good performance indicator of the heuristic technique.

![Figure 10. Comparison of privacy for K-anonymity and heuristic anonymization technique](image)

**Execution Time:** The computational complexity of both K-anonymity and heuristic anonymization technique are evaluated in terms of their execution time as illustrated in Figure 11. It is evident that the heuristic anonymization technique spent less execution time than K-anonymity. The embedding of other models such as L-diversity and (α, k)-anonymity in the heuristic anonymization technique to ensuring the strong privacy without affecting on execution time.
6. Conclusion

This paper provides the basic concept of pilot study on anonymization model and demonstrates that the synergy of different privacy preserving models can provide a strong privacy balance with the information utilization. The detailed correlation between cloud computing and privacy preserving of published data is exemplified. With the materialization of CC technologies, various organizations attempted to leverage the Internet-based paradigm. The main aim is to optimize the utility of computing resources in a flexible and scalable manner. Despite the increased use of cloud-based platforms in diverse business areas, the privacy and security requirements have prevented their adoption in the healthcare and related areas. Actually, the privacy concerns (data utilization requirements) of different data providers (consumers) are quite different. Over the years, numerous techniques are developed to ensure the privacy of sensitive data publication on cloud environment. Yet, no precise and efficient solution for privacy is achieved. To surmount this limitations, present paper proposed a data privacy preserving technique useful for cloud publishing. To achieve high-quality generalization, a novel anonymization technique is developed by unifying the techniques such as heuristic K-anonymity, heuristic L-diversity and (α, k)-anonymity. The need of hybrid solution via the innovative integration of different privacy protection models is justified. The experimental results obtained using the proposed anonymization technique revealed enhanced privacy protection against various attacks and offered the obligatory data for research analysis.

7. References

[1] X. Ding, Q. Yu, J. Li, J. Liu, and H. Jin, “Distributed anonymization for multiple data providers in a cloud system,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2013, vol. 7825 LNCS, no. PART 1, pp. 346–360.
[2] H. Kargupta, S. Datta, Q. Wang, and K. S. K. Sivakumar, “On the privacy preserving properties of random data perturbation techniques,” Third IEEE Int. Conf. Data Min., 2003.
[3] L. Sweeney, “Privacy-Preserving Surveillance using Selective Revelation,” no. October, 2005.
[4] M. Prakash and G. Singaravel, “An approach for prevention of privacy breach and information leakage in sensitive data mining,” Comput. Electr. Eng., pp. 1–7, 2015.
[5] T. Ristenpart and E. Tromer, “Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds,” … Conf. Comput. …, 2009.
[6] S. Kamara, C. Papamanthou, and T. Roeder, “Dynamic searchable symmetric encryption,” … 2012 ACM Conf. …, pp. 965–976, 2012.
[7] R. J. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” in Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, 2005, 2005, no. Icde, pp. 217–228.
[8] L. Sweeney, “k- Anonymity: A Model For Protecting Privacy,” Int. J. Uncertainty, Puzziness Knowledge-Based Syst., vol. 10, no. 5, pp. 557–570, 2002.

[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,” Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 754–759, 2006.

[10] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “–Diversity,” ACM Trans. Knowl. Discov. Data, vol. 1, no. 1, p. 3–es, 2007.

[11] N. Li, T. Li, and S. Venkatasubramanian, “Closeness: A new privacy measure for data publishing,” IEEE Trans. Knowl. Data Eng., vol. 22, no. 7, pp. 943–956, 2010.

[12] X. Xiao and Y. Tao, “M-invariance: towards privacy preserving re-publication of dynamic datasets,” SIGMOD Conf., pp. 689–700, 2007.

[13] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing,” ACM Comput. Surv., vol. 42, no. 4, pp. 1–53, Jun. 2010.

[14] P. Jurczyk and L. Xiong, “Distributed anonymization: Achieving privacy for both data subjects and data providers,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5645 LNC, pp. 191–207, 2009.

[15] W. Jiang and C. Clifton, “A secure distributed framework for achieving k-anonymity,” VLDB J., vol. 15, no. 4, pp. 316–333, 2006.

[16] S. Zhong, S. Zhong, Z. Yang, Z. Yang, R. N. Wright, and R. N. Wright, “Privacy-enhancing k-anonymization of customer data,” Proc. twenty-fourth ACM SIGMOD-SIGACT-SIGART Symp. Princ. database Syst., p. 147, 2005.

[17] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, “k-Anonymous data mining: a survey,” Adv. Database Syst., pp. 105–136, 2008.

[18] B. B. Patil and A. J. Patankar, “Multidimensional k-anonymity for protecting privacy using nearest neighborhood strategy,” in 2013 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2013, 2013, p. IEEE Podhigai; IEEE Signal Processing, Computation.

[19] C. Tai, P. S. Yu, and M. Chen, “Privacy Preserving Frequent Pattern Mining on Multi-cloud Environment,” 2013 Int. Symp. Biometrics Secur. Technol., pp. 235–240, Jul. 2013.

[20] C. Tai, P. S. Yu, and M. Chen, “k-Support Anonymity Based on Pseudo Taxonomy for Outsourcing of Frequent Itemset Mining,” Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 473–482, 2010.

[21] X. Zhang, C. Liu, C. Yang, W. Dou, and J. Chen, “Combining Top-Down and Bottom-Up : Scalable Sub-Tree Anonymization over Big Data using MapReduce on Cloud,” 2013.

[22] 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,” J. Comput. Syst. Sci., vol. 80, no. 5, pp. 1008–1020, Aug. 2014.

[23] X. Zhang, L. T. Yang, C. Liu, and J. Chen, “A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud,” IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 2, pp. 363–373, 2014.

[24] J. J. Yang, J. Q. Li, and Y. Niu, “A hybrid solution for privacy preserving medical data sharing in the cloud environment,” Futur. Gener. Comput. Syst., vol. 43–44, pp. 74–86, 2014.

[25] A. Gkoulalas-Divanis and G. Loukides, “Anonymization of electronic medical records to support clinical analysis,” pp. 9–30, 2013.

[26] Y. Xiao, C. Lin, Y. Jiang, X. Chu, and F. Liu, “An Efficient Privacy-Preserving Publish-Subscribe Service Scheme for Cloud Computing,” 2010 IEEE Glob. Telecommun. Conf. GLOBECOM 2010, vol. 1, pp. 1–5, Dec. 2010.

[27] A. Sachan, D. Roy, and P. V Arun, An Analysis of Privacy Preservation Techniques in Data Mining. 2013.

[28] Z. Xu and X. Yi, “Classification of Privacy-preserving Distributed Data Mining protocols,” in 2011 Sixth International Conference on Digital Information Management, 2011, pp. 337–342.
[29] Hamza, “Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing,” J. Inf. Secur., vol. 2013, no. April, pp. 101–112, 2013.

[30] X. Xiao, “The Hardness and Approximation Algorithms for,” arXiv:0912.5426v1, 2010.

[31] S. Moro and R. M. S. Laureano, “Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology,” Eur. Simul. Model. Conf., no. Figure 1, pp. 117–121, 2011.

[32] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” Decis. Support Syst., vol. 62, pp. 22–31, 2014.

[33] K. Bache and M. Lichman, “UCI machine learning repository.” 2013.

[34] H. A. Elsalamony, “Bank Direct Marketing Analysis of Data Mining Techniques,” Int. J. Comput. Appl., vol. 85, no. 7, pp. 12–22, 2014.

[35] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, “A Systematic comparison and evaluation of k-Anonymization algorithms for practitioners,” Trans. Data Priv., vol. 7, no. 3, pp. 337–370, 2014.

Acknowledgements
Author is thankful to UCI (University of California at Irvine) Machine Learning Repository for providing the dataset.