International Journal of Trend in Scientific Research and Development (IJTSRD)

International Open Access Journal

ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 6

# A Technological Survey on Privacy Preserving Data Publishing

**Rajshree Srivastava**
Department of Computer Science and Engineering
Jaypee Institute of Information Technology
Noida-128, (U.P.), India

**Kritika Rani**
Department of Computer Science and Engineering
Jaypee Institute of Information Technology
Noida-128, (U.P.), India

## ABSTRACT

There is an enormous collection of records of a particular individual in a system. These data needs to be highly secured. Due to easily availability of these records, the records or information is on high risk. These records are being used for the business purpose and as well as for the decision-making in the respective domain. However, any data that is in the raw form comes in the category of sensitive record as it contains the complete information of each particular individual. In the present scenario while publishing, the data mainly depend on the rules, policies and guidelines so that only the required information are published based on the agreements. Hence, privacy preserving and data publishing can be defined as tool and methods for publishing information while preserving privacy of the records. In this paper, there is a survey of various techniques and algorithm designed so far in order to preserve privacy of the data.

***Keywords***: *k-nonymity, privacy preserving data publishing, l-diversity, slicing, anatomizations*

## I. INTRODUCTION

The original data of the individual is very sensitive so, there is risk of information breach. Original data of any particular individual can be his name, age, gender, employee code, company name, address, contact number, salary, etc. Therefore, the main challenge for the privacy of the data is to design methods and tools for publishing data in a risk-free environment.

In privacy preserving data publishing, there are two stages to successfully complete the process of publishing the record, these include data collection and data publisher. First task is to collect all data from the required domain of the owner and second task is to publish those collected data in the public with the security of the data such that the individual identity cannot be identified. Further, data publisher is the one who is having the original set of records, after collection of records the anonymiztion technique is applied in this phase only. After that is being, carry forward to the data recipient who publishes the data in the public. There are various anonymization techniques which are applied on the original table to make it more secure, these include generalization, suppression, etc.

Anonymization is defined as a process that removes or replaces the identity of the individual from any record. The original table or the original set of records must satisfy the any of the anonymization technique. In generalization the values or record are being generalized. For e.g. age 10-20 are being grouped into one, age 20-30 are being grouped into one and so on. While in the case of suppression, either the values or records are being deleted or it is being replaced with less distinct values in order to maintain uniformity.

## II. RELATED WORK

Privacy is considered one of the essential factors for publishing the data for preserving the data in the effective manner. In order to preserve the data there are various techniques which are being introduced in Privacy-preserving data publishing (ppdp). The first technique is k-anonymity, which states that if there is any given person specific detail it must produce a released data that guarantees that the individuals who are suspected for the data cannot be re-identified and

data remain useful [1]. Any published data is said to have adhering the property of k-anonymity if the information for each person contained in the released that cannot be identified from at least k-1 individuals

present [2]. This technique has some of the vulnerabilities due to which a new technique is being proposed known as l-diversity. A q* block is said to l-diverse if it contains at least l "well represented" values for the sensitive attributes(S)[5]. Any table is said to be l-diverse if any only if every q* block is diverse. It is easy to achieve but practical it is very difficult to achieve. The key idea is to limit the disclosure risk, for this to achieve there is requirement to measure the disclosure risk of the anonymized table. The limitation of l-diversity is that it is limited upto some extent on its assumption of the adversial knowledge. T-closenesss proposed a novel notation that formalizes the idea of background knowledge [7].

In order to calculate the distance between the values of the sensitive attributes Earth movers distance is considered best. Any equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in the present class and the distribution of the attribute present is the complete table is not more than the threshold value. Slicing is one of the new techniques which is introduced that partitions data both horizontally and vertically [11].

In slicing, there is random grouping of data due to which we does not have a clear scenario. There is information loss further in this case. An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it. [13]To address the limitation of slicing, overlapping slicing is being introduced that handless the data attributes on the concept of fuzzy clustering. It ensures the utility of the published data by adding attribute in the column so that attributes, which are duplicate, are combined to get better correlation value. It fails to support high dimensional data log with the multiple sensitive data. [16]In order of overcome the drawback of overlapping slicing Anatomization technique is being introduced. In this technique, the limitation of the overlapping slicing is being solved and further the loss of information is low. It strictly follows the property of k-anonymity as well as l-diversity. The approach of anatomy states that if there are two tables with join attributes and it goes for publishing, then it will correspond to those two tables that come in the category of lossy. Different types of techniques , are being discussed and compared in the table.

**TABLE I**

**COMPARATIVE PAPER STUDIED [2002- 2016]**

| YEAR | NAME OF THE PAPER | AUTHOR | TECHNIQUE USED | ALGORITHM APPLIED | DISADVANTAGES | ADVANTAGES |
|---|---|---|---|---|---|---|
| **2002** | Achieving k-Anonymity Privacy Protection Using Generalization and Suppression | L.Sweeney | MinGen | Global, bottom-up, complete, impractical | The two types of attack: background knowledge and homogeneity is possible.<br>*the problem is NP-hard<br>*the search space is exponential in this case. | It supports identity disclosure. |
| **2004** | Bottom-Up Generalization: A Data Mining Solution to Privacy Protection | Wang | Bottom-up-generalization | Global, bottom-up, greedy | *in this user may not any point or time can stop and can obtain a generalized table which is fully satisfying the property of anonymity. | *shows the transformation of specific data to less specific data.<br>*it shows better scalability |
| **2005** | Top-Down Specialization for information and privacy preservation. | Fung | Top-Down Specialization | Global, top-down, greedy | It does not handle numerical attributes. | *preserves both information utility and also privacy of the individual.<br>* in this user can stop at any point or time and can obtain a generalized table which is fully satisfying the property of anonymity. |
| **2005** | Incognito: Efficient full-domain k-anonymity | LeFevre | Incognito | Global, bottom-up, hierarchy-based, complete | In this the surely of minimality is not present. In this guarantee of minimal data is not possible. | *It is feasible for large databases<br>*in this the problem of Min. Gen. Algorithm is been solved. |
| **2006** | Mondrian Multidimension ak-anonymity | LeFevre | Mondrian | Local, top-down, partition-based, greedy | *in this anonymization is quite sensitive if there will be any small change in the synthetic data there will be variation in the result. | *in this also the problem of Min. Gen Algorithm is solved. |
| **2006** | l-diversity | A.Machanavajjhala | Bayes Optimal Privacy | | *attribute and record linkage is possible.<br>*limitation is in the practice as multiple adversaries have different levels of knowledge.<br>*extension of handling multiple sensitive | *It protects from the background knowledge attack.<br>*it supports identity disclosure. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | attributes , and to develop methods for continuous sensitive attributes | |
| **2006** | Anatomy: a simple approach for preserving data | X.Xiao,Y.Tao | Anatomy technique | Linear algorithm | | *obeys l-diversity<br>*overcome the problem of generalization. |
| **2007** | T Closeness: Privacy Beyond k-anonymity and l-diversity | N.Li,T.Li and S.Venkatasubramanian, | For the calculation the distance EMD is been used. | | *vulnerable to probabilistic attack<br>*vulnerable to similarity and skewness attack. | *protects against identity disclosure. |
| **2010** | A survey : Recent Development on Privacy-preserving and Data publishing | B.C.M.Fung, R.Chen, P.S.Yu | | | . | *outline the publishing technologies and the model.<br>*it ehncances the knowledge about advantages and limitation of all privacy preserving data publishing techniques defined so far. |
| **2012** | Slicing: A new Approach to privacy Preserving Data Publishing | Tiancheng Li, Ninghui Li, Jian Zhang, Ian Molloy | New anonymization technique known as slicing technique. | Slicing algorithm | *they have considered slicing where only one attribute is in one particular column only.<br>*random grouping is not best way of completing the task.<br>*the values of data in the column of the bucket are arranged randomly. | *it overcomes the limitation of generalization and bucketization.<br>*it protects against privacy threats. |
| **2013** | Data Slicing Technique to Privacy Preserving And Data Publishing | Alphonsa Vedangi, V.Anandam | Slicing technique | Slicing algorithm | *they have considered slicing where only one attribute is in one particular column only.<br>*random grouping is not best way of completing the task.<br>*the values of data in the column of the bucket are arranged randomly. | *it overcomes the limitation of generalization and bucketization.<br>*it protects against privacy threats. |
| **2014** | A Data | J.Yang, Z.Liu, | Decompositi- | Overlappin | *it cannot handle | *Data security is |

| | | | | | | |
|---|---|---|---|---|---|---|
| | Anonymous Method based on Overlapping Slicing | J.Zhang | on technique | g algorithm | multiple sensitive attribute data. | preserved in a better way.<br>*attribute duplication is more in one column.<br>*the problem of the slicing is solved in this case. |
| **2016** | Anatomization with slicing: a new privacy preservation approach for multiple sensitive attributes | V.Shyamala Susan ,T.Christopher | Anonymization Technique | Advanced new clustering algorithim and Minkowsi distance measure algorithm | *it is not applicable on quasi-identifier table(QIT)and sensitive table(SA) | *It can handle multiple sensitive attribute data. |

**TABLE II**

**COMPARITIVE STUDY OF PRIVACY MODELS AND THE TECHNIQUES**

| MODELS OF PRIVACY | K-ANONYMITY | L-DIVERSITY | T-CLOSENESS | SLICING |
|---|---|---|---|---|
| **Initiation** | [2002] | [2006] | [2007] | [2012] |
| **Attacks** | Background attack and as well as homogeneity attack | Skewness attack and attack of similarity. | Skewness attack and as well as attack of similarity. | It is not vulnerable to this type of attack. |
| **Disclosure** | Fails to support attribute disclosure but supports identity disclosure. | does not support attribute disclosure. | It does not prevent. | It prevents attribute disclosure risk. |
| **Ease** | It is simple and easy to understand and can be achieved. | It may be difficult to achieve. | It is somewhat easy to achieve in comparison to l-diversity. | It is easy to achieve. |
| **Categorization** | It is a privacy model. | It comes under privacy-preserving and data-publishing categorization. | It comes under privacy-preserving and data publishing categorization. | It comes under the category of privacy-preserving and data publishing(ppdp) categorization. |
| **Privacy against attacker** | There is no guarantee of privacy against attacker using background knowledge. | It also does not guarantee. | It also does not guarantee. | It may and mot guarantee against the privacy against attacker. |
| **Probabilistic Attack** | In this probabilistic attack is not possible. | In this probabilistic attack is not possible. | Probabilistic attack is possible. | Probabilistic attack is not possible. |
| **Record Linkage** | In this record linkage is possible. | Record linkage is possible. | In this record linkage is possible | In this record linkage is not possible. |
| **Attribute Linkage** | In this attribute linkage is possible. | In this attribute linkage is possible. | In this attribute linkage | In this attribute linkage is also not possible. |
| **Monotonocity Property** | It satisfies the property of monotonicity property. | It also satisfies the property of monotonicity property. | It may or may not satisfy the monotonicity property. | It satisfies the property of monotonicity. |
| **Categorization** | It does not have any categorization. | It has been categorized into 3 types. | It does not have any category-zation. | It has been categorized into two categories: slicing and overlapped slicing. |
| **Record Linkage** | In this record linkage is possible. | Record linkage is possible. | In this record linkage is possible. | In this record linkage is not possible. |
| **Data Utility loss** | Is medium | Is medium | In this data utility loss is high. | Is very low in the case of high dimensionality. |
| **Membership Disclosure** | Yes it is possible | Yes it is possible | No it is not possible | No it is not possible. |
| **High dimensionality** | It cannot be applied on high dimensionality. | It does not support high dimensionality. | it also does not support high dimensionality. | It supports high dimensionality. |

## III. CONCLUSION

In this paper, there is survey of the different techniques which are being introduced till now in order to preserve privacy of the record. The last technique i.e. Anatomization is considered one of the best technique so far among all the techniques. Future work can be introduction to new technique where loss of information is less and is more highly secured. A new technique should strongly follow the property of k-anonymity , l-diversity and as well as t-closeness , so that the co-relation among the attributes which are less can be grouped together and then anonymization technique can be applied.

## IV. REFERENCES

[1] L.Sweeney, "K-anonymity: a model for protecting privacy", International Journal Uncertain Fuzz, 10(5):577-570, 2002(a)

[2] L.Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression", International Journal Uncertain Fuzz, 10(6): 571-588, 2002(b)

[3] K.LeFevre, D.J.DeWitt and R.Ramakrishnan, "Incognito: Efficient full domain k-anonymity", In SIGMOD, pages 49-60, 2005

[4] A.Machanavajjhala, J.Gehrke, D.Kifer and M.Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity. In Proc. 22nd International Conference Data Engg(ICDE), page 24, 2006

[5] A.Machanavajjhala, Kifer,D.,Gehrke, J.,and Venkitasubramaniam.M, "L-diversity: privacy beyond k-anonymity", ACM Trans Knowledge Discover Data 1(1):1-52, 2007

[6] X.Xiao and Y.Tao, "Anatomy: simple and effective privacy preservation", In VLDB '06: Proceedings of the 32nd International conference on Very large data bases, pages 139-150, VLDB Endowment, 2006

[7] N.Li, T.Li, and S.Venkatasubramanian, "t-closeness: privacy beyond k-anonymity and l-diversity", Proc. IEEE International Conferenece data Eng.(ICDE), pp. 106-115, 2007

[8] D.J.Martin, D.Kifer, A.Machanavajjhala, J.Gehrke and J.Y.Halpern, "Worst-case background knowledge for privacy-preserving data publishing", In ICDE, pages 126-135, 2007

[9] H.Tian, et al. W.Zhang, "Extending l-diversity for better Data Anonymization.", Sixth International Conference on Information Technology: New Generations, 2009

[10] B.C.M.Fung, K.Wang, R.Chen and P.S.Yu, "Privacy-preserving data publishing: A survey on recent developments", ACM Computing Surveys, 2010

[11] T.Li,N.li, J.Zhang and L.Molloy, "Slicing: a new approach for privacy preserving data publishing", In IEEE Transactions on knowledge and Data Engineering, volume 24, page 561-574, 2012

[12] A.Vedangi, V.Anandam, "Data slicing technique to privacy preserving and data publishing", International Journal of Research in Engineering and Technology, 2(10):120-6, 2013

[13] J.Yang, Z.Li,J.Zhang, "A data anonymous method based on overlapping slicing", Proceedings of the IEEE 18th International Conference o Computer Supported Cooperative Work in Design, 2014

[14] A.A.Dhaigude, P.Kumar, "Improved Slicing Algorithm for greater utility in privacy preserving data publishing", International Journal of Data Engineering (IJDE), Volume (5): Issue (2); 2014

[15] M.D.Kamalesh, B.Bharathi, "Slicing an efficient Transaction Data Publication and For Data Publishing", International Journal of Science and Technology, Volume 8(S8), 306-309, 2015

[16] V.Shyamala Susan and T.Christopher, "Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes", Springer Plus (2016)5:964