© 2019 IJSRSET | Volume 6 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET19629

A Study on Models and Techniques of Anonymization in Data Publishing

Shipra Sharma, Naveen Choudhary, Kalpana Jain
Department of CSE, College of Technology and Engineering, Udaipur, Rajasthan, India

ABSTRACT

In the era where world runs online the storing and publishing of data online has also increased to a great extent. In this era a large amount of information is collected and published to a network which is publically available. With the exposure of data comes the risk of information leakage of an individual while publishing the data online. Hence for the same we need a security system for preserving the privacy of individual and here the concept of preserving privacy in data publishing came into existence. To achieve this privacy different privacy models and techniques have been proposed which gives different levels of resistance against different attacks by adversaries. In this paper we will discuss about these models and techniques and have a comparative study among them.

Keywords : Privacy Models, Anonymization Techniques, Data Publishing, Privacy Preservation.

I. INTRODUCTION

The publishing of data involves providing the data for public use for further research, study or surveys. But when the data is published the identity of individuals must be preserved to maintain the privacy. This procedure of maintaining the privacy results in loss of information of data and decreases its utility. So the major challenge in this field is to preserve the privacy with minimum data loss.

During the publishing of data we modify the data in such a way that it does not lead to identity leak of an individual and make it anonymous is a process called anonymization. But before anonymization of data we need to understand different type of data which exists.

1. Identifier: The fields or values which uniquely identify an individual are called Identifier. For example name, social security number.

2. Quasi Identifier: The values which do not directly identify an individual but when linked with external data set it can lead to identity disclosure as shown in fig. 1.

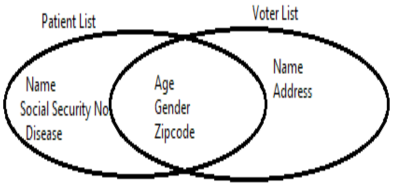
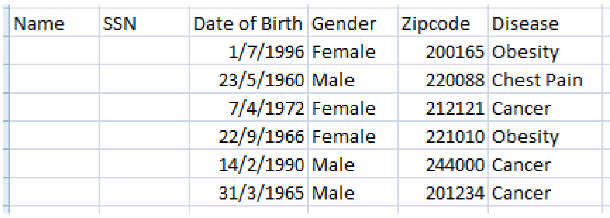


Fig. 1: Quasi identifier linkage example.

3. Sensitive Attribute: The values which a person doesn't want to disclose or share. For example disease or salary.

4. Non Sensitive Attribute: The details even if leaked won't harm the individual are non sensitive attribute.

Hence in anonymization we remove the identifier field from the data set so that no direct identification of individual can be possible. Then we modify the quasi identifier to prevent from linkage attack before publishing the data. Table 1 shows an example of anonymization in which the data is anonymized before publishing. In this example the identifiers which are name and Social Security Number (SSN) are removed. The quasi identifiers in below table are date of birth, gender and zip code which are modified before publishing. The sensitive field is disease which a person does not want to disclose. Here the main focus is not to hide the sensitive data but to hide the identity of individual whose data is being published.



Name	SSN	Date of Birth	Gender	Zipcode	Disease
		1/7/1996	Female	200165	Obesity
		23/5/1960	Male	220088	Chest Pain
		7/4/1972	Female	212121	Cancer
		22/9/1966	Female	221010	Obesity
		14/2/1990	Male	244000	Cancer
		31/3/1965	Male	201234	Cancer

Table 1 : Anonymization example

II. PRIVACY MODELS

To describe a metrics for privacy or the risk of disclosure of identity privacy models were proposed. Privacy models describe how much the system is capable of handling the attacks made for identity leak. Privacy preserving in data publishing has three models as described below.

A. k-Anonymity:

Sweeney and Samarati[1] proposed an approach for protecting the data from record linkage attack. The record linkage attack means when the attacker can infer the knowledge from linking the published data to some externally available data and extract the required information from the table. In table 2 we can see that the there are two database table. In this the attacker can link both the table with common parameters like Zip code and DOB and identify the disease (sensitive attribute) of particular adversary.

Table 2 : Medical Data Table

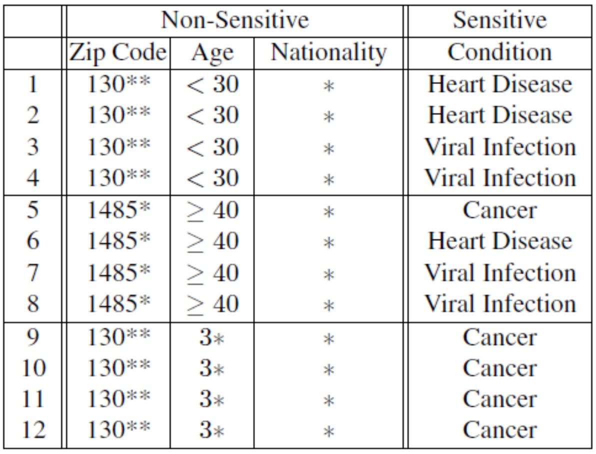
Name	DOB	Gender	Zip code	Disease
	2/3/98	Female	221100	Cancer
	15/1/70	Male	213200	Ulcer
	4/6/88	Male	211500	Obesity
	31/8/90	Female	221100	Cancer
	21/11/98	Female	221100	Cancer
	8/2/66	Male	213200	Ulcer
	1/5/73	Male	211500	Obesity
	6/2/90	Female	221100	Cancer

Table 3 : Voter List

Name	Address	City	Zip Code	Gender	…
….	….	…	….	…	…
John	8/2/66	Delhi	221100	Male	…

This scenario was handled by k-anonymity. k-anonymity proposes the dataset should be divided into equivalent classes such that for a record with given attribute there should be k-1 attributes which match those attributes. So basically, it says table is divided into equivalence classes where the quasi identifier of k rows are indistinguishable in each class. Table 4 shows an example of k anonymous table where the quasi identifiers such as zip code, age and nationality of an equivalence class are similar.

Table 4 : 4-anonymous table



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

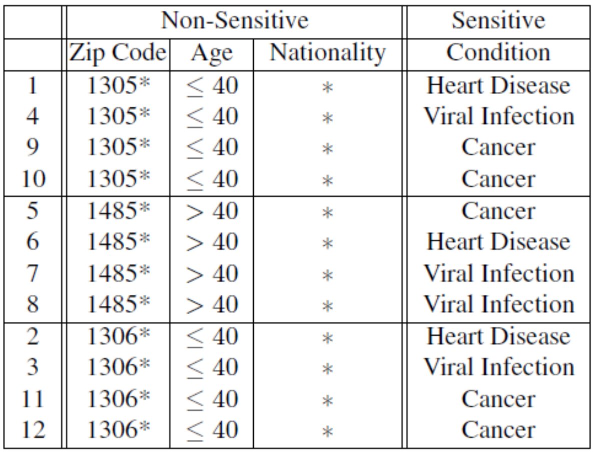
Attacks on k-anonymity

Although protecting from record linkage attack k-anonymity does not protect from attribute linkage attack. Attribute linkage attack is the attack in which the if all the members of a group shares a same value and then the attacker would need not to precisely know the exact record, he can conclude from that certain value. As shown in table 4 there is a 4-anonymous table [2]. In this table the attacker knows that Bob lives in zip code 13053 and is of age 35 then the attacker knows that Bob's record number is 9,10,11,12. So the attacker need not to know exactly which is Bob's record, the attacker can conclude Bob is suffering from Cancer.

B. l-diversity

Prone to attribute linkage attack a new privacy model known as l-diversity was proposed Machanavajjhala et al. [2] which can handle the attacks of k-anonymity. It states that for a data set to be l diverse then in an equivalence class there should be l "well represented" values for sensitive attributes in an equivalence class. The basic understanding of term well represented is that in each equivalence class there should be l different values of sensitive attributes related to each equivalence class.

Table 5 : 3 diverse table[2]



	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

In the table 5 with each equivalence class there are three different type or we can say 3 well represented values of sensitive attribute (here disease) are associated. As k-anonymity worked on quasi identifiers, l diversity is an extended form of k anonymity which works sensitive attributes.

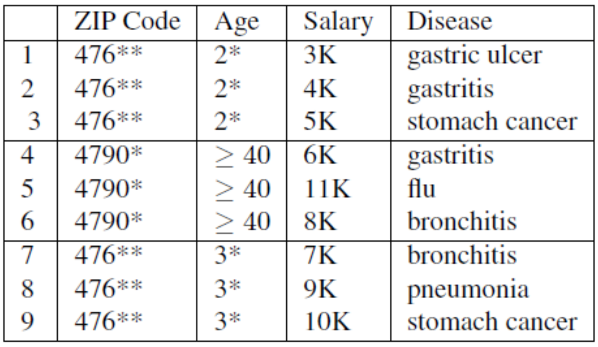
Attacks on l diversity

Li et al.[3] in there paper stated that l diversity may be difficult and unnecessary to achieve. This was explained by taking different examples. Suppose a medical data set of test result of virus was taken where the result was only sensitive attribute and it can be either positive or negative. It was also found out of population of 10,000 only 1% of population gets infected by the virus that is only 1% gets positive result. Then people with positive record will not want to disclose this information and people with negative class will not mind it the information is leaked. So for equivalence with all the negative records 2-diversity will be unnecessary to achieve. Now for making 2 diverse with respect to result there can be at most 10000*1% = 100 equivalence classes which further result into information loss and hence makes it difficult to achieve.

Skewness Attack: When the distribution of sensitive attribute is skewed in overall table then even though l-diversity is satisfied it fails in preventing attribute disclosure. Considering above example of virus result if a 2-diverse equivalence class has 50% percent of people suffering from that disease then any person who is in this equivalence class has a risk of 50% having that disease, which is 1% compared to overall table which also leads to high risk of privacy breach. Now we make an equivalence class with 50 members and out of which 49 are positive and 1 negative then 98% of people have result as positive while overall table has just 1%. Second if we make an equivalence class which has 49 negative records and 1 positive record then both the classes will satisfy 2-diverse l-diversity. But both the classes will have different levels of privacy as former will have higher risk of privacy breach and later will have comparatively very less possibility.

Similarity Attack: Attack in which values of an attribute are distinct but similar in meaning or semantics. Consider an example as shown in table 6[3]. In given example if a attacker knows that the person he is looking for lives in zip code 47653 and is of age 25 then the attacker will know the person belong to equivalence class one and from this he can infer the person has stomach related disease. He can also infer person's salary lies in range of 3K-5K.

Table 6. 3-diverse table of medical record.



	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

C. t-closeness : For an equivalence class to have t-closeness to make sure the distance between an equivalence class and overall table for a sensitive attribute should not be more than a threshold t. For a table to have t-closeness all the equivalence class should have t-closeness.the value of t is measured by formula of Earth Movers Distance (EMD). There are two other distance measures also available called Kullback -Leibler and variation distance but they fail to consider semantic distance hence EMD is used[4]. EMD takes into account the minimum amount of work done which is required to move a piece of earth into hole.

Limitations of t-closeness:

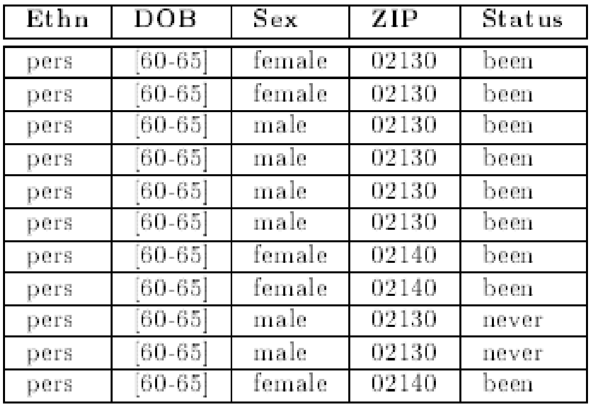
Different sensitive values require different levels of protection. t-closeness lacks in providing that flexibility.In case of numerical sensitive attributes it does not prevent from attribute linkage attack[5].It also degrades the utility has because distribution sensitive attribute should be same in all equivalence classes.

III. ANONYMIZATION TECHNIQUES

A. Generalization

In this anonymization technique a specific value of a quasi-identifier is replaced by a general value which is less specific but maintains the semanticity of value which further makes it difficult for the adversary to directly identify the targeted record. Table 7[1] shows an example of generalization whee quasi identifiers such as age and ethnicity is generalized to generic range and value.

Table 7 : Generalization

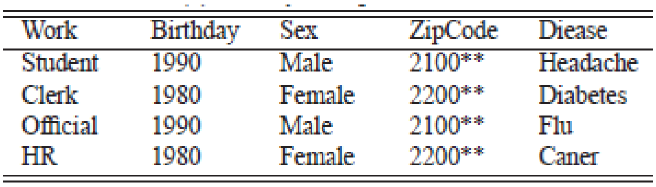


Ethn	DOB	Sex	ZIP	Status
pers	[60-65]	female	02130	been
pers	[60-65]	female	02130	been
pers	[60-65]	male	02130	been
pers	[60-65]	male	02130	been
pers	[60-65]	male	02130	been
pers	[60-65]	male	02130	been
pers	[60-65]	female	02140	been
pers	[60-65]	female	02140	been
pers	[60-65]	male	02130	never
pers	[60-65]	male	02130	never
pers	[60-65]	female	02140	been

B. Suppression

A technique in which a value is replaced by any special character like '*' or any other character so that the replaced value does not gets disclose. Table 8 [6] shows and example of suppression in which zip code field is suppressed to two digits. Various authors have proposed various techniques for suppression. It can be done replacing a whole record or by suppressing a given value in data set. As per requirement we can also suppress the complete attribute or if required than for particular value could suppress only a cell [6].

Table 8. Suppression example.

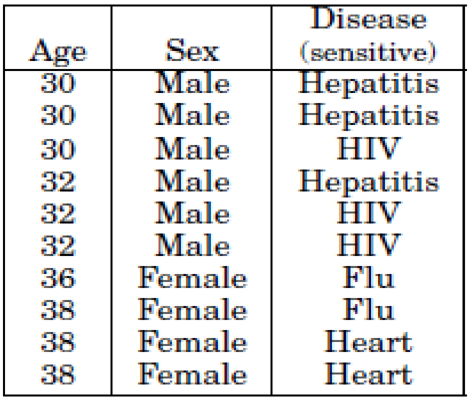


Work	Birthday	Sex	ZipCode	Diease
Student	1990	Male	2100**	Headache
Clerk	1980	Female	2200**	Diabetes
Official	1990	Male	2100**	Flu
HR	1980	Female	2200**	Caner

C. Anatomization

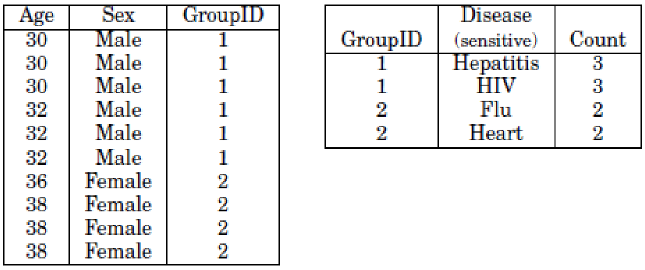
It is a technique in which the table is divided into two tables. One table contains quasi identifier and other table contain sensitive attribute. These both tables are linked together using a common group id. As the data is not changed or replaced it is more beneficial for data miners. Table 9(1) a table is divided into two tables with common group id in table 9(2) [7].

Table 9 : Anatomization Example



Age	Sex	Disease (sensitive)
30	Male	Hepatitis
30	Male	Hepatitis
30	Male	HIV
32	Male	Hepatitis
32	Male	HIV
32	Male	HIV
36	Female	Flu
38	Female	Flu
38	Female	Heart
38	Female	Heart

[a]



Age	Sex	GroupID
30	Male	1
30	Male	1
30	Male	1
32	Male	1
32	Male	1
32	Male	1
36	Female	2
38	Female	2
38	Female	2
38	Female	2

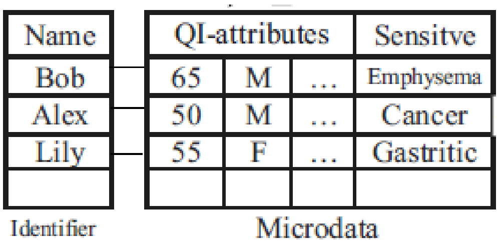
GroupID	Disease (sensitive)	Count
1	Hepatitis	3
1	HIV	3
2	Flu	2
2	Heart	2

[b]

D. Permutation

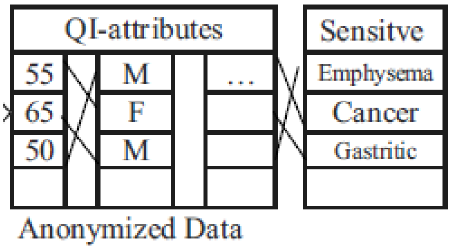
It is an advanced approach of anatomization. It shuffle attributes present in a quasi identifier group. This provides a stronger privacy than anatomization. Table10(1) shows an example of permutation[9]. Here Bob is 65 years male suffering from Emphysema, Alex is 50 year male suffering from Cancer and Lily is 55 year female suffering from Gastritic. Table10(2) shows the result after permutation is applied to table where Quasi identifier i.e. age and gender are shuffled. Now it will be difficult for the adversary to correlate the correct record.

Table 10 : Permutation Example.



Name	QI-attributes			Sensitve
Bob	65	M	…	Emphysema
Alex	50	M	…	Cancer
Lily	55	F	…	Gastritic

[a]



QI-attributes			Sensitve
55	M	…	Emphysema
65	F		Cancer
50	M		Gastritic

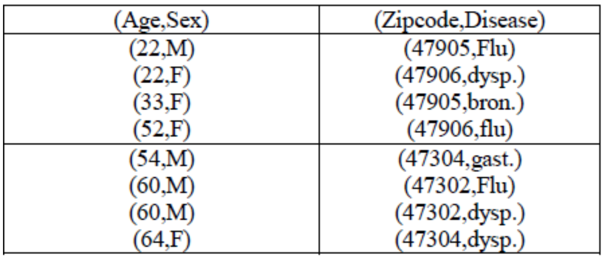
[b]

E. Slicing

A method called slicing was introduced based on vertical and horizontal partitioning of data. Vertical partitioning includes the attributes are grouped into columns, this grouping is done based on the correlation between attributes.

In horizontal partitioning phases the data is divided into buckets trying to put highly correlated data together. Table 11 shows example of slicing after horizontal and vertical partitioning of data[11].

Table 11 : Slicing example.



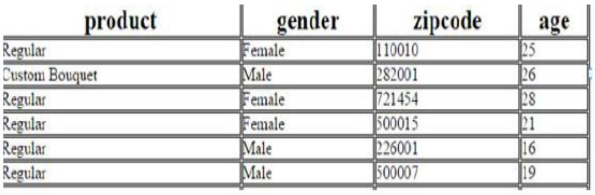
(Age,Sex)	(Zipcode,Disease)
(22,M) (22,F) (33,F) (52,F)	(47905,Flu) (47906,dysp.) (47905,bron.) (47906,flu)
(54,M) (60,M) (60,M) (64,F)	(47304,gast.) (47302,Flu) (47302,dysp.) (47304,dysp.)

Slicing hence breaks the correlation between cross columns and maintains the correlation inside each column [10].

F. Perturbation

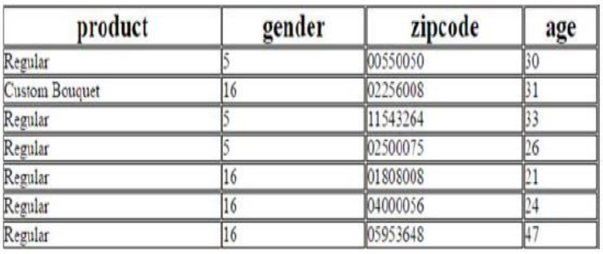
It is technique in which a data is replaced by some other synthetic data by applying some noise to the data. The data modification is done in a manner in which the statistical of data remains unchanged. Table 12(1) shows original data set where table 12(2) shows a data set in which zip code , age and gender has been applied some noise and values has been replaced[12].

Table 12: Perturbation Example



product	gender	zipcode	age
Regular	Female	110010	25
Custom Bouquet	Male	282001	26
Regular	Female	721454	28
Regular	Female	500015	21
Regular	Male	226001	16
Regular	Male	500007	19

[a]



product	gender	zipcode	age
Regular	5	00550050	30
Custom Bouquet	16	02256008	31
Regular	5	11543264	33
Regular	5	02500075	26
Regular	16	01808008	21
Regular	16	04000056	24
Regular	16	05953648	47

[b]

Techniques	Generalization	Suppression	Anatomization	Permutation	Slicing	Perturbation
Advantages	This is useful as it gives different schemes for generalization	Offers strong privacy as hides the data where there is risk of information leak	Does not alters the data hence very efficient for data miners	Overcomes weakness of anatomization by shuffling the data without modifying it	Reduces the dimensionality of data and maintain the correlation between attribute which is good for utility.	It preserves the statistical information even after being modified.
Attacks/ Limitations	The more generalized data the more information loss and has of linkage attack.	High in information loss	Dividing data into two table does not provide good privacy.	Shuffling of data will remove the correlation in attributes.	It is prone to linkage of data types of attacks.	Just adding noise won't be good for privacy and will lead to high information loss.

IV. CONCLUSION

Thus with this paper we tried to represent importance of privacy models and techniques to achieve them. In data publishing we understood the fact that to protect the data from unsolicitated disclosure of information we need to protect them and maintain their privacy. To achieve this privacy there is a considerable amount of loss in information, so we need a system where we can achieve maximum privacy without much loss in information. This paper discusses various privacy models which show the how achieving privacy model can protect from specified attacks. It also shows the drawbacks of each model and also shows which attacks are still possible on those models. This paper shows a stepwise evolution of privacy model and why the later models were formed. This paper also discusses the anonymization techniques available and through which we can achieve different privacy models. Each anonymization technique has some pros and some cons which has also been discussed in this paper.

V. REFERENCES

[1]. P. Samarati and L. Sweeney "Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through

Generalization and Suppression" Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[2]. Ashwin Machanavajjhala ,Johannes Gehrke and Daniel Kifer " l-Diversity: Privacy Beyond k-Anonymity" 22nd International Conference on Data Engineering (ICDE'06) pp. 24-35, April 2006.

[3]. Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity" 2007 IEEE 23rd International Conference on Data Engineering pp. 106-115, April 2007.

[4]. B.Santhosh Kumar and K.V.Rukmani "Novel Privacy notion t-closeness: Privacy Preserving Data Mining" NCACEIT'11 pp. 1-5 , January 2011

[5]. Avinash Kumar Singh, Narayan P. Keer and Anand Motwani "A Review of Privacy Preservation Technique" International Journal of Computer Applications vol. 90 pp.17-20, February 2014.

[6]. Yang Xu,Tinghuai Ma , Meili Tang and Wei Tian

[7]. "A Survey of Privacy Preserving Data Publishing using Generalization and Suppression" Applied Mathematics & Information Sciences An International Journal vol.8 pp.1103-1116, May 2014.

[8]. Benjamin C. M. Fung, Ke Wang, Rui Chen and Philip S. Yu "Privacy-Preserving Data Publishing: A Survey of Recent Developments" ACM Computing Surveys vol. 42 pp. 1-53 June 2010.

[9]. Xianmang He, Yanghua Xiao, Yujia Li and Qing Wang "Permutation Anonymization: Improving Anatomy for Privacy Preservation in Data Publication" New Frontiers in Applied Data Mining PAKDD pp. 111-123 , May 2011

[10]. Dong Li, Xianmang He, LongBin Cao and Huahui Chen "Permutation anonymization" Journal of Intelligent Information Systems vol. 47, pp.427-445 August 2015.

[11]. Tiancheng Li, Ninghui Li, Jian Zhang and Ian Molloy "Slicing: A New Approach to Privacy Preserving Data Publishing" IEEE Transactions on Knowledge and Data Engineering.vol. 24 pp. 561-574 March 2012.

[12]. Vijay R. Sonawane, Kanchan S. Rahinj "A New Data Anonymization Technique used For Membership Disclosure Protection" International Journal of Innovative Research in Science, Engineering and Technology vol. 2 pp.1230-1233, April 2013

[13]. Arshveer Kaur "A Hybrid Approach of Privacy Preserving Data Mining using Suppression and Perturbation Techniques" International Conference on Innovative Mechanisms for Industry Applications (ICIMIA 2017) pp. 306-311, February 2017.

[14]. P. Samarati. "Protecting respondent's privacy in microdata release" IEEE Transactions on Knowledge and Data Engineering, vol. 13 pp. 1010–1027, November 2001.

Cite this article as :

Shipra Sharma, Naveen Choudhary, Kalpana Jain, "A Study on Models and Techniques of Anonymization in Data Publishing", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), ISSN : 2456-3307, Volume 6 Issue 2, pp. 84-90, March-April 2019. Available at doi : https://doi.org/10.32628/IJSRSET19629
Journal URL : http://ijsrset.com/IJSRSET19629