Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
SURVEY REPORT
Available Online at www.ijarcs.info

Preserving Privacy during Big Data Publishing using K-Anonymity Model – A Survey

Divya Sadhwani
Department of Computer Science & Engg.
UIT-RGPV
Bhopal, India

Dr. Sanjay Silakari
Department of Computer Science & Engg.
UIT-RGPV
Bhopal, India

Mr. Uday Chourasia
Department of Computer Science & Engg.
UIT-RGPV
Bhopal, India

Abstract: The advancements in technology have led to a massive increase in the amount of data that is generated now a days. This huge amount of data that is generated from multiple sources is known as big data. This big data is used for analysis by Businesses, Healthcare Organizations, Government, etc. As big data also contains user-specific information, directly releasing this data for analysis can pose serious threats to user's privacy. So to preserve privacy during big data publishing, Anonymization techniques are used. Anonymization use generalization and suppression techniques to preserve privacy of an individual. There are many privacy models, like k-anonymity, l-diversity, t-closeness, that are used to anonymize the data. There are many algorithms that are used to implement k-anonymity model. This survey paper will first explain privacy models that are used to anonymize the data and then gives an overview of the various algorithms that are used to implement k-anonymity.

Keywords: Big Data, Anonymization, Generalization, Suppression, k-anonymity

I. INTRODUCTION

Due to recent technological developments, a continuous growth in data generation has been observed. This data is generated from multiple sources such as social networking sites, healthcare applications, sensors, etc. This huge amount of data that is generated from multiple sources in multiple formats with very high speed is known as big data. There are five characteristics of big data that are usually reflected by 5 V's, viz., Volume, Velocity, Variety, Value and Veracity [1]. This big data is mainly used for analysis purposes by Businesses, Governments, Healthcare Organizations, etc. Big data also aid in scientific research. Big data contains browsing histories, data collected from laptops and mobile phones, data obtained from healthcare organizations, etc. This data also contains personal information. As big data is mainly used for analysis purposes, directly releasing big data to the third parties can pose a serious threat to the privacy of an individual. For e.g., Amazon, Flipkart can learn our shopping preferences. Google too makes a record of our browsing history. YouTube also recommends videos to its users based on their search history and watched history. Big data enables organizations to gather personal information of users and use it for their own profit [2]. So privacy of an individual should be preserved before big data is published to third parties.

II. BIG DATA PRIVACY

i. PRIVACY OBJECTIVES

There are mainly two privacy objectives that should be achieved when data is anonymized [4]:

1.UNIQUE IDENTITY DISCLOSURE: If data is published then there should not be any record that can identify an individual.

2. SENSITIVE ATTRIBUTE DISCLOSURE: Attackers won't be able to learn about sensitive attribute of an individual via disclosed attributes.

ii. TYPES OF ATTRIBUTES

A record in a data set consists of four types of attributes:

1. IDENTIFIERS: These are the attributes which can directly identifyan individual. So these attributes are removed before publishing the data. For e.g. name, social security number, etc.
2. QUASI-IDENTIFIERS: These are the attributes which cannot identify an individual directly but if they are linked with publicly available data then they can identify an individual easily. For e.g. zip code, age, sex, etc.
   An Equivalence Class is a set of records that have same value for all the quasi-identifier attributes.
3. SENSITIVE ATTRIBUTES: These are the attributes which an individual wants to hide from others. For e.g. disease.
4. NON-SENSITIVE ATTRIBUTES: Any attributes other than the above three mentioned are known as Non-Sensitive Attributes.

iii. LINKING ATTACK

Before publishing data, Identifiers (Name of the individual, Social Security Number) are removed. But still there are many attributes (quasi-identifiers) that if combined with external data can identify an individual easily. For e.g., suppose table 1 has to be released for further analysis.

TABLE 1: ORIGINAL DATASET

NAME	AGE	SEX	ZIP CODE	DISEASE
SMITH	25	M	13001	FLU
CLARK	25	M	12057	HIV
DAVID	26	M	14599	BRONCHITIS
ANA	26	F	13001	PNEUMONIA
ROSY	27	F	17000	HEPATITIS

Before releasing table 1, Identifiers (in this case Name attribute) must be removed from thetable.

TABLE 2: DATASET AFTER REMOVING IDENTIFIERS

AGE	SEX	ZIP CODE	DISEASE
25	M	13001	FLU
25	M	12057	HIV
26	M	14599	BRONCHITIS
26	F	13001	PNEUMONIA
27	F	17000	HEPATITIS

Now suppose there is an external data which is available to the attacker. Following table shows the external data which is a Voter Registration List available to the attacker.

TABLE 3: PUBLICLY AVAILABLE DATASET

NAME	AGE	SEX	ZIP CODE
DAVID	26	M	14599
ANA	26	F	13001
CLARK	25	M	12057
SMITH	25	M	13001
ROSY	27	F	17000

On comparing Table 2 and Table 3, the attacker will get to know that Clark is suffering from HIV. So even after removing the identifiers, an individual can be re-identified with the help of data available publicly.

Combining data of the released table with the data of the publicly available table is known as LINKING ATTACK.

iv. ANONYMIZATION

Data Anonymization is a technique which is used to preserve privacy when big data is published to third parties. Anonymization refers to hiding sensitive and private data of the users. Anonymization can make use of many techniques, viz. generalization, suppression, perturbation, anatomization and permutation [2]. Mostly generalization and suppression techniques are used for anonymizing data because data anonymized using generalization and suppression still have high utility. So this data can be used further by researches [4]. Generalization refers to replacing a value with more generic value. For e.g., dancer can be replaced with artist. Suppression refers to hiding the value by not releasing it at all. The value is replaced by a special character, e.g. @, *. Both generalization and suppressionresults in loss of information. Generalization impacts all the tuples while suppression impacts a single tuple [14].

v. PRIVACY MODELS

There are many privacy models that are used to prevent attacks on the privacy of the published data, viz. k-anonymity, l-diversity, t-closeness.

1. **k-anonymity**: This privacy model is used to prevent Linking Attacks. Sweeney and Samarati proposed the k-anonymity principle [9]. According to k-anonymity principle, a tuple in the published data set is indistinguishable from k-1 other tuples in that data set. Therefore, an attacker who knows the values of quasi-identifier attributes of an individual is not able to distinguish his record from the k-1 other records [3]. k-anonymity uses generalization and suppression techniques to hide the identity of an individual [4]. For e.g., Table 4 is 2-anonymous table, i.e., two tuples have same values in the quasi-identifier attributes (in this case, Age, Sex and Zip Code).

TABLE 4: 2-ANONYMOUS TABLE

AGE	SEX	ZIP CODE	DISEASE
[20-40]	M	18***	HIV
[20-40]	M	18***	HIV
[41-50]	F	120**	CANCER
[41-50]	F	120**	HEART DISEASE

Although k-anonymity can solve the problem of identity disclosure attack, it cannot solve the problem of attribute disclosure attack. For e.g., if the sensitive attribute lack diversity in values and attacker is only interested in knowing the value of sensitive attribute then the aim of attacker is achieved. This type of attack is known as Homogeneity Attack.

For e.g., if an attacker has Table 5 available as an external data, then he can link the table 4 and table 5 and can come to a conclusion that Andrew is suffering from HIV.

TABLE 5: EXTERNAL DATA AVAILABLE TO AN ATTACKER

NAME	AGE	SEX	ZIP CODE
ANDREW	31	M	18601
CLARKE	27	M	18555
ROSY	49	F	12001
ANA	42	F	12456

Another kind of attack which k-anonymity cannot prevent is Background Attack. This model assumes that attacker has no additional background knowledge. Suppose, if attacker knows that Ana has low chance of Cancer, then after combining table 4 and 5, attacker can conclude that Ana is suffering from a heart disease.

2. **l-diversity**: To prevent attribute disclosure attack, it was the next privacy model which was proposed. According to l-diversity model, an equivalence class must have l "well-represented" values for sensitive attributes. It is also known as Distinct l-diversity. For e.g., following table is 2-diverse, i.e., each equivalence class contains two distinct values for sensitive attributes.

TABLE 6: 2-DIVERSE TABLE

AGE	SEX	ZIP CODE	DISEASE
[21-30]	M	140**	FLU
[21-30]	M	140**	BRONCHITIS
[31-50]	F	17***	PNEUMONIA
[31-50]	F	17***	HIV

Distinct l-diversity model suffers from probabilistic inference attacks. For e.g. consider the following table.

TABLE 7: TABLE FOR EXPLAINING PROBABILISTIC ATTACK

AGE	SEX	ZIP CODE	DISEASE
[21-30]	M	120**	HIV
[21-30]	M	120**	FLU
[21-30]	M	120**	FLU
[21-30]	M	120**	PNEUMONIA
[21-30]	M	120**	FLU
[21-30]	M	120**	FLU
[21-30]	M	120**	FLU

There is only one equivalence class in table 7. The table is 3-diverse because it contains three distinct values in the sensitive attribute Disease. But five out of seven records contain Flu in the Disease attribute. So an attacker can affirm that disease of target person is Flu with accuracy of 70%.

There is other version of l-diversity which is known as Entropy l-diversity. According to Entropy l-diversity, entropy of the distribution of values of sensitive attributes in each equivalence class should be at least log (l).

l-diversity model is difficult to achieve. Moreover, this model also suffers from skewness and similarity attacks.

Consider following table for understanding the concept of skewness attack. There are two sensitive attributes Salary and Disease.

TABLE 8: EXAMPLE TABLE FOR L-DIVERSITY

AGE	ZIP CODE	SALARY	DISEASE
24	12889	2K	GASTRIC ULCER
26	12110	3K	GASTRITIS
28	12005	4K	STOMACH CANCER
31	15601	6K	FLU
33	15666	7K	BRONCHITIS
37	15689	9K	CANCER
43	19123	11K	HEART DISEASE
45	19765	12K	GASTRITIS
49	19303	14K	PNEUMONIA

Following table is the 3-diverse version of the above table.

TABLE 9: 3-DIVERSE VERSION OF TABLE 8

AGE	ZIP CODE	SALARY	DISEASE
[21-30]	12***	2K	GASTRIC ULCER
[21-30]	12***	3K	GASTRITIS
[21-30]	12***	4K	STOMACH CANCER

[31-40]	156**	6K	FLU
[31-40]	156**	7K	BRONCHITIS
[31-40]	156**	9K	CANCER
[41-50]	19***	11K	HEART DISEASE
[41-50]	19***	12K	GASTRITIS
[41-50]	19***	14K	PNEUMONIA

Now suppose that attacker knows that Alice has low salary (2K-4K). Then he can conclude that Alice is suffering from some stomach disease. This is known as Similarity Attack because there is some kind of similarities in the values of sensitive attribute Disease [16].

Now consider following table to understand concept of skewness attack. Suppose there are 1,00,000 records of a virus and that virus attacks only 1% of the population. Third equivalence class consists of equal number of positive and negative records. In other words, everyone in that class has 50% chance of having the class which is much higher than the real distribution [16].

TABLE 10: TABLE TO EXPLAIN SKEWNESS ATTACK

AGE	ZIP CODE	SALARY	DISEASE
[11-20]	12***	2K	NEGATIVE
[11-20]	12***	3K	NEGATIVE
[21-30]	156**	4K	NEGATIVE
[21-30]	156**	6K	NEGATIVE
[31-40]	19***	7K	NEGATIVE
[31-40]	19***	9K	POSITIVE
[41-50]	170**	11K	NEGATIVE
…	…	…	…
[81-90]	170**	12K	NEGATIVE

3. **t-closeness**: This model overcomes the weaknesses of l-diversity model. According to this model, distribution of values of sensitive attribute in each equivalence class must be close to that of the overall dataset.

Information gain of the attacker is the measure of the privacy. Before the table is released, attacker has some prior belief B_0 about the value of the sensitive attribute of an individual. Then the attacker's belief is influenced by Q, the distribution of the value of sensitive attribute in the whole table. This is posterior belief of the attacker and is denoted by B_1. This Q is the public information. Finally the anonymized table is given to the attacker. As attacker knows the quasi-identifier values of the individual, he can simply identify the equivalence class to which the individual belongs. Then the attacker learns the distribution P of value of sensitive attribute in that equivalence class. Now the belief of the attacker changes to B_2. We cannot limit the gain between B_0 and B_1 but we can limit the gain from B_1 to B_2 by limiting the distance between P and Q. Requiring P and Q to be close decreases the utility of the information [16].

III. k-ANONYMITY ALGORITHMS

1. DATAFLY ALGORITHM

This algorithm is the first practical implementation of k-anonymity model. This algorithm achieves k-anonymity by using full domain generalization. Full domain generalization generalize all attribute values to the same level. For e.g., if attribute value is 12345, then this value will be generalized to 123** in all its occurrences in the table.Datafly algorithm follows some steps to preserveprivacy. First of all it constructs the frequency of all theunique values in quasi-identifier attributes. It also stores the number of occurrence of each sequence. Then it starts generalizing data starting with the attribute which has highest frequency. It performs generalization recursively until required level of k or less tuples have distinct sequences in frequency. At the end, the algorithm suppresses those tuples which have frequency of less than k [4].

Datafly algorithm results in loss of information and this is the biggest disadvantage of Datafly algorithm.

TABLE 11: ORIGINAL DATASET FOR DATAFLY ALGORITHM

BIRTH DATE	SEX	ZIP CODE	NO. OF OCCURS	TUPLE NO.
12/01/1984	M	4601	1	T1
04/04/1988	F	4888	1	T2
19/09/1989	F	4601	1	T3
27/02/1990	M	4700	1	T4
13/03/1984	M	4601	1	T5
17/05/1990	M	4700	1	T6
6	2	3		

Following table shows the tuples that are grouped together based on the value of quasi-identifier attribute. Domain generalization is done on birthdate attribute.

TABLE 12: GENERALIZED TABLE

BIRTH DATE	SEX	ZIP CODE	NO. OF OCCURS	TUPLE NO.
1984	M	4601	2	T1, T5
1990	M	4700	2	T4, T6
1988	F	4888	1	T2
1989	F	4601	1	T3

Following table shows the final 2-anonymous table. As the last two rows of table 12 have frequency less than 2 so these two rows are removed from the final published table.

TABLE 13: FINAL OUTPUT OF DATAFLY ALGORITHM

BIRTH DATE	SEX	ZIP CODE
1984	M	4601
1984	M	4601
1990	M	4700
1990	M	4700

2. **µ-ARGUS**

This is the second implementation of k-anonymity. It also makes use of generalization and suppression techniques to anonymize the data. It assign values to each of the attributes in the table. The values which are assigned are in between 0 and 3 and corresponds to not identifying, most identifying, more identifying and identifying. Then it makes a combination by testing 2 and 3 combinations of attributes. The combinations which are not safe are eliminated by generalizing attributes within combinations and by cell suppression. It does not erase the entire tuples as is done by Dataflyalgorithm. It suppresses the value at the cell level. So, the output of µ-Argus contains more tuples as compared to Datafly algorithm [4], i.e. µ-Argus results in less data distortion. But µ-Argus may not always provide k-anonymity because it only tries 2 and 3 combinations of attributes [15].

3.**OPTIMAL K-ANONYMITY**

It is one of the practical method of k-anonymity model which determines optimal anonymization of a given dataset. An optimal anonymization is one which anonymize the data as little as possible, so as to minimize the information loss. Datafly algorithm and µ-Argus starts from the original dataset and then generalize the dataset greedily so as to obtain k-anonymous dataset. Unlike Datafly algorithm and µ-Argus, Optimal k-anonymity starts with the fully generalized dataset and then it specializes the dataset to obtain optimal k-anonymous dataset, i.e., It starts with the most suppressed value and then it generalizes the value which minimizes suppression and information loss [4].

4. INCOGNITO

This algorithm exploits the monotonicity property regarding the frequency of tuples in the lattice. There are two monotonicity properties:

I. GENERALIZATION PROPERTY (ROLLUP) - According to this property, if at some level k-anonymity holds then it also holds for any ancestor nodes.
II. SUBSET PROPERTY (APRIORI) – According to this property, if for a set of quasi-identifier attributes k-anonymity does not hold then it does not hold for any of its superset.

Incognito algorithm starts by checking single attribute subsets of the quasi-identifiers. It then iterates and checks k-anonymity with respect to increasingly large subsets [12].

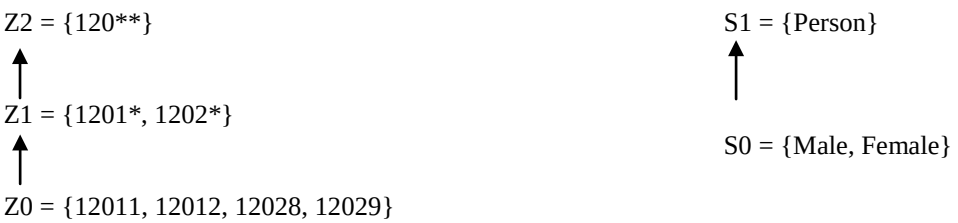

FIGURE 1: DOMAIN GENERALIZATION HIERARCHIES

Figure 1 [12] shows the domain generalization hierarchies for zip code and sex.

Incognito algorithm starts with one dimensional quasi-identifiers as shown below.

FIGURE 2: STEP 1 OF INCOGNITO ALGORITHM

Then it moves to two dimensional marginal as shown below.

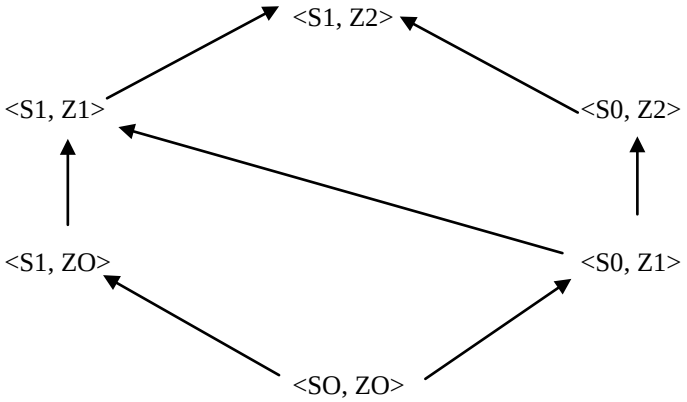

FIGURE 3: STEP 2 OF INCOGNITO ALGORITHM

5. MONDRIAN ALGORTIHM

Mondrian is a multidimensional k-anonymity algorithm. This algorithm is very fast, scalable and produces better results. This algorithm uses strict partitioning and relaxed partitioning methods which further results in better data utility [4]. Partitioning technique maps each tuple of the dataset into a multidimensional space. Then the generalization of the dataset equals to the partitioning of the corresponding multidimensional space. A partition of the multidimensional space corresponds to unique anonymization result. If partitions are not intersecting or overlapping with one another then it is known as Strict Partitioning. And if the partitions are overlapping with one another then it is known as Relaxed partitioning [9]. Relaxed Partition is much better than Strict Partition. For e.g., suppose we have to strictly partition table in figure 4 [9] into at least two regions then at least there would be one region which will contain no more than two tuples. Therefore, if 3-anonymity or more is required then there would be no strict partition based algorithm which can serve our purpose [9].

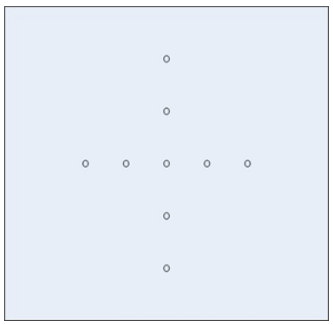

FIGURE 4: DATA IN 2-DIMENSION PLANE

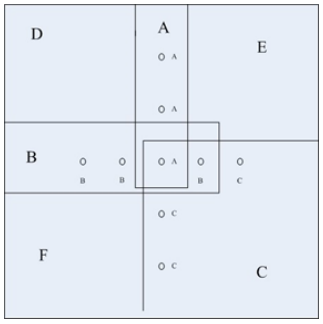

FIGURE 5: RELAXED PARTITION

But we can use relaxed partition based algorithm to partition the table into six regions, namely A, B, C, D, E, F as shown in figure 5 [9]. Region A, B and C contain three tuples while region D, E and F

contain no tuples. Each tuple in the intersection of regions belongs to only one region [9].

Mondrian Algorithm uses two different approaches, viz. Global Recoding and Local Recoding, for generalization and suppression techniques [4].

Datasets which are anonymized using global recoding technique generalizes or suppresses all the attributes equally for all the entries. In other words, a value of an attribute generalizes to another value for all of its occurrences. For e.g., a zip code value 12345 will be generalized to 123** for all of its occurrences. Global Recoding is of two types: Single Dimensional Global Recoding and Multidimensional Global Recoding. The advantage of Global Recoding is that the table that is anonymized contains homogeneous set of values. The disadvantage of Global Recoding is that it results in more information loss.

Datasets which are anonymized using local recoding technique suppresses attributes on per cell basis. In other words, local recoding map individual data values to generalized values. The advantage of Local Recoding is that it results in less information loss. Local Recoding has better utility than Global Recoding [4].

Consider the following tables to understand the concept of global recoding and local recoding.

TABLE 14: TABLE FOR ILLUSTRATING GLOBAL AND LOCAL RECODING

AGE	SEX	ZIP CODE	DISEASE
23	M	37023	CANCER
27	M	37001	FLU
38	F	34001	BRONCHITIS
39	M	34789	FLU

TABLE 15: RESULT OF GLOBAL RECODING ON TABLE 14

AGE	SEX	ZIP CODE	DISEASE
[21-30]	PEOPLE	[34001-37023]	CANCER
[21-30]	PEOPLE	[34001-37023]	FLU
[31-40]	PEOPLE	[34001-37023]	BRONCHITIS
[31-40]	PEOPLE	[34001-37023]	FLU

TABLE 16: RESULT OF LOCAL RECODING ON TABLE 14

AGE	SEX	ZIP CODE	DISEASE
[21-30]	M	[37001-37023]	CANCER
[21-30]	*	[37001-37023]	FLU
[31-40]	*	[34001-34789]	BRONCHITIS
[31-40]	M	[34001-34789]	FLU

Table 14 shows a dataset. Table 15 shows the result of global recoding on the dataset shown in table 14 whereas Table 16 shows the result of local recoding on the dataset shown in table 14.

6. BOTTOM-UP GENERALIZATION

It is one of the ways to fulfill sub-tree anonymization [7]. Sub-tree data anonymization technique is a widely used technique to anonymize the data. In sub-tree anonymization technique, all child values of a non-leaf node in the domain hierarchy are generalized to the node's value [7]. Bottom up generalization changes specific data to less specific data. Data is generalized in such a way that it remains useful for research and analysis. A good generalization should focus on preserving data for future use while achieving k-anonymity [13].

7. TOP-DOWN SPECIALIZATION

Top-down specialization is another way to fulfill sub-tree anonymization. Sub-tree anonymization achieves good trade-off between data utility and information loss [7]. In top-down specialization, a dataset is anonymized by performing specialization operation on it. A specialization operation replaces a value with its child value.

IV. LITERATURE SURVEY

REF. NO	AUTHORS	TITLE OF PAPER	YEAR	OBJECTIVE
[3]	Zakerzadeh, Hessam, Charu C. Aggarwal, and Ken Barker	Privacy-preserving big data publishing	2015	This paper addresses the issue of scalability in privacy algorithms of big data. This paper first introduces two privacy models, namely, k-anonymity and l-diversity. This paper then proposes an algorithm which is based on MapReduce framework and can handle the issue of

				scalability.
[4]	Russom, Yohannes	Privacy preserving for Big Data Analysis	2013	This thesis identify the risk of disclosure of sensitive information of an individual without hiding it. This thesis first studies k-anonymity model which is one of the models needed for preserving privacy. Then it gives an overview of three practical algorithms of k-anonymity, namely, Datafly, µ-Argus and Optimal k-Anonymity. It then studies Mondrian algorithm in detail and later provides practical implementation of it.
[5]	Zhang, Xuyun, et al.	A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud.	2014	This paper focusses on the issue of anonymizing large scale data sets. This paper initially gives a brief overview of cloud computing paradigm and privacy issue in big data. Then it describes some basic terms of anonymization. Then it describes top-down specialization (TDS) algorithm, which is one of the k-anonymity algorithms. This algorithm offers a good tradeoff between data privacy and data utility. Later this paper highlights the issue of scalability in this algorithm. It then proposes a two phase top-down specialization approach for anonymization of large data sets. This approach makes use of MapReduce framework.
[6]	Zhang, Xuyun, et al.	A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud	2013	This paper focuses on the scalability problem in finding the median of values. This paper first gives introduction of data anonymization technique which is used to preserve privacy when data is published to third parties. This paper then discusses about the multidimensional anonymization scheme. This scheme is more flexible and causes less data distortion. Later, it proposes a scalable multidimensional anonymization approach for data stored on cloud.
[7]	Zhang, Xuyun, et al.	Combining top-down and bottom-up: scalable sub-tree anonymization over big data using MapReduce on cloud	2013	This paper first discusses the two ways of fulfilling sub-tree anonymization, namely, top-down specialization (TDS) and bottom-up generalization (BUG). But the existing sub-tree anonymization schemes are facing scalability problem. BUG in not preferred if k is large and TDS is not preferred if k is small. This paper later proposes a hybrid approach by combining TDS and BUG so as to solve the scalability issue.
[8]	LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan	Mondrian multidimensional k-anonymity	2006	This paper first discusses concept of k-anonymity. This paper proposes a multidimensional model. This new model is more efficient and provides higher quality results as compared to other single-dimensional models. Moreover, Multidimensional recoding model can be applied to numerical as well categorical data. This paper explains strict multidimensional partitioning and relaxed multidimensional partitioning. This paper later proposes a greedy algorithm for preserving privacy using k-anonymity privacy model. The proposed algorithm is similar to kd-tree.
[9]	Tang, Qingming, et al.	Improving Strict Partition for Privacy Preserving Data Publishing	2010	This paper studies partition based algorithms. Partition based algorithms are used to preserve privacy of data sets. Partition algorithms map each tuple of the dataset into a multidimensional space. Partitions are of two types, namely, strict partition and relaxed partition. As strict partition based algorithms results in high information loss, so this paper proposes a hybrid approach which further partitions a region generated by strict partitioning algorithm into smaller

				intersecting regions.
[10]	Basu, Anirban, et al.	k-anonymity: Risks and the Reality	2015	This paper focusses on the concept of k-anonymity. k-anonymity is a well-known technique of privacy preserving data publishing. K-anonymity reduces the worst case probability of re-identification of individual based on quasi-identifiers to 1/k. This paper later evaluates the risk as background knowledge can increase the probability of re-identification.
[11]	Kavitha, S., S. Yamini, and Raja Vadhana	An evaluation on big data generalization using k-Anonymity algorithm on cloud	2015	This paper first studies data anonymization which is a technique to preserve privacy of an individual while data publishing. Later it studies top-down specialization which is one of the algorithms of k-anonymity. As scalability is one of the challenges of big data, this paper later studies two phase top-down specialization technique.
[12]	LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan	Incognito: Efficient full-domain k-anonymity	2005	This paper begins with the introduction of joining or linking attacks. Then it provides definitions of some basic terms. Then this paper explains domain generalization and value generalization. Then this briefly describes already existed full domain generalization algorithms. Then it describes a new algorithm Incognito which is based on roll-up subset property. Incognito algorithm is based on full domain generalization.
[13]	Wang, Ke, Philip S. Yu, and Sourav Chakraborty	Bottom-up generalization: A data mining solution to privacy protection	2004	The paper begins by explaining the privacy problem and the generalization technique. This paper evaluates bottom-up generalization which is one of the algorithm for achieving k-anonymity. Bottom up generalization generalizes a value to a less specific but consistent value in order to preserve privacy of an individual.
[14]	Zhu, Yan, and Lin Peng	Study on k-anonymity models of sharing medical information	2007	This paper first analyses the need of sharing of data among organizations. It then illustrates linking attack. It then formulates a new anonymity model, namely k-anonymity, which makes use of generalization and specialization technique to preserve privacy of an individual. Later it formulates l-diversity anonymity model which is an extension of k-anonymity model.
[15]	Sweeney, Latanya	Achieving k-anonymity privacy protection using generalization and suppression	2002	This paper first talks about k-anonymity. Later it gives an overview of MinGen algorithm which combines generalization and suppression techniques to achieve k-anonymity. MinGen is a theoretical algorithm. This paper later compares MinGen to Datafly and µ-Argus. Both Datafly and µ-Argus are the practical implementation of k-anonymity.

V. CONCLUSION

In this survey paper, a brief overview of privacy preservation in big data publishing is presented. Privacy of an individual should be preserved before big data is published to a third party because big data also consists of user-specific information. Description of the three main privacy models, namely, k-anonymity, l-diversity and t-closeness is given in this survey paper. Then algorithms which are used to implement k-anonymity model are also explained.

VI.REFERNCES

[1] Gahi, Youssef, MouhcineGuennoun, and Hussein T. Mouftah. "Big Data Analytics: Security and privacy challenges." *Computers and Communication (ISCC), 2016 IEEE Symposium on*. IEEE, 2016.

[2] Mehmood, Abid, et al. "Protection of big data privacy." *IEEE access* 4 (2016): 1821-1834.

[3] Zakerzadeh, Hessam, Charu C. Aggarwal, and Ken Barker. "Privacy-preserving big data publishing." *Proceedings of the 27th International Conference on Scientific and Statistical Database Management*. ACM, 2015.

[4] Russom, Yohannes. *Privacy preserving for Big Data Analysis*. MS thesis. University of Stavanger, Norway, 2013.

[5] Zhang, Xuyun, et al. "A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud." *IEEE Transactions on Parallel and Distributed Systems* 25.2 (2014): 363-373.

[6] Zhang, Xuyun, et al. "A MapReduce based approach of scalable multidimensional anonymization for big data privacy preservation on cloud." *Cloud and Green Computing (CGC), 2013 Third International Conference on*. IEEE, 2013.

[7] Zhang, Xuyun, et al. "Combining top-down and bottom-up: scalable sub-tree anonymization over big data using MapReduce on cloud." *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on*. IEEE, 2013.

[8] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Mondrian multidimensional k-anonymity." *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*. IEEE, 2006.

[9] Tang, Qingming, et al. "Improving Strict Partition for Privacy Preserving Data Publishing." *Networking and Distributed Computing (ICNDC), 2010 First International Conference on*. IEEE, 2010.

[10] Basu, Anirban, et al. "k-anonymity: Risks and the Reality." *Trustcom/BigDataSE/ISPA, 2015 IEEE*. Vol. 1. IEEE, 2015.

[11] Kavitha, S., S. Yamini, and Raja Vadhana. "An evaluation on big data generalization using k-Anonymity algorithm on cloud." *Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on*. IEEE, 2015.

[12] LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishnan. "Incognito: Efficient full-domain k-anonymity." *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005.

[13] Wang, Ke, Philip S. Yu, and Sourav Chakraborty. "Bottom-up generalization: A data mining solution to privacy protection." *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE, 2004.

[14] Zhu, Yan, and Lin Peng. "Study on k-anonymity models of sharing medical information." *Service Systems and Service Management, 2007 International Conference on*. IEEE, 2007.

[15] Sweeney, Latanya. "Achieving k-anonymity privacy protection using generalization and suppression." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 571-588.

[16] Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007.

