ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 7, Issue 3, pp: (150-156), Month: July - September 2019, Available at: www.researchpublish.com

K-Anonymity: Tool for preserving Privacy

Shrinkhala Shinghai [1], Somesh Dewangan [2,]Rahul Mishra[3]

[1]M.Tech Scholar RSR-RCET, Bhilai, Chhattisgarh,India

[2,3]Assistant Professor RSR-RCET, Bhilai, Chhattisgarh,India

Abstract: **The boost in data storage capacity and computing power due to advancement in cloud computing and big data have extended the reach of information to the third party such as census data, medical data, and personal consumption data. However, privacy protection is one of the most concerned issues in big data and cloud. As big data and clouds hold user-specific information and therefore, we can't directly share an individual's data for analysis as this will lead to threats to user's privacy. Hence, to preserve privacy during big data publishing and cloud computing we can use Anonymization techniques. To anonymize the data, the data privacy models such as k-anonymity, l-diversity, t-closeness are used. This survey paper first describes the privacy models used to anonymize the data and further share the concept of the different algorithms used to implement k-anonymity.**

Keywords: **Anonymization, K-anonymity, l-diversity, t-closeness., k-anonymity algorithm.**

1. INTRODUCTION

In today's era, due to recent technological developments and networked society places great demand on the collection and sharing of person-specific data for many new uses. This happens at a time when more and more historically public information is electronically available. Due to this factor, continuous growth in data generation has been observed. Micro data are being published by many organizations for various purposes such as business, demographic research, public health research, etc. Personal records of people are progressively being collected by numerous government and company establishments and published in the cloud for data analysis. It is facilitated by various organizations to publish sufficiently private ideas over this information that are collected. Now a day's cybercrime is the greatest threat to society. 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. 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 profit. So the privacy of an individual should be preserved before big data is published to third parties. This published data can put the privacy of an individual at risk. To protect the anonymity of the individual, the data holders encrypt or eliminate the explicit identifiers such as name, contact numbers, email address and addresses. However, if we combine other attributes like sex, date of birth, zip code, race, etc with publicly released reports it can be used to determine the anonymous individuals. The generous information that is easily available today, when combined with the heightened computational power available to the attackers, make such attacks a thoughtful problem.

PRIVACY OBJECTIVES

While anonymizing the data, following two privacy objectives should be achieved:

Unique identity disclosure: If data is published then there should not be any record that can identify an individual.

Sensitive attribute disclosure: Attackers won't be able to learn about sensitive attribute of an individual via disclosed attributes.

DATA ANONYMIZATION

Anonymization is a term defined in the Oxford dictionary as 'unknown'. Anonymization refers to hiding sensitive and private data of the users. Anonymization makes an object different from other objects. It can be done by removing personally identifying information (PII) like Name, Social Security number, Phone number, Email, Address, etc.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 7, Issue 3, pp: (150-156), Month: July - September 2019, Available at: www.researchpublish.com

Whenever any data is published to third parties, the Data Anonymization technique is used to preserve the privacy of the data. Anonymization uses many techniques such as generalization, suppression, perturbation, anatomization, and permutation [2]. Often generalization and suppression techniques are used for anonymizing data because data anonymized using generalization and suppression have high utility [4]. Generalization can be defined as replacing a value with more generic value. E.g. we can replace a dancer with an artist. Suppression can be termed as hiding the value by not releasing it at all where the value is replaced by a special character such as @, *. Both generalization and suppression results in loss of information. Generalization impacts all the tuples while suppression impacts a single tuple [14].

2. PRIVACY MODELS

Several privacy models are used to prevent attacks on the privacy of the published data, viz. K-anonymity, l-diversity, t-closeness.

1. K-anonymity: Sweeney and Samarati proposed the K-anonymity principle [9]. This privacy model is used to prevent Linking Attacks. According to the 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 are 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]. E.g., Table 1.1 is a 2-anonymous table, i.e., two tuples have the same values in the quasi-identifier attributes (in this case, Age, Sex and Zip Code).

Although K-anonymity can resolve the problem of identity disclosure attack, it cannot solve the problem of attribute disclosure attack. 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.

TABLE 1.1-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

TABLE 1.2: EXTERNAL DATA AVAILABLE TO AN ATTACKER

NAME	AGE	SEX	ZIP CODE
ALEX	31	M	18601
BOB	27	M	18555
CHRISTY	49	F	12001
ROSE	42	F	12456

E.g., if an attacker has Table 1.2 available as external data, then he can link the Table 1.1 and Table 1.2 and can conclude that ALEX is suffering from HIV.

Another sort of attack which K-anonymity cannot stop is Background Attack. This model believes that an attacker has no additional background knowledge. Suppose, if the attacker knows that Rose has a low chance of Cancer, then after combining Tables 1.1 and 1.2, an attacker can conclude that Rose is suffering from heart disease.

2. l-diversity: Another privacy model that was endeavoured to prevent attribute disclosure attack is l-diversity. According to the l-diversity model, an equivalence class is an obligation to have “well-represented” values for sensitive attributes. It is also recognized as a distinct l-diversity.

One more version of l-diversity is known as Entropy l-diversity. The entropy l-diversity model is defined as "the entropy of the distribution of values of sensitive attributes in each equivalence class should be at least log (l)". The l-diversity model is tough to achieve. Furthermore, this model also suffers from skewness and similarity attacks.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 7, Issue 3, pp: (150-156), Month: July - September 2019, Available at: www.researchpublish.com

TABLE 2: EXAMPLE TABLE FOR L-DIVERSITY

AGE	ZIP CODE	SALARY	DISEASE
24	12882	2K	ULCER
26	12111	3K	HIV
28	12006	4K	BRONCHITIS
31	15604	6K	CANCER
33	15669	7K	FLU
37	15686	9K	STOMACH CANCER

TABLE 2.1-DIVERSE VERSION OF TABLE 2

AGE	ZIP CODE	SALARY	DISEASE
[21-30]	12***	2K	ULCER
[21-30]	12***	3K	HIV
[21-30]	12***	4K	BRONCHITIS
[31-40]	15***	6K	CANCER
[31-40]	15***	7K	FLU
[31-40]	15***	9K	STOMACH CANCER

3. t-closeness: According to this model, the allocation of values of a sensitive attribute in each equivalence class must be close to that of the overall dataset. The knowledge gain by the attacker is the measure of privacy. Before the table is released, the attacker has some prior belief B0 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 a sensitive attribute in the whole table. This is the posterior belief of the attacker and is denoted by B1 anonymized table is given to the attacker. As the 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 the value of a sensitive attribute in that equivalence class. Now the belief of the attacker changes to B2 but we cannot limit the gain between B0 and B1 but we can limit the gain from B1 to B2 by limiting the distance between P and Q. Requiring P and Q to be close decreases the utility of the information [16]. This model overcomes the weaknesses of the l-diversity model.

3. K-ANONYMITY ALGORITHMS

1. DATAFLY ALGORITHM

Datafly algorithm is the first effective enforcement of the K-anonymity model that attains K-anonymity by promoting full domain generalization in which all attribute values are generalized to the same level. E.g., if the attribute value is 45678, then this value will be generalized to 456** in all its events in the table. Datafly algorithm practices some steps to preserve privacy. Firstly, it builds the frequency of all the unique values in quasi-identifier attributes, and then it saves the number of occurrences of each sequence. Then it begins generalizing data starting with the attribute that has the highest frequency. It performs generalization recursively till the required level of k or fewer tuples have distinct sequences in frequency. In the end, the algorithm suppresses those tuples which have a frequency of less than k [4].

The biggest drawback of the Datafly algorithm is that it results in a suppression of information.

The following table explains the tuples that are grouped based on the value of the quasi-identifier attribute. Domain generalization is performed on the birth date attribute.

TABLE 3.1: ORIGINAL DATASET FOR DATAFLY ALGORITHM

BIRTH DATE	SEX	ZIP CODE	NO. OF OCCURS	TUPLE NO.
12/02/1985	M	4600	1	T1
04/05/1988	F	4886	1	T2
19/01/1989	F	4602	1	T3
27/02/1999	M	4701	1	T4
13/03/1985	M	4600	1	T5
17/05/1999	M	4701	1	T6

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 7, Issue 3, pp: (150-156), Month: July - September 2019, Available at: www.researchpublish.com

TABLE 3.2: GENERALIZED TABLE

BIRTH DATE	SEX	ZIP CODE	NO. OF OCCURS	TUPLE NO.
1985	M	4600	2	T1, T5
1999	M	4701	2	T4, T6
1988	F	4886	1	T2
1989	F	4602	1	T3

The resulting table depicts the final anonymous table. As the last two rows of table 3.2 have a frequency smaller than 2 so these two rows are excluded from the final published table.

TABLE 3.3: FINAL OUTPUT OF DATAFLY ALGORITHM

BIRTH DATE	SEX	ZIP CODE
1985	M	4600
1985	M	4600
1999	M	4701
1999	M	4701

2. µ-ARGUS

µ-ARGUS is the second implementation of the K-anonymity algorithm. It practices generalization and suppression methods to anonymizes the data. It allocates values to each of the attributes in the table. The values which are allocated are lies within 0 and 3 and correspond to not identifying, most identifying, more identifying and identifying. Then it makes a sequence of testing 2 and 3 combinations of attributes. The combinations which are not safe are removed by generalizing attributes within combinations and by cell suppression. It does not eliminate the entire tuples as the Datafly algorithm does. It only suppresses the value at the cell level. So, the output of µ-Argus contains more tuples as compared to a 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

To determine optimal anonymization in a given dataset in K-Anonymity this method is applied. Optimal anonymization anonymizes the data as limited as possible to lessen the information loss. Datafly algorithm and µ-Argus begins from the original dataset and then generalize the dataset greedily, to obtain the k-anonymous data set. Unlike the Datafly algorithm and µ-Argus, Optimal K-anonymity starts with the fully generalized dataset and then it specializes the dataset to obtain the 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].

S2 = {120}**

S1 = {1201*, 1202*}

S0 = {12011, 12012, 12028, 12029}

4. INCOGNITO

INCOGNITO algorithm attains the monotonicity property regarding the frequency of tuples in the lattice. The two monotonicity properties are as follows:

- Generalization Property (Rollup) - Generalization property states that, if at some level K-anonymity is obtained, then it also achieved for any ancestor nodes.

• Subset Property (Apriori) – This property states that, if for a set of quasi-identifier attributes K-anonymity does not exist, then it does not hold for any of its supersets.

The incognito algorithm starts by checking the single attribute subsets of the quasi-identifiers. It then iterates and checks K-anonymity concerning increasingly larger subsets. [12].

Z1 = {Person}

Z0 = {Male, Female}

5. MONDRIAN ALGORTIHM

MONDRIAN ALGORITHM is a multidimensional K-anonymity algorithm that is extremely quick, scalable and yields more reliable results. This algorithm utilizes strict partitioning and relaxed partitioning methods which further results in better data utility [4]. The partitioning technique outlines each tuple of the dataset into a multidimensional space. Then the generalization of the dataset is equivalent to the partitioning of the corresponding multidimensional space. A partition of the multidimensional space resembles a unique Anonymization result. If partitions are not intersecting or overlapping with one another than it is known as Strict Partitioning. And if the partitions are overlapping with one another than it is known as relaxed partitioning [9]. Relaxed Partition is far better than Strict Partition.

Two different approaches are used in Mondrian Algorithm i.e., 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

6. BOTTOM-UP GENERALIZATION

It is one of the ways to fulfil 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 fulfil sub-trees 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.

4. LITERATURE SURVEY

Kavitha, S., S. Yamini, and Raja Vadhana "An evaluation on big data generalization using K-anonymity algorithm on a cloud, 2015". This paper shares the concept of data Anonymization. It examines the top-down specialization algorithm. As big data have the obstacle which is scalability, so this paper proposes a two-phase top-down specialization technique.

Basu, Anirban, et al. "K-anonymity: Risks and the Reality, 2015". This paper shares the knowledge of K-anonymity which is a broad method to preserve privacy while data publishing. K-anonymity lowers the probability of re-identification of individuals in worst-case based on quasi-identifiers to 1/k. It also assesses the probability of re-identification because of background knowledge.

Zakerzadeh, Hessam, Charu C. Aggarwal, and Ken Barker "Privacy-preserving big data, 2015".This paper discusses the problem of scalability in privacy algorithms of big data. It introduces two privacy models, particularly, k-anonymity and l-diversity and introduces an algorithm based on the Map Reduce framework and can handle the scalability issue.

Zhang, Xuyun, et al "A scalable two-phase top-down specialization approach for data Anonymization using map-reduce on a cloud,2014 ".This paper deal with large scale data set used for anonymizing by sharing a concise view of the cloud computing paradigm and big data privacy issues. It describes the top-down specialization (TDS) algorithm and some basic terms of Anonymization. This paper emphasizes the issue of scalability in this algorithm. It then introduces a two-phase, top-down specialization approach for the Anonymization of large data sets by using the Map-Reduce framework.

Russom, Yohannes"Privacy-preserving for Big Data Analysis. MS thesis. University of Stavanger, Norway, 2013". This thesis recognizes the risk of disclosure of sensitive data of an individual without hiding it. It provides a summary of three practical algorithms of k-anonymity, particularly, Datafly, µ-Argus, and Optimal k-Anonymity. It also studies the Mondrian algorithm in particular and then presents a practical implementation of it.

Zhang, Xuyun, et al, "Map-Reduce based approach of scalable multidimensional Anonymization for big data privacy preservation on a cloud, 2013". This paper aims to find the median values on the scalability problem. It gives a concise description of various data anonymization techniques that are used to preserve privacy when data is issued to third parties. This paper also explains the multidimensional anonymization scheme as this scheme is more flexible and less data is destroyed. Further, it suggests a scalable multidimensional anonymization approach for data stored in the cloud.

Zhang, Xuyun, et al. "Combining top-down and bottom-up: scalable sub-tree Anonymization over big data using Map-Reduce on a cloud, 2013". This paper explains the two modes of fulfilling sub-tree Anonymization, particularly, top-down specialization (TDS) and bottom-up generalization (BUG). However, the existing sub-tree Anonymization schemes are suffering from scalability problems. We can't use BUG if k is large and TDS if k is small. So, this paper recommends a hybrid strategy to resolve the scalability issue by coupling TDS and BUG.

Tang, Qingming, et al." Improving Strict Partition for Privacy-Preserving Data Publishing, 2010". This paper focuses on partition-based algorithms which are used to preserve the privacy of data sets. In Partition algorithms, each tuple of the dataset mapped into a multidimensional space. Partition algorithms are of two kinds, i.e, strict partition, and relaxed partition. Strict partition-based algorithms lead to a high loss in information, so this paper suggests a hybrid approach. This approach partitions a region created by a strict partitioning algorithm into smaller intersecting regions.

Zhu, Yan, and Lin Peng "Study on K-anonymity models of sharing medical information, 2007". This paper outlines the necessity for the sharing of data amidst organizations. Due to the sharing of data the chance of the linking attack may occur. So, it designs the K-anonymity model in such a way that makes use of generalization and specialization techniques to preserve the privacy of an individual. Further, it describes the l-diversity anonymity model which is an extension of the K-anonymity model.

LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishna "Mondrian multidimensional K-anonymity,2006". This paper explains the theory of K-anonymity and the idea of a multidimensional model. This multidimensional model is more economical and produces better quality results as compared to other single dimensional models. Furthermore, we can use the Multidimensional recoding model to solve numerical as well as categorical data. This paper describes strict multidimensional partitioning and relaxed multidimensional partitioning and introduces a greedy algorithm for preserving privacy using the K-anonymity privacy model. The proposed algorithm is alike to KD-tree.

LeFevre, Kristen, David J. DeWitt, and Raghu Ramakrishna "Incognito: Efficient full domain K-anonymity, 2005". This paper researches various notions such as joining or linking attacks, domain generalization and vague generalization. Further, this paper concisely illustrates already existed full domain generalization algorithms and new algorithm known as Incognito algorithm. An incognito algorithm works on full domain generalization.

Wang, Ke, Philip S. Yu, and Sourav Chakraborty" Bottom-up generalization: A data mining solution to privacy protection, 2004". The paper studies the generalization technique and problems in privacy preservation. Later, this paper shares the concept of bottom-up generalization. Bottom-up generalization generalizes a consistent value but less specific to preserve the privacy of an individual.

Sweeney, Latanya "Achieving K-anonymity privacy protection using generalization and suppression,2002." This paper shares the concept of the K-anonymity and the MinGen algorithm. MinGen is a theoretical algorithm that combines generalization and suppression techniques to achieve K-anonymity. Moreover, it analyzes MinGen to Datafly and µ-Argus. Both Datafly and µ-Argus are the working implementation of K-anonymity.

5. CONCLUSION

This paper shares the idea of privacy preservation in a data-intensive environment. Before publishing data to a third party, its privacy should be preserved because big data and cloud contain user-specific information. This survey paper shares the concept of three privacy models ,i.e, K-anonymity, l-diversity, and t-closeness. Further, it studies the algorithms used to implement the K-anonymity model.

REFERENCES

[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 Map-Reduce 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 bottomup: scalable sub-tree Anonymization over big data using Map-Reduce 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 kanonymity." 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 kAnonymity algorithm on cloud."Trustcom/BigDataSE/ISPA, 2015 IEEE. Vol. 1. IEEE, 2015. 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 kanonymity." 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.