Survey on Anonymization of Privacy Preserving Data Publishing

[1]Anushree Raj, [2]Rio G.L. D'Souza

[1]Research Scholar, [2]Professor
Computer Science and Engineering Department,
St Joseph Engineering College Mangalore, Karnataka, India

Abstract: **Privacy-preserving data mining has numerous applications which are naturally supposed to be "privacy-violating" applications. The key is to design methods which continue to be effective, without compromising security. Data mining is the process of analyzing data. Data Privacy is collection of data and distribution of data. Privacy issues arise in different area such as health care, intellectual property, biological data, financial transaction etc. Protection of data is a very challenging task while data transfer. Sensitive information needs protection. There are two kinds of major attacks against privacy namely record linkage and attribute linkage attacks. Research have proposed some methods namely k-anonymity, ℓ-diversity, t-closeness for data privacy. k-anonymity method preserves the privacy against record linkage attack alone.**

***IndexTerms*: Anonymization, Privacy Preserving, k-anonymity, PPDM, PPDP**

I. INTRODUCTION

Gigantic volume of itemized individual information is consistently gathered and imparting of these information is ended up being gainful for data mining application. Privacy preserving data mining (PPDM) is a method of protecting the privacy of data without sacrificing the utility of the data. Data mining or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. This paper provides a detailed review of different types of privacy preservation, various privacy preserving Data Mining approaches, techniques and algorithms used in Privacy Preserving Data Mining (PPDM).

II. PRIVACY PRESERVING DATA MINING

Privacy Preserving Data Mining (PPDM) is a field of Data Mining which is used for the extraction of useful knowledge from large amount of data [1], while protecting the sensitive information simultaneously. Data Mining refers to extracting or mining knowledge from large amounts of data. Privacy preserving [2] is said to be done when the attacker is not able to learn anything extra from the given data even with the presence of his background knowledge obtained from other sources.

The process of Privacy Preserving Data Publishing comprise of two important phases, data collection and data publishing. The main objective of PPDM is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process [3].

Privacy preserving in data mining (PPDM) is a new area of research in Data Mining process. Its ultimate goal is to allow one to extract relevant knowledge from large amount of data and provide accurate data mining result, while prevent sensitive information from disclosure or inference. In PPDM, new techniques are invented to provide privacy for the knowledge discovered in Data Mining. It also takes care that knowledge discovery process should not be banned because of privacy reason.

a. ***Types of privacy preservation*** **:** Privacy preservation is divided into following types: -

● Privacy preservation data mining (PPDM): PPDM uses tools and techniques of data mining [4]. It modifies data to mask the sensitive information. In this Data recipient could be an adversary. It directly hides sensitive data and Fails to preserve the truthfulness at record level.

● Privacy preservation data publishing (PPDP): In PPDP, sensitive data is not hidden but hides the identity of an individual by anonymizing the data.

● Privacy preserving distributed data mining (PPDDM): Data mining task is done by different placed by different parties and then combined the data and then published [5].

● Privacy preserving social network data publication (PPSNDP): Social network such as Facebook, LinkedIn etc. are published while preserving data owner privacy.

b. Types of Attack on published data

○ Linkage attack: A linkage attack attempts to re-identify individuals in an anonymised dataset by combining that data with another dataset. The 'linking' uses indirect identifiers also known as quasi-identifiers

- Record linkage attack: It require a dataset against which to link the anonymized data

- Attribute linkage attack: It require a identifier against which to link the anonymized data
- Table linkage attack: It require a database against which to link the anonymized data

○ Probabilistic attack: Probabilistic attack occurs when adversary knows some background information about the victim and he is able to get new information from the table about victim [6].

III. CLASSIFICATION OF PPDM APPROACHES

The Privacy preserving data mining techniques can be broadly classified into six approaches [7]. The fig 1 shows the hierarchy of various PPDM approaches and techniques involved.

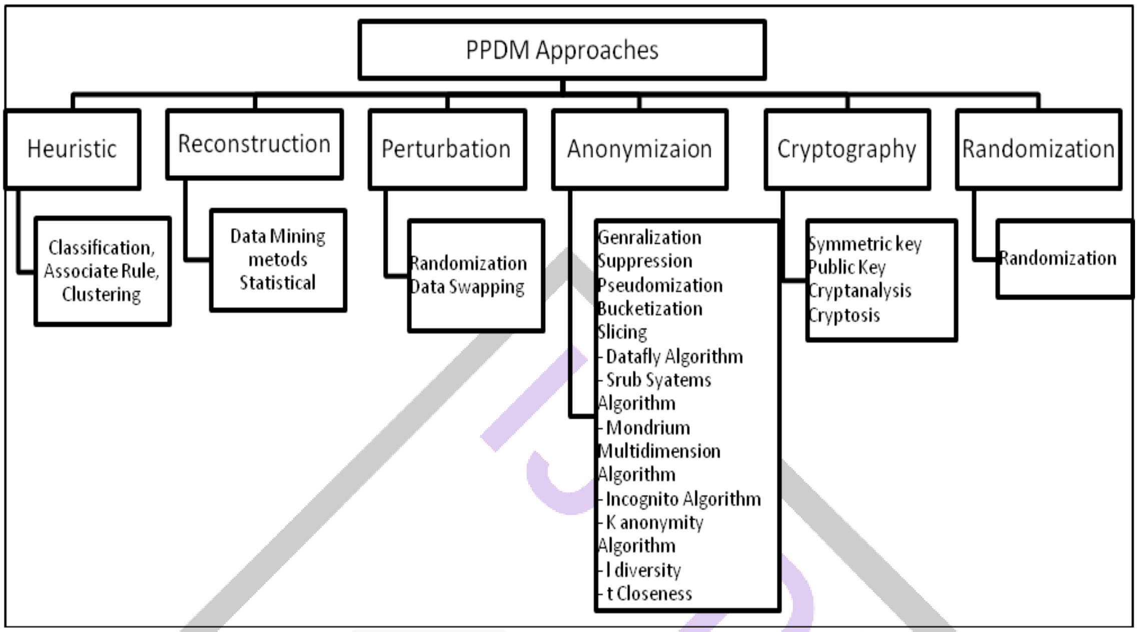

Fig 1 Hierarchy of various PPDM approaches and techniques involved.

a. ***Heuristic Approach****:* Heuristic method is used for centralized database, here two varieties of data is viewed, which is, raw knowledge and aggregated information. Over each forms of knowledge Classification, Association rule mining, Clustering methods are applied, after that hiding procedures are used over the effect of them to preserve it from incorrect utilization.

b. ***Reconstruction approach:*** Reconstruction approach is also used for centralized database, but here, only one type of data is used, which is, raw data. The data mining methods are applied over the raw data. Whatever the outcome comes, the statistical distributed based method is used over them.

c. ***Anonymization approach*****:** Anonymization method aims at making the individual record be indistinguishable among a group records by utilizing techniques of generalization and suppression [8]. Anonymization refers to a methodology where character or/and delicate data about record holders are to be covered up. It even accepts that delicate data should be retained for analysis. There are four sort of quality of fundamental type of data [9]:

- ○ Explicit Identifiers is a situated of properties containing information that recognizes a record manager explicitly, for example, name, percentage and so forth.
- ○ Quasi Identifiers is a situated of properties that could potentially recognize a record manager when combined with publicly available data.
- ○ Sensitive Attributes is a situated of properties that contains touchy individual particular information, for example, illness, salary and so forth.
- ○ Non-Sensitive Attributes is a situated of properties that makes no problem if revealed even to conniving gatherings.

d. ***Randomization Approach:*** The randomized response approach [10] is a manner to mask the original information by adding some random data or noise in it, so One are not able to say that knowledge from a person contains genuine know- how or now not. The added random data or noise must be as big as possible hence the data about someone cannot be recovered by the un-trusted one. This is statistical approach first proposed by Warner. The randomized response process is done in two phases.

e. ***Perturbation approach:*** The perturbation approach modified the normal information values with synthetic information values, in order that the data computed from the perturbed data does now not distinguish from the know-how computed from original data [7]. The perturbation approach is of two types.
Additive perturbation: In additive type, random noise is added to the original data.
Multiplicative perturbation: In multiplicative type, random rotation method is used to perturb data.

Randomization: The randomization method is a simple technique which can be easily implemented at data collection time, because the noise added to a given record is independent of the behavior of other data records. This is also a weakness because outlier records can often be difficult to mask [10].
Data Swapping: In Data Swapping the values across different records are swapped in order to perform the privacy-preservation [11]. One advantage of this technique is that the lower order marginal totals of the data are completely preserved and are not perturbed at all. Therefore certain kinds of aggregate computations can be exactly performed without violating the privacy of the data.

f. ***Cryptographic approach*:** Cryptographic procedures are ideally meant for situations where multiple parties collaborate to compute outcome or share non sensitive mining outcome and thereby averting disclosure of sensitive knowledge [12]. Cryptography approach basically works on distributed database, which is the one, where data is stored in different places. The data which is being stored may be raw data or aggregated data or both. On applying data mining methods on each type of data some results will come, on them encryption technique will be used.
Cryptographic Technique - (SMC) Secure Multiparty Computation
Symmetric key – [13] Includes Encryption/decryption, single key
Public key – [14] Includes Encryption/decryption, digital signature, key exchange
Cryptanalysis – [15] Includes Cipher text encryption/decryption.
Cryptosys - Cryptosys [16] is a suite of cryptographic algorithms needed to implement a particular security service, for achieving confidentiality. It consists of three algorithms: one for key generation, one for encryption, and one for decryption

IV. ANONYMIZATION TECHNIQUES

Anonymization refers to a methodology where character or/and delicate data about record holders are to be covered up. It even accepts that delicate data should be retained for analysis [17].

3.1 Different Anonymization techniques are:

1) ***Suppression:*** Suppression [18] involves not releasing a value at all. It replaces tuple or attribute values with special symbol "**" that is instead of the original value we replace it with some anonymous value throughout the database. A table generated by suppression in which the value of age and the type of disease is replaced with an "**"

Gender	***Zip code***	***Age***	***Disease***
Female	575001	**	**
Female	575015	**	**
Female	575015	**	**
Female	575006	**	**

Table 1: A published data by suppression

2) ***Generalization:*** Generalization [19] involves replacing (or recoding) a value with a less specific but semantically consistent value. It replaces attribute values with semantically unvarying but less particular value. Due to this replacement, many records have same QI values. Generalization replaces exact values with a more general description to hide the details of attributes, making the QIDs less identifying as shown in Table 2.

Name	***Gender***	***Zip code***	***Age***
Anisha	Person	575002	35-45
Rekha	Person	575016	35-45
Malini	Person	575003	35-45
Sowmya	Person	575005	35-45

Table 2: A published table by generalization

It is easy to find a semantically unvarying value which can be used to implement generalization but it fails on high-dimensional data due to the curse of dimensionality and it also causes too much information loss.

3) ***Bucketization:*** Bucketization is the process of defining the several records grouping based on their sensitive values. The apparent sensitive values of the attributes are identified and sorted based on the frequencies in ascending order [20]. After sorting is done, the contiguous sensitive values are grouped into the similar bucket. Only the buckets contain at least ℓ distinct sensitive values which are kept after bucketing process completion as shown in table 3. The purpose of bucketizing the attributes is to establish the least cost splitting up of a multidimensional data set into B set of buckets (where the value of B is smaller than the number of data points in the dataset).

Gender	***Zip code***	***Age***	***GID***
Female	575001	35	1
Female	575015	36	2
Female	575015	48	3
Female	575006	55	4

a) QId table

Disease	***QID***
Asthma	1
Malaria	2
Tumor	3
Cancer	4

b) Sensitive table

Table 3: A published data by bucketization

4) ***Perturbation:*** Under perturbation [21], a value can be changed to any arbitrary value. For example, Male can be changed to Female and vice versa. Table 4 shows an example with perturbation. Drawback is, it reduces data utility.

Name	*Gender*	*Zip code*	*Age*
Anisha	Male	575001	35
Rekha	Male	575015	36
Malini	Male	575015	48
Sowmy a	Male	575006	55

Table 4: A published data by perturbation

5) ***Slicing:*** Slicing handles high-dimensional data. To deal with problems which occur in generalization and bucketization, Slicing [22] is a new technique to preserve privacy of published data**.** Slicing is a novel data anonymization technique which improves the current state of the art. Slicing partitions the data set both vertically and horizontally. The basic idea of slicing is to break the association cross columns, but to preserve the association within each column as shown in table 5.

Gender , Disease	*Zip code, Age*
Female , Asthama	575001,35
Female , Malaria	575015,36
Female , Tumor	575015,48
Female , Cancer	575006,55

Table 5: A published data by perturbation

V. k- ANONYMIZATION TECHNIQUES

k-Anonymization is a technique that prevents joining attacks by generalizing and/or suppressing portions of the released micro-data so that no individual can be uniquely distinguished from a group of size *k*. In k-anonymity techniques [23], we reduce the granularity of representation of these pseudo-identifiers with the use of techniques such as generalization and suppression. In the method of generalization, the attribute values are generalized to a range in order to reduce the granularity of representation.

1) **Mondrian Multidimensional k Anonymity:** It provides a multidimensional partitioning is NP-hard, and efficient greedy approximation algorithm for several general purpose quality metrics [24]. the multidimensional model, which provides an additional degree of flexibility leads to higher-quality anonymizations, as measured both by general-purpose metrics and more specific notions of query answerability. Integrates an anticipated query workload directly into the anonymization algorithms.

2) ***Scrub System***: The Scrub systems approach provides a methodology for removing personally identifying information in medical records so that the integrity of the medical information remains intact even though the identity of the patient remains confidential [25].

3) ***Datafly*****:** Datafly [26] maintains anonymity in medical data by automatically generalizing, substituting, and removing information as appropriate without losing many of the details found within the data. Decisions are made at the field and record level at the time of database access, so the approach can be used on the fly in role-based security within an institution, and in batch mode for exporting data from an institution.

4) ***The µ-Argus System*****:** In µ -Argus, [27] the data holder provides a value for *k* and specifies which attributes are sensitive by assigning a value between 0 and 3 to each attribute. These correspond to "not identifying," "most identifying," "more identifying," and "identifying," respectively. µ -Argus then identifies rare and therefore unsafe combinations by testing 2- and 3-combinations of attributes. Unsafe combinations are eliminated by generalizing attributes within the combination and by cell suppression. Rather than removing entire tuples, µ -Argus suppresses values at the cell-level. The resulting data typically contain all the tuples and attributes of the original data, though values may be missing in some cell locations.

5) ***Incognito algorithm****:* The Incognito algorithm [28] generates the set of all possible k-anonymous full-domain generalizations of T, with an optional tuple suppression threshold. Based on the subset property, the algorithm begins by checking single-attribute subsets of the quasi-identifier, and then iterates, checking k-anonymity with respect to increasingly large subsets.

6) ***k-Anonymity:*** k-anonymity provides privacy protection by guaranteeing that each released record will relate to at least k individuals even if the records are directly linked to external information [29]. K anonymity is a combination of generalization and suppression. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value. Suppression involves not releasing a value at all.

Find sets of OSC and return nodes with least information loss.

7) ***l- Diversity****:* An equivalence class is said to have l-diversity if there are at least l-“well-represented” values for the sensitive attribute. A table is said to have l-diversity if every equivalence class of the table has l-diversity[30]. A q⋆-block is ℓ-diverse if contains at least ℓ “well-represented” values for the sensitive attribute S. A table is ℓ-diverse if every q⋆ -block is ℓ-diverse.

8) ***t- Closeness:*** The t-closeness model is a further enhancement on the concept of l-diversity [31]. One characteristic of the l-diversity model is that it treats all values of a given attribute in a similar way irrespective of its distribution in the data. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t.

VI. COMPARISON BETWEEN DIFFERENT ANONYMITY ALGORITHMS

Privacy Model	***Sanitization method***	***Description***	***Advantages and Disadvantages***
Mondrian Multi-dimensional	Partitioning algorithm	It is an optimal multidimensional partitioning is NP-hard, but simple and efficient greedy approximation algorithm	It uses the binary search to obtain the solution in less time.
Scrub systems	Generalization	locates personally-identifying information in letters between doctors and notes written by clinicians;	de-identifies information and cannot guarantee anonymity
Data fly	Generalization	Produces a resulting database whose information matches the anonymity level set by the user with respect to the recipient profile.	It checks very few nodes for k-anonymity, hence results very fast. But, algorithm skips many nodes, and resulting data is very generalized and provides very little information.
µ-Argus System	Suppression	It's a project to develop specialized software for disclosing public-use data such that the identity of any individual contained in the released data cannot be recognized.	The µ-Argus systems combinations of fields adhere to a minimal bin size, and by so doing, confidentiality can be maintained and we can even provide anonymous data for public use.
Incognito	Generalization	This algorithm produces all the possible k-anonymous full-domain generalizations of a relation(say T), with an optional tuple suppression threshold.	The algorithm always locates the optimal solution. But, uses breadth first search method which takes a lot of time to traverse the solution space.
k-anonymity	Generalization, Suppression	Anonymity is guaranteed by the existence of atleast other k-1 records for each record in a database.	The probability of uniquely inferring the sensitive information of individual known by the adversary is less than *1/k*, thus, it can safeguard individuals' privacy to a large extent.
l-Diversity	Generalization, Bucketization,	Expands the k-anonymity model by requiring every equivalence class to have atleast l "well-represented" value for the sensitive attributes.	It guarantees l- distinct values for the delicate trait in every equivalence class, But doesn't includes the measures that combine distance-estimation properties
t-Closeness	Generalization, Suppression	Solves the l-diversity problem of skewed sensitive values distribution by requiring that the distribution of the sensitive values in each equivalence class to be "close" to the corresponding distribution in the original table, where close means upper bounded by a threshold t.	Distribution of the attribute in the whole table is no more than a threshold t. But, doesn't includes the measures that combine distance-estimation

VII. CONCLUSIONS

This paper tries to summarize the basics about Privacy Preserving Data Mining with its various classifications. Then it introduces to Data Anonymization. Privacy is a key issue in publish data because improper disclosure of certain data assets will harm the prospects. Popular approaches of data anonymization like generalization, suppression, bucketization, slicing and perturbation have been discussed. The evolution in the research of the anonymization techniques like Mondrian Multi-dimensional, Scrub systems, Data fly, µ-Argus System, Incognito, k-anonymity, l- Diversity, t-Closeness, presented in the paper shows how different types of attacks can compromise privacy and how different techniques can be applied to protect from these invasions. Overall this paper tries to show conceptive form of Data Anonymization.

REFERENCES

[1] Charu C. Aggarwal, Philip S. Yu ,"A General Survey of Privacy-Preserving Data Mining Models and Algorithms", Springer, 2008.

[2] A.S.Shanthi, , Dr. M. Karthikeyan” A Review on Privacy Preserving Data Mining “IEEE International Conference on Computational Intelligence and Computing, 2012

[3] Stan Matwin, “Privacy Preserving Data Mining Techniques: Survey and Challenges”, Discrimination and Privacy in the information Society, Springer, pp.209-222, 2013.

[4] C. C. Aggarwal, Data Mining: The Textbook, Springer International Publishing Switzerland, 2015.

[5] Kavita Rodiya and Parmeet Gill, "A Review on Anonymization Techniques for privacy preserving data publishing" IJRET: International Journal of Research in Engineering and Technology, November 2015

[6] C.Saravanabhavan, Dr.R.M.S.Parvathi"An Efficient Approaches for Privacy Preserving In Microdata Publishing Using Slicing and Partitioning Technique "International Journal of Engineering Research and Applications (IJERA), August 2013

[7] Hina Vaghashia and Amit Ganatra, “A Survey: Privacy Preservation Techniques in Data Mining”, International Journal of Computer Applications (0975 – 8887) Volume 119 – No.4, June 2015

[8] P. Samarati and L. Sweeney, “Protecting privacy when disclosing information:-anonymity and its enforcement through generalization and suppression,” SRI International, SRI-CSL-98-04, 1998.

[9] Seema Kedar1, Sneha Dhawale2, Wankhade Vaibhav3,Pavan Kadam4 , Siddharth Wani5 , Pavan Ingale, “Privacy Preserving Data Mining“, IJARCCE Vol. 2, Issue 4, April 2013

[10] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1):146, 2007.

[11] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and Ian Molloy. Slicing: A New Approach for Privacy Preserving Data Publishing.

[12] Helger Lipmaa,” Cryptographic Techniques in PrivacyPreserving Data Mining”, University College London, Estonian Tutorial 2007.

[13] The Blocak Cipher Rijndael, Joan Daeman and Vincent Rijmen, CARDIS 2000, LNCS 1820, pp 277-284 2000, @ Springer-Verlag Berlin eidenlberg 2000

[14] A Method for Obtaining Digital Signatures and Public-Key Cryptosystems R.L. Rivest, A. Shamir, and L. Adleman, February 1978 Volume 2

[15] Cryptography and Cryptanalysis: A Review Gangadhar Tiwari, Debashis Nandi, Madhusudhan Mishra. International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 10, October - 2013

[16] A Survey of Public-Key Cryptosystems Neal Koblitz Dept. of Mathematics, Box 354350 Univ. of Washington, Seattle, Alfred J. Menezes, Waterloo, Ontario N2L 3G1 Canada ajmeneze@uwaterloo.ca August 7, 2004

[17] Mahesh Dhande, N.A.Nemade and Yogesh Kolhe, 2013, “Privacy Preserving in K- Anonymization Databases Using AES Technique”.

[18] Yang Xu, Tinghuai Ma, Meili Tang and Wei Tian. A Survey of Privacy Preserving Data Publishing using Generalization and Suppression. Applied Mathematics & Information Sciences.2014. Pages 1103-1116

[19] P. Samarati, “Protecting Respondent’s Privacy in Microdata Release,” IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pages 1010-1027, Nov./Dec. 2001.

[20] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, “Worst-Case Background Knowledge for Privacy-Preserving Data Publishing,” Proc. IEEE 23rd Int’l Conf. Data Eng. (ICDE), pages 126-135, 2007

[21] H. Kargupta, S. Datta, Q.Wang, and K. Sivakumar, “On the privacy preserving properties of random data perturbation techniques,” in Proceedings of the International Conference on Data Mining (ICDM), p. 99, 2003.

[22] Tiancheng Li, Ninghui Li, “Slicing: A New Approach for Privacy Preserving Data Publishing” IEEE Transactions On Knowledg

[23] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[24] Mondrian Multidimensional K-Anonymity Kristen LeFevre David J. DeWitt Raghu Ramakrishnan

[25] Sweeney, L. Replacing personally-identifying information in medical records, the Scrub system. Proceedings, American Medical Informatics Association. Washington: Hanley & Belfus, 1996.

[26] Sweeney, L. Guaranteeing anonymity when sharing medical data, the datafly system. Proceedings, American Medical Informatics Association. Nashville: Hanley & Belfus, Inc, 1997.

[27] Three Computational Systems for Disclosing Medical Data in the Year 1999 Latanya Sweeney

[28] Incognito: Efficient Full-Domain K-Anonymity Kristen LeFevre David J. DeWitt Raghu Ramakrishnan

[29] L. Sweeney, “K-Anonymity: A model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[30] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. **l**-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), March 2007

[31] N.Li,T.Li,and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and "- diversity. In Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), Istanbul, Turkey, April 2007.