International Journal of Computer Applications (0975 - 8887)
Volume 182 - No.28, November 2018

A Survey of Data Hiding Techniques

Kshitij Pathak
Department of Computer Science and Engineering
UIT, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M.P.), India

Sanjay Silakari
Professor
Department of Computer Science and Engineering
UIT, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal (M.P.), India

Narendra S. Chaudhari
Professor
Department of Computer Science & Engineering
Indian Institute of Technology, Indore (M.P.) India

ABSTRACT

This paper introduces the privacy, data privacy - Stakeholders and classifications of attributes for data hiding techniques. It also throws the light on various data hiding techniques such as randomization, k-anonymity, l-diversity, t-closeness and tokenization. Also, the importance of balancing privacy and utility is discussed.

General Terms

Data Hiding in Database, Data Privacy

Keywords

Randomization, k-anonymity, l-diversity, Tokenization, t-closeness

1. INTRODUCTION

Thousands of women in Uttar Pradesh (India) reported that they are getting fake calls. When the case was investigated, it was known that their information including their mobile numbers are getting sold for Rs. 50 to Rs. 500.
In another case, Indian Government has sent notices to Chinese and other mobile device makers when found Suspicious of hacking and theft of information from smartphones to provide the outline and procedures followed for data security. 21 phone makers including leading Chinese brands Oppo, Vivo, Xiaomi and Gionee, have been enquired to give "detailed, structured written response" on how they secure data and ensure its safety and security, a government order said. (16th August 2017; news in http://www.newindianexpress.com/business/2017/aug/16/).
There are a lot of such incidents reported in last few years not only in India but worldwide. When such type of incidents reported, company face financial loss, loss of reputation and loss of their customers.

2. DATA PRIVACY - STAKEHOLDERS IN AN ENTERPRISE

An organization has various stakeholders [19] of data privacy as shown in Figure 1. Their description is as follows:

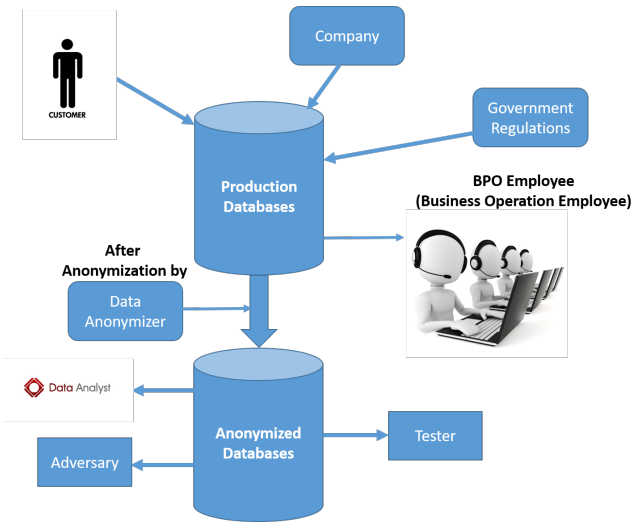

Fig. 1. **Data Privacy - Stakeholders**

(1) **Company:** Company keeps a significant amount of information including employee record, client details. Some of the data related to banking and medical detail are very sensitive, and it is their responsibility to protect such crucial data. The company holds a record of customer/data owner, under government compliance and regulations. If such information is leaked, it may cost companies any legal action from the government. At the same time, trust over the enterprise or reputation of the company is degraded publicly.

(2) **Government:** Another stakeholder in data privacy is Government. In India, Company must comply with Information Technology act 2000. All these laws have to be strictly followed by an organization.

(3) **Data Analyst:** It is the person whose primary task is to extract hidden information from a large amount of data (Big Data). Here input is not original data; first data is anonymised by data anonymizer as per government regulation for protecting sensitive information, and then it is being released for a data analyst for data mining.

(4) **Data Anonymizer:** Person or group of individuals whose task is to anonymize the database before releasing it for data sharing, knowledge discovery, prediction and so on. Various techniques related to anonymity will be discussed in section 4.

(5) **Tester:** In software development life cycle, β-testing is performed by sending software to loyal companies/customers. For efficient testing to be executed, along with software, good quality data must be supplied which also includes customer sensitive information. Therefore, test team will also be provided anonymised data after performing anonymization on production databases.

(6) **Business Operation Employee:** BPO employee or Business operation employee handle live databases. So, they granted to work on production databases to support customer requirements whereas data analysts and tester work on static data which is generated by suitable privacy preserving method.

(7) **Adversary/Data Snooper:** An adversary is one who tries to theft data which may be company's employee or external person (other than organization employee). Hence, anonymization should be secure enough that an attacker/adversary could not identify any individual by performing mining of data.

3. CLASSIFICATION OF ATTRIBUTES

[19] and [16] classified attributes of data present in tables/databases as follows:

i) **Explicit identifiers (EI)**: Attributes which uniquely identify an individual/record owner from datasets. Example : Aadhar Number, PAN Number etc.

ii) **Quasi-identifiers (QI)**: Attributes which are available publicly. Adversary/data snooper can theft these data and make an effort to identify the individual. Therefore, anonymization must be applied on Quasi-Identifiers to thwart data snooper effort to identify customer/individual in the database.

iii) **Sensitive data (SD) and Nonsensitive data (NSD)**: It is defined as attributes or set of attributes that contain trustworthy information related to individual/record owner, such as bank account number, balance, diabetic, hypertension etc. Nonsensitive are the attributes with no sensitivity for the given context. A logical representation of data table is shown in Table 1.

4. TECHNIQUES OF DATA HIDING

4.1 The Randomization Method

It is a prevalent method of preserving privacy in the database. In this method, noise is either added or multiplied to records to mask the value of records. Initial work can be found in [20, 12]. It is first probed when data has been collected by survey through question-answer method from individuals. To remove answer bias, randomization method is used to distort data by probability distribution methods. Although, data can be reconstructed by removing noise. This is discussed in [2]. Randomization can be explained by an example below: Consider a set of data values $\{A1, A2, A3 ... A_n\}$ then it is distorted by additive strategy by adding noise generated from probability distribution $\{N1, N2....N_n\}$ to produce output as $\{A1 + N1, A2 + N2 .....A_n + N_n\}$. The variance of the noise added is taken large so that no prediction or guessing of original values can be done. In multiplicative strategy, noise is multiplied by data values. Randomization can be extended to various data mining tasks such as classification as done in [2] and association rule mining as done in [5]. Features of Randomization method are :

- Simple.
- Does not depend on the distribution of records.
- Applied only at data collection time.
- Attacks possible as challenging to mask outlier records.
- Can be implemented to association rules as well as on classifier for privacy preservation.

4.2 Group-Based Anonymization : k-Anonymity

Randomization method can be applied only at data collection time, so there is need of method or an approach if privacy-preservation cannot be applied at data collection time. Also, the randomization method is weak when outlier records are present. To overcome all these limitations, group-based anonymization methods are constructed. Various approaches that come under the category of group-based anonymization are k-anonymity, personalized privacy preservation, utility-based privacy preservation, l-diversity and t-closeness. k-anonymity apply generalization and suppression to hide sensitive data.

4.2.1 Basic Terminologies : Generalization and Suppression Technique. Generalization technique can be applied at the level of:

(1) **Attribute (AG):** Applying generalization on an attribute means generalizing all the values of that attribute. For example, Let city attribute is to be generalized, then it can be generalized by district, state or country.

(2) **Cell (CG):** Applying generalization technique on a single cell means affecting a specific value in the specific column. Consider Date_of_joining column, a cell of this column can be generalized as containing month and year rather than exact date or containing the only year.

Suppression technique can be applied at the level of:

(1) **Tuple (TS):** Applying suppression technique on tuple means remove a particular tuple from the table to achieve k-anonymity.

(2) **Attribute (AS):** Applying suppression technique on attribute, i.e., to disturb all the values of a particular column/attribute.

(3) **Cell (CS):** suppression on cell means to disturb certain cells of a particular attribute.

4.2.2 Classification of generalization and suppression techniques. The various combinations of generalization and suppression technique [4] are shown in the Table 2 below:

Table 2. **Classification of Techniques**

	Suppression			
Generalization	**Attribute**	**Cell**	**Tuple**	**None**
Attribute	AG_AS	AG_CS	AG_TS	AG_
Cell	CG_AS (Not Applicable)	CG_CS (Not Applicable)	CG_TS	CG_
None	_AS	_CS	_TS	

(1) **AG_AS:** In this technique, generalization, as well as suppression both, are applied at the attribute level. Although, no approach has been reported for this model. The reason behind it is, once the generalization is applied at attribute level there is no need to apply suppression. So, we can conclude that AG_AS model is equivalent to AG_ model.

Table 1. **Customer Data**

EI	EI	QI	QI	QI	SD
Aadhar_number	Name	Address	Gender	Contact_No	Account_No
Aadhar1	Anil	478001	M	35000	acc_no1
Aadhar2	Rohan	462241	M	23000	acc_no2
Aadhar3	Rudu	462241	M	54000	acc_no3
Aadhar4	Mayank	478001	M	11000	acc_no4
Aadhar5	Mitali	418023	F	90000	acc_no5
Aadhar6	Abhinav	416023	M	89000	acc_no6
Aadhar7	Shyam	400023	M	89000	acc_no7

(2) **AG_CS:** In this model, generalization is applied on column or attribute whereas Suppression is being implemented on a particular cell of the table. This model has been investigated in [6, 8, 7] and Datafly [17].

(3) **AG_TS:** In this model, generalization is applied on column or attribute whereas suppression is applied on a tuple or record of the table. This model is based on work reported in [15]. This technique provides a right balance between runtime complexity and data privacy. Various algorithms based on this model are developed in [3, 9, 10, 18, 21].

(4) **AG_ :** In this classification, Suppression is not applied, whereas generalization is applied on column or attribute. It is same as model AG_AS.

(5) **CG_TS :** In this model, generalization is applied on a cell and suppression is applied on the tuple.

(6) **CG_ :** In this model, only generalization is applied, no suppression. An algorithm based on CG_ is proposed in [23]. This model is equivalent to CG_CS.

(7) **_AS:** In this model, generalization is not considered, only Suppression is applied on a column of the table. No recognized work has been reported in this model. It can be viewed as a reduction of _AG in which the generalization hierarchies are at height 1.

(8) **_CS :** In this model, generalization is not considered, only suppression is applied at cell level in the table. An algorithm based on _CS is proposed [14]. It can be viewed as a reduction of _AG in which the generalization hierarchies are at height 1.

(9) **_TS :** In this model, generalization is not considered, only Suppression is applied at tuple level. It can be viewed as a reduction of AG_TS in which the generalization hierarchies are at height 0. Algorithms build on the _TS model have polynomial runtime complexity and also solution generated is unique.

4.2.3 Generalization hierarchy (GH). Hierarchies related to each attribute are assumed to exist, where leaves consist of the data that could be found in Private Table (PT), and the rest of the levels are a generalization of these data accordingly.
To illustrate, consider a table PT, having five attributes which keep track of employee medical details, whether they have diabetes. Let table consist of following attributes: Aadhar_number, sex, marital_status, office_location, diabetes.
First of all, it is clear that Aadhar_number will not be released as an individual can be identified. Now among the four remaining attributes, a combination of values should be updated in such a way that individual identity cannot be determined. Quasi-Identifiers is an attribute or set of attributes in PT that, in conjunction, can be allied with external information to re-identify an individual to whom the information refers. k-anonymity technique updates the table in such a way that combination of values must be identical with at least k-tuples. The outcome of the group by clause on attributes (SEX, MARITAL_STATUS, OFFICE_LOCATION, DIABETES) with aggregate function count is shown in Table 3.

Table 3. **Result of Aggregate Function on Employee Table**

Sex	Marital_Status	Office_Location	Diabetics	Count
F	divorced	Indore	Y	1
F	married	Ujjain	N	4
F	married	Ujjain	Y	2
F	single	Indore	N	2
F	single	Indore	Y	1
M	divorced	Indore	N	12
M	divorced	Indore	Y	9
M	divorced	Ujjain	N	2
M	divorced	Ujjain	Y	3
M	married	Ujjain	N	3

Table 4. **Result of Aggregate Function on Modified Employee Table**

Sex	Marital_Status	Office_Location	Diabetics	Count
Any	divorced	Indore	N	12
Any	divorced	Indore	Y	10
Any	divorced	Ujjain	N	2
Any	divorced	Ujjain	Y	3
Any	married	Ujjain	N	7
Any	married	Ujjain	Y	2
Any	single	Indore	N	2
Any	single	Indore	Y	1

It can be identified from Table 3 that a divorced woman at office location Indore has diabetes. Generalization technique generalizes the value of the attribute to hide her identity. Let SEX attribute be generalized as 'any-sex' rather than 'M' or 'F.' Now 'Divorced' peoples at office location Indore suffering from diabetes count is updated to 10 as shown in Table 4 . So an individual cannot be identified.

4.2.4 k-anonymity : Samarati Algorithm. Samarati algorithm [15] work on attribute set in conjunction with domain generalization hierarchy. In this algorithm, the concept of domain generalization is applied on tuple domain. Tuple domain in association with domain generalization hierarchy is a lattice. In this lattice, every vertex is representing generalized table which is obtained by generalizing the associated attributes and suppressing few tuples. The following is a summary of Samarati's algorithm:

(1) Let PT be a private table to be generalized, given set of attributes, i.e., quasi-identifiers.

(2) Initially, search area is the whole lattice

(3) Pick the area of search at middle height (Mh).

(4) Now consider all node at height Mh and check if at Mh, is there exists at least one node that satisfies k-anonymity with minimum suppression.
   a) If not the minimum, specify the upper half of Mh as the new area of search.
   b) If minimum, specify the lower half of Mh as the new area of search.

(5) Repeat step 3 till search area consists of more than one level in the lattice else, return a solution at this level.

4.2.5 Bayardo-Agrawal Algorithm. k-optimize algorithm is proposed by Bayardo and Agrawal [3] which is based on attribute generalization. This method associates an integer value, i.e., index to values of attributes of quasi-identifier.
Consider the example described in subsubsection *4.2.3*, let quasi-identifiers are sex, marital_status, and office_location, so the index will be assigned to different values of attributes in sequence. Since sex is first attributed in the ordered set so sex = 'F' is assigned index 1, sex = 'M' is assigned index 2. Next attribute in the ordered set is marital_status which have three values married, divorced and single so index assigned to values married, divorced and single is 3, 4 and 5 respectively. Similarly next quasi-identifier attribute office_location= 'Ujjain' is assigned index 6 and office_location='Indore' is assigned index 7. k-optimize algorithm then builds an enumeration tree over a set of index values.

4.2.6 Incognito Algorithm. Incognito algorithm [10] generates all the possible full-domain generalizations (k- anonymous) of a table or relation, with an optional tuple suppression threshold. Algorithm starts by checking for single attributes which are a subset of quasi-identifier and then iterates by considering larger subsets of quasi-identifier.

4.2.7 Drawback of k-anonymity

(1) k-anonymity suffers from homogeneity attacks.

(2) Optimal k-anonymization is NP-hard.

(3) Algorithm performance is reduced with high-dimensional data and large record sizes.

(4) Achieving a balance between privacy versus utility is hard. A higher value of k provides great privacy and low utility whereas a lower value of k provides high utility and low privacy.

(5) The use of suppression leads to high information loss or low utility and using the only generalization leads to a highly generic table having very low utility.

4.3 l-diversity

k-anonymity methods suffer from homogeneity attack as the value of a sensitive attribute in a block of k-record is same. The outcome of k-anonymity does not satisfy diversity. Therefore, l-diversity is proposed to diverse the value of a sensitive attribute in a block of k-records, i.e., it not only keeps the minimum group size of k (k-anonymity) but also maintains the diversity of attributes.
Consider an example shown in Table 5 below :
First applying k-anonymity by suppressing the value of gender and generalizing the age attribute. The resultant table is shown in Table 6. Now the table is k-anonymous where k = 4. However, the value of a sensitive attribute within a group of 4 tuples is same ('**,' '<60', 'Ujjain,' 'Diabetes'). An adversary can quickly identify their diseases via homogeneity attack. Now applying l-diversity on Table 6, the outcome is shown in Table 7
l-diversity model of privacy is defined as follows:

Table 5. **Health_details**

Gender	Age	City	Disease
M	53	Indore	Thyroid
F	34	Indore	Swine FLU
M	31	Indore	Thyroid
M	39	Ujjain	Diabetes
F	48	Ujjain	Diabetes
M	56	Ujjain	Diabetes
F	58	Ujjain	Diabetes
F	50	Ujjain	Diabetes
M	29	Indore	Diabetes

Table 6. **Health_details After k-anonymity**

Gender	Age	City	Disease
**	<60	Indore	Thyroid
**	<40	Indore	Swine FLU
**	<40	Indore	Thyroid
**	<40	Ujjain	Diabetes
**	<60	Ujjain	Diabetes
**	<60	Ujjain	Diabetes
**	<60	Ujjain	Diabetes
**	<60	Ujjain	Diabetes
**	<40	Indore	Diabetes

Table 7. **Health_details After l-diversity**

Gender	Age	City	Disease
**	<60	Indore	Thyroid
**	<40	Indore	Swine FLU
**	<40	Indore	Thyroid
**	<40	Ujjain	Diabetes
**	<60	Ujjain	Swine FLU
**	<60	Ujjain	Diabetes
**	<60	Ujjain	Diabetes
**	<60	Ujjain	Thyroid
**	<40	Indore	Diabetes

Let a q*-block be a set of tuples such that its non-sensitive values generalize to q*. A q*-block is l-diverse if it contains l "well represented" values for the sensitive attribute S. A table is l-diverse, if every q*-block in it is l-diverse [13].
Other methods for applying l-diversity is proposed in [1] and [22]. Disadvantages of l-diversity are skewness attack, similarity attack as discussed in [11].

4.4 t-Closeness

t-closeness model enhances the concept of l-diversity [11]. The distribution of data has not been taken into consideration by l-diversity which is very important in real life situations. An attacker utilizes the old data and can make assumptions about confidential and sensitive values in data, For example, an attribute corresponding to the birthmarks of an employee may be sensitive. In [11], t-closeness model was proposed which take care of distribution of data present in form of tables and generalized tables be at most t.

4.5 Tokenization

Anonymization technique is applied to static data. So, there is the necessity of data privacy algorithms which work on dynamic data.

Tokenization is formed to apply privacy on real-time data, or we can say tokenization is introduced for privacy protection during the runtime of an application as shown in Figure 2 .

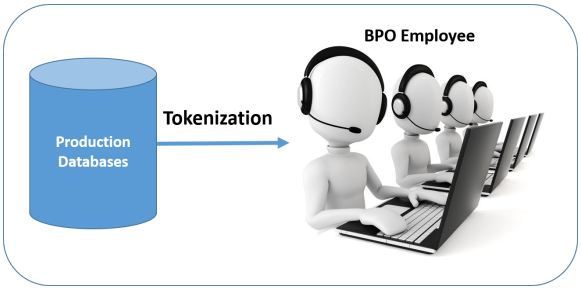

Fig. 2. **Data Privacy - Tokenization**

Tokenization technique replaces the sensitive data with a random string of characters known as a token. Token generated is not dependent on sensitive data, so it provides high-level privacy protection. This method is mainly used in Payment Card Industry (PCI) in which credit/debit card number are replaced with a token. This approach also supports user view of data as token look likes the original data. In section 2, it is mentioned that business operation employee work on production data since they need to know their customers to provide support and services as well as to handle the queries. General information related to clients is passed as it as is as there is no need to hide their customer's identity, but they will be provided tokenized data to protect crucial information of clients like a credit card, health, etc.

Example:

Consider an End-user requested for a customer data stored in cloud having personal as well as his financial details. Financial details include his bank account number and currently drawn salary. With tokenization method, a random string is generated, i.e., the token is generated for his/her current salary drawn as well as bank account number and rest information is forwarded as plaintext. The mapping between original value and token is stored in the token vault.

4.5.1 Features and Drawbacks of Tokenization

i) There is no mathematical relationship between original data and tokenized data, so there is no key to generate original data from tokenized data.

ii) It is consistent, cheap and provides a high level of security.

iii) Format preserving as it supports different data types as well as tokenized data looks like the original data.

iv) Compatible with other Technologies such as NFC payments and ACH transfers.

v) Tokenization makes it easier for merchants to become PCI compliant.

vi) Original data never leaves the origin which satisfies specific government compliance and regulations.

vii) Performance degrades with large databases.

viii) Useful only for structured fields like aadhar number and payment details.

5. IMPORTANCE OF BALANCING DATA PRIVACY AND UTILITY

Companies use data for testing, extracting knowledge, so large-scale data has to be shared for mining and research. However, at the same time, sensitive data has to be protected. Now one more issue will be arised after applying privacy-preservation.

Is released data can still be utilized?

If not, data is of no use. As discussed in the section 3, EI attributes like SSN, Aadhar_number are completely removed from the database to ensure individual privacy. On QI and SD attributes, anonymization is applied for preserving privacy. However, anonymization of data should not be up to such an extent that data utility vanishes. Let us take an example. Consider table 'Salary' shown in Table 8 with attributes Aadhar_number, name, zipcode, gender, salary. Before releasing the table for data sharing, EI attributes (Aadhar_number in the example) must be removed entirely. Now on remaining attributes, certain updates need to be done for protecting sensitive information as shown in Table 9. As we can

Table 8. **Salary table**

Aadhar_number	name	zipcode	gender	salary
1234-XXXX-3XXX	Ajay	458001	M	35000
1234-XXXX-5XXX	Sunil	465441	M	23000
1234-XXXX-2XXX	Sumiti	465441	F	5000
1234-XXXX-9XXX	Varun	458001	M	11000
1234-XXXX-1XXX	Rahul	400023	M	90000
1234-XXXX-4XXX	Paridhi	400023	F	89000

see that names have been changed, zip code has been modified to protect that high-income peoples are at zip code 400023. Gender updated to Male. Salary kept in original form. This update protect privacy, but the utility of the data has been reduced since no mining on female employees. If cryptographic techniques are used for

Table 9. **Updated Salary table**

name	zipcode	gender	salary
Anil	478001	M	35000
Rohan	462241	M	23000
Rudu	462241	M	5000
Mayank	478001	M	11000
Aarush	411023	M	90000
Abhinav	411023	M	89000

data privacy, then the outcome of the technique will provide high privacy and low utility when encryption is done. High utility and low privacy will be encountered after decryption is applied so it is clear that if we measure privacy and utility in a range of 0 to 1 then resultant with cryptographic technique on privacy and utility is 1 and 0 respectively after encryption and 0 and 1 respectively after decryption.

We can conclude that either privacy achieved or utility achieved. To have proper balance between privacy and utility, anonymization is performed which controls the level of utility and privacy. Released data after anonymization is used at different places for mining and research so, right anonymization technique keeps a proper balance by putting privacy and utility in 'Shades of Gray.' This is explained in the diagram shown in Figure 3.

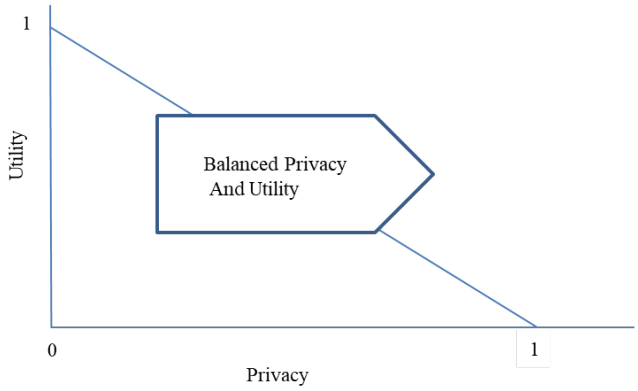

Fig. 3. **Balance between privacy and utility**

6. CONCLUSION

This paper introduces the privacy, data privacy - Stakeholders and classifications of attributes for data hiding techniques. A comprehensive description of various data hiding techniques such as Randomization, k-anonymity, l-diversity, t-closeness and tokenization is discussed. Their advantages and disadvantages are listed. Also, the importance of balancing privacy and utility is discussed. This discussion help researcher to pick the correct data hiding technique which provides the data privacy as well as maintain balance between privacy and utility.

7. REFERENCES

[1] Charu C Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, pages 901–909. VLDB Endowment, 2005.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *ACM Sigmod Record*, volume 29, pages 439–450. ACM, 2000.

[3] Roberto J Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 217–228. IEEE, 2005.

[4] V Ciriani, S De Capitani di Vimercati, S Foresti, and P Samarati. k-anonymity. security in decentralized data management. *jajodia S., Yu T., Springer*, 2006.

[5] Alexandre Evfimievski, Ramakrishnan Srikant, Rakesh Agrawal, and Johannes Gehrke. Privacy preserving mining of association rules. *Information Systems*, 29(4):343–364, 2004.

[6] Luisa Franconi and Silvia Polettini. Individual risk estimation in mu-argus: A review. *Lecture notes in computer science*, 3050:262–272, 2004.

[7] A Hundepool, A Van deWetering, R Ramaswamy, L Franconi, A Capobianchi, PP DeWolf, J Domingo-Ferrer, V Torra, R Brand, and S Giessing. μ-argus version 3.2 software and user's manual. statistics netherlands, 2003.

[8] Anco Hundepool and LCRJ Willenborg. μ-and τ-argus: Software for statistical disclosure control. In *Third International Seminar on Statistical Confidentiality*, 1996.

[9] Vijay S Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288. ACM, 2002.

[10] Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60. ACM, 2005.

[11] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007.

[12] Chong K Liew, Uinam J Choi, and Chung J Liew. A data distortion by probability distribution. *ACM Transactions on Database Systems (TODS)*, 10(3):395–411, 1985.

[13] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 24–24. IEEE, 2006.

[14] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228. ACM, 2004.

[15] Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[16] Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information. In *PODS*, volume 98, page 188, 1998.

[17] Latanya Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium*, page 51. American Medical Informatics Association, 1997.

[18] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[19] Nataraj Venkataramanan and Ashwin Shriram. *Data Privacy: Principles and Practice*. CRC Press, 2016.

[20] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[21] William Winkler. Using simulated annealing for k-anonymity. Technical report, Research Report 2002-07, US Census Bureau Statistical Research Division, 2002.

[22] Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd international conference on Very large data bases*, pages 139–150. VLDB Endowment, 2006.

[23] A Zhu. Approximation algorithms for k-anonymity. *Journal of Privacy*, 2005.