ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

Anonymity Preserving Privacy

[I]Trasha Gupta, [II]Monika Gupta
[I]Dept. of Computer Science, Deen Dayal Upadhyaya College, University of Delhi, Delhi, India
[II]Data Engineering Group, Indraprastha Institute of Information Technology, Delhi, India

Abstract

In this paper we focused on identification disclosure of individuals. To prevent this we can anonymize public dataset to limit disclosure of the data records by using different generalization techniques. Many organizations & institutes use public data for their personal interest, it leads to violation of data privacy of some individuals .There are many cases that even after removing private data, such as Name, Address, individual privacy can be comprised by combining attributes from the database. These joined attributes are named as Quasi-identifier. Now, we apply generalization on the Quasi-identifier to make them resemble with each other.

Keywords

Quasi-identifier, Suppression, t-Closeness, K-Anonymity, L-Diversity

I. Introduction

Many organizations are increasing their data continuously that contains un-aggregated information about individuals. Such as information that include medical, voter registration, census, and customer data. These valuable sources of information are used for research, and trend analysis. By this individuals can be uniquely identified by using micro-data and their private information would be disclosed. This is unacceptable with respect to user privacy. To avoid the identification of records in micro-data, unique information like Names and Income is anonymized from the table. However, this ensures the privacy of individuals in the data but even after removing this private information we can re- identified individual records using Quasi -identifiers.

A. PURPOSE: Privacy Preservation of Data

Data mining technology has its own side-effects on privacy. It has been noticed that during research, new things using records of individuals, their privacy can be compromised. So, it has recently become a key research issue and is receiving a growing attention from the research community. However, despite such efforts, a common understanding of what is meant by privacy is still missing, so we are trying to maintain privacy of individuals.

B. SCOPE

Social networks do not currently anonymize their datasets so we can use Quasi-identifiers to breach privacy, therefore all social network data i.e. profile data of the user must be considered as quasi-identifiers, and provide security to social network API calls which allow queries to specify individuals. Even in mobile network to preserve identity of its customer, service provider can use this concept.

II. Methodology

1. Basic Terminology

We have tried to implement a data security paper concept i.e. Incognito: Efficient Full Domain Anonymity. Earlier papers have some disadvantages where data mining results were not accurate, because by using the background knowledge information about the sensitive attribute, when combined with Quasi identifiers can be exploited. In particular, we focus on individual privacy, which is concerned with the anonymity of individuals. So we apply this technique. According to this, we choose a quasi-Identifier Attribute set Q which is a minimal set of attributes that are present in our table and then analyze the frequency set of table with respect to another quasi identifier k which is a mapping from each unique combination and finally apply the generalization. Property relation T is said to satisfy the anonymity property, with respect to attribute set Q if every count in the frequency set of T with respect to Q is greater than or equal to k.

2. Terms and Abbreviations

- Generalization: A method of obtaining anonymity that involves replacing or re-coding a value with a less specific but semantically consistent value. E.g. For attribute Gender, Male and Female can be its two values, now after the process of generalization these two can be replaced with a more generalized term Person.
- Quasi-identifier: A set of indirect identifiers that in combination can uniquely identify individuals but that can't do so in isolation. E.g. if private identifiers such as name and address is deleted, combination of race and gender of a person can also reveal some information about the individual.
- Micro data: This data file store the value of the attribute in the form of attribute codes and see record-level data. Ex {1} {48} {1} {1} {10} {83}
- Suppression: This method is used for obtaining K-anonymity by not releasing a value (e.g. in a table, suppressing several cells) or an entire record (in releases of record-level data.
- Record-level data: Data at the level of an individual person, for example record-level data need not directly identify the data subject but are more vulnerable to re-identification than are aggregate data. Also referred to as micro-data or sometimes considered a quasi- identifier.
- Risk analysis: Evaluating the disclosure risks of each record in anonymized data based on the fact that adversary has some background knowledge & using this knowledge he can still exploit the privacy.

3. Approach

Anonymization of the data is done by the following steps. Initially on original data we apply chosen generalization technique and then check for its utility i.e. whether generalized has distorted original data. After this risk can be evaluated i.e. is it possible for an attacker to reveal some information from the generalized data using some of its background knowledge. If there is a risk, data is further generalized. Finally if it is found that data has been completely anonymized then it is considered as an output.

Our aim is to publish micro-data– tables that contain aggregated information about individuals. Our tables include Age, Gender, Race, Education attributes. We use micro-data that is a valuable source of information for the allocation of anonymous information and trend analysis. However, if individuals can be uniquely identified in the micro-data such that their personal information could be disclosed then this is unacceptable.

To avoid the identification of records in micro data, attributes like Gender, Race, Education salary that can be linked with external data to uniquely identify individuals in the population are called quasi-identifiers. To counter linking attacks using quasi identifiers, we can use a definition of privacy called L- diversity. K-anonymity says that every record in the table is indistinguishable from at least k-1 other records with respect to every set of quasi-identifier attributes; such a table is called a k-anonymous table. Hence, for every combination of values of the quasi-identifiers in the k-anonymous table, there are at least k- Records that share those values. This ensures that individuals cannot be uniquely identified by linking attacks. But L-diversity says that each class in a table should have l-distinct values.

Example: As shown in table 1, we have records from any organization and the table contains no uniquely identifying attributes like "Name, Income" etc. In this example, we divide the attributes into two groups: the "Sensitive" attributes and the "Non-sensitive" attributes at the run time according to the choice of the user. By applying the techniques of L-diversity we are preserving the privacy of our data or providing protection against insider attacks. "Age= 23" means that the age is in the range [20–30].

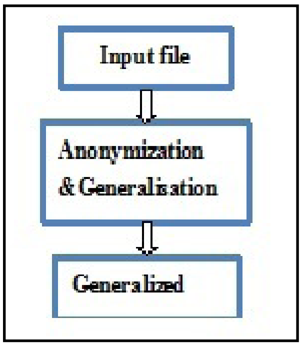


Fig. 1: Process of Anonymization

Table 1: Example

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer gastritis
2	47602	22	4K	stomach cancer
3	47678	27	5K	gastritis flu,bronchitis
4	47905	43	6K	bronchitis pneumonia
5	47909	52	11K	stomach cancer
6	47906	47	8K	
7	47605	30	7K	
8	47673	36	9K	
9	47607	32	10K	

4. Reason for using L-Diversity

While K-anonymity does not protects against identity homogeneity attack and the background knowledge attack, to overcome this problem we can used q*-block to be the set of tuples, due to this limitation of K-anonymity we are using the principle of L-diversity.

5. Reason for using t-Closeness

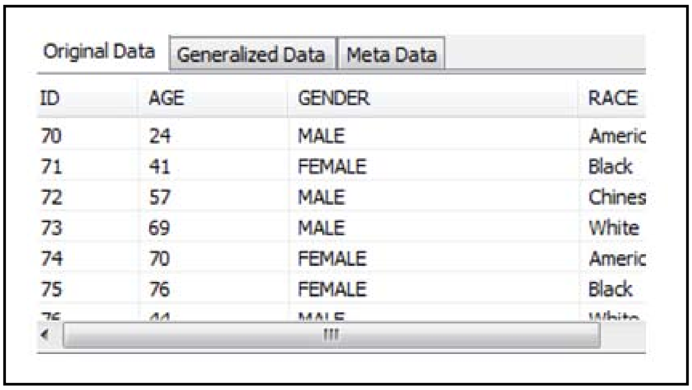
While l-diversity does not protects against attribute disclosure. To overcome this problem we can T-closeness that requires the distribution of a sensitive attribute in any eq. class to be close to the distribution of a sensitive attribute in the overall table.

6. Process Used

By clicking on the "File" first load the data, users can also load the metadata file. Once loaded, the tuples in the dataset will be shown in the "Original Data" table in the upper-Left panel of the user interface are appeared. In addition, users need to choose quasi-identifiers (QI) and sensitive attribute (SA) by clicking "File Initialize". Multiple Quasi identifiers can be chosen, at a given time and then generalized it.

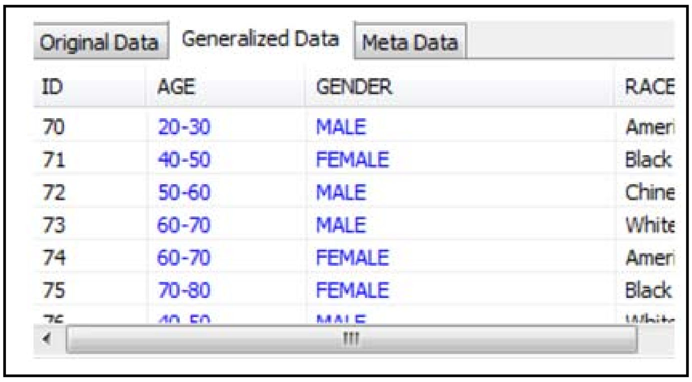
Before Generalization the original data is shown in figure 2. And after generalization the data is shown in figure 3.

Figure 2: Original data



ID	AGE	GENDER	RACE
70	24	MALE	Americ
71	41	FEMALE	Black
72	57	MALE	Chines
73	69	MALE	White
74	70	FEMALE	Americ
75	76	FEMALE	Black
76	44	MALE	White

Fig. 2: Original data



ID	AGE	GENDER	RACE
70	20-30	MALE	Ameri
71	40-50	FEMALE	Black
72	50-60	MALE	Chine
73	60-70	MALE	White
74	60-70	FEMALE	Ameri
75	70-80	FEMALE	Black
76	40-50	MALE	White

Fig. 3: Generalized data

The data need to be anonymized before its public release. The first step of Anonymization is to remove explicit identifiers, which replaces quasi-identifier values with values that are less-specific but semantically consistent & define an equivalence class. In other words, L-diversity requires that each equivalence class contains at least L-1 records, from the figure 2, as shown above.

III. Analysis

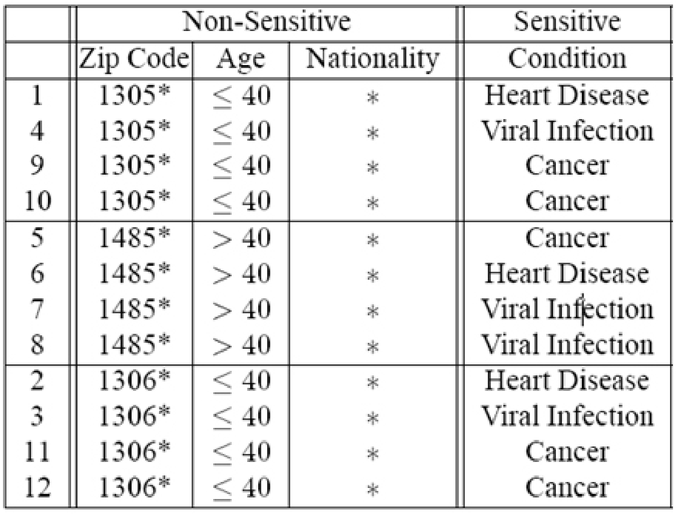
We have done comparative analysis on different types of Anonymization techniques namely K-Anonymization, l-diversity and t-closeness. This has been done by studying different types of attacks possible on each stated technique.

1. Need for L-Diversity

The need for L-diversity is explained for various cases as below.

Attacks against K-Anonymity

Even when sufficient care is taken to identify the quasi-identifier, a solution that adheres to K-Anonymity can still be vulnerable to attacks. Consider the example of a hospital which gathers large amounts of detailed data (micro-data) about patients. Such data can be mined in order to derive certain disease patterns, and can benefit in medical research. However, releasing the micro-data introduces a privacy threat even after removing all directly identifying information, such as name or Income, the data still contain quasi-identifier (QID) attributes (e.g., Age, Education) that can help an attacker to learn the identity of individuals whose personal information is included in the micro-data. To prevent disclosure of sensitive information, the K-anonymity paradigm requires each published record to be indistinguishable with respect to the QID attribute set among an anonymized group of k − 1 other records and removed.



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

Homogeneity Attack

Suppose Robin and Swati are antagonistic neighbors. One day Robin falls ill and is taken by ambulance to the hospital. Having seen the ambulance, Swati sets out to discover what disease Robin is suffering from. Swati discovers the 4-anonymous table of current inpatient records published by the hospital, and so she knows that one of the records in this table contains Robin's data. Since Swati is Robin's neighbor; she knows that Bob is a 31-year-old American male who Education is Post graduates. Therefore, Swati knows that Robin's record number is {9, 10, 11, or 12}.Now, all of those patients have the same medical condition (cancer), and so Swati concludes that Robin has cancer.

Background Knowledge attack: Alice has a pen friend name Vivek who is admitted to the same hospital as Robin, and whose patient records also appear in the table. Swati knows that Vivek is a 21 year old male who currently graduates student. Based on this information, Swati learns that Robin's information is contained in record number 1, 2, 3, or 4. Without additional information, Swati is sure whether Vivek caught a virus or has heart disease.

The solution to the problems above is L-Diversity in which each class has L-Distinct values unlike K-Anonymity.

2. Need for t-closeness

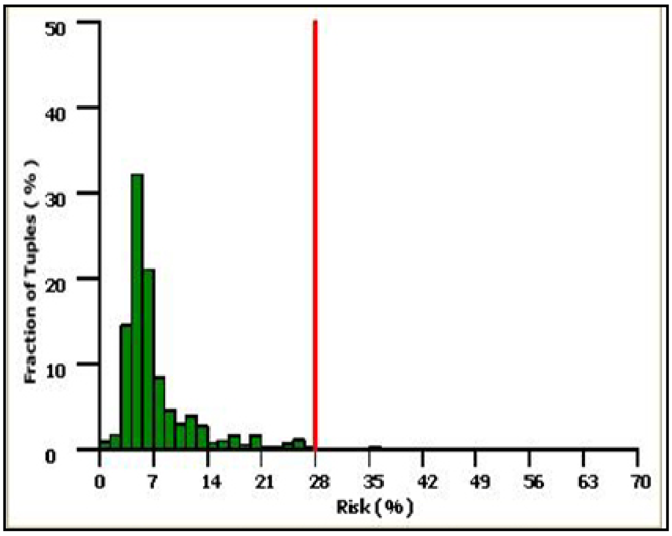
There exist several cases where L-diversity is insufficient to prevent attribute disclosure, so t-closeness is required.

Skew-ness Attack:

When the overall distribution is skewed, satisfying l-diversity does not prevent attribute disclosure because anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population. Now consider an equivalence class that has 49 positive records and only 1 negative record. It would be distinct 2-diverse and has higher entropy than the overall table (and thus satisfies any Entropy diversity that one can impose), even though anyone in the equivalence class would be considered 98% positive, rather than 1% percent. In fact, this equivalence class has exactly the same diversity as a class that has 1 positive and 49 negative records, even though the two classes present very different levels of privacy risks.

3. Risk Analysis

Risk Analysis is defined as evaluating the disclosure risks of each record in anonymized data based on user-specified assumptions about the adversary's background knowledge. In addition, the distribution of the disclosure risks of all records in the dataset can be illustrated in a histogram. To enhance the Anonymization quality, users can eliminate those records with high risk.



IV. Discussion

We have discussed the following issues:

- K-anonymity suffers from background knowledge attack and is thus not safe for Anonymization of data.
- L-Diversity no longer requires knowledge of the full distribution of the sensitive and non-sensitive attributes.
- L-Diversity does not even require the data publisher to have as much information as the adversary.
- The larger the value of ℓ, the more information is needed by adversary to rule out possible values of the sensitive attribute.
- Different adversaries can have different background knowledge leading to different inferences.
- But L-diversity is not able to protect attribute disclosure thus we need t-Closeness for this.
- T-closeness is not able to protect from identity disclosure.

V. Conclusion

In this tool we have shown the functioning of anonymized dataset. It prevents from some strong attacks by apply different techniques in the sensitive attributes. We have introduced a framework that ensures privacy by using Anonymization. Second, although privacy and utility are duals of each other, privacy has received much more attention.

VI. Future Work

There are several avenues for future work. First, we want to extend our initial ideas for handling multiple sensitive attributes, and propose to consider dynamic generalization where an adversary learns access statistics. Also, an examination of social networking "Profile data" as Anonymization functions and try to enforcing l-diversity is necessary. This could lead to stronger notions of anonymity and to notions which can measure the effectiveness of introducing dummy data or dummy queries to further enhance the security of personal data.

Author's Profile

Trasha Gupta received her B. Sc(Hons.) degree in Computer Science from University Of Delhi in 2007. She received her M.Sc degree in 2009. Currently, she is teaching as Assistant Professor in Deen Dayal Upadhyaya College, University Of Delhi. Her research interest includes Data Mining, Natural Language Processing.