Closeness: A New Privacy Measure for Data Publishing

B.SANTHOSH KUMAR*1, M.JOHN SATHYA *2, N.NAVEEN *3

*1 PG Scholar, Anna University of Technology, Tiruchirappalli.
*2 b.santhoshkumar@csice.edu.in
*2 UG SCHOLAR, C.S.I. College of Engineering, Ketti.
*3 mjohnsathya@gmail.com
*2 UG SCHOLAR, C.S.I. College of Engineering, Ketti.
*3 naveenagaraj@gmail.com

ABSTRACT

We show that l-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. Motivated by these limitations, we propose a new notion of privacy called “closeness.” We present the base model t-closeness, which requires that the distribution of a multiple sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). We then propose a more flexible privacy model called on; t-closeness that offers higher utility. We describe our desiderata for designing a distance measure between two probability distributions and present two distance measures. The k-anonymity privacy requirement for publishing micro data requires that each equivalence class (i.e., a set of records that are indistinguishable from each other with respect to certain “identifying” attributes) contains at least k records. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure. The notion of l-diversity has been proposed to address this; l-diversity requires that each equivalence class has at least ‘well-represented values for each multiple sensitive attribute. We propose a novel privacy notion called “closeness.” We formalize the idea of global background knowledge and propose the base model t-closeness which requires that the distribution of a multiple sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t).

INTRODUCTION

GOVERNMENT agencies and other organizations often need to publish micro data, e.g., medical data or census data, for research and other purposes. Typically, such data are stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories: 1) Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number. 2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip code, Birth-date, and Gender. 3) Attributes that are considered sensitive, such as Disease and Salary.

When releasing micro data, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in the literature: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. The privacy preserving data mining problem has gained considerable importance in recent years because of the vast amounts of personal data about individuals stored at different commercial vendors and organizations.

Privacy-preserving data mining has emerged as a very important issue to be addressed in recent times. This is because of the ability to store data of users had increased , use of social networks helps in yielding personal information , sophisticated data mining algorithms and high computational powers available with the adversary. All this makes it
possible to leverage this information. Although most of the applications rest remove the records having sensitive information like name, social security numbers (or any other unique identification number), other kind of attributes like sex, age, pin codes, profession can be combined to form a pseudo-identifier and the sensitive information can then be retrieved from public data records like census which contain all records.

### TABLE 1
Original Patients Table

<table>
<thead>
<tr>
<th>ZIP Code</th>
<th>Age</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47677</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>2</td>
<td>47602</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>3</td>
<td>47678</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>4</td>
<td>47905</td>
<td>Flu</td>
</tr>
<tr>
<td>5</td>
<td>47909</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>6</td>
<td>47906</td>
<td>Cancer</td>
</tr>
<tr>
<td>7</td>
<td>47605</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>8</td>
<td>47673</td>
<td>Cancer</td>
</tr>
<tr>
<td>9</td>
<td>47607</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

Privacy-preserving data mining has emerged as a very important issue to be addressed in recent times. This is because of the ability to store data of users had increased, use of social networks helps in yielding personal information, sophisticated data mining algorithms and high computational powers available with the adversary. All this makes it possible to leverage this information. Although most of the applications first remove the records having sensitive information like name, social security numbers (or any other unique identification number), other kind of attributes like sex, age, pin codes, profession can be combined to form a pseudo-identifier and the sensitive information can then be retrieved from public data records like census which contain all records.

### K-ANONYMITY

The k-anonymity model requires that within any equivalence class of the micro data there are at least k records. In other words we should not be able to make ANY query to the database which returns less than k matches. Achieving k-anonymity is provided by use of generalization relationships between domains and between values that attributes can assume. Suppression is a complementary approach to providing k-anonymity.

Today's globally networked society places great demand on the dissemination and sharing of information, which is probably becoming the most important and demanded resource. While in the past released information was mostly in tabular and statistical form (macro data), many situations call today for the release of specific data (micro data). Micro data, in contrast to macro data reporting precomputed statistics, provide the convenience of allowing the final recipient to perform on them analysis as needed. To protect the anonymity of the entities, called respondents, to which Micro data undergoing public or semipublic release refer, data holders often remove or encrypt explicit identifiers such as names, addresses, and phone numbers. De-identifying data, however, provides no guarantee of anonymity. Released information often contains other data, such as race, birth date, sex, and ZIP code, which can be linked to publicly available information to re-identify (or restrict the uncertainty about) the data respondents, thus leaking information that was not intended for disclosure.

The large amount of information easily accessible today, together with the increased computational power available to the attackers, makes such linking attacks a serious problem. Indeed, the restricted access to information and its expensive processing, which represented a form of protection in the past, do not hold anymore.
Information about us is collected every day, as we join associations or groups, shop for groceries, or executes most of our common daily activities; the amount of privately owned records that describe each citizen's finances, interests, and demographics is increasing every day. Information bureaus such as TRW, Equifax, and Trans Union hold the largest and most detailed databases on American consumers. Most municipalities sell population registers that include the identities of individuals along with basic demographics; examples include local census data, voter lists, city directories, and information from motor vehicle agencies, tax assessors, and real estate agencies.

The k-anonymity model assumes that person-specific data are stored in a table (or a relation) of columns (or attributes) and rows (or records). The process of anonymizing such a table starts with removing all the explicit identifiers, such as name and SSN, from it. However, even though a table is free of explicit identifiers, some of the remaining attributes in combination could be specific enough to identify individuals. For example, as shown by Sweeney, 87% of individuals in the United States can be uniquely identified by a set of attributes such as {ZIP, gender, date of birth}. This implies that each attribute alone may not be specific enough to identify individuals, but a particular group of attributes could be. Thus, disclosing such attributes, called quasi-identifier, may enable potential adversaries to link records with the corresponding individuals.

**Definition 1. (Quasi-identifier)** A quasi-identifier of table T, denoted as QT, is a set of attributes in T that can be potentially used to link a record in T to a real-world identity with a significant probability. The main objective of the k-anonymity problem is thus to transform a table so that no one can make high-probability associations between records in the table and the corresponding entity instances by using quasi-identifier.

**Definition 2. (k-anonymity requirement)** Table T is said to be k-anonymous with respect to quasi-identifier QT if and only if for every record r in T there exist at least (k − 1) other records in T that are indistinguishable from r with respect to QT. By enforcing the k-anonymity requirement, it is guaranteed that even though an adversary knows that a k-anonymous table T contains the record of a particular individual and also knows the quasi-identifier value of the individual, he cannot determine which record in T corresponds to the individual with a probability greater than 1/k. The k-anonymity requirement is typically enforced through generalization, where real values are replaced with “less specific but semantically consistent values”. Given a domain, there are various ways to generalize the values in the domain. Commonly, numeric values are generalized into intervals, and categorical values into a set of distinct values (e.g., \{USA, Canada\}) or a single value that represents such a set (e.g., North-America). A group of records that are indistinguishable from each other is often referred to as an *equivalence class*.

**T-closeness: A new privacy measure**

Intuitively, privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the multiple sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief. The novelty of our approach is that we separate the information gain into two parts: that about the whole population in the released data and that about specific individuals.

**Limitations of T-closeness**

There is no computational procedure to enforce t-closeness followed in. There is effective way till now of combining with generalizations and suppressions or slicing. Lost co-relation between different attributes: This is because each attribute is generalized separately and so we lose their dependence on each other. Utility of data is damaged if we use very small t. (And small t will result in increase in computational time.)

**Earth mover’s distance**

In probability theory, the *earth mover's distance* (EMD) is a measure of the distance between two probability distributions over a region $D$. Informally, if the distributions are interpreted as two different ways of piling up a certain amount of dirt over the region $D$, the EMD is the minimum cost of turning one pile into the
other; where the cost is assumed to be amount of dirt moved times the distance by which it is moved.

The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. Intuitively, one distribution is seen as a mass of earth spread in the space and the other as a collection of holes in the same space. EMD measures the least amount of work needed to fill the holes with earth. A unit of work corresponds to moving a unit of earth by a unit of ground distance. EMD can be formally defined using the well-studied transportation problem. Let

\[ P = (p_1, p_2, \ldots, p_m), Q = (q_1, q_2, \ldots, q_m), \]

and \( dij \) be the ground distance between element \( i \) of \( P \) and element \( j \) of \( Q \). We want to find a flow \( F \) such that:

\[ \text{WORK}(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}, \]

subject to the following constraints:

1. \( f_{ij} \geq 0 \)
2. \( 1 \leq i \leq m, 1 \leq j \leq m \)
3. \( p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i, \quad 1 \leq i \leq m \)
4. \( \sum_{i=1}^{m} \sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1. \)

The EMD is defined as the minimum amount of work needed to change one signature into the other. The notion of "work" is based on the user-defined ground distance which is the distance between two features. The size of the two signatures can be different. Also, the sum of weights of one signature can be different than the sum of weights of the other (partial match). Because of this, the EMD is normalized by the smaller sum.

**RELATED WORK**

In this section, we briefly survey existing literature that addresses data privacy. Instead of providing a comprehensive survey, we discuss various aspects of data privacy. Note that Ensuring privacy in published data has been a difficult problem for a long time, and this problem has been studied in various aspects. In, Lambert provides informative discussion on the risk and harm of undesirable disclosures and discusses how to evaluate a dataset in terms of these risk and harm.

In, Dalenius poses the problem of re-identification in (supposedly) anonymous census records and firstly introduces the notion of “quasi-identifier”. He also suggests some ideas such as suppression or encryption of data as possible solutions. Data privacy has been extensively addressed in statistical databases, which primarily aim at preventing various inference channels. One of the common techniques is data perturbation which mostly involves swapping data values or introducing noise to the dataset. While the perturbation is applied in a manner which preserves statistical characteristics of the original data, the transformed dataset is useful only for statistical research. Another important technique is query restriction, which restricts queries that may result in inference. In this approach, queries are restricted by various criteria such as query-set-size, query-history, and partitions. Although this approach can be effective, it requires the protected data to remain in a dedicated database at all time.

**CONCLUSION AND FUTURE ENHANCEMENTS**

We propose two instantiations: a base model called t-closeness and a more flexible privacy model called (n-t)-closeness. We explain the rationale of the (n-t)-closeness model and show that it achieves a better balance between privacy and utility. To incorporate semantic distance, we choose to use the Earth Mover Distance measure. We also point out the limitations of EMD, present the desiderata for designing the distance measure, and propose a new distance measure that meets all the requirements. Finally, through experiments on real data, we show that similarity attacks are a real concern and the (n-t)-closeness model better protects the data while improving the utility of the released data.

(n-t)-closeness allows us to take advantage of anonymization techniques other than generalization of quasi-identifier and suppression of records. For example, instead of suppressing a whole record, one can hide some sensitive attributes of the record; one advantage is that the number of records in the anonymized table is accurate, which may be useful in some applications. Because this technique does not affect quasi-identifiers, it does not help achieve k-anonymity, and hence, has not been considered before. Removing a sensitive value in a group reduces diversity, and therefore, it does
not help in achieving ‘-diversity. However, in t-
closeness, removing an outlier may smooth a
distribution and bring it closer to the overall
distribution. Another possible technique is to
generalize a sensitive attribute value, rather than
hiding it completely. An interesting question is how
to effectively combine these techniques with
generalization and suppression to achieve better
data quality.

REFERENCES

Data Bases (VLDB), pp. 901-909,
2005.
Khuller, R. Panigrahy, D. Thomas, and A. Zhu,
“Achieving Anonymity via Clustering,” Proc. ACM
Network Flows: Theory, Algorithms, and
through Optimal k- Anonymization,” Proc. Int’l
Conf. Artificial Intelligence (AAAII), pp. 602-608,
“Secure Anonymization for Incremental Datasets,”
Proc. VLDB Workshop Secure Data Management
“Privacy Skyline: Privacy with Multidimensional
Large Data Bases (VLDB), pp. 770-781, 2007.
[8] G.T. Duncan and D. Lambert, “Disclosure-
Down Specialization for Information and Privacy
Wasserstein Metrics for Probability Distributions,”
Michigan Math J., vol. 31, pp. 231-240,
1984.
Zhang, “Aggregate Query Answering on
Information and Sufficiency,” Annals of Math.
Harm,” J. Official Statistics, vol. 9, pp. 313-331,
1993.