INTERNATIONAL JOURNAL OF RESEARCH IN COMMERCE, IT & MANAGEMENT

IJRCM

A Monthly Double-Blind Peer Reviewed (Refereed/Juried) Open Access International e-Journal - Included in the International Serial Directories

Indexed & Listed at:

Ulrich's Periodicals Directory ©, ProQuest, U.S.A., EBSCO Publishing, U.S.A., Cabell's Directories of Publishing Opportunities, U.S.A.,

Open J-Gage, India [link of the same is duly available at Inflibnet of University Grants Commission (U.G.C.)],

Index Copernicus Publishers Panel, Polandwith IC Value of 5.09 &number of libraries all around the world.

Circulated all over the world & Google has verified that scholars of more than 2401 Cities in 155 countries/territories are visiting our journal on regular basis.

Ground Floor, Building No. 1041-C-1, Devi Bhawan Bazar, JAGADHRI – 135 003, Yamunanagar, Haryana, INDIA

http://ijrcm.org.in/

CONTENTS

CALL FOR MANUSCRIPTS

Weinvite unpublished novel, original, empirical and high quality research work pertaining to recent developments & practices in the area of Computer, Business, Finance, Marketing, Human Resource Management, General Management, Banking, Education, Insurance, Corporate Governance and emerging paradigms in allied subjects like Accounting Education; Accounting Information Systems; Accounting Theory & Practice; Auditing; Behavioral Accounting; Behavioral Economics; Corporate Finance; Cost Accounting; Econometrics; Economic Development; Economic History; Financial Institutions & Markets; Financial Services; Fiscal Policy; Government & Non Profit Accounting; Industrial Organization; International Economics & Trade; International Finance; Macro Economics; Micro Economics; Monetary Policy; Portfolio & Security Analysis; Public Policy Economics; Real Estate; Regional Economics; Tax Accounting; Advertising & Promotion Management; Business Education; Management Information Systems (MIS); Business Law, Public Responsibility & Ethics; Communication; Direct Marketing; E-Commerce; Global Business; Health Care Administration; Labor Relations & Human Resource Management; Marketing Research; Marketing Theory & Applications; Non-Profit Organizations; Office Administration/Management; Operations Research/Statistics; Organizational Behavior & Theory; Organizational Development; Production/Operations; Public Administration; Purchasing/Materials Management; Retailing; Sales/Selling; Services; Small Business Entrepreneurship; Strategic Management Policy; Technology/Innovation; Tourism, Hospitality & Leisure; Transportation/Physical Distribution; Algorithms; Artificial Intelligence; Compilers & Translation; Computer Aided Design (CAD); Computer Aided Manufacturing; Computer Graphics; Computer Organization & Architecture; Database Structures & Systems; Digital Logic; Discrete Structures; Internet; Management Information Systems; Modeling & Simulation; Multimedia; Neural Systems/Neural Networks; Numerical Analysis/Scientific Computing; Object Oriented Programming; Operating Systems; Programming Languages; Robotics; Symbolic & Formal Logic and Web Design. The above mentioned tracks are only indicative, and not exhaustive.

Anybody can submit the soft copy of his/her manuscript **anytime** in M.S. Word format after preparing the same as per our submission guidelines duly available on our website under the heading guidelines for submission, at the email address: infoijrcm@gmail.com.

GUIDELINES FOR SUBMISSION OF MANUSCRIPT

1. **COVERING LETTER FOR SUBMISSION**:

DATED: _____________

THE EDITOR
IJRCM

Subject: **SUBMISSION OF MANUSCRIPT IN THE AREA OF.**

(e.g. Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)

DEAR SIR/MADAM

Please find my submission of manuscript entitled '____' for possible publication in your journals.

I hereby affirm that the contents of this manuscript are original. Furthermore, it has neither been published elsewhere in any language fully or partly, nor is it under review for publication elsewhere.

I affirm that all the author (s) have seen and agreed to the submitted version of the manuscript and their inclusion of name (s) as co-author (s).

Also, if my/our manuscript is accepted, I/We agree to comply with the formalities as given on the website of the journal & you are free to publish our contribution in any of your journals.

NAME OF CORRESPONDING AUTHOR:
Designation:
Affiliation with full address, contact numbers & Pin Code:
Residential address with Pin Code:
Mobile Number (s):
Landline Number (s):
E-mail Address:
Alternate E-mail Address:

NOTES:

a) The whole manuscript is required to be in ***ONE MS WORD FILE*** only (pdf. version is liable to be rejected without any consideration), which will start from the covering letter, inside the manuscript.
b) The sender is required to mentionthe following in the **SUBJECT COLUMN** of the mail:
New Manuscript for Review in the area of (Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)
c) There is no need to give any text in the body of mail, except the cases where the author wishes to give any specific message w.r.t. to the manuscript.
d) The total size of the file containing the manuscript is required to be below **500 KB**.
e) Abstract alone will not be considered for review, and the author is required to submit the complete manuscript in the first instance.
f) The journal gives acknowledgement w.r.t. the receipt of every email and in case of non-receipt of acknowledgment from the journal, w.r.t. the submission of manuscript, within two days of submission, the corresponding author is required to demand for the same by sending separate mail to the journal.

2. **MANUSCRIPT TITLE**: The title of the paper should be in a 12 point Calibri Font. It should be bold typed, centered and fully capitalised.

3. **AUTHOR NAME (S) & AFFILIATIONS**: The author (s) **full name**, **designation**, **affiliation** (s), **address**, **mobile/landline numbers**, and **email/alternate email address** should be in italic & 11-point Calibri Font. It must be centered underneath the title.

4. **ABSTRACT**: Abstract should be in fully italicized text, not exceeding 250 words. The abstract must be informative and explain the background, aims, methods, results & conclusion in a single para. Abbreviations must be mentioned in full.

5. **KEYWORDS**: Abstract must be followed by a list of keywords, subject to the maximum of five. These should be arranged in alphabetic order separated by commas and full stops at the end.

6. **MANUSCRIPT**: Manuscript must be in ***BRITISH ENGLISH*** prepared on a standard A4 size ***PORTRAIT SETTING PAPER***. It must be prepared on a single space and single column with 1" margin set for top, bottom, left and right. It should be typed in 8 point Calibri Font with page numbers at the bottom and centre of every page. It should be free from grammatical, spelling and punctuation errors and must be thoroughly edited.

7. **HEADINGS**: All the headings should be in a 10 point Calibri Font. These must be bold-faced, aligned left and fully capitalised. Leave a blank line before each heading.

8. **SUB-HEADINGS**: All the sub-headings should be in a 8 point Calibri Font. These must be bold-faced, aligned left and fully capitalised.

9. **MAIN TEXT**: The main text should follow the following sequence:

   INTRODUCTION

   REVIEW OF LITERATURE

   NEED/IMPORTANCE OF THE STUDY

   STATEMENT OF THE PROBLEM

   OBJECTIVES

   HYPOTHESES

   RESEARCH METHODOLOGY

   RESULTS & DISCUSSION

   FINDINGS

   RECOMMENDATIONS/SUGGESTIONS

   CONCLUSIONS

   SCOPE FOR FURTHER RESEARCH

   ACKNOWLEDGMENTS

   REFERENCES

   APPENDIX/ANNEXURE

   It should be in a 8 point Calibri Font, single spaced and justified. The manuscript should preferably not exceed ***5000 WORDS***.

10. **FIGURES &TABLES**: These should be simple, crystal clear, centered, separately numbered & self explained, and **titles must be above the table/figure**. **Sources of data should be mentioned below the table/figure**. It should be ensured that the tables/figures are referred to from the main text.

11. **EQUATIONS**:These should be consecutively numbered in parentheses, horizontally centered with equation number placed at the right.

12. **REFERENCES**: The list of all references should be alphabetically arranged. The author (s) should mention only the actually utilised references in the preparation of manuscript and they are supposed to follow **Harvard Style of Referencing**. The author (s) are supposed to follow the references as per the following:

- All works cited in the text (including sources for tables and figures) should be listed alphabetically.
- Use (**ed.**) for one editor, and (**ed.s)** for multiple editors.
- When listing two or more works by one author, use --- (20xx), such as after Kohl (1997), use --- (2001), etc, in chronologically ascending order.
- Indicate (opening and closing) page numbers for articles in journals and for chapters in books.
- The title of books and journals should be in italics. Double quotation marks are used for titles of journal articles, book chapters, dissertations, reports, working papers, unpublished material, etc.
- For titles in a language other than English, provide an English translation in parentheses.
- The location of endnotes within the text should be indicated by superscript numbers.

PLEASE USE THE FOLLOWING FOR STYLE AND PUNCTUATION IN REFERENCES:

BOOKS

- Bowersox, Donald J., Closs, David J., (1996), "Logistical Management." Tata McGraw, Hill, New Delhi.
- Hunker, H.L. and A.J. Wright (1963), "Factors of Industrial Location in Ohio" Ohio State University, Nigeria.

CONTRIBUTIONS TO BOOKS

- Sharma T., Kwatra, G. (2008) Effectiveness of Social Advertising: A Study of Selected Campaigns, Corporate Social Responsibility, Edited by David Crowther & Nicholas Capaldi, Ashgate Research Companion to Corporate Social Responsibility, Chapter 15, pp 287-303.

JOURNAL AND OTHER ARTICLES

- Schemenner, R.W., Huber, J.C. and Cook, R.L. (1987), "Geographic Differences and the Location of New Manufacturing Facilities," Journal of Urban Economics, Vol. 21, No. 1, pp. 83-104.

CONFERENCE PAPERS

- Garg, Sambhav (2011): "Business Ethics" Paper presented at the Annual International Conference for the All India Management Association, New Delhi, India, 19–22 June.

UNPUBLISHED DISSERTATIONS AND THESES

- Kumar S. (2011): "Customer Value: A Comparative Study of Rural and Urban Customers," Thesis, Kurukshetra University, Kurukshetra.

ONLINE RESOURCES

- Always indicate the date that the source was accessed, as online resources are frequently updated or removed.

WEBSITES

- Garg, Bhavet (2011): Towards a New Natural Gas Policy, Political Weekly, Viewed on January 01, 2012 http://epw.in/user/viewabstract.jsp

A NEW NOTION PROXIMITY FOR DATA PUBLISHING WITH PRIVACY PRESERVATION

S.BOOPATHY
ASST. PROFESSOR
SNS COLLEGE OF ENGINEERING
COIMBATORE

P.SUMATHI
ASST. PROFESSOR (SG)
SNS COLLEGE OF ENGINEERING
COIMBATORE

ABSTRACT

Publishing data about individuals without revealing sensitive information about them is an important problem. k-anonymity has been proposed as a mechanism for protecting privacy in microdata publishing. In recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure. To address this limitation the notion of l-diversity has been proposed, which requires the distribution of a sensitive attribute in each equivalence class has at least l well represented values. k-anonymity and l-diversity make it harder for the attacker to figure out private associations. But they still give away some knowledge and they do not give any guarantees on the amount of data being disclosed. Here, our analysis shows that k-anonymity and l-diversity has a number of limitations like homogeneity attack and background knowledge of the attacker. Motivated by these limitations in k-anonymity and l-diversity, we propose a novel notion of privacy called "Proximity", in which we first present the base model "t-closeness", which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. Then secondly the flexible privacy model called (n, t)- Proximity is proposed. Finally we describe the desiderata for designing the distance measure and the Earth Mover's Distance measure is used to measure the distance between two probabilistic distributions.

KEYWORDS

privacy preservation, data publishing, data security, data anonymization.

1. INTRODUCTION

Agencies and many other organizations often need to publish microdata – tables that contain unaggregated information about individuals. These tables can include medical data, voter registration data, census data and customer data. Microdata is a valuable source of information for the research and allocation of public funds, trend analysis and medical research. Typically, such data is stored in a table, and each record or row corresponds to one individual. Each record has a number of attributes, which can be classified into explicit identifiers, quasi-identifiers and sensitive attributes.

When releasing microdata, it is necessary to prevent the sensitive information of individuals from being disclosed. Two types of information disclosure have been identified in the literature [1], [2]: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed. While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove the explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers.

We need to measure the disclosure risk of an anonymized table to effectively limit disclosure. Samarati [3] and Sweeney [4] introduced k-anonymity, which requires that each equivalence class contains at least k-records. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. To address this limitation k-anonymity, Machanavajjhala [5] introduced a new privacy, called l-diverstiy, which requires that the distribution of a sensitive attribute in each equivalence class has at least l well-represented values. The problem with the l-diverstiy is that it is limited in its assumption of adversarial knowledge. Meaning, it is possible for an adversary to gain information about a sensitive attribute as long as he has information about the global distribution of this attribute. This assumption generalizes the specific background and homogeneity attacks used to motivate l-diverstiy. In general, a problem with privacy preserving methods, is that they effectively assume all attributes to be categorical; the adversary either does or does not learn something sensitive.

Here, in this paper, we propose a novel a new privacy notion called "proximity." We first formalize the idea of global background knowledge and use as the base model as t-closeness. This model requires that the distribution of sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table. The distance between the two distributions should be no more than a threshold t. This effectively limits the amount of individual-specific information an observer can learn. However, an analysis on data utility shows that t-closeness substantially limits the amount of useful information that can be extracted from the released data. Based on the analysis, we propose a more flexible privacy model called (n, t)-proximity, which requires that the distribution in any equivalence class is close to the distribution in a large-enough equivalence class with respect to the sensitive attribute. That is the distance between the two distributions should be no more than a threshold t and the equivalence class contains at least n records. This limits the amount of sensitive information about individuals while preserves features and patterns about large groups. Our analysis shows that (n, t)-proximity achieves a better balance between privacy and utility than existing privacy models such as k-anonymity, l-diversity and t-closeness.

Measuring the values of distance between sensitive attributes is a core task. Here we use the Earth Mover Distance metric [6] to measure the distance between two distributions. We also show that Earth Mover Distance has its limitations and describe our desiderata for designing the distance measure. Finally, we evaluate the effectiveness of the (n, t)-proximity model in both privacy protection and utility preservation through experiments on a real data set.

2. K-ANONYMITY AND L-DIVERSITY

The protection k-anonymity provides is simple and easy to understand. If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence greater than 1/k. While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. This has been recognized by several authors [5], [7], [8]. Two attacks were identified in [5]: the homogeneity attack and the background knowledge attack. Let us first show the two attacks with an example to give the intuition behind the problems with k-anonymity.

Example 1. Table 1 shows medical records from a fictitious hospital located in upstate New York. Note that the table contains no uniquely identifying attributes like name, social security number, etc. In this example, we divide the attributes into two groups: the sensitive attributes (consisting only of medical condition/disease) and the non-sensitive attributes (zip code, age). An attribute is marked sensitive if an adversary must not be allowed to discover the value of that attribute for any individual in the dataset. Attributes not marked sensitive are non-sensitive. Furthermore, let the collection of attributes {zip code, age} be the quasi-identifier for this dataset. Table 2 shows a 3-anonymous table derived from the table 1 (here "*" denotes a suppressed value so, for example, "zip code = 1485*" means that the zip code is in the range (14850 − 14859) and "age=3*"means the age is in the range (30 – 39)). Note that in the 3-anonymous table, each tuple has the same values for the quasi-identifier as at least two other tuples in the table.

TABLE 1: ORIGINAL INPATIENT MICRODATA

	Non-Sensitive		Sensitive
	Zip Code	Age	Disease
1	98677	29	Heart Disease
2	98602	22	Heart Disease
3	98678	27	Heart Disease
4	98905	43	Flu
5	98909	52	Heart Disease
6	98906	47	Cancer
7	98605	30	Heart Disease
8	98673	36	Cancer
9	98607	32	Cancer

TABLE 2: A 3-ANONYMOUS INPATIENT MICRODATA

	Non-Sensitive		Sensitive
	Zip Code	Age	Disease
1	986*	2*	Heart Disease
2	986*	2*	Heart Disease
3	986*	2*	Heart Disease
4	9890*	≥ 40	Flu
5	9890*	≥ 40	Heart Disease
6	9890*	≥ 40	Cancer
7	986*	3*	Heart Disease
8	986*	3*	Cancer
9	986*	3*	Cancer

2.1 ATTACKS ON K-ANONYMITY

Homogeneity Attack: Suppose John knows that James is a 27-year man living in ZIP 98678 and James's record is in the table. From Table 2, John can conclude that James corresponds to one of the first three records, and thus, must have heart disease.

Observation 1 k-Anonymity can create groups that leak information due to lack of diversity in the sensitive attribute.

Background Knowledge Attack: Suppose that by knowing Roy's age and zip code, John can conclude that Roy corresponds to a record in the last equivalence class Table 2. Furthermore, suppose John knows that Roy has a very low risk for heart disease. This background knowledge enables John to conclude that Roy most likely has cancer.

Observation 2 k-Anonymity does not protect against attacks based on background knowledge.

To address these limitations of k-anonymity, Machanavajjhala [5] introduced l-diverstiy as a stronger notion of privacy.

2.2 THE L-DIVERSITY PRINCIPLE

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. Machanavajjhala [5] gave a number of interpretations of the term "well represented" in this principle: Distinct l-diversity, Probabilistic l-diversity, Entropy l-diversity and Recursive (c, l)-diversity.

3 THE L-DIVERSITY LIMITATIONS

While the l-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it has several shortcomings that we now discuss.

l-diversity may be difficult to achieve and may not provide sufficient privacy protection.

Example 2. Suppose that the original data have only one sensitive attribute: the test result for a particular virus. It takes two values: positive and negative. Further, suppose that there are 10000 records, with 99 percent of them being negative, and only 1 percent being positive. Then, the two values have very different degrees of sensitivity. One would not mind being known to be tested negative, because then one is the same as 99 percent of the population, but one would not want to be known/considered to be tested positive. In this case, 2-diversity does not provide sufficient privacy protection for an equivalence class that contains only records that are negative. In order to have a distinct 2-diverse table, there can be at most 10000 x 1% =100 equivalence classes and the information loss would be large. Also, observe that because the entropy of the sensitive attribute in the overall table is very small, if one uses entropy l-diversity, l must be set to a small value.

l-diversity is insufficient to prevent attribute disclosure. Below, we present two attacks on l-diversity.

Skewness attack: When the overall distribution is skewed, satisfying that l-diversity does not prevent attribute disclosure. Consider again Example 2. Suppose that one equivalence class has an equal number of positive records and negative records. It satisfies distinct 2-diversity, entropy 2-diversity, and any recursive (c, 2)-diversity requirement that can be imposed. However, this presents a serious privacy risk, because anyone in the class would be considered to have 50 percent possibility of being positive, as compared with the 1 percent 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 l-diversity that one can impose), even though anyone in the equivalence class would be considered 98 percent 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

Similarity attack: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. Consider the following example:

Example 3. Table 3 is the original table, and Table 4 shows an anonymized version satisfying distinct and entropy 3-diversity. There are two sensitive attributes: Salary and Disease. Suppose one knows that James's record corresponds to one of the first three records, then one knows that James's salary is in the range [3K-5K] and can infer that James's salary is relatively low. This attack applies not only to numeric attributes like "Salary," but also to categorical attributes like "Disease." Knowing that James's record belongs to the first.

TABLE 3: ORIGINAL SALARY/DISEASE MICRODATA

	Non-Sensitive		Sensitive	
	Zip Code	Age	Salary	Disease
1	98677	29	3K	Gastric Ulcer
2	98602	22	4K	Gastritis
3	98678	27	5K	Stomach Cancer
4	98905	43	6K	Gastritis
5	98909	52	11K	Flu
6	98906	47	8K	Bronchitis
7	98605	30	7K	Bronchitis
8	98673	36	9K	Pneumonia
9	98607	32	10K	Stomach Cancer

TABLE 4: A 3-DIVERSE SALARY/DISEASE MICRODATA

	Non-Sensitive		Sensitive	
	Zip Code	Age	Salary	Disease
1	986**	2*	3K	Gastric Ulcer
2	986**	2*	4K	Gastritis
3	986**	2*	5K	Stomach Cancer
4	9890*	≥ 40	6K	Gastritis
5	9890*	≥ 40	11K	Flu
6	9890*	≥ 40	8K	Bronchitis
7	986**	3*	7K	Bronchitis
8	986**	3*	9K	Pneumonia
9	986**	3*	10K	Stomach Cancer

equivalence class enables one to conclude that James has some stomach-related problems, because all three diseases in the class are stomach-related. This leakage of sensitive information occurs because while l-diversity requirement ensures "diversity" of sensitive values in each group, it does not take into account the semantical closeness of these values. In short, distributions that have the same level of diversity may provide very different levels of privacy, because there are semantic relationships among the attribute values, because different values have very different levels of sensitivity, and privacy is also affected by the relationship with the overall distribution.

4 A NEW PRIVACY MEASURE PROXIMITY

Intuitively, privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the 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 new thing of our approach is that we separate the information gain into two parts: that about the population in the released data and about specific individuals.

4.1 BASE MODEL: t-CLOSENESS

To motivate our approach, let us perform the following thought experiment: First, an observer has some prior belief B_0 about an individual's sensitive attribute. Then, in a hypothetical step, the observer is given a completely generalized version of the data table where all attributes in a quasi-identifier are removed (or, equivalently, generalized to the most general values). The observer's belief is influenced by Q, the distribution of the sensitive attribute values in the whole table, and changes to belief B_1. Finally, the observer is given the released table. By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in, and learn the distribution P of sensitive attribute values in this class. The observer's belief changes to B_2.

The l-diversity requirement is motivated by limiting the difference between B_0 and B_2 (although it does so only indirectly, by requiring that P has a level of diversity). We choose to limit the difference between B_1 and B_2. In other words, we assume that Q, the distribution of the sensitive attribute in the overall population in the table, is public information. We do not limit the observer's information gain about the population as a whole, but limit the extent to which the observer can learn additional information about specific individuals. To justify our assumption that Q should be treated as public information, we observe that with generalizations, the most one can do is to generalize all quasi-identifier attributes to the most general value.

Thus, as long as a version of the data is to be released, a distribution Q will be released. We also argue that if one wants to release the table at all, one intends to release the distribution Q and this distribution is what makes data in this table useful. In other words, one wants Q to be public information. A large change from B_0 to B_1 means that the data table contains a lot of new information, e.g., the new data table corrects some widely held belief that was wrong. In some sense, the larger the difference between B_0 and B_1 is, the more valuable the data is. Since the knowledge gain between B_0 and B_1 is about the population the data set is about, we do not limit this gain. We limit the gain from B_1 to B_2 by limiting the distance between P and Q. Intuitively, if P = Q, then B_1 and B_2 should be the same. If P and Q are close, then B_1 and B_2 should be close as well, even if B_0 may be very different from both B_1 and B_2.

4.2 A FLEXIBLE PRIVACY MODEL (N, T)-PROXIMITY

First, we illustrate that t-closeness limits the release of useful information through the following example.

Example 4. Table 5 is the original data table containing 3000 individuals, and Table 6 is an anonymized version of it. The Disease attribute is sensitive and there is a column called Count that indicates the number of individuals. The probability of cancer among the population in the data set is 700 / 3 = 0.23, while the probability of cancer among individuals in the first equivalence class is as high as 300 / 600 = 0.5. Since 0.5 - 0.23 > 0.1, the anonymized table does not satisfy 0.1-closeness.

To achieve 0.1-closeness, all tuples in Table 5 have to be generalized into a single equivalence class. This results in substantial information loss. If we examine the original data in Table 5, we can discover that the probability of cancer among people living in zip code 986** is as high as 500 / 1000 = 0.5, while the probability of cancer among people living in zip code 989** is only 200 / 2000 = 0.1. The important fact that people living in zip code 986** have a much higher rate of cancer will be hidden if 0.1-closeness is enforced.

TABLE 5: ORIGINAL INPATIENTS MICRODATA

	Non-Sensitive		Sensitive	Non-Sensitive
	ZIP Code	Age	Disease	Count
1	98673	29	Cancer	100
2	98674	21	Flu	100
3	98605	25	Cancer	200
4	98602	23	Flu	200
5	98905	43	Cancer	100
6	98904	48	Flu	900
7	98906	47	Cancer	100
8	98907	41	Flu	900
9	98603	34	Cancer	100
10	98605	30	Flu	100
11	98602	36	Cancer	100
12	98607	32	Flu	100

TABLE 6: AN ANONYMOUS VERSION OF TABLE 5 (0.1-CLOSENESS)

	Non-Sensitive		Sensitive	Non-Sensitive
	ZIP Code	Age	Disease	Count
1	986**	2*	Cancer	300
2	986**	2*	Flu	300
3	989**	4*	Cancer	200
4	989**	4*	Flu	1800
5	986**	3*	Cancer	200
6	986**	3*	Flu	200

The (n, t)-Proximity Principle: An equivalence class E_1 is said to have (n, t)-proximity if there exists a set E_2 of records that is a natural superset of E_1 such that E_2 contains at least n records, and the distance between the two distributions of the sensitive attribute in E_1 and E_2 is no more than a threshold t. A table is said to have (n, t)-proximity if all equivalence class have (n, t)-proximity.

The intuition is that it is okay to learn information about a population of a large-enough size (at least n). One key term in the above definition is "natural superset". Assume that we want to achieve (1000, 0.1)-proximity for the above example. The first equivalence class E_1 is defined by (zip code = "986**", 20 ≤ Age ≤ 29) and contains 600 tuples. One equivalence class that naturally contains it would be the one defined by (zip code = "986**", 20 ≤ Age ≤ 39). Another such equivalence class would be the one defined by (zip code = "98***", 20 ≤ Age ≤ 29). If both of the two large equivalence classes contain at least 1000 records, and E_1's distribution is close to (i.e., the distance is at most 0.1) either of the two large equivalence classes, then E_1 satisfies (1000, 0.1)-proximity.

In the above definition of the (n, t)-proximity principle, the parameter n defines the breadth of the observer's background knowledge. A smaller n means that the observer knows the sensitive information about a smaller group of records. The parameter t bounds the amount of sensitive information that the observer can get from the released table. A smaller t implies a stronger privacy requirement. In fact, Table 6 satisfies (1000, 0.1)-proximity. The second equivalence class satisfies (1000, 0.1)-proximity because it contains 2000 > 1000 individuals, and thus, meets the privacy requirement (by setting the large group to be itself). The first and the third equivalence classes also satisfy (1000, 0.1)-proximity because both have the same distribution (the distribution is (0.5, 0.5)) as the large group which is the union of these two equivalence classes and the large group contains 1000 individuals.

Choosing the parameters n and t would affect the level of privacy and utility. The larger n is and the smaller t is, one achieves more privacy and less utility. By using specific parameters for n and t, we are able to show the relationships between (n, t)-proximity with existing privacy models such as k-anonymity and t-closeness.

Finally, there is another natural definition of (n, t)-proximity, which requires the distribution of the sensitive attribute in each equivalence class to be close to that of all its supersets of sizes at least n. We point out that this requirement may be too strong to achieve and may not be necessary. Consider an equivalence class (50 ≤ Age ≤ 60, Sex = "Male") and two of its supersets (50 ≤ Age ≤ 60) and (Sex = "Male"), where the sensitive attribute is "Disease." Suppose that the Age attribute is closely correlated with the Disease attribute but Sex is not. The two supersets may have very different distributions with respect to the sensitive attribute: the superset (Sex = "Male") has a distribution close to the overall distribution but the superset (50 ≤ Age ≤ 60) has a very different distribution. In this case, requiring the distribution of the equivalence class to be close to both supersets may not be achievable. Moreover, since the Age attribute is highly correlated with the Disease attribute, requiring the distribution of the equivalence class (50 ≤ Age ≤ 60, Sex = "Male") to be close to that of the superset (Sex = "Male") would hide the correlations between Age and Disease.

5 ANONYMIZATION ALGORITHMS AND DISTANCE MEASURES

One challenge is designing algorithms for anonymizing the data to achieve (n, t)-proximity. In this section, we describe how to adapt the Mondrian [9] multidimensional algorithm for our (n, t)-proximity model. The algorithm consists of three components: 1) Choosing a dimension on which to partition, 2) Choosing a value to split and 3) Checking if the partitioning violates the privacy requirement.

5.1 COMPONENTS 1 AND 2; CHOOSING A DIMENSION AND A VALUE TO SPLIT (TOP-DOWN GREEDY ALGORITHM FOR STRICT MULTIDIMENSIONAL PARTITIONING)

```
 Anonymize(partition)
if (no allowable multidimensional cut for partition)
return ϕ : partition → summary
else
dim ←choose dimension()
fs ←frequency_set(partition, dim)
splitVal ←find_median(fs)
lhs ←{t Є partition : t.dim ≤ splitV al}
rhs ←{t Є partition : t.dim > splitV al}
return Anonymize(rhs) U Anonymize(lhs)
```

5.2 Component 3; Checking if the partitioning violates the privacy requirement

```
Let P be a set of tuples
P is partitioned into r partitions { P_1,  P_2, ….., P_r }
for every p_i
if P_i(1 ≤ i ≤ r)
find = true
if P_i(1 ≤ i ≤ r)
```

```
find = false
for every Q Є Parent(P) and |Q| ≥ n
if D[P_i, Q] ≤ t
find = true
if find == false
return false
return true
```

5.3 DISTANCE MEASURES

The Earth Mover's Distance(EMD) is used to measure the distance between two probabilistic distributions, which 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, p_3, ... p_m), Q = (q_1, q_2, q_3, ... q_m), and d_{ij} be the ground distance between element i of P and element j of Q. We want to find a flow F = [f_{ij}], where f_{ij} is the flow of mass from element i of P to element j of Q that minimizes the overall work.

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

Subject to the following constraints,

$f_{ij} \geq 0$, $1 \leq i \leq m$, $1 \leq j \leq m$, (e1)

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i,\ 1 \leq i \leq m, \qquad \text{(e2)}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1. \qquad \text{(e3)}$$

These three constraints guarantee that P is transformed to Q by the mass flow F. Once the transformation problem is solved, the EMD is defined to be the total work.

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

More generally, the EMD is the work divided by the total flow. However, since we are calculating distance between two probability distributions, the total flow is always 1, as shown in e3. Now we derive formulas for calculating Earth Mover's Distance for the special cases that we need to consider. The EMD for numerical attributes: let $r_i = p_i - q_i$, (i = 1, 2, . . . , m), then the distance between P and Q can be calculated as

$$D[P, Q] = \frac{1}{m-1}\left(|r_1| + |r_1 + r_2| + ... + |r_1 + r_2 + . . . r_{m-1}|\right)$$

$$= \frac{1}{m-1}\sum_{i=1}^{i=m}\left|\sum_{j=1}^{j=i} r_j\right|$$

and the EMD for categorical attributes can be calculated as

$$D[P, Q] = \frac{1}{2}\sum_{i=1}^{i=m}|p_i - q_i| = \sum_{p_i \geq q_i}(p_i - q_i) = -\sum_{p_i \langle q_i}(p_i - q_i)$$

5.4 PROXIMITY WITH EMD – ANALYSIS

Going back to the example 3,

Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k},

P_1 = {3k, 4k, 5k} and P_2 = {6k, 8k, 11k }

Now we calculate D[P_1, Q] and D[P_2, Q] using EMD.

Let v_1 = 3k, v_2 = 4k, v_3 = 5k, v_4 = 6k, v_5 = 7k, v_6 = 8k, v_7 = 9k, v_8 = 10k, v_9 = 11k, the distance between v_i and v_j to be |i - j|/8; that is one optimal mass flow that transforms P_1 to Q is to move 1/9 probability mass across the following pairs: (5k → 11k), (5k → 10k), (5k → 9k), (4k → 8k), (4k → 7k), (4k → 6k), (3k → 5k), (3k → 4k) the cost of this is 1/9 x (6 + 5 + 4 + 4 + 3 + 2 + 2 + 1)/8 = 27/72 = 3/8 = 0.375.

Thus the maximal distance is 1. We have D[P_1, Q] = 0.375 and D[P_2, Q] = 0.167.

TABLE 7: INPATIENT MICRODATA WITH T-CLOSSNESS AND PROXIMITY WITH RESPECT TO SALARY AND DISEASE

	Non-Sensitive		Sensitive	
	Zip Code	Age	Salary	Disease
1	9867*	≤ 40	3K	Gastric Ulcer
2	9867*	≤ 40	5K	Stomach Cancer
3	9867*	≤ 40	9K	Pneumonia
4	9890*	≥ 40	6K	Gastritis
5	9890*	≥ 40	11K	Flu
6	9890*	≥ 40	8K	Bronchitis
7	9860*	≤ 40	4K	Gastritis
8	9860*	≤ 40	7K	Bronchitis
9	9860*	≤ 40	10K	Stomach Cancer

The Table 7 shows the anonymized version of Table 3. It has 0.167 closeness with respect to Salary and 0.278 closeness with respect to the Disease. The similarity attack is prevented in the above version of table. So here we note that t-closeness and (n, t)-proximity protect against attribute disclosure. But it do not deal with the identity disclosure. Thus, it may be desirable to use both (n, t)-proximity and k-anonymity. Further it should be noted that (n, t)-proximity deals with the homogeneity and background knowledge attacks.

6 CONCLUSIONS

While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The notion of l-diversity attempts to solve this problem. We have shown that l-diversity has a number of limitations and especially presented two attacks on l-diversity. Motivated by these limitations, we have proposed a new privacy notion called “proximity.” We propose two instantiations: a base model called t-closeness and a more flexible privacy model called (n, t)-proximity. We explain the rationale of the (n, t)-proximity model and show that it achieves a better balance between privacy and utility.

7 REFERENCES

1. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-Diversity: Privacy Beyond k-Anonymity,” Proc. Int’l Conf. Data Eng. (ICDE), p. 24, 2006.
2. D. Lambert, “Measure of Disclosure Risk and Harm,” J. Official Statistical, vol. 9, pp. 313-331, 1993.
3. G.T. Duncan and D.Lambert, “Disclosure-Limited Data Dissemination,” J. Am. Statistical Assoc., vol. 81, pp. 10-28, 1986.
4. K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Mondrian Multidimensional k-Anonymity,” ICDE, p. 25,2006
5. L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” Int’l J.Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.
6. P. Samarati, “Protecting Respondent’s Privacy in Microdata Release,” IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.
7. T.M. Truta and B. Vinay, “Privacy Protection: P-Sensitive k-Anonymity Property,” Proc. Int’l Workshop Privacy Data Management (ICDE Workshops), 2006.
8. X. Xiao and Y. Tao, “Personalized Privacy Preservation,” Proc. ACM SIGMOD, pp. 229-240, 2006.
9. Y. Rubner, C. Tomasi, and L.J. Guibas, “The Earth Mover’s Distance as a Metric for Image Retrieval,” Int’l J. Computer Vision, vol. 40, no. 2, pp. 99-121, 2000.

REQUEST FOR FEEDBACK

Dear Readers

At the very outset, International Journal of Research in Commerce, IT and Management (IJRCM) acknowledges & appreciates your efforts in showing interest in our present issue under your kind perusal.

I would like to request you to supply your critical comments and suggestions about the material published in this issue as well as on the journal as a whole, on our E-mail i.e. **infoijrcm@gmail.com** for further improvements in the interest of research.

If you have any queries please feel free to contact us on our E-mail **infoijrcm@gmail.com**.

I am sure that your feedback and deliberations would make future issues better – a result of our joint effort.

Looking forward an appropriate consideration.

With sincere regards

Thanking you profoundly

Academically yours

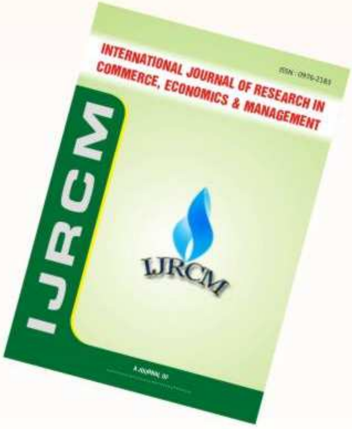
Sd/-

Co-ordinator

ABOUT THE JOURNAL

In this age of Commerce, Economics, Computer, I.T. & Management and cut throat competition, a group of intellectuals felt the need to have some platform, where young and budding managers and academicians could express their views and discuss the problems among their peers. This journal was conceived with this noble intention in view. This journal has been introduced to give an opportunity for expressing refined and innovative ideas in this field. It is our humble endeavour to provide a springboard to the upcoming specialists and give a chance to know about the latest in the sphere of research and knowledge. We have taken a small step and we hope that with the active co-operation of like-minded scholars, we shall be able to serve the society with our humble efforts.

Our Other Journals


INTERNATIONAL JOURNAL OF RESEARCH IN COMMERCE, ECONOMICS & MANAGEMENT
IJRCM
IJRCM

INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION & MANAGEMENT
IJRCM
IJRCM