iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** XII **Month of publication:** December 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Distributed Privacy preserving and Handling Privacy information leakage by using k-anonymity algorithm

Padmapriya.G[1], Dr.M.Hemalatha[2]
[1]Research Scholar, [1,2]Dept. of. Computer science, Karpagam University, Coimbatore, India

Abstract*—There is increasing pressure to share health information and even make it publicly available. However, such disclosures of personal health information raise serious privacy concerns. To alleviate such concerns, it is possible to anonymize the data before disclosure. One popular anonymization approach is k-anonymity. There have been no evaluations of the actual re-identification probability of k-anonymized data sets. Through a simulation, we evaluated the re-identification risk of k-anonymization and three different improvements on three large data sets. Re-identification probability is measured under two different re-identification scenarios. Information loss is measured by the commonly used discernability metric. For one of the re-identification scenarios, k-Anonymity consistently over-anonymous data sets, with this over-anonymization being most pronounced with small sampling fractions. Over-anonymization results in excessive distortions in the data (i.e., high information loss), making the data less useful for subsequent analysis. We found that a hypothesis testing approach provided the best control over re-identification risk and reduces the extent of information loss compared to baseline k-anonymity. Guidelines are provided on when to use the hypothesis testing approach instead of baseline k-anonymity.
Keywords*— k-anonymization,re-identification,high information loss.

I. INTRODUCTION

The sharing of raw research data is believed to have many benefits, including making it easier for the research community to confirm published results, ensuring the availability of original data for meta-analysis, facilitating additional innovative analysis on the same data sets, getting feedback to improve data quality for on-going data collection efforts, achieving cost savings from not having to collect the same data multiple times by different research groups, minimizing the need for research participants to provide data repeatedly, facilitating linkage of research data sets with administrative records, and making data available for instruction and education. Consequently, there are pressures to make such research data more generally available. For example, in January 2004 Canada was a signatory to the OECD Declaration on Access to Research Data from Public Funding. This is intended to ensure that data generated through public funds are publicly accessible for researchers as much as possible. To the extent that this is implemented, potentially more personal health data about Canadians will be made available to researchers world wide. The European Commission has passed a regulation facilitating the sharing with external researchers of data collected by Community government agencies. There is interest by the pharmaceutical industry and academia to share raw data from clinical trials.

Researchers in the future may have to disclose their data. The Canadian Medical Association Journal has recently contemplated requiring authors to make the full data set from their published studies available publicly online. Similar calls for depositing raw data with published manuscripts have been made recently. The Canadian Institutes of Health Research (CIHR) has a policy, effective on 1st January 2008, that requires making some data available with publications. The UK MRC policy on data sharing sets the expectation that data from their funded projects will be made publicly available. The UK Economic and Social Research Council requires its funded projects to deposit datasets in the UK Data Archive (such projects generate health and lifestyle data on, for example, diet, reproduction, pain, and mental health). The European Research Council considers it essential that the raw data be made available, preferably immediately after publication, but not later than six months after publication. The NIH in the US expects investigators seeking more than $500,000 per year in funding to include a data sharing plan (or explain why that is not possible). Courts, in criminal and civil cases, may compel disclosure of research data.

Such broad disclosures of health data pose significant privacy risks. The risks are real, as demonstrated by recent successful re-identifications of individuals in publicly disclosed data sets (see the examples in [triangle]). One approach for protecting the identity of individuals when releasing or sharing sensitive health data is to anonymize it. A popular approach for data anonymization is k-anonymity. With k-anonymity an original data set containing personal health information can be transformed so that it is difficult for an intruder to determine the identity of the individuals in that data set. A k-anonymized data

set has the property that each record is similar to at least another k-1 other records on the potentially identifying variables. For example, if k = 5 and the potentially identifying variables are age and gender, then a k-anonymized data set has at least 5 records for each value combination of age and gender. The most common implementations of k-anonymity use transformation techniques such as generalization, global recoding, and suppression.

Any record in a k-anonymized data set has a maximum probability An external file that holds a picture, illustration, etc.

Object name is 627.S1067502708001047.si2.jpg of being re-identified. In practice, a data custodian would select a value of k commensurate with the re-identification probability they are willing to tolerate—a threshold risk. Higher values of k imply a lower probability of re-identification, but also more distortion to the data, and hence greater information loss due to k-anonymization. In general, excessive anonymization can make the disclosed data less useful to the recipients because some analysis becomes impossible or the analysis produces biased and incorrect results.

Thus far there has been no empirical examination of how close the actual re-identification probability is to this maximum. Ideally, the actual re-identification probability of a k-anonymized data set would be close to An external file that holds a picture, illustration, etc.

Object name is 627.S1067502708001047.si3.jpg since that balances the data custodian's risk tolerance with the extent of distortion that is introduced due to k-anonymization. However, if the actual probability is much lower than An external file that holds a picture, illustration, etc.

Object name is 627.S1067502708001047.si4.jpg then k-anonymity may be over-protective, and hence results in unnecessarily excessive distortions to the data.

In this paper we make explicit the two re-identification scenarios that k-anonymity protects against, and show that the actual probability of re-identification with k-anonymity is much lower than An external file that holds a picture, illustration, etc.

Object name is 627.S1067502708001047.si5.jpg for one of these scenarios, resulting in excessive information loss. To address that problem, we evaluate three different modifications to k-anonymity and identify one that ensures that the actual risk is close to the threshold risk and that also reduces information loss considerably. The paper concludes with guidelines for deciding when to use the baseline versus the modified k-anonymity procedure. Following these guidelines will ensure that re-identification risk is controlled with minimal information loss when using k-anonymity..

II. THE TWO RE-IDENTIFICATION SCENARIOS FOR A K-ANONYMIZED DATA SET

The concern of k-anonymity is with the re-identification of a single individual in an anonymized data set. There are two re-identification scenarios for a single individual:

Re-identify a specific individual (known as the prosecutor re-identification scenario). The intruder (e.g., a prosecutor) would know that a particular individual (e.g., a defendant) exists in an anonymized database and wishes to find out which record belongs to that individual.

Re-identify an arbitrary individual (known as the journalist re-identification scenario). The intruder does not care which individual is being re-identified, but is only interested in being able to claim that it can be done. In this case the intruder wishes to re-identify a single individual to discredit the organization disclosing the data.pattern viewer is the second component that shows recurring patterns discovered. The third component is an extractor module that extracts favored information from similar Web pages based on the extraction rule selected by the user.

A. Re-identification Risk under the Prosecutor Scenario

The set of patients in the file to be disclosed is denoted by s.Before the file about s can be disclosed, it must be anonymized. Some of the records in the file will be suppressed during anonymization, therefore a different subset of patients, s', will be represented in the anonymized version of this file. Let the anonymized file be denoted by ζ.There is a one-to-one mapping between the records in ζ and the individuals in s'.Under the prosecutor scenario, a specific individual is being re-identified, say, a VIP. The intruder will match the VIP with the records in ζ on *quasi-identifiers*. Variables such as gender, date of birth, postal code, and race are commonly used quasi-identifiers. Records in ζ that have the same values on the quasi-identifiers are called an*equivalence class*. Let the number of records in ζ that have exactly the same quasi-identifier values as the VIP be f. The re-identification risk for the VIP is then $1/f$. For example, if the individual being re-identified is a 50 year old male, then f is the number of records on 50 year old males in ζ. The intruder has a probability $1/f$ of getting a correct match. Since the data custodian does not know, a priori, which equivalence class a VIP will match against, one can assume the worse case scenario. Under the worse case scenario, the intruder will have a VIP who matches with the smallest equivalence class in ζ, which in a k-

anonymized data set will have a size of at least k. Hence the re-identification probability will be at most $1/k$ .Therefore, under the prosecutor re-identification scenario k-anonymity can ensure that the re-identification risk is approximately equal to the threshold risk, as intended by the data custodian. This, however, is not the case under the journalist re-identification scenario.

B. Re-identification Risk under the Journalist Scenario

We assume that there exists a large finite population of patients denoted by the set U. We then have $s' \subseteq s \subseteq U$. An intruder would have access to an *identification database* about the population U, and uses this identification database to match against the patients in ζ. The identification database is denoted by Z, and the records in Z have a one-to-one mapping to the individuals in U.In the example of ► we have a data set about 14 individuals that needs to be disclosed. This data set is 2-anonymized to produce the anonymized data set, ζ. After 2-anonymization, there are only 11 records left in ζ since three had to be suppressed.

Original Database to Disclose

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS		
	ID	Name	Gender	Year of Birth	Test Result
1	1	John Smith	Male	1959	+ve
.	2	Alan Smith	Male	1962	-ve
.	3	Alice Brown	Female	1955	-ve
.	4	Hercules Green	Male	1959	-ve
.	5	Alicia Freds	Female	1942	-ve
.	6	Gill Stringer	Female	1975	-ve
.	7	Marie Kirkpatrick	Female	1966	+ve
.	8	Leslie Hall	Female	1987	-ve
.	9	Bill Nash	Male	1975	-ve
.	10	Albert Blackwell	Male	1978	-ve
.	11	Beverly McCulsky	Female	1964	-ve
.	12	Douglas Henry	Male	1959	+ve
.	13	Freda Shields	Female	1975	-ve
|s|	14	Fred Thompson	Male	1967	-ve

2-Anonymization

	QUASI-IDENTIFIERS		
	ID	Gender	Decade of Birth	Test Result
1	1	Male	1950-1959	+ve
.	2	Male	1960-1969	-ve
.	4	Male	1950-1959	-ve
.	6	Female	1970-1979	-ve
.	7	Female	1960-1969	+ve
.	9	Male	1970-1979	-ve
.	10	Male	1970-1979	-ve
.	11	Female	1960-1969	-ve
.	12	Male	1950-1959	+ve
.	13	Female	1970-1979	-ve
|s'|	14	Male	1960-1969	-ve

Disclosed (k-Anonymized) Database (ζ)

Matching

Identification Database (Z)

	IDENTIFYING VARIABLE	QUASI-IDENTIFIERS	
	ID	Name	Gender	Year of Birth
1	1	John Smith	Male	1959
.	2	Alan Smith	Male	1962
.	3	Alice Brown	Female	1955
.	4	Hercules Green	Male	1959
.	5	Alicia Freds	Female	1942
.	6	Gill Stringer	Female	1975
.	7	Marie Kirkpatrick	Female	1966
.	8	Leslie Hall	Female	1987
.	9	Bill Nash	Male	1975
.	10	Albert Blackwell	Male	1978
.	11	Beverly McCulsky	Female	1964
.	12	Douglas Henry	Male	1959
.	13	Freda Shields	Female	1975
.	14	Fred Thompson	Male	1967
.	15	Joe Doe	Male	1961
.	16	Mark Fractus	Male	1974
.	17	Lillian Barley	Female	1978
.	18	Jane Doe	Female	1961
.	19	Nina Brown	Female	1968
.	20	William Cooper	Male	1973
.	21	Kathy Last	Female	1966
.	22	Deitmar Plank	Male	1967
.	23	Anderson Hoyt	Male	1971
.	24	Alexandra Knight	Female	1974
.	25	Helene Arnold	Female	1977
.	26	Anderson Heft	Male	1968
.	27	Almond Zipf	Male	1954
.	28	Alex Long	Female	1952
.	29	Britney Goldman	Female	1956
.	30	Lisa Marie	Female	1988
|U|	31	Natasha Markhov	Female	1941

Table 1:A hypothetical example of the three databases assumed in the k-anonymity privacy model under the journalist re-identification scenario. The intruder performs the matching while the data custodian performs the 2-anonymization.

An intruder gets hold of an identification database with 31 records. This is the Z database. The intruder then attempts re-identification by matching an arbitrary record against the records in ζ on year of birth and gender. In our example, once an arbitrary individual is re-identified, the intruder will have that individual's test result. he discrete variable formed by cross-classifying all values on the quasi-identifiers in ζ can take on J distinct values. Let $X_{\zeta,i}$ denote the value of a record i in the ζ data set. For example, if we have two quasi-identifiers, such as gender and age, then we may have $X_{\zeta,1} = "MALE, 50"$, $X_{\zeta,1} = "MALE, 53"$, and so on. Similarly let $X_{Z,i}$ denote the value of record i in the Z data set. The sizes of the different equivalence classes are given by $f_j = \sum_{i \in s'} I(X_{\zeta,i} = j), \quad j = 1, \ldots, J,$ where f_j is the size of a ζ equivalence class and $I(\cdot)$ is the indicator function. Similarly we have $F_j = \sum_{i \in U} I(X_{Z,i} = j), \quad j = 1, \ldots, J,$ where F_j is the size of an equivalence class in Z .nder the journalist re-identification scenario, the probability of re-identification of a record in an equivalence class j is $1/F_j$.However, a smart intruder would focus on the records in equivalence classes with the highest probability of re-

identification. Equivalence classes with the smallest value for F_j have the highest probability of being re-identified, and therefore we assume that a smart intruder will focus on these. The probability of re-identification of an arbitrary individual by a smart intruder is then given by:

$$\theta_{\max} = 1 \Big/ \min_j (F_j) \qquad (1)$$

If we consider again, the 2-anonymized file had the age converted into 10 year intervals. In that example we can see that $\theta_{\max} = 0.25$ because the smallest equivalence class in Z has 4 records (ID numbers 1, 4, 12, and 27). With 2-anonymization the data custodian was using a threshold risk of 0.5, but the actual risk of re-identification, $\theta_{\max}$, was half of that. This conservatism may seem like a good idea, but in fact it has a large negative impact on data quality. In our example, 2-anonymization resulted in converting age into ten year intervals and the suppression of more than one fifth of the records that had to be disclosed (3 of 14 records had to be suppressed). By most standards, losing one fifth of a data set due to anonymization would be considered extensive information lossNow consider another approach: k-map. With k-map it is assumed that the data custodian can k-anonymize the identification database itself (and hence directly control the F_j values). Let's say that the Z identification database is k-anonymized to produce Z'. The k-map property states that each record in ζ is similar to at least k records in Z'. This is illustrated in . Here, the data custodian 2-anonymizes the identification database directly, and then implements the transformations to the data set to be disclosed. In this example $\theta_{\max} = 0.5$ because the smallest equivalence classes in Z' for records 1 to 14 have two records. Also, the extent of information loss is reduced significantly: there are no records suppressed in the ensured that the actual re-identification risk is what the data custodian intended and we have simultaneously reduced information disclosed data set and the age is converted into 5 year intervals rather than 10 year intervals. By using the k-map property we have loss.

Identification Database (Z)

ID	IDENTIFYING VARIABLE Name	QUASI-IDENTIFIERS Gender	Year of Birth
1	John Smith	Male	1959
2	Alan Smith	Male	1962
3	Alice Brown	Female	1955
4	Hercules Green	Male	1959
5	Alicia Freds	Female	1942
6	Gill Stringer	Female	1975
7	Marie Kirkpatrick	Female	1966
8	Leslie Hall	Female	1987
9	Bill Nash	Male	1975
10	Albert Blackwell	Male	1978
11	Beverly McCulsky	Female	1964
12	Douglas Henry	Male	1959
13	Freda Shields	Female	1975
14	Fred Thompson	Male	1967
15	Joe Doe	Male	1961
16	Mark Fractus	Male	1974
17	Lillian Barley	Female	1978
18	Jane Doe	Female	1961
19	Nina Brown	Female	1968
20	William Cooper	Male	1973
21	Kathy Last	Female	1966
22	Deitmar Plank	Male	1967
23	Anderson Hoyt	Male	1971
24	Alexandra Knight	Female	1974
25	Helene Arnold	Female	1977
26	Anderson Heft	Male	1968
27	Almond Zipf	Male	1954
28	Alex Long	Female	1952
29	Britney Goldman	Female	1956
30	Lisa Marie	Female	1988
31	Natasha Markhov	Female	1941

Anonymization

Anonymized Identification Database (Z')

ID	IDENTIFYING VARIABLE Name	QUASI-IDENTIFIERS Gender	Year of Birth
~~27~~	~~Almond Zipf~~	~~Male~~	~~1954~~
1	John Smith	Male	1955-1959
4	Hercules Green	Male	1955-1959
12	Douglas Henry	Male	1955-1959
2	Alan Smith	Male	1960-1964
15	Joe Doe	Male	1960-1964
14	Fred Thompson	Male	1965-1969
22	Deitmar Plank	Male	1965-1969
26	Anderson Heft	Male	1965-1969
16	Mark Fractus	Male	1970-1974
20	William Cooper	Male	1970-1974
23	Anderson Hoyt	Male	1970-1974
9	Bill Nash	Male	1975-1979
10	Albert Blackwell	Male	1975-1979
5	Alicia Freds	Female	1940-1944
31	Natasha Markhov	Female	1940-1944
~~28~~	~~Alex Long~~	~~Female~~	~~1952~~
3	Alice Brown	Female	1955-1959
29	Britney Goldman	Female	1955-1959
11	Beverly McCulsky	Female	1960-1964
18	Jane Doe	Female	1960-1964
7	Marie Kirkpatrick	Female	1965-1969
19	Nina Brown	Female	1965-1969
21	Kathy Last	Female	1965-1969
~~24~~	~~Alexandra Knight~~	~~Female~~	~~1974~~
6	Gill Stringer	Female	1975-1979
13	Freda Shields	Female	1975-1979
17	Lillian Barley	Female	1975-1979
25	Helene Arnold	Female	1975-1979
8	Leslie Hall	Female	1985-1989
30	Lisa Marie	Female	1985-1989

2-Map

Disclosed Database (ζ)

ID	QUASI-IDENTIFIERS Gender	Year of Birth	Test Result
1	Male	1955-1959	+ve
2	Male	1960-1964	-ve
3	Female	1955-1959	-ve
4	Male	1955-1959	-ve
5	Female	1940-1944	-ve
6	Female	1975-1979	-ve
7	Female	1965-1969	+ve
8	Female	1985-1989	-ve
9	Male	1975-1979	-ve
10	Male	1975-1979	-ve
11	Female	1960-1964	-ve
12	Male	1955-1959	+ve
13	Female	1975-1979	-ve
14	Male	1965-1969	-ve

Table .2 An illustration of the k-map approach whereby the data custodian does have access to the identification database. The crossed out values are suppressed records.

In practice, the k-map model is not used because it is assumed that the data custodian does not have access to an identification database, but that an intruder does.Therefore, the k-anonymity model is used instead.

There are good reasons why the data custodian would not have an identification database. Often, a population database is expensive to get hold of. Plus, it is likely that the data custodian will have to protect multiple populations, hence multiplying the expense. For example, the construction of a single profession-specific database using semi-public registries that can be used for re-identification attacks in Canada costs between $150,000 to $188,000. Commercial databases can be comparatively costly. Furthermore, an intruder may commit illegal acts to get access to population registries. For example, privacy legislation and the Elections Act in Canada restrict the use of voter lists to running and supporting election activities. There is at least one known

case where a charity allegedly supporting a terrorist group has been able to obtain Canadian voter lists for fund raising. A legitimate data custodian would not engage in such acts.

However, a number of methods have been developed in the statistical disclosure control literature to estimate the size of the equivalence classes in Z from a sample. If these estimates are accurate, then they can be used to approximate k-map. Approximating k-map will ensure that the actual risk is close to the threshold risk, and consequently that there will be less information loss. Three such methods are considered below.

III. PROPOSED IMPROVEMENTS TO K-ANONYMITY UNDER THE JOURNALIST RE-IDENTIFICATION SCENARIO

We consider three alternative approaches to reduce the extent of over-anonymization under the journalist re-identification scenario. These three approaches extend k-anonymity to approximate k-map.

A. Baseline (D1)

Baseline k-Anonymization algorithms apply transformations, such as generalization, global recoding, and suppression until all equivalence classes in ζ are of size k or more.

B. Individual Risk Estimation (D2)

The actual re-identification risk for each equivalence class in ζ, $1/\hat{F}_j$, can be directly estimated.
Subsequently, the k-anonymization algorithm should ensure that all equivalence classes meet the following condition $1/\hat{F}_j \leq 1/k$. One estimator for $1/\hat{F}_j$ has been studied extensively and was also incorporated in the mu-argus tool (which was developed by the Netherlands national statistical agency and used by many other national statistical agencies for disclosure control purposes), but it has never been evaluated in the context of k-anonymity. To the extent that this individual risk estimator is accurate, it can ensure that the actual risk is as close as possible to the threshold risk.

C. Hypothesis Testing Using a Poisson Distribution (D3)

One can use a hypothesis testing approach for determining if $F_j > k$.[56,71] If we assume that the size of the sample equivalence classes f_j follow a Poisson distribution, we can construct the null Poisson distribution for $H_0 : F_j < k$ and determine which observed value of f_j will reject the null hypothesis at $\alpha = 0.1$. Let's denote this value as k'. Then the k-anonymity algorithm should ensure that the following condition $f_j \geq \min(k,k')$ is met for all equivalence classes.

D. Hypothesis Testing Using a Truncated-at-zero Poisson Distribution (D4)

In practice, we ignore equivalence classes that do not appear in the sample, therefore, the value of f_j cannot be equal to zero. An improvement in the hypothesis testing approach above would then be to use a truncated-at-zero Poisson distribution to determine the value of k'. The k-anonymity algorithm should ensure that the condition $f_j \geq \min(k,k')$ is met for all equivalence classes.

IV. EVALUATION

For each k-anonymized data set the actual risk is measured as $\theta_{\max}$ and the information loss is measured in terms of the discernability metric. Averages were calculated for each sampling fraction across the 1000 samples.

The results are presented in the form of three sets of graphs:

Risk. This shows the value of $\theta_{\max}$ against sampling fraction for each of the four approaches.

Information Loss. Because the discernability metric is affected by the sample size (and hence makes it difficult to compare across differing sampling fractions), we normalize it for D2, D3, and D4 by the baseline value. For example, a value of 0.8 (or 80%) for D2 means that the information loss for D2 is 80% of that for the baseline k-anonymity approach. The graph shows the normalized discernability metric for these three approaches against the sampling fraction. The value for D1 will by definition always be 1 (or 100%).

Suppression. Because the extent of suppression is an important indicator of data quality by itself, we show graphs of the percentage of suppressed records against sampling fraction for the four approaches.

V. CONCLUSION

There is increasing pressure to disclose health research data, and this is especially true when the data has been collected using public funds. However, the disclosure of such data raises serious privacy concerns. For example, consider an individual who participated in a clinical trial having all of their clinical and lab data published in a journal web site accompanying the article on the trial. If it was possible to re-identify the records of that individual from this public data it would be a breach of privacy. Such an incident could result in fewer people participating in research studies because of privacy concerns, and if it happened in Canada, would be breaking privacy laws.It is therefore important to understand precisely the types of re-identification attacks that can be launched on a data set and the different ways to properly anonymize the data before it is disclosed.Anonymization techniques result in distortions to the data. Excessive anonymization may reduce the quality of the data making it unsuitable for some analysis, and possibly result in incorrect or biased results. Therefore, it is important to balance the amount of anonymization being performed against the amount of information loss.In this paper we focused on k-anonymity, which is a popular approach for protecting privacy. We considered the two re-identification scenarios that k-anonymity is intended to protect against. For one of the scenarios, we showed that actual re-identification risk under the baseline k-anonymity is much lower than the threshold risk that the data custodian assumes, and that this results in an excessive amount of information loss, especially at small sampling fractions. We then evaluated three alternative approaches and found that one of them consistently ensures that the re-identification risk is quite close to the actual risk, and always has lower information loss than the baseline approach.It is recommended that data custodians determine which re-identification scenarios apply on a case-by-case basis, and anonymize the data before disclosure using the baseline k-anonymity model or our modified k-anonymity model accordingly.

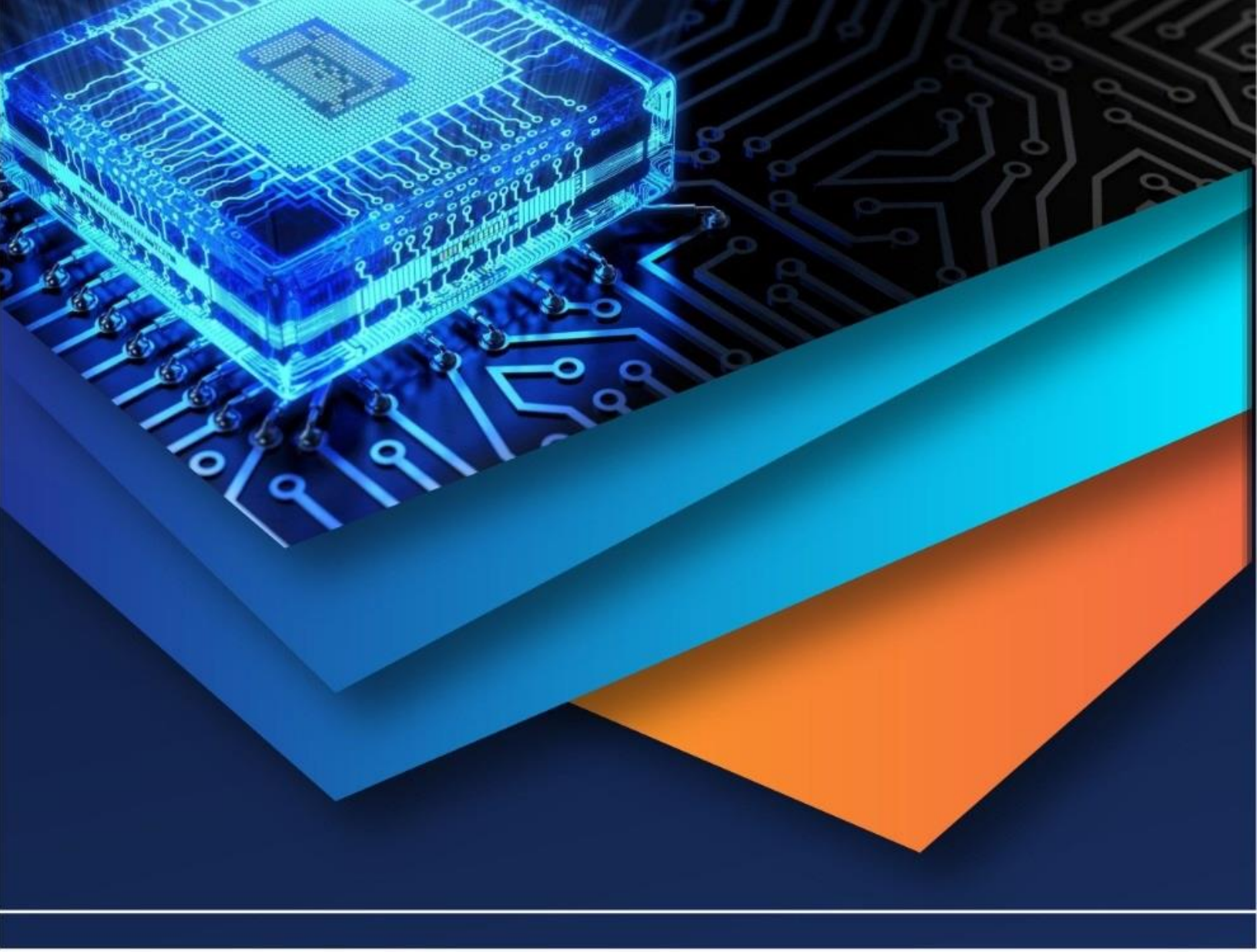
REFERENCES

[1]. Polettini S. A note on the individual risk of disclosureIstituto nazionale di statistica (Italy); 2003.
[2]. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. L-Diversity: Privacy Beyond k-AnonymityInternational Conference on Data Engineering; 2006.
[3]. Little R, Rubin D. Statistical Analysis With Missing DataJohn Wiley & Sons; 1987.
[4]. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables Am J Epidemiol 1991;134(8):895-907. [PubMed]
[5]. Kim J, Curry J. The treatment of missing data in multivariate analysis Soc Methods Res 1977;6:215-240.
[6]. Willenborg L, de Waal T. Statistical Disclosure Control in PracticeSpringer-Verlag; 1996.
[7]. Bethlehem J, Keller W, Pannekoek J. Disclosure control of microdata J Am Stat Assoc 1990;85(409):38-45.
[8]. Skinner C, Holmes D. Estimating the re-identification risk per record in microdata J Off Stat 1998;14(4):361-372.
[9]. Chen G, Keller-McNulty S. Estimation of identification disclosure risk in microdata J Off Stat 1998;14(1):79-95.
[10]. Zayatz L. Estimation of the percent of unique population elements on a microdata file using the sampleUS Bureau of the Census; 1991.
[11]. Greenberg B, Voshell L. The geographic component of disclosure risk for microdataBureau of the Census; 1990.
[12]. Willenborg L, Mokken R, Pannekoek J, Microdata and disclosure risks. Proceedings of the Annual Research Conference of US Bureau of the Census, 1990;167–180.
[13]. Skinner G, Elliot M. A measure of disclosure risk for microdata J Royal Stat Soc (Ser B) 2002;64(Part 4):855-867.
[14]. Aggarwal C. On k-anonymity and the curse of dimensionality Proceedings of the 31st VLDB Conference 2005.
[15]. Du Y, Xia T, Tao Y, Zhang D, Zhu F. On Multidimensional k-Anonymity with Local Recoding Generalization. IEEE 23rd International Conference on Data Engineering, 2007;1422–4.
[16]. Xu J, Wang W, Pei J, Wang X, Shi B, Fu A. Utility-based anonymization using local recodingACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2006.
[17]. Wong R, Li J, Fu A, Wang K. (alpa,k)-Anonymity: An enhanced k-anonymity model for privacy-preserving data publishingACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2006.
[18]. Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A. Approximation algorithms for k-anonymity J Priv Technol 2005.

AUTHOR PROFILE

Padmapriya.G received the first degree in B.C.A from Bharathiyar University in 2004, Tamilnadu, India. She obtained her master degree in Computer Communication from Bharathiyar university and She obtained her master degree in Computer Applications from Bharathyiar University in 2012, Tamilnadu, India. She is currently pursuing her Ph.D. degree Under the guidance of Dr. M.Hemalatha, Head, Dept of Software Systems, Karpagam University, Tamilnadu, India.

Dr. M. Hemalatha completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women's University, Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science in Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several National and International Journals.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY