An Extensive Study on Data Anonymization Algorithms Based on K-Anonymity

Ms. Simi M S[1], Mrs. SankaraNayaki K[2], Dr.M.Sudheep Elayidom[3]
1. PG Student Dept of Computer Science, Adishankara Institute of Engineering and Technology, Kalady, Kerala, India-683574 .
2Asst: Prof, Dept of Information Technology, Adishankara Institute of Engineering and Technology, Kalady,Kerala, India-683574
3. Professor, Dept of Computer Science School of Engineering, CUSAT, Kochi-22, Kerala, India

Abstract: For business and research oriented works engaging Data Analysis and Cloud services needing qualitative data, many organizations release huge microdata. It excludes an individual’s explicit identity marks like name, address and comprises of specific information like DOB, Pin-code, sex, marital status, which can be combined with other public data to recognize a person. This implication attack can be manipulated to acquire any sensitive information from social network platform, thereby putting the privacy of a person in grave danger. To prevent such attacks by modifying microdata, K-anonymization is used. With potentially increasing data, the effective method to anonymize it stands challenging. After series of trails and systematic comparison, in this paper, we propose three best algorithms along with its efficiency and effectiveness. Studies help researchers to identify the relationship between the values of k, degree of anonymization, choosing a quasi-identifier and focus on execution time.

Keywords: Generalization, Incognito Algorithm, K-anonymity, Microdata, Quasi-Identifier, Samarati's Algorithm, Suppression, Sweeney’s Algorithm.

1. INTRODUCTION

Microdata is being published by numerous organizations for many different purposes such as business, demographic research, public health research and so on. This published data can risk the privacy of an individual [1]. To protect the anonymity of the entities, the data holders encrypt or remove the explicit identifiers such as name, phone numbers, aadhar number and addresses. However, other attributes like sex, date of birth, zip code, race etc, when joined together with publicly released information, can be utilized to identify the anonymous individuals. The large amount of information that is easily accessible today, together with the increased computational power available to the attackers, make such attacks a serious problem [2].

Data about us is gathered on an everyday premise, as we join organizations, gatherings or are admitted in a hospital, search for basic supplies, or execute our normal day-to-day exercises. The measure of exclusive records portraying every native's finance, interests, and demographics is expanding each day. Numerous districts offer populace registers that incorporate the identities of people alongside fundamental demographics. This information, which is often publicly disseminated on the other hand, sold, can be utilized for connecting personalities with re-identified data [3]. This kind of circumstance has brought particular concerns in the medical [4] and financial fields, where microdata, which is progressively discharged for circulationor research, can be or have been liable to be manhandled, prone to abuses, thereby, threatening the privacy of individuals [2].

The problem we investigate in this paper is how to protect microdata from implication attack, and compare three Anonymization algorithms in terms of its relationship between the value of 'k' and the execution time.Removing the unique identifiers, for example, Name, Id from a table can't ensure privacy. Re-identification is conceivable by utilizing a set of attributes and another database containinga similar set of attributes [5]. At times, this approach can additionally release delicate data about a person. An illustration portraying the attack is demonstrated as follows:

Table I- Implication attack

DOB	**Sex**	**Pin code**	**Disease**
27/03/1980	Female	684578	Hepatitis
12/04/1988	Female	684674	Cancer
03/08/1991	Male	689643	Heart Disease
18/05/1985	Female	684987	Interstitial cystitis

Name	**DOB**	**Sex**	**Pin code**
Aida	27/03/1980	Female	684578
Camilla	12/04/1988	Female	684674
Smith	03/08/1991	Male	689643
Carissa	18/05/1985	Female	684987

Hospital Patient Data (a)Vote Enrollment Data (b)

An attacker can simply join the data from hospital patient data and vote enrollment information. In Table I, by coordinating the attributes like DOB, Sex and Pin-code the attacker canderive that Camilla is suffering from Cancer, which is an extremely sensitive data related to a person. As it is quite apparent that concealing the name, phonenumber or other explicit identifiers does not ensure the security of sensitive data of an individual, in this way, we require more effective procedures to accomplish our objective. Thus by using an appropriate algorithm for anonymizing data one can release maximum amount of data and can ensure that privacy of no individual is being put in danger due to the released data.

2. Background and Related Work

Recently, manhas proposed several algorithms for anonymizing Microdata [20]. The principle objective of thesealgorithms is to satisfy specific privacy requirements, while guaranteeing that theanonymized information stays valuable for analysis.

2.1 Privacy Models

Normally there are three types of attributes structured in the data(i) key-attributes (explicit attributes) (ii) quasi-identifiers and (iii) sensitive attributes. Key attributes remarkably identify people, example of these properties are names and aadhar numbers. Data like pin code, sex, marital status, date of birth, does not uniquely identify a record owner, but their combination, called the quasi-identifier (QID) [6], does. Sensitive attributes are those that people are not willing to be connected with; examples are disability status, psychiatric problems and so on.

The privacy requirements are done with certain privacy models and its implementation is performed by data transformation to preserve information utility.The larger part of privacy models focus on blocking three critical threats to individual privacy: (a) identity revelation, which happens when a person is connected to their record in the publicly available data, (b) feature disclosure (or vulnerable data revelation) which happens when a patient is connected with sensitive data (e.g., HIV-positive), and (c) table linkage, in this the attacker looks at the released table to decides the presence or absence of the victim's record[7],[9].In all three types of threats, we assume that theadversary knows the QID of the victim. Besides, data transformation is applied by means of two techniques: Generalization and Suppression [8]. In Generalization the values are replaced with more general values, whereas in Suppression, it omits certain data values. Multiple privacy models have been proposed to offer individual privacy concerns. All these are developed by considering the various attacking scenarios of the data.Table II summarizes the privacy models against various threats, among these we focus on *k*-Anonymity because it is more vulnerable to attacks. [27]

(i) k-Anonymity- Model against identity Revelation

Table II-privacy models

Privacy Model	**Identity revelation**	**Feature disclosure**	**Table linkage**
k-Anonymity[9]	✓		
MultiR k-Anonymity[10]	✓		
***l*-Diversity[11]**	✓		
(*k, e*)-Anonymity [12]		✓	
(*α, k*)-Anonymity [13]	✓	✓	
***t*-Closeness[14]**		✓	
(*X, Y*)-Privacy [15]	✓	✓	
ϵ-Differential Privacy[16]			✓
(*d, γ*)-Privacy[17]			✓
Distributional Privacy[18]			✓

k-anonymity is the well established model for identity revelation. It was the primary model proposed for microdata anonymization and it is the base from which further expansion have been developed. The definition of *k*-anonymity is as follows [19]: *Let RT(A1,...,An) be a table and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT[QI_{RT}] appears with at least k occurrences in RT[QI_{RT}].*

Table III-anonymous Hospital patient data (k=3)

DOB	**Sex**	**Pin-code**	**Disease**
1980-89	Female	684***	Hepatitis
1980-89	Female	684***	Cancer
1990-99	Male	689***	Heart Disease

To avert record linkage through QID, Samarati and Sweeney propose the thought of k- anonymity [8]: If one record in the table has some value QID, at least k -1 different records likewise have the value QID. In other words, the base equivalence group size on QID is in any event, k. A table fulfilling this necessity is called k-anonymous. In a k-anonymous table, every record is indistinguishable from at least k −1 different records with respect to QID. Thus the likelihood of connecting an individual to a particular record through QID is at most 1/k [20].

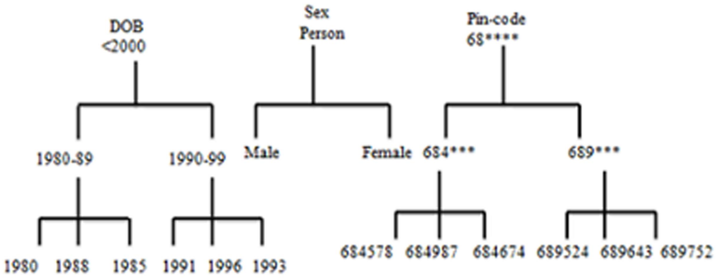

Fig. 1 Taxonomy trees

Table III shows a 3-anonymous table by generalizing QID = DOB, Sex, Pin-code from Table II using the taxonomy trees in Fig. 1. It has two distinct groups on QID, namely" 1980-89, Female, 684***" and 1990-99, Male, 649***". Since each group contains at least 3 records, the table is 3-anonymous, value of k=3.

If we link the records in vote enrollment database to the records in Table III through QID, each record is linked to either no record or at least 3 records in Table III.

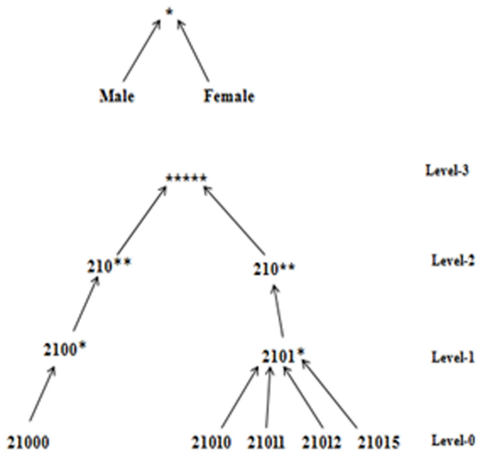

Fig. 2 Domain Generalization Hierarchy

2.2 Anonymization Procedures

An anonymization algorithm can utilize distinctive operations to accomplish the desired level of security. Among these, deterministic mechanisms represent a more appropriate choice when the point is to protect the honesty of the information [2]. One of these techniques is generalization and suppression [8],[9],[19],[20]. Sincethe algorithms assessed in our study utilize generalization and suppression, we give more insights about these operations.

(i) Generalization

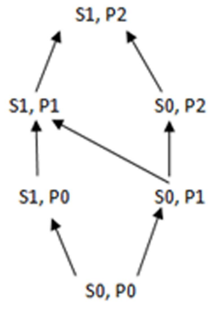

Fig. 3 Generalization lattice

Generalization consists of substituting the values of a given attribute with more generalvalues [8], [19]. To this purpose, the notion of domain (i.e., the set of values that an attribute canassume) is replaced with a set of generalized domains. In other words, Conversion of any value to a more general form is the process of generalization. E.g. "Male" and "Female"can be generalized to "Person". Generalization can be applied on the attribute level (column) and also in cell level.

Generalization Hierarchy [8], [9]**-:** Generalization Hierarchy can be defined as a graph or a grid structure. The nodes of this graph are achieved by generalizingvarious combinations of attributes together at different levels.

Consider two aspects "Gender" and "ZIP Code" of a relation T. Value of attribute Gender at level-0 of generalization can be "Male" and "Female". To achieve level 1 of Generalization with respect to attribute Gender we must generalize the values. We can generalize these two values "Male" and "Female" to another value, say, "Person" or adding asterisk up to a level defined by us. By generalizing the values of attribute Sex to "Person" we achieve level 2 generalization with respect to Gender.Similar is for ZIP code but the level can be increased. By combining different levels of generalization of different attributes we can form the Domain GeneralizationHierarchy:

(ii) Suppression

Suppression means, removing any value totally from an information table [9],[19],[20]. Suppression prompts more information loss when comparedto generalization, because suppression takes awayall the attribute value in the cell. Since it takes off all the details, suppression isapplied just for key characteristics. If we apply Suppression ofthe data we can guarantee that no dangerous implication attack will be carried out. It can be applied to the row level column level and in the complete cell.

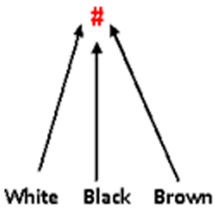

Fig. 4 Suppression of Race

3 Anonymization Algorithms

There are a number of algorithms based on different models of k-anonymity to accomplish k-anonymity. In our relative study, we have chosen three k-anonymization algorithms utilizing generalization and suppression [8], [9], [19]. We have picked these in view of the accompanying reasons: (1) these algorithms have been broadly referred to in the literature(2) these algorithms use different strategies of anonymization permitting a more complete assessment, (3) a public implementation of these algorithms is accessible and (4) these can be assessed inside a similar framework, taking into consideration an all the more reasonable correlation [28]. In the following section, we describe the algorithms applicable to the scope of this workwe likewise show a schematic representationalso, a case for each of the algorithms, with the target of making themeffortlessly conceivable for specialists. (a)Incognito Algorithm (b) Samarati's Algorithm (c) Sweeney's Algorithm.

3.1 Incognito Algorithm

Incognito algorithm [23] produces the set of all conceivable k-anonymous full-domain generalizations of relation T, with anoptional tuple suppression threshold. In the algorithm each iteration consists of two parts. It starts by checking single-attribute subsets of the quasi-identifier, and afterward repeats, checkingk-anonymitywith respect to larger subsets of quasi-identifiers.

3.2 Samarati's Algorithm

This algorithm scans for the conceivable k- anonymous solutions by seizing various levels in Domain Generalization Hierarchy (DGH). It utilizes the binary search to acquire the solution in less time. [24] Samarati makes the assumption that great solutions are the ones where end results in a table have minimal generalizations. Therefore, her algorithm is intended to look at the cross section and distinguish the least level on which at least one solution vector is discovered (the generalizations that fulfill k-anonymity with minimal suppression). This algorithm executes the AG_TS model, generalization is applied in the level of column and suppression is applied at the level of row. *MaxSup* is the greatest number of tuples that are permitted to be suppressed to accomplish k-anonymity.

3.3 Sweeney's Algorithm- Datafly

Datafly algorithm is an algorithm for offering anonymity of Electronic Health Records [25].Anonymization is carried out by means of mechanically generalizing, substituting, inserting and removing statistics without losing details for Demographic research.

3.4 Analogy of Algorithms

Table IV- Analogy of Algorithms

	Algorithm	**Advantages**	**Disadvantages**
1	Incognito [23]	1.The algorithm finds all the k-anonymous full domain generalizations 2. Optimal solution can be selected according to different criteria	1.The algorithm uses breadth first search method which takes a lot of time to traverse the solution space
2	Samarati's [24]	1. Uses the binary search to obtain the solution in less time. 2. Looks for the solution with the least generalization. 3.samarati's output dependably has an opportunity to be an optimal solution 4. Great result when compared to Datafly	1.The chance to get an optimal solution dramatically varies with k, MaxSup lattice size.
3	Sweeney-Datafly [25]	1.The algorithm checks very few nodes for k-anonymity due to which it is able to give results very fast 2.It is a greedy approach that generates frequency lists and iteratively generalizes those combinations with less than k occurrences 3.Practically implementable	1. The algorithm skips many nodes, therefore, resulting data is much generalized and sometimes this released data may not be suitable for research purpose as it Provides very little information. 2.Suppressing all values within the tuple

4 RESULT AND DISCUSSION

In this section we exhibit the outcomes for the tests led utilizing the algorithms clarified in the past section. For our investigation we utilized the Adult datasets, obtained from the UCIrvine Machine Learning Repository [26].The parameters changed in this trials are-:

1. *k-value*: Defines the protection level that must be fulfilled by the anonymization calculation.
2. *Dataset size*: Corresponds to the quantity of records in the dataset.

Varied k-value and Data size -: we provide various k values then observe the result, and find the Execution time of each algorithm, using a configuration of QID=5. Fig. 5, Fig. 6 and Fig. 7shows the result of trials. From the graph obviously, as the number of k value expands the time taken for anonymization is increments, because when k increment the time needing for anonymization is also increases for preserving the privacy of data. In the case of varied data size the anonymization time hasa huge spike increment. As can be seen in figures, execution time of Incognito algorithm has minimal variation with the k value and data size. In the case of Sweeney's algorithm there is large variation of execution time.As the data size increases the curve goes to smooth because with the data size the anonymization operations are decremented.

At the point when contrasting these two algorithms andSamarati's algorithm it has exceptionally insignificant variation and also the execution time is comparatively low, and also when the data size increases it hasn’t any detectable impact in the execution time. So we can conclude that among these three algorithms of anonymization Samarati’s is the best one for anonymization.

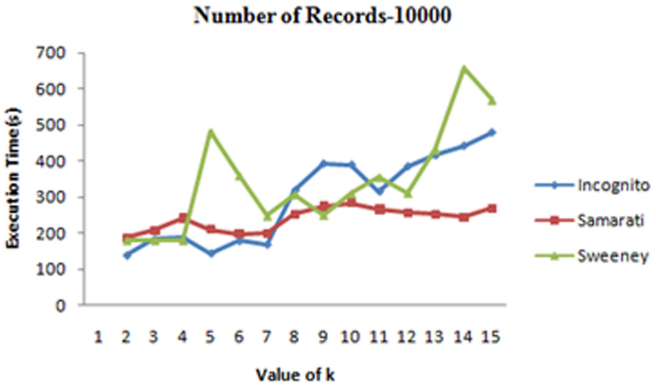

Fig. 5 Dataset with 10000 data’s

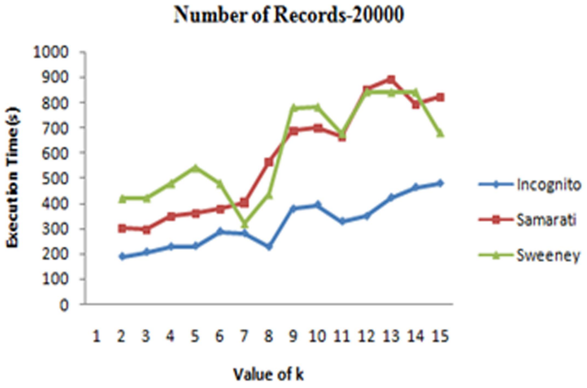

Fig. 6Dataset with 20000 data’s

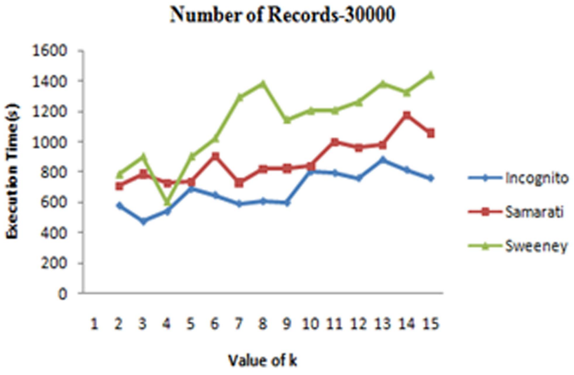

Fig. 7 Dataset with 30000 data’s

5 Conclusion

From the above results, we can understand that “The level of anonymization is directly proportional to the number of records”, the k value has to be chosen in such a way it bridges the gap between the privacy and the released microdata. The choice of algorithm thus plays a crucial role. Experimental results prove that Samarati algorithm is consistent even when the data size increases and would provide better anonymization. Many research extensions are possible with the obtained results, the important and challenging one is being able to find the optimal K value for the Quasi identifier(s) based on the nature of the dataset by maintaining the Anonymization-Privacy Preserving ratio.

REFERENCES

[1] J. Gantz and D. Reinsel, “The digital universe in 2020: Big Data, Bigger Digital Shadows, and
Biggest Growth in the Far East”, Technical report, IDC, sponsored by EMC, 2012.

[2] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala, “Privacy-Preserving Data Publishing.Foundations and Trends in Databases”, 2(1–2):1–167, 2009.
[3] Xiao, X., Tao, Y.: “Personalized privacy preservation”, In: SIGMOD, pp. 229–240, 2006
[4] ArisGkoulalas-Divanis, GrigoriosLoukides, ”Medical Data PrivacyHandbook” ,2015, springer.
[5] Poulis, G., Loukides, G., Gkoulalas-Divanis, A., Skiadopoulos, S.: “Anonymizing data with relational and transaction attributes”, In: ECML/PKDD (3), pp. 353–369, 2013
[6] T. Dalenius, “Finding a needle in a haystack - or identifying anonymous census record. Journal of Official Statistics”, 2(3):329–336, 1986.
[7] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu and Philip S. Yu, “Introduction to Privacy Preserving Data Publishing Concepts and Techniques”,August 2010
[8] P. Samarati and L. Sweeney, “Generalizing data to provide anonymity when disclosing information”, In Proc. of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS), page 188, Seattle, WA, June 1998.
[9] Sweeney, L, “k-anonymity: a model for protecting privacy”, Int. J. Uncertainty Fuzziness Knowledge Based Syst. 10, 557–570 (2002).
[10]M. ErcanNergiz, C. Clifton, and A. ErhanNergiz, “Multirelational k-anonymity. In Proc. Of the 23rd International Conference on Data Engineering (ICDE)”, pages 1417–1421, Istanbul, Turkey, 2007.
[11]A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam." l-diversity: Privacy beyond k-anonymity”, ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), March 2007.
[12]A.Annamalai and G. Deora, “Theoretical Diversity Improvement in GSC (N,L) Receiver With Nonidentical Fading Statistics”, IEEE Transactions on Communications, Volume52 Issue8, 2004.
[13]R. C. W. Wong, J. Li., A. W. C. Fu, and K. Wang, “(α,k)-anonymity: An enhanced k anonymity model for privacy preserving data publishing”, In Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 754–759,Philadelphia, PA, 2006.
[14]N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity”. In Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), Istanbul, Turkey, April 2007.
[15]K. Wang and B. C. M. Fung, “Anonymizing sequential releases. In Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)”, pages 414–423, Philadelphia, PA, August 2006.
[16] C. Dwork, “Differential privacy”, In Proc. of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), pages 1–12, Venice, Italy, July 2006.
[17]V. Rastogi, D. Suciu, and S. Hong, “The boundary between privacy and utility in data publishing” In Proc. of the 33rd International Conference on Very Large Data Bases (VLDB), pages 531–542, Vienna, Austria,September 2007.
[18]A. Blum, K. Ligett, and A. Roth, “A learning theory approach to non-interactive database privacy”, In Proc. of the 40th annual ACM Symposium on Theory of Computing (STOC), pages 609–618, Victoria, Canada, 2008.
[19] L. Sweeney, “Achieving K-anonymity Privacy Protection Using Generalization and Suppression”, Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5):571–588, 2002.
[20]Samarati, P., “Protecting respondents’ identities in microdata release”, IEEE Trans. Knowl. DataEng. 13(6), 1010–1027 (2001).

[21]Kristen LeFevre , David J. DeWitt , Raghu Ramakrishnan, “Incognito: efficient full-domain K-anonymity”, Proceedings of the 2005 ACM SIGMOD international conference on Management of data, June 14-16, 2005, Baltimore, Maryland.

[22]Ciriani V., De Capitiani di Vimercati S., Foresti S., SamaratiP,”k-Anonymity. Security in Decentralized Data Management”, ed. Jajodia S., Yu T., Springer, 2006.

[23] L. Sweeney, “Datafly: a system for providing anonymity in medical data. In Database Security”, XI: Status and Prospects, IFIP TC11 WG11.3 11th Int'l Conf. on Database Security, 356-381, 1998.

[24]K. Bache and M. Lichman. UCI Machine Learning Repository, 2013.

[25]J. Soria-Comas, J. Domingo-Ferrer, D. S´anchez, and S. Mart´ınez, ”Improving the Utility of Differentially Private Data Releases via k-Anonymity”, In Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TRUSTCOM-13, pages 372–379, 2013.

[26]Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. J. Biomed. Inform. 50, 4–19 (2014)