International Journal of Engineering & Technology, 7 (3.4) (2018) 24-27

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



K-Anonymity Versus L-Diversity: A Comparative Analysis on Data Anonymization Techniques

Dr. Sowmyarani C N[1], Dr. Dayananda P[2]

1Associate Professor, Department of CSE, RVCE, Bengaluru 560059,India
2Associate Professor, Department of ISE,JSSATE, Bengaluru 560060,India
**Corresponding author E-mail: dayanandap@gmail.com*

Abstract

The main aim of data publishing is to make the data utilized by the researchers, scientists and data analysts to process the data by analytics and statistics which in turn useful for decision making. This data in its original form may contain some person-specific information, which should not be disclosed while publishing the data. So, privacy of such individuals should be preserved. Hence, privacy preserving data publishing plays a major role in providing privacy for person-specific data. The data should be published in such a way that, there should not be any technical way for adversary to infer the information of specific individuals. This paper provides overview on popular privacy preserving techniques. In this study, a honest effort shows that, concepts behind these techniques are analyzed and justified with suitable examples, drawbacks and vulnerability of these techniques towards privacy attacks are narrated.

Keywords: *privacy preserving data publishing, anonymization, anonymity, background knowledge, privacy disclosure, k-anonymity, l-diversity, t-closeness, similarity attack, Background knowledge attack*

1. Introduction

Apart from traditional privacy preserving techniques such as perturbation [1, 2], inference control [3, 4, 5] etc., a popular technique such as data anonymization [6] evolved as a major technique for privacy preserving data publishing. It further extended by researchers into different types of anonymization techniques. One of the major technique which became popular and basis for other anonymization techniques is K-Anonymity [7].

2. K-Anonymity

This technique is applied on data which is in table form. It categorizes the attributes in a table into 3 categories.

i. Data identifier attributes: These are the attributes, which will directly identifies the person's record and leaves to privacy disclosure [12] such as name, phone number and address of a person.

ii. Quasi-identifier attributes: These are the attributes, by determining these attribute values, the individual record can be identified. Quasi-identifiers are selected based on the type of attributes and values stored in the table.
For example, in a hospital database, where the patients records will be stored in table form, original data may contain Name, age, zip-code and disease of a patient. Age and zip-code are the attributes, where the values of these two attributes combined will relatively disclose the disease value in that record. Particular patient's record will get disclosed. Such attributes are called as quasi-identifiers.

While selecting the quasi-identifier attributes, the selection should follow criteria that, any individual should not be able to know the value from that attribute by using any external available data. If that is the case, then the anonymized data will be difficult for intruder to infer any details of particular individual.

iii. Sensitive attribute: In above example, in a hospital data base, disease information of individual is considered as sensitive information. So, the values of disease attributes are sensitive.
The attribute which reveals sensitive information about an individual is called as sensitive attribute. The value of this attribute is not allowed to determine by an intruder by relating to a particular record which belongs to individual.

k-anonymity Property : The table T(a1, a2, … ,an), and satisfies k-anonymity, if and only if its its Quasi-identifier (Qid), the sequence of values in T[Q-id] contains at least k number of occurrences in T[Qid] where T[Qid] denotes the projection of attributes in Qid, maintaining with duplicate tuples.
In k-anonymity, if the table is said to be k-anonymous, the records in the table should be in such a way that, each equivalence class should have k number of records. Table 2 shows the k-anonymized data for k=3. Therefore, it is said to be 3-anonymous data. Depending on the k value, number of records in the each equivalence class throughout table will vary.

Table 1 Original patterns table

Age	Zip code	Disease
45	15677	Heart Attack
43	15602	Heart Attack
46	15678	Heart Attack
66	15905	Stomach ulcer
57	15909	Heart Attack
74	15906	Lung Cancer
52	15605	Heart Attack

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

58	15673	Lung Cancer
53	15607	Lung Cancer

Table 2. 3-Anonymous version of Table1

Age	**Zip Code**	**Disease**
4*	156**	Heart Attack
4*	156**	Heart Attack
4*	156**	Heart Attack
>=50	156**	Stomach ulcer
>=50	156**	Heart Attack
>=50	156**	Lung Cancer
5*	156**	Heart Attack
5*	156**	Lung Cancer
5*	156**	Lung Cancer

K-anonymity is the basic anonymization technique where lot of research work is still being carried out. But achieving k-anonymity with minimum information loss is very difficult. Meyerson and Williams [13] have shown that, for any public database D, achieving k-anonymity on D with minimum number of suppressed entries is NP-Hard [14]. Major analysis is carried out on k-anonymity [19, 20] techniques. There is complete (alpha, k)-anonymity [15] technique which reduces the data distortion using clustering method in privacy preserving data mining [16, 17], P-cover k-anonymity [27] which protects multiple sensitive attributes and personalized (a,k)-anonymity [18] uses local recoding to keep information loss low.

Table 3. Inpatient Micro data

Non-Sensitive			**Sensitive**
Country	Age	Zip-Code	Health Condition
Canada	27	15308	Hep-B
USA	28	15347	Hep-B
Canada	26	15364	Lung Cancer
USA	25	15369	Lung Cancer
Japan	41	13853	HIV
China	48	13874	Tuberculosis
India	45	13864	Sinusitis
India	42	13862	Heart Attack
Canada	33	15362	Cough
USA	37	15304	Cough
USA	36	15305	Cough
Canada	35	15368	Gastric ulcer

Table 4. 2-Anonymous table view of Table 3

Non-Sensitive attribute			Sensitive Attribute
Country	Age	Zipcode	Disease
America	<=30	153**	Hep-B
America	<=30	153**	Hep-B
America	<=30	1536*	Lung cancer
America	<=30	1536*	Lung cancer
Asia	>=40	138**	HIV
Asia	>=40	138**	Tuberculosis
Asia	>=40	138**	Sinusitis
Asia	>=40	138**	Heart Attack
America	3*	1536*	Cough
America	3*	1536*	Cough

Table 3 shows health records of patients registered under some hospital. This contains country, age zip code and disease information. This original data contained in table should be anonymized using k-anonymity. Observing the attributes and the values contained, country, age and zipcode are considered as non-sensitive attributes, in turn chosen as quasi-identifiers. Disease is considered as sensitive attribute. Table 4 shows the 2-anonymous version of table 3.

Table 5. External Information available

Name	Age	Country	Zip Code
Rima	27	USA	15364
Nicky	43	India	13864
Raghu	26	Canada	15369
Wang	44	Japan	13874

Generalization and suppression is done to achieve k-anonymity in table 4. The country attribute is having country names as its values. These are generalized to name of the continent to which the country belongs to. For example, in table 4, the values such as Canada and USA are replaced by continent as generalized value and Japan, china and India are replaced by Asia as generalized value for country attribute. The values for zip code and age attributes are generalized and suppressed as well. For example, zip codes starting with 153 are suppressed as 153**. Some age attribute values are generalized and some are suppressed. If age value is greater than 30 and less than 40, it will be suppressed as 3*. If it is any value less than or equal to 30, it is generalized as <=30. Table 5 shows the external information for intruder by which he can gain background knowledge about these individuals say, Rima, Nicky, Raghu and Wang. Intruder who has this background knowledge relates this with the 2-anonymous table 4; he can finally conclude that, Rima and Raghu are having Lung cancer leading to background knowledge attack.

If the same intruder get to know about some information regarding an individual say X as, X lives in area with zip code starting from 153 and his age is greater than 30. Intruder relates this information with 2-anonymous table 4. He will get to know that, X record belongs to last equivalence class. Since both the records are having homogenous value for sensitive attribute disease as cough, he can conclude that, X is having Cough. This way of attack on privacy is called as homogeneity attack. k-anonymity protects identity disclosure [21]. But it is vulnerable to attribute disclosure [11, 22] which is in turn homogeneity attack [20] and background knowledge attack [43]. To overcome these drawbacks, another model came into existence called p-sensitive k- anonymity [22, 23, 24, and 25]. But, K-anonymity is most popular and simple way of anonymizing the data. Hence it has been used in the different fields for preserving privacy. This technique is effectively being used in Location based services in mobile computing and healthcare[39] services.

L-DIVERSITY

Definition: l-diversity technique [8, 9] states that, a table T satisfies l-diversity, if and only if each of the equivalence classes in a table must contain l well represented values for sensitive attribute.

Table 6. Original pattern

Zip Code	**Age**	**Disease**
15056	42	Gastritis
15024	44	Headache
15262	56	Flu
15345	59	Flu

Table 7. 2-diversity micro data

Zip Code	**Age**	**Disease**
150**	4*	Gastritis
150**	4*	Headache
152**	5*	Flu
152**	5*	Flu

Table 6 shows the original pattern of the data that has been collected and Table 7 gives the 2-diverse pattern after applying the l-diversity technique. l-diversity [31, 34, 35, 40, 41] is difficult to achieve and it may not provide the sufficient protection as it is subject to similarity attack.

Table 8. Original Table

Age	**Zip Code**	**Salary**	**Disease**
49	34677	3K	Blood cancer
42	34602	4K	Lung cancer
47	34678	5K	Stomach Cancer
53	34905	6K	Lung cancer
52	34909	11K	Flu
57	34906	8K	Phthisis
20	34605	7K	Phthisis
26	34673	9K	Anemia
22	34607	10K	Stomach Cancer

Table 8 shows the original data, and Table 9 shows the data after applying l-diversity technique which leads to anonymized data satisfying 3-diversity. In the original table 8, two attributes are identified as sensitive attributes namely Salary attribute and Disease attribute. If an intruder get the published data with shown in table 9, by knowing the detail of some person X such as age is in 40's, he can identify that, his record belongs to first equivalence class and easily conclude by analyzing salary attribute values in that equivalence class and comparing those values with other values in the whole table that, person X is having low salary causing privacy disclosure. T-closeness [10] is a technique where it evolved to improve the privacy preserved in the data sets. But, it hampered the correlation between the Qids.

Table 9. 3-Diverse Version of Table 10

Age	Zip Code	Salary	Disease
4*	346**	3K	Blood cancer
4*	346**	4K	Lung cancer
4*	346**	5K	Stomach Cancer
>50	349**	6K	Lung cancer
>50	349**	11K	Flu
>50	349**	8K	Phthisis
2*	346**	7K	Phthisis
2*	346**	9K	Anemia
2*	346**	10K	Stomach Cancer

3. Comparative Analysis

The comparative analysis shows the strengths and drawbacks of K-anonymity and l-diversity techniques as depicted in the below Table 10.

4. Future Scope And Conclusions

Many anonymization techniques [37, 38, 39] evolved to protect private information and provide the utility of data while preserving privacy. There is always a tradeoff existing between preserving privacy and providing utility. The techniques are under one or the other type of privacy attack [36]. The majority of privacy attacks are not being overcome by any of the techniques. If the technique should not be vulnerable to any of privacy attack, the utility will be reduced in a higher level. There is need of robust technique which preserves privacy and as well as provide utility by not being vulnerable to privacy attacks. As the adversary gathers more and more information on target person, inferring the target person detail from published data may become easy for him. Techniques which do not allow adversary to predict the target person information with background knowledge should evolve to provide more privacy.

Table 10. Comparative Analysis

Anonymization Technique	Drawbacks	Strengths
K-Anonymity	➢ It is not difficult to achieve in contrast with l-diversity ➢ There is no level of sensitivity described ➢ Inefficient to prevent homogeneity attack ➢ There is improved k-anonymity models such as p-sensitive k-anonymity[23, 24], (p, alpha)-sensitive k-anonymity [26]	➢ K-Anonymity is popular and basis for many data anonymization techniques
l-diversity	➢ It is very difficult to achieve l-diversity if the sensitive attribute values are not more distinct ➢ There will be different levels of sensitivity for single sensitive attribute value ➢ Achieving l-diversity may be unnecessary ➢ L-diversity[25] is inefficient to prevent attribute disclosure ➢ This type of disclosure not only occurs for numerical sensitive attribute, it also occurs for categorical sensitive attributes. ➢ There is an improved l-diversity model [28] and fast p-sensitive l-diversity model [29] to preserve numerical sensitive attribute values and enhanced and extended l-diversity [31, 32, 33, 34] models in the literature, but limitations of l-diversity technique is overcome by (*a*, *d*)-diversity [30].	➢ L-diversity preserves nore privacy as it includes diverse attribute values within equivalence classes

References

[1] Duncan, G. and Fienberg, S. Obtaining information while preserving privacy: A Markov perturbation method for tabular data. In Statistical Data Protection, 1998, 351–362.

[2] Xiao-Bai Li; Sarkar, S., "A Tree-Based Data Perturbation Approach for Privacy-Preserving Data Mining," Knowledge and Data Engineering, IEEE Transactions on , vol.18, no.9, pp.1278,1283, Sept. 2006

[3] Haibing Lu; Yingjiu Li; , "Practical Inference Control for Data Cubes," Dependable and Secure Computing, IEEE Transactions on , vol.5, no.2, April-June 2008, pp.87-98.

[4] Yingjiu Li; Haibing Lu; Deng, R.H.; , "Practical inference control for data cubes," Security and Privacy, 2006 IEEE Symposium, 21-24 May 2006, pp.6-120.

[5] R. Brand. "Microdata protection through noise addition" In inference control in statistical databases, From theory to practice, London UK, 2002, Pages 97-116.

[6] L. Sweeney, "k-anonymity: a model for protecting privacy", International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 2002, pp. 557-570.

[7] L. Sweeney, “k-anonymity: a model for protecting privacy”, International Journal on Uncertainty, Fuzziness and Knowledge based Systems, 2002, pp. 557-570.

A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,” l-diversity: Privacy beyond k-anonymity”. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), 2006, pp1- 24.

[8] P. Samarati. Protecting respondents’ identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027, 2001.

[9] Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian: “Closeness: A New Privacy Measure for Data Publishing”, IEEE Trans. Knowl. Data Eng. 22(7): 943-956 (2010)

[10] L. Willenborg and T. DeWaal. Elements of Statistical Disclosure Control. Springer-Verlag, New York, 2001.

[11] Machanavajjhala, A.; Kifer, D.; Abowd, J.; Gehrke, J.; Vilhuber, L.; , "Privacy: Theory meets Practice on the Map," Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference, 7-12 April 2008, pp.277-286.

A. Meyerson and R. Williams. “On the complexity of optimal k-anonymity [C]”. In: Proc of the ACM SIGMOD Int’l Conf on Principles of DB Systems. New York: ACM Press, 2004.pp. 223-228.

[12] Bayardo, R.J.; Agrawal, R., "Data privacy through optimal k-anonymization," Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference, 5-8 April 2005, pp.217-228.

[13] Han Jian-min; Yu Hui-qun; Yu Juan; Cen Ting-ting, "A Complete (alpha,k)-Anonymity Model for Sensitive Values Individuation Preservation," Electronic Commerce and Security, 2008 International Symposium, 3-5 Aug. 2008, pp.318-323.

[14] Jian Wang; Yongcheng Luo; Yan Zhao; Jiajin Le, "A Survey on Privacy Preserving Data Mining," Database Technology and Applications, 2009 First International Workshop , , 25-26 April 2009, pp.111-114.

[15] Saranya, K.; Premalatha, K.; Rajasekar, S.S., "A survey on privacy preserving data mining," Electronics and Communication Systems (ICECS), 2015 2nd International Conference, 26-27 Feb. 2015, pp.1740-1744

[16] Xiaojun Ye; Yawei Zhang; Ming Liu, "A Personalized (a,k)-Anonymity Model," Web-Age Information Management, 2008. WAIM '08. The N.inth International Conference, 20-22 July 2008, pp.341-348.

[17] Vijayarani, S.; Tamilarasi, A.; Sampoorna, M., "Analysis of privacy preserving K-anonymity methods and techniques," Communication and Computational Intelligence (INCOCCI), 2010 International Conference, 27-29 Dec. 2010,pp.540-545.

[18] L. Sweeney " k-ANONYMITY: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002, Pages 557-570.

[19] MIKLAU, G. AND SUCIU, D. 2004. A formal analysis of information disclosure in data exchange. In Proceedings of the ACM SIGMOD. ACM, New York, pp. 575–586.

[20] Xiaoxun Sun; Hua Wang; Lili Sun “ Extended K-Anonymity Models Against Attribute Disclosure” Third International Conference on Network and System Security, 2009. NSS '09. pp. 130 – 136.

[21] Xiaoxun Sun; Hua Wang; Jiuyong Li; Ross, D. “Achieving P-Sensitive K-Anonymity via Anatomy” ICEBE '09. IEEE International Conference on e-Business Engineering, 2009, pp. 199 – 205.

[22] Truta T M , Vinay B, “Privacy protection: p-Sensitive k-anonymity property,” Proceedings of the 22nd on Data Engineering Workshops, IEEE Computer Society, Washington Dc, 2006.

[23] Sowmyarani C N and Dr. G N Srinivasan. Article: Survey on Recent Developments in Privacy Preserving Models. International Journal of Computer Applications 38(9), January 2012. Published by Foundation of Computer Science, New York, USA, pp. 18-22.

[24] Xiaoxun Sun, Hua Wang, Jiuyong Li, Truta, T.M, Ping Li, “(p+, α) -sensitive k-anonymity: A new enhanced privacy protection model” , 2008, pp.59-64.

[25] Yingjie Wu; XiaowenRuan; Shangbin Liao; Xiaodong Wang; “P-Cover K-anonymity model for Protecting Multiple Sensitive Attributes”, The 5th International Conference on Computer Science & Education Hefei, China. August 24–27, 2010

[26] Jianmin Han; Huiqun Yu; Juan Yu “An improved l-diversity model for numerical sensitive attributes” Third International Conference on Communications and Networking in China, 2008. ChinaCom 2008. Year: 2008, pp. 938 – 943.

[27] Tripathy, B.K.; Maity, A.; Ranajit, B.; Chowdhuri, D., "A fast p-sensitive l-diversity Anonymisation algorithm," Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE, 22-24 Sept. 2011, pp.741-744.

[28] Qian Wang; Xiangling Shi “(a, d)-Diversity: Privacy Protection Based on l-Diversity WRI World Congress on Software Engineering”, WCSE '09, Volume: 3, 2009, pp. 367 – 372.

[29] Machanavajjhala, A.; Kifer, D.; Abowd, J.; Gehrke, J.; Vilhuber, L.; , "Privacy: Theory meets Practice on the Map," Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference, 7-12 April 2008, pp.277-286.

[30] Gaoming Yang; Jingzhao Li; Shunxiang Zhang; Li Yu “An enhanced l-diversity privacy preservation 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2013, pp. 1115 – 1120.

[31] Hongwei Tian; Weining Zhang, "Extending l-Diversity for Better Data Anonymization," Information Technology: New Generations, 2009. ITNG '09. Sixth International Conference, 27-29 April 2009, pp.461,466.

[32] Lefevre, K.,Dewitt, D. J., And Ramakrishnan , “ Workload-aware anonymization”. In Proceedings of the 12th ACM SIGKDD. ACM, New York.2006.PP 277-286.

[33] Machanavajjhala, Kifer, D., Gehrke, J., And Venkitasubramaniam, M,” l-diversity: Privacy beyond k-anonymity”, ICDE '06. Proceedings of the 22nd International Conference on Data Engineering, 2006.

[34] Vora, P.L., "An Information-Theoretic Approach to Inference Attacks on Random Data Perturbation and a Related Privacy Measure," Information Theory, IEEE Transactions on , vol.53, no.8, Aug. 2007, pp.2971-2977.

[35] J. Wang, Y. Li, D. Yang, H. Gao, G. Luo and J. Li, "Achieving Effective -Anonymity for Query Privacy in Location-Based Services," in IEEE Access, vol. 5, 2017, pp. 24580-24592.

[36] Y. Zhang, W. Tong and S. Zhong, "On Designing Satisfaction-Ratio-Aware Truthful Incentive Mechanisms for k -Anonymity Location Privacy," in IEEE Transactions on Information Forensics and Security, vol. 11, no. 11, Nov. 2016, pp. 2528-2541.

A. Aristodimou, A. Antoniades and C. S. Pattichis, "Privacy preserving data publishing of categorical data through k-anonymity and feature selection," in Healthcare Technology Letters, vol. 3, no. 1, 2016, pp. 16-21.

[37] K. Oishi, Y. Tahara, Y. Sei and A. Ohsuga, "Proposal of l-diversity algorithm considering distance between sensitive attribute values," 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, 2017, pp. 1-8.

[38] Y. Sei, H. Okumura, T. Takenouchi and A. Ohsuga, "Anonymization of Sensitive uasi-Identifiers for l-diversity and t-closeness," in IEEE Transactions on Dependable and Secure Computing, 2017 vol. PP, no. 99, pp. 1-1.