A Study on k-anonymity, l-diversity, and t-closeness Techniques of Privacy Preservation Data Publishing

Twinkle Patel
Assistant Professor
Sal College of Engineering, India

Dr Kiran Amin
Principal
Ganpat University, India

Abstract

Most organizations are dealing with massive amounts of information collection and are stored in large databases. Personal health record (PHR) is an emerging patient-centered model of exchange of health information, often outsourced for third-party storage, such as cloud providers. There have been wide-ranging issues regarding privacy, however, as personal health data may be exposed to those third party servers and unauthorized parties. This work aims to highlight three of the popular strategies for clinical anonymization, namely k-anonymity, l-diversity, and t-closeness. There is also a summary of the benefits and weaknesses of these strategies. Extensive analytical and experimental findings are presented showing our proposed scheme's security, scalability and performance.

Keywords: health records; data anonymization; k-anonymity; l-diversity; t-closeness

I. Introduction

While technology is expanding rapidly, so are the numerous cybercrimes, such as internet phishing, in which confidential information is compromised, which raises concerns regarding data privacy and security among consumers and businesses globally. In addition, the use of social networking sites, electronic healthcare systems, online trading, etc. has created a large number of large data sets. In addition, the use of social networking sites, electronic healthcare systems, online trading, etc. has created a large number of large data sets. There is therefore a high demand for privacy-preserving information publishing (PPDP) for internet-protected data sharing. Several de-identification and anonymization strategies were applied to protect the information proprietor's privacy before the data is disclosed to the public or for secondary use[10].

In this paper, the flow of the study will be as follows. This paper will elaborate on PPDP and data anonymization in Section II. Section III will include literature reviews on data anonymizations techniques that have been adopted in the medical field in chronological order (based on the published date).

II. Privacy preserving data publishing

In short, PPDP sanitizes personal data (e.g. digital health records) extremely likely to be made accessible to organizations or the government. The intruder can be anyone (data recipient) who gets personal information about an entity as shown in Figure 1 below. Therefore, it is the data publisher's essential duty to implement various privacy protection measures to control the information released by changing it prior to publication[3][4].

A myriad of highly complex data mining techniques have been developed as a solution to alleviating data breaches and preventing fines from government agencies. During each stage of data mining, it is important to enforce techniques that allow data privacy and secure information exchange. The techniques used to secure sensitive data are the restriction of access and distorting data. Encryption techniques such as Identity-Based Encryption (IBE) and Attribute-Based Encryption (ABE) are well known in the data depository process to secure when data is stored in the cloud provider or clinical server[2][9]. The PPDP is acquired mainly in the Big Data Analytics data processing phase that will be the subject of this study. Data anonymisation, also known as data masking or desensitization, is used to hide or conceal any sensitive data about an individual, thus preventing the re-identification of the individual[7][17].

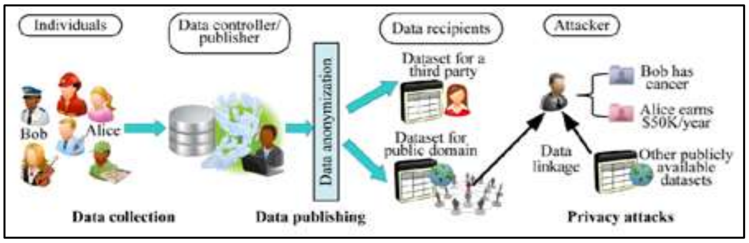
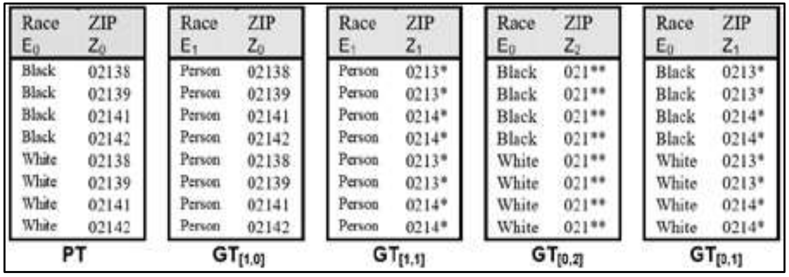


Fig. 1: Outline of Privacy Preserving Data Publishing (PPDP) [4]

Table - 1
Generalization with suppression within a private table (PT)



Race E_0	ZIP Z_0	Race E_1	ZIP Z_0	Race E_1	ZIP Z_1	Race E_0	ZIP Z_2	Race E_0	ZIP Z_1
Black	02138	Person	02138	Person	0213*	Black	021**	Black	0213*
Black	02139	Person	02139	Person	0213*	Black	021**	Black	0213*
Black	02141	Person	02141	Person	0214*	Black	021**	Black	0214*
Black	02142	Person	02142	Person	0214*	Black	021**	Black	0214*
White	02138	Person	02138	Person	0213*	White	021**	White	0213*
White	02139	Person	02139	Person	0213*	White	021**	White	0213*
White	02141	Person	02141	Person	0214*	White	021**	White	0214*
White	02142	Person	02142	Person	0214*	White	021**	White	0214*
PT		$GT_{[1,0]}$		$GT_{[1,1]}$		$GT_{[0,2]}$		$GT_{[0,1]}$	

PPDP uses methods of anonymization, such as generalization and denial, to protect the data by changing it to conceal the authentic sensitive data. Also used to anonymize huge data networks are graph-based methods such as restricted perturbation. Such methods are further broken down into k-anonymity, l-diversity, grouping, clustering, law of association, condensation and cryptography[2]. In this paper focusing on the healthcare field, some of these approaches will be explained in detail. High-dimensional information obtained from heterogeneous sources is considered to have medical data, and these health care data are recently digitized to minimize costs and improve efficiency. This results in the development of Electronic Health Records (EHRs), which store patient-specific information in a centralized repository, ranging from their statistics to health criteria, and recently cloud-based medical data storage, becoming widespread. To collect personal information, be vulnerable to numerous cyberattacks.

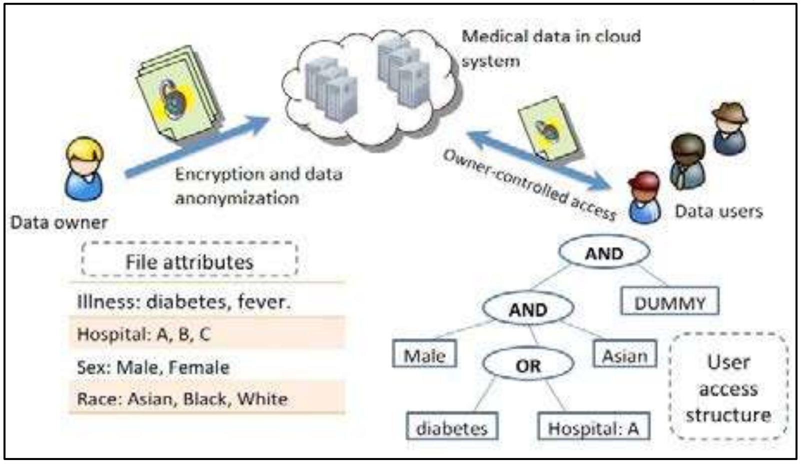


Fig. 2: Anonymization of medical data [13]

III. Literature review

Data publishers such as healthcare institutions have a greater responsibility to share patient information without allowing adversaries to identify a person and preserve confidentiality using the k-anonymity technique. K-Anonymity is a security concept that guarantees that it is not possible to identify different records in a single dataset. The datasets are said to be k-anonymous only when a single row is identical to a minimum of (k – 1) rows. Therefore, k-anonymity can be used to prevent database linkages [4].

This paper illustrates how generalization and denial are used to achieve k-anonymity. In generalisation, the original value is modified to fill in a general yet syntactically stable value, whereas the values are expressed in asterisk'*' in suppression. When k-anonymity is applied, the power to connect a person who distinguishes data with others through quasi-identifiers is limited. Examples are a zip code, age, date of birth, name, address, etc. This paper elucidates various algorithms and theorems. The original quality is generalized in suppression by removing those elements but preserving its fidelity.For instance, if the zip code, Z is 02139 (ground domain, Z0) it is changed to 0213* (Z1) where the last digit is removed (suppressed) or it can be a zero. As the hierarchy increases, the zip code becomes more generalized till it achieves maximum suppression (*****). Another method used is a generalization of a table at attribute (column) level. Here, the generalized table contains tuples (rows) whereby each tuple is similar to a minimum of (k-1) other tuples within the same table. Lastly, minimal distortion in a table is related to minimal generalization using several theoretical algorithms [16].

Having identified that k-anonymity is prone to certain privacy attacks, which can be prevented when applied with l-diversity technique. Due to the limitations in k-anonymity and Bayesian optimal privacy such as the background knowledge of the attacker and probability of the principles of l-diversity for increasing the data privacy. This principle states that a generalized quasi-identifier (q*)-block is l-diverse if it contains a minimum of 'l' properly depicted values under the sensitive attribute (S). If every q*-block is l-diverse, then the table meets the l- diversity concept. There are two variations of l-diversity by incorporating examples from medical microdata. Entropy l-diversity is used to counter the uniformity of values in a table. If there are two values in a medical dataset, which are healthy and not healthy, then the healthy value is represented as don't-care sets, which can be handled by entropy l-diversity. For recursive (c,l)-diversity, the table meets the requirements if every q*-block agrees to the following function:

$$ri < c\,(rl + rl{+}1 + \ldots + rm)$$

where c is a constant and r i is the repetition of the sensitive value that appears in the q*-block

Ultimately, this paper shows multiple algorithms to prove l-diversity's ability to recognize imperfections in k-anonymity to solve homogeneity attack and background information.

Table – 2
Use of l-diversity (3-diverse table) in patient microdata

	Postcode	*Age*	*Disease*
1	*570***	*3**	*Stomach ulcer*
2	*570***	*3**	*Stomach cancer*
3	*570***	*5**	*Stomach ulcer*
4	*433***	*2**	*Dengue*
5	*433***	*3**	*Flu*
6	*433***	*2**	*Sinus*
7	*432***	*6**	*Calcium deficiency*
8	*432***	*5**	*Diabetic*
9	*432***	*2**	*Flu*

T-closeness is a privacy protection strategy introduced to overcome the limitations of existing methods of k-anonymity and l-diversity. In l-diversity, it is presumed that if the distribution of the attribute is known, which is a drawback of this approach, the adversary will acquire knowledge on a sensitive attribute. However, most strategies for protecting confidentiality presume that the attributes have definite, i.e. categorical, values. Distribution skewness and semantic similarity of the sensitive values in the equivalence class are possible attacks faced by the l-diversity technique. The principle of t-closeness is defined as, if the distance between the sensitive attribute of an equivalence class and that of the whole table is less than or equals to a threshold, t then the equivalence class possess t-closeness. This reduces the risk of the opponent learning unique information of an individual. The distribution distance between the sensitive attributes is measured using the metric called Earth Mover's Distance (EMD) which takes into account the feature/attribute values' semantic proximity. However, this technique retains the transparency of the feature, but it also exposes the identity. K-anonymity and t-closeness will both work together to safeguard the confidentiality of published data[11].

Another issue was raised in 2008 in preserving the privacy of string data such as genomic and biological data. On pseudo-data, an alternative of k-anonymity called condensation was used to hide the actual record values without affecting multi-dimensional statistical data. The anonymization is done here by summarizing statistical data sets that are used to create pseudo-strings. Such pseudo-data are similar to the original strings generated from the symbols ' distribution information containing the probabilistic measurements. These strings are studied for several aggregate enumerations like the consistency of the structure, alignment of distance within the strings and accuracy of mining algorithms like classification. Condensation approach also reveals that classification precision is strongly maintained and decreases slightly with larger group sizes, while retaining statistical data's originality. This method is very useful in the medical field to recognize the disease patterns or physical features (e.g. eye color, hair quality) that are due to DNA string sections, where the individual record can be uncovered. However, the data are pseudonymized, which provides another level of security to the underlying information [1].

Transferring medical information to multiple nodes and data providers through wireless sensor networks presents many privacy issues. Clearly detecting data is not enough to secure the privacy of the attacker's personal data. For example, if the patient is alone in a hospital room, the attacker will relate the node signal that transmits the unique ID to medical parameters such as a patient's heart rate. To prevent this, it is possible to apply the generalization method under k-anonymity to make the node IDs indistinguishable from one another. The less descriptive the feature, the higher the anonymization level. In k-anonymity, to make it less recognizable, the ground value (specific, original value) is mapped to a generalized value. Some of the safety concerns posed in this paper regarding health sensor networks are eavesdropping transmitted data information, modifying data in the sensor after receiving data, and tracking traffic. The data collected from different sensors are grouped into clusters that increase energy efficiency and reduce channel interference, even with large data sets. If the node numbers in a cluster do not match the pre-set' k' threshold value, then false data is entered before the signal is sent to the main station to satisfy the k-value. This results in the adversary from assuming the patient identity as the clusters meet the k-value [5].

IV. DISCUSSION

A. k-Anonymity

Many information owners, including government agencies and hospitals, believe that data, such as medical records, may remain anonymous if specific details such as name, address and telephone number are withheld before the rest of the documents are revealed. Re-identification of the person, however, by linking the data with other published data. When adding noise to the dataset such as false values and scrambling may provide anonymity, this will result in inaccurate statistical results in tuples when conducting data mining and analysis. In 1998, Samarati and Sweeney formalized a technique called k-anonymity, which uses methods of generalization and suppression to allow data disclosure in a controlled manner. Quasi-identifiers are special characteristics that identify the birth date and sex of a person. A table containing such quasi-identifiers is said to follow k-anonymity if at least' k' times reappear per tuple value of the quasi-identifiers, making the tuple distinguishable from each other[14][16].

1) *Principle of k-Anonymity*

If each value is indistinguishable from a minimum of (k-1) records from the same table in a given dataset, then the table is said to be k-anonymous. The higher the price of k, the higher the protection of privacy[8].

2) *Generalization*

Generalization is a method used to represent the values of the attributes in a table to make tuple recognition less discreet. The original attribute is defined as a ground domain in this process, and with that generalization the domain value increases. Quasi-identifiers are mapped from Z0 (02123, 02126) to Z1 (02120, 02120) like zip code to generalize the values and at the same time not lose the truthfulness of the data. This is called the hierarchy of generalization of the domain. K-minimum generalization is used for private tables with specific values if the table has already exceeded k-anonymity within the table[14][16]. The drawback of this approach is that if there are smaller outliers, i.e. tuples that occur less than k-times, there will be a need for a high level of generalization[14].

3) *Suppression*

Suppression is used with generalization to support k-anonymity. Suppression is a technique used in the quasi-identifiers to cover those values[14]. The suppressed value is defined with an asterisk(*) which can be applied to hierarchies for the generalization of domain and quality. Considering the example in the generalization section, the mapped value can be suppressed as Z1 (0212*, 0212*) and further suppressed to Z2 (021**), then reach maximum suppression (*****) [16].

4) *Pros of k-Anonymity*

- It defends against disclosure of identity by inhibiting connections to a database with values below ' k. ' This prevents the opponent from linking sensitive data to external data[8][15].
- Compared to other anonymity methods such as cryptographic solution[5], the cost of developing this system is considerably lower.
- There is extensive use of k-anonymity algorithms such as Datafly, Incognito, and Mondrian, particularly in PPDP.

5) *Cons of k-Anonymity*

This technique has found many drawbacks, mainly attacks such as unsorted matching, complementary release, minimality, and temporal attacks[8][9][16]. Certain drawbacks include this method, which can cause a high loss of usefulness if used in high-dimensional data and extraordinary precautions are needed if the published information has already been anonymised. However, in this research two of the well-known attacks on k-anonymity will be briefed below.

- Homogeneity attack: This can produce clusters which disclose information when there is insufficient diversity in the sensitive attributes. Suppose A and B are competitors, and A knows that B lives in a certain zip code and is of a specific age, and wants to know the medical status of B. So, with A’s insight on B, A can identify that the information matches with a number of medical records and all these records have the same medical condition (sensitive attribute), i.e. cancer. Thus, the k- anonymous table should be further sanitized by diversifying the sensitive values within the tuples that share similar values of their quasi-identifiers [8] [12].
- Background knowledge attack: In this type of attack, the attacker has a proven knowledge of the individual and the critical characteristics of the individual can be leaked with additional logical reasoning. See A and C as friends, and A would like to believe the personal information of C found in the same list of patients as B. As A says, C is a 45-year-old Asian woman living in a particular zip code. Nonetheless, the record shows that C may have any of the three diseases— cancer, heart disease, and viral infection. Based on the background information from A that C avoids high-calorie meals and has low blood pressure, A infers heart disease from C. Hence, k-anonymity is prone to background knowledge attack [8] [12].

B. l-Diversity

To overcome the drawbacks of k-anonymity, l-diversity was suggested. They also implemented a novel approach as an extension to k-anonymity that can ensure data privacy even without revealing the background knowledge of the adversary to prevent disclosure of attributes. This approach revolves around the idea that each category has "well-represented" responsive attributes. This technique is a modification of k-anonymity by incorporating the k- anonymity principle [11] [12].

1) *Principle of l-Diversity*

A k-anonymous table is said to be l-diversely if, for each sensitive attribute, the equivalence class in the table has at least ' l ' well-represented values[6][12]. You should elucidate the word "well-represented" according to the following principles:

Distinct l-diversity: within the equivalence group, a value occurs more recurrently than other values.

Entropy l-diversity: For each equivalence group, the entire table must have at least log(l) as entropy. In the case of low entropy of the whole table, this strategy may be too prohibitive if only a few values are the same.

Recursive (c, l)-diversity: if the sensitive values in each equivalence group do not occur either too often or too rarely, a table is said to adhere to this rule. This notion is better than the above two notions[11][12].

2) *Pros of l-Diversity*

- Provides greater distribution within the community of critical attributes, thus improving data protection.
- Protects against disclosure of information, an expansion of the technique of k-anonymity.

- Thanks to faster pruning by the l- diversity algorithm[6][11][12], the quality of l-diversity is slightly better than k-anonymity.

3) *Cons of l-Diversity*

- l-diversity can be redundant and laborious to achieve.
- Prone to attacks such as distorted attack and similarity attack because it is ineffective to avoid exposure to attributes due to the semantic relationship between the relevant attributes[9][11].

C. *t-Closeness*

Improving l-diversity is a technique of t-closeness by increasing the granularity of the information interpreted. The level of information on a particular data by the researcher is limited whereas the knowledge is not restricted to the overall table comprising the datasets. Therefore, this reduces the correlation between the quasi-identifier attributes and the sensitive attributes. Earth Mover's Distance (EMD) is used to measure the distance between distributions.

1) *Principle of t-Closeness*

The equivalence class t-closeness is reached when the distance of the responsive attribute in this category is not greater than the limit, t with the distance of the attribute in the entire table. If all equivalence groups have t-closeness[11], the table is recognized as having t-closeness.

2) *Pros of t-Closeness*

- It prevents the disclosure of attributes that protect the privacy of data.
- Protects against homogeneity and attacks of background knowledge in k-anonymity.
- This defines the attribute's linguistic proximity, a constraint of l-diversity.

3) *Cons of t-Closeness*

- Using the measure of Earth Mover's Distance (EMD) in t-closeness, the relation between t-value and the knowledge gained is difficult to identify.
- Needs the responsive attribute to be similar to that in the overall table[9][11] in the equivalence group.

V. CONCLUSION

Data privacy protection has arisen as a definite condition to data privacy preservation. The spike in cybercrimes has resulted in a severe risk of breach of privacy. This has contributed to the emergence of various techniques for anonymization. This paper addressed these growing issues in PPDP and translated into the healthcare domain, which provides greater opportunities for disclosure per person. Based on the academic literature devoted to PPDP, a number of anonymization approaches applied to medical data are outlined here to counteract this. In addition, the scope of this work is limited to the methodology of k-anonymity with its extended modifications, i.e. l-diversity and t-proximity. Each of these methods has been further illuminated with concepts and associated references. The analysis of these three strategies ' advantages and disadvantages has also been rationalized. Finally, this document is committed to providing a brief summary of the current trends in anonymization techniques which direct medical information in achieving privacy protection under PPDP.

REFERENCES

[1] Aggarwal, C.C. and Yu, P.S. (2008) ‘A framework for condensation-based anonymization of string data’, Data Mining and Knowledge Discovery, 16(3), pp. 251–275. doi: 10.1007/s10618-008-0088-z.

[2] Aldeen, Y.A.A.S., Salleh, M. and Razzaque, M.A. (2015) ‘A comprehensive review on privacy preserving data mining’, SpringerPlus, 4(1). doi: 10.1186/s40064-015-1481- x.

[3] Allard, T., Nguyen, B. and Pucheral, P. (2013) ‘METAAP: Revisiting privacy-preserving data publishing using secure devices’, Distributed and Parallel Databases, 32(2), pp. 191–244. doi: 10.1007/s10619-013-7122-x.

[4] Ayala-Rivera, V., Mcdonagh, P., Cerqueus, T. and Murphy, L. (2014) ‘A systematic comparison and evaluation of k- anonymization Algorithms for practitioners’, TRANSACTIONS ON DATA PRIVACY, 7(3), pp. 337–370.

[5] Belsis, P. and Pantziou, G. (2012) ‘A k-anonymity privacy- preserving approach in wireless medical monitoring environments’, Personal and Ubiquitous Computing, 18(1), pp. 61–74. doi: 10.1007/s00779-012-0618-y.

[6] Casas-Roma, J., Herrera-Joancomartí, J. and Torra, V. (2016) ‘A survey of graph-modification techniques for privacy-preserving on networks’, Artificial Intelligence Review,. doi: 10.1007/s10462-016-9484-8.

[7] Emam, K.E. (2007) Data Anonymization practices in clinical research. [Online] Available at: http://www.ehealthinformation.ca/wp-content/uploads/2014/07/2006-Data-Anonymization- Practices.pdf (Accessed: 2 February 2017).

[8] Hussien, A.A., Hamza, N. and Hefny, H.A. (2013) ‘Attacks on Anonymization-Based privacy-preserving: A survey for data mining and data publishing’, Journal of Information Security, 4(2), pp. 101–112. doi: 10.4236/jis.2013.42012.

[9] Jain, P., Gyanchandani, M. and Khare, N. (2016) ‘Big data privacy: A technological perspective and review’, Journal of Big Data, 3(1). doi: 10.1186/s40537-016-0059-y.

[10] Li, F., Zou, X., Liu, P. and Chen, J.Y. (2011) ‘New threats to health data privacy’, BMC Bioinformatics, 12(Suppl 12), p. S7. doi: 10.1186/1471-2105-12-s12-s7.

[11] Li, N., Li, T. and Venkatasubramanian, S. (2007) ‘T- closeness: Privacy beyond k-anonymity and l-diversity’, ICDE 2007 IEEE 23rd International Conference on Data Engineering, doi: 10.1109/icde.2007.367856.

[12] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. (2007) ‘L -diversity: privacy beyond k-anonymity’, ACM Transactions on Knowledge Discovery from Data, 1(1). doi: 10.1145/1217299.1217302.

[13] Nabeel, M., Shang, N. and Bertino, E. (2013) ‘Privacy preserving policy-based content sharing in public clouds’, IEEE Transactions on Knowledge and Data Engineering, 25(11), pp. 2602–2614. doi: 10.1109/TKDE.2012.180.

[14] Samarati, P. and Sweeney, L. (2007) Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression. [Online] Available at: https://epic.org/privacy/reidentification/Samarati_Sweeney_ paper.pdf (Accessed: 2 February 2017).

[15] Singh, A.P. and Parihar, D. (2013) ‘A review of privacy preserving data publishing technique’, International Journal of Emerging Research in Management &Technology, 2(6), pp. 32–38.

[16] Sweeney, L. (2002) ‘Achieving k-Anonymity Privacy Protection Using Generalization and Suppression’, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), pp. 571–588. doi: 10.1142/s021848850200165x.

[17] Vinogradov, S. and Pastsyak, A. (2012) Evaluation of data Anonymization tools. [Online] Available at: http://www.epiuse.co.in/brochure/E4236_P5KPL-AM_SE_u.pdf (Accessed: 2 February 2017).