A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data

Manoj Jayabalan
Asia Pacific University of Technology and Innovation
19 PUBLICATIONS **13** CITATIONS

A Study on k-anonymity, l-diversity, and t-closeness Techniques focusing Medical Data

Keerthana Rajendran
School of Computing & Technology
Asia Pacific University of Technology
& Innovation, Technology Park
Malaysia 57000 Bukit Jalil,
Kuala Lumpur, Malaysia

Manoj Jayabalan
School of Computing & Technology
Asia Pacific University of Technology
& Innovation, Technology Park
Malaysia 57000 Bukit Jalil,
Kuala Lumpur, Malaysia

Muhammad Ehsan Rana
School of Computing & Technology
Asia Pacific University of Technology
& Innovation, Technology Park
Malaysia 57000 Bukit Jalil,
Kuala Lumpur, Malaysia

Summary
In today's world, most organizations are facing data accumulation in massive amounts and storing them in large databases. Myriad of them, the particular healthcare industry has recognized the potential use of these data to make informed decisions. Data from the Electronic Health Records (EHRs) system are prone to privacy violations, especially when stored in healthcare medical servers. Privacy Preserving Data Publishing (PPDP) caters means to publish useful information while preserving data privacy by employing assorted anonymization methods. This paper provides a discussion on several anonymity techniques designed for preserving the privacy of microdata. This research aims to highlight three of the prominent anonymization techniques used in medical field, namely k-anonymity, l-diversity, and t-closeness. The benefits and limitations of these techniques are also reviewed.

Key words:
PPDP; data anonymization; k-anonymity; l-diversity; t-closeness

1. Introduction

As the technologies are rapidly expanding, so are the various cyber crimes such as internet phishing where the confidential information is violated and this raises concerns of preserving data privacy and security among users and enterprises globally. Besides, the use of social network sites, electronic healthcare systems, online trading, etc. has generated a large number of datasets that constitutes big data. The analysis performed on patients' data causes privacy and security concerns at stages such as data collection, storage, and processing. Thus, there is a high demand for Privacy Preserving Data Publishing (PPDP) for protected data sharing via the internet. To protect the privacy of the data proprietor, multitude de-identification and anonymization techniques are applied before the data is released 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). This will be followed by a detailed focus on the overview, principle, uses, and limitations of three anonymization techniques which are *k*-anonymity, *l*-diversity and *t*-closeness under Section IV with regards to healthcare microdata. This paper will end with a concluding remark on the mentioned techniques and the notion whether the aim of preserving data privacy is achieved from the research that has been done for PPDP in the recent years in Section V.

2. Privacy Preserving Data Publishing

In short, PPDP sanitizes personal data (e.g. electronic health records) that are highly susceptible to making them available to agencies or public. As depicted in Figure 1 below, the attacker can be anyone (data recipient) who obtain the personal information about an individual. Thus, it is a vital duty of the data publisher to apply various privacy preserving measures to control the information of the released data by modifying it before publication [3] [4].

A myriad of data mining algorithms with high complexity has been introduced as an approach to mitigate breach of data and to avoid the penalties from government agencies. During every stage in data mining, it is crucial to implement techniques that render data privacy and allows the safe exchange of information. Access confinement and distorting data are methods used to protect the sensitive data. At the data depository phase, encryption techniques such as Identity-Based Encryption (IBE) and Attribute-Based Encryption (ABE) are well-known apart to protect while data stored in the cloud vendor or medical server [2] [9]. The PPDP is mainly acquired in the data processing step of the big data analytics which will be the focus of this research. Data anonymization, also known as data masking or data desensitization, is used to obfuscate or conceal any sensitive data about an individual, thus limiting the person's re-identification [7] [17].

Manuscript received December 5, 2017
Manuscript revised December 20, 2017

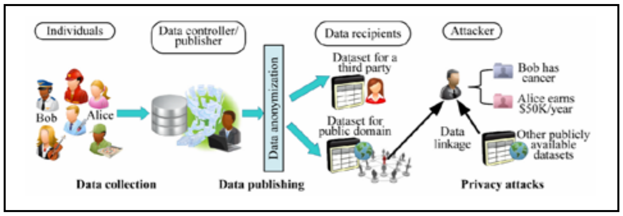
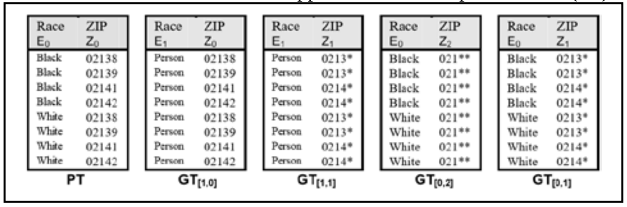


Fig. 1. Outline of Privacy Preserving Data Publishing (PPDP) [4]

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



Race	ZIP	Race	ZIP	Race	ZIP	Race	ZIP	Race	ZIP
E_0	Z_0	E_1	Z_0	E_1	Z_1	E_0	Z_2	E_0	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 anonymization methods like generalization and suppression to safeguard the data by modifying them to conceal the authentic sensitive data. Graph-based techniques such as constrained perturbation are also used to anonymize huge data networks. These methods are further divided into *k*-anonymity, *l*-diversity, classification, clustering, association rule, condensation and cryptographic [2] [6] [9]. Some of these techniques will be explained in detail in this paper centering the healthcare domain. Medical data are known to have high dimensional information obtained from heterogeneous sources and recently these healthcare data are digitalized to reduce expense and enhance quality. This result in the evolution of Electronic Health Records (EHRs) which stores patient-specific information ranging from their demographics to health parameters in a centralized repository and recently cloud-based medical data storage is gaining popularity. Being prone to multiple cyber-attacks for retrieving personal information, EHRs system is compelled to employ anonymization techniques to prevent the medical data from misuse, abuse or patient privacy violation as illustrated in Figure 2 [10].

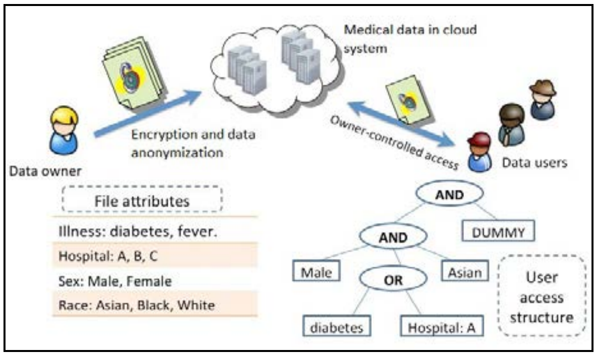


Fig. 2. Anonymization of medical data [13]

3. Literature Review

It is a greater responsibility of the data publishers such as healthcare institutions to share patient information without allowing the adversaries to identify an individual and maintain the privacy using the *k*-anonymity technique. *k*-anonymity is a privacy preserving paradigm to ensure distinct records in a specific dataset cannot be identified. 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 highlights how *k*-anonymity is obtained using generalization and suppression. In generalization, the original value is altered to fill in a general but syntactically constant value whereas, in suppression, the values are represented in asterisk '*'. The power to link a person distinguishing data with others via quasi-identifiers is restricted when *k*-anonymity is applied. Examples of these identifiers are a zip code, gender, birth date, name, address, etc. Various algorithms and theorems are elucidated in this paper. In suppression, the original attribute is generalized by removing some elements but retaining its faithfulness. 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. The 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:

$$r_i < c\ (r_l + r_{l+1} + \ldots + r_m)$$

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

Overall, this paper displays several algorithms in proving the ability of *l*-diversity to acknowledge the imperfections in *k*-anonymity to overcome homogeneity assault and background information. This technique is also realistic and easily comprehensible [12].

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 preserving technique proposed to address the limitations in the existing k-anonymity and *l*-diversity methods. In *l*-diversity, it is assumed that the adversary can attain knowledge on a sensitive attribute if the attribute distribution is known, which is a limitation of this method. In addition, most privacy preserving methods assume that the attributes have definite values, i.e. categorical. 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) that considers the semantic proximity of the feature/attribute values. However, this technique does preserve feature disclosure but identity is still disclosed. So, *k*-anonymity and *t*-closeness can work together to preserve the privacy of published data [11].

In 2008, another issue in preserving the privacy of string data such as genomic and biological data was raised. An alternative of *k*-anonymity called condensation was used on pseudo-data to conceal the actual values of the records without affecting the multi-dimensional statistical data. Here, the anonymization is done by summarizing statistical datasets that will be used to create pseudo-strings. These pseudo-data are similar to the original strings, which are created from the distribution information of the symbols containing the probabilistic measures. 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 shows that classification precision is highly maintained and decreases moderately with bigger group sizes, thus retaining the originality of the statistical data. 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].

Conveying medical information over wireless sensor networks to multiple nodes and data suppliers poses many privacy problems. Simply performing data deidentification is not sufficient to protect the anonymity of personal data from the attacker. For instance, the attacker can link the node signal that transmits the unique ID with medical parameters such as heart rate of a patient if the patient is alone in a hospital room. To avoid this, generalization method under *k*-anonymity can be applied to make the node IDs indistinguishable from each other. The lesser the attribute becomes descriptive, the higher the level of anonymization of the records. In *k*-anonymity, the ground value (specific, original value) is mapped to a generalized value to make it less identifiable. Some of the security concerns raised in this paper with respect to health sensor networks are eavesdropping transmitted data information, data modification in the sensor after receiving the data and traffic monitoring. In 2012, Belsis and Pantziou proposed *k*-anonymity technique using clustering. The collected data from various sensors are organized into clusters which improve energy productivity and reduces disruption between channels, even with large datasets. If the node numbers in a cluster do not meet the pre-set '*k*' threshold value, then false data are inputted to meet the *k*-value before sending the signal to the main station. This results in the adversary from assuming the patient identity as the clusters meet the *k*-value [5].

4. Discussion

4.1 k-Anonymity

Most of the data holders including the government agencies and hospitals misunderstand that the data, e.g. medical records, will remain anonymous if the explicit details such as name, address and phone number are concealed before disclosing the rest of the records. Nevertheless, re-identification of individual by linking the data with other published data, e.g. voter's list, will result in loss of anonymity. Though adding noise to the dataset such as false values and scrambling might provide anonymity, but this will give inaccurate statistical results within tuples when performing data mining and analysis. To address these problems, Samarati and Sweeney formalized a technique called *k*-anonymity in 1998, which use generalization and suppression methods to allow data revelation in a controlled manner while securing the value integrity of tuples. Quasi-identifiers are unique attributes that recognize an individual such as birth date and gender. A table containing these quasi-identifiers is said to meet *k*-anonymity if each tuple value of the quasi-identifiers recurs at least '*k*' times, thus making the tuple distinguishable from one another [14] [16].

4.1.1 Principle of k-Anonymity

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

4.1.2 Generalization

Generalization is a technique used to represent the attribute values in a table to make the identification of tuples less discrete. In this method, the original attribute is represented as a ground domain and the domain value increases with increasing generalization. Quasi-identifiers like zip code are mapped from Z0 (02123, 02126) to Z1 (02120, 02120) to generalize the values and at the same time not lose the truthfulness of the data. This is called domain generalization hierarchy. For private tables with specific values, *k*-minimum generalization is used if the table has already achieved k-anonymity within the table [14] [16]. The limitation of this method is that there will be a need for high level of generalization when there are lesser outliers, i.e. tuples that occur less than *k*-times [14].

4.1.3 Suppression

To complement *k*-anonymity, suppression is used with generalization. Suppression is a technique that is used to mask certain values in the quasi-identifiers [14]. The suppressed value is represented with an asterisk (*) and this can be applied to both domain and value generalization hierarchies. 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].

4.1.4 Pros of k-Anonymity

- It preserves against identity disclosure by inhibiting the links to a dataset with less than '*k*' values. This prevents the adversary from connecting a sensitive data with an external data [8] [15].
- The cost of incurred in establishing this method is considerably lesser compared to the cost of another anonymity method such as cryptographic solution [5].
- Algorithms of *k*-anonymity such as Datafly, Incognito, and Mondrian are used extensively, especially in PPDP. It is also mentioned that clustering is incorporated in *k*-anonymity to enhance privacy preservation [4].

4.1.5 Cons of k-Anonymity

There are many limitations that have been identified in this technique, mainly attacks such as unsorted matching, complementary release, minimality and temporal attacks [8] [9] [16]. Other disadvantages include this technique can cause high utility loss if it is employed in high-dimensional data and exceptional measures are needed if the released data has already undergone anonymization more than once [15]. However, in this research two of the well-known attacks on *k*-anonymity will be briefed below.

- Homogeneity attack: When there is inadequate heterogeneity in the sensitive attributes, this can generate clusters that expose information. Suppose A and B are opponents and A knows that B lives in a particular zip code and is of a particular age, and wants to know B's medical status. 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 adversary has a known knowledge about the individual and with additional logical reasoning, individual's sensitive attributes can be leaked. Consider A and C are acquaintances and A would like to infer C's personal data which is found in the same patient record as B. As A knows that C is a 45-year old Asian female living at a particular zip code. Nevertheless, the record shows that C can have any of the three diseases - cancer, heart disease and viral infection. Based on A's background information that C prevents high-calorie meals and has low blood pressure, A infers that C has heart disease. Hence, *k*-anonymity is prone to background knowledge attack [8] [12].

4.2 l-Diversity

l-diversity was proposed to conquer the limitations of *k*-anonymity. As an extension to k-anonymity, they have introduced a novel method, which can ensure data privacy even without identifying the enemy's background knowledge to avoid attribute disclosure. This approach revolves around the notion that the sensitive attributes in each group are "well-represented". This technique is a modification of k-anonymity by incorporating the k-anonymity principle [11] [12].

4.2.2 Principle of l-Diversity

A *k*-anonymous table is said to be *l*-diverse if each of equivalence class in the table has at least '*l*' "well-represented" values for each sensitive attribute [6] [12]. The term "well-represented" can be elucidated as per the following principles:

Distinct *l*-diversity: A value appears more recurrently than other values within the equivalence class. The downfall in this is that the attacker can infer that this value is likely to represent the entity based on the probability of occurrence.

Entropy *l*-diversity: The entire table must have at least log(l) as entropy to be able to meet entropy *l*-diversity for every equivalence class. This technique may be too prohibitive in the case of low entropy of entire table when only a few values are the same.

Recursive (c, l)-diversity: A table is said to agree to this principle if the sensitive values in each equivalence class do not occur either too frequently or too rarely. This notion is stronger than the previous two notions mentioned above [11] [12].

4.2.3 Pros of l-Diversity

- Provides a greater distribution of sensitive attributes within the group, thus increasing data protection.
- Protects against attribute disclosure, an enhancement of *k*-anonymity technique.
- The performance of *l*-diversity is slightly better than *k*-anonymity due to faster pruning by the *l*-diversity algorithm [6] [11] [12].

4.2.4 Cons of l-Diversity

- *l*-diversity can be redundant and laborious to achieve.
- Prone to attacks such as skewness attack and similarity attack as it is inadequate to avoid attribute exposure due to the semantic relationship between the sensitive attributes [9] [11].

4.3 t-Closeness

A betterment of *l*-diversity is a *t*-closeness technique by decreasing the granularity of the interpreted data. The observer's extent of knowledge on a specific data is limited while the knowledge is not limited to the overall table containing the datasets. Therefore, this reduces the correlation between the quasi-identifier attributes and the sensitive attributes. The distance between the distributions is measured using Earth Mover's Distance (EMD). For a categorical attribute, EMD is used to measure the distance between the values in it according to the minimum level of generalization of these values in the domain hierarchy [11].

4.3.1 Principle of t-Closeness

t-closeness of an equivalence class is attained when the sensitive attribute distance in this class is not greater than the threshold, *t* with the attribute distance in the whole table. The table is acknowledged to have t-closeness if all equivalence classes have t-closeness [11].

4.3.2 Pros of t-Closeness

- It interrupts attribute disclosure that protects data privacy.
- Protects against homogeneity and background knowledge attacks mentioned in *k*-anonymity.
- It identifies the semantic closeness of attributes, a limitation of *l*-diversity.

4.3.3 Cons of t-Closeness

- Using Earth Mover’s Distance (EMD) measure in *t*-closeness, it is hard to identify the closeness between *t*-value and the knowledge gained.
- Necessitates that sensitive attribute spread in the equivalence class to be close to that in the overall table [9] [11].

5. Conclusion

Preservation of data privacy has transpired as a definite prerequisite in privacy preserving data publishing. The increase in cyber crimes has caused severe risk of privacy breach. This has prompted the manifestation of various anonymization techniques. This paper has discussed on these rising concerns in PPDP, converging into the healthcare domain, which poses greater chances of disclosure of personal and sensitive data. To circumvent this, a range of anonymization methods applied on medical data were summarized here based on the academic literature dedicated to PPDP. Furthermore, the scope of this research is limited to the *k*-anonymity technique with its extended modifications, which are *l*-diversity and *t*-closeness. Each of these techniques was illuminated in detail with principles and related references. The comparison of the advantages and disadvantages of these three methods were also rationalized. Overall, this research is committed to providing a brief on the existing trends of anonymization techniques orientating medical data in achieving privacy preservation 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) ‘MET𝔸P: 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).