https://doi.org/10.52326/jes.utm.2022.29(3).05
UDC 004.056(669)

APPLICATION OF PRIVACY-PRESERVING DATA PUBLISHING IN TERTIARY INSTITUTIONS OF KEBBI STATE USING GENERALIZATION AND SUPPRESSION

Anas Shehu[1*], ORCID: 0000-0002-5307-1457,
Alhassan Salihu[1], ORCID: 0000-0003-1926-6813,
Abubakar Sani[2], ORCID: 0000-0001-9616-0647

[1]*Department of Computer Science Kebbi State Polytechnic Dakin-gari, Dakin-gari, 862106, Nigeria.*
[2]*Department of Computer Science Yusuf Maitama Sule University, Kano, 700214, Nigeria.*
*Corresponding author: Anas Shehu, *anasshehu8@gmail.com*

Received: 06. 14. 2022
Accepted: 08. 02. 2022

Abstract. The research was conducted in the field of publishing data to preserve confidentiality. Several educational datasets have been used to address privacy and utility. The sample questionnaires served to investigate the level of privacy awareness and enforcement in the records of students in tertiary institutions in Kebbi State, Nigeria. The benchmark datasets were obtained from Kebbi State Polytechnic Dakin-gari. K-anonymity and l-diversity models were used with *k* configurations and suppression limits of 10 and 50% in the ARX 3.9.0 de-anonymization environment. The work evaluates data privacy, quality, and execution time for each k value and suppressions limit. Experimental results demonstrate that the higher the suppression the more balanced exists between privacy and utility. It was observed that suppression of 50% provides less anonymization time irrespective of *k* compared to *k* values in suppression = 10%. This was proved to be due to less time it takes anonymization to be completed Also, from respondents, 92% of students' records were kept permanently in plain and, issued to third parties like that—with no privacy guarantee. This poses privacy threats to datasets.

Keywords: *Arx de-anonymization tool, Dakin-gari, k-anonymity, privacy, quality, utility.*

Rezumat. Cercetarea a fost efectuată în domeniul publicării datelor pentru păstrarea confidențialității. Au fost folosite câteva seturi de date educaționale pentru a aborda confidențialitatea și utilitatea. Chestionarele eșantion au servit pentru a investiga gradul de conștientizare a confidențialității și aplicarea acesteia în dosarele studenților din instituțiile terțiare din statul Kebbi, Nigeria. Seturile de date care au servit drept reper au fost obținute de la Kebbi State Polytechnic Dakin-gari. Modelele de K-anonimitate și l-diversitate au fost utilizate cu configurații *k* și limite de suprimare de 10 și 50% în mediul de de-anonimizare ARX 3.9.0. Lucrarea evaluează confidențialitatea datelor, calitatea și timpul de execuție pentru fiecare valoare k și limită de suprimare. Rezultatele experimentale demonstrează, că o suprimare este mai mare induce echilibru între intimitate și utilitate. S-a observat, că suprimarea de 50% oferă mai puțin timp de anonimizare indiferent de k comparativ cu valorile *k* în suprimare = 10%. Acest lucru se datorează faptului că anonimizarea durează mai puțin pentru a fi finalizată. De asemenea, din partea respondenților, 92% din dosarele studenților au fost păstrate în mod permanent neconfidențial, fiind eliberate astfel unor terți, fără garanție de confidențialitate. Acest lucru reprezintă amenințări de confidențialitate a seturilor de date.

Cuvinte cheie: *instrument de dezanonimizare, Arx, Dakin-gari, k-anonimitate, confidențialitate, calitate, utilitate.*

1. Introduction

In Computer Science & Information Technology, privacy could be seen as control over the disclosure of Personally Identifiable Information (PII), or quasi-identifiers (QI). This PII or QI helps in establishing a user profile when combined with a publically available dataset that leads to personality being watched, profiled, and make unwanted revelations that resulted in physical and economic harm. Privacy ought to be guaranteed when sensitive biomedical data is shared for any reason [1], though the most common datasets use are biomedical and demographic data [2]. Notwithstanding, that did not limit other datasets to be used as individuals and industries carry out research from multiple and disparate domains day in and day out where attributes of individual should be protected using industry acceptable techniques. With the current growth of information technologies, various organizations such as hospitals, financial houses, educational institutions are constantly collecting information about individuals and keep it in their databases for future use. These volumes of data increase exponentially [3] as a result of this, privacy becomes the subject of hot debate as it requires models, privacy risks for protecting it as well as providing utility [2]. On this note, this work intends to explore the available resources to apply privacy to student datasets before sharing them with researchers. To protect privacy, recommended data transformation models should be used in the process. Examples of such models are Global recoding, full-*domain generalization* Plus *record suppression* [1], user defines hierarchy is always useful for generalization as it dictates transformation rules that minimize attributes precision in a hierarchical pattern. While full domain generalization makes an attribute generalized on an equal level of associated hierarchy. Refer to figure 1 for generalization hierarchy level 0 of gender and LGA are more specific compared to level 1 and of course level 2 presents the B/Kebbi highest level that cannot be recognized.

As for suppression, the original attribute value is replaced by a symbol such as '*', '#' and so on to detach meaning from it [3].

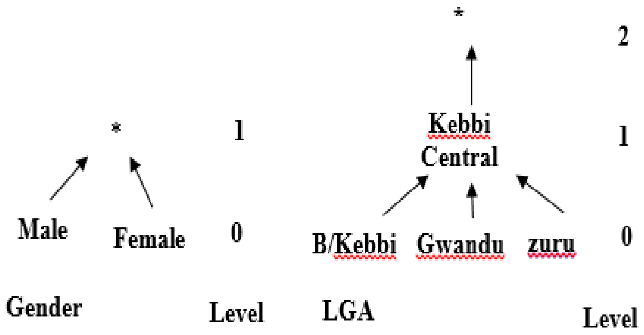

Figure 1. Generalization hierarchy adopted from [4].

1.1. Privacy Models

Privacy models were developed aimed at mitigating the risk of linkage attacks taking quasi identifier (QI) as a target [2], that QI cannot be eliminated from the dataset as they are important and needed for analyses. We formally defined QI as *attributes A_1...... A_d* in table *T* that can be joined with external public data to re-identify individual records such as student matric no., application no., gender, zip code, date of birth, age, etc. K-anonymity is a commonplace model used in preserving QI privacy. For more detail about k-anonymity, refer to [2].

Also, attributes are sensitive if an individual may not want to be linked with it, for example in our case, student registration fee, student department, occupation, salary, and disease in the biomedical domain. To protect sensitive attributes, *l-diversity,* and *t-closeness* as prevalent models are being utilized [5]. Figure 2 below shows the taxonomy of the privacy model.

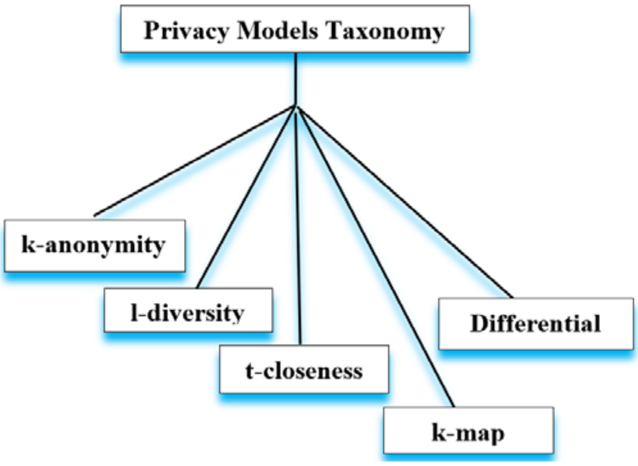

Figure 2. Taxonomy of Privacy Models.

1.2. Contributions

In summary, this work will present the following contributions: 1. Presentation of questionnaire and its response 2. Presentation and analyses of survey results concerning privacy in the educational domain. 3. Application of Student datasets in the field of PPDP using the ARX tool. 4. Extensive evaluation of the anonymized dataset concerning different values and two different suppression limits. And finally, we present the experimental setup used in the work.

1.3. Survey Results

The questionnaire was designed for the survey work to ascertain the level of data privacy, information awareness, and its application in all tertiary institutions in Kebbi State. Each institution was administered 30 questionnaires with the targeted respondents of Level Coordinators, Management and information Units, Bursary and Registry among others, and below are the name of the institutions:

1. Waziri Umaru Federal Polytechnic Birnin Kebbi **(WUFP)**,
2. Kebbi State Polytechnic Dakingari **(K/S Pol. Dakingari),**
3. Collage of Education Argungu **(COE Argungu)**,
4. Health Technology Jega **(Health Tech.Jega)**,
5. Aleiro University of Science and Technology **(AUST)**,
6. Federal University Birnin Kebbi **(FUBK)**.

The table below provides samples of questionnaires administered and the associated responses per each institution.

Table 1

Sample of questionnaire responses

Question	WUFP l	K/S Pol. Dakingari	COE. Argungu	Health Tech. Jega	AUST	FUBK
What is the total number of students in your institution?	Above 5000-(66.66%)	1000-2000-(66.66%)	4000-5000-(46.66%)	Above 5000-(63.66%)	Above 5000-(60.66%)	Above 5000-(60.66%)
Does your institution keep student records?	Yes – (80%)	Yes – (100%)	Yes – (100%)	Yes – (73.33%)	Yes – (100%)	Yes – (100%)
How long does your institution keep the student's record?	Forever-(68.96%)	Forever-(73.33%)	Forever-(90%)	Forever-(20%)	Forever-(100%)	Forever-(100%)
Which of the student's details do you consider sensitive?	Account No.-(36.66%)	Account No.-(93.33%)	Account No.-(50%)	Account No.-(50%)	Account No.-(20%)	Account No.-(20%)
Does your institution use a computing platform in keeping student records?	Yes.-(100%)	Yes.-(86.66%)	Yes.-(86.66%)	Yes.-(86.66%)	Yes.-(93.33%)	Yes.-(100%)
Does the student's record keep in plain text?	Yes.-(56.66%)	Yes.-(60%)	Yes.-(56.66%)	Yes.-(76.66%)	Yes.-(80%)	Yes.-(80%)
How simple it is to identify individual records?	Very Simple.-(63.33%)	Very Simple.-(66.66%)	Very Simple.-(53.33%)	Very Simple.-(66.66%)	Very Simple.-(96.66%)	Very Simple.-(96.66%)
Does your institution give out student data to a third party?	Yes.-(23.33%)	Yes.-(66.66%)	Yes.-(56.66%)	Yes.-(56.66%)	Yes.-(26.66%)	Yes.-(26.66%)
Are you aware of information privacy and data protection law?	Yes.-(46.66%)	Yes.-(86.66%)	Yes.-(36.66%)	Yes.-(53.33%)	Yes.-(80%)	Yes.-(80%)
Does the institution prevent students' data from any attack?	No.-(100%)	No.-(100%)	No.-(100%)	No.-(100%)	No.-(100%)	No.-(100%)

From Table 1 above, the results provided are consolidated for the whole six (6) higher institutions of learning in the state. We chose to use a one-sided response as it is the majority and provides insight into what the research wants. The most interesting things to note from the table are: (1). 92.22% of the student records were collected for the entire institution only, and 75.38% were kept for eternity. Though 92% of the record were stored in computing platforms used by various institutions, 68.33% of the total record were kept as plain text-(as is collected). This shows the extent of privacy threats faced by the record. (3). Also as indicated in the table above, 73.88% of all student records in the entire school are prone to internal attack due to the simplicity of identifying individual records with less effort. For these, we can attest to the fact that the entire records for the whole institutions of Kebbi state are being faced with privacy threats as figures shown in table 1 due to the absence of any privacy protection techniques applied to the information. Even though respondents claimed to be aware of information privacy and data protection law.

2. Methodology

In this section, we will present the methodology used in the conduct of this work such as the experiment framework, the dataset used, the Experiment setup, the toolbox used, and the results in discussion. Figure 3 below is the entire work activity diagram.

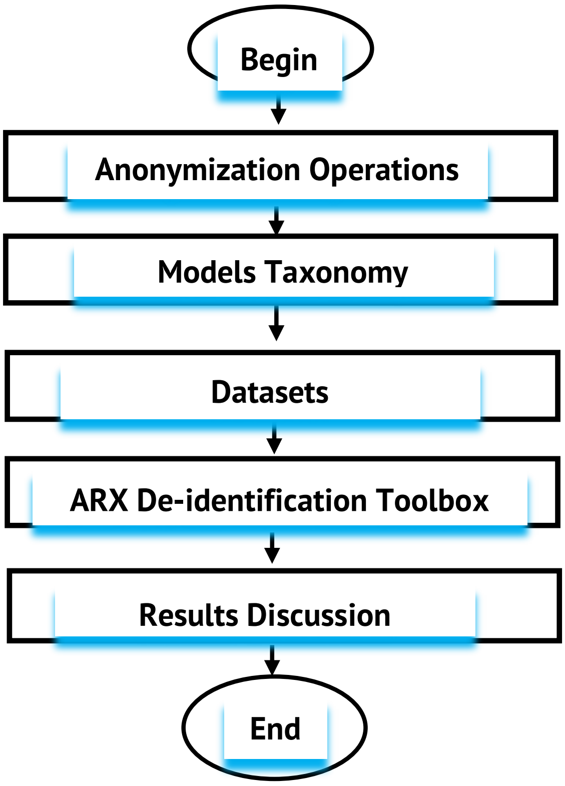

Figure 3. Activity Diagram.

Anonymization operations and taxonomy tree has been explained in the previous section above.

2.1. Experiment Framework

Figure 4 below shows the framework for experimenting. The first process involved in the framework is *New Project* where a user must provide the name of the newly created project before the ARX environment becomes enabled. *Importing data* is a process also where ARX user brings in .csv datasets for the anonymization process and will only be enabled if a project is created. The *configuration* enables the user to create and edit rules, define privacy guarantees, parameterize the coding model and configure utility measures. While *anonymizing* is a process of performing data transformation. Filtering, analyzing the solution space, and organizing transformations are done through *Explore results*. The user keeps doing this process until the anonymized data suits his needs. If the final results are acceptable then, *Analyze Results* process is used where the main analysis takes place to compare and analyze the input and output such as attribute analysis, equivalence class analysis, performing local recoding, and final results summary. Lastly, the final results are stored for further use and analysis.

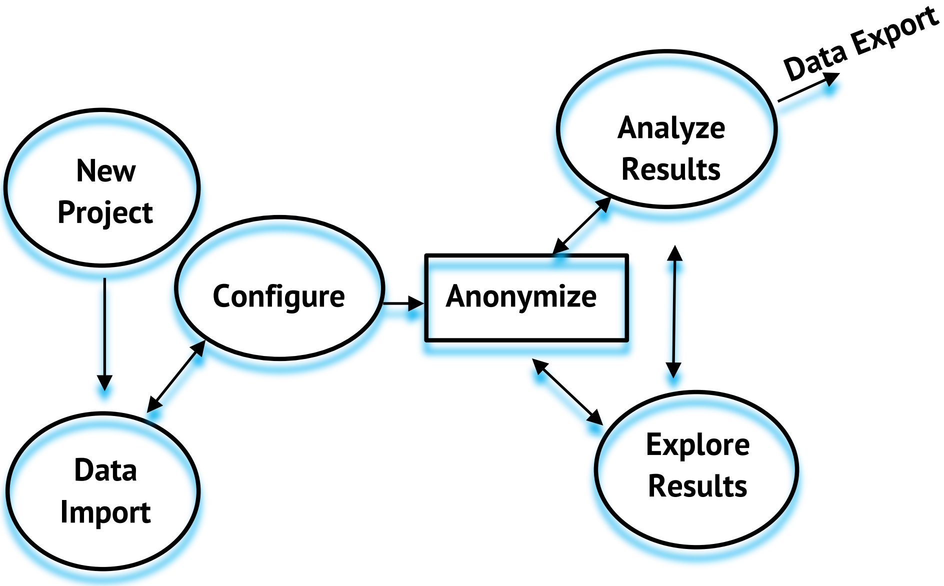

Figure 4. Experiment Framework.

2.2. Dataset

To the best of our knowledge, no dataset benchmark is set for the educational domain and since the ARX anonymization toolbox works with any dataset we chose to use a dataset collected from one of the institutes of higher learning in Kebbi State, Nigeria, known as Polytechnic Dakingari. Initially, the dataset contained 260 records which after data cleanup became 180 records only. Tables 2 and 3 show the overview of the datasets.

Table 2

Overview of the datasets

Dataset	Quasi-Identifier	Records	Highest Transformation	Size (KB)
Student	8	180	1,223,040	2

Table 3

Overview of the attributes in the datasets

Dataset	Quasi-Identifier (height of Hierarchy)	SA (Distinct Values)
Student	Sex (2), Matric Number (13), invoice (11), Application number (9), State (7), local govt. (12), session (1), status (1)	Department (23)

2.3. Experimental Settings

In this work, the experiments were conducted on a laptop computer running 64-bit Windows 8 (6.2, Build 9200) with AMD E-300 APU with Radeon (TM) processor at 1.3GHZ clock speed with 4 GB RAM. As for the five models, this work uses the ARX anonymization toolbox, to be explained next. Moreover, all the five models and the metrics are implemented in the toolbox. The research did not perform any pre-computation in the toolbox that can give an advantage to some models over others.

2.4. Parameter Value

Parameter values of k used in the experiment were recommended as the best configurations in[5]. As for parameter *L* values also cannot exceed the distinct values of SA for a good result, refer to [5] and [2], thus, this research takes care of that. Our work use recursive (c, l) diversity, where *c* stands to be constant and *l* "well represented" sensitive value. Table 4 below summarizes the configurations used in the experiments carried out.

Table 4

Experimental Configurations

Experiment	Parameter Settings	Datasets (Size)

Varied Parameter values
Suppression limit = 10%
	[k-value = 3, c =4, l=3
	k-value = 5, c =4, l=3
	k-value =7, c =4, l=3
	k-value = 9, c =4, l=3
	k-value = 11, c =4, l=3]
		Student (180)

Varied Parameter values
Suppression limit =50%
	[k-value = 3, c =4, l=3
	k-value = 5, c =4, l=3
	k-value =7, c =4, l=3
	k-value = 9, c =4, l=3
	k-value = 11, c =4, l=3]	

2.5. ARX Anonymization Toolbox

ARX - Powerful Anonymization Toolbox is a comprehensive open-source software for anonymizing sensitive personal data. It supports full-domain generalization, record suppression, local recoding, and microaggregation [6]. It was developed within three years by five computer scientists in Germany, refer [7]. For ARX graphical interface refer to [2].

3. Results and Discussion

In this section, the results obtained during the experiments using the configuration and student dataset above are going to be analyzed and explained about certain quality metrics such as Granularity, Non-uniform entropy, and Discernibility. Also, some transformations and anonymization time per run will be presented. The best score is the one with the lowest score [8].

Granularity. This model collects and presents the granularity of the output dataset. From the first set of four bars in figure 5 we can see how this model displays two different sets of results as the suppression limit is 10%. As *k* = 3 and 5, almost 90% of the output dataset cannot be identified due to a high level of anonymization. This indicates that when this data is shared for research purposes, it will provide little utility and hardly achieve research purposes due to high privacy. Similarly, as the *k* value increased from 3 to 5, the same results were obtained with no effect. On the other hand, when the *k* value moved to 7, 9, and 11, we can observe the slightest increase from 91.11% to 95% all through. This no doubt affects the attribute quality more and made it unworkable by researchers, though privacy became higher than 3 and 5. But, the effect of the *k* value became constant as observed.

In figure 6 below displays results as suppression limit = 50%, indicating attributes level details are clearer than when suppression was 10%. All the returned results indicate 61% down. That proves that privacy and quality were balanced.

Non-uniform entropy. This model measures information loss based on common information in a dataset that measures the amount of information that can be obtained about the original values of variables in the input dataset by observing the values of variables in the output dataset. However, the metric makes this quantification for an individual attribute in the dataset. In the second four bars of figure 5 below, as suppression limit = 10%, we can also see that as the *k* value keep increasing from 3 through 7, information loss for the datasets keeps decreasing, though, with different values of 16.58%, 8.88% and 6.03 % respectively. However, 6.03% remains constant from *k* = 7 through 11. Meaning that the datasets cannot be de-anonymized more than *k* = 7 and, these values provides minimum loss.

On the other hand, figure 6 presents results as a suppression limit = 50%. It is evident that as *k* = 3 and 5, distortion was not much compared to the same values as suppression = 10%. When *k* = 9, loss of information is almost the same as its counterpart in 10% above. On the other hand, in the 10% limit, *k* = 7 and 11 outperformed their counterparts in the 50% limit.

Discernibility. This measures how identical a record is to others within each equivalence class by assigning an additional penalty to it equal to the size of the equivalence class it belongs. For detail refer to [9]. As indicated in the third group of bars in figure 5 as suppression = 10%, the best scores are when *k* = 7, 9, and 11 which showed the highest identicality of the records in the output dataset. And that indicates higher privacy than quality. But in figure 6 where suppression = 50%, we can also observe the third group of bars with different scores all less than in figure 5. This indicates not much additional penalty as there are fewer equivalence classes.

Anonymization Time. This quantifies the time taken to complete transformation per run and, it measures in seconds. From the last group of four bars in figure 5, we can observe that as the suppression limit is 10%, the last time was when *k=9*, followed by *k=5*. That should not be unconnected with search space until a global optimum solution was returned. And in these two values, the time to return was small. We can also see that as *k=11*, anonymization time was the longest, because of the time taken to return the global optimum.

In figure 6, when the suppression limit was 50 % we can deduce that *k=3* returned the least anonymization time compared to its counterpart in suppression limit 0f 10 %. This happened because the privacy has been relaxed the more and returning to global optimum will not take much time. Also, the rest of the *k* values here outperformed their counterparts above with the increase of values even though they maintain consistent values among themselves. That could be understood that as the suppression limit is relaxed to 50, the increase of *k* values has little or no effect on anonymization time unlike when suppression is tight to 10% which showed different timing due to stricter privacy and suppression.

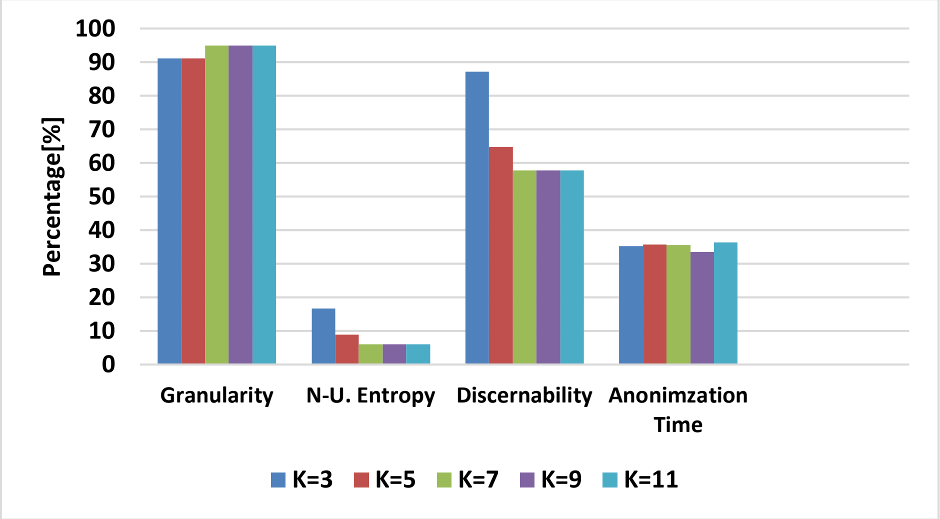

Figure 5. Suppression limit of 10%.

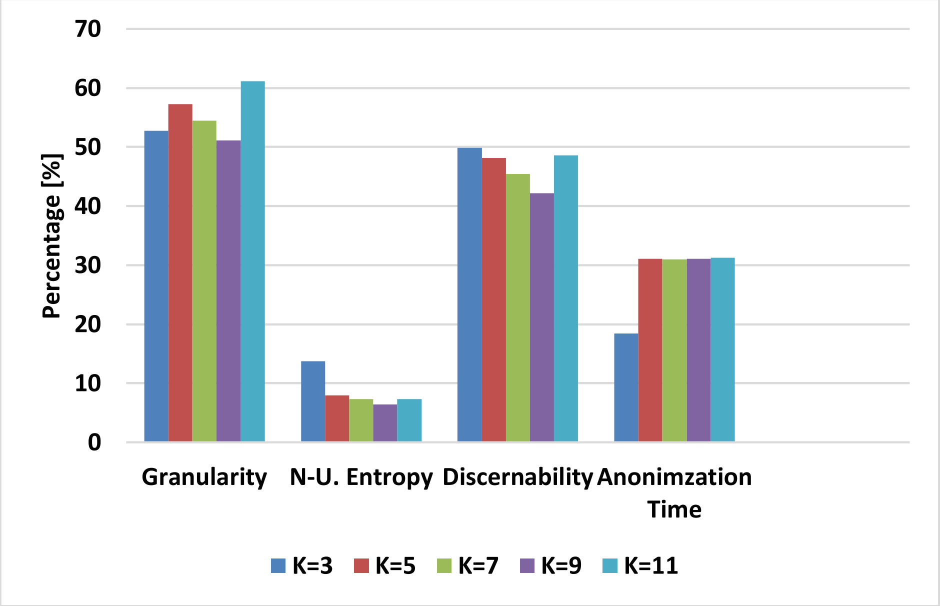

Figure 6. Suppression limit of 50%.

In [10] three privacy models were compared based on information loss metrics. The experiment was conducted using three datasets of which the largest among them contains 16, 422 tuples. In their work, it was concluded that t-closeness has better utility compared to k-anonymity and l-diversity. Yadav compared only two models and the dataset was unknown [11]. Execution time was measured, and it was concluded that *k*-anonymity outperformed l-diversity. In the work of [8], five privacy models were compared out of which one of them-slicing is the anonymization technique and not the privacy model [12]. Furthermore, only one benchmark dataset was used in the work but with a larger size (640,000 records). It was reported that k-anonymity outperformed the rest in terms of execution time. On the other hand, slicing was the worst performer. Prasser et al. present a comprehensive theoretical review of the three most prominent privacy models in big data. The advantages and limitations of these models were stated therein. Though, their proposed solutions can only work where there is only one sensitive attribute in the dataset.

In a model proposed by [13] that data utility can be increased and maintain significant privacy based on the outlier equivalence class. *K*-anonymity and l-diversity were used but, with the single configuration of 5 and 2 respectively. In their work, two datasets were used with a suppression limit of 100%. However, their work was conducted using ARX 3.5.1 environment. Also [14], proposed a model based on superclass substitution for utility improvement on k-anonymity. Their model proved better quality than the other two. Furthermore, a student admission dataset was used. In a similar research effort by [15], four privacy models were used made in a single framework-ARX. The beauty of this work is that various parameter values were used to ascertain the correctness and validity of the result. Though the metric used during the analysis was also four, the dataset is non-educational, and the factor of study is information loss as parameter values changes. The authors in [16] used adult dataset from UCI machine learning repository which was partitioned into five groups from 40000 to 640000 records. On each set of group, five different privacy models were run against execution time and data utility. Though from their work non-of the model outperformed others from all angles.

Based on this literature, we can confirm that none of the work mentioned above has categorically used a dataset from the educational domain, and none used the quality model of *Granularity, Non-uniform Entropy,* and *Discernibility*. Also, none of them used this set of configurations in the ARX environment based on suppression limits of 10 and 50% respectively.

4. Conclusions

In this research, it could be concluded that the higher the suppression limit the more balance exists between privacy and utility. Also, it was observed that the suppression limit of 50% provides less anonymization time in respective *k* values compared to *k* values in suppression = 10%. This was proved to be due to less time it takes anonymization to search and return a globally optimum solution. Conclusively, we can say that the suppression limit of 10% does not provide a balance between privacy and quality. However, the work observed that 92% of the students, records faced privacy threats as there was no privacy policy implemented on data at rest or during sharing with a third party in all Kebbi State educational institutions. Additionally, none of the respondents has a clear view of what privacy is all about. We note also that all the respondents misunderstood data privacy with data confidentiality. Therefore, there is a need for stakeholders in all the institutions to educate data holders about privacy and privacy-enhancing technologies.

Conflict of interest. The authors declare no conflict of interest.

References

1. Prasser, F.; Kohlmayer, F.; Kuhn, K.A. Efficient and effective pruning strategies for health data de-identification. *BMC Medical Informatics and Decision Making* 2016, pp. 1 - 14.
2. Kohlmayer, F.; Prasser, F.; Kuhn, K.A.; Spengler, B.E. Lightning: Utility-Driven Anonymization of High-Dimensional Data. *Transactions on data privacy* 2016, 9, pp. 161 - 185.
3. Narula, D.; Kumar, P.; Upadhayaya, S. Performance Evaluation of K-Anonymization Algorithms for Generalized Information Loss. *IJCTA* 2016, pp. 227 - 235.
4. Shehu, A.; Salihu, A.; Sani, A. Application of Privacy-Preserving Data Publishing on Students' Data in Tertiary Institutions of Kebbi State Using K-Anonymity. *International Journal of Innovative Science and Research Technology* 2022, 7, pp. 1083 - 1091.
5. Kohlmayer, F.; Prasser, F.; Kuhn, K.A. The cost of quality : Implementing Generalization and Suppression for Anonymizing Biomedical Data With Minimal Information Loss. *Journal of Biomedical Informatics* 2015, 58, pp. 37 - 48.

6. Prasser, F.; Kohlmayer, F.; Kuhn, K. A Benchmark of Globally-Optimal Anonymization Methods for Biomedical Data. In Proceedings of the IEEE 27th International Symposium on Computer-Based Medical Systems, New York, USA, May., 2014, pp. 66 - 71.
7. Kohlmayer, F.; Prasser, F.; Kuhn, K.A.; Spengler, E.B. A Tool for Optimizing De-Identified Health Data for Use in Statistical Classification. In Proceedings of the IEEE 30th International Symposium on Computer-Based Medical Systems, Thessaloniki, Greece, June, 2017, pp. 169 - 174.
8. Kohlmayer, F.; Prasser, F.; Kuhn, K.A. The Importance of Context: Risk-based De-identification of Biomedical Data. *Methods of Information in Medicine* 2016, 55, pp. 347 - 355.
9. Jain, P.; Gyanchandani, M.; Khare, N. Big Data Privacy: A Technological Perspective and Review. *Journal of Big Data* 2016, 3, pp. 1 - 25.
10. Emam, A. Data Privacy on E-health Care Yystem. *International Journal of Engineering, Business and Enterprise Application* 2013, 3, pp. 89 - 99.
11. Yadav, D. Secure Techniques of Data Anonymization for Privacy Preservation. *International Journal of Advanced research in Computer Science* 2017, 8, pp. 1693 - 1695.
12. Kohlmayer, F.; Prasser, F.; Kuhn, K.A. A Flexible Approach to Distributed Data Anonymization. *journal of Biomedical informatics* 2014, 50, pp. 62 - 76.
13. Yilmaz, V.; Murat, A. A New Approach to Utility-Based Privacy Preserving in Data Publishing. In Proceedings of the 2017 IEEE International Conference on Computer and Information Technology (CIT), 2017, pp. 204 - 209.
14. Ravindra, T.; Maheshwari, P.; Binod, K. A Comparative Study of Data Suppression Technique for Privacy of Individuals. *International Journal of Current Trends in Engineering & Technology* 2018, 4, pp. 230 - 241.
15. Moein, A.M.; Taha, S.R.; Noman, M.; Hadi, H. The risk-utility tradeoff for data privacy models. In Proceedings of the 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 2016, pp. 1 - 5
16. Antony, P.J.; Selvadoss, T.A. Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data. *Advances in Computational Sciences and Technology* 2017, 10, pp. 973 - 984.

Citation: Shehu, A.; Salihu, A.; Sani, A. Application of privacy-preserving data publishing in tertiary institutions of Kebbi state using generalization and suppression. *Journal of Engineering Science* 2022, 29 (3), pp. 59-69. https://doi.org/10.52326/jes.utm.2022.29(3).05.

Publisher's Note: JES stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submission of manuscripts: jes@meridian.utm.md