ISSN No:-2456-2165

Application of Privacy-Preserving Data Publishing on Students' Data in Tertiary Institutions of Kebbi State Using K-Anonymity

Anas Shehu, Alhassan Salihu
Department of Computer Science Kebbi State
Polytechnic Dakin-gari, Dakin-gari, 862106, Nigeria

Abubakar Sani
Department of Computer Science Yusuf Maitama
Sule University, Kano, 700214, Nigeria

Abstract. The research was conducted in privacy-preserving data publishing, to our knowledge only a few used educational datasets to address privacy and utility. This research used sample questionnaires to investigate the awareness of privacy and its application to student records and also applying privacy to students' datasets of all tertiary institutions of Kebbi State, Nigeria. Student datasets were obtained from Kebbi State Polytechnic Dakin-gari which we used as a benchmark. 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 two variants of 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 when suppression = 10%. This was proved to be due to less time it takes anonymization to be completed. Also, our work ranks six institutions from 1st through 6th based on some parameters obtained via questionnaire/responses on privacy threats. The work however established that all students' records are faced with serious privacy threats as no institution employ any privacy-preserving techniques. Consequently, the research proposed a privacy framework for all six schools to deploy for better preservation.

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

I. 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 individuals should be protected using industry-acceptable techniques. With the current growth of information technologies, various organizations such as hospitals, financial houses, and educational institutions are constantly collecting information about individuals and keeping 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 isalways 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/Kebbihighest level that cannot be recognized.

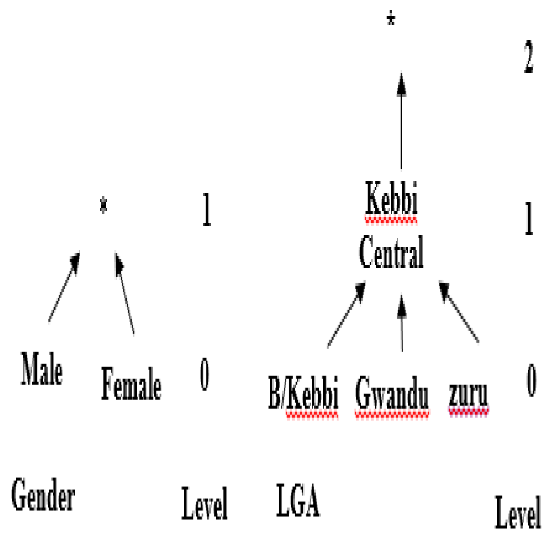

Fig. 1: Generalization hierarchy

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

ISSN No:-2456-2165

II. 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* $A_1 \ldots\ldots A_d$ 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 [4]. Figure 2 below shows the taxonomy of the privacy model.

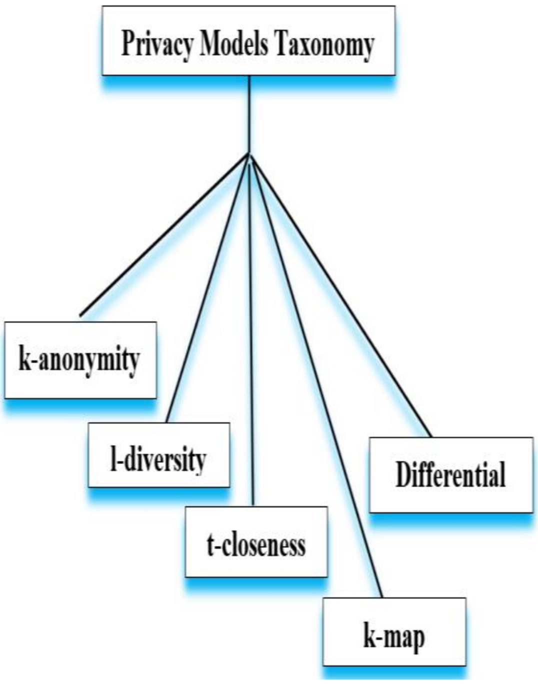

Fig. 2: Taxonomy of Privacy Models

III. CONTRIBUTIONS

This work presents contributions as follows: 1. Presentation and analyses of survey results on information privacy and data protection in the educational domain. 2. Application of Student datasets in the field of PPDP using the ARX tool and complex configurations. 3. Extensive evaluation of the anonymized dataset with disparate k values and different suppression limits. The work rank six institutions based on the privacy threats they possess. And finally, we proposed a privacy framework that will serve as a working guide for these institutions.

IV. SURVEY RESULTS

This work was designed to ascertain the level of data privacy awareness and its application in all tertiary institutions in Kebbi State, The designed questionnaire was administered to all six (6) institutions, regarding the protection of students' data privacy and the associated risk. Each institution was administered 30 questionnaires, and below are the name of the institutions:

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

The table 1 and 2 below provide samples of questionnaires administered and the associated responses per each institution.

Sample Of Questionnaire Responses			
Question	WUFP l	K/S Pol. Dakin-gari	COE. Argungu
What is the total number of students in your institution?	Above 5000-(66.67%) 4000-5000 (26.67%) 3000-4000 (6.66%)	Above 5000- (6.67%) 4000-5000 (13.33%) 2000-3000- (6.67%) 1000-2000- (73.33%)	Above 5000- (20%) 4000-5000- (46.67%) 3000-4000- (23.33%) 2000-3000- (10%)
Does your institution keep student records?	Yes (80%) No (13.33%) Not Sure (6.67%)	Yes (100%)	Yes (100%)
How long does your institution keep the student's record?	Forever-(68.96%) 4-Acd. Sess. (24.14%) 2-Acd. Sess. (6.90%)	Forever -(73.33%) 4-Acd. Sess. (20%) 2-Acd. Sess. (6.67%)	Forever- (90%) 4-Acd. Sess. (10%)
Which of the student's details do you consider sensitive?	Account No.-(36.67%) Account Name-(23.33%) Bank Name-(6.67%) Not Sure.-(33.33%)	Account No.-(93.33%) Account Name- (6.67%)	Account No.-(50%) Account Name-(33.33%) Bank Name-(10%) Not Sure.-(6.67%)
Does your institution use a computing platform in keeping student records?	Yes.- (100%)	Yes.- (86.67%) Not Sure.- (13.33%)	Yes.- (86.67%) Not Sure.- (13.33%)
Does the student's record kept in plain text?	Yes.- (56.67%) No.- (6.67%) Not Sure.- (36.66%)	Yes.- (60%) No.- (13.33%) Not Sure.- (26.67%)	Yes.- (56.67%) No.- (10%) Not Sure.- (33.33%)
How simple it is to identify individual records?	Very Simple.-(63.33%) Simple.-(36.67%)	Very Simple.-(66.67%) Simple.- (20%) Less Simple.-(6.67%) Nil.- (6.66%)	Very Simple.-(53.33%) Simple.- (46.67%)
Does your institution gives out student data to a third party	Yes.- (23.33%) No.- (20%) Not Sure.- (56.67%)	Yes.- (66.66%) No.- (6.67%) Not Sure.- (26.67%)	Yes.- (56.66%) Not Sure.- (43.34%)
Are you aware of information privacy and data protection law?	Yes.- (46.67%) No.- (33.33%) Not Sure.- (20%)	Yess.- (86.66%) No.- (6.67%) Not Sure.- (6.67%)	Yes.- (36.66%) No.- (63.34%)
Does the institution prevent students' data from any attack?	Yes.- (56.67%) No.- (43.33%)	Yes.- (80%) No.- (20%)	Yes.- (90%) No.- (10%)

Table 1: Sample of questionnaire responses of first three Institutions

Sample Of Questionnaire Responses			
Question	Health Tech. Jega	AUST	FUBK
What is the total number of students in your institution?	Above 5000- (63.33%) 4000-5000- (20%) 1000-2000- (16.67%)	Above 5000- (60%) 3000-4000- (30%) 1000-2000- (10%)	Above 5000- (60%) 3000-4000- (30%) 1000-2000- (10%)
Does your institution keep student records?	Yes (73.33%) Not Sure (26.67%)	Yes (100%)	Yes (100%)
How long does your institution keep the student's record?	Forever- (20%) 4-Acd. Sess. (13.33%) 3-Acd. Sess. (20%) 1-Acd. Sess. (46.67%)	Forever-(100%)	Forever-(100%)
Which of the student's details do you consider sensitive?	Account No.-(50%) Account Name-(33.33%) Bank Name-(10%) Not Sure.-(56.67%)	Account No.-(20%) Account Name-(23.33%) Not Sure.-(33.33%)	Account No.-(20%) Account Name- (23.33%) Not Sure.- (56.67%)
Does your institution use a computing platform in keeping student records?	Yes.- (80%) Not Sure.- (20%)	Yes.- (93.33%) Not Sure.- (6.67%)	Yes.- (100%)
Does the student's record keep in plain text?	Yes.- (76.67%) No.- (3.33%) Not Sure.- (20%)	Yes.- (80%) No.- (20%)	Yes.- (80%) No.- (20%)
How simple it is to identify individual records?	Very Simple.-(66.67%) Simple.- (20%) Difficult.-(13.33%)	Very Simple.-(96.66%) Nil.- (3.34%)	Very Simple.-(96.66%) Nil.- (3.34%)
Does your institution gives out student data to a third party	Yes.- (56.66%) Not Sure.- (43.34%)	Yes.- (26.67%) No.- (10%) Not Sure.- (63.33%)	Yes.- (26.67%) No.- (10%) Not Sure.- (63.33%)
Are you aware of information privacy and data protection law?	Yes.- (53.33%) No.- (26.67%) Not Sure.- (20%)	Yes.- (80%) Not Sure.- (20%)	Yes.- (80%) No.- (20%)
Does the institution prevent students' data from any attack?	Yes.- (100%)	Yes.- (70%) No.- (30%)	Yes.- (70%) No.- (30%)

Table 2: Sample of questionnaire responses of last three Institutions

The analysis provided here covers the average of the total score against response for all institutions. From table 1 above, the results provide a detailed individual score for each higher institution of learning in the state. The most interesting things to note from the table are: (1). 92.22% of the student records were collected and kept for all six institutions, and 75.38% were kept for eternity as indicated. Interestingly, 91.11% of the record were stored in computing platforms used by various institutions but, 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. (2). As shown in the table 68.33% of all student records of all schools are prone to internal attack due to the simplicity of identifying individual records with less effort. (3). from the table also we can see that all responses from the schools show protection of data privacy with high scores that amounted to 77.78. But on the contrary, the techniques employed against the record are for confidentiality and not privacy protection. (4). 42.77% shows how student records are faced with a strong external attack based on the fact that schools do give out data to the third party for one reason or the other in plain format; even though, respondents claimed knowledge of information privacy and data protection law. For these, we can attest to the fact that the entire records for the whole institutions of Kebbi state are being faced with privacy issues threats as figures indicated in table 1 because of the absence of any privacy protection techniques applied to the information. Table 3 below will provide an individual ranking of privacy threats—highest to lowest per institution.

V. INSTITUTION PRIVACY THREATS

Table 1 and 2 analyses were consolidated for all schools and therefore, in this sub-section, we shall discuss privacy threats per individual institutions according to some parameters derived from the questionnaire and responses even though, there is no existing parameter that ranks privacy threats of industry or institution. The parameters involved here are five, Plus their total score from respondents of each school and are as follows:

- Keeping student's record
- Using a computing platform for record-keeping
- Format of the record
- The simplicity of student identification within the record
- Measures employed to prevent an attack

We use the formula as Sum of Parameters / total number In our case, the sum of parameters / 5. Based on this, we have the following ranking of highest privacy threats to lowest per school.

Privacy Threats Ranking Per Institution				
Name of Institution	**Parameter Calculation**	**Score (%)**	**Ranking**	**Students Affected**
WUFP l	80+100+56.67+63.33+56.67 / 5	71.33	5th	>5000
K/S Pol. Dakin-gari	100+86.67+60+66.67+80 / 5	65.87	6th	2000
COE. Argungu	100+86.67+56.67+53.33+90 / 5	77.33	4th	5000
Health Tech. Jega	73.33+80+76.67+66.67+100 / 5	79.33	3rd	>5000
AUST	100+93.33+80+96.66+70 / 5	87.99	2nd	>5000
FUBK	100+100+80+96.66+70 / 5	89.33	1st	>5000

Table 3: Privacy Threats Ranking Per Institution

Table 3 above revealed the extent of the privacy threat for each school. However, this focuses only on the responses got from the respondents as indicated in table 1 and concentrates on the highest score. The sixth column of the table presents the total number of students that might be vulnerable to privacy risk. Consequently, we can see that none of the institutions is safe about privacy measures. Figure 3 below shows the proposed framework for the six institutions to guide them in preventing an attack.

As student data is collected, it ought to be cleaned in the first place before anything. The prepared data then be kept in any storage medium for the anonymization process. Schools are at the liberty to choose from available tools or techniques that satisfy their anonymity and utility, refer [1], [2], and [3] for anonymization techniques and tools. With this framework, data users can issue out de-identified data to the third party with less fear and harm.

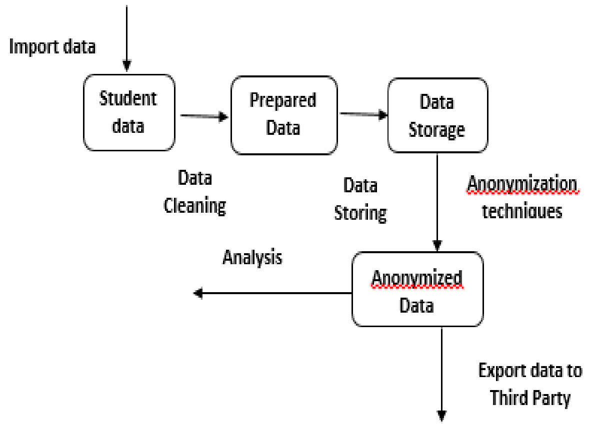

Fig. 3: Privacy Framework

VI. 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 4 below is the entire work activity diagram.

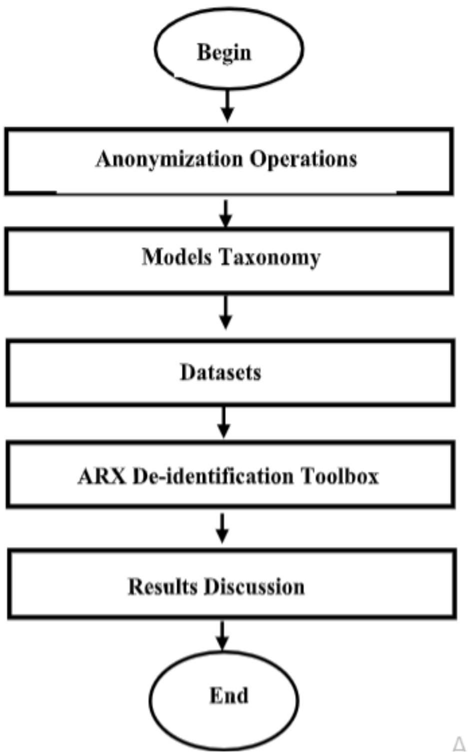

Fig. 4: Activity Diagram

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

A. *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.

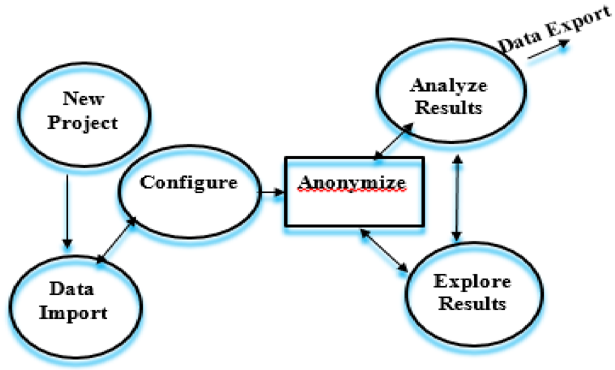

Fig. 5: Experiment Framework

B. *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 4 and 5 show the overview of the datasets.

Overview of the Dataset				
Dataset	**Quasi-Identifier**	**Records**	**Highest Transformation**	**Size (KB)**
Student	8	180	1,223,040	2.

Table 4: Overview of the Dataset

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)

Table 5: Overview of the attributes in the datasets

C. *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 1.3GHZclock 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.

a) Parameter Value
Parameter values of k used in the experiment were recommended as the best configurations in[4]. As for parameter *L* values also cannot exceed the distinct values of SA for a good result, refer to [4] and [2], thus, this research takes care of that. Table 6 below summarizes the configurations used in the experiments carried out.

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]	

Table 6: Experimental Configurations

D. 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 micro aggregation [5]. It was developed within three years by five computer scientists in Germany, refer [6]. For ARX graphical interface refer to [2].

VII. RESULTS AND DISCUSSIONS

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 [7].

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, 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 [8]. 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.

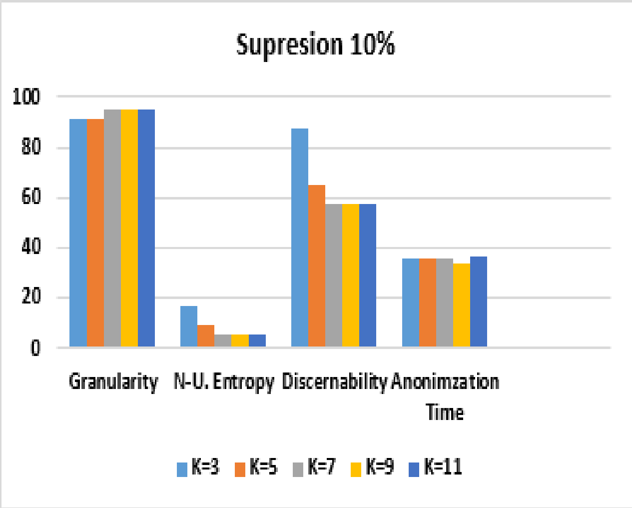

Fig. 6: Suppression limit 10% for all k values

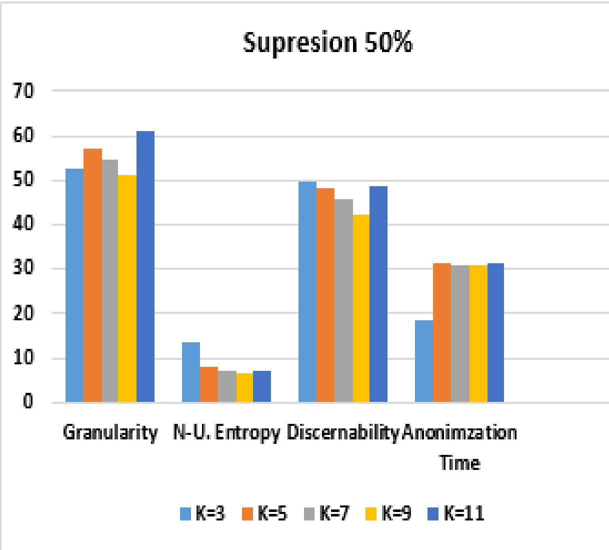

Fig. 7: Suppression limit 50% for all k values

VIII. RELATED WORK

In [9] 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. [10] compared only two models and the dataset was unknown. Execution time was measured, and it was concluded that k-anonymity outperformed l-diversity. In the work of [7], five privacy models were compared out of which one of them-***slicing*** is the anonymization technique and not the privacy model [11]. 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. [12] presents 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 super class 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 groups, 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.

IX. CONCLUSION

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 found that none of the schools employ any privacy technique as such, and all of them are faced with privacy threats. Moreover, the work rank the institutions' privacy threats from 1st through 6th based on some parameters recorded from questionnaire and responses. Additionally, none of the respondents has a clear view of what privacy is all about, as such; they all misunderstood data privacy with data confidentiality. Therefore, tor the work proposed a privacy framework for the six institutions to employ to mitigate the threats. More importantly, there is a need for stakeholders in all the institutions to educate data holders about privacy and privacy-enhancing technologies.

REFERENCES

[1.] F. Prasser, F. Kohlmayer, and K. A. Kuhn, “Efficient and effective pruning strategies for health data de-identification,” *BMC Medical Informatics and Decision Making*, pp. 1-14, 2016.

[2.] Prasser, Bild, Eicher, Spengler, Kohlmayer, and Kuhn, “Lightning: Utility-Driven Anonymization of High-Dimensional Data,” *Transactions on data privacy*, pp. 161-185, 2016.

[3.] N. D, K. P, and U. S, “Performance Evaluation of K-Anonymization Algorithms for Generalized Information Loss,” *IJCTA*, pp. 227-235, 2016.

[4.] Kohlmayer, Prasser, and Kuhn, “The cost of quality : Implementing Generalization and Suppression for Anonymizing Biomedical Data With Minimal Information Loss,” *Journal of Biomedical Informatics*, pp. 37-48, Sep., 2015.

[5.] Prasser, Kohlmayer, Lautenschläger, and Kuhn, “ARX - A Comprehensive Tool for Anonymizing Biomedical Data,” in in AMIA Annual Symposium, 2014, pp. 984-992.

[6.] Prasser, Eicher, Bild, Spengler, and Kuhn, “A Tool for Optimizing De-Identified Health Data for Use in Statistical Classification,” in International Symposium on Computer-Based Medical Systems 2017, pp. 169-174.

[7.] Prasser, Kohlmayer, and Kuhn, “The Importance of Context: Risk-based De-identification of Biomedical Data,” *Journal of Biomedical Data*, pp. 347-355, 2016.

[8.] P. Jain, M. Gyanchandani, and N. Khare, “Big data privacy: a technological perspective and review,” *Journal of Big Data*, pp. 1-25, 2016.

[9.] AlShwaier, and Emam, “Data privacy on e-health care system,” *International Journal of Engineering, Business and Enterprise Application*, pp. 89-99, 2013.

[10.] Dubli, and Yadav, “Secure Techniques of Data Anonymization for Privacy Preservation,” *International Journal of Advanced research in Computer Science*, vol. 8, pp. 1693-1695, Jun., 2017.

[11.] F. Kohlmayer, F. Prasser, and K. A. kuhn, “A Flexible Approach to Distributed Data Anonymization,” *journal of Biomedical informatics*, vol. 50, pp. 62-76, Aug., 2014.

[12.] F. Prasser, F. Kohlmayer, and K. Kuhn, “A Benchmark of Globally-Optimal Anonymization Methods for Biomedical Data,” in International Symposium on Computer-Base Medical System, 2014, pp. 66-71.

[13.] Y. Vural, and M. Aydos, "A New Approach to Utility-Based Privacy Preserving in Data Publishing." pp. 204-209.

[14.] T. Ravindra, P. Maheshwari, and K. Binod, “A Comparative Study of Data Suppression Technique for Privacy of Individuals,” *International Journal of Current Trends in Engineering & Technology*, vol. 04, no. 01, 2018.

[15.] M. M. Almasi, T. R. Siddiqui, N. Mohammed, and H. Hemmati, "The risk-utility tradeoff for data privacy models." pp. 1-5

[16.] J. A. P, and T. A. Selvadoss, “Comparison and Analysis of Anonymization Techniques for Preserving Privacy in Big Data,” *Advances in Computational Sciences and Technology ISSN*, pp. 0973-6107, 2017.