Privacy Preserving Publication of Educational Data

Tamanna Rahaman
Department of Computer Science & Engineering
Bangladesh Army University of Engineering & Technology
Rajshahi, Bangladesh
is.rahaman08@gmail.com

Md. Muktar Hossain
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
muktar.cse11.ruet@gmail.com

Fahmida Haque Mim
Department of Computer Science & Engineering
Bangladesh Army University of Engineering & Technology
Rajshahi, Bangladesh
fhmimu09@gmail.com

Farah Wahida
Department of Computer Science & Engineering
Rajshahi University of Engineering & Technology
Rajshahi, Bangladesh
mewki@gmail.com

2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI) | 978-1-6654-3843-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ACMI53878.2021.9528126

***Abstract*—The need for securing data privacy has increased as the online publication of data has grown by leaps and bounds in recent years. This situation puts a lot of people at risk of a security breach. K-anonymity, which was the first published method of providing data anonymity is no longer as effective to keep up with the demands of various types of dynamic data, like- medical, transactional, trajectory, educational etc. Research into data privacy has gained more focus to keep up with the demands of time, but educational data is still relatively unexplored. As more of the schooling system is moving online, it caused an increase of research into student's results. This published academic information can cause a breach of privacy for the students. In our paper, we compare educational data with transactional data and modify the anony method to suit the safe publication of educational data.**

Keywords—Educational Data, Data Publication, Privacy Breach, Inference Attack

I. INTRODUCTION

The rise in technology has made all kinds of information available to anyone and everyone. While easily accessible data has made our lives more effortless, it also has a downside. With so much of our personal information online, the privacy of many individuals is compromised. A single piece of information that might seem harmless can be used as background knowledge by an enemy for more sinister purposes through data mining and composition attack[11]. An attacker can put together different information from different sources to infer that the individual did not want to be known to the public. That is why we need more secure methods of data publication.

So, research into data publication has become more prevalent nowadays. The earlier work [1][10] was solely for static one-time data publication, but now new methods are being discovered every day for dynamic data sets and multiple publications [8][9]. Most of these study was done with medical data [14] as it is considered highly private. Transaction [6] and trajectory [12] being a close second in terms of the number of research. This paper focuses on the academic data of various students. This data type is mainly unexplored but necessary, as every year, students' data is released for statistical study. While working on data publication of transactional data [6], we noticed a similarity of characteristics in our school's published result. This prompted us to apply the already established algorithm on the result data set. Still, we discovered that as the data is not wholly similar, there remains a significant risk of identity exposure if someone possesses some background knowledge. When we consider a scenario of attack, it is usually carried out by a known person to the victim. So it is also safe to presume the attacker has some previous knowledge about the victim.

TABLE 1: STUDENT RESULT

	Name	ID	Gender	Semester	CGPA	Fail
1	Ana	16204001	F	2	2.98	Math, History
2	John	16204002	M	2	3.45	
3	Peter	16204003	M	2	3.24	Biology
4	Sam	16204004	M	2	3.95	
5	Bobby	16204005	M	2	2.30	Math, Chemistry, Biology, Physics
6	Jo	16204006	F	2	3.78	
7	Ellen	16204007	F	2	3.53	
8	Charlie	16204008	F	2	3.66	
9	Dave	16204009	M	2	2.72	Physics, Biology
10	Ben	16204010	F	2	3.19	Math

TABLE 2: ANONYMIZED STUDENT RESULT

	ID	Gender	Year	CGPA	Fail
1	162040**	M	2	3.24	Biology
2	162040**	M	2	3.95	
3	162040**	M	2	2.30	Math, Chemistry, Biology, Physics
4	162040**	M	2	2.72	Physics, Biology
5	162040**	M	2	3.45	
6	162040**	F	2	3.19	Math

978-1-6654-3843-8/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: UNIVERSITY TEKNOLOGI MALAYSIA. Downloaded on April 11,2022 at 07:40:13 UTC from IEEE Xplore. Restrictions apply.

7	162040**	F	2	3.53	
8	162040**	F	2	3.66	
9	162040**	F	2	2.98	Math, History
10	162040**	F	2	3.78	

Example 1: Table 1 represents a few students' results in the same year. Here anyone can identify the students with their name and ID. The gender and year are quasi-identifiers. The CGPA and List of subjects the student failed at is a sensitive attribute. Table 2 shows the anonymized student result. In the anonymized table, the name attribute is removed, the ID attribute is k-anonymized, and the data set is shuffled. The name is a direct identifier, so that's why it was deleted. The same can be said about the ID, but it contains other properties that will compromise the data utility. So, the ID is generalized instead of removed. The same goes for gender attributes. The real result was published ID wise, so the anonymized table is shuffled at random to stop someone from linking the ID to individual rows.

Now table 2 should have been anonymized, but background knowledge can help an adversary link result to individual students. This depends on what type of knowledge the attacker has. Consider different scenarios where an attack can happen.

Scenario 1: The attacker knows Ana is female and has failed in math. Only 2 female students have failed in math, so the probability of guessing wright is 1/2 = 50%. If the attacker knows Ana has failed in history, the probability moves to 100% because only one female student has failed in history.

Scenario 2: The attacker knows Dave has failed in 2 subjects but not which subjects. There are 2 tuples with the number of fail 2, and one of them is male and the other female. So, the attacker can say with surety that row 7 must belong to Dave. The same can be said about Peter and Ben. They both only failed in one subject, and one is female and one male.

Scenario 3: The attacker knows Bobby got the lowest grade in the class and wants to know which subject he failed. Here the lowest grade is row 3, and it must belong to Bobby.

Similarly, other student's records are also at risk of an inference attack. The background knowledge is not hard to obtain, and with this knowledge, the probability of identity disclosure is very high.

II. Motivation

The whole world is moving forward in terms of technology. Even a few years ago, so much business moving online was inconceivable, but today it's our reality. Especially the education system has undergone a massive change. Now students are not restricted to classrooms. Online learning has become the norm. Every year many institutes publish educational data that includes details about students, their performance records, attendance, course feedback, and educators' details. This supports the study of academic growth in different areas, the environmental effect on schooling, statistics of failure rate, the effectiveness of teaching methods, etc. This study of educational data has opened up a new field of study called Learning Analytics (LA). According to [13], LA was defined as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. "But openly sharing these data which contain sensitive information about students, can cause harm. The thought of improving the situation motivated us to research safer modes of data publication.

III. Related Work

Using data analytics, we can analyze data to make better decisions, verify models and theories that help us come to more well-researched conclusions. As a result, many academics have concluded that large-scale data collaboration and big data are the future of learning analytics, fueling greater education [2] [3]. Already, studies [4] are underway to cumulate online education data into a sole combined dataset and then analyze that data set to find elements affecting learner retention and development. However, open sharing of data, no matter how well-meaningly it is done, can cause privacy breaches, as shown by Sweeney in [5] and singer in [15]. Many more studies in the Learning Analytics field show concern about the ethical and confidentiality implications of the publication of educational data. So, our work focuses on mitigating some of these concerns.

IV. Dataset

A data set is a group of information systematized as a stream of bytes in a logical record. Here, we used a simple academic result data set, and it contains ID, Registration number, Session, Semester, Gender, CGPA, Fail. Table 1 provides a simplified section of the data set. Firstly, we remove all personal information (name, address, phone number, parent's name) from the dataset. Here, we used some general identifiers e.g., ID, Registration Number, Session), some QI (Semester, Gender), and some sensitive terms (CGPA, Fail). After reviewing the data set, anyone can quickly identify who has failed in a subject or more than a subject, gender, and result. If we use ** to replace the last 2 digits of the id, it will be quite challenging to identify the person within their generalized group.

V. Data Pre-processing

The data set we used needed to be pre-processed before we could run experiments with it. The original data set was taken from our own university result of one semester. There were a lot more attribute in the original data. The student's personal information like- name, phone number, address etc. was removed. We selected the relevant columns for our study and discarded the rest. The missing values in selected attributes was first replaced with null values and later replaced with the max value of that column. The noise was minimum after data selection and replacing the missing values. The few rows with noisy data was removed. The number of rows with noisy data was small so their removal did not cause that much change in the result.

Authorized licensed use limited to: UNIVERSITY TEKNOLOGI MALAYSIA. Downloaded on April 11,2022 at 07:40:13 UTC from IEEE Xplore. Restrictions apply.

VI. METHODOLOGY

Our proposed method, *E-anony*, has two main steps these are-

1. **Clustering:** The input data set is clustered into smaller sections based on some arbitrary condition, like – result, grade point or number of failed subject. The cluster record number has to be equal or more than 2. The cluster should not be so large that the data utility is lost.

2. **Sanitization:** Each cluster is sanitized meaning that no one record of a cluster can be linked to one specific individual. This is achieved by merging the sensitive attributes of a cluster into a global bag belonging to that cluster. After every cluster is sanitized the clusters are combined to produce an output that can be safely published without privacy risks. If an attribute already exists in the global bag than the superscript number over that attribute will indicate how many times the attribute is mentioned.

Algorithm: E-anony

Input: **T** table with attribute set $\{A_1, A_2, \ldots, A_n\}$, *where* $A_i \in T$

Output: Anonymizes table T^i

1. Suppress Direct identifiers: name, mobile number etc.
2. K-anonymize QID, quasi Identifiers
3. Create smaller cluster C_i where every $C_i \in A_i$
4. Sanitize clusters by pushing Sensitive attribute S in global bag Cg
5. While A_i != 0, where i = number of cluster
6. Push S in C_g
7. If S already in C_g,
8. $S = S^{j++}$
9. End if
10. End while
11. Output new $T = T^i$

To further understand the algorithm we need to be familiar with some of the definitions mentioned below-

Definition 1: Attributes

The characteristics column of the data set is called attributes. Let A_i denote the attributes. The data set, denoted by *T*, is made up of several attributes. They are represented by a set of $\{A_1, A_2, \ldots, A_n\}$ where each *A* is a different attribute. There are three types of attributes; these are direct identifiers, quasi-identifier, and sensitive attributes.

Definition 2: Direct identifiers

The student ID, roll number, registration number are direct identifiers. These attributes each have unique values, and they can be used to identify any student directly. Let D_{ij} denote a direct identifier, and all values of this attribute belong to A_i. So, set of $\{D_{11}, D_{12}, \ldots, D_{1n}\} \in A_1$.

Definition 3: Quasi-identifiers

The attributes that are not unique to an entity but can help identify an individual when joined with other identifiers. In our data set, gender, semester, session year, etc., are quasi-identifiers. Let Q_{ij} denote quasi-identifiers, and Q_{ij} belongs to A_i. So, set of $\{Q_{11}, Q_{12}, \ldots, Q_{1n}\} \in A_1$.

Definition 4: Sensitive attribute

The attributes that are private to each individual is called a sensitive attribute. We are attempting to modify the data set in such a way that no sensitive attribute can be linked to a single individual. *S* denotes the sensitive attribute.

In student database result, CGPA, Score points, Fail subjects, etc. are considered a sensitive attribute.

Definition 5: Cluster

Cluster is smaller sets of input data denoted by C_i where every C_i is a part of the input table T. The input is divided into finite set of clusters $\{C_1, C_2, \ldots, C_n\}$.

Definition 6: Global Bag

The sensitive attributes of a cluster is merged into one column and that is called the global bag of that cluster. It is denoted by C_g.

VII. EXPERIMENTAL RESULTS

Table 3 is the result of applying our proposed method on table 1. If we try to apply the same scenario attacks on the anonymized data set we will see different result now.

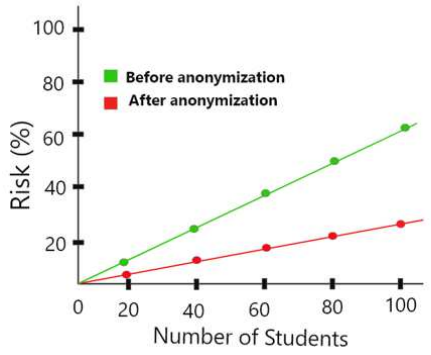
TABLE 3: ANONYMIZED DATA SET

	ID	Gender	Year	CGPA	Fail
Cluster 1: C_1					
1	162040**	M	2	3.24	Math, Chemistry, Biology^3, Physics^2
2	162040**	M	2	2.72	
3	162040**	M	2	2.30	
Cluster 2: C_2					
4	162040**	M	2	3.95	
5	162040**	M	2	3.45	
Cluster 3: C_3					
6	162040**	F	2	3.78	
7	162040**	F	2	3.53	
8	162040**	F	2	3.66	
Cluster 4: C_4					
9	162040**	F	2	2.98	Math^2, History
10	162040**	F	2	3.19	

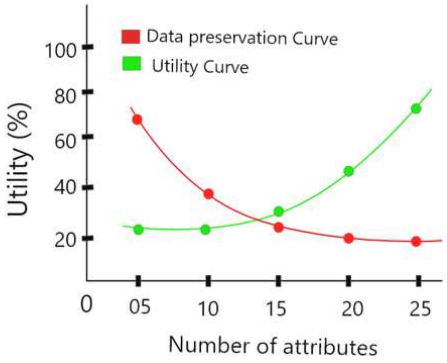
Scenario 1: Even if the attacker knows Ana has failed in history, the probability remains 50% because both record 9 and 10 corresponds to a female student failing in history.

Scenario 2: The attacker knows Dave has failed in 2 subjects but he can be anyone among records 1, 2 and 3. So, Dave's record cannot be linked to any one sensitive attribute.

Authorized licensed use limited to: UNIVERSITY TEKNOLOGI MALAYSIA. Downloaded on April 11,2022 at 07:40:13 UTC from IEEE Xplore. Restrictions apply.



(a) Risk of inference attack vs. Number of students



(b) Utility of data vs. Number of attributes

Fig. 1: Experiment Results

Scenario 3: The attacker knows Bobby got the lowest grade in the class and wants to know which subject he failed in. Here the lowest grade is row 3, and it must belong to Bobby but the attacker cannot link the sensitive attribute of subjects he failed at with his record thus his privacy is protected.

In this section we also aim to show 1. Results of inference attack on the data set before and after applying anonymizing method and 2. The effect of attribute number on both data privacy and utility through a plotted graph. Before applying the E-anony method the risk of identity leakage was higher. The proposed method mitigated that risk but did not totally erase it. We hope to improve that result in further study. The results are shown in Fig 1(a). On another hand we found while data pre-processing if we remove some attributes the chance of privacy breach drops dramatically. As an example consider the Gender column in Table 1, without the knowledge of the sex of the individual the chances of background knowledge attack is considerable less. Also if the result is only shown in CGPA and the failed list removed over 50% of the data could not be linked with an individual. But removing these columns means compromising the data utility. The information researches need will not be present in the data set so utility will drop significantly. This is shown in Fig 1(b) with a diagram.

VIII. Conclusion

Research into data publication with preserved data security is an ongoing effort. Finally, we can say that the proposed method has increased the privacy for the publication of educational data by modifying the privacy method *E-anony*. The risk of identity disclosure from an inference attack is significantly reduced but not eliminated completely. Removing one of the sensitive attributes will give the data set much-improved privacy, but that will compromise the data's integrity too much to be of use any more. In the future, we hope to develop the method further and apply it to other types of data.

Reference

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

[2] V. Mayer-Schonberger and K. Cukier, Learning with Big Data: The Future of Education, Boston, MA, USA:Houghton Mifflin.

[3] G. Siemens and R. S. d Baker, "Learning analytics and educational data mining: Towards communication and collaboration", Proc. 2nd Int. Conf. Learn. Analytics Knowl., pp. 252-254, Apr. 2012.

[4] P. Ice, "The PAR framework proof of concept: Initial findings from a multi-institutional analysis of federated postsecondary data", J. Asynchronous Learn. Netw., vol. 16, no. 3, pp. 63-86, 2012.

[5] L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon Data Privacy Working Paper 3, 2000.

[6] Juyong Li,Michael Bewong, Jixue Liue, Lin Liu, J iuyong Li and Kim-Kwang Raymond Choo, “A Relative Privacy Model for Effective Privacy Preservation in Transactional Data”, In: IEEE Trustcom/BigDataSE/ICESS.2017 ,PP.394-401.

[7] M.Al Karim, A.Karim, S.Azam, E.Ahmed, F. De Boer, A.Islam, and F. N. Nur, “Cognitive Learning Environmentand Classroom Analytics (CLECA): A Method Based on Dynamic Data Mining Techniques,” Innovative DataCommunication Technologies and Application, pp. 787–797, 2021

[8] Raymond Chi-Wing Wong, Ada Wai-Chee Fu,Jia Liu,Ke Wang,Yabo Xu . “Global Privacy Guarantee in Serial Data Publishing”, In: Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, pp.995-959

[9] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu, Jian Pei. “Anonymity for Continuous Data Publishing”,In : EDBT 2008, 11th International Conference on Extending Database Technology.

[10] Benjamin C.M. Fung, Ke Wang, RUI CHEN, PHILIP S. YU. “Privacy-Preserving Data Publishing: A Survey of Recent Developments”,In : ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010

[11] A H M Sarowar Sattar, Sumyea Helal. In: Privacy Risk Against Composition Attack, International Journal of Innovative Research in Computer Science & Technology (IJIRCST),Volume-6, Issue-2, March 2018K. Wang and B. C. M. Fung. Anonymizing sequential releases. In KDD, 2006.

[12] Alastair R Beresford and Frank Stajano. Location privacy in pervasive computing. IEEE Pervasive computing, 2(1):46–55, 2003.

[13] G. Siemens and P. Long, "Penetrating the Fog: Analytics in learning and education", EDUCAUSE Rev., vol. 46, no. 5, 2011

[14] F.H.Semantha, S.Azam, K.C.Yeo, and B. Shanmugam, “A Systematic Literature Review on Privacy byDesign in the Healthcare Sector,” Electronics, vol. 9, no. 3, p. 452, Mar. 2020.

[15] N. Singer, "InBloom student data repository to close", New York Times, Apr. 2014.

Authorized licensed use limited to: UNIVERSITY TEKNOLOGI MALAYSIA. Downloaded on April 11,2022 at 07:40:13 UTC from IEEE Xplore. Restrictions apply.