JOURNAL OF INTERACTIVE MEDIA IN EDUCATION

Kyritsi, KH, et al. 2019. The Pursuit of Patterns in Educational Data Mining as a Threat to Student Privacy. *Journal of Interactive Media in Education,* 2019(1): 2, pp. 1–10. DOI: https://doi.org/10.5334/jime.502

ARTICLE

The Pursuit of Patterns in Educational Data Mining as a Threat to Student Privacy

Kyriaki H. Kyritsi*, Vasilios Zorkadis†, Elias C. Stavropoulos* and Vassilios S. Verykios*

Recent technological advances have led to tremendous capacities for collecting, storing and analyzing data being created at an ever-increasing speed from diverse sources. Academic institutions which offer open and distance learning programs, such as the Hellenic Open University, can benefit from big data relating to its students' information and communication systems and the use of modern techniques and tools of big data analytics provided that the student's right to privacy is not compromised. The balance between data mining and maintaining privacy can be reached through anonymisation methods but on the other hand this approach raises technical problems such as the loss of a certain amount of information found in the original data. Considering the learning process as a framework of interacting roles and factors, the discovery of patterns in that system can be really useful and beneficial firstly for the learners and furthermore, the ability to publish and share these results would be very helpful for the whole academic institution.

Keywords: privacy; learning analytics; distance learning; data publishing; anonymization; statistical disclosure control

Introduction

Major breakthroughs in technology have exponentially reduced the cost of collecting and storing data and as a consequence the amount of collected data has been enormously extended. In computer terminology these vast amounts of divergent data that are produced at an impressive speed are called 'big data' and are best described using a number of dimensions or V's (see, for example, Firican, 2017). The three most commonly used dimensions are: **volume**, which describes the vast amount of data that are constantly generated in our digitised world, changing the storage unit from gigabytes (10^9 bytes) to zettabytes (10^{21} bytes); **velocity**, which corresponds to the speed at which data are created, especially in real-time applications; and **variety**, which describes the different forms of generated data such as text, images, voice and geospatial data. In addition to these dimensions some more have recently been introduced: **veracity**, which signifies the quality of data; **valence**, which describes the connectedness of data in the form of graphs which look like atoms; and finally **value**, which describes the opportunities lying in big data analytics along with the practical use and the benefits from it.

In particular, the analysis of big data is considered as a way of capitalising on commercial activities and on manufacturing, retail and financial services, etc., because it offers reliable insights into actual problems by finding trends, patterns, correlations and other features that exist among data. Nevertheless, there are ethical and legal issues concerning the privacy of individuals, as there are questions about how personal data are being made available, for example, in the case of social media where users seem willing to sacrifice their privacy in favour of adopting new trends in communication. In an article published in 2004, Helen Nissenbaum (2004, p. 118), an associate professor at New York University, introduced the concept of privacy as a 'contextual integrity' to explain that almost everything people do happens in a context that cannot easily be categorised as only private or only public, and often these two may overlap. For that reason, data privacy should always be considered with respect to the context into which the data is being used and analyzed.

It becomes obvious, that data and especially personal data provide information and facts that carry a powerful source of knowledge for many aspects of our social and economic life, while at the same time the privacy of data subjects cannot be overlooked. Consequently to that limitation, modern societies apply strict regulations and laws to protect individuals and their rights. In May 2018, the EU brought into force the General Data Protection Regulation (GDPR), which is by far the most complete regulatory framework applying to all EU citizens in relation to data protection. The GDPR introduced the concept of 'privacy by design' and key changes in the area of human rights, as it emphasises the 'right to privacy' and enforces the notion of consent. In addition, the regulation brought into law

* Hellenic Open University, GR

† Hellenic Data Protection Authority, GR

Corresponding author: Elias C. Stavropoulos (estavrop@eap.gr)

the obligation to announce a data breach within 72 hours of the occurrence of it and, finally, made it possible to apply serious penalties in the case of non-compliance.

In this study, the intention was to experimentally test the opportunities of learning analytics provided the privacy of learners is protected and at the same time to look into the threats and losses upon the utility of information. The rest of the paper is organised as follows. Related work, below, gives a brief background insight; the section on key concepts and definitions refers to the basic concepts and definitions of data privacy-preserving methods; and the section following that describes *k*-anonymity and *l*-diversity, the two methods of anonymisation that have been applied for the purposes of this study. The implications of this are analysed in the subsequent case study using educational data from the Hellenic Open University (HOU). Finally under Discussion and Conclusions we present the results of the study, considering how they could become applicable in the learning process.

Related work

While the term 'big data analytics' is used to describe all the methods applied to the analysis of big data, the term 'learning analytics' is used to describe 'the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs' (Long and Siemens, 2011, p. 34). Learning analytics is an area of research that builds upon ideas from other fields such as process mining, business intelligence, data processing, information retrieval, technology enhanced learning, educational data mining and data visualisation (Scheffel, 2015, p. 2).

In education, learning analytics dashboards (LADs) are considered to be important tools for visualising the results of analyses and conveying information to learners, teachers, course administrators and other stakeholders. The utilisation and evaluation of LADs is, for example, the subject of the research of Gkontzis et al. (2017, p. 17), who concentrated on three tools of the HOU: (1) the forum graph report, which depicts the interactions among students and between students and tutors in an online platform (forum) which facilitates the communication in distance learning courses; (2) the course dedication block, a tool which measures for each student, the time spent in the online platform for the purposes of the certain course; and (3) the analytics graphic block, a descriptive tool containing all available information for each learner such as name, courses, grades and other pieces of information.

Research on data privacy has been developed based upon two approaches or scenarios. The first scenario involves privacy-preserving data publishing, which actually means sharing data with third parties without violating the privacy of those individuals whose (potentially) sensitive information is in the data. This approach is also referred to as 'non-interactive systems' (Gursoy et al., 2017, p. 68). The second scenario involves privacy-preserving data mining or disclosure control, and is also called 'interactive anonymisation systems' (Davis and Osoba, 2016, p. 2).

Following the trend for open data in the U.S., a team of researchers from Harvard University and Massachusetts Institute of Technology announced in May 2014 the release of an open data set containing student records from 16 courses that ran during the first year after the launch of the edX MOOC platform. In the paper describing the release of this data and exploring the implications and challenges that arose, Daries et al. (2014) considered how the value of data is affected by privacy-preserving methods. As a means of measuring the utility loss between the de-identified and the original data, they note the difference in the amount of correlation observed on the de-identified data compared to the initial correlation found on the original data.

Finally, in order to have a wider understanding of up-to-date methods in data privacy and especially in the field of privacy-preserving data mining we studied association rule hiding, a method that belongs to the subfield of knowledge hiding. This technique ensures that only the useful part of information is mined and that sensitive information is excluded from the mining operation. Kagklis et al. (2014, pp. 2–4 & pp. 6–8) synthesised some certain taxonomy of frequent item set hiding techniques. Especially on the category of linear programming-based hiding techniques, the authors presented an analytical case study of the more frequently used algorithms and came to the conclusion that there is a trade-off between time complexity/scalability and the side effects of the hiding process.

Key concepts and definitions

As noted previously, there are two approaches to data privacy: privacy-preserving data publishing and privacy-preserving data mining. Certain aspects are common to both methods, and it is therefore useful to define the discrete characteristics present in every data set and categorise them according to the distinct type of information that they contain.

- Direct identifiers are those features and the data used to describe them that can uniquely identify an individual.
- Quasi-identifiers are attributes that when used alone do not necessarily disclose an individual's identity, but in combination with external databases can identify a data subject.
- Sensitive attributes are attributes containing private information that people normally do not feel willing to share with others or reveal in public.
- Auxiliary information is data that bears no privacy risk and does not fit into any of the above categories.

Secondly, if a data breach occurs it is possible the intruder or anyone who subsequently gains access to the data may be able; a. to identify individuals, b. to find new facts about them or c. to make inferences on certain personal characteristics. This can happen in three ways (Templ et al., 2014, pp. 2–3):

- Identity disclosure occurs when a known individual is associated with a released data record.
- Attribute disclosure occurs when an intruder is able to determine new attributes of an individual based on the information available in the released data, for example the medical condition, educational level and others.
- Inferential disclosure occurs when an intruder is able to determine the value of some characteristic of an individual more accurately with the released data than otherwise would have been possible, for example the average annual earnings of a certain profession.

Finally, given the scope of this paper it is useful to define and distinguish de-identification from anonymisation (Elliot et al., 2016, pp. 15–16):

- De-identification refers to the process of removing or masking direct identifiers such as a person's name, address and any other unique characteristic associated with a person.
- Anonymisation refers to the process of ensuring that the risk of somebody being identified in data is negligible. Anonymisation not only protects individuals from being directly re-identified, but also ensures that re-identification may not occur indirectly too.

Privacy through *k*-anonymity and *l*-diversity

In the privacy-preserving data publishing scheme which is also called a non-interactive system the goal is to transform a data set in order to enforce a certain definition of privacy before sharing it with third parties or publishing it openly.

Assuming that unique values in a data set are more likely to be re-identified than the values that are not unique in the given data set, then one way to protect confidentiality is to ensure that each distinct pattern of key variables is possessed by at least a minimum number of records; let us call it a k number of records in the sample (Templ et al., 2014, pp. 5–7). More specifically, the method of *k*-anonymity (Samarati and Sweeney, 1998, pp. 4–5; Sweeney, 2002, pp. 8–10) makes the assumption that if enough entries (rows) are indistinguishable from each other, then the privacy of subjects will be preserved since each subject's data would be associated with a group of persons as opposed to the individual in question.

In the relevant terminology, a 'key value' is a certain combination of values from different variables of a given data set which represents a pattern that is considered important to the scope of each data analysis. These values are not pre-defined.

Let f_k be the frequency or the count of records with pattern *k* in the sample. A record is called a sample unique or a unique record in the sample if it has a pattern k of which the $f_k = 1$.

Let F_k denote the number of units in the population having the same pattern k. A record is called a population unique or a unique record in the population if $F_k = 1$.

In order to achieve *k*-anonymity, a typical practice is to set k = 3, which ensures that the same pattern of the key variables is possessed by at least three records in the sample and by that notation 3-anonymity is achieved; meaning $f_k \geq 3$ for all records.

For example, consider a summary statistics table (see **Table 1**) representing educational data from two basic courses of a university program. For the purpose of this example, assume that the key value is the set of all the attributes appearing in the table, that is, the 'Course Code', 'Grade', 'Gender' and 'Age' of the students. Each row represents a distinct record in the data set (or sample) and the last column, named 'Frequency', describes the frequency or the count of records with the same pattern of the key value in the data set. In brief, this produces the following summary statistics.

From this data it can be determined that:

- In the case of course 'PL10', it is easy to determine that all female students (records 2 and 3 of the table) passed the course. It is also clear that the only male student failed to pass the course.
- In the case of course 'PL11', anyone who knows the gender and age of the data subjects can find out the attribute of the grade: female students in rows 4 and 8 have unique values and the same stands for each of the records of the male students (records 5, 6, 7); meaning that the combination of gender, age and grade turns these records into unique records in the sample (the frequency is 1/8 for each one of them).
- For the four attributes 'Course Code', 'Grade', 'Gender' and 'Age', all records represent unique values in the data. The threat is that the more unique values in a dataset, the higher the risk of re-identification.

It is possible to achieve *k*-anonymity through applying generalisation or suppression, or a combination of these two methods.

Generalisation is a method applied to categorical data in order to recode categories with few observations into a single category with larger frequency counts. Applying generalisation to continuous variables means discretising the variable into levels representing an interval of the

Table 1: Example of data containing unique records in the sample.

	Course code	Grade	Gender	Age	Frequency (for the set of all the attributes)
1	PL10	4	Male	22	1/8
2	PL10	5	Female	24	1/8
3	PL10	5	Female	28	1/8
4	PL11	7	Female	28	1/8
5	PL11	8	Male	21	1/8
6	PL11	8	Male	23	1/8
7	PL11	6	Male	22	1/8
8	PL11	5	Female	23	1/8

original values. In the above example, generalisation is applied to the attribute of age by recoding into two new intervals, respectively [21, 24] and [25, 28]. Then the frequency counts with respect to the remaining attributes are calculated once again. The results of the generalisation are shown in **Table 2**.

Examining the column 'Frequency (for the attribute of age)' demonstrates that the threshold of 2-anonymity has been achieved. Nevertheless, the frequency counts in the last column, 'Frequency (for the set of all the attributes)', show that 2-anonymity is not achieved for the whole data set and as a consequence the above table could not be released. This leaves the options of either dropping out at this point or continuing with the anonymisation process by applying further generalisation or suppression.

Suppression is applied if unique values of key variables remain after recoding. It can be achieved by injecting missing values to replace any values that contain a high disclosure risk and are thus considered unsafe in the *k*-anonymity model. An example of suppression is given in **Table 3**. To begin with, for the variable of 'Gender' all values with frequency counts less than the threshold of the value 2 (2-anonymity) are replaced. The frequency counts for 'Age' and 'Gender' are then calculated once again. (Remember, the goal in this example is to achieve *k*-anonymity for the set of all variables or attributes).

Following the above process, the data set is 2-anonymous with respect to 'Age' and 'Gender', although it has partially lost information relating to the attribute of 'Gender' in some records. However, the data set as a whole remains not anonymous as there are four records with a frequency count of 1/8. There are two options for continuing: either to apply suppression on the records with a frequency count of 1/8 and lose a significant part of the value of the data or apply further generalisation to the attribute of 'Age' and the attribute of 'Grade'. In the case of 'Age' more generalisation would mean total loss of the information because the two intervals would have to be recoded into one.

The *k*-anonymity method has the limitation that even if a group of observations fulfils *k*-anonymity, an intruder can still discover sensitive information. To address that problem, the notion of *l*-diversity has been developed as a means of diversifying the specific sensitive attribute(s) and achieving stronger privacy standards. According to Machanavajjhala et al. (2007, pp. 14–23), the purpose of *l*-diversity is to create an *l*-diverse group of observations, or in other words, a group of observations that contains l 'well-represented' values for the sensitive variable.

To demonstrate the *l*-diversity method, suppose that in the example of the educational data the attribute of 'Grade' is considered as a sensitive variable and the set of values 'Course Code', 'Gender' and 'Age' as the key values. Suppose that each of the records is *k*-anonymous with respect to the key variables, but with respect to the sensitive variable, an intruder can discover new information. The data set and the sensitive attribute of 'Grade' are shown in **Table 4**.

The first, the fourth and the eighth record of the table (records 1, 4, 8) are not distinct *l*-diverse. The options are either to remove (suppress) these records in order to achieve a minimum of 2-distinct values in the k-anonymous group of observations or to make further transformations in the original data. Applying the first option creates the following *2*-anonymous and *2*-diverse table (**Table 5**).

From the above example, it is obvious that the loss of information or utility loss is rather large as there has been a suppression of three out of eight records of the data set (37.5% of the values of the data set).

A Case Study on Educational Data

The original data consisted of five (5) datasets coming from a module of an undergraduate program of the HOU and contained: a set of administrative data and four sets of forum activity data. The set of administrative data included the scores of students in written assignments, online quizzes and other evaluation projects, the score on the final test and finally some other information such as student's name and surname, student's registration number and e-mail address. The other datasets contained the system log files that were generated throughout an academic year by the use of four (4) different fora of communication. In total, there were 105.604 system log files.

The Hellenic Open University divides the total number of students in groups of approximately 30 students each for better learning results. For each group, one tutor is appointed responsible with responsibilities,

Table 2: Example of applying generalisation to achieve *k*-anonymity.

	Course code	Grade	Gender	Age interval	Frequency (for the attribute of age)	Frequency (for the set of all attributes)
1	PL10	4	Male	21–24	6/8	1/8
2	PL10	5	Female	21–24	6/8	1/8
3	PL10	5	Female	25–28	2/8	1/8
4	PL11	7	Female	25–28	2/8	1/8
5	PL11	8	Male	21–24	6/8	2/8
6	PL11	8	Male	21–24	6/8	2/8
7	PL11	6	Male	21–24	6/8	1/8
8	PL11	5	Female	21–24	6/8	1/8

Table 3: Example of applying suppression to achieve *k*-anonymity (before the suppression).

	Course code	Grade	Gender	Age interval	Frequency (for age and gender)	Frequency (for the set of all attributes)
1	PL10	4	Male	21–24	4/8	1/8
2	PL10	5	Female	21–24	2/8	1/8
3	PL10	5	Female	25–28	2/8	1/8
4	PL11	7	Female	25–28	2/8	1/8
5	PL11	8	Male	21–24	4/8	2/8
6	PL11	8	Male	21–24	4/8	2/8
7	PL11	6	Male	21–24	4/8	1/8
8	PL11	5	Female	21–24	2/8	1/8

Example of applying suppression to achieve *k*-anonymity (after the suppression).

	Course Code	Grade	Gender	Age interval	Frequency (for the set of all attributes)
1	PL10	4	****	21–24	1/8
2	PL10	5	****	21–24	2/8
3	PL10	5	****	25–28	2/8
4	PL11	7	****	25–28	1/8
5	PL11	8	male	21–24	2/8
6	PL11	8	male	21–24	2/8
7	PL11	6	male	21–24	1/8
8	PL11	5	****	21–24	1/8

Table 4: Example of applying the *l*-diversity method to *k*-anonymous educational data (first step of *l*-diversity).

	Course code	Gender	Age interval	Frequency (for the set of all attributes)	Sensitive attribute: grade	Distinct *l*-diversity
1	PL10	****	21–24	3/8	4	1
2	PL10	female	21–24	2/8	5	2
3	PL10	female	****	2/8	5	2
4	****	female	****	4/8	7	1
5	PL11	male	21–24	3/8	8	2
6	PL11	male	21–24	3/8	8	2
7	PL11	male	21–24	3/8	6	1
8	****	female	21–24	4/8	5	2

throughout the academic year, such as correcting the assignments, giving guidance and teaching students in face-to-face meetings.

Methodology

In order to measure the disclosure risk and the utility loss of data, three of the above five big data sets were utilised and to these there were applied firstly data management, secondly data analysis and thirdly data anonymisation techniques.

More specifically, the data contained:

- almost 90,000 data logs recording all the forum activity generated by students and tutors containing the date and time of the data log, IP address of the user, identification number of the program's module, the identification number of the data log along with the label of the activity relating respectively to: viewing the lesson, viewing a discussion, creating a discussion or a post and adding some content to a discussion or a post
- data containing registration information (e.g. ID number, email addresses) – that is, personally identifiable information

Table 5: Example of suppression on a *k*-anonymous data set in order to achieve *2*-diversity (second step of *l*-diversity).

	Course code	Gender	Age interval	Frequency (for the set of all attributes)	Sensitive attribute: grade	Distinct *l*-diversity
1	PL10	female	21–24	2/5	5	2
2	PL10	female	****	2/5	5	2
3	PL11	male	21–24	2/5	8	2
4	PL11	male	21–24	2/5	8	2
5	****	female	21–24	2/5	5	2

- scores of students in written assignments, online quizzes and the final test

The first step of the methodology used MS Excel to perform data cleaning, complex computations and data transformations in order to produce a single database containing all the necessary information while at the same time hiding the direct identifiers. In addition, two new variables were created; the first, representing the overall forum activity in number of data logs for each student and for each tutor; and the second, representing the number of days a student or tutor was active on the forum. This database was then subjected to the same analysis before and after anonymisation.

Analysis before anonymisation

The second step of the methodology, using the SPSS Statistical Package, applied descriptive statistics in order to search whether, and to what extent, each one of the forum activities was correlated to the grades of students in three major performance tests; the first containing the average grade of the written assignments taken throughout the academic year, the second containing the average grade in the online quizzes and the third and most important; the final test taken in the end of the program. For that purpose, we computed the correlation coefficients for each one of these three scores (and the forum activity variables) and tested the significance of the coefficients by using statistical tests.

More specifically, hypothesis testing is a statistical tool that measures the probability of an assumption (called the null hypothesis) to be true or false according to a predecided level of probability. This level of probability is called the p-value and signifies the pre-decided tolerance for the Type I error (the null hypothesis being rejected when it is actually true) and for the Type II error (the null hypothesis being accepted when it is actually false). Following this process, the results of the tests were used to decide on which forum activities act as direct identifiers, as quasi-identifiers and as sensitive attributes.

Analysis after anonymisation

In the third step of the methodology, two methods were applied: one offered as a default method by the ARX Anonymization Tool which creates 'generalisation hierarchies'; and an empirical method that used recoding of categorical as well as continuous variables with a more intuitive setting of the range of the data intervals, using SPSS.

With the ARX tool, after manually setting the corresponding parameters for *k*-anonymity, the tool performs certain transformation leading to the creation of 'generalisation hierarchies' and then calculates the re-identification risk of the anonymised data set.

With the empirical method, the data set was recoded into a new one consisting of bigger groups of records having values within the same frequency intervals instead of having numerous records with discrete values, thus having records with unique frequency counts in the data set. The method is empirical because it is mostly decided by the analyst using the statistical tool, meaning that the analyst sets the range of the intervals and not vice versa. Then, the new data set is loaded into the ARX tool in order to calculate the re-identification risk and prove it is *k*-anonymous or repeat the process, if not, until it becomes anonymous.

Finally, the utility loss of information after the anonymisation process was measured for both methods by computing the same statistical test as in the second step of the methodology; that is calculating the correlation coefficients among the forum activities and the score (for the three performance tests; average of the written assignments, average of online quizzes and the final test). The purpose of this test was to measure whether there existed a significant change in the correlation among the above mentioned quantities (the forum activities and the grades) prior and after the anonymisation process; a decrease in the amount of the coefficient should indicate a significant information loss.

Results

In this section, the most important results from the methodology and the analysis of the pre-anonymised as well as the post-anonymised data are being briefly presented.

Figure 1 is a representative graph (scatter plot) showing how the forum activity called 'discussion created', represented on the y-axis, plots against the score achieved by students in the 'final test', represented on the x-axis. From the plot, it is obvious that the majority of students do not create many discussions because the highest value for this variable is 5 discussions. Nevertheless, it can be clearly observed that the creation of higher numbers of discussions is more common among students who obtained higher scores on the 'final test' even though this correlation does not imply a strong linear relationship among these two variables.

The same stands for the forum activity called 'post created' when considered with respect to the score on the 'final test', as represented in **Figure 2**. The majority of

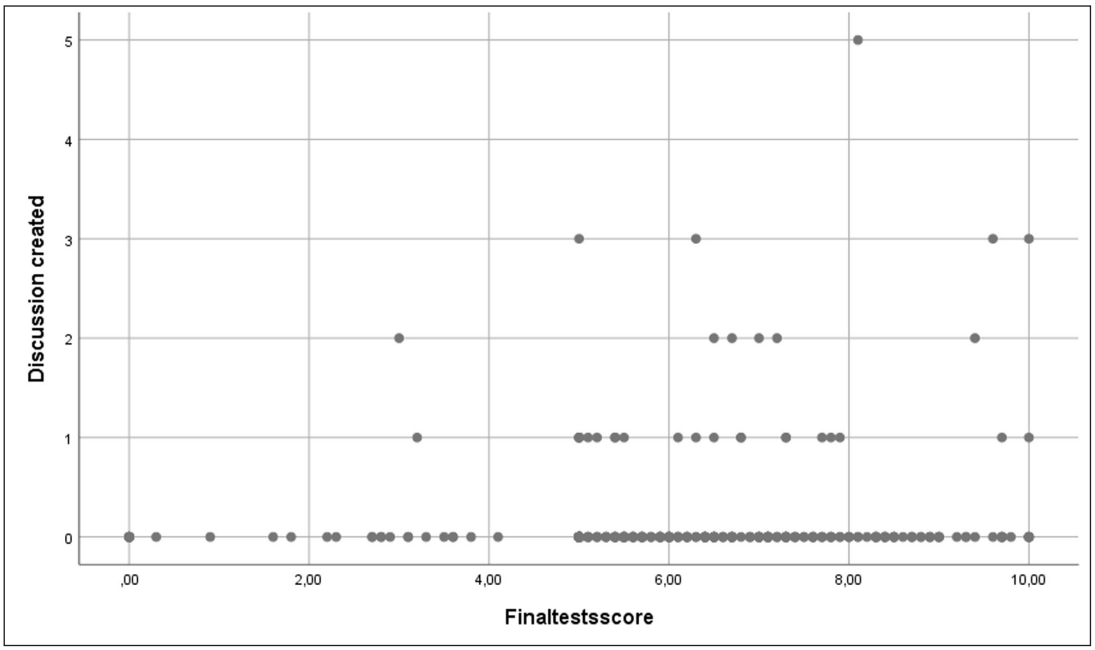

Figure 1: Scatter plot of the 'discussion created' forum activity and the score achieved in the 'final test'.

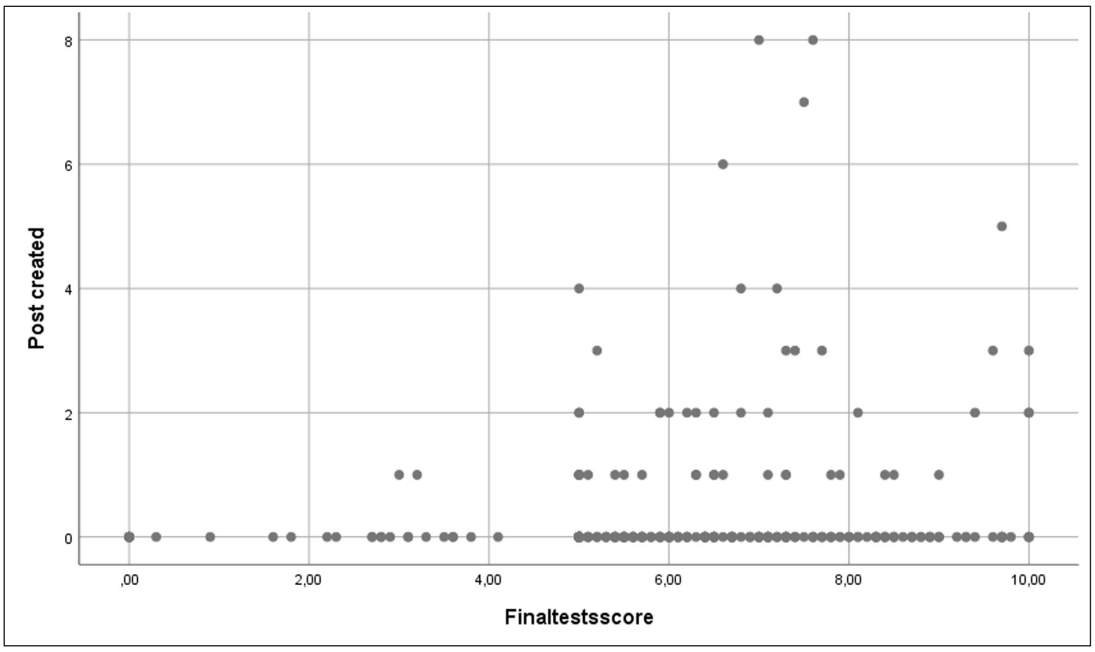

Figure 2: Scatter plot of the 'post created' forum activity and the score achieved in the 'final test'.

students create at most two posts during the online academic year. Nevertheless, there seems to be a positive correlation between the creation of posts and achieving a higher score on the 'final test'.

In order to test the above observations, we calculated the Pearson's correlation coefficients for each of the forum activity types (see **Table 6**) and the score achieved by students on the three performance tests (see Methodology) and then performed the relevant statistical test to determine whether the resulting correlation coefficients were significant. The coefficients can take positive or negative values in the [−1, +1] interval. In order to have a significant correlation, the computed p-value must be smaller than the level of significance, which we took as the 0.01 level. In **Table 6**, the correlation coefficients and the statistical tests among the forum activity types and the 'final test' are presented.

Table 7 shows how the risk of re-identification, as computed by the ARX tool, helped to create a ranking among the forum activities as possible quasi-identifiers.

Finally, **Table 8** presents one of the recoded variables and more specifically the new variable 'new number of days active' coming from recoding the original values of the 'number of days active' into seven categories of activity (from extremely low to extremely high activity) according to the respective frequency counts of each value. It becomes obvious that these data transformations and the recoding of the data records resulted in a really significant decrease of the re-identification risk. For example, from the initial 7.317% (**Table 7**) percentage of unique records

Table 6: The correlation coefficients among the score on the 'final test' and forum activity.

	Final test score		
		p-value	
Forum activity	**Pearson's correlation coefficient**	**Significant at the 0.01 level (*)**	**Not significant correlation**
Number of days active	+ 0.269	(<0.001)*	
Discussion viewed	+ 0.248	(<0.001)*	
Discussion created	+ 0.103		(0.069)
Post created	+ 0.148	(0.009)*	
Total number of logs	+ 0.260	(<0.001)*	

Table 7: The risk of re-identification and the ranking of the quasi-identifiers.

Quasi-identifiers	Re-identification risk	Unique records in the sample	Unique records in the population	Risk ranking (from 1 the highest, to 8)
Number of days active	21.951%	7.317%	0.473%	4
Lesson viewed	28.048%	13.902%	2.272%	3
Discussion created	1.463%	0.487%	0.040%	7
Discussion viewed	31.219%	12.439%	0.848%	2
Post created	2.195%	0.731%	0.092%	6
Some content has been posted	3.170%	0.975%	0.125%	5
Total number of logs	43.658	24.390%	3.481%	1
Gender	0.487%	0%	0%	8

Table 8: The recoded variable 'new number of days active'.

	Frequency	Percent	Cumulative percent
Extremely low activity – values up to 5 days	129	31.5	31.5
Very low activity – values up to 20 days	122	29.8	61.2
Low activity – values up to 35 days	54	13.2	74.4
Medium activity – values up to 50 days	39	9.5	83.9
High activity – values up to 66 days	27	6.6	90.5
Very high activity – values up to 100 days	29	7.1	97.6
Extremely high activity – values more than 100 days	10	2.4	100.0
Total	410	100.0	

found in the original 'number of days active', it is apparent that the values in the new intervals have higher frequency counts (**Table 8**) and there are no unique values in the data set.

Discussion and Conclusions

No doubt, there are great benefits to the social sciences from the utilization of big data and big data analytics provided we use the proper tools and have the right skills to mine and analyze it. On the flip side, there are certain constrains in the use of personal information arising from ethical issues relating to privacy rights especially when data has been collected without consent or without a clear scientific, historical or similar purpose.

The privacy-preserving data publishing methods and techniques have been developed to meet those ends and provide a safe place in data science offering the means to make data useful and respect the privacy of the data subjects. However, this is a quite complex problem consisting of various factors such as the context of privacy, the form of data, the goals of the project and many other components which interact to create a level of difficulty. Moreover, what is more intriguing is the search for methods to maintain the utility of data in values with statistical

significance so that academic research on educational data offers conclusions which are both useful and assure the privacy and anonymity of the data subjects.

Using educational data from the HOU, this study experimented with the anonymisation process that must be applied before sharing data containing personal and/or sensitive information. The first step of the process, creating the database, requires analytical skills and the second step of the methodology using the SPSS, helps to determine the variables that act as direct identifiers and/or as quasi-identifiers; as a consequence that is very helpful for the anonymisation process.

Furthermore, the second step of the above analysis is by far the most interesting because through it we can discover how the variables are interrelated with each other and we can use these findings as a decision tool for the utility of the forum in the learning process. Different variables play a diverse role and testing the significance of their correlations provides a new perspective on the knowledge we want to explore and mine from the data.

In the anonymisation phase of the experiment, it was observed that there existed a ranking among the different quasi-identifiers with respect to the importance of the estimated re-identification risk for each one of them, so consequently before publishing the data set we must check any given combination of quasi-identifiers. In addition to that, it was observed that adding more quasi-identifiers into a combination always increased the disclosure risk of the whole data. On the contrary, adding more records to the data set decreased the re-identification risk of the whole data set as this became bigger and contained less unique records in it.

Also in the anonymisation phase of the experiment, we explored the utility loss of data after the anonymisation process, and concluded that *k*-anonymity, by its structure, can lead to a level of information loss that in an extreme case means the data are no longer usable. The 'trade-off' between preserving privacy and preserving the value of information or preventing utility loss upon data anonymisation is a challenge, requires numerous data records and of course, analysts with certain skills.

Summing up, the analysis of the case study led to the conclusion that the use of forum activities were correlated to the scores, even though not very strongly; as the coefficients were closer to 0 than the value of +1. Nevertheless, there is still a significant correlation to the scores thus it can be clarified that the use of the forum actually helps students on the assignments, quizzes and most important on the final test.

Especially in the scores on the 'final test', which is the most crucial for completing the module; the average score among students not using the forum was 2.94 whereas the average score among students using the forum was 6.23 when the minimum grade that qualifies passing the course is 5.

In the future, it would be interesting to conduct further research on the interacting parts of the online communication platform in order to identify more trends and patterns within this system and provide the facts which will support suggestions and improvements for the overall learning performance. These may relate to the expectations of students, the requirements of the academic staff, certain skill indexes which apply globally or just in the EU, and many other factors that all together build a big educational model perceived as a dynamic construction rich in different types of ever-increasing information.

Competing Interests

The authors have no competing interests to declare.

References

Daries, JP, Reich, J, Waldo, J, Young, EM, Whittinghill, J, Seaton, DT, Ho, AD and **Chuang, I.** 2014. 'Privacy, anonymity, and big data in the social sciences', *ACM Queue*, 12(7) [Online]. Available at https://queue.acm.org/detail.cfm?id=266164 (Accessed 1 March 2019).

Davis, JS and **Osoba, OA.** 2016. *Privacy Preservation in the Age of Big Data: A Survey* [Online]. Available at https://www.rand.org/content/dam/rand/pubs/working_papers/WR1100/WR1161/RAND_WR1161.pdf (Accessed 1 March 2019).

Elliot, M, Mackey, E, O'Hara, K and **Tudor, C.** 2016. *The Anonymisation Decision-Making Framework* [Online]. Available at http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf (Accessed 1 March 2019).

Firican, G. 2017. 'The 10 Vs of big data', *Upside*, 2 August [Online]. Available at https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx (Accessed 1 March 2019).

Gkontzis, AF, Karachristos, CV, Lazarinis, F, Stavropoulos, EC and **Verykios, VS.** 2017. 'A holistic view on academic wide data through learning analytics dashboards.' In: Ubachs, G and Konings, L (eds.), *Conference Proceedings: The Online, Open and Flexible Higher Education Conference, 25–27 October 2017*, Maastricht, European Association of Distance Teaching Universities (EADTU), pp. 12–27 [Online]. Available at https://eadtu.eu/home/policy-areas/lifelong-learning/publications/430-online-open-and-flexible-higher-education-conferences (Accessed 1 March 2019).

Gursoy, ME, Inan, A, Nergiz, ME and **Saygin, Y.** 2017. 'Privacy-preserving learning analytics: challenges and techniques', *IEEE Transactions on Learning Technologies*, 10(1): 68–81 [Online]. Available at https://ieeexplore.ieee.org/document/7563858 (Accessed 1 March 2019). DOI: https://doi.org/10.1109/TLT.2016.2607747

Kagklis, V, Verykios, VS, Tzimas, G and **Tsakalidis, AK.** 2014. 'Knowledge sanitization on the web'. In: *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14)*, 1–11. Thessaloniki, 2–4 June. New York, USA, ACM Press. [Online]. DOI: https://doi.org/10.1145/2611040.2611044 (Accessed 1 March 2019).

Long, P and **Siemens, G.** 2011. 'Penetrating the fog: analytics in learning and education', *EDUCAUSE Review*, 46(5): 31–40. Available at https://er.educause.

edu/~/media/files/article-downloads/erm1151.pdf (Accessed 1 March 2019).

Machanavajjhala, A, Gehrke, J, Kifer, D and **Venkitasubramaniam, M.** 2007. '*l*-diversity: privacy beyond *k*-anonymity', *ACM Transactions on Knowledge Discovery from Data*, vol. 1, March 2007, article no.3 [Online] https://desfontain.es/PDFs/PhD/LDiversityPrivacyBeyondKAnonymity.pdf (Accessed 6 April 2019). DOI: https://doi.org/10.1145/1217299.1217302

Nissenbaum, H. 2004. 'Privacy as contextual integrity', *Washington Law Review*, 79: 100–39 [Online]. Available at https://crypto.stanford.edu/portia/papers/RevnissenbaumDTP31.pdf (Accessed 1 March 2019).

Samarati, P and **Sweeney, L.** 1998. 'Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression'. In: *Proceedings of the IEEE Symposium on Research in Security and Privacy* [Online]. Available at https://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf (Accessed 6 April 2019).

Scheffel, M. 2015. *A Framework of Quality Indicators for Learning Analytics (Learning Analytics Review 2)*, Bolton, LACE [Online]. Available at http://www.laceproject.eu/publications/learning-analytics-quality-indicators.pdf (Accessed 4 March 2019). DOI: https://doi.org/10.1145/2723576.2723629

Sweeney, L. 2002. '*k*-anonymity: a model for protecting privacy', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5): 557–570 [Online]. DOI: https://doi.org/10.1142/S0218488502001648 (Accessed 4 March 2019).

Templ, M, Meindl, B and **Kowarik, A.** 2014. *Introduction to Statistical Disclosure Control (SDC)*, IHSN [Online]. Available at https://www.semanticscholar.org/paper/Introduction-to-Statistical-Disclosure-Control-(-)-Templ-Meindl/311d73e8e4b5b1dca49901d929f383bbd0817a7c (Accessed 4 March 2019).

How to cite this article: Kyritsi, KH, Zorkadis, V, Stavropoulos, EC and Verykios, VS. 2019. The Pursuit of Patterns in Educational Data Mining as a Threat to Student Privacy. *Journal of Interactive Media in Education,* 2019(1): 2, pp. 1–10. DOI: https://doi.org/10.5334/jime.502

Submitted: 22 December 2018 **Accepted:** 07 February 2019 **Published:** 27 May 2019

]u[*Journal of Interactive Media in Education* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS