Privacy-Preserving Learning Analytics: Challenges and Techniques

Mehmet Emre Gursoy, Ali Inan, Mehmet Ercan Nergiz and Yucel Saygin

Abstract—Educational data contains valuable information that can be harvested through learning analytics to provide new insights for a better education system. However, sharing or analysis of this data introduce privacy risks for the data subjects, mostly students. Existing work in the learning analytics literature identifies the need for privacy and pose interesting research directions, but fails to apply state of the art privacy protection methods with quantifiable and mathematically rigorous privacy guarantees. This work aims to employ and evaluate such methods on learning analytics by approaching the problem from two perspectives: (1) the data is anonymized and then shared with a learning analytics expert, and (2) the learning analytics expert is given a privacy-preserving interface that governs her access to the data. We develop proof-of-concept implementations of privacy preserving learning analytics tasks using both perspectives and run them on real and synthetic datasets. We also present an experimental study on the trade-off between individuals' privacy and the accuracy of the learning analytics tasks.

Index Terms—Data mining, data privacy, learning analytics, learning management systems, protection.

1 Introduction

The low cost of handling data along with the technological advances in data mining and big data have led service providers to collect, process, and analyze huge amounts of data in the hope of discovering the great value within. Educational data is no exception. There is nowadays a wide variety of digital information available to educational institutions about learners, including performance records, educational resources, attendance to course activities, feedback on course materials, course evaluations and social network data of students and educators. New educational environments, technologies and regulations are being designed to further enrich the types of information made available to institutions [40]. With all the diverse set of data types and sources of information, we face loosely-structured and complex data in educational systems [24].

Rich educational data sources, the need for a better understanding of how students learn, and the goal of enhancing learning and teaching have led to the new field of Learning Analytics (LA). In [44], 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". Surely, educational data and LA have great potential value. Analytics performed on past data can benefit future teaching practices [37]. Predictive models that characterize the current performance of a student can help forecast performance in the future (possibly to prevent failures and/or promote success). The more data available about learners, the better the learning process can be analyzed, and the more effective the cooperative and collaborative learning groups will become [23]. Visualization of learners' data can lead to better and more timely feedback.

While there are clear benefits in collecting, utilizing and sharing educational data, the sensitive nature of the data raises legitimate privacy concerns. Many initiatives and regulations protect personal data privacy in domains such as health, commerce, communications and education [11], [18], [51], [55]. Most regulations do not enforce absolute confidentiality which would cause more harm than good [5], [33], but rather protect 'individually identifiable data' that can be traced back to an individual with or without external knowledge. This gave rise to a wide range of studies primarily focusing on de-identifying private data with as little harm to its information content as possible, in an attempt to preserve both the privacy and usefulness of the data.

It is difficult to give a broad definition of data privacy without a specific context [33]. Privacy in the context of education should be considered with respect to various scenarios. Research on data privacy has formally defined and enforced privacy primarily in two scenarios: (1) Sharing data with third parties without violating the privacy of those individuals whose (potentially) sensitive information is in the data. This is often called *privacy-preserving data publishing*. Research in this area can also enrich the open data initiatives for learning analytics, e.g., [7]. (2) Mining data without abusing the individually identifiable and sensitive information within. This is often called *privacy-preserving data mining* or *disclosure control*.

In this paper, we study appropriate methods for both scenarios, bearing in mind the requirements of educational data and learning analytics. Our contributions, in this paper,

- *M. E. Gursoy is with the College of Computing, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: memregursoy@gatech.edu.*
- *A. Inan is with the Computer Engineering Department, Adana Science and Technology University, Adana, Turkey. E-mail: ainan@adanabtu.edu.tr.*
- *M. E. Nergiz is with Acadsoft Research, Gaziantep, Turkey. E-mail: nergiz@gmail.com.*
- *Y. Saygin is with the Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey. E-mail: ysaygin@sabanciuniv.edu.*
- *This research was funded by The Scientific and Technological Research Council of Turkey (TUBITAK) under grant number 114E261.*

Corresponding author: Ali Inan.
Manuscript received ?; revised ?.

can be summarized as follows:

- We show how the aforementioned, semi-structured and complex educational data can be modelled with a hierarchical data structure.
- We assess the applicability of existing data privacy methods to educational data and learning analytics. This requires a critical study on the pros and cons of various methods, due to the unique nature and characteristics of educational data.
- We present the problem of *privacy-preserving learning analytics (PPLA)*, and extend some of the well-known privacy methods to educational data to offer solutions to the PPLA problem. We present technical detail regarding how these privacy methods can be enforced in practice.
- We provide proof-of-concept implementations of interesting learning analytics tasks to experimentally demonstrate the trade-off between privacy and utility.

2 Related Work

Data analytics is about examining data in order to draw conclusions for better decision making or to verify models and theories. In that regard, many academics see large-scale *data collection* and *big data* as promising fields of research. According to Mayer-Schonberger and Cukier, the future of learning analytics lies in big data [26]. There have already been studies (e.g., the PAR project [19]) that aggregate online educational data into a single, federated dataset and then analyze the dataset for factors affecting student retention, progression, and completion. On the other hand, several academics criticize the hype for big data technologies and their consequences. In [3], [4] and [14], authors point out that data may contain hidden bias, e.g., only wealthy schools have computerized education and the data collected from these schools do not accurately represent the whole population. In social sciences, great effort is spent on the collection of data - especially when trying to pick a representative sample of a population. Many of the new algorithms on big data omit such careful consideration. Therefore conclusions reached by these algorithms should not be blindly trusted. Danaher critically refers to this as *algocracy* (i.e., being ruled by algorithms) in [6].

In this work, we focus on the problem of *privacy*. In [47], [48] and [49], Solove discusses why it is difficult to formulate what is private and what constitutes a privacy violation. He provides a taxonomy of privacy violations, and compares and contrasts different elements of his classification. Within his taxonomy, our work is concerned with *privacy in information dissemination*, i.e., sharing the data or the information extracted from the data while preserving privacy. In [16], Gurses provides a bird's eye view on privacy and describes the roles of engineers in building systems that operate on private data. Two important aspects from her study are confidentiality and control. The latter is concerned with individuals' right to control how their data is used and disseminated. The author argues that a quantifiable and open privacy protection mechanism is valuable. Our work aims to formulate such mechanisms in the realm of learning analytics.

Without a doubt, educational data contains private and sensitive information. Recent LA papers call for collaboration and open learning initiatives [43], which can benefit researchers from all over the world. However, sharing sensitive information requires extra care with regards to privacy. Careless attempts to collect and share data in other domains have led to privacy problems. In [52], Sweeney showed that using simple demographic information, one can uniquely identify the majority of the US population. In [29], Narayanan and Shmatikov showed that an insufficiently anonymized Netflix movie rating database can compromise the identity of a user. A recent incident from the LA community is the InBloom disaster [45]: InBloom was a non-profit corporation offering to warehouse and manage student data, e.g., attendance and grades. It had to shut its doors because parents complained about privacy, e.g., they found some data on InBloom too intimate and were not comfortable with a third-party vendor acquiring this data. The major issue regarding InBloom was its data collection and storage policies, whereas in this work we focus on data and information dissemination.

Several works in the LA domain discuss the ethical and privacy implications of educational data. In [17], Heath points out that institutions are collecting educational big data, and offers philosophers', lawyers' and education specialists' perspectives on privacy. In [46], Slade and Prinsloo emphasize the definition of consent and students' ability to opt-out of data collection. In [38], Prinsloo and Slade evaluate the policy frameworks of two large-distance education institutions according to a set of considerations including who benefits under what conditions, consent, de-identification and opting-out. Authors conclude that current policies are mostly concerned with academic analytics (data security, integrity of demographic data) and not with learning analytics (learners' data at the course and departmental level). As such, there is a pressing need for enforcing privacy during the processes of obtaining and sharing LA results. These two can be achieved using privacy-preserving data mining and data publishing techniques, respectively.

Studies concerning the legal and ethical questions in LA specifically target *consent*. In [50], Solove emphasizes that everyone has the right to manage how his/her data is stored and processed, i.e., individual consent should be central in the analysis of private data. In [39], Prinsloo and Slade study the consent and data collection policies of three popular MOOC providers. The authors find that although MOOC providers explicitly inform their users on what data is collected, opting out is not an option. In [41], Sclater and Bailey cite the Data Protection Act of the UK, and argue that students should have the right to view their data or LA results that use their data. In [37] and [41], authors briefly touch upon the need for anonymization to de-identify private data, but do not give a detailed, technical perspective. We do so in this work. Further, we believe that consent is tightly connected to privacy and anonymity: most people would not want their data to be used or shared, unless they are assured that they will remain anonymous.

Anonymity and anonymization in LA have also been discussed in [8] and [54]. In [54], Swenson asserts that data should not be used by any party without proper anonymization, and even anonymized data can be de-anonymized

in a way that can violate privacy. In [8], Drachsler and Greller present DELICATE, a checklist for trusted learning analytics. Although the checklist is broad, one of the main points raised is the need for anonymization. The Learning Analytics Community Exchange (LACE) project has recently put great emphasis on privacy and anonymization for learning analytics. Their report [15] provides a list of ethical and privacy issues concerning LA, together with the current policies and regulations. The report concludes by stating that the trust of students and staff is essential for the adoption of LA, and a great way of fostering trust is to embed support for privacy in LA tools. To achieve this, we believe that technical solutions, such as the ones we present in this work, are necessary.

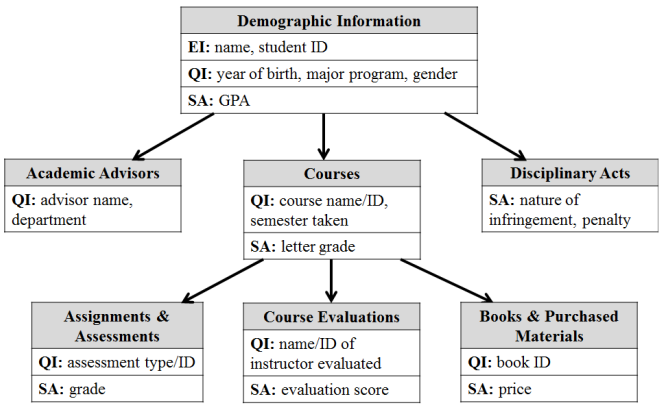


Fig. 1: Schema for student data records

3 Key Concepts and Definitions

This section introduces the notions related to the collection, conceptual modeling, storage and mining of educational data, motivates the need for privacy, and briefly outlines the system architecture.

3.1 Actors Involved and Their Roles

In the process of data collection and mining, the following *roles* can be identified:

- **Data subjects** are persons and entities whose (potentially sensitive) data is collected and analyzed. In this work, we focus on students as the primary data subjects, but instructors' and schools' sensitive information should also be protected.
- The **data owner/curator** is the party that collects and stores data regarding the subjects. The data curator often decides whether data should be shared with third parties, in what manner and using which privacy measures. School administrators can be regarded as the data curator.
- **Data analysts and recipients** include all parties that are given access to the data, e.g., third-party LA experts, data scientists. In the case where data is published (e.g., made available on the Web) the public can be seen as the data recipient.

Actors are those parties that interact with the data collection and learning analytics system. There is no clear-cut mapping between actors and their roles. Certain actors often need to have multiple roles: a course instructor needs to access and modify parts of the data to grade her students. Yet, she should be prohibited from viewing students' grades in other classes, or which student gave her a bad evaluation. This information should be made available only in aggregate or anonymized form, if at all.

3.2 Types of Information

A database contains several *attributes* (i.e., data fields). For example, in tabular data, a column corresponds to an attribute and each cell in that column contains a *value* for that attribute. In terms of privacy, attributes can be divided into four categories:

- **Explicit Identifiers (EI)** are attributes that uniquely and explicitly identify a data subject. Names, student IDs and social security numbers are examples of EIs. Removal of these attributes is necessary but not sufficient to ensure privacy.
- **Quasi-Identifiers (QI)** are attributes that do not necessarily disclose individuals' identity when used alone, but can be used in combination and/or together with external databases to single out data subjects. Examples include gender, date of birth and courses taken.
- **Sensitive Attributes (SA)** are private information such as GPA, letter grades etc. that data subjects are usually not willing to share with third parties. A privacy-preserving scheme should prohibit an adversary to make inferences regarding subjects' sensitive information.
- **Auxiliary Information** is data that bears no privacy risk and does not fit into any of the categories above. This data is often useful for LA, e.g., a course's learning outcomes and objectives defined by the instructor.

One aspect of privacy is *contextual integrity* [33], [34]. This is a conceptual framework that ties privacy to specific contexts (e.g., healthcare, education) and argues that information gathering and dissemination should be specific to that context. For instance, accessing and viewing a student's health record is not a privacy violation in the context of healthcare, but it is a violation in the context of education. To comply with the ideas of contextual integrity, the data curator needs to be careful in deciding what constitutes a quasi-identifier, what is sensitive and what is not.

3.3 Student Data Records (SDRs)

We say that a student's education-related information at an institution is collected into a single data record we call a *student data record (SDR)*. One record per student is maintained. Each SDR should follow a similar, but loosely-defined schema. A sample schema is provided in Fig. 1. We place no constraints on SDRs apart from the ability to model them using a tree-like (i.e., hierarchical) data structure. For example, the data curator can have the flexibility of choosing which attributes are QIs, SAs etc. at each level, according to the type of data he has. This data model can also trivially support tabular and set-valued data. As part of our funded

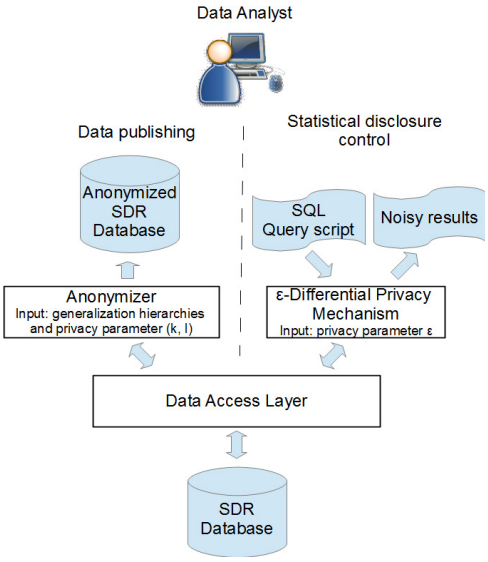


Fig. 2: Overview of the system architecture

research, we obtained datasets from universities and high schools that resemble the schema in Fig. 1. Sample SDRs are given in 3a. Notice that we write QIs within the vertices of the trees and SAs right outside the vertices.

3.4 System Architecture

General architecture of a PPLA system is outlined in Fig. 2. An LA expert will have to choose between two privacy protection mechanisms: data publishing (further discussed in Sec. 4), and statistical disclosure control (further discussed in Sec. 5).

We placed an abstract data access layer between an SDR database and the outside world, that handles all issues related to accessing and fetching SDRs, e.g., in distributed vs. centralized database environments. [1]

The two privacy-preserving techniques are *data publishing (anonymization)* and *statistical disclosure control*. Data publishing relies on privacy definitions such as k-anonymity [53] and ℓ-diversity [25]. The data curator picks a proper privacy definition and decides on a value for the privacy parameter (i.e., k for k-anonymity and ℓ for ℓ-diversity). Then an anonymizer algorithm accesses the SDR database and transforms certain attribute values in such a way that the output (which is now called an anonymized SDR database) conforms with the privacy definition. Such conformance is assumed to imply that the association between an anonymized SDR and the corresponding data subject is sufficiently broken - an adversary cannot determine, above a confidence threshold, which student an anonymized SDR corresponds to. Therefore the anonymized database can be shared with a data analyst for further processing. In this scenario, the data analyst will obtain a modified but truthful version of the original SDR database.

Statistical disclosure control techniques restrict direct access to data. The data analyst can only access the database through a disclosure control layer. The state-of-the-art mechanism for this purpose is based on ε-differential privacy [9], [10]. ε-differential privacy ignores queries that fetch non-statistical data from the database. Statistical queries such as the count, minimum, maximum or average of groups of SDRs that satisfy a predicate condition are answered. The true answer to these statistical queries are protected against privacy disclosures through the addition of random noise.

An important question is whether the system architecture we define here can be supported by existing commercial software. The data model and system architecture we assume are quite generic and compatible with many technologies. From a databases point of view, the advent of NoSQL databases have greatly helped storage of unstructured and semi-structured data. Markup languages such as XML and JSON are also prime candidates to represent and manage hierarchical data objects. These can be readily used to store SDRs. From a learning analytics point of view, there exist standards, e.g., Caliper and xAPI, that record students' data in a Learning Record Store (LRS). LRSs are data stores that serve as repositories holding learning records. Relevant works discuss guidelines in selecting which LRS to use and the analytics that can be performed on that LRS [1], [21]. LRSs can communicate learner data with other systems. Then, in Fig. 2 we can replace the SDR database with an LRS, and program the data access layer to fetch learners' records. We note that some standards (e.g., xAPI) already use JSON, which makes it easy to transfer data between an LRS and a PPLA system.

We note that interoperability with existing LRSs is more of an issue for statistical disclosure control, where a privacy layer must sit between a database and the analyst. The implementation of this would be LRS-dependent. On the other hand, in data publishing, a straightforward method is to move the desired data to a trusted location in a desired format, run the anonymizer, and then publish the results.

Also, depending on the setting and the choice of LRS, data can be kept in a centralized manner or distributed across multiple servers. For instance, the server at the university's registrar's office may hold all demographic information related to students, and departmental servers may hold students' courses and grades. These can be merged later using explicitly identifying information (e.g., student IDs) on demand. Furthermore, data from multiple institutions and LRSs can be merged as in PAR [19]. In such cases, the existence of a common standard across these institutions would be beneficial, but we must account for a certain degree of difference and freedom. Hence we choose to stick with the abstract representations of SDRs rather than focusing deeply on one technology.

4 Privacy Through Anonymization

We first study the *data publishing* scenario. In this scenario, the data curator *anonymizes* the dataset and then shares the anonymized data with the data recipient. After the data is shared, the data curator has no control over what the recipient decides to do with the data.

Anonymization refers to the data privacy approach that seeks to hide the identity and/or sensitive information of data subjects. That is, a data recipient armed with certain

1. If SDRs are stored in a centralized relational database management system, there is no need for the data access layer.

background knowledge should not be able to infer (with high confidence) either the SDR or the sensitive information of one or more subjects. We outline the prominent ways of defining background knowledge and adversarial inference below. In any case, anonymization involves a trade-off between data utility and privacy: if absolutely no inferences can be made using a published dataset, then the dataset is essentially useless for LA purposes. For example, if the school owner publishes completely random data about students, students' privacy is perfectly preserved but an LA expert can harvest no useful relationship from the data. If the published data is very specific, though, it helps not only the LA expert to build accurate models but also the adversary to make inferences.

4.1 Defining Adversaries and Privacy Notions

Adversarial Background Knowledge. In anonymization, it is vital to properly define the background knowledge (i.e., the power) of the adversary. The literature [13] assumes that an adversary has knowledge that his victim's record will be in the published dataset and complete knowledge regarding the QIs of his victim. The adversary also knows all links (i.e., edges) in his victim's SDR, e.g., Alice submitted an evaluation of 8/10 for course CS201, not for CS301. On the other hand, adversarial background knowledge is limited to QIs, and does not include SAs. The privacy notions we will soon define will cover negative knowledge (e.g., Alice did not take CS205) as well as positive knowledge (e.g., Alice took CS201).

Despite these widely accepted assumptions, educational data and learning analytics present unique challenges, some of which are discussed below:

- The adversary could be an *insider*, e.g., a course instructor. Although the data privacy literature assumes that an adversary has no knowledge regarding SAs of data subjects, an instructor will know what grades she gave to each student. This makes the course instructor a stronger adversary.
- Knowing that a student has faced *some* disciplinary action (or, failed *some* class) is sometimes more important to an adversary than knowing *which* disciplinary action (or, *which* class) it was. Although anonymization can protect against the question of *which*, it cannot guarantee hiding the fact that *some* event has happened.
- Having multiple records per data subject in a database is complicating, since these records are often correlated. E.g., a student may attend two universities, and the Ministry of Education collects data from these two universities. Then the ministry should pre-process the data by merging SDRs that belong to the same student before anonymization, a step that can easily be overlooked.
- Continuous and sequential releases based on anonymization are problematic. Assume that the government publishes educational data every two years, e.g., in 2012 and 2014. Consider Bob, a sophomore student actively enrolled in college during the 2012 release. Bob will have taken more classes by 2014. The 2012 release will contain Bob's information thus far. The 2014 release needs to contain data from before 2012, which has already been included in the 2012 release. If the overlapping data between the two releases is anonymized differently, this can lead to identity disclosure [58].

Privacy Definitions. The goal of anonymization is to transform a dataset to enforce a certain definition of privacy. We now survey the literature for the prevalent definitions of privacy.

k**-anonymity** is the most popular definition, and states that each record in the published dataset needs to be indistinguishable from at least $k-1$ other records with respect to QI values [53]. QI-wise groups of indistinguishability are called **equivalence classes**. The main criticism of k-anonymity is that it does not consider the distribution of sensitive values [56], e.g., all records in an equivalence class may contain the same sensitive value. For records in such equivalence classes, the adversary can infer a sensitive value with 100% confidence. Two notions were developed to address this issue: ℓ**-diversity** asserts that every equivalence class should contain ℓ *well-represented* values for each SA [25]. (There can be different interpretations of *well-represented*. A widely accepted definition is to bound the frequency of a sensitive value in an equivalence class by $1/\ell$ [59].) t**-closeness** asserts that the distance between the distribution of sensitive values in an equivalence class and the whole data should differ by no more than a threshold t [22]. Finally, **anatomy** preserves privacy by disassociating QIs and SAs, and releasing them in separate datasets [60].

We illustrate these privacy notions on tabular educational data, in Table 1. The data curator wishes to publish students' grades in a particular class, where the attributes *age*, *gender* and *major* are QIs, and *grade* is sensitive. The original dataset T is given in Table 1a. A 2-anonymous version and a 3-anonymous version are given in 1b and 1c, respectively. Each equivalence class is highlighted using a different color. Neither 1b nor 1c are 2-diverse. In both, the first equivalence class violates 2-diversity. (E.g., from Table 1c, an adversary knowing that Bob is a 21 year-old male Computer Science student can infer that Bob is in the gray equivalence class. Thus, he concludes with probability $2/3$ that Bob got an A-.) 1d is, however, 2-diverse. Anatomy is used in 1e, and data is divided into two tables: one for the QIs and one for the SA. For each sensitive value, only a count is given. Unlike the previous models, there is no explicit link between a record and its SA.

4.2 Extensions to SDRs

The privacy notions given so far have been developed for tabular data anonymization. However, they have applicability outside that domain. (See [13] for a survey.) Next, we will discuss their application to SDRs (i.e., the data model we define in Section 3.3). We give formal mathematical definitions in [36], whereas here we present verbal explanations and intuition.

We say that two SDRs R_1 and R_2 are QI-isomorphic if they share the same structure (vertices and edges) and QIs (labels within vertices). This definition of isomorphism is analogous to the definition of *tree isomorphism*, which requires two trees to appear the same. The only exception in

TABLE 1: Anonymizing tabular educational data

(a) Private dataset T

Age	Gender	Major	Grade
21	male	Computer Science	A-
22	male	Computer Science	A-
19	male	Electrical Engr.	B+
19	female	Industrial Engr.	C
20	female	English	A
23	female	Art History	B

(b) 2-anonymous T

Age	Gender	Major	Grade
20-25	male	Computer Science	A-
20-25	male	Computer Science	A-
19	*	Engineering	B+
19	*	Engineering	C
20-25	female	Arts & Humanities	A
20-25	female	Arts & Humanities	B

(c) 3-anonymous T

Age	Gender	Major	Grade
15-25	male	Science & Engineering	A-
15-25	male	Science & Engineering	A-
15-25	male	Science & Engineering	B+
15-25	female	*	C
15-25	female	*	A
15-25	female	*	B

(d) 2-diverse T

Age	Gender	Major	Grade
15-25	*	Science & Engineering	A-
15-25	male	Science & Engineering	A-
15-25	male	Science & Engineering	B+
15-25	*	Science & Engineering	C
20-25	female	Arts & Humanities	A
20-25	female	Arts & Humanities	B

(e) Publishing T via anatomy

Age	Gender	Major	Group-ID
21	male	Computer Science	1
22	male	Computer Science	1
19	male	Electrical Engr.	1
19	female	Industrial Engr.	2
20	female	English	2
23	female	Art History	2

Group-ID	Grade	Count
1	A-	2
1	B+	1
2	A	1
2	B	1
2	C	1

our case are the SAs: we place no restrictions on them (yet). We say that a set of larger than k SDRs is k^{SDR}-anonymous and forms an equivalence class, if all records in the set are pairwise QI-isomorphic. In other words, all records look the same in terms of structure and QIs.

As outlined in the previous section, an inherent shortcoming of k^{SDR}-anonymity is its ignorance towards SAs. We therefore propose ℓ^{SDR}-diversity. We say that an equivalence class of SDRs is ℓ^{SDR}-diverse if for all vertices in that equivalence class that look the same, the frequency of occurrence of a sensitive value is at most $1/\ell$. In other words, ℓ^{SDR}-diversity takes the definition of ℓ-diversity for tabular data, and applies it to all vertices of SDRs.

A published dataset of SDRs is k^{SDR}-anonymous [ℓ^{SDR}-diverse] if all records in the dataset belong to a k^{SDR}-anonymous [ℓ^{SDR}-diverse] equivalence class. The literature stops at ℓ-diversity, i.e., there is no work that applies t-closeness to SDRs, or tree-structured data in general. In our ongoing research we would like to extend the likes of t-closeness and anatomy to SDRs.

We explain the rationale behind k^{SDR}-anonymity and ℓ^{SDR}-diversity using examples. In Fig. 3a, we present two SDRs, where *major*, *birth year* and *gender* are QIs at the root vertex, each class is drawn as a child of the root vertex, and evaluations for classes are drawn as children of the corresponding classes. In course evaluations, the instructor is treated as the QI and his/her evaluation score (for simplicity, we assume that this is an integer out of 10) is treated as the SA. If the records in Fig. 3a are published without anonymization, any of the following pieces of adversarial background knowledge may cause the adversary to distinguish one record from the other: (1) The victim was born in 1995. (2) The victim took CS305. (3) The victim took CS201 from Prof. Harry. (4) The victim took three classes. (1) demonstrates that attacks due to demographic information are possible, similar to attacks on tabular data. (2), (3) and (4) demonstrate attacks that are unique to SDRs. (2) and (3) show that QI information that is not only located *within* the root vertex but also *connected* to it may cause leakage. (4) shows that an adversary may use *structural* knowledge (with or without QIs) to perform attacks. In this particular case, knowing the number of classes a student took will yield his SDR. Notice that information that is shared by both SDRs (e.g., victim is majoring in Computer Science, or victim has evaluated Prof. Bloggs in *some* class) is not sufficient to distinguish records.

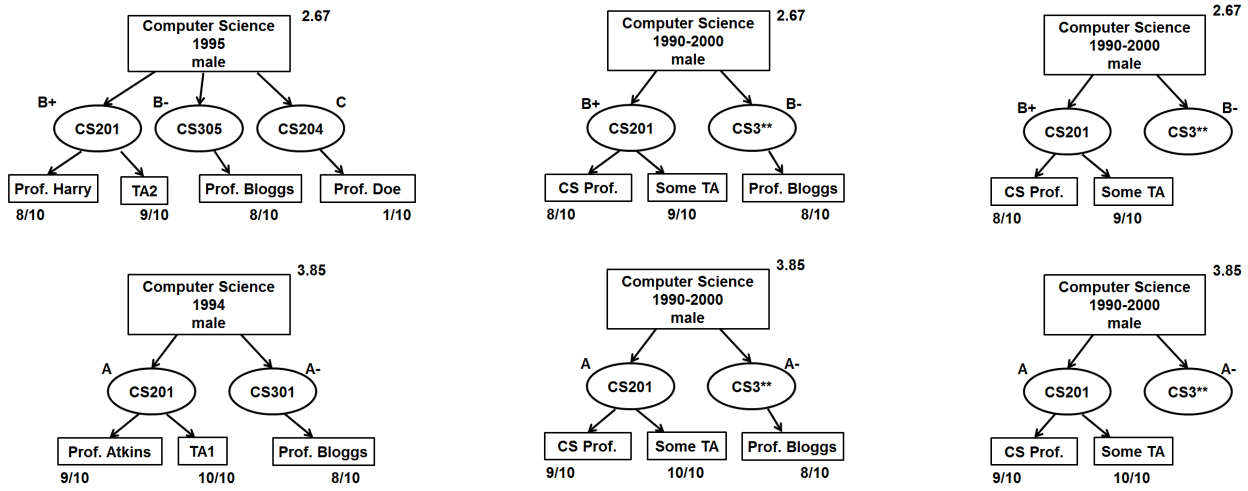
k^{SDR}-anonymity solves the privacy problems above by providing indistinguishability with respect to QIs *and* structure. For example, the SDRs in Fig. 3b are 2-anonymous, and no background knowledge of QIs or structure may help the adversary distinguish these records. This is because k^{SDR}-anonymity either invalidated or obfuscated the adversary's background knowledge. For example, by removing the fact that the first student took CS204, both students are shown having taken 2 classes, and the (4)th attack is now invalid. For attacks (1), (2) and (3), there are now 2 SDRs that satisfy these constraints, e.g., knowing that the victim was born in 1995 no longer singles out the first record because both students' SDRs say that they were born between 1990-2000.

ℓ^{SDR}-diversity builds on k^{SDR}-anonymity and enforces diversity for *every* vertex in an equivalence class. A privacy problem in Fig. 3b, for instance, is that both students gave Prof. Bloggs 8/10. The adversary does not need to distinguish one record from the other to infer this sensitive information, due to lack of diversity. (An interesting note: the adversary's background knowledge does not even have to include that these students evaluated Prof. Bloggs in order to infer this sensitive information. Attacks (1), (2), (3) could as well be sufficient.) The ℓ^{SDR}-diversity definition stops such inferences by making *every* vertex diverse, e.g., as in Fig. 3c. We do understand, however, that in some scenarios diversity in *some* vertices might be needed, but not others. (E.g., the data curator decides GPAs and course grades should be diverse but lack of diversity in evaluation scores is okay.) Then, the definition of ℓ^{SDR}-diversity can be modified to specify those vertices where diversity must be enforced, and leave others unattended.

It is not a good idea to convert SDRs to tabular format and then run tabular k-anonymity and ℓ-diversity algorithms on them, since this approach leads to serious privacy problems. Due to space constraints we omit further discussion on this topic, and refer the interested reader to [31] and [36].

4.3 Tools for Anonymization

Given a database of SDRs, one aims to produce a k^{SDR}-anonymous/ℓ^{SDR}-diverse version with as little modification as possible, so that the data is authentic and accurate, but also ensures an adequate level of privacy. The choice of



(a) Two SDRs prior to anonymization (b) 2-anonymous SDRs (c) 2-diverse SDRs

Fig. 3: Anonymization of tree-structured student data records

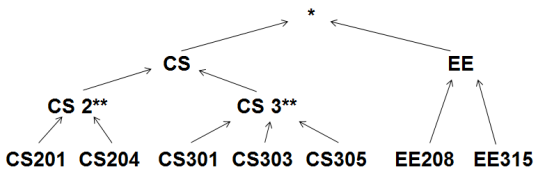


Fig. 4: Generalization hierarchy for courses

privacy notion and the exact values for k or ℓ are left to the data curator (or dictated by the social norms, requirements of the data recipient etc.). Higher k and ℓ offer better protection, but are harder to satisfy.

The two conventional ways to modify datasets to achieve privacy are *generalization* and *suppression* [13]. Generalizations replace specific values by more general ones, e.g., course ID "CS305" can be replaced by "CS 3rd year" or "CS3**". Generalizations occur with the help of a *generalization hierarchy*, where data values that are semantically closer are grouped together and generalized to the same value. A small generalization hierarchy is given in Fig. 4. Generalization hierarchies can be inputs to the anonymization procedure (i.e., specified by the data curator) or inferred automatically from the data (e.g., when the data is numeric). Suppressions conceal information by deleting it: information that exists in the original data is removed from the final output. The deletion of CS204 in Fig. 3b is an example.

Both generalizations and suppressions incur a *cost*: they decrease the potential utility of the LA and data mining techniques. For instance, if all third year CS classes are generalized to "CS3**", the data miner can measure student success in 3rd year CS classes, but not in individual classes. Also, in Fig. 3b, by suppressing CS204, we lose the useful information that this student did poorly in CS204 potentially because he did not like his instructor Prof. Doe. This information could have been harvested via the application of appropriate learning analytics methods.

We would like to note that in cases where answering detailed questions is necessary (e.g., "How many students took CS301?") but the data is too generalized to answer such questions (e.g., all CS 3rd year classes have been generalized to "CS3**", therefore no "CS301" classes are directly observed in the output) one can make use of *data reconstruction* [32]. In data reconstruction, we assume that the *actual* value of an *observed* value in the output is equally likely to be any of the leaves that lie under that *observed* value in the generalization hierarchy. For example, "CS3**" could be any of CS301, CS303 or CS305 according to Fig. 4. In the presence of additional statistics (e.g., how often each class is taken) each candidate can be given an appropriate weight instead of assuming that they are equally probable.

Finally, an important advantage of anonymization is that fake or noisy information is not added to the released data. Apart from the uncertainty and analytical utility loss discussed above, anonymization preserves truthfulness. Noise can be a significant drawback when dealing with sensitive information, e.g., if we add bogus information to Alice's SDR that she has faced a disciplinary problem where she actually has not, this can have serious consequences for her.

5 Privacy Through Statistical Disclosure Control

Next, we study the *privacy-preserving data mining* scenario, where data is kept by the data curator (in un-anonymized form) and never published. Instead, the data analyst receives a privacy-preserving interface, using which he can run various statistical analyses and learning analytics. We emphasize that this is different than access control, in the sense that the data analyst can run any statistical function/query on the data he desires, i.e., his access is not limited to certain portions of the data. But, the answers returned by the database incorporate subjects' privacy and can hence be limited and noisy.

The state-of-the-art method in statistical disclosure control is *differential privacy* [9], [10]. In this model, only statistical queries are allowed, and answers to these queries are perturbed with (random) noise. Unlike anonymization where properly defining adversarial knowledge and the privacy notion is required, differential privacy protects against

all types of background information (e.g., attributes need not be divided into QI/SA etc.). In that sense, some of the problems discussed in Section 4.1 can be avoided using this technique. In the next section, we briefly introduce the basics of differential privacy.

5.1 Differential Privacy

In cryptography, some encryption mechanisms can guarantee that the ciphertext (i.e., encrypted data) reveal no information at all about the plaintext (i.e., raw data). This notion is called "semantic security". Dwork proves in [9] that a guarantee similar to "semantic security" cannot be achieved in disclosure control. That is, any means of access to a database that contains sensitive data automatically implies a (non-zero) risk of disclosure. Additionally, revealed information may not even pertain to individuals/data subjects whose data is in this database. Even if one does not participate in the database as a data subject by providing his/her record, he or she is still under risk of disclosure. Consequently, a disclosure control mechanism should not be after preventing privacy leaks. ε-differential privacy instead tries to control/bound the risk of disclosure.

Consider a single individual, say Alice, who is trying to make a decision on participating in a database. Let the database version that contains (resp. does not contain) Alice's record be referred to as $\mathcal{D}$ (resp. $\mathcal{D}'$). All such $\mathcal{D}$ and $\mathcal{D}'$ will differ in only one record and any such pair of databases are called *neighboring databases*. Differential privacy tries to bound the probability that the response to a query set/algorithm remains the same in the two worlds for all possible data subjects including Alice[2]. Def. 1 formalizes this notion.

Definition 1 (ε-Differential Privacy (ε-DP)). *A randomized algorithm $\mathcal{A}$ gives ε-DP if for all neighboring datasets $\mathcal{D}$,$\mathcal{D}'$ and for all possible outcomes of the algorithm $S \subset Range(\mathcal{A})$,*

$$Pr[\mathcal{A}(\mathcal{D}) \in S] \leq e^{\varepsilon} \times Pr[\mathcal{A}(\mathcal{D}') \in S]$$

where the probabilities are over the randomness of $\mathcal{A}$.

In the definition, $\mathcal{A}$ includes the disclosure control layer which is supposed to satisfy ε-differential privacy. S is a transcript - a vector of outputs, e.g., statistical query results. The probability that $\mathcal{A}$ produces the output S on $\mathcal{D}$ and $\mathcal{D}'$ is bounded by e^{ε} for all possible S.

Differential privacy assumes that the output of an algorithm does not overly depend on one record. In other words, there is a significant probability (controlled by parameter ε) that the same result could have been obtained if the algorithm was run on a neighboring database. If an algorithm produces the same outcome with and without Alice's SDR in $\mathcal{D}$, then including Alice in $\mathcal{D}$ does not bear any privacy risk for her. This gives a stronger incentive for data sharing and data subjects can be reassured that they are safe. ε is often regarded to be small, e.g., $\varepsilon = 0.1, \ln 2$.

A typical way to achieve ε-DP is to model the learning analytics task as a function (represented in terms of a set of statistical queries). Then, the true answer of this function (i.e., the queries) is computed. Finally, random noise is added to the true answer. The amount of noise depends on the privacy parameter ε and the *sensitivity* of the function.

Definition 2 (Sensitivity). *Let $f : \mathcal{D} \rightarrow \mathbb{R}^d$ be a function that maps a database into a fixed-size vector of real numbers. The sensitivity of f is defined as:*

$$\Delta f = \max_{\mathcal{D},\mathcal{D}'} ||f(\mathcal{D}) - f(\mathcal{D}')||_1$$

for all neighboring databases $\mathcal{D}$,$\mathcal{D}'$, where $||.||$ denotes the L_1 norm.

In the definition, f is the function that models the data analytics task. The domain of the function is a data set $\mathcal{D}$ and its range is a d-dimensional vector of real numbers (the value d depends on the number of queries in f). Δf is computed on all possible neighboring database pairs. Based on this definition, sensitivity measures the maximum difference in the output of f that can be caused by changing one record in the database.

For example, the sensitivity of computing the answer to: "how many students are majoring in CS?" is 1, because a change in one record either (i) increases the count of CS majors by 1, e.g., by making a non-CS major a CS major, (ii) decreases the count by 1, e.g., by making a CS major a non-CS major, or (iii) does not change the count of CS majors. Computing the answer to: "f: what is the minimum GPA observed?" has sensitivity 4, because there may exist a database $\mathcal{D}$ where all students have perfect GPAs (i.e., 4.0) where $f(\mathcal{D}) = 4.0$. But, by modifying one student's GPA and making it 0.0 in $\mathcal{D}'$, it is possible to obtain $f(\mathcal{D}') = 0.0$. Hence, $\Delta f = 4.0 - 0.0 = 4.0$.

The output of a statistical query can either be numeric (i.e., real or integer-valued) or categorical (i.e., discrete valued). For example, the query: "how many students took CS301?" is numeric, whereas the query: "which CS class was taken the most?" is categorical. For each of these, there exist simple mechanisms that satisfy differential privacy.

Laplace Mechanism [10]. For numeric queries, the Laplace mechanism generates ε-DP outputs as follows: given a dataset $\mathcal{D}$ and a function f, it first computes the true output of the function $f(\mathcal{D})$, and then perturbs the true output by adding noise. The noise is sampled from a Laplace distribution with mean 0 and scale $\Delta f/\varepsilon$. That is, the mechanism that returns the noisy answer $f(\mathcal{D}) + Lap$, where Lap is noise, is ε-DP.

Exponential Mechanism [28]. The exponential mechanism can be used for statistical queries with a categorical answer. It is useful for selecting a discrete output r from a domain R in a differentially private manner. For this purpose, the mechanism employs a utility function (i.e., quality criterion) q that associates each output $r \in R$ with a probability of being selected. This probability should be non-zero for each $r \in R$.

The exponential mechanism first computes the sensitivity of the quality criterion, Δq. Then, it computes the quality score of each output on the database $\mathcal{D}$, $q(\mathcal{D}, r)$. Finally, instead of deterministically choosing the output with highest $q(\mathcal{D}, r)$, it probabilistically samples one $r \in R$, where the probability of being chosen for each r is proportional to $e^{\frac{\varepsilon q(\mathcal{D},r)}{2\Delta q}}$.

2. The probability is bounded from above and below, since Def. 1 is symmetric for $\mathcal{D}$ and $\mathcal{D}'$.

For example, assume that each student can take a class once, and let "Which CS class was taken the most?" be the statistical query we would like to answer using differential privacy. The possible outputs are $R = \{\text{CS201}, \text{CS204}, \text{CS301}, \text{CS303}, \text{CS305}\}$. We can define the quality criterion q to be "the number of students that took the class". Then, Δq would be 1. For each candidate, we would first compute $q(\mathcal{D}, r)$ by querying the database for the true counts, and then compute $e^{\frac{\varepsilon q(\mathcal{D},r)}{2\Delta q}}$. Finally, the exponential mechanism returns one candidate at random, where the probability of being selected for each candidate is proportional to its $e^{\frac{\varepsilon q(\mathcal{D},r)}{2\Delta q}}$ score.

Composition Properties [27]. The two mechanisms above compute the output of one statistical query. A learning analytics task, however, requires repeated application of several statistical queries. The composition properties of differential privacy, namely *sequential composition* and *parallel composition*, enable the data analyst to run multiple queries in parallel or in succession. Therefore, any complicated learning analytics task that can be broken down into a query-by-query "recipe" can be implemented in a differentially private manner. Parallel composition is applied in cases where a statistical query partitions the database in disjoint sets of records (i.e., changing a single record does not affect multiple answers). Sequential composition is applied when a successive query's query region has an overlap with the previous query.

5.2 Critique of Differential Privacy

Differential privacy obviously has advantages. It is not limited to adversarial background knowledge, and can hence defend against adversaries that are much stronger than those in the anonymization literature. The adversary does not see the bulk of the data, so many of the issues and privacy threats described in Section 4.1 are no longer of concern. Also, the amount of information disclosure can be theoretically bounded. Because of these, some academics argue that differential privacy makes anonymization-based privacy obsolete. However, several shortcomings of differential privacy have also been recognized [2]. Some of these shortcomings, in the context of learning analytics, are:

- Differential privacy is based on noise addition. As argued at the end of Section 4.3, adding noisy or false information might cause problems in certain sensitive situations.
- Since differential privacy is a relatively new notion, a lot of useful data mining methods that can be used in LA have not yet been implemented using differential privacy (e.g., some clustering algorithms, association rules, unsupervised learning). However, all types of queries and LA methods can be run on an anonymized and published dataset.
- Differential privacy is used via statistical queries. An LA task must be expressed in terms of statistical queries so that it can be run via a differentially private interface. For a not tech-savvy user, expressing a high-level LA task as a collection of low-level data queries might be difficult.
- The choice of ε is an open question. k-anonymity and ℓ-diversity have semantic meanings that can easily be interpreted and enforced. However, due to the probabilistic nature of ε-DP, there is no clear link between ε and the observed output. Even a fine-tuned ε might not guarantee a semantic privacy requirement.
- Application of ε-DP to tabular data is more straightforward and well-studied, but an off-the-shelf application of ε-DP to SDRs does not exist (yet).

6 Experiments and Discussion

In this section we provide privacy-preserving proof-of-concept implementations of two LA tasks and illustrate the trade-off between the utility of these methods (measured by their accuracy) and data subjects' privacy. We used the Java programming language for implementation, and ran the experiments using a laptop with commodity hardware.

Datasets. We obtained two datasets from two different universities in Turkey. Both datasets share a similar schema to that in Fig. 1 with minor modifications (e.g., in one dataset, instead of the students' year of birth, their age was reported). We name the datasets *synthetic* and *real* for reasons described next.

For the synthetic dataset, we obtained data regarding students from a university's Computer Science undergraduate program. The data contained 30 students and their grades in introductory-level classes and some upper-level classes. Using this sample and according to the university's graduation requirements, we generated a dataset of 1000 students. We simulated GPA values using a normal distribution, where the mean and the standard deviation were determined by the GPA scores of our sample. According to the university's graduation requirements, we ensured that all students took the obligatory and introductory-level classes. To each student, we randomly assigned a fixed number of classes from the pool of core classes, and a varying number of technical area electives. Students' grades in each class were determined by their GPA, the type of class and the distribution of grades in that class in our sample (depending on availability - we had a distribution for all introductory classes but only some of the upper-level classes).

The real dataset contains 3162 students from another university, majoring in different areas. Compared to the synthetic dataset, it has a wider array of classes and more discrepancy between the number of classes students take (e.g., there exist students that took only 1-2 classes, as well as students that took 60 classes). The real dataset was used as is, apart from trivial pre-processing (e.g., removal of duplicate records).

Experimental Setup. We implemented privacy via anonymization and statistical disclosure control separately. We used the anonymization algorithms in [31] and [36] to obtain k^{SDR}-anonymous and ℓ^{SDR}-diverse databases respectively. Both algorithms are for handling hierarchical/tree-structured data, and are therefore directly applicable to SDRs. Algorithms were run using the parameters suggested by their authors. When needed, data was reconstructed probabilistically (see Section 4.3) without assuming the existence of additional statistics.

For statistical disclosure control, we implemented a differentially private interface to run the learning analytics

tasks. We made a simplifying assumption that each class is taken only once. The datasets we had were already modified such that if a student took a class more than once, only the highest grade would be retained and the others were dropped. Hence, this is a reasonable assumption. Also, differentially private real-valued answers to integer-valued functions (e.g., count queries) were rounded to the nearest non-negative integer, wherever applicable.

Data reconstruction after anonymization and noise addition for ε-DP are probabilistic. Therefore each experiment was conducted 200 times and results were averaged.

6.1 Query Processing

One of the main goals of LA is personalizing and adapting the learning process and content, ensuring that each student receives resources and teaching that are appropriate to their knowledge state [42]. Instructors should aim to provide learning opportunities that are tailored according to students' background and needs. For example, students' background, current performance and interests can be used to recommend courses or majors, so that every student gets the most out of their education.

Most universities offer a similar set of classes for their first year students. For example, all engineering majors take courses in calculus, physics, biology etc. Learning analytics can be used to recommend suitable majors for students based on their performance in these introductory classes. After all, certain classes are more relevant for certain majors, e.g., a biology course is probably not very useful for a Computer Science student, but it is crucial for a Biochemistry student.

A simple recommendation system can be built by finding correlations between students' success in a major and their grades in first-year classes. Then, for a new student, given his/her grades thus far, one can determine their suitability for different majors: what is the probability that Alice becomes a successful Biochemistry major, given that her grade in introductory biology (BIO101) was low? To be concrete, let us model the question as follows:

$$Pr(\text{major = Biochem} \cap \text{GPA} \geq 3.50 \mid \text{BIO101} \leq \text{C+})$$

$$= \frac{Pr(\text{major = Biochem} \cap \text{GPA} \geq 3.50 \cap \text{BIO101} \leq \text{C+})}{Pr(\text{BIO101} \leq \text{C+})}$$

$$= \frac{\text{COUNT the \# of Biochem majors in the database w/ GPA} \geq 3.50 \text{ and BIO101} \leq \text{C+}}{\text{COUNT the \# of students w/ BIO101} \leq \text{C+}}$$

As shown above, a computation of the recommendation system can be reduced to statistical database queries (e.g., count queries). The parameters can easily be changed, i.e., one can repeat the process above for different majors, for different GPA thresholds and one or more first-year classes. (Note that here the "$\leq$" sign, when used with course grades, should not imply lexicographical ordering but rather be interpreted as "worse than".)

In the first experiment, we generated queries in the form above, and ran them on the original data, k^{SDR}-anonymous and ℓ^{SDR}-diverse data, and via ε-DP; for various values of k, ℓ and ε. We ran a total of 90 queries on the synthetic dataset and 50 queries on the real dataset. We measured the *average relative error (AvRE)* of these queries as follows: let X_i be the answer obtained when the ith query is issued on the original data and Y_i be the answer obtained when the same query is issued after a privacy definition is enforced. The AvRE of N queries is:

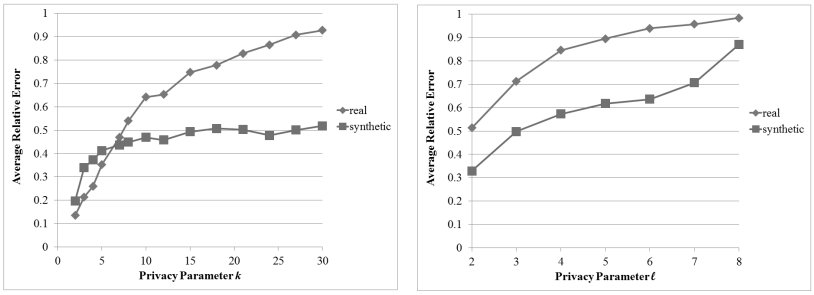
$$AvRE = \frac{\sum_{i=1}^{N} \frac{|Y_i - X_i|}{X_i}}{N}$$

For k^{SDR}-anonymity and ℓ^{SDR}-diversity, we obtained the results in Fig. 5. As expected, the amount of error increases as privacy requirements get stricter, i.e., k and ℓ increase. An interesting observation is that there is a crossover between the error curves of the synthetic and real datasets in Fig. 5a, when k=6. We believe that this is because of the content of the two datasets: since all students in the synthetic dataset are Computer Science majors, they have taken more or less the same classes and have similar SDRs. Therefore it is easier to find k students to group together. On the other hand, when k is large, it is difficult to find k similar students in the real dataset. This can be a factor when choosing an appropriate k for anonymization.

We observe that ℓ^{SDR}-diversity often causes higher AvRE than k^{SDR}-anonymity. Even k=12 has less error than ℓ=3. We remind the reader that k^{SDR}-anonymity is a prerequisite for ℓ^{SDR}-diversity, hence this is an expected result. An important factor in ℓ^{SDR}-diversity is the amount of diversity in sensitive values: say that the MATH101 instructor decides to grade very generously and gives everyone an A or A- in the class. Then, 3-diversity for MATH101 cannot be achieved no matter the algorithm, due to the simple fact that there is no 3rd sensitive value for MATH101 in the database. Generalizations and suppressions will have to be performed, which potentially introduce additional error.

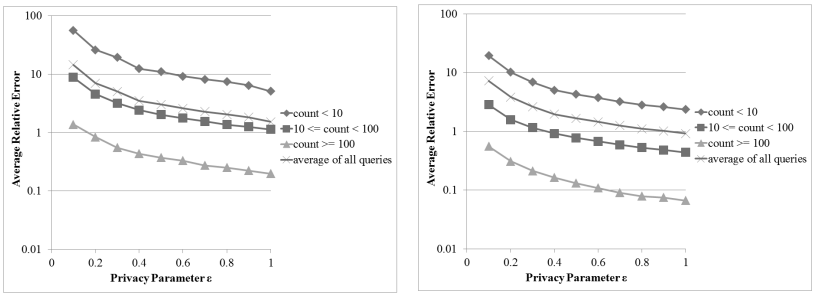
We issue the same queries on the original data via an ε-DP interface, and report the AvRE in Fig. 6 for the synthetic and real datasets separately. Notice that in these figures, the y axes are in logarithmic scale. We divide our queries into 3 categories: queries with small answers (i.e., less than 10), queries with large answers (i.e., greater than 100) and queries in between. We compute the AvRE and draw an error curve for each category, and finally one curve for the average of all queries. This is to emphasize the relationship between a query's answer and its relative error (relevant for ε-DP but not anonymization). Also, in differential privacy, higher ε permits higher amount of information disclosure, and hence provides less privacy. As expected, when ε is increased in Fig. 6, privacy requirements are relaxed, and this causes a decrease in AvRE.

As explained in Section 5, ε-DP adds noise to the output of each query to satisfy privacy. The noise is sampled from a distribution that does not depend on the true answer of the query. If the true answer is small relative to the noise added, then the error rate is high. If the true answer is large and the same amount of noise is added, then the error will be negligible. Our experiments clearly demonstrate this. With $\varepsilon \geq 0.5$ and queries with large answers, we obtain AvRE close to 0.1, which is significantly better than anonymization. But for all other queries (or for $\varepsilon < 0.5$), the



(a) k^{SDR}-anonymity (b) ℓ^{SDR}-diversity

Fig. 5: Query processing on anonymized data



(a) Real dataset (b) Synthetic dataset

Fig. 6: Query processing via ε-DP

error obtained from ε-DP is much higher (at least 2-3 times higher) than the error in anonymization.

6.2 GPA Prediction

Another interesting task in LA is to build predictive models that characterize current student performance and help forecast future success/failure. Data mining methods, especially supervised learning and classification, are useful for building predictive models.

Some courses are good indicators of a student's success in college. For instance, it is very uncommon that a student does poorly in calculus but eventually graduates as a math major with a stellar GPA. If one can successfully predict students' success based on such classes, this can serve as an indicator (or warning) that a student will do well (or have trouble graduating) from their choice of major. This scenario can also go hand in hand with the scenario in the previous section. One can predict a student's GPA in different majors using his/her first-year courses, and advise the student to choose a major that will maximize their success.

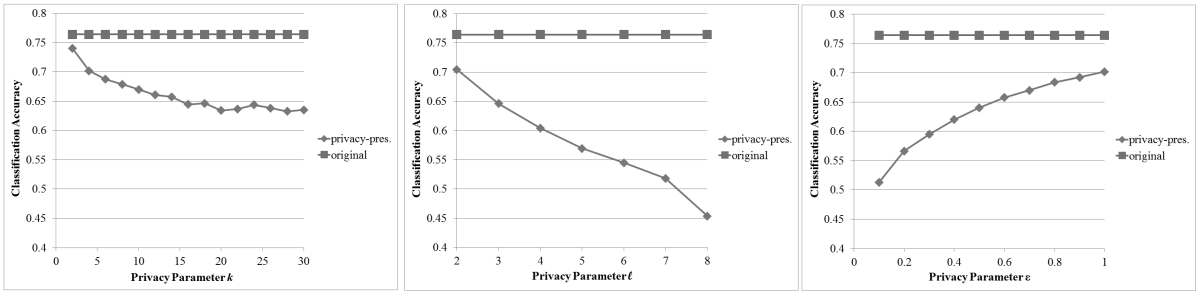
To build a predictive model, a data analyst needs to access the data. Instead of granting access to the original data, a privacy-preserving method is to anonymize the data first and then grant access. Alternatively, the model can be built through a differentially private interface, which will add noise to the model every time a data access occurs. For example, there exist differentially private algorithms for Naive Bayes and decision tree classification [12], [57].

In this experiment, we used a Naive Bayes Classifier (NBC) to implement a predictive model. We chose to use NBCs because they are easy to implement using differential privacy [57], and they are widely accepted and used as baseline classifiers. A recent study in learning analytics has also used NBCs to predict academically at risk students [20].

The goal of the NBC is to predict students' GPA at the end of their fourth year based on their performance in introductory classes. We ran this experiment only on the synthetic dataset because the real dataset did not have a common set of classes that every student needs to take. Among the classes in the synthetic dataset, we chose three first-year and two second-year obligatory classes as predictors. We discretized the GPA range [0-4] into four groups: 0.0-0.99, 1.0-1.99, 2.0-2.99 and 3.0-4.0. We used 5-fold cross validation when building and testing the classifier.

We first build the classifier on the original data, without anonymization or ε-DP. The accuracy of the classifier turns out to be 76%, which is reasonably high. (In contrast, the accuracy of a classifier that outputs random results would be around 25%. A more informed classifier that takes into account the mean and standard deviation of students' GPA could be roughly 40-50% accurate, depending on the standard deviation.) We then build classifiers on k^{SDR}-anonymous and ℓ^{SDR}-diverse data, and using ε-DP. We graph the classification accuracy of these classifiers, with the aim of quantifying the decrease in accuracy after a privacy metric is applied.

We graph the results in Fig. 7. In all three figures, the y-axis shows the ratio of students that were correctly classified, i.e., number of students whose GPA was correctly predicted divided by the total number of students. In Fig. 7a and 7b, there is a steady decrease in classification accuracy with increasing k and ℓ. In the case of k^{SDR}-anonymity, after k=20, classification accuracy stays roughly the same at around 63%. The drop is more significant (approximately linear) in Fig. 7b, where ℓ^{SDR}-diversity is used. Contrary



(a) k^{SDR}-anonymity (b) l^{SDR}-diversity (c) ε-DP

Fig. 7: Classification after privacy-preserving techniques

to anonymization-based methods, there is a positive correlation between the privacy parameter ε and classification accuracy in differential privacy. Again, this is because higher ε implies less privacy.

6.3 Discussion

We start our discussion by re-stating that in both the query processing experiment and the GPA prediction experiment, we witness the obvious trade-offs between privacy and utility. Increasing the level of privacy (i.e., increasing k and ℓ, or decreasing ε) comes at the cost of reduced accuracy and higher error in the LA tasks. In addition, in the majority of the experiments we observe that anonymization methods (i.e., k^{SDR}-anonymity and ℓ^{SDR}-diversity) offer results with higher utility and accuracy than those obtained using ε-DP. Although the privacy guarantees of anonymization and ε-DP are fundamentally different, we believe that this is a significant finding that could guide system designers in choosing one method over the other in cases where both are applicable.

An interesting argument is that we use the anonymized data to classify subjects, and this itself is a privacy violation. First, we note that after our privacy protection methods, personal identifiers are removed, and a certain quantifiable privacy notion is met. If we cannot build accurate classifiers using this data, then the data has become useless. This is undesirable from a LA point of view. After all, data is either most useful or perfectly anonymous, but never both [35], [37]. Second, the classifiers are trained using private data, but are used by students or trusted school officials (on behalf of students) that wish to use the classifier, and for doing so they need to voluntarily feed their private data (e.g., course grades) to the classifier.

A compelling research direction is to implement k^{SDR}-anonymity, ℓ^{SDR}-diversity and ε-DP into existing learning management systems (LMSs). We feel that the technical difficulties in doing so are context-specific (i.e., specific for each LMS and institution). We choose to mitigate these issues by placing a data access layer (see Fig. 2) that collects data from multiple LMSs, merges them into one SDR etc. The abstract SDR representation suits many technologies and therefore maximizes applicability. It can be customized by data owners to suit particular LMSs and data formats.

The choice of a privacy protection mechanism depends on various factors. Although we have discussed and presented experimental results for the *utility* and *accuracy* factors, the *cost*, *scalability* and *performance* factors are also important, as they determine how convenient it is to employ one mechanism over the other.

We start by analyzing the performance of the proposed mechanisms. The performance overhead of ε-DP is low, since ε-DP relies mainly on data-independent noise addition. To return an answer to any query, the system needs to get the true answer of the query from the LMS, which is a step necessary regardless of whether ε-DP is used. Then, noise addition for ε-DP is just a single step of additional computation. On the other hand, the performance overhead for anonymization is often high. For medium-sized datasets (e.g., the ones we experiment on) the current algorithms used for SDR anonymization take considerable time (tens of minutes or hours) [36]. However, anonymization often leads to a one-time data release (i.e., data publishing) which can be done overnight and has no bearing on a real-time system. Continuous anonymization is also possible, but incurs a *delay* not only for the de-identification step but also to increase the utility of its output.

We now comment on the scalability and cost factors of the proposed mechanisms. Algorithms for k^{SDR}-anonymity and ℓ^{SDR}-diversity are quadratic in the number of SDRs and exponential in the height/levels of SDRs. That is, if we want to anonymize 10 times more SDRs, then we have to pay 100 times more computational cost. For larger datasets with thousands of SDRs, the cost becomes large. This is a problem for scalability. Again, since the noise in ε-DP is often data-independent, it is scalable. Furthermore, as many LMSs run in the cloud, the costs for data manipulations are critical. ε-DP fares better in this regard too, as anonymization techniques need to make many data accesses and manipulations before outputting a final result.

There can be cases where anonymization would be preferable to ε-DP in terms of cost and performance overhead, too. Say that there are many data analysts interested in a university's database stored on a cloud, and they all wish to run long LA tasks. The university has two choices: (1) Anonymize and publish the database once, where each analyst receives a copy of the anonymized database. This has a large one-time cost for the university, but since the analysts now have a copy of the database, they no longer need to access the university's cloud. They will perform computations on the published database (which they can store locally). This may yield less computation on the cloud in the long run (and hence less cost and performance overhead for the university). (2) Each analyst is given a ε-DP

access to the cloud, and is forced to retrieve information through this access. In this case the only one-time overhead is to establish a ε-DP interface, which does not require much computation power. But, if the LA tasks are costly and contain a lot of data accesses, the university might be worse off using this strategy.

7 Conclusion

In this paper we studied the application of state-of-the-art privacy-preserving data publishing and mining methods to learning analytics. Despite its detailed and technical discussion, this paper is not meant to be conclusive. Rather, we hope that it sparks interest in the area, especially in applying privacy protection mechanisms to existing learning analytics methods, and to adjust these methods so that the added privacy does not destroy their utility.

Our analysis shows that there are trade-offs between the proposed privacy mechanisms, and there is no single technical solution to the *privacy* problem. Anonymization is easy to understand, extensively studied and applicable to many types of data (e.g., tree-structured SDRs, tabular and graph data). In contrast, ε-DP, the state-of-the-art in statistical disclosure control, offers protection against stronger adversaries and is more scalable; but comes at the cost of utility and convenience. The major issues adversely affecting its convenience are: (1) the need to reduce a data analysis task to a set of low-level queries, and (2) the absence of ε-DP implementations on different types of data. Yet, considering that anonymization has recently come under fire from academics [30], [35], we can expect a shift towards ε-DP; and some of the major issues concerning ε-DP can be solved via implementing a readily available, privacy-integrated tool for LA. This can be an interesting area for future work.

Data privacy is a difficult problem. Despite technical solutions, there are still complexities in defining privacy and inherent limitations of privacy-preserving mechanisms. For example, how do we adequately define adversarial background knowledge for anonymization? Will the data owner's definition be sufficient, or can a stronger adversary be present? Furthermore, students' promiscuity and carelessness in sharing personal information is a risk that cannot be addressed by a privacy mechanism enforced by an institution. For example, people nowadays are happy to share their location data (e.g., location check-ins on Foursquare). Students share each others' posts and information on social media platforms. Are they aware of the privacy implications of these? A learning institution's enforcement of students' privacy means very little if the students themselves are not aware of their privacy. Therefore, we conclude by stating that technical solutions for privacy are most beneficial if there is a common demand from all parties, i.e., academics, practitioners, data and system owners, and students.

References

[1] Berking, P. (2015, November). Choosing a Learning Record Store (LRS). Retrieved from http://adlnet.gov/adlassets/uploads/2015/11/Choosing-an-LRS.pdf.

[2] Clifton, C., & Tassa, T. (2013, April). On syntactic anonymity and differential privacy. In *29th IEEE International Conference on Data Engineering Workshops (ICDEW 2013)*, (pp. 88-93). IEEE.

[3] Crawford, K. (2011). Six provocations for big data. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431.

[4] Crawford, K. (2013). The hidden biases in big data. *HBR Blog Network*, 1.

[5] Crook, M. A. (2013). The risks of absolute medical confidentiality. *Science and Engineering Ethics*, 19(1), 107-122.

[6] Danaher, J. (2014). Rule by algorithm? Big data and the threat of algocracy. *Institute for Ethics and Emerging Technologies*.

[7] Drachsler, H., Dietze, S., Herder, E., d'Aquin, M., & Taibi, D. (2014, March). The learning analytics & knowledge (LAK) data challenge 2014. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 289-290). ACM.

[8] Drachsler, H., & Greller, W. (2016, April). Privacy and analytics: It's a DELICATE issue - A checklist for trusted learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics And Knowledge* (pp. 89-98). ACM.

[9] Dwork, C. (2006). Differential privacy. In *ICALP 2006. LNCS, vol. 4052* (pp. 112). Springer, Heidelberg.

[10] Dwork, C. (2008). Differential privacy: A survey of results. In *Theory and applications of models of computation* (pp. 1-19). Springer Berlin Heidelberg.

[11] EUP. (2002) Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector. European Union: European Parliament.

[12] Friedman, A., & Schuster, A. (2010, July). Data mining with differential privacy. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 493-502). ACM.

[13] Fung, B., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (CSUR)*, 42(4), 14.

[14] Gangadharan, S. P. (2012). Digital inclusion and data profiling. *First Monday*, 17(5).

[15] Griffiths, D., Drachsler H., et al. (2016). Is privacy a show-stopper for learning analytics? A review of current issues and solutions. *Learning Analytics Review 6*.

[16] Gurses, S. (2014). Can you engineer privacy?. *Communications of the ACM*, 57(8), 20-23.

[17] Heath, J. (2014). Contemporary Privacy Theory Contributions to Learning Analytics. *Journal of Learning Analytics*, 1(1), 140-149.

[18] House, T. W. (2012). Consumer data privacy in a networked world. Retrieved April, 13, 2013.

[19] Ice, P., Diaz, S., Swan, K., Burgess, M., Sharkey, M., Sherrill, J., ... & Okimoto, H. (2012). The PAR Framework Proof of Concept: Initial Findings from a Multi-Institutional Analysis of Federated Postsecondary Data. *Journal of Asynchronous Learning Networks*, 16(3), 63-86.

[20] Jayaprakash, S. M., Moody, E. W., Laura, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1), 6-47.

[21] Kitto, K., Cross, S., Waters, Z., & Lupton, M. (2015, March). Learning analytics beyond the LMS: the connected learning analytics toolkit. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 11-15). ACM.

[22] Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE International Conference on Knowledge and Data Engineering 2007 (ICDE 2007)*, pp. 106-115. IEEE.

[23] Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & dApollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research*, 66(4), 423-458.

[24] Macfadyen, L. P., Dawson, S., Pardo, A., & Gasevic, D. (2014). Embracing big data in complex educational systems: The learning analytics imperative and the policy challenge. *Research & Practice in Assessment*, 9.

[25] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 3.

[26] Mayer-Schonberger, V., & Cukier, K. (2014). *Learning with Big data: The future of education*. Houghton Mifflin Harcourt.

[27] McSherry, F. D. (2009, June). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, (pp. 19-30). ACM.

[28] McSherry, F., & Talwar, K. (2007, October). Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science, 2007. (FOCS'07)*. pp. 94-103. IEEE.
[29] Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy, 2008*. (pp. 111-125). IEEE.
[30] Narayanan, A., & Felten, E. W. (2014). No silver bullet: De-identification still doesn't work. White Paper.
[31] Nergiz, M. E., Clifton, C., & Nergiz, A. E. (2009). Multirelational k-anonymity. *IEEE Transactions on Knowledge and Data Engineering*, 21(8), 1104-1117.
[32] Nergiz, M. E., & Clifton, C. (2007). Thoughts on k-anonymization. *Data & Knowledge Engineering*, 63(3), 622-645.
[33] Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, 79, 119.
[34] Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
[35] Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57, 1701.
[36] Ozalp, I., Gursoy, M. E., Nergiz, M. E., & Saygin, Y. (2016). Privacy-preserving publishing of hierarchical data. *ACM Transactions on Privacy and Security (TOPS)*, forthcoming.
[37] Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438-450.
[38] Prinsloo, P., & Slade, S. (2013, April). An evaluation of policy frameworks for addressing ethical considerations in learning analytics. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 240-244). ACM.
[39] Prinsloo, P., & Slade, S. (2015, March). Student privacy self-management: implications for learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 83-92). ACM.
[40] (n.d.). Retrieved February 09, 2016, from https://en.wikipedia.org/wiki/Fatih_project
[41] Sclater, N., & Bailey, P. (2015). Code of practice for learning analytics.
[42] Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., Ferguson, R., ... & d Baker, R. S. Open Learning Analytics: an integrated & modularized platform.
[43] Siemens, G., & d Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252-254). ACM.
[44] Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. EDUCAUSE review, 46(5), 30.
[45] Singer, N. (2014). InBloom student data repository to close. *New York Times*.
[46] Slade, S., & Prinsloo, P. (2013). Learning analytics ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510-1529.
[47] Solove, D. J. (2002). Conceptualizing privacy. *California Law Review*, 1087-1155.
[48] Solove, D. J. (2004). *The digital person: Technology and privacy in the information age*. NYU Press.
[49] Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review*, 477-564.
[50] Solove, D. J. (2012). Introduction: Privacy self-management and the consent dilemma. *Harvard Law Review*, 126, 1880.
[51] Standard for privacy of individually identifiable health information. (Aug. 14 2002). Federal Register, 67(157):5318153273.
[52] Sweeney L. (2000). *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon, Data Privacy Working Paper 3.
[53] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05), 557-570.
[54] Swenson, J. (2014, March). Establishing an ethical literacy for learning analytics. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 246-250). ACM.
[55] The Family Educational Rights and Privacy Act (FERPA). (2002) 20 U.S.C., 1232g; 34 CFR Part 99
[56] Truta, T. M., & Vinay, B. Privacy protection: p-sensitive k-anonymity property. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*.
[57] Vaidya, J., Shafiq, B., Basu, A., & Hong, Y. (2013, November). Differentially private naive bayes classification. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, (vol. 1, pp. 571-576). IEEE.
[58] Wang, K., & Fung, B. (2006, August). Anonymizing sequential releases. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 414-423). ACM.
[59] Xiao, X., Yi, K., & Tao, Y. (2010, March). The hardness and approximation algorithms for l-diversity. In *Proceedings of the 13th International Conference on Extending Database Technology*, (pp. 135-146). ACM.
[60] Xiao, X., & Tao, Y. (2006, September). Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, (pp. 139-150). VLDB Endowment.

Mehmet Emre Gursoy received the MS degree from University of California Los Angeles (UCLA) and BS degree from Sabanci University, Turkey, both in Computer Science. He is currently pursuing his PhD in the College of Computing, Georgia Institute of Technology. His research interests include data privacy, data mining and network security.

Ali Inan is an Assistant Professor of the Computer Engineering Department at Adana Science and Technology University, Adana, Turkey. He received his Ph.D. in Computer Science from the University of Texas at Dallas and his B.S. and M.S. degrees in Computer Science from Sabanci University, Istanbul, Turkey. His research interests are in database systems, data mining and security, and privacy issues related to the management of data. He is particularly interested in privacy preserving data mining.

Mehmet Ercan Nergiz is the founding CEO of Acadsoft Research. He received his B.S. degree in Computer Engineering from Bilkent University, Ankara, Turkey, in 2003. He received his M.S. degree (2005) and Ph.D. degree (2008) in Computer Science from Purdue University. His research interests include privacy-preserving databases, secure multiparty computation, and anonymization-based privacy protection.

Yucel Saygin is a Professor of Computer Science with the Faculty of Engineering and Natural Sciences at Sabanci University in Turkey. He received his B.S., M.S., and PhD. degrees from the Department of Computer Engineering at Bilkent University in 1994, 1996, and 2001, respectively. His main research interests include data mining, and privacy preserving data management. Yucel Saygn has published in international journals like ACM Transactions on Database Systems, VLDB Journal, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Engineering Management, and in proceedings of international conferences. He co-chaired various conferences and workshops in the area of data mining and privacy preserving data management. He was the coordinator of the MODAP (Mobility, Data Mining, and Privacy) project funded by EU FP7 under the Future and Emerging Technologies Program.