JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 8(2), 83–92. https://doi.org/10.18608/jla.2021.7353

# De-identification is Insufficient to Protect Student Privacy, or—What Can a Field Trip Reveal?

Elad Yacobson[1], Orly Fuhrman[2], Sara Hershkowitz[3], Giora Alexandron[4]

**Abstract**

Learning analytics have the potential to improve teaching and learning in K–12 education, but as student data is increasingly being collected and transferred for the purpose of analysis, it is important to take measures that will protect student privacy. A common approach to achieve this goal is the de-identification of the data, meaning the removal of personal details that can reveal student identity. However, as we demonstrate, de-identification alone is not a complete solution. We show how we can discover sensitive information about students by linking de-identified datasets with publicly available school data, using unsupervised machine learning techniques. This underlines that de-identification alone is insufficient if we wish to further learning analytics in K–12 without compromising student privacy.

**Notes for Practice**

- De-identified educational data may reveal sensitive information about students when linked with publicly available data sets.
- Learning analytics and complete student privacy are competing goals that cannot be settled by technology alone.
- Solutions should combine technology, sensible data policies that are backed up by regulations and legal agreements, and appropriate training for teachers and other educational stakeholders.

**Keywords**

Learning analytics, privacy, re-identification

**Submitted:** 02/09/20 — **Accepted:** 20/05/21 — **Published:** 03/09/21

Corresponding author [1] *Email: elad.yacobson@weizmann.ac.il Address: Department of Science Education, Weizmann Institute of Science, Herzl St 234, Rehovot, Israel. ORCID ID: https://orcid.org/0000-0002-9816-4235*

[2] *Email: orlyf@cet.ac.il Address: The Center for Educational Technology, Klausner St 16, Tel-Aviv, Israel. ORCID ID: https://orcid.org/0000-0003-4270-9512*

[3] *Email: sarah@cet.ac.il Address: The Center for Educational Technology, Klausner St 16, Tel-Aviv, Israel.*

[4] *Email: giora.alexandron@weizmann.ac.il Address: Department of Science Education, Weizmann Institute of Science, Herzl St 234, Rehovot, Israel. ORCID ID: https://orcid.org/0000-0003-2676-6912*

## 1. Introduction

Modern digital learning environments collect rich data that can be used to improve the design of these environments and to develop "intelligent" mechanisms for personalized learning (Siemens, 2013). This potential is starting to realize itself, and we are witnessing rapid growth in the number and variety of learning analytics applications (Roy & Singh, 2017). This development relies on data and an ecosystem of interoperable technologies for processing it, so advances in the field mean that data is circulated between more and more third-party applications (Zeide & Nissenbaum, 2018). Together, the increasing collection of student data, its transfer between entities, and its use to make various decisions make the issue of student privacy and ethical use of student data a major concern (Khalil & Ebner, 2016). The disruptive effect of the Coronavirus pandemic on K–12 educational systems worldwide—moving education to the digital space (EDP, 2020; UNESCO, 2020)—may have a long-lasting effect (Li & Lalani, 2020). In the context of privacy, this means that students will spend a larger portion of their time in learning contexts that are subject to automatic data collection, increasing the risk. Thus, safeguards for privacy are essential for the use of big data in education (Reidenberg & Schaub, 2018). Privacy is a broadly defined term. We follow the definition of Pardo and Siemens (2014) and refer to privacy as the "regulation of how personal digital information is being

observed by the self or distributed to other observers," with personal digital information interpreted as "the information about persons captured by any means and then encoded in digital format." "A Taxonomy of Privacy" by Solove (2005) lists various types of harms that may arise from infringements on privacy. Its application to K–12 education is discussed in Reidenberg and Schaub (2018), which, among other things, refers to secondary uses or inappropriate disclosure of student information, as well as to students' and parents' fear that personal information may be tracked.

Means for privacy protection should eliminate or reduce such potential harms, and solutions should be developed along the technological, legal, organizational, and cultural dimensions (Hoel & Chen, 2016). Finding the weak spots in such solutions helps to improve them, reducing privacy risks, driving the development of better technological solutions, and helping improve data-sharing policies (Sweeney, 2015). In many cases, decisions on privacy-related issues are not only based on the objective level of threat but also affected by public opinion, which is influenced by high-profile data breaches (Krueger & Moore, 2015). The failure of inBloom, the $100 million student data collection project funded by the Gates Foundation, remains a warning sign for how strong this effect may be (Strauss, 11 April 2014).

Since learning analytics research relies on large and complex educational data sets, there is an inherent tension between data-driven research and innovation, and full protection (Peterson, 2016). The main legal reference for student privacy is probably FERPA (Family Educational Rights and Privacy Act; see 20 U.S.C. 1232g), though many data streams used for learning analytics are likely to be excluded from its protection (Reidenberg, 2015). For example, FERPA implicitly refers to quasi-identifiers—pieces of information that are not of themselves unique identifiers but can be combined with other quasi-identifiers to form such a unique identifier (OECD, 2005): "Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty." However, the Department of Education does not provide guidance to educational institutions on how to identify such quasi-identifiers (Daries et al., 2014). Overall, while the learning analytics literature recognizes potential privacy risks, there is little discussion of the concrete ways in which privacy and learning analytics conflict (Rubel & Jones, 2016). One attempt in this direction, with the goal of advancing education research through public releases of data, is the work of Daries and colleagues (2014), who study means to openly share educational data. Their approach aligns with FERPA's definition and also considers quasi-identifiers to address the risk that de-identified data may be re-identified by linking data sets via quasi-identifiers, as shown by Sweeney (2000). However, a strict level of de-identification has been shown to potentially interfere with the analysis of the data (Khalil & Ebner, 2016).

While most of the issues reviewed so far are relevant for various educational contexts, many of them are amplified in K–12 education. Because most K–12 students are minors, the ethical considerations and legal consequences may be more severe (Hoel, Griffiths, & Chen, 2017), and regulations regarding issues such as parental consent and deletion of private data are sometimes unclear or flawed (Peddy, 2017). At the same time, while many educational technology companies serving the K–12 sector fail to implement even basic means for protecting student privacy (Singer, 2015), the centralized nature of many school systems enables wide application that scales the risk (Hoel et al., 2017).

To conclude, there is an inherent tension between learning analytics and student privacy that may be amplified in the K–12 context. This study takes a practical approach to studying this conflict between the competing goals of protecting K–12 student privacy while keeping the data useful for learning analytics. We do so by exploring the risk of re-identifying personal information from de-identified student interaction data. The rationale is that student interaction patterns form an individual fingerprint, which can be linked with publicly available data to form quasi-identifiers that may reveal sensitive information about learners and schools. Specifically, the research question that we study is whether we can re-identify group information—the schools, and the classes within them, which worked with an intelligent tutoring system (ITS)—based on de-identified logs and publicly available information. While classes and schools are not personal information per se, it may still be considered sensitive, for example, if certain schools do not wish to reveal student achievements. More generally, this relates to the idea of group privacy: in many cases, we are analyzed into clusters and influenced based on group membership, but in most cases where people are addressed collectively, no protection will apply (Taylor, Floridi, & Sloot, 2017).

While we present a case study demonstrating how group information can be revealed, it is not totally unlikely that methods like ours—which basically combine knowledge on the domain and data-mining algorithms to generate linkages between data sources—could be used to re-identify even individual students, as was shown in other domains (Sweeney, 2000; Barbaro & Zeller, 2006; Henriksen-Bulmer & Jeary, 2016). Since revealing personal information carries much more risk for the individual, it is thus not inconceivable that methods like this could potentially be used maliciously with more work.

To the best of our knowledge, this is the first learning analytics paper that actually demonstrates how personal properties can be re-identified from de-identified data arriving from K–12 digital learning applications. Our hope is that this will contribute to our understanding of how to design educational technologies and data policies that will enable learning analytics to flourish while not compromising student privacy.

The rest of this paper is organized as follows. First, we describe the methodology; then we present the results; and, last, we discuss their implications.

## 2. Methods

### 2.1 Experiment Setup

The data for this study was taken from a pilot of a new adaptive reading comprehension ITS, which was conducted by the company that developed the ITS under the regulations of Israel's Ministry of Education. The ITS contained about 20 exercises, each consisting of a short text followed by a dozen closed-ended questions. It was used by approximately 600 fifth grade students in 16 different classes, from five schools who chose to participate in the pilot. Each class used the ITS for a period of about two months, once or twice a week, during regular school hours, with additional activity at home after school hours.

### 2.2 Data

The ITS collected student interaction data, including responses to interactive assessment items, navigation between web pages, scrolling, and video events (e.g., play, pause, seek). The activity stream was captured as xAPI statements[1] and was stored on a Learning Locker open source Learning Record Store (LRS)[2]. Among this data, the study used only the interactive assessment events, and within these, only the *time* and *correctness* properties. The data was de-identified prior to its transfer to the research team and was protected by a data-sharing agreement and the institute's regulations for conducting such research.

### 2.3 Process

The rationale behind the process is that the temporal characteristics of student interaction data are more likely to be closer among students who belong to the same class and school. Thus, clustering based on user similarity measures that are computed on such characteristics can reveal these connections.

The process works as follows. (1) We build for each student the list of time intervals in which the student worked in the system. (2) Next we compute the similarity between every pair of students, yielding (3) a user-user similarity matrix. (4) We then cluster the students into (what we interpret as) classes, based on their similarity, using standard clustering techniques (K-means and gap statistic). (5) Then we identify the class routine and look for deviations from it. (6) Finally, we intersect these deviations with publicly available information, to match clusters with schools. Below we describe each step in more detail.

**Step 1: Building students' time-interval lists.** First, for each student $s$, we built a list $l_s$ of the time intervals in which $s$ used the tutor. To build $l_s$, we iterated over the time-sorted events of $s$. Events that are over 45 minutes apart are considered to belong to different intervals (we used 45 as a threshold because this is the duration of a lesson).

**Step 2: Computing the Jaccard index for each pair of students.** After creating the list of time intervals for each student, we computed the similarity between the interval lists of each pair of students, using the *Jaccard similarity index*. The Jaccard index, also known as *intersection over union*, measures the size of the intersection divided by the size of the union and is defined as

$$\text{Jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

Here, $A$ and $B$ are the lists of time intervals of two students $i$ and $j$, denoted $l_i$ and $l_j$, respectively. The intersection $|l_i \cap l_j|$ is defined as the amount of time both $i$ and $j$ worked simultaneously with the tutor, whereas the union $|l_i \cup l_j|$ is defined as the overall amount of time in which either $i$ or $j$ was working with the tutor. This is illustrated in Figure 1. The overall time in which *either* student $i$ or student $j$ was working with the tutor is 34 minutes, and the amount of time in which *both* students were working with the tutor is 9 minutes. Therefore, the Jaccard index for $i$ and $j$ is 0.26.

However, computing the Jaccard index in this manner to measure the similarity between two users is liable to be misleading, since it does not take into account the difference in the lengths of the time intervals and can therefore overemphasize the importance of a single long interval (Kabir, Wagner, Havens, Anderson, & Aickelin, 2017).

In order to address this issue, we also tried a discrete approach to computing the Jaccard index: we checked for each pair of students $i$ and $j$ the number of overlapping time intervals within $l_i$ and $l_j$, out of the total number of time intervals contained in $l_i$ and $l_j$. Two time intervals are considered as overlapping if at least half of each interval is contained in the other. The rationale is that each time interval represents a lesson, and partially overlapping intervals may represent two students who participate in the same lesson but may progress differently within the tutor. This is illustrated in Figure 2. Among the interval lists of students $i$ and $j$, there are two overlapping intervals out of six intervals in total (overlapping intervals are considered as the same interval), so the Jaccard index for students $i$ and $j$ is 0.33. We note that the discrete approach does not form a metric because the Jaccard distance that is based on it does not satisfy the triangle inequality. However, it seems that this can rarely affect results on real data, while the advantages of the discrete approach may make it superior to the continuous one. In the case reported below, both approaches yielded the same results.

[1] https://xapi.com
[2] https://docs.learninglocker.net

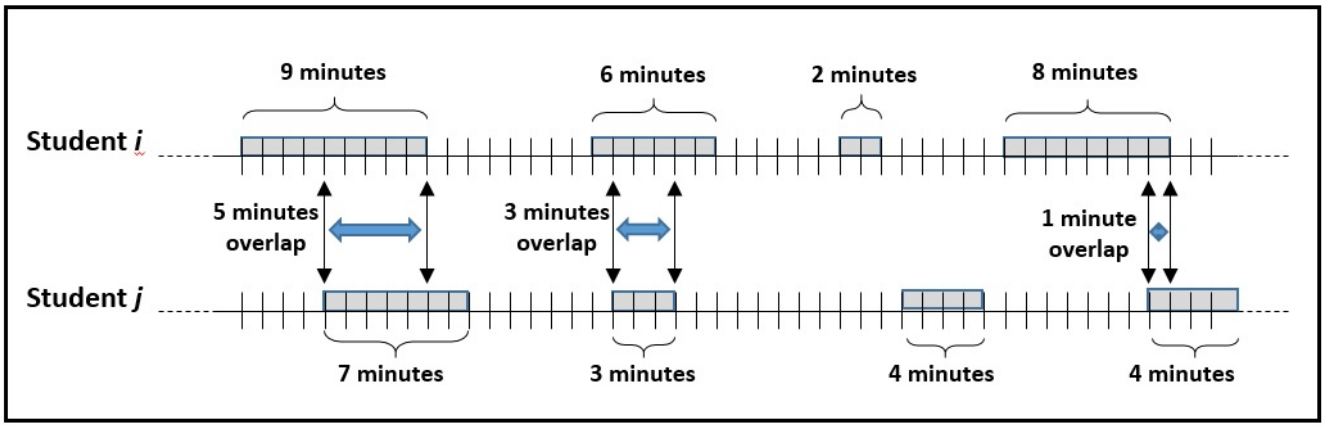


**Figure 1.** Time intervals of two students $i$ and $j$—continuous method.

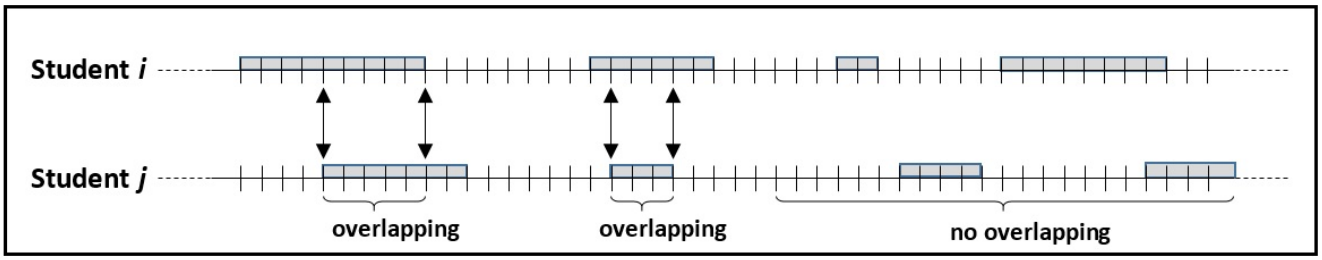


**Figure 2.** Time intervals of two students $i$ and $j$—discrete method.

**Step 3: Building a user-user similarity matrix.** Using the Jaccard index computed for each pair of students, we build a *user-user* similarity matrix $M$, where $M[i, j]$ = the Jaccard index of student $i$ and student $j$. Since Jaccard$(i, j)$ = Jaccard$(j, i)$, $M$ is symmetric.

**Step 4: Clustering the students into classes.** To cluster the students, we apply K-means clustering on $M$. The number of clusters is determined using the gap statistic (Tibshirani, Walther, & Hastie, 2001).

**Step 5: Identifying class attributes.** After clustering the students into classes, we examined several group characteristics of the classes. First, we looked into the distribution of student performance within each class, where a student's performance is operationalized as "fraction correct on first attempt" (we operationalized performance this way since considering more than one attempt provides less information on the differences between the clusters, because eventually students tend to get most items correct). In addition to class achievements in the tutor, we searched for temporal patterns—the weekly time slots in which each class tended to work, whether there were deviations from this weekly routine, the date of the first and last interaction of each class, etc.

**Step 6: Re-identifying classes.** This phase includes linking the class attributes discovered in Step 5 with publicly available information found on the Internet, to identify the physical location of classes and schools. Steps 5 and 6 are the non-automatic part of the process, which requires considerable legwork.

## 3. Results

### 3.1 Clustering

As described earlier, we used K-means clustering to group the students into classes. In order to determine the number of clusters for the K-means algorithm, we used the gap statistic. Initially this yielded 14 clusters (see Figure 3), but after clustering the students to classes using K-means with 14 clusters, some of those clusters contained more than 60 students. Knowing that this is not a reasonable number of students for one class in K–12 education, we tried running K-means with 15 clusters and 16 clusters. With 15 clusters, the same result occurred-—some of the clusters contained more than 50 students, which again was not reasonable. When using 16 clusters, all of the clusters contained fewer than 40 students, which is the maximum number of students allowed in one class in our school system. Thus, we used 16 as the number of clusters.

ISSN 1929-7750 (online). 

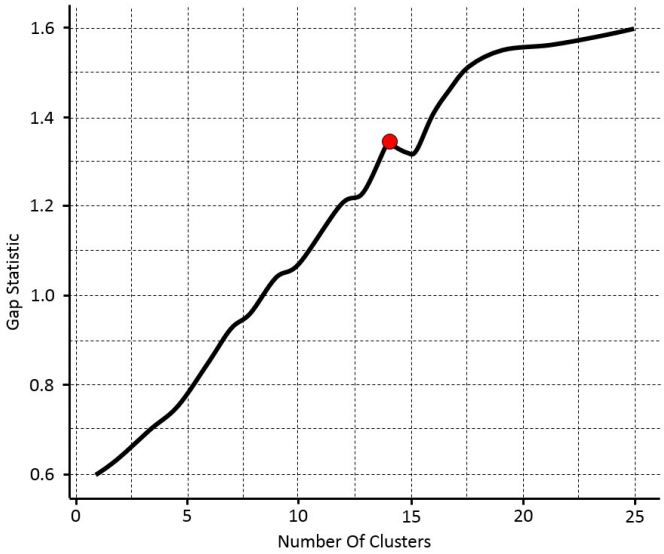


**Figure 3.** Gap statistic.

After clustering using 16 clusters, we measured the goodness-of-fit of the clustering with the adjusted Rand index (ARI) (Hubert & Arabie, 1985) against the ground truth (list of student IDs within each class). The ARI for the K-means clustering was 0.943. (The ARI was similar for both the continuous and the discrete Jaccard index methods. See Step 2 in the Methods subsection for more details.)

In order to check whether a similar result could be obtained with a more naive algorithm, we compared our similarity-based approach to a *union-find* (*connected components*) algorithm: every two students $i$, $j$ are considered as belonging to the same cluster if at least $x\%$ of their time intervals overlap. If students $i$ and $j$ are found to belong to the same cluster, the clusters of $i$ and $j$ are merged. We examined the quality of the clustering for different values of $x$. The optimal value for $x$ was 28%. With $x = 28\%$, the ARI for the union-find algorithm was 0.67, significantly lower than the user similarity–based approach. Thus, we concluded that the union-find approach fails to translate the similarities into clusters. For the remainder of the experiment, we used the clustering obtained by the K-means algorithm.

### 3.2 Identifying Target Classes

For each cluster obtained in the previous step, we examined student achievements in the tutor, operationalized as a fraction of correct on first attempt. We focused our attention on two clusters with significantly higher/lower performance (see Figure 4). The considerable differences suggested that these clusters did not represent regular classes, but rather special classes for gifted children (the high-performing class) and children with learning disabilities (the low-performing class). Hereafter, we refer to these clusters as gifted/disabilities classes.

### 3.3 Identifying Weekly Class Routine

Having decided to focus on these two clusters, we checked the time intervals in which the students in the two clusters were active in the tutor. We found that the start and end dates of using the tutor were the same for both clusters. Since the start and end dates of the other clusters were different, we hypothesized that both clusters represented classes belonging to the same school. This hypothesis was based on our assumption that it is more likely to find such similarity among classes that belong to the same school than is merely coincidental.

We performed a short web search and found that there were only three schools in the country that had both a gifted children class and a special class for children with disabilities within the same school.

Following this, we looked for patterns in the weekly routine of these two clusters, and whether there were deviations from these routines. Checking the time slots in which the cluster representing the low-performing class used the system, we found no irregularities. However, when we looked at the high-performing cluster, we found that it used the tutor twice a week mostly following a regular pattern, except for two dates: once in the fourth session in which this cluster used the tutor, and once in the final session. This is illustrated in Figure 5.

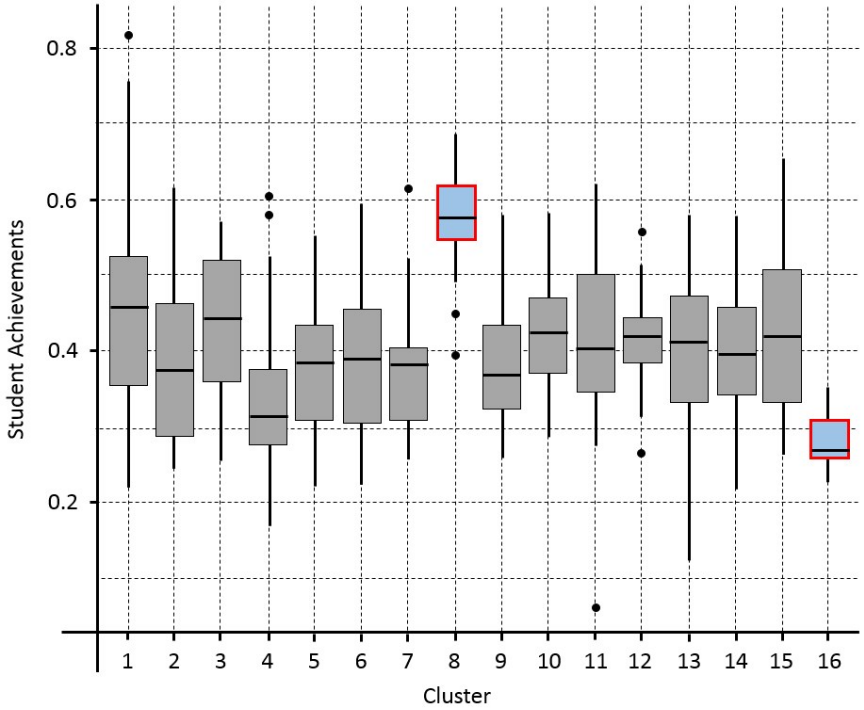


**Figure 4.** Proportion of correct on first attempt, by cluster.

| | Date1 | Date2 | Date3 | Date4 | Date5 | Date6 | Date7 | Date8 | Date9 |
|---|---|---|---|---|---|---|---|---|---|
| Hour #1 | | | | | | | | | |
| Hour #2 | | | | | | | | | |
| Hour #3 | | | | | | | | | |
| Hour #4 | | | | | | | | | |

**Figure 5.** Gifted class timetable.

### 3.4 Linking Temporal Patterns with Publicly Available Information

After identifying the aforementioned irregularities, we looked for information that could explain them. We checked the three schools' webpages, looking for information about special events during the dates in which the deviations from the regular weekly pattern occurred. We also looked in archives of the local newspapers of the three towns in which these schools reside.

This search into publicly available information yielded that on the morning of one of these dates, one of the three schools went out on a short field trip. Thus, we hypothesized that the gifted children class (represented by the high-performing cluster) belonged to that school. *Checking with the company confirmed that our identification was correct.*

## 4. Discussion

In K–12 education, the context and practices determine the ways in which students interact with online learning environments and are naturally reflected in their interaction data. For example, in K–12, students tend to learn in regular school hours, to have weekly routines, and to belong to classes. When a class learns a certain topic together from an ITS, student activity time will overlap, and students will most likely interact with similar items. Such characteristics of the context imply constraints that restrict the multivariate distribution of the data in ways that can be used to discover hidden properties. In the Results section, we show how we utilize this idea to identify class membership and then discover the physical location of some of the classes. This is based on an unsupervised learning technique that clusters students based on similarities in their temporal behaviour and then links the clustering with the physical location using information that is publicly available on the web. Thus, the temporal patterns served as sort of quasi-identifiers that enabled us to link between data sources and reveal new information.

Identifying schools and classes may eventually affect students as individuals. It may reveal information that the schools do not wish to disclose, and it is not totally unlikely that it can be used to make determinations about individual students. For example, the group information can be used to target students with various types of advertisements, or a college may decide to reject a student based on class or school characteristics. In addition, such group profiling increases the risk of re-identification, since it may provide new ways to link between data sets. Kitto and Knight (2019) referred to data linking as an example of an important opportunity to improve learner models that may be underused due to ethical concerns. While we agree with their conclusion that harms and benefits should be balanced, our use case demonstrates that the risk in linking is not theoretical; in particular, the time and correctness of student attempts—the events that our algorithm exploits for the re-identification—are among the most basic events used by many learning analytics applications, from teacher dashboards to adaptive engines. Therefore, learning analytics applications that receive access to this data should be regarded as containing sensitive information even if the data was "cleaned" from any direct identifiers, such as name, address, phone number, and IP address.

Theoretically, our findings may also indicate a potential risk of re-identifying individuals. The idea that underlies personalization in education is that students may differ in their preferences and needs. Learning analytics research seeks to identify digital traces that are useful for specifying such differences; namely, it operates under the assumption that student data, and especially process data of the type used in this study, contains behavioural patterns that are unique among individual students or learner profiles. If this assumption holds, it means that learning data can potentially hold individual "fingerprints" that can be used to link between learners across contexts, eventually enabling the accumulation of information about them, even if each system is secured by itself. If such educational fingerprints exist, the conclusion is that cleaning the data from any attribute that can reveal personal information about the learners will severely limit the ability to compute such personal learning fingerprints and use them to provide individualized learning. Or, in Paul Ohm's words, "Data can either be useful or perfectly anonymous but never both" (2009).

The important question that arises, in the light of these findings, is how to protect student privacy from possible infringements while still maintaining the possibility of a useful and reliable analysis. This study suggests that the inherent tension between the utility of educational data and the risk to student privacy cannot be settled merely by de-identification techniques. A possible solution is applying technological means such as injecting some amount of noise into the data, an approach known as *differential privacy* (Dwork, 2008). However, although such a solution is likely to succeed in preventing the process of re-identification, it is also likely that it will affect analytics on the individual student, rendering it unsuitable for applications aimed at personalized learning. A personal user model (Kay & Kummerfeld, 2019) that is responsible for storing and providing access to student raw data may be another valuable direction for addressing privacy risks, which can allow the researcher to follow a learner over time and link data from various sources and products (Kitto & Knight, 2019). However, it is hard to see how delivering responsibility to the user may work with K–12 learners.

And since both the value of the data and the risk are hard to discern in advance, judgeful decisions would have to be made based on partial information, requiring policy makers to make difficult choices. Narayanan and Felten (2014) argue that there is no "silver bullet" (in data privacy). Not surprisingly, this holds for educational technology as well. Therefore, the solution cannot rely solely on technological means but should rather combine them with clear regulation that can back up trustable data sharing policies and cultural and organizational processes that raise the awareness and capability of K–12 educators and institutions to understand, minimize, and manage privacy risks. This highlights the role of the *context* in addressing privacy issues. The most recognized theory for contextual privacy is Nissenbaum's contextual integrity (2009). This framework suggests examining privacy as an appropriate flow of information, where what is considered "appropriate" depends on contextual information norms. These norms depend on the roles of the sender and the receiver, on the type of information being shared, and on how information is transmitted. Norms may be context dependent, and as Hoel and Chen pointed out (2019), using western privacy concepts to inform privacy engineering for a cross-cultural market may not work well. Thus, this theory provides a conceptual framework for examining the issue of privacy and learning analytics in K–12 education through a prism that both underlines the role of organizational processes and different stakeholders and balances benefit and risk by defining appropriate norms.

While it is beyond the scope of this paper to detail solutions, it is worthwhile to point out that there seems to be lack of knowledge on the matter among learning analytics practitioners and educational stakeholders, limiting their ability to make knowledgeable decisions. Addressing this may be low-hanging fruit. As was suggested by Macfadyen (2017), learning analytics practitioners should learn about data ethics and privacy issues, including the risks and benefits of learning analytics tools and strategies. In this regard, since practitioners tend to assume that de-identification is a data privacy strategy that provides complete protection, it is important that they be aware of its limitations. In particular, as is pointed out in EDUCAUSE (2015), not all types of data and datasets have the same level of risk. For example, process data or forum data requires careful consideration because it contains information that may be more distinguishable among individuals. This can be addressed through various types of on-the-job training that should be offered to learning analytics practitioners. K–12 educators as well should be familiar with the issue of data privacy on some level. Teachers and school staff govern the access of children to

ISSN 1929-7750 (online). 

educational applications and may also share information about educational activities on school webpages and social media. Providing them with some basic training on privacy could help them make more informed decisions on school and class matters that may involve privacy issues.

### 4.1 Limitations

Concerning the overall process, while we do argue that the re-identification procedure was quite simple—a relatively basic computational step followed by a bit of legwork, we definitely do not argue that it is a standard process that can uncover each class that used the ITS. While the clustering step can probably be generalized (with some caveats; see below), the linking part was ad hoc, exploiting context-dependent properties and the external data that we could link to the ITS data.

Regarding the clustering, our technique is based on interpreting overlap in temporal patterns of student behaviour as evidence of same-class membership. With more students, the probability of overlap by chance will obviously increase. From a different direction, schools do not always organize learning in rigid classes, especially in personalized learning scenarios where flexibility is desired, which will obviously render our method less effective. While it is comforting to know that in such cases big data and personalized learning may hinder re-identification rather than render it, we actually expect that by adding additional similarity features (e.g., topics students are learning), class (and other properties) re-identification could still be possible.

Last, concerning the results, one may argue (and we may agree) that revealing students' class/school is not a severe privacy breach. However, we believe that it is not totally theoretical that similar methods can be used to re-identify more personal traits.

## 5. Summary and Conclusions

This paper studies potential threats to student privacy in K–12 education and how that privacy may compete with learning analytics. Specifically, it does so using a case study that demonstrates how student information can be re-identified from de-identified data by linking publicly available information with the results of a temporal analysis–based clustering algorithm that has access to very simple student interaction data—time and correctness of student attempts. While the growing ecosystem of data-driven educational research and development carries a lot of potential for K–12 education, it amplifies the risk to student privacy. By revealing potential risks that learning analytics processes may pose to student privacy, we hope to contribute to the development of better technological solutions and data-sharing policies, which will eventually help reduce the risk and help learning analytics fulfill its potential for learners.

## Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The research of EY was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center. The work of GA is also supported by the Willner Family Leadership Institute for the Weizmann Institute of Science.

## References

Barbaro, M., & Zeller, T. (2006, 01). A face is exposed for AOL searcher no. 4417749. *New York Times*. (Accessed May 20, 2021) Retrieved from http://shawndra.pbworks.com/f/A+Face+Is+Exposed+for+AOL+Searcher+No.+4417749+-+New+York+T.pdf

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., … Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, *57*(9), 56–63. https://doi.org/10.1145/2643132

Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation* (TAMC 2008), 25–29 April, Xi'an, China (pp. 1–19). Springer. https://doi.org/10.1007/978-3-540-79228-4-1

EDP. (2020). *Education During COVID-19; Moving towards e-Learning.* (EUROPEAN DATA PORTAL; accessed May 12, 2021) Retrieved from https://www.europeandataportal.eu/en/impact-studies/covid-19/education-during-covid-19-moving-towards-e-learning

EDUCAUSE. (2015). *Guidelines for Data De-identification or Anonymization.* (Accessed May 12, 2021) Retrieved from https://www.educause.edu/focus-areas-and-initiatives/policy-and-security/cybersecurity-program/resources/information-security-guid

Henriksen-Bulmer, J., & Jeary, S. (2016). Re-identification attacks—A systematic literature review. *International Journal of Information Management*, *36*, 1184–1192. https://doi.org/10.1016/j.ijinfomgt.2016.08.002

Hoel, T., & Chen, W. (2016). Privacy-driven design of learning analytics applications: Exploring the design space of solutions for data sharing and interoperability. *Journal of Learning Analytics*, *3*(1), 139–158. https://doi.org/10.18608/jla.2016.31.9

Hoel, T., & Chen, W. (2019). Privacy engineering for learning analytics in a global market: Defining a point of reference. *The International Journal of Information and Learning Technology*, *36*(4), 288–298. https://doi.org/10.1108/IJILT-02-2019-0025

Hoel, T., Griffiths, D., & Chen, W. (2017). The influence of data protection and privacy frameworks on the design of learning analytics systems. In *Proceedings of the Seventh International Conference on Learning Analytics and Knowledge* (LAK 2017), 13–17 March 2017, Vancouver, BC, Canada (pp. 243–252). New York: ACM. https://doi.org/10.1145/3027385.3027414

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. https://doi.org/10.1007/BF01908075

Kabir, S., Wagner, C., Havens, T. C., Anderson, D. T., & Aickelin, U. (2017). Novel similarity measure for interval-valued data based on overlapping ratio. In *Proceedings of the 2017 IEEE International Conference on Fuzzy Systems* (FUZZ-IEEE 2017), 9–12 July 2017, Naples, Italy (pp. 1–6). IEEE. https://doi.org/10.1109/FUZZ-IEEE.2017.8015623

Kay, J., & Kummerfeld, B. (2019). From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology*, *50*(6), 2871–2884. https://doi.org/10.1111/bjet.12878

Khalil, M., & Ebner, M. (2016). De-identification in learning analytics. *Journal of Learning Analytics*, *3*(1), 129–138. https://doi.org/10.18608/jla.2016.31.8

Kitto, K., & Knight, S. (2019). Practical ethics for building learning analytics. *British Journal of Educational Technology*, *50*(6), 2855–2870. https://doi.org/10.1111/bjet.12868

Krueger, K. R., & Moore, B. (2015). New technology "clouds" student data privacy. *Phi Delta Kappan*, *96*(5), 19–24. https://doi.org/10.1177/0031721715569464

Li, C., & Lalani, F. (2020). *The COVID-19 Pandemic Has Changed Education Forever. This Is How.* (Accessed May 12, 2021) Retrieved from https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/

Macfadyen, L. (2017). What does a learning analytics practitioner need to know? In *Proceedings of the Workshop on Methodology in Learning Analytics and the Workshop on Building the Learning Analytics Curriculum* (LAK 2017), 13–17 March 2017, Vancouver, BC, Canada.

Narayanan, A. R. V., & Felten, E. W. (2014). *No Silver Bullet: De-identification Still Doesn't Work.* (Accessed May 12, 2021) Retrieved from http://www.randomwalker.info/publications/no-silver-bullet-de-identification.pdf

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press. https://doi.org/10.1515/9780804772891

OECD. (2005). *Glossary of Statistical Terms: Quasi-identifier.* (Accessed May 12, 2021) Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=6961

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*, 1701–1777.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450. https://doi.org/10.1111/bjet.12152

Peddy, A. M. (2017). Dangerous classroom "app"-titude: Protecting student privacy from third-party educational service providers. *Brigham Young University Education and Law Journal*, *2017*(1), 125–159. (Accessed May 12, 2021) Retrieved from https://digitalcommons.law.byu.edu/cgi/viewcontent.cgi?article=1395context=elj

Peterson, D. (2016). Edtech and student privacy: California law as a model. *Berkeley Technology Law Journal*, *31*, 961–996. Retrieved from https://btlj.org/data/articles2016/vol31/31$_a r/0961_0 996_P eterson_W EB.pdf$

Reidenberg, J. R. (2015). Hearing testimony on how emerging technology affects student privacy. In *Hearing before the U.S. Congress, House Committee on Education and the Workforce, Subcommittee on Early Childhood, Elementary and Secondary Education, 114th Congress,* 12 February 2015, Washington, DC, USA. Retrieved from https://www.govinfo.gov/content/pkg/CHRG-114hhrg93208/pdf/CHRG-114hhrg93208.pdf

Reidenberg, J. R., & Schaub, F. (2018). Achieving big data privacy in education. *Theory and Research in Education*, *16*(3), 263–279. https://doi.org/10.1177/1477878518805308

Roy, S., & Singh, S. N. (2017). Emerging trends in applications of big data in educational data mining and learning analytics. In *Proceedings of the Seventh International Conference on Cloud Computing, Data Science Engineering—Confluence,* 12–13 January 2017, Noida, India (pp. 193–198). IEEE. https://doi.org/10.1109/CONFLUENCE.2017.7943148

Rubel, A., & Jones, K. (2016). Student privacy in learning analytics: An information ethics perspective. *The Information Society*, *32*, 143–159. https://doi.org/10.1080/01972243.2016.1130502

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, *57*(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Singer, N. (2015). *Data security gaps in an industry student privacy pledge.* (Accessed May 12, 2021) Retrieved from https://bits.blogs.nytimes.com/2015/02/11/data-security-gaps-in-an-industry-student-privacy-pledge/?$_r = 0$

Solove, D. (2005). A taxonomy of privacy. *University of Pennsylvania Law Review*, *154*(3), 477–564. https://doi.org/10.2307/40041279

Strauss, V. (11 April 2014). $100 million Gates-funded student data project ends in failure. *Washington Post*. (Accessed May 12, 2021) Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2014/04/21/100-million-gates-funded-student-data-project-ends-in-failure/

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, *671*, 1–34. Retrieved from https://doi.org/10.1184/R1/6625769.v1

Sweeney, L. (2015). Only you, your doctor, and many others may know. *Technology Science*. Retrieved from https://techscience.org/a/2015092903/

Taylor, L., Floridi, L., & Sloot, B. (2017). *Group Privacy: New Challenges of Data Technologies*. Springer. https://doi.org/10.1007/978-3-319-46608-8

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

UNESCO. (2020). *National Education Responses to COVID-19: Summary Report of UNESCO's Online Survey.* (Accessed May 12, 2021) Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000373322

Zeide, E., & Nissenbaum, H. (2018). Learner privacy in MOOCs and virtual education. *Theory and Research in Education*, *16*(3), 280–307. https://doi.org/10.19173/irrodl.v21i4.4643