Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Too Detailed to Share? Towards Risk-Based Privacy Protection of Fine-Grained Educational Data

Hibiki ITO[a*], Chia-Yu HSU[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*hibiki.itoo@gmail.com

Abstract: Due to the recent increase in the use of digital learning platforms, fine-grained digital trace data has been growing in the education sector. However, despite the potential of such micro-level log data for understanding and personalising individual learning processes, its secondary use is limited due to privacy concerns. A key to advancing data sharing for the secondary use while protecting individual privacy is effective risk assessments. Nevertheless, prior research predominantly focuses on privacy risks of structured tabular data, leaving fine-grained digital trace data underexplored. To fill this gap, we conduct a comprehensive risk analysis of fine-grained educational data using the unicity framework. Employing two real-world datasets reflecting on secondary and higher education settings and two open datasets on self-paced language learning, we demonstrate that fine-grained educational data is highly susceptible to re-identification through timestamps. In addition, we show that the effectiveness of naïve coercing of timestamps depends on the number of students in the dataset and the diversity of educational contexts where the data is collected. Our findings help practitioners to make risk-based decisions to choose appropriate privacy protection strategies.

Keywords: Data sharing, privacy, fine-grained data, unicity

1. Introduction

The last few decades have witnessed the rapid increase in the use of digital platforms in education, leading to growing data about learning and its environment. Particularly, educational activities that were once transient and confined to the involved leaners and educators now leave fine-grained traits of log data that reveal their learning profiles. Educational researchers in learning analytics (LA) and educational data mining (EDM) are urging to take advantage of these micro-level data to understand individual learning processes and personalise learning.

However, despite the potential of increasing fine-grained digital trace data in education, the secondary use of such data by researchers has been limited due to privacy concerns (Fischer et al., 2020), leaving sensitive, but potentially useful data in enclaves (Baker & Hutt, 2025). Although protecting the privacy of learners is quintessential in the use of educational data, data sharing is also required for advancing open science and developing reliable educational technologies (Baker et al., 2024).

To ensure an appropriate trade-off between the preservation of individual privacy and the pursuit of collective societal benefits, careful risk assessments and choosing appropriate privacy protection techniques are of paramount importance (Joksimović et al., 2022). Prior research in the field of LA/EDM has investigated privacy risks associated with educational data such as the re-identification risk (Prasser & Kohlmayer, 2015). To mitigate these risks, privacy-preserving data sharing methods such as k-anonymity (Angiuli et al., 2015), synthetic data generation (Liu et al., 2025) and differential privacy (Gursoy et al., 2017) have been studied in the context of education. However, these studies primarily focus on structured

tabular data, and less attention has been paid to fine-grained, set-valued educational data such as log data from learning platforms and behavioural sensors—the gap we aim to fill. Particularly, the risk assessment of fine-grained log data is missing in the literature, which is essential for advancing the privacy-aware secondary use.

To this end, this paper focuses on the re-identification risk of fine-grained educational data, providing a first comprehensive risk assessment through the unicity framework (de Montjoye et al., 2013) in the realm of education. Informally, the unicity metric estimates how many evet-level records are sufficient to single out individual data subjects in a set-valued dataset. This intuitive and realistic metric of privacy risk arguably offers useful evidence for data custodians to make risk-based decisions on what privacy protection methods to apply for sharing educational data.

2. Related works

2.1 Risk assessment on educational data

Effective risk assessment is key to balancing individual privacy concerns with the societal benefits derived from the secondary use of educational data (Joksimović et al., 2022). An inaccurate estimation of privacy risk—be it an underestimation or overestimation—can lead to the implementation of suboptimal protective strategies. Such misalignment may result in either latent vulnerabilities or an unwarranted loss in data utility. The former undermines stakeholder trust and potentially disrupts educational activities as symbolised by the failure of inBloom (Bulger et al., 2017), while the latter may produce severe outcomes, including harmful intervention results (Fredrikson et al., 2014).

Among various types of information-theoretic privacy risks to consider when choosing appropriate anonymisation techniques, we focus on the re-identification risk of pseudonymous data because such data is practically common and beneficial in education research, and the re-identification risk is an area of active research in the domain of education while leaving fine-grained data underexplored. It should be noted that by information-theoretic risk we mean to pay particular attention to quantifiable re-identification risks such as uniqueness of records in a dataset; otherwise, assessing the feasibility of attackers who have access to anonymised data would necessarily depend on the context.

In the literature, the ARX (Prasser & Kohlmayer, 2015) is perhaps the most used assessment tool of re-identification risk of educational data. Having been developed primarily for health data, it allows for estimating re-identification risk defined through the uniqueness of records within specific populations. For example, Kyritsi et al. (2019) employed the ARX tool to estimate the re-identification risk of aggregated tabular data of LMS logs. The study showed that total number of logs for each learner exhibited the highest re-identification risk with 24.39% records unique in the sample dataset and 3.48% unique in the population, according to ARX's definition of population uniqueness.

Another tool for assessing re-identification risk is the Re-identifier Risk Ready Reckoner (R4) developed at the Commonwealth Scientific and Industrial Research Organisation (CSIRO, 2019). Unlike the algorithm of ARX, it quantifies re-identification risk using Markov models accounting for not only uniqueness but also uniformity, i.e., consistency of individuals throughout the course (Vatsalan et al., 2022).

Nonetheless, both these tools take as input data in tabular form and are not applicable to set-valued data. Although in the secondary use, set-valued educational data is usually aggregated into features like the total number of logs in the above example, aggregating and conducting risk assessment every time when a third-party analyst creates a request for data sharing is burdensome and not a sustainable solution. Therefore, evaluating the re-identification risk of fine-grained data would play a crucial role in allowing for a wider range of analyses in the secondary use of the data, which is missing in the literature of the education domain.

2.2 Unicity framework

Proposed by de Montjoye et al. (2013), *unicity* is a framework to evaluate the re-identification risk based on the uniqueness of individuals in a set-valued dataset. Using large-scale location data, de Montjoye et al. (2013) demonstrated that given four geographical points, on average over 95% of individuals can be uniquely determined. It should be noted that re-identifiability and uniqueness are distinct concepts, and unicity is a metric of uniqueness in the first place, potentially overestimating re-identification risk (Barth-Jones et al., 2015). However, as mentioned before, we focus on the re-identification risk of educational data quantified by unicity as other factors such as how an adversary gains auxiliary information necessarily depend on the context.

There has also been a critique that unicity should be estimated based on the uniqueness in the population as the number of individuals in a dataset potentially impacts its unicity (Sánchez et al., 2016). A rebuttal to this claim by Farzanehfar et al. (2021) demonstrated that unicity remains high for larger, population-level location data. That said, unlike location data, educational data are typically collected, stored and used within a single institution or even a single module by a single teacher. Hence, we argue that assessing the unicity of micro-level data collected within a specific context such as a single institution and a single module (see Section 3.1) is indeed meaningful in the domain of education. Overall, our research question (RQ) is formulated as follows:

RQ: *What is the re-identification risk evaluated by the unicity framework in fine-grained educational data?*

3. Materials and methods

3.1 Data

For the experiments we use two private real-world datasets and two open datasets publicly available online. Table 1 shows the summary of the datasets used in the experiments. All data in our experiments are pseudonymous.

Table 1. *Description of the datasets*

Dataset	# records	# students	Period	Description
BookRoll University	66,259	51	4 m	Reading behaviour in a bachelor-level academic reading module
BookRoll Secondary	6,486,986	752	4 m	Reading behaviour in multiple secondary school classes (mostly mathematics and English)
Duolingo	12,854,226	115,222	2 w	Duolingo vocabulary lessons for multiple learning languages
EdNet-KT4	131,441,538	297,915	1y 3m	English reading and listening exercises

To empirically assess the re-identification risk in sharing fine-grained educational data, we employed log data collected via BookRoll, an e-book system through which learning materials are distributed to students as PDF files (Ogata et al., 2015). As shown in Figure 1, interaction logs (e.g. open/close materials, highlight texts, create handwritten notes) are sent to the learning record store (LRS) in the form of the xAPI standard. Each log includes the timestamp, the actor, the xAPI verb, the context ID indicating the class/module and the device type on which the event was operated.

Figure 1. An example of xAPI log collected by the BookRoll system

Additionally, each xAPI log is associated with a BookRoll-specific operation name, which indicates more granular description than xAPI verbs. Table 2 shows some examples of operation names and corresponding xAPI verbs.

Table 2. *Examples of BookRoll operations and xAPI verbs*

xAPI verb	Operation Name	Function
`read`	`NEXT`	Go forward to the next page
	`PREV`	Go back to the previous page
`noted`	`ADD_MEMO`	Add a note
	`DELETE_MEMO`	Delete a note
`highlighted`	`ADD_MARKER`	Add a marker highlight
	`DELETE_MARKER`	Delete a marker highlight

We use two datasets consisting of BookRoll logs. First, the BookRoll University dataset contains log data of undergraduate students within an academic reading module at a Japanese public university. The context is specific to a single module with medium class size (n=51), and the duration of data collection ranges over a semester, thus being a common data unit for LA/EDM analyses. We also employ this dataset to reflect the scope of a typical primary use of educational data at higher education institutions, as we are interested in sharing such data for the secondary use. Typical analyses on this dataset would include temporal learning processes or collaborative learning within a specific context.

Second, the BookRoll Secondary dataset consists of Japanese secondary school students' log data (n=752) across multiple classes (mostly mathematics and English). The scope of the data is limited to a single school, which is a typical unit of educational data in K-12 settings for primary use. Again, this is because we are interested in the risk of sharing such data that are often not released due to privacy concern. As the dataset is larger than the previous dataset in terms of volume and represents more contextual diversity, it would be suitable for studying cross-context learning behaviour in a secondary school setting. For fair comparison, we set the period of data collection to a single semester, the same as for the BookRoll University dataset. This serves as a lower bound of more longitudinal data since the uniqueness of individual trajectories generally increases as data becomes more longitudinal.

To further generalise the results on the previous two private datasets that represent more formal curriculum-based education for the young, we conduct additional experiments with two public datasets that reflect on more self-paced life-long learning. The Duolingo dataset (Settles, 2017) is open data for replication of a study on second language learning by Settles and Meeder (2016). It contains Duolingo vocabulary lesson results for each learner, where each record includes the timestamp, the lexeme (i.e. the target word) and the learning language. Although the dataset spans only a two-week period, it reflects more diverse self-paced learning behaviours, as learners engage with Duolingo lessons independently—unlike

the BookRoll data, where learner activity is shaped by regular, scheduled classes in formal educational settings.

Lastly, the EdNet-KT4 (call EdNet for short) dataset is a large-scale open dataset consisting of exercise results on Santa, a multi-platform self-study app for preparation of the TOEIC (Test of English for International Communication) test (Choi et al., 2020). The data provides more longitudinal learning processes of a larger number of learners, complementing the Duolingo dataset. The results derived from these public datasets also exemplifies a common form of educational data that are gathered and stored by educational technology companies, yet such data often remain inaccessible to researchers. These findings could serve as evidence to encourage companies to adopt effective privacy protection methods for data sharing, which is essential for advancing the broader use of educational data (Fischer et al., 2020).

3.2 Methods

We apply the unicity framework (de Montjoye et al., 2013) to the four datasets[1]. Algorithm 1 shows the process of calculating the unicity, given a dataset and a set of quasi-identifiers (QIs). Note that in each dataset, every *student* contributes multiple *events*. We estimate the unicity by taking the average of the outputs over ten random seeds. For the BookRoll University and Secondary datasets, we include the entire datasets in each sample (i.e. the sample size *m* equals the number of students in each dataset). For the Duolingo and EdNet datasets, we set *m*=2500.

ALGORITHM 1: CALCULATE UNICITY

Input: $\mathfrak{D}$ (dataset), Q (set of QIs), ε (number of events available to the attacker), m (sample size)
Output: *Unicity*

$D \leftarrow$ subset $\mathfrak{D}$ by quasi-identifiers in Q // only quasi-identifiers are used
$S \leftarrow$ randomly choose m students from D
Unicity $\leftarrow 0$
for t in S do // go through every student in the sample
  Trajectory $\leftarrow$ all events of t in D
  Observations $\leftarrow$ randomly choose ε events in *Trajectory*
  Candidates $\leftarrow$ choose from D students whose events include *Observations*
  if $|Candidates| = 1$ do // check if the target is unique
    Unicity $\leftarrow$ *Unicity* $+1$
  end if
end for
return *Unicity* $/ m$

In our experiments, we focus on timestamps as a quasi-identifier, as they are common in set-valued log data and also play an important role in temporal analyses of learning processes (Knight et al., 2017), thereby being a key to balance privacy and utility. One may consider a re-identification attack in the example scenario as illustrated in Figure 2. Having access to pseudonymous log data, an adversarial third-party analyst can search for auxiliary information e.g. on online social media and link the information to record in the dataset, re-identifying the target student. Here, auxiliary information does not have to be public, but can be, for instance, physical observations or inferred from other information, and we assume that the attacker knows that the target student is in the dataset. In this case, the private information regarding the target student's learning behaviour is inadvertently disclosed to a third party without detection. In addition, all student information associated with the same ID, including possibly sensitive information such as final marks of every module, would be revealed to the third-party analyst. The unicity framework provides a means of evaluating the average-case likelihood of this type of privacy violation occurring.

[1] Code available at https://github.com/hibiki-i/too-detailed-to-share.

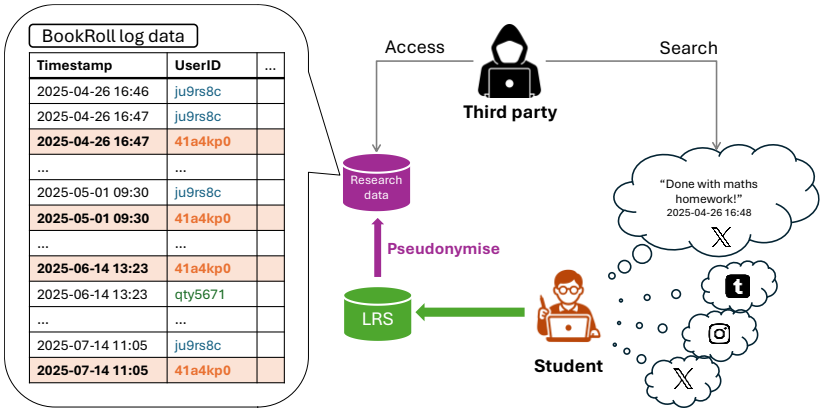

Figure 2. Example re-identification attack by the uniqueness of log trajectory

4. Results

4.1 Timestamps are personally identifiable information (PII)

Figure 3 shows the estimated unicity for different datasets with 95% confidence intervals (CI) computed by bootstrapping (the subsequent plots also show CIs in the same way). Here, ε represents the number of observations available to the attacker and the timestamps are the only quasi-identifier. That is, an attacker only knows whether a target student is active on the learning platform at certain time points (i.e. one-minute time windows). For example, the unicity of the BookRoll Secondary dataset for ε=4 is 0.913, meaning that given four time points an attacker can determine on average 91.3% of the students in the dataset. For smaller ε, the BookRoll Secondary exhibits larger unicity than the BookRoll University, while they converge to the high unicity as ε increases. This is perhaps because the BookRoll University dataset reflects less diversity by focusing on a specific context, where students regularly attend lectures at the same time, reducing the uniqueness of individual trajectories. Nonetheless, both BookRoll datasets exhibit high unicity with a few observations available to the attacker, implying that pseudonymous log data of learning behaviour is highly susceptible to re-identification.

In addition, despite learners' potential behavioural diversity due to the nature of self-paced learning, the Duolingo and EdNet datasets exhibit lower unicity compared to the BookRoll datasets, converging below 0.6 and 0.3, respectively. This is probably due to the larger numbers of individuals in these datasets. Nonetheless, these values must be interpreted with caution, as the unicity metric only evaluates the *average-case* re-identification risk. In other words, the unicity framework quantifies the risk averaged over all students, without capturing the variability in individual vulnerability—some students face higher re-identification risks than others. When protecting individual privacy, we are typically interested in the *worst-case* vulnerability—sometimes referred to as the prosecutor scenario. From this perspective, even though the unicity of the Duolingo and EdNet datasets are relatively small, sharing these datasets would still require strong security and privacy protection measures.

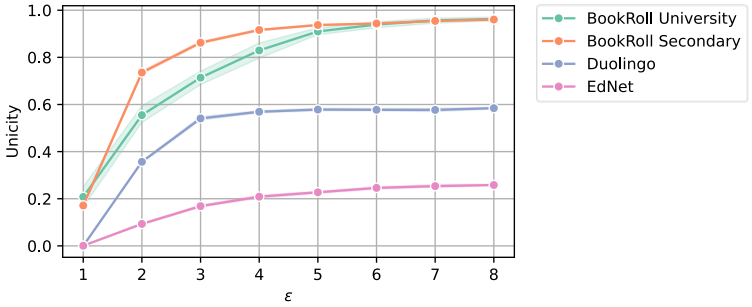

Figure 3. The unicity of different datasets with one-minute timestamps

As the literature suggests that the dataset size impacts its unicity (Barth-Jones et al., 2015; Farzanehfar et al., 2021), Figure 4 plots the unicity of each dataset for ε=4, for which we grouped the Duolingo dataset by learning languages and the unicity is separately estimated for each of them. Here, and in the subsequent plots as well, we fix ε=4, because Figure 3 tells us that unicity by and large converges at this point for all the datasets. Though it can be observed that unicity tends to decrease with the number of individuals in a dataset with a mostly convex curve (Farzanehfar et al., 2021), the trend cannot be generalised over different contexts. For example, while the BookRoll Secondary dataset consists of more students than the BookRoll University, the former shows higher unicity perhaps due to the cross-contextual nature of log data, promoting the uniqueness of each student's learning behaviour. Additionally, the EdNet dataset entails exceptionally low unicity for its number of students included. This gives us a practical implication that dataset size does not solely determine the vulnerability.

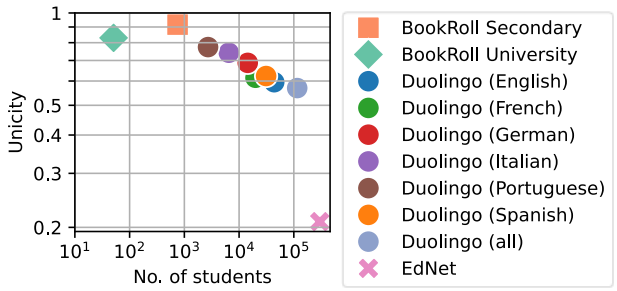

Figure 4. The unicity of different datasets for ε=4

4.2 Generalisation of timestamps

As fine-grained timestamps can be strong PII, we apply naïve coercing: the unit of timestamps are generalised to longer time windows such as quarters, hours and dates (see Figure 5). Here we assume that an attacker only has rough observation that a target student is active on the learning platform within a time window and that a time window is attributed as active if there are one or more logs within that window in the dataset. For example, if there is only one student whose (possibly multiple) BookRoll logs fall in a certain time window, the student can be re-identified by observing that the student uses BookRoll within that time window.

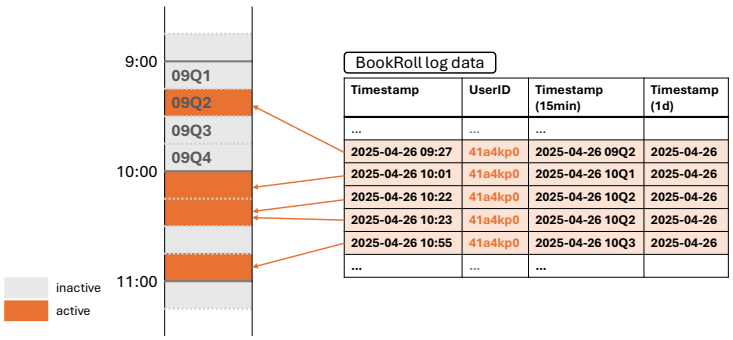

Figure 5. Generalisation of timestamps

Figure 6 illustrates the changes in unicity of each dataset when generalised timestamps with different levels (one-minute, quarter and date) are used as a quasi-identifier. Overall, the effectiveness of naïve coercing depends on the nature of a target dataset. For the BookRoll University dataset, several students are not protected from re-identification with a few timestamps even if the timestamps are generalised to dates. This is probably because students do not engage with the learning materials every day, increasing the uniqueness of

dates when each student is active on BookRoll. On the other hand, for the other three datasets, the generalisation of timestamps to dates protects almost all learners from re-identifying by at least up to eight dates, implying that this naïve coercing effectively mitigates the re-identification risk (i.e. unicity near zero).

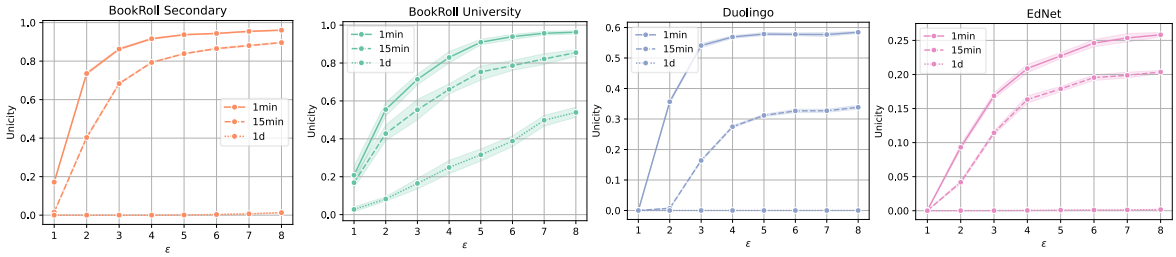

Figure 6. The unicity of different datasets with timestamp generalisation

4.3 Combining auxiliary information

While timestamps entail high risk of re-identification as a quasi-identifier, the risk can be even higher when an attacker have more auxiliary information that can be linked to the log data in question. Figure 7 illustrates how additional quasi-identifiers impact the unicity of each dataset. For the BookRoll datasets, timestamps combined with xAPI verbs would more easily determine unique students. The effect of adding the xAPI verbs to the set of quasi-identifiers is greater than adding device and context information (e.g. 10th grade maths). Moreover, adding the operation, more granular description of a log event than the verb, increases the unicity of the BookRoll University to 0.982 for ε=4. That is, almost all students are uniquely identified given four data points with timestamps and operation names.

The unicity of the Duolingo and EdNet datasets also increases with auxiliary information. Especially, the effect of identifying which target word a learner is engaged at the timestamp is remarkable, raising the unicity over 0.8. Furthermore, the information of platform (either mobile or web) and actions (like xAPI verbs) of EdNet records contributes to the elevated re-identification risk, albeit to a comparatively small extent.

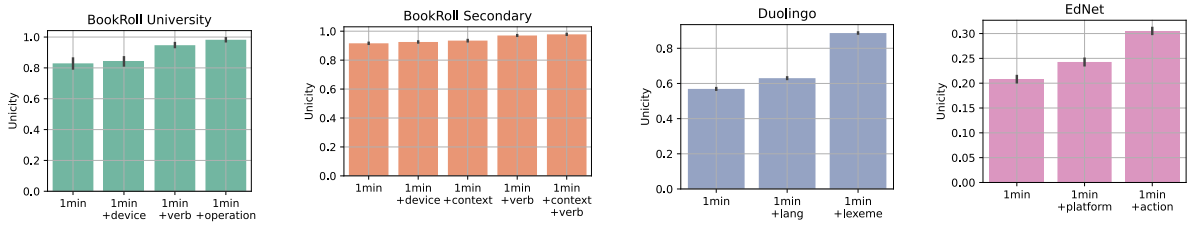

Figure 7. Unicity for ε=4 with different sets of quasi-identifiers

5. Discussion

5.1 Implications

Employing the unicity framework, we investigated the re-identification risk of micro-level educational data. Our contributions are twofold: First, the results demonstrate that educational digital trace data of relatively small size (up to several hundred individuals) are highly vulnerable to re-identification using timestamps as a quasi-identifier while larger datasets also remain unsafe to share by mere pseudonymisation. Second, naïve coercing is potentially effective for cross-contextual or large-scale data but remains ineffective for small-scale, context-specific data.

In practice, our findings help data custodians to make risk-based decisions on choosing proper security and privacy-protection measures. Data administrators and analysts

must recognise the high re-identification risk associated with pseudonymous fine-grained educational data—underscored by the unicity results in this study—as mere pseudonymisation is very common in the secondary use of educational data (Baker et al., 2024). Additionally, our findings suggest that appropriate privacy protection measures depend on the dataset size and whether the data reflects multiple educational contexts. Specifically, for both large-scale fine-grained educational data and smaller datasets covering multiple contexts, generalising timestamps can be an effective privacy-preserving strategy, provided that timestamps are the only quasi-identifier. However, our results illustrate that smaller datasets remain vulnerable despite the generalisation of timestamps, and that other information about each log such as xAPI verbs increases the vulnerability to re-identification, thereby requiring stronger privacy protection techniques like noise addition and synthetic data generation.

5.2 Limitations

A technical remark on our results is that the unicity for these datasets should be seen as lower bounds for re-identification risk. As attacker's observations are chosen from each individual's log trajectory uniformly at random (see Algorithm 1), this unicity estimation potentially biases attacker's observations toward *popular* time windows when many people are active on the learning platform, thereby possibly underestimating unicity (Achara et al., 2015). Nonetheless, our findings demonstrate that even the lower bound for unicity is high, supporting the claim that fine-grained educational data carries high re-identification risk.

Another limitation inherent to the unicity framework is that it only accounts for the uniqueness of individual trajectories. It might be possible to re-identify individuals by, for example, analysing correlations or patterns of log trajectories. Thus, in practice, unicity should not be the sole information to make risk-based decision for sharing educational data in privacy-aware manner. Nonetheless, since uniqueness is a major factor enabling direct re-identification, the assessment of unicity plays a pivotal role in the risk analysis of fine-grained educational data.

Finally, although our datasets reflect typical digital trace data in education, the generalisability of our findings should be tested in future research with various educational contexts. Particularly, data curators are encouraged to apply the unicity framework to assess the re-identification risk by themselves, when sharing pseudonymous log data with third parties.

6. Conclusion

Overall, this paper provides a first comprehensive analysis of the re-identification risk of fine-grained educational data through the unicity framework. The findings demonstrate that, despite pseudonymisation, educational log data is highly susceptible to re-identification through timestamps, and that the effectiveness of timestamp generalisation as a privacy protection strategy depends on dataset size and contextual diversity. Acknowledging the limitations of the unicity framework, our work contributes to inform the community about the re-identification risk of fine-grained educational data, encouraging data custodians to make risk-based decisions to share digital-trace data in privacy-preserving manner.

Acknowledgements

This work was partly supported by Council for Science, 3rd SIP JPJ012347 and JSPS KAKENHI Grant Number 23H00505, 25KJ1515.

References

Achara, J. P., Acs, G., & Castelluccia, C. (2015). On the Unicity of Smartphone Applications. *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*. CCS'15: The 22nd ACM Conference on Computer and Communications Security, Denver Colorado USA. https://doi.org/10.1145/2808138.2808146

Angiuli, O., Blitzstein, J., & Waldo, J. (2015). How to de-identify your data. *Communications of the ACM*, *58*(12), 48–55.

Baker, R. S., & Hutt, S. (2025). MORF: A Post-Mortem. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, (pp. 797–802).

Baker, R. S., Hutt, S., Brooks, C. A., Srivastava, N., & Mills, C. (2024). Open science and Educational Data Mining: Which practices matter most? In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 279–287). International Educational Data Mining Society. https://doi.org/10.5281/ZENODO.12729816

Barth-Jones, D., El Emam, K., Bambauer, J., Cavoukian, A., & Malin, B. (2015). [Review of *Assessing data intrusion threats*]. *Science (New York, N.Y.)*, *348*(6231), 194–195.

Bulger, M., McCormick, P., & Pitcan, M. (2017). *The Legacy of inBloom*. https://datasociety.net/wp-content/uploads/2017/02/InBloom_feb_2017.pdf

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., & Heo, J. (2020). EdNet: A large-scale hierarchical dataset in education. In *Lecture Notes in Computer Science* (pp. 69–73). Springer International Publishing.

CSIRO. (2019, April 4). *Re-Identification Risk Quantification*. Privacy Technology Research Group. https://research.csiro.au/isp/research/privacy/r4/

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*(1), 1376.

Farzanehfar, A., Houssiau, F., & de Montjoye, Y.-A. (2021). The risk of re-identification remains high even in country-scale location datasets. *Patterns (New York, N.Y.)*, *2*(3), 100204.

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130–160.

Fredrikson, M., Lantz, E., Jha, S., Lin, S. M., Page, D., & Ristenpart, T. (2014). Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. *USENIX Security Symposium*, *2014*, 17–32.

Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2017). Privacy-preserving learning analytics: Challenges and techniques. *IEEE Transactions on Learning Technologies*, *10*(1), 68–81.

Joksimović, S., Marshall, R., Rakotoarivelo, T., Ladjal, D., Zhan, C., & Pardo, A. (2022). Privacy-driven learning analytics. In *Manage Your Own Learning Analytics* (pp. 1–22). Springer International Publishing.

Knight, S., Friend Wise, A., & Chen, B. (2017). Time for change: Why learning analytics needs temporal analysis. *Journal of Learning Analytics*, *4*(3), 7–17.

Kyritsi, K. H., Zorkadis, V., Stavropoulos, E. C., & Verykios, V. S. (2019). The Pursuit of Patterns in Educational Data Mining as a Threat to Student Privacy. *Journal of Interactive Media in Education*, *2019*(1), 2.

Liu, Q., Shakya, R., Jovanovic, J., Khalil, M., & de la Hoz-Ruiz, J. (2025). Ensuring privacy through synthetic data generation in education. *British Journal of Educational Technology: Journal of the Council for Educational Technology*. https://doi.org/10.1111/bjet.13576

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-book-based learning analytics in university education. *IEEE International Conference on Consumer Electronics*, 401–406.

Prasser, F., & Kohlmayer, F. (2015). Putting statistical disclosure control into practice: The ARX data anonymization tool. In *Medical Data Privacy Handbook* (pp. 111–148). Springer International Publishing.

Sánchez, D., Martínez, S., & Domingo-Ferrer, J. (2016). Review of *Comment on 'Unique in the shopping mall: On the reidentifiability of credit card metadata'*. *Science (New York, N.Y.)*, *351*(6279), 1274.

Settles, B. (2017). *Replication data for: A trainable spaced repetition model for language learning* [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/N8XJME

Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1848–1858.

Vatsalan, D., Rakotoarivelo, T., Bhaskar, R., Tyler, P., & Ladjal, D. (2022). Privacy risk quantification in education data using Markov model. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, *53*(4), 804–821.