Privacy of Sequential Data for Learning Analytics

Anailys Hernández Julián[1,†] , Mercedes Rodríguez-García[1*,†]and Juan Manuel Dodero[1*,†]

[1] *Universidad de Cádiz, Escuela Superior de Ingeniería, Av. Universidad de Cádiz, 10, 11519 Puerto Real, Cádiz, Spain*

Abstract

Sequential data from multimodal learning experience and data source could provide the risk of background knowledge and also be exposed to third parties who may use them for malicious purposes, such as identity theft. This problem has been treated for Learning Analytics researchers with a general focus prioritizing the removal of direct identifiers over collected data. Nonetheless, the issue of collecting too much data and source anonymization methods for collecting sequential data with the aim of limiting sequential information had not been addressed until now. This research addresses the issue of collecting sequential data in a scalable manner and with a trade-off between privacy, accuracy and utility using sketching methods and differential privacy.

Keywords

Privacy, Sequential Data, Learning Analytics, Sketching, Differential Privacy

1. Goals and Research Questions

The development of new educational platforms has increased the use of student-related data of sequential nature, coming from sensors, webcams, and some other sources. Sequences are important types of data that are present in medical, security, business and some other areas, for example individual humans behave through various temporal activities [1]. We recognize that sequential data constitutes a significant portion of the information utilized for *Learning Analytics* (LA) purposes. A systematic review conducted by [2] on performance prediction in programming learning showcases various behavioral features commonly observed in such studies, including clickstream, engagement, and programming data reflecting programming behaviors.

LASI Europe 2024 DC: Doctoral Consortium of the Learning Analytics Summer Institute Europe 2024, May 29-31 2024, Jerez de la Frontera, Spain

* Corresponding author.

† These authors contributed equally.

anailys.hernandez@gm.uca.es (A. Hernández Julián); mercedes.rodriguez@gm.uca (M. Rodríguez-García); juanma.dodero@gm.uca.es (J.M. Dodero)

0009-0005-6451-5702 (A. Hernández Julián); 0000-0002-0803-4139 (M. Rodríguez-García); 0000-0002-4105-5679 (J.M.Dodero)

In [3] a review of what has been happening in *MMLA* (Multimodal Learning Analytics) shows that several research are approached empirically and in many of them the data used are audio, biometrics, video, eye tracking, log files and so on, which have sequential characteristic. Even though sequential data is not directly considered as identifying, they are still vulnerable in terms of privacy preservation. In [4], sequential biometric data is used for identification while in [5] the privacy preservation through sequential biometric keyboard data is addressed. Both studies demonstrate that sequential data of a biometric nature (such as keyboard biometrics or facial recognition) can compromise the identity of users who generate them. Another example of sequential sensitive data is mentioned in [6], which recognizes eye movement/gaze pattern as unique as an iris or fingerprint and can be used for user authentication.

Although related to the large amount of sequential data generated from LA tech scientists exposed as a privacy issue the collection of sensitive data and collecting too much data, along with challenges in remote and local storage and processing of sensitive data [7]. Nonetheless, some studies related to LA use the information related with the frequency of certain characteristics in the data and not only with its sequential nature, [8], [9], [10]. Those concerns are closely related to the issue that data should be private but at the same time useful for the purpose it is collected.

Research questions

RQ1. What are the existing privacy techniques for sequential data over LA?

RQ2. Which privacy techniques are suitable for collecting sequential data in LA without compromising students´ privacy?

RQ3. Is it possible to propose a method for collecting sequential data in a safe manner using its frequency, its sequential nature, or both?

Therefore, this work aims to achieve the following overall objective: Propose a privacy solution for collecting sequential data for LA purposes. Related to that, the following specific objectives are proposed:

1. Study privacy techniques over sequential data.
2. Propose scalable and cost-saving privacy techniques for source collecting sequential data for LA.

Based on the objectives the following research hypotheses are proposed:

1. It is possible to propose a source anonymization method for collecting sequential data for LA, in a scalable manner and with a trade-off between privacy, accuracy and utility.

2. State of Art

The MMLA research shows that all sequential data is not analyzed using only its sequential properties. In [10] certain students' interactions (click events) are quantified for applying self-regulation, in this case time is not important, just the quantity of interactions occurred. In sequential data such as *SCR*s (Skin Conductance Responses), researchers can analyze arousal levels. It is possible to differentiate features such as amplitude and rise time of peaks, but they can also be used to study temporally unfolding events by counting the number of occurring peaks per minute [11]. In [12] those metrics are used, considering both sequential and frequency characteristics of the signal. Another very well extended sequential data is eye tracking data. In [8] metrics related with the number of times students look at certain *AOI*s (Areas of Interest) are related to time management, information processing and so on. In this case the frequency characteristic is only used. On the other hand, in [13], AOI metrics such as AOI hits and dwell time are used to analyze learning styles. In this case the stored values are the current AOI observed over time and the total amount of time spent over a specific AOI, using the sequential characteristic of data. In [6], [14] the metric *JVA* (Joint Visual Attention), is known as a strong predictor of the quality of a group's interaction and success. This metric is obtained by computing the number of times a pair or team achieves joint visual attention. Also, in [14] researchers highlight they expect that a series of advanced eye-tracking data privacy-preserving mechanisms for erasing sensitive data from the raw dataset, aggregating data, enabling *DP* (Differential Privacy), and providing AOI metrics and summary data of eye movement events could be developed in the future.

Considering the privacy preserving issue in LA, solutions had been proposed with a general focus not specifically related to sequential data. The solution proposed in [15], develops a dataset, which contains data from courses presented at the Open University (OU). It contains demographic data together with aggregated clickstream data of students' interactions in the Virtual Learning Environment. The privacy solution used of ARX tool [16], which is an open-source software that implement a wide range of anonymization techniques and evaluates potential risk of deidentification. This kind of solution requires for researchers advanced knowledge about privacy preserving techniques. In [17], propose the use of DP and generalization methods. Its solution considers tabular data, but it is flexible and can be applied to unstructured and semi-structured data. [18] propose sharing synthetic data as a solution for privacy in the context of training learning models. The results demonstrate that models trained with synthetic data still have poor results when validated with the original data. Another example of privacy preserving in LA is [19], where an analysis between *k-anonimity* and DP is done, using machine learning technique for decision, but do not consider the different kind of data used in multimodal analysis. [20] proposed a privacy-protecting infrastructure for *MOOC* (Massive Open Online Courses) to facilitate secure and replicable research over data. The study addresses the challenge of balancing the need for data sharing and analysis in educational research. This solution is based on the use a of hole framework, which is a complex compared with anonymization tasks over raw data. In [21] devices were attached to students to capture physiological, spatial, and audio data, along with video recordings of sessions. To address privacy issues, identifying information was removed from the spatial and audio data dataset, and non-verbal features were utilized to preserve participants' privacy. Nonetheless authors consider the increasing use of sensors in learning analytics, calling for an open and regular discussion about potential unintended issues.

Privacy adapted to new challenges and types of sequential data in LA is an open issue. Sequential data can be mainly treated to compute the frequency of certain events, compute sequential characteristics or both. We propose to deal with this issue with the use of DP-based and sketch count methods. DP is a well-known privacy mechanism used for sequential and heterogeneous data types [22], [23], [24], [25], [26], and sketches are used to summarize data using hash functions and providing useful information about frequency and other time-related characteristics [27], [28], [29], [30]. They have the advantage of being implemented on the user´s side, adding privacy during data collection. It is also a scalable and cost-saving solution that reduce data size and storage requirements. Combinations of both methods were exposed in [31], [32], [33].

These techniques guarantee privacy preserving, utility, scalability, and the possibility of saving space for collecting sequential data. Although these works apply these techniques over general sequential data [33], [34], they have not been implemented and explored for LA-related purposes.

3. Methodology

Our proposal entails designing, developing, and evaluating privacy-preserving methods for LA sequential data, focusing specifically on techniques rooted in sketching and DP. The research methodology proposed is based on DBR (*Design-Based Research*), which includes designing, developing, and evaluating prototypical solutions [35]. DBR studies have characteristics such as the occurrence of multiple iterations and the focus on the design and evaluation of solution artifacts. For proving the proposal, we will select metrics from students' interaction data logs in *LMS* (Learning Management Systems) and eye tracking AOI, because of the multiple metrics computed over it using its frequency and sequential characteristics. (**Figure *1*** Research Methodology (DBR))

Stage 1:

- Review of experiments collecting interaction students' data logs.
- Select metrics from interaction students' data logs used in LA with frequency characteristics.
- Review of sketch methods for finding frequency items.
- Design and implement a sketch method for collecting interaction students' data logs.
- Evaluation and adjustment of the implemented privacy techniques according to utility metrics.

Stage 2:

- Review of experiments collecting eye tracking data.
- Select eye tracking metrics for AOI used in LA with frequency characteristics.
- Design and implement a sketch method for collecting eye tracking data and compute the selected AOI metrics.
- Evaluation and adjustment of the implemented privacy techniques according to utility metrics.

Stage 3:

- Select eye tracking metrics for AOI used in LA with frequency and sequential characteristics.
- Design DP and sketch methods for collecting and computing selected metrics.
- Evaluation and adjustment of the privacy techniques according to privacy and utility metrics.

Final reflections: Conduct a general analysis of the results obtained in each cycle to establish the possibility to extend the solution to other sequential data such as behavioral interactions.

4. Current status of the work and results achieved

Conducting a literature review, participating in a research stay in Coimbra, and planning future experiments using eye tracking data and data from interaction with LMS.

References

[1] G. Dong and J. Pei, Sequence data mining, vol. 33. Springer Science & Business Media, 2007.

[2] W. C. Choi, C. T. Lam, and A. J. Mendes, "A Systematic Literature Review on Performance Prediction in Learning Programming Using Educational Data Mining," in Proceedings - Frontiers in Education Conference, FIE, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/FIE58773.2023.10343346.

[3] P. Prinsloo, S. Slade, and M. Khalil, "Multimodal learning analytics—In-between student privacy and encroachment: A systematic review," British Journal of Educational Technology, vol. 54, no. 6, pp. 1566–1586, Nov. 2023, doi: 10.1111/bjet.13373.

[4] P. Delgado-Santos, R. Tolosana, R. Guest, R. Vera-Rodriguez, F. Deravi, and A. Morales, "GaitPrivacyON: Privacy-preserving mobile gait biometrics using unsupervised learning," Pattern Recognit Lett, vol. 161, pp. 30–37, Sep. 2022, doi: 10.1016/j.patrec.2022.07.015.

[5] A. Morales et al., "Keystroke Biometrics in Response to Fake News Propagation in a Global Pandemic," in Proceedings - 2020 IEEE 44th Annual Computers, Software, and Applications Conference, COMPSAC 2020, Institute of Electrical and Electronics Engineers Inc., Jul. 2020, pp. 1604–1609. doi: 10.1109/COMPSAC48688.2020.00-26.

[6] Y. Wang, S. Lu, and D. Harter, "Eye Tracking and Learning Analytics for Promoting Proactive Teaching and Learning in Classroom: A Survey," in ACM International Conference Proceeding Series, Association for Computing Machinery, Nov. 2020, pp. 156–160. doi: 10.1145/3439147.3439161.

[7] Q. Liu and M. Khalil, "Understanding privacy and data protection issues in learning analytics using a systematic review," British Journal of Educational Technology, vol. 54, no. 6, pp. 1715–1747, Nov. 2023, doi: 10.1111/bjet.13388.

[8] W. Chango, R. Cerezo, M. Sanchez-Santillan, R. Azevedo, and C. Romero, "Improving prediction of students' performance in intelligent tutoring systems using attribute

selection and ensembles of different multimodal data sources," J Comput High Educ, vol. 33, no. 3, pp. 614–634, Dec. 2021, doi: 10.1007/s12528-021-09298-8.

[9] A. A. Mubarak, H. Cao, and S. A. M. Ahmed, "Predictive learning analytics using deep learning model in MOOCs' courses videos," Educ Inf Technol (Dordr), vol. 26, no. 1, pp. 371–392, Jan. 2021, doi: 10.1007/s10639-020-10273-6.

[10] L. Silva, A. Gomes, and A. Mendes, "Investigating Students' Usage of Self-regulation of Learning Scaffoldings in a Computer-based Programming Learning Environment," in Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1, New York, NY, USA: ACM, Mar. 2024, pp. 1244–1250. doi: 10.1145/3626252.3630885.

[11] M. E. Dawson, A. M. Schell, and D. L. Filion, "The electrodermal system," in Handbook of Psychophysiology, Fourth Edition, Cambridge University Press, 2016, pp. 217–243. doi: 10.1017/9781107415782.010.

[12] H. J. Pijeira-Díaz, H. Drachsler, P. A. Kirschner, and S. Järvelä, "Profiling sympathetic arousal in a physics course: How active are students?," J Comput Assist Learn, vol. 34, no. 4, pp. 397–408, Aug. 2018, doi: 10.1111/jcal.12271.

[13] D. Bittner, F. Hauser, V. K. Nadimpalli, L. Grabinger, S. Staufer, and J. Mottok, "Towards Eye Tracking based Learning Style Identification," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jun. 2023, pp. 138–147. doi: 10.1145/3593663.3593680.

[14] Y. Wang, S. Lu, and D. Harter, "Towards Collaborative and Intelligent Learning Environments Based on Eye Tracking Data and Learning Analytics: A Survey," IEEE Access, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 137991–138002, 2021. doi: 10.1109/ACCESS.2021.3117780.

[15] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Data Descriptor: Open University Learning Analytics dataset," Sci Data, vol. 4, Nov. 2017, doi: 10.1038/sdata.2017.171.

[16] "ARX Data Anonymization Tool," https://arx.deidentifier.org/.

[17] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, "Privacy-Preserving Learning Analytics: Challenges and Techniques," IEEE Transactions on Learning Technologies, vol. 10, no. 1, pp. 68–81, Jan. 2017, doi: 10.1109/TLT.2016.2607747.

[18] B. Flanagan, R. Majumdar, and H. Ogata, "Fine Grain Synthetic Educational Data: Challenges and Limitations of Collaborative Learning Analytics," IEEE Access, vol. 10, pp. 26230–26241, 2022, doi: 10.1109/ACCESS.2022.3156073.

[19] M. Ivanova, I. Trifonova, and G. Bogdanova, "Privacy Preservation in eLearning: Exploration and Analysis," in 2022 20th International Conference on Information Technology Based Higher Education and Training, ITHET 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ITHET56107.2022.10031904.

[20] S. Hutt, R. S. Baker, M. M. Ashenafi, J. M. Andres-Bray, and C. Brooks, "Controlled outputs, full data: A privacy-protecting infrastructure for MOOC data," British Journal of Educational Technology, vol. 53, no. 4, pp. 756–775, Jul. 2022, doi: 10.1111/bjet.13231.

[21] L. Zhao et al., "Modelling Co-located Team Communication from Voice Detection and Positioning Data in Healthcare Simulation," in ACM International Conference

Proceeding Series, Association for Computing Machinery, Mar. 2022, pp. 370–380. doi: 10.1145/3506860.3506935.

[22] Y. Zhao and J. Chen, "A Survey on Differential Privacy for Unstructured Data Content," ACM Comput Surv, vol. 54, no. 10 s, Sep. 2022, doi: 10.1145/3490237.

[23] Q. Ye, H. Hu, N. Li, X. Meng, H. Zheng, and H. Yan, "Beyond value perturbation: Local differential privacy in the temporal setting," in Proceedings - IEEE INFOCOM, Institute of Electrical and Electronics Engineers Inc., May 2021. doi: 10.1109/INFOCOM42981.2021.9488899.

[24] Z. Wang, W. Liu, X. Pang, J. Ren, Z. Liu, and Y. Chen, "Towards Pattern-aware Privacy-preserving Real-time Data Collection; Towards Pattern-aware Privacy-preserving Real-time Data Collection," 2020.

[25] J. Steil, I. Hagestedt, M. X. Huang, and A. Bulling, "Privacy-aware eye tracking using differential privacy," in Eye Tracking Research and Applications Symposium (ETRA), Association for Computing Machinery, Jun. 2019. doi: 10.1145/3314111.3319915.

[26] G. Elkoumy, A. Pankova, and M. Dumas, "Differentially private release of event logs for process mining," Inf Syst, vol. 115, p. 102161, 2023.

[27] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine, "Synopses for massive data: Samples, histograms, wavelets, sketches," Foundations and Trends in Databases, vol. 4, no. 1–3, pp. 1–294, 2011, doi: 10.1561/1900000004.

[28] G. Cormode and S. Muthukrishnan, "An improved data stream summary: The count-min sketch and its applications," Journal of Algorithms, vol. 55, no. 1, pp. 58–75, Apr. 2005, doi: 10.1016/j.jalgor.2003.12.001.

[29] R. Gribonval, A. Chatalic, N. Keriven, V. Schellekens, L. Jacques, and P. Schniter, "Sketching Data Sets for Large-Scale Learning: Keeping only what you need Sketching Data Sets for Large-Scale Learning: Keeping only what you need. IEEE Signal Processing Magazine Sketching Datasets for Large-Scale Learning-Keeping Only What You Need," vol. 38, no. 5, 2021, doi: 10.1109/MSP.2021.3092574ï.

[30] L. Melis, G. Danezis, and E. De Cristofaro, "Efficient Private Statistics with Succinct Sketches," 2016.

[31] R. Balu and T. Furon, "Differentially private matrix factorization using sketching techniques," in IH and MMSec 2016 - Proceedings of the 2016 ACM Information Hiding and Multimedia Security Workshop, Association for Computing Machinery, Inc, 2016, pp. 57–62. doi: 10.1145/2909827.2930793.

[32] J. Wang and X. Li, "Secure Medical Data Collection in the Internet of Medical Things Based on Local Differential Privacy," Electronics (Switzerland), vol. 12, no. 2, Jan. 2023, doi: 10.3390/electronics12020307.

[33] Y. Li, X. Lee, B. Peng, T. Palpanas, and J. Xue, "PrivSketch: A Private Sketch-Based Frequency Estimation Protocol for Data Streams," in International Conference on Database and Expert Systems Applications, Springer, 2023, pp. 147–163.

[34] Q. Yang, F. Ji, and F. Liu, "An Efficient and Differential Privacy-Based Scheme for Aggregating Mobility Datasets," J Adv Transp, vol. 2024, 2024, doi: 10.1155/2024/5374764.

[35] S. Barab, K. Squire, S. Barab, and K. Squire, "Design-based research: putting a stake in the ground design-based research: putting a stake in the ground, Vol. 8406." 2009.

5. Appendices

A. Design Base Research Scheme Methodology

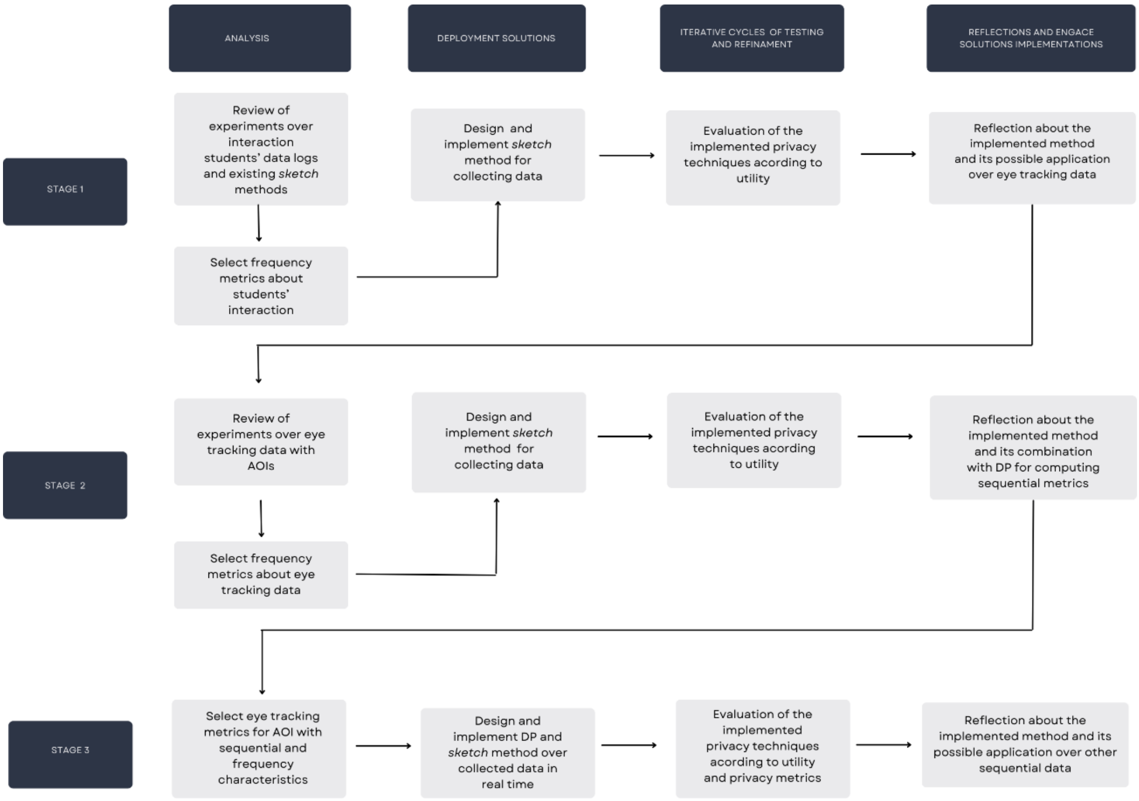
ANALYSIS
DEPLOYMENT SOLUTIONS
ITERATIVE CYCLES OF TESTING AND REFINAMENT
REFLECTIONS AND ENGACE SOLUTIONS IMPLEMENTATIONS
STAGE 1
Review of experiments over interaction students' data logs and existing *sketch* methods
Select frequency metrics about students' interaction
Design and implement *sketch* method for collecting data
Evaluation of the implemented privacy techniques acording to utility
Reflection about the implemented method and its possible application over eye tracking data
STAGE 2
Review of experiments over eye tracking data with AOIs
Select frequency metrics about eye tracking data
Design and implement *sketch* method for collecting data
Evaluation of the implemented privacy techniques according to utility
Reflection about the implemented method and its combination with DP for computing sequential metrics
STAGE 3
Select eye tracking metrics for AOI with sequential and frequency characteristics
Design and implement DP and *sketch* method over collected data in real time
Evaluation of the implemented privacy techniques acording to utility and privacy metrics
Reflection about the implemented method and its possible application over other sequential data

Figure 1 Research Methodology (DBR)