Towards Decentralised Learning Analytics (Positioning Paper)

Audrey Ekuban
audrey.ekuban@open.ac.uk
Knowledge Media Institute, Open University
Milton Keynes, UK

John Domingue
john.domingue@open.ac.uk
Knowledge Media Institute, Open University
Milton Keynes, UK

ABSTRACT

When students interact with an online course, the routes they take when navigating through the course can be captured. Learning Analytics is the process of measuring, collecting, recording, and analysing this Student Activity Data. Predictive Learning Analytics, a sub-field of Learning Analytics, can help to identify students who are at risk of dropping out or failing, as well as students who are close to a grade boundary. Course tutors can use the insights provided by the analyses to offer timely assistance to these students. Despite its usefulness, there are privacy and ethical issues with the typically centralised approach to Predictive Learning Analytics. In this positioning paper, it is proposed that the issues associated with Predictive Learning Analytics can be alleviated, in a framework called EMPRESS, by combining 1) self-sovereign data, where data owners control who legitimately has access to data pertaining to them, 2) Federated Learning, where the data remains on the data owner's device and/or the data is processed by the data owners themselves, and 3) Graph Convolutional Networks for Heterogeneous graphs, which are examples of knowledge graphs.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Security and privacy** → **Human and societal aspects of security and privacy**.

KEYWORDS

learning analytics, federated learning, blockchain, heterogeneous knowledge graph

ACM Reference Format:
Audrey Ekuban and John Domingue. 2023. Towards Decentralised Learning Analytics (Positioning Paper). In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543873.3587644

1 INTRODUCTION

Learning Analytics, a socio-technical practice, started as a research area in 2011. It is growing both in universities and other institutions, with the business sector poised to grow globally by $4.19 billion, a Compound Annual Growth Rate (CAGR) of 23%, during 2021-2025 [17]. In universities, it remains a research discipline and has also become a "Business As Usual" activity. Learning Analytics can be defined as the process of measuring, collecting, and analysing students' learning activity data to provide actionable insights. These insights can then be used to improve teaching and learning. Learning activity data include the substantial amount of digital traces a student leaves behind when interacting with a course in an online system.

Predictive Learning Analytics (PLA) is a sub-field of Learning Analytics [14]. PLA provides the required insight that enables a lecturer to intervene and offer assistance to a student, who might be at risk of not successfully completing an online module or course. It uses Machine Learning (ML), a sub-field of Artificial Intelligence (AI), to predict, for example, the likelihood that a student will fail or drop out of a course or module. While the typically centralised approach to PLA has many benefits, there are some issues which need to be addressed. These include concerns about ethics, privacy and security, and difficulties associated with its application to a student's lifelong learning. A major risk factor with PLA is its requirement for Student Activity Data to be linked to individual students. PLA utilises Supervised Machine Learning, a subcategory of Machine Learning. In Supervised Machine Learning, algorithms that learn from data are used to create Machine Learning Models, in a process called Machine Learning Training.

2 CHALLENGES IN PREDICTIVE LEARNING ANALYTICS

Due to the volume and growth of students' learning activity data, Learning Analytics can be described as the "rise of Big Data in Education" [6]. Inherent in Big Data practices are increased concerns about privacy, security, and governance of the large volume of data [11]. In addition to the Learning Analytics practice inheriting these concerns, there is additional risk in PLA. For digital trace data to provide the required insights, there is a need for the trace data to be linked to individual students. This can expose students to potentially unethical or privacy-infringing practices. Specifically, the core ethical and privacy issues in Learning Analytics being discussed globally include "transparency; data ownership and control; accessibility of data; validity and reliability of data; institutional responsibility and obligation to act; communications; cultural values; inclusion; consent; student agency and responsibility" [16].

The Open University's policy on the Ethical use of Student Data for Learning Analytics [18] identified eight principles to guide the ethical use of student data in its policy statement.

- Principle 1: Learning Analytics should be an ethical practice
- Principle 2: The university has a responsibility to all stakeholders to use and extract meaning from student data for the benefit of students.

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587644

- Principle 3: Students should not be wholly defined by their visible data or interpretation of that data.
- Principle 4: The purpose and the boundaries regarding the use of Learning Analytics should be well-defined and visible.
- Principle 5: The University should be transparent regarding data collection, and provide students with the opportunity to update their own data and consent agreements at regular intervals.
- Principle 6: Students should be engaged as active agents in the implementation of Learning Analytics (for example, informed consent, personalised learning paths, and interventions).
- Principle 7: Modelling and interventions based on the analysis of data should be sound and free from bias.
- Principle 8: Adoption of Learning Analytics requires broad acceptance of the values and benefits and the development of appropriate skills.

One socio-technical concern not mentioned above, and the main motivation for a decentralised approach, is that of surveillance or "dataveillance" [2]. There have been discussions as to whether there is a difference between Learning Analytics and Learning Surveillance, with some taking the view that Learning Analytics cannot exist without surveillance [3]. There are also questions surrounding consent. The DELICATE checklist provides guidelines for data collection, which include seeking consent before collecting data, using clear and understandable consent questions with Yes/No options, and offering the option to opt-out of data collection without any negative consequences [4].

The authors of [16] question the legitimacy of many of the current approaches to consent. In considering the EU's GDPR, it had been suggested that, if data were not considered sensitive, consent would not be required for data collection, on the basis that there is a legitimate interest [13]. However, it is possible that sensitive information could be gleaned from non-sensitive data. The authors of [15] also question that consent is not required for non-sensitive data, arguing that data which may be non-sensitive in one context at a particular time, may be sensitive in another context and/or another time. The guidance that legitimate interest could be used by UK Higher Education Institutions (HEIs), has since been updated to state the need for a Data Protection Impact Assessment (DPIA), a process that helps an organisation to identify the data protection risks of a project. These risks include the rights and freedoms of individuals.

There have been some recent studies that have involved students. At Ulster University [7], some misgivings were expressed about a university finding out information that a student might not want the university to know. The findings from that study suggest that there are three groups of students. The authors named these groups "(1) naive and trusting, (2) cautious and compromising, and (3) enlightened and demanding". Group 1 consisted of those who did not see a problem with the university using their data. Group 2 consisted of those who understood how activity data was used but felt it would be nice to have more transparency. Group 3 consisted of those who felt strongly that students should be allowed to opt out of data collection. Whilst, at University of Michigan [8], the researchers found that female students were more likely to trust the institution and use of their data by instructors, but were also more generally concerned about data collection practices. Black students indicated lower levels of institutional trust. White students were over-represented, while Black students were under-represented among people who made a consent decision to data collection. It could be argued that the collection of a student's activity data in an online course, where explicit consent is not granted, could impact the way in which a student interacts with the course, especially if the student does not trust the institution.

2.1 Data Protection

Data Protection Regulations govern how organisations use personal information. The EU and UK GDPR have similar definitions for Data Controllers and Data Processors.

Data Controllers (Art. 24 of the EU GDPR) are the decision-makers when it comes to the purposes and means of processing personal data. It will therefore be the responsibility of the Data Controllers in an HEI to decide to adopt Decentralised Learning Analytics, and to ensure that any DPIAs are conducted, as is the case now.

A Data Processor (Art 28 of the EU GDPR) processes personal data on behalf of the Data Controller. In the case of Decentralised Learning Analytics, the Data Processors are data scientists and developers, who may or may not be in-house. This is the same as it is in centralised data. However, there may be differences in the technical knowledge required when the data is decentralised.

Both the EU and UK GDPR have the notion of Data protection by design and by default. In the UK it is a legal requirement.

2.2 Privacy By Design

There are seven principles which govern Privacy By Design [1], which if adhered to, satisfy "Data protection by design and default".

- Proactive not Reactive; Preventative not Remedial: Privacy By Design helps to prevent the invasion of privacy, by anticipating the risk of that invasion happening. Linking different database instances to explore an individual's unique fingerprint is a significant privacy threat, and should therefore not occur.
- Privacy as the Default: In the Privacy by Design approach the default privacy is the one which gives the most privacy protection. This suggests that, in the case of informed consent, a student should be presumed to not have opted into Learning Analytics until the student has decided to opt-in.
- Privacy Embedded into Design: Privacy should be embedded into the design of IT systems and business practices.
- Full Functionality—Positive-Sum, not Zero-Sum: The Privacy by Design approach seeks to allow all legitimate interests to be taken into account.
- End-to-End Lifecycle Protection: "Privacy by Default" and "Privacy Embedded into Design", ensures that the Privacy by Design approach extends throughout the entire lifecycle.
- Visibility and Transparency ("trust but verify!"): All stakeholders should be assured that all stated promises and objectives, subject to verification, are being adhered to.
- Respect for User Privacy ("Keep it user-centric!"): The interests of the individual should be paramount.

3 EDUCATIONAL MINING WITH PRIVACY RIGHTS AND ETHICS FOR STUDENT SELF-SOVEREIGNTY (EMPRESS)

It is envisaged that EMPRESS will be a Decentralised Machine Learning Pipeline which will contain several mechanisms to enable, as much as possible, the eight principles in The Open University's policy on the Ethical use of Student Data for Learning Analytics [18]. In the first instance, it will focus on Principle 5 and Principle 6. Principle 5 focuses on transparent data collection. Principle 6 states that students should be engaged as active agents in the implementation of Learning Analytics, with informed consent being an example of this. It is envisaged that student engagement would alleviate any mistrust that students have for Learning Analytics.

A Privacy by Design approach will be adopted as follows:

- Proactive not Reactive; As a solution, the aim of EMPRESS is to eliminate the need to collect digital trace data.
- Privacy as the Default: In EMPRESS, a student would be presumed to not have opted into Learning Analytics until the student has decided to opt-in.
- Privacy Embedded into Design: EMPRESS is being designed with privacy as a key factor.
- Full Functionality—Positive-Sum, not Zero-Sum: EMPRESS will seek to demonstrate that it is possible for students to have both privacy and the benefits from Predictive Learning Analytics.
- End-to-End Lifecycle Protection: "Privacy by Default" and "Privacy Embedded into Design" will ensure that the Privacy by Design approach extends throughout the entire lifecycle of EMPRESS and any data involved.
- Visibility and Transparency ("trust but verify!"): The EMPRESS framework will utilise tools or techniques to implement Input and Output Verification, and Flow Governance. It is anticipated that these tools / techniques will give assurance of any stated promises and objectives.
- Respect for User Privacy ("Keep it user-centric!"): In the case of the EMPRESS framework the individual is the student. By focusing on student self-sovereignty, EMPRESS will be inherently student-centric.

By focusing on student privacy and associated privacy rights concerns that are inherently linked to Learning Analytics, EMPRESS aims to reduce the socio-technical concerns previously identified, as it relates to mainstream Predictive Learning Analytics (PLA) in Higher Education Institutions (HEIs). In addition, EMPRESS will facilitate the integration of learning taken outside of the HEI and beyond. It is proposed to combine the following (see Figure 1:

- Self-sovereign data, where humans have ownership and control over their data. "Data ownership and control" is one of the ethical and privacy issues in Learning Analytics identified in [16]. In Predictive Learning Analytics, not only is the student activity data created by a student, but it is also linked to that student. Self-Sovereign Learning Analytics will allow students to control who legitimately has access to data pertaining to them.
- Federated Learning, a concept introduced by Google in 2016 [9]. Federated Learning enables Machine Learning models to be trained without the data owners having to share their data. Federated Learning is used to develop predictive text models while maintaining data privacy. With Federated Learning, multiple devices or entities can collaboratively train a model without sharing raw data. Each device trains a local model on its data and sends the model's parameters to a central server, which aggregates the parameters and sends updated parameters back to each device. This process is repeated iteratively until the model converges to a desirable level of accuracy.
- Relational Graph Convolutional Neural Networks [12], which provides a way of training heterogeneous graphs. An heterogeneous graph is composed of multiple node types and edges with various relation types. A Knowledge Graph can be considered to be a type of heterogeneous graph. Nodes are knowledge graph entities which are labelled with their types. Edges between two nodes capture relationships between the knowledge graph entities.
  The use of Heterogeneous Graphs in Learning Analytics [10] is an emerging area of research. We envisage that the hierarchical relationships between courses, topic links and assessments could be considered. This would enable EMPRESS to employ "Graph Embedding Based Recommendation Techniques" [5].

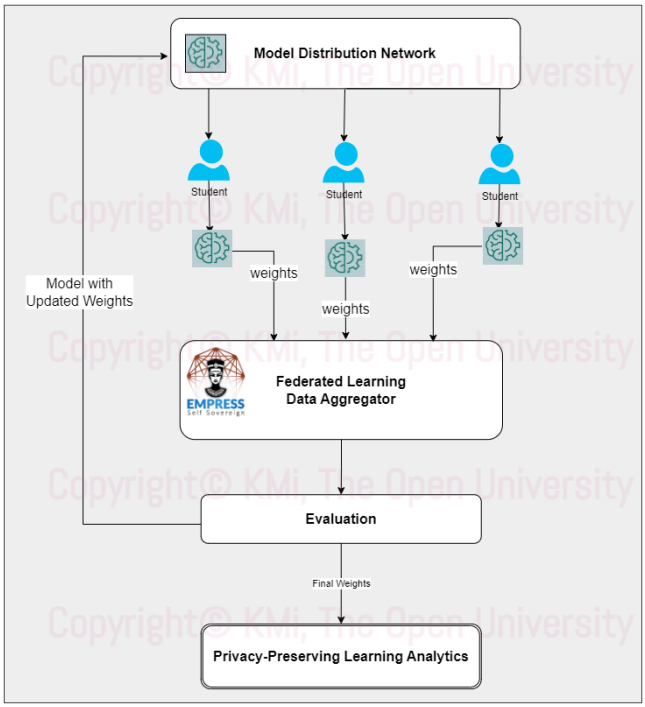

Figure 1: Proposed solution based on work to date

3.1 Technical Aspects

There are 3 parts to EMPRESS: EMPRESS Administrator, EMPRESS Processor, and EMPRESS Learners. These represent the Data Controller, Data Processor, and students, respectively.

3.1.1 EMPRESS Administrator. As a consent mechanism, it is suggested that the EMPRESS Administrator implements an incentivisation method based on cryptocurrency, a digital currency that is not reliant on a central authority. A cryptocurrency wallet, which can be a device or a program, stores the private and public keys. Crypto tokens are managed by smart contracts, called a Token Contract, and have a name, and a symbol. Smart contracts consist of code which implements a set of rules, and which run on a blockchain. They are also known as self-enforcing agreements. The token contract has a contract owner. Both the Token Contract and the Token Contract Owner have a public address, hence all currency transactions and the contract itself, can be verified, if necessary. An example of a decentralised Application Infrastructure is shown in Figure 2.

The above approach to a consent mechanism will facilitate storing a student's consent decision on a Blockchain, and reading the student's consent decision from the Blockchain. The default student consent will be "Opt-Out".

3.1.2 EMPRESS Processor. As privacy mechanisms are added to EMPRESS, the blockchain can also be used to indicate which EMPRESS processors offer which privacy mechanisms. In figure 1, the EMPRESS Processor has provided a Federated Learning Data Aggregator. Future work will address further privacy mechanisms, as well as decentralised predictions.

3.1.3 EMPRESS Learners. EMPRESS Learners are responsible for their consent preferences and the privacy mechanisms they wish to use. Future work will enable EMPRESS Learners to collaborate with each other.

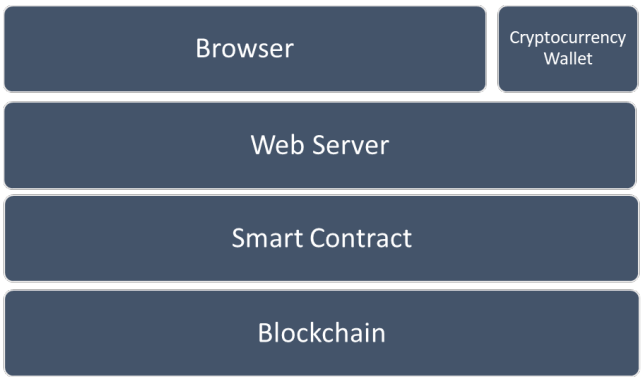

Figure 2: Blockchain Application Layers

4 CONCLUSION

This positioning paper outlines socio-technical concerns surrounding Predictive Learning Analytics, which are being currently discussed globally. The major discussions are those surrounding informed consent and surveillance concerns. As the field of Predictive Learning Analytics continues to grow, it is important for educational institutions and policymakers to carefully consider the opportunities as well as the challenges presented.

This paper introduced EMPRESS, a decentralised Machine Learning pipeline, with Blockchain technology as a supporting role. EMPRESS combines solutions for 1) Self-sovereign data, where students have ownership and control over their data. 2) Federated Learning, where data scientists are able to build Machine Learning models using data that is not visible to them, and 3) Graph Convolutional Networks for Heterogeneous graphs, which are examples of knowledge graphs.

Decentralised Predictive Learning Analytics gives control back to students, alleviates the perception that Learning Analytics is a surveillance tool, and reduces the risks associated with Centralised Predictive Learning Analytics. Additionally, the use of blockchain technology provides transparency to student consent decisions.

REFERENCES

[1] Ann Cavoukian. 2010. Privacy by design: the definitive workshop. A foreword by Ann Cavoukian, Ph. D. *Identity in the Information Society* 3, 2 (2010), 247–251.
[2] Roger Clarke. 1988. Information technology and dataveillance. *Commun. ACM* 31, 5 (1988), 498–512.
[3] Matt Crosslin. 2019. *Is Learning Analytics Synonymous with Learning Surveillance, or Something Completely Different?* Retrieved February 12, 2023 from https://www.edugeekjournal.com/2019/10/30/is-learning-analytics-synonymous-with-learning-surveillance-or-something-completely-different/
[4] Hendrik Drachsler and Wolfgang Greller. 2016. Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge*. 89–98.
[5] László Grad-Gyenge, Attila Kiss, and Peter Filzmoser. 2017. Graph embedding based recommendation techniques on the knowledge graph. In *Adjunct publication of the 25th conference on user modeling, adaptation and personalization*. 354–359.
[6] Kyle ML Jones. 2019. Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. *International Journal of Educational Technology in Higher Education* 16, 1 (2019), 1–22.
[7] Paul Joseph-Richard and James Onohuome Uhomoibhi. 2021. Ethics in Predictive Learning Analytics: An Empirical Case Study on Students Perceptions in a Northern Irish University. In *Advancing the Power of Learning Analytics and Big Data in Education*. IGI Global, 86–107.
[8] Warren Li, Kaiwen Sun, Florian Schaub, and Christopher Brooks. 2022. Disparities in students' propensity to consent to learning analytics. *International Journal of Artificial Intelligence in Education* 32, 3 (2022), 564–608.
[9] Brendan McMahan and Daniel Ramage. 2017. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog* 3 (2017). Retrieved March 13, 2023 from https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
[10] Ahmed A Mubarak, Han Cao, Ibrahim M Hezam, and Fei Hao. 2022. Modeling students' performance using graph convolutional networks. *Complex & Intelligent Systems* 8, 3 (2022), 2183–2201.
[11] R Rawat and R Yadav. 2021. Big data: Big data analysis, issues and challenges and technologies. In *IOP Conference Series: Materials Science and Engineering*, Vol. 1022. IOP Publishing, 012014.
[12] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 593–607.
[13] Niall Sclater. 2016. Developing a code of practice for learning analytics. *Journal of Learning Analytics* 3, 1 (2016), 16–42.
[14] Niall Sclater, Alice Peasgood, and Joel Mullan. 2016. Learning analytics in higher education. *London: Jisc. Accessed February* 8, 2017 (2016), 176.
[15] Sharon Slade and Paul Prinsloo. 2013. Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist* 57, 10 (2013), 1510–1529.
[16] Sharon Slade and Alan Tait. 2019. Global guidelines: Ethics in learning analytics. (2019).
[17] Technavio. 2021. *Global Learning Analytics Market 2021-2025*. Technavio. Retrieved February 10, 2023 from https://www.researchandmarkets.com/reports/5509402/global-learning-analytics-market-2021-2025
[18] Open University. 2014. *Policy on Ethical use of Student Data for Learning Analytics*. Open University. Retrieved February 12, 2023 from https://help.open.ac.uk/documents/policies/ethical-use-of-student-data/files/22/ethical-use-of-student-data-policy.pdf