Learning Analytics and Privacy: A Library Perspective

Soo-yeon Hwang <syhwang@shsu.edu>

Michael Hanson <hansonm@shsu.edu>

Sam Houston State University

Corresponding author's contact information:

Soo-yeon Hwang <syhwang@shsu.edu>

Sam Houston State University, Box 2179, Huntsville TX 77341

Abstract

This column discusses learning analytics in higher education, including the involvement of academic libraries, privacy concerns, and some elements to consider when developing and deploying learning analytics systems ethically.

Keywords: learning analytics, higher education, academic library, big data analytics, privacy, EZproxy

For the past several years, one of the most transformative aspects of our lives has been the manipulation of enormous amount of data generated by our interaction with the ubiquitous computing that permeates our existence by large institutions and companies. Educational institutions have also willingly participated in this phenomenon. Students and faculty generate data from interacting with learning software and other activities at school, as well as being heavy users of social media. Schools and companies seek to use that data in various ways that benefit themselves. This data practice around education has been dubbed Learning Analytics (LA), which Siemens defines as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs (Siemens, 2013, p. 1382)."

In the higher education setting, LA is employed for numerous reasons, such as identifying students at risk to fail a course, suggesting students to see a professor or seek academic advising (e.g. changing major), adapting courses according to student learning behaviors, providing personalized program and course suggestions, providing feedback to professors on course or instruction effectiveness, and identifying needs of distinct student populations (Jones et al., 2020).

LA is increasingly popular in academic libraries as well. Academic libraries have long harbored a good amount of data about students' use of library resources such as circulation statistics and electronic resource usage. Lately, a growing number of libraries have been participating in their parent institutions' LA efforts using the library patron data (e.g. Kabo et al., 2021; LeMaistre et al., 2018; Nurse et al., 2018). Academic libraries have much to gain from doing so: Libraries can prove their value to the university by demonstrating quantitatively how library use contributes to student success and other strategic goals of the university. Libraries can identify underserved groups and improve overall library services. The analyzed data can be used to assist librarians to improve instruction and reference practices. To these ends, there is a great incentive to broadly collect data.

With the ubiquitous computing environment came nearly ubiquitous collection of data on those using the information technology. Big data mining practices often start with systems collecting data first, and then administrators returning later to look for patterns from the data. Libraries follow a similar pattern. Library patron data comes from multiple sources: Systems such as Integrated Library System or Library Management System (circulation data), library website and LibGuides (Google Analytics and Springshare statistics), electronic databases (usage statistics from the vendors), authentication software such as EZproxy or OpenAthens (electronic resource access logs), and RFID/space counters (building access logs) all generate a substantial amount of data. Furthermore, librarians gather much data around instruction and reference services (chat texts, emails, and phone calls).

Many library systems have been broadly gathering data. Sometimes this was to prevent illegal use of "paid for" library resources and looked to be able to identify improper use to meet resource provider expectations. Other times it is a recognition that the more data collected, the more opportunity there is to discover patterns that could assist patrons. Kabo et al. (2021) and LeMaistre et al. (2018) analyzed the authentication software (EZproxy) access logs and found that the use of online library resources was a significant predictor of student success. These logs can be expansive in their data collection. EZproxy, especially, captures all of the access activities through proxy, not just at the time of login. Moreover, we had a chance to examine some sample lines of EZproxy logs at our institution and found that the patron's university ID was visible in every line. This would make it easy to associate it with other university data which share the same unique identifier (university ID). When combined with requiring proxy authentication for every electronic resource access whether on- or off-campus, one can gather a wealth of usage data. However, we also found the amount of personally identifiable information collected in the logs disconcerting. It was difficult not to feel that this much data as a breach of patron privacy.

Librarians have long held ethics that protect library patrons' rights to find, consume, and share information. Many librarians are concerned that the data mining practices in LA conflict with some ethical principles in the American Library Association's Code of Ethics, particularly patron privacy, intellectual freedom, and intellectual property rights (Jones & Salo, 2018). Protecting patron information is perhaps easier to justify when an outside entity is seeking the information such as the U.S. government's attempt to get information from libraries under the USA Patriot Act than gathering and employing it for ourselves or our parent institution.

Although it is tantalizing to gather as much data as possible with the intention to positively improve library services, autonomy and privacy of the library patrons cannot be overridden. We need to question if it is ethical to collect this much data in this granularity (i.e. individual level). Rather than sweeping up all the patron data we can and then worrying about how we manage and use it afterwards, librarians should ask mindful, ethical questions first, and then decide what and how we collect and manage data to answer our questions, allowing for ethical practices along the way.

First, we should consider the data granularity. Will we be able to answer the same questions that we have now with group-level data, instead of individual-level data? In other words, can we remove personally identifiable information at the time of data collection and still achieve the main goals of LA? We will be limited to the questions we asked at the beginning of the data collection (e.g. the types of patron groups such as those based on diversity, equity, and inclusion measures), but there will be less invasion of individual privacy.

One of the most touted features of LA is individual interventions such as nudging at-risk students for extra advising. However, there are ethical questions with interventions. Advocates of intervention argue institutions have the responsibility to analyze their student's data and employ it to improve student success, including both passive and active/intrusive interventions. Those less comfortable with intervention worry that students will lose agency or will feel the data is affirming that

they will not succeed in college (Fritz & Whitmer, 2019). If group-level data collection is adopted to protect privacy, the lack of granularity may make providing personalized interventions harder or impossible to accomplish.

Moreover, key to any LA system should be highly visible privacy features and educating users about these options, as well as clearer privacy policies. The most ethical practice would be to deploy LA system in an opt-in model; in other words, instead of opting in everyone by default, only those users who opt-in voluntarily should be included in LA practices. In addition, if users choose to, it should be easy to opt out and/or delete themselves completely from the LA system. They should also be able to download and take their own data with them before deletion.

Lastly, a student advisory board or committee should be involved in the development and deployment of LA on their campus because the largest user base of higher education systems, from whence LA data is extracted, is students. Their active involvement and oversight will help the LA system developers and administrators to create and operate systems that are easier to use and protect user privacy better.

Jones and Salo (2018) suggest that librarians should be deeply involved in the development and deployment of LA not just for pragmatic gains such as proving libraries' value or improving library service, but for ethical shaping of technology and policies across the university. The LA trend is here to stay. The question is how librarians can influence its course of development in a way that benefits all the actors involved (the university, library, faculty, and students) and yet protects user privacy and intellectual freedom.

References

Fritz, J., & Whitmer, J. (2019). Ethical learning analytics: "Do no harm" versus "do nothing." *New Directions for Institutional Research*, *2019*(183), 27–38. https://doi.org/10.1002/ir.20310

Jones, K. M. L., Briney, K. A., Goben, A., Salo, D., Asher, A., & Perry, M. R. (2020). A comprehensive primer to library learning analytics practices, initiatives, and privacy issues. *College & Research Libraries*, *81*(3), 570–591. https://doi.org/10.5860/crl.81.3.570

Jones, K. M. L., & Salo, D. (2018). Learning analytics and the academic library: Professional ethics commitments at a crossroads. *College & Research Libraries*, *79*(3), 304–323. https://doi.org/10.5860/crl.79.3.304

Kabo, F., Paulson, N., Bradley, D., Varnum, K. J., & Teasley, S. (2021, March 17). *Associations between library usage and undergraduate student GPA, 2016-2019*. Library Assessment Conference. https://doi.org/10.7302/794

LeMaistre, T., Shi, Q., & Thanki, S. (2018). Connecting library use to student success. *Portal: Libraries and the Academy*, *18*(1), 117–140. https://doi.org/10.1353/pla.2018.0006

Nurse, R., Baker, K., & Gambles, A. (2018). Library resources, student success and the distance-learning university. *Information and Learning Science*, *119*(1/2), 77–86. https://doi.org/10.1108/ILS-03-2017-0022

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