Some Thoughts on Privacy and Security for Educational Data

Ryan S. Baker

University of Pennsylvania

NORTHEAST BIG DATA INNOVATION HUB

Penn

Penn Center for Learning Analytics

- Conducting research on the data becoming available from online learning

Selected Projects

- Replicating findings about success in MOOCs across dozens of MOOCs (Andres et al., in press a, in press b)
- Connecting performance and behavior in MOOCs to participation in community of practice (Wang et al., 2014, 2016)
- Connecting performance and behavior in middle school mathematics homework to college enrollment and major (San Pedro et al., 2013, 2015)

Common Thread Across Many of our Projects

- Connecting fine-grained data at time A
- With outcome data at time B

- Requires integrating across data sources
- Important to do so in a fashion that is both secure and protects privacy

Value of Longitudinal Research

- The educational practices that are effective in the short-term are not always effective in the long-term

- Example: Cramming for the test
  - Leads to better performance on the test!
  - Leads to much more forgetting after the test (Tigner, 1999; Kornell, 2009)

Value of Longitudinal Research

- Only by integrating data on performance and behavior during learning
- With data on long-term outcomes

- Can we understand which behaviors and strategies are most important for student long-term success

Value of Longitudinal Research

- If we can’t link to longitudinal and external outcomes in some fashion
- Automated optimization algorithms will end up optimizing for within-system performance
- Probably hurting long-term student outcomes

Privacy Issues in Educational Data

- Certain types of educational data are protected under federal law – FERPA
  - Specific types of Personally Identifiable Information (PII)

- Education now generating a lot of data not clearly covered under existing law
  - Online learning data
  - Discussion forum data

Deidentification

- Essentially impractical to fully deidentify discussion forum data

“Hi everyone! I’m [name] and I’m a certified public accountant in [town]. My boss down at [business] suggested I take a look at this course, and I have to say I’ve found it very useful.”

Deidentification

- There ***is*** a question whether this learner ever meant their identity to be private, but that’s a different story…
- And who wants their discussion forum posts from when they are 19 following them forever?

Deidentification

- Even online learning data with no obvious identifiers can sometimes be reidentified

Real-World Example

- Student made unusual error “74” in online math homework
- Student tweeted about their unusual error “74”
- By combining the value “74” and the time in the interaction log data, it was possible to determine exactly who the student was

Real-World Example

- Student made unusual error “74” in online math homework
- Student tweeted about their unusual error “74”
- By combining the value “74” and the time in the interaction log data, it was possible to determine exactly who the student was
- And also to reidentify the school identifier for a lot of other students, giving more converging evidence on them as well

That said…

- There isn’t huge risk in figuring out which students are doing better or more poorly in their math homework…

That said…

- There isn’t huge risk in figuring out which students are doing better or more poorly in their math homework… or is there?

That said…

- There isn’t huge risk in figuring out which students are doing better or more poorly in their math homework… or is there?

- Is it possible that students who show specific disengaged behaviors during high school learning may eventually be less likely to get a college loan?

Parental Concern

- Online learning data will be used to advertise commercial services

- A real concern?

Parental Concern

- Online learning data will be used to advertise commercial services

- A real concern?

- It really happens… some university-level learning management systems recommend commercial tutoring services to struggling students
  - OK or not?

Concern

- Still relatively few reports of educational data breaches or harm from educational data breaches
  (Bienkowski et al., in press)

Concern

- Still relatively few reports of educational data breaches or harm from educational data breaches
  (Bienkowski et al., in press)
  - But example: DC Public schools accidentally posted disability status for 12,000 children

K-12 Parental Concern

- A great deal of parental concern about this in some places

- We’re seeing the emergence of a movement very concerned with student privacy

K-12 Parental Concern

- A great deal of parental concern about this in some places
- We’re seeing the emergence of a movement very concerned with student privacy
- Led to disbanding of InBloom initiative

Emergence of organizations

- Such as one “school privacy consortium” organization whose leadership is predominantly made up of security consulting firms (4)

- Recommends very restrictive contract to schools that – for example – bars use of data for research or enhancement of educational quality
- Recommends security audits to schools and compliance certification of vendors
- Non-profits and university-based free learning software being barred from schools

Summary

- Creating high-quality online learning is greatly facilitated by linked longitudinal data
- There are real reasons for concern about data privacy
- But the steps being taken do not always match the risks

Some directions

Legal agreements not to attempt to re-identify data

- Increasingly adopted by online learning systems that share data for scientific research

Link data through trusted brokers

- Create brokers who have PII, who can link together data sets for use in longitudinal outcome research

- One example of this is the Pittsburgh Science of Learning Center DataShop (Koedinger et al., 2010), which conducts this service for researchers using their LearnLab school sites

Conduct analyses on secure servers

- Conduct analyses on secure servers, where the data identifiers are present and can be used to link data, but cannot be directly accessed

- Can be possible to hack, but probably acceptable for data where risk is relatively minimal anyways

MORF
MOoc Replication Framework

- Project just getting started at UPenn where researchers can submit if-then questions or code to be run on our MOOC data

- Used to replicate 15 research questions across 29 MOOCs, using external researcher’s code (Andres et al., in press)

This community has a lot to contribute

- I’m first and foremost a scientific researcher with educational data, although I ***do*** manage UPenn’s efforts to use MOOC data for research

- Please let me know how I can help connect you to the developers and researchers who could use your expertise to protect student privacy while enhancing their learning

Thank you!

Penn Center for
Learning Analytics

twitter.com/BakerEDMLab

Baker EDM Lab

WE ARE RECRUITING A POSTDOC
“Big Data and Education” on edX, June 18
All lab publications available online – Google “Ryan Baker”

NSF
DEPARTMENT OF EDUCATION UNITED STATES OF AMERICA
ARL
LearnLab
Pittsburgh Science of Learning Center
BILL & MELINDA GATES foundation