Harvard Journal of Law & Technology
Volume 28, Number 2 Spring 2015

EDUCATIONAL PRIVACY IN THE ONLINE CLASSROOM: FERPA, MOOCS, AND THE BIG DATA CONUNDRUM

*Elise Young**

TABLE OF CONTENTS

I. INTRODUCTION

Massively open online courses ("MOOCs") are virtual classrooms that run on the Internet. In addition to their educational functions, MOOCs collect, centralize, and analyze massive amounts of information about their students. This information can include education records, student performance, and even how, when, and where a student clicks each time she logs in. Such widespread information collection and analysis is colloquially known as "big data."

* Harvard Law School, J.D. 2014. Thanks to Amy Rossignol, for her comments and revisions, and to my Article Editor, Jennifer Chung, for her patience and critical feedback. Thanks to Harvard Law School's Summer Academic Fellowship for support during writing this piece. Finally, thanks to the editors of the Harvard Journal of Law & Technology for their dedication and support throughout this process. All opinions are the author's.

Big data is shorthand for the ability to store so much information that even trivial details can be kept and analyzed for emergent trends in areas such as consumer preferences, economic development, and crime mapping. Despite its large scale, big data raises privacy concerns on an individual level because it also excels at revealing unexpected correlations that may disclose not only someone's identity but some new "fact" about that person. Big data thus has the potential to actually create personally identifiable information without affirmative action on the part of the user whose data was collected. This dynamic may violate certain statutory privacy protections.

The Family Educational Rights and Privacy Act ("FERPA")[1] is one such privacy statute. However, FERPA is so dated that when confronted with a technology that can collect and use big data, like MOOCs, the statute practically breaks down. This Note examines the individual privacy concerns implicated by big data in general, assesses whether the privacy language of FERPA can address big data collection and analysis in the MOOC context, and provides broad suggestions for updating FERPA so that it may better adapt to big data privacy concerns.

Registering for a MOOC goes something like this: To take a course — perhaps Astrophysics or Introduction to Philosophy — you must first create an account with your selected MOOC provider. One example is edX, a nonprofit founded by Harvard and MIT in 2012.[2] To register with edX, you create a username and password and provide the following information: your email address, full name, country, and, optionally, your gender, year of birth, highest level of education completed, and reason for registering.[3] To complete registration, you must also agree to edX's terms of service and honor code.[4]

None of these requirements are particularly unique for website registration. But once you begin your selected MOOC, you create a new set of data points over the duration of your participation, collected by the MOOC. Among other things, edX logs when you access a module (the virtual equivalent of a classroom unit), how often you come back to the module, how long it takes for you to complete a

1. 20 U.S.C. § 1232g (2012).

2. *See MIT and Harvard Announce edX*, HARVARD GAZETTE (May 2, 2012), http://news.harvard.edu/gazette/story/2012/05/mit-and-harvard-announce-edx.

3. *See Register*, EDX, https://courses.edx.org/register (last visited May 8, 2015). EdX also allows you to log in using your Google or Facebook profile. Using this method still requires submitting your name, email, username, and country. *Id.* In addition, if you register with Facebook, edX can access certain public information, including your profile picture, gender, age range (e.g., 21+), and information regarding your friends. *See Data Policy*, FACEBOOK, https://www.facebook.com/policy.php (last visited May 8, 2015); *What Is Public Information?*, FACEBOOK, https://www.facebook.com/help/203805466323736 (last visited May 8, 2015).

4. *See Register*, EDX, https://courses.edx.org/register (last visited May 8, 2015).

quiz, your quiz scores, how many times you watch a video, and whether you stop visiting a course, resume it, and then stop again.[5] In all, edX collects approximately twenty gigabytes of user data per course[6] — the equivalent of millions of physical pages of information.[7]

This collection is completely unremarkable in the online context. Indeed, it is ubiquitous across essentially all services on the Internet. Advertisers and other service providers can track a person's web browsing using cookies[8] or the more sophisticated "canvas fingerprint."[9] Other software applications can log keystrokes or record behavior by observing how a user's mouse moves across a webpage.[10] Such tracking is not the only way to collect data. Many individuals willingly offer real-time updates on their activities through Twitter posts, Instagram photographs (which can be geo-tagged[11]), and Facebook statuses. Personal profiles, whether created by individuals or data brokers,[12] are increasingly detailed, transferable, predictive, and profitable.[13] Currently, most of this data is put to benign uses such as

5. *See* Jon Daries, *The HarvardX-MITx Person-Course Dataset AY2013*, HARVARD DATAVERSE NETWORK 3–4 (May 27, 2014), http://dx.doi.org/10.7910/DVN/26147 (select "Data & Analysis" tab, then download "Person Course Documentation.pdf") (explaining the structure of the public dataset). User behavior is categorized as registered, viewed, explored, or certified. *Id.* The dataset includes interactions, such as the number of unique days a student accessed the course and the number of forum posts created. *Id.* More specific personal data, such as individual quiz grades, are excluded. *See id.* The public dataset contains far less information than is actually collected. *See* Jon Daries, *Person-Course De-Identification Process*, HARVARD DATAVERSE NETWORK 6 (May 27, 2014), http://dx.doi.org/10.7910/DVN/26147 [hereinafter Daries, *De-Identification*] (select "Data & Analysis" tab, then download "Person Course Deidentification.pdf").

6. *See* A. D. Ho et al., *HarvardX and MITx: The First Year of Open Online Courses* 5 (HarvardX and MITx, Working Paper No. 1, 2014), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263.

7. *See* Riley v. California, 573 U.S. ___, 134 S. Ct. 2473, 2489 (2014) ("Sixteen gigabytes translates to millions of pages of text, thousands of pictures, or hundreds of videos.").

8. Cookies are text files that your computer downloads from a website; the purpose of a cookie is, in most instances, to facilitate ease of interaction with frequently-visited websites, such as Amazon or Yelp. WebWise Team, *What Are Cookies?*, BBC (Oct. 10, 2012), http://www.bbc.co.uk/webwise/guides/about-cookies.

9. Canvas fingerprinting is essentially a new type of cookie that uses a unique image rather than a text file to track users, thus circumventing traditional privacy settings. *See* Julia Angwin, *Meet the Online Tracking Device that Is Virtually Impossible To Block*, PROPUBLICA (July 21, 2014), http://www.propublica.org/article/meet-the-online-tracking-device-that-is-virtually-impossible-to-block.

10. *See* Julia Angwin, *The Web's New Gold Mine: Your Secrets*, WALL ST. J. ONLINE, July 30, 2010, *available at* FACTIVA, Doc. No. WSJO000020100731e67u003h1.

11. *What Is a Geotag?*, INSTAGRAM, https://help.instagram.com/411058025616750?sr=22&query=geo-tag&sid=1klGrlFwBp2dRp1Mz (last visited May 8, 2015).

12. Data brokers are companies that collect, then resell or share consumer information. FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY i (2014), *available at* https://www.ftc.gov/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014 [hereinafter FTC].

13. *See id.* at 23 (stating that data brokers earn approximately $426 million per year in revenue from their marketing and risk mitigation services). One data broker, Acxiom, claims to have "[m]ulti-sourced insight into approximately 700 million consumers world-

marketing, which is not quite the stuff of Orwellian nightmares. But as processing power increases and data capabilities improve, the insights into and applications of online user data may evolve to have a more serious impact.[14]

Who regulates this activity? Increasingly, government entities are attempting privacy regulation. From the White House[15] to state attorneys general,[16] privacy legislation and enforcement actions are "in." However, some have questioned whether it is feasible to expect today's Congress to enact privacy legislation.[17] Existing federal privacy statutes are themselves something of a "patchwork,"[18] and only FERPA relates to education.[19] Finally, MOOCs do not easily fall within FERPA's ambit,[20] if at all. Where FERPA may apply to MOOCs, it does so in an ad hoc fashion.

---

wide" and "[o]ver 3,000 propensities for nearly every U.S. consumer." ACXIOM CORP., ACXIOM CORPORATION ANNUAL REPORT 8 (2013), *available at* http://www.acxiom.com/wp-content/uploads/2013/09/2013-Annual-Report.pdf.

14. Beyond the marketing context, current big data uses and discoveries include the identification of a severe side effect when patients used two popular drugs, more efficient allocation of energy using "smart grids," and predictive crime mapping. *See* Jules Polonetsky & Omer Tene, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 245–48 (2013); *see also* Andrew G. Ferguson, *Predictive Policing and Reasonable Suspicion*, 62 EMORY L.J. 259, 265 (2012) (describing predictive policing as a generic term to denote use of "computer models that predict areas of future crime locations from past crime statistics and other data").

15. *See, e.g.*, The White House Office of the Press Secretary, Fact Sheet: Safeguarding American Consumers & Families (Jan. 12, 2015), http://www.whitehouse.gov/the-press-office/2015/01/12/fact-sheet-safeguarding-american-consumers-families (describing President Obama's proposed legislation to limit the uses of educational data, among other privacy initiatives).

16. *See, e.g.*, N.Y. State Office of the Attorney Gen., A.G. Schneiderman Proposes Bill To Strengthen Data Security Laws, Protect Consumers from Growing Threat of Data Breaches (Jan. 15, 2015), http://www.ag.ny.gov/press-release/ag-schneiderman-proposes-bill-strengthen-data-security-laws-protect-consumers-growing.

17. *See* Hanni Fakhoury, *Why Wait for Congress? States Passing Electronic Privacy Legislation*, ELEC. FRONTIER FOUND. (June 3, 2013), https://www.eff.org/deeplinks/2013/05/why-wait-congress-states-passing-electronic-privacy-legislation.

18. Daniel J. Solove & Chris J. Hoofnagle, *A Model Regime of Privacy Protection*, 2006 U. ILL. L. REV. 357, 357, 401 (2006) ("[P]rivacy protections in the United States are riddled with gaps and weak spots. Although most industrialized nations have comprehensive data protection laws, the United States has maintained a sectoral approach [to privacy] where certain industries are covered and others are not."). Federal statutes that protect privacy include FERPA, the Health Insurance Portability and Accountability Act ("HIPAA"), the Fair Credit Reporting Act ("FCRA"), and the Privacy Act of 1974, which addresses government collection of data. *See Existing Federal Privacy Laws*, CTR. FOR DEMOCRACY & TECH. (Nov. 30, 2008), https://cdt.org/insight/existing-federal-privacy-laws.

19. The Children's Online Privacy Protection Act (COPPA) may also cover some educational data, but it is limited to children under the age of 13. 15 U.S.C. § 6501 (2012). Thus, COPPA is typically not relevant to MOOCs, which are directed at older audiences. *See, e.g.*, *MIT and Harvard Announce edX*, HARVARD GAZETTE (May 2, 2012), http://news.harvard.edu/gazette/story/2012/05/mit-and-harvard-announce-edx (discussing edX as a university initiative to enhance undergraduate education).

20. *See infra* Part V.A.

The challenges of applying existing privacy legislation to big data practices are rooted in antiquated conceptions of data records, varying definitions of what counts as identifying information, and a tendency to equate protecting privacy with achieving anonymity.[21] These issues are not limited to FERPA; rather, they reflect a twentieth-century codification of privacy incommensurate with the dramatic changes created by advances in computing.

Many recent incidents have highlighted privacy issues arising from big data, such as Edward Snowden's leak of National Security Agency ("NSA") cell phone metadata collection[22] and the outcry over data-based educational resources.[23] Such worrisome programs are rooted in the power of big data — massive collections of diverse information paired with the computational power to analyze and mine them for predictive insights and conclusions.[24] Big data has immensely exciting potential, offering increased business efficiency,[25] more rapid disease detection,[26] and other benefits. But there are correspondingly significant downsides: Conclusions gleaned from analyzing data en masse could reveal specific personal identities[27] or enable more concrete injuries, such as erroneous government classification of individuals as terrorists.[28]

---

21. *See infra* Part II; *see also* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1740 (2010) (noting that many statutes assume that anonymization protects privacy).

22. *See* Glenn Greenwald, *NSA Collecting Phone Records of Millions of Verizon Customers Daily*, GUARDIAN (June 6, 2013, 6:05 AM), http://www.theguardian.com/world/2013/jun/06/nsa-phone-records-verizon-court-order.

23. Olga Kharif, *Privacy Fears over Student Data Tracking Lead to InBloom's Shutdown*, BLOOMBERG BUSINESSWEEK (May 1, 2014), http://www.businessweek.com/articles/2014-05-01/inbloom-shuts-down-amid-privacy-fears-over-student-data-tracking.

24. *See* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE ix (2014), http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf [hereinafter PCAST, TECHNOLOGY]. Another common formulation of big data is Doug Laney's "3V" model for managing data through three attributes: volume, velocity, and variety. *See, e.g.*, Doug Laney, *3D Data Management: Controlling Data Volume, Velocity, and Variety*, META GROUP (Feb. 6, 2001), http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf; Svetlana Sicular, *Gartner's Big Data Definition Consists of Three Parts, Not To Be Confused with Three "V"s*, FORBES (Mar. 27, 2013, 8:00 AM), http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/ ("Big data is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.") (internal quotation marks omitted).

25. *See, e.g.*, James Manyika et al., *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, MCKINSEY & CO. 5 (May 2011), http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

26. *See* Polonetsky & Tene, *supra* note 14, at 246.

27. *See* Neil M. Richards & Jonathan H. King, *Three Paradoxes of Big Data*, 66 STAN. L. REV. ONLINE 41, 43 (2013).

28. *See, e.g.*, David Gray & Danielle Citron, *The Right to Quantitative Privacy*, 98 MINN. L. REV. 62, 67, 81 (2013) (describing how Maryland state police used nationwide data centers to erroneously classify individuals, including two nuns and a local political candidate,

A common thread among big data harms is the discovery or exploitation of personally identifiable information ("PII"). Broadly speaking, PII is data that could directly or indirectly identify an individual. Significantly, big data analytics can use predictive inferences to generate new PII from an anonymized dataset.[29]

For instance, Target's product-prediction model designated a shopper as pregnant based on her buying habits. Target subsequently mailed her coupons and advertisements for pregnancy-related items.[30] While Target's activity seemed innocuous, the shopper was a teenage girl whose outraged father called the store to complain about the advertisements without knowing his daughter was actually pregnant.[31] Notably, the data supporting Target's classification of "pregnant" was unrelated to the teenager's identity. Instead, Target based its conclusion on the teenager's purchases from a group of twenty-five products correlated with pregnant shoppers, such as unscented lotion and vitamin supplements.[32] Although such data standing alone might not reveal the shopper's identity, Target's big data prediction system derived a "creepy"[33] and likely unwelcome inference from her shopping patterns. Furthermore, the categorization constituted an additional data point about this shopper: By predicting that she was pregnant, Target narrowed the pool of possible patrons she could be. Generating PII without an individual's affirmative action in providing the data creates a privacy problem because it bypasses traditional direct information collection methods (which are, at least in some areas, regulated) and thus eviscerates a basic privacy protection: an individual's ability to consent.[34]

PII is any information that can be used to distinguish or trace an individual.[35] However, different statutes treat PII differently. Some, like the Health Insurance Portability and Accountability Act ("HIPAA")[36] provide for express PII elements,[37] thus clearly stating

---

as terrorists, and shared this classification with federal agencies without notifying the targets).

29. *See* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 98 (2014).

30. *See* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

31. *See id.*

32. *See id.*

33. *See* Omer Tene & Jules Polonetsky, *Introducing a Theory of Creepy*, RE/CODE (Apr. 18, 2014, 4:00 AM), http://recode.net/2014/04/18/introducing-a-theory-of-creepy/.

34. *See* Daniel Solove, *Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1880 (2013); *see also* Crawford & Schultz, *supra* note 29, at 98.

35. Erika McCallister et al., *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, NAT'L INST. OF STANDARDS & TECH. 2-1 (Apr. 2010), http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf.

36. Pub. L. No. 104-191, § 1, 110 Stat. 1936, 1936 (1996).

37. Privacy of Individually Identifiable Health Information, 45 C.F.R. § 164.514(b) (2014) (outlining conditions that must be met before information is considered not to be

what information could identify a person. FERPA, on the other hand, uses a more catch-all approach, designating some express PII elements but also information that "in combination" could identify the protected party.[38] The PII triggers for these statutes dictate the manner in which sensitive data should be protected: by removing or obscuring PII via de-identification methods.[39] The statute-specific PII triggers also inform how data may be used: either through exceptions or after the removal of any PII.[40]

However, privacy-protection methods based on using PII as a trigger are increasingly problematic for several reasons. First, PII, as statutorily defined, may be so vague as to make compliance with existing privacy legislation an impracticably difficult task.[41] Second, equating privacy protection to appropriate handling of PII implies that any data that is not PII does not create a risk of individual identification.[42] In other words, the current statutory approach tends to overlook indirect identifiers: data points that do not explicitly identify a person but can reveal one's identity when combined with other data.[43] In-

---

PII). One benefit of using this express PII approach is that it provides a safe harbor for any company that successfully removes these statutorily-mandated elements.

38. FERPA defines "PII" as including, but not limited to:

> (a) The student's name;
> (b) The name of the student's parent or other family members;
> (c) The address of the student or student's family;
> (d) A personal identifier, such as the student's social security number, student number, or biometric record;
> (e) Other indirect identifiers, such as the student's date of birth, place of birth, and mother's maiden name;
> (f) Other information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty; or
> (g) Information requested by a person who the educational agency or institution reasonably believes knows the identity of the student to whom the education record relates.

Family Educational Rights and Privacy, 34 C.F.R. § 99.3 (2014).

39. *Id.* at § 99.31(b) (setting out the de-identified information exception). Under this exception, "[a]n educational agency or institution, or a party that has received education records or information from education records under this part, may release the records or information without the consent required by [FERPA] after the removal of all personally identifiable information" provided that "the educational agency or institution or other party has made a reasonable determination that a student's identity is not personally identifiable, whether through single or multiple releases, and taking into account other reasonably available information." *Id.* at § 99.31(b)(1).

40. *See id.* at § 99.31 (setting forth exceptions to FERPA's consent requirement for disclosure of PII).

41. *See infra* Part V.A.

42. *See* Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 NYU L. REV. 1814, 1816 (2011).

43. *See* Latanya Sweeney, K-*Anonymity: A Model for Protecting Privacy*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYS., 557, 563 (2002) (defining quasi-identifiers as "attributes that in combination can uniquely identify individuals such as birth date and gender").

deed, indirect identifiers lie at the heart of the privacy problems raised by big data: Sufficient quantities of innocuous, non-PII information can yield detailed and targeted information about an individual because they ultimately function as quasi-identifiers.[44] The safe harbors in statutes relying on PII erroneously imply that once one removes PII, one has successfully protected the individual's privacy.[45] Finally, PII functions to ground privacy in protection of individual anonymity, rather than in protecting spaces for autonomous expression and development,[46] or any others of the myriad conceptions of privacy. Privacy statutes do so by specifying that an individual's privacy is protected when PII has been removed from her data — when the data has been "anonymized."[47] Rather than being dictated by privacy jurisprudence or theory, this de-identification process has been the dominant method for privacy protection over the last century because it functionally protected those rights. However, due to big data, evolving social norms, and the ease with which information can now be shared, this regime has transformed from functional to a merely illusory protection of individual privacy at the expense of data utility.[48]

Recent suggestions for protecting privacy focus on regulating the use of personal data rather than limiting its collection.[49] This shift reflects a changed conceptualization of the harms presented by big data, where the collection of data is not inherently harmful. Instead, the harm is created by the possible conclusions derived from big data. In these situations, the statutory privacy trigger would not be the col-

---

44. *See id.*; Arvind Narayanan & Edward W. Felten, No Silver Bullet: De-identification Still Doesn't Work (July 9, 2014) (unpublished manuscript), *available at* http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf (providing an overview of several studies showing how quasi-identifiers can identify or lead to the identification of an individual).

45. *See* Sweeney, *supra* note 43, at 558 (conducting an early re-identification effort showing the ease with which individuals could be identified through seemingly anonymized data); *see also infra* text accompanying notes 224–226.

46. *See* Julie E. Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904, 1906–07 (2013).

47. *See* Family Educational Rights and Privacy, 34 C.F.R. § 99.31(b) (2014) (outlining the de-identification exception to FERPA's consent requirements); *Protecting Student Privacy While Using Online Educational Services: Requirements and Best Practices*, PRIVACY TECHNICAL ASSISTANCE CENTER (Feb. 25, 2014), *available at* http://ptac.ed.gov/document/protecting-student-privacy-while-using-online-educational-services ("Metadata that have been stripped of all direct and indirect identifiers are not considered protected information under FERPA because they are not PII" — that is, this data does not require FERPA protection because it is sufficiently anonymous); *see also* Ohm, *supra* note 21, at 1740.

48. *See* Jon P. Daries et al., *Privacy, Anonymity, and Big Data in the Social Sciences*, ASSOC. FOR COMPUTING MACH. (Aug. 14, 2014), http://queue.acm.org/detail.cfm?id=2661641 (discussing the degree to which de-identified datasets differ from original datasets and the impact this may have on the analysis of that data); Justin Brickell & Vitaly Shmatikov, *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing, in* 2008 KNOWLEDGE DISCOVERY & DATA MINING CONF. 70, 70 (2008), *available at* http://www.cise.ufl.edu/~nemo/anonymity/papers/jlbrick_kdd2008.pdf (stating that "even modest privacy gains require almost complete destruction of the data-mining utility").

49. *See* PCAST, TECHNOLOGY, *supra* note 24, at xiii.

lection of PII, but rather, some determination based on a conclusion derived from big data correlations, such as the Target shopper "outed" as pregnant to her family.[50] In such cases, the initial existence of PII is irrelevant because it is not the failure to de-identify that causes harm, but rather the ability to correlate non-PII elements to draw a creepy (arguably privacy-violating) conclusion that the analyzing party may act on. A data use-oriented approach to regulating privacy reflects the understanding that what is far more important than the ability to identify a user is the real-world impact of decisions made about the user — like whether to increase or lower the user's credit limit.[51]

These problems are best illustrated by delving into how an existing privacy statute that relies on PII for its privacy protection handles big data. This Note explores these issues through an in-depth analysis of FERPA and how it relates to MOOC providers, revealing several important lessons. First, it emphasizes the limited applicability of federal privacy statutes by assessing whether FERPA even applies to MOOC providers. This limitation derives from a historical vision of PII as siloed within particularly sensitive areas like health records and education — a balkanization that big data eviscerates. Second, it underscores that a statute's PII and record definitions break down when confronted with a multiplicity of data sources that can be aggregated to generate inferences. Third, it argues that by-the-book PII protection measures can succeed in producing sufficiently anonymous datasets, but these measures may dramatically reduce the utility and accuracy of the collected data.[52] In other words, we cannot have both anonymity and optimally useful data.

Part II discusses big data and privacy, with a focus on the PII/de-identification privacy standard adopted by existing statutes in this area. It then examines the inadequacy of this method of privacy protection and suggests a use-oriented approach. Part III gives an overview of FERPA and the context of its adoption. Part IV discusses the advent of MOOCs. Part V assesses whether FERPA would apply to MOOCs, how this application would impact a MOOC provider's behavior, whether this regime is desirable, and how legislators could modify FERPA to better comport with shifting notions of privacy and desire for innovation. Finally, Part VI explores the takeaways from

50. *See* Duhigg, *supra* note 30; *see also* Gray & Citron, *supra* note 28, at 81 (discussing erroneous classification of Catholic nuns as terrorists by a Maryland police algorithm).

51. *See* Joseph W. Jerome, *Buying and Selling Privacy: Big Data's Different Burdens and Benefits*, 66 STAN. L. REV. ONLINE 47, 51 (Sept. 2013) (discussing how a man's credit limit was lowered from $10,800 to $3,800 because he had used his credit card at stores where other consumers had poor repayment track records).

52. *See* Daries et al., *supra* note 48; *see also* Jane Yakowitz, *The Tragedy of the Data Commons*, 25 HARV. J. L. & TECH. 1, 8 (2011) (noting that "the legal minimum anonymization requires some of the utility of a dataset to be lost through redaction and blurring in order to ensure that no subject has a unique combination of indirect identifiers").

this case study and provides several suggestions for improving privacy regulation in a data-driven age.

II. BIG DATA AND PRIVACY

Big data, a current darling of the technology world,[53] is increasingly scrutinized for its benefits and harms with respect to privacy. Edward Snowden's leak of the NSA's phone metadata collection activities considerably inflamed concerns over the extent of modern data collection methods.[54] In an effort to deal with the fallout, President Obama called for a "comprehensive review of big data and privacy" in order to consider:

> [H]ow the challenges inherent in big data are being confronted by both the public and private sectors; whether we can forge international norms on how to manage this data; and how we can continue to promote the free flow of information in ways that are consistent with both privacy and security.[55]

The resulting report recommended numerous changes to privacy regulations, including revisions to FERPA.[56]

Big data is a somewhat amorphous concept. It encapsulates two significant components: huge quantities of varied data and large-scale analytics.[57] Both are made possible by cheap data storage (resulting in increased data collection and retention) and increased computer processing power (enabling analysis of massive amounts of data).[58] This Note uses "big data" as a catchall phrase for this process: both large datasets and the analytical capability to draw inferences from them.

---

53. *See* Gary Marcus & Ernest Davis, *Eight (No, Nine!) Problems with Big Data*, N.Y. TIMES (Apr. 6, 2014), http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html Marketing research firm Gartner argued as of 2013 that big data was at the height of the "Hype Cycle" and would soon enter the "Trough of Disillusionment." Arik Hesseldahl, *Has Big Data Reached Its Moment of Disillusionment?*, ALLTHINGSD (Jan. 24, 2013, 4:36 PM PT), http://allthingsd.com/20130124/has-big-data-reached-its-moment-of-disillusionment.

54. *See* Glenn Greenwald et al., *Edward Snowden: The Whistleblower Behind the NSA Surveillance Revelations,* GUARDIAN (June 11, 2013, 9:00 AM), http://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance?CMP=twt_gu; *see also* Greenwald, *supra* note 22.

55. The White House Office of the Press Secretary, Remarks by the President on Review of Signals Intelligence (Jan. 17, 2014), http://www.whitehouse.gov/the-press-office/2014/01/17/remarks-president-review-signals-intelligence.

56. *See* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 64 (2014), http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf [hereinafter PCAST, VALUES].

57. *See* PCAST, TECHNOLOGY, *supra* note 24, at ix.

58. *See* Polonetsky & Tene, *supra* note 14, at 240.

The allure of big data lies in the insights and breakthroughs that can be gleaned from analysis of massive data troves, such as discovering drug side effects and quantifying crime waves.[59] These insights can also have substantial economic benefits. For example, the McKinsey consulting firm estimates that effective use of big data could provide $300 billion annually to the U.S. health care system.[60]

To achieve both the touted benefits and anticipated harms of big data, one first needs a significant quantity of data. The dataset may be narrowly focused (as with health research) or broad (as with consumer marketing.) In either situation, huge quantities of information on numerous subjects are collected or cross-linked from other databases for analysis. Data brokers have some of the most detailed repositories, using information from such sources as public records, consumer transactions, and public social profiles.[61] Beyond data brokers, any service an individual uses can collect the details of that interaction, for example, shopping on Amazon or watching movies on Netflix.

Notably, big data creates privacy concerns not just because it involves massive collections of individual information, but because those collections can be linked, aggregated, and then analyzed in unforeseen ways. Data may be collected from many sources, such as information volunteered on Facebook, cookies on websites, metadata from phone calls and emails,[62] and supermarket discount cards.[63] Data from these individual sources seem innocuous when standing alone, but when these data collections are aggregated they can reveal a surprisingly complete picture of a person and may even generate unexpected inferences, such as whether someone is pregnant[64] or has a gambling addiction.[65] These conclusions often have a predictive flavor, such as projections of an individual consumer's future actions given past behavior or current state, calling to mind Philip K. Dick's

59. *See id.* at 245–48. McKinsey identifies five ways in which big data creates value: creating transparency, enabling experimentation, customizing services for individual consumers, using automated algorithms to augment/replace human decision-making, and innovating new business models. Manyika et al., *supra* note 25, at 5.

60. *See* Manyika et al., *supra* note 25, at 2.

61. *See* FTC, *supra* note 12, at 11–14 (providing an overview of data broker sources).

62. *See* Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 402, 407 (2014). PCAST defines metadata as "ancillary data that describe properties of the data such as the time the data were created, the device on which they were created, or the destination of a message." PCAST, TECHNOLOGY, *supra* note 24, at xi.

63. COMM. ON COMMERCE, SCI., & TRANSP., A REVIEW OF THE DATA BROKER INDUSTRY: COLLECTION, USE, AND SALE OF CONSUMER DATA FOR MARKETING PURPOSES 2, 16 (2013), http://www.commerce.senate.gov/public/?a=Files.Serve&File_id=0d2b3642-6221-4888-a631-08f2f255b577 (discussing the types and sources of data collected by data brokers).

64. *See* Duhigg, *supra* note 30.

65. PCAST, TECHNOLOGY, *supra* note 24, at 12–13 (listing a number of insights big data may yield).

prescient mutants in *Minority Report*.[66] As more data is collected and linked, there are greater chances of deriving unexpected correlations or conclusions.[67]

The opportunities and innovations enabled by big data strike some as "creepy."[68] The basis for this creepiness lies in two simple points: (1) most people do not realize what or how much data they are providing, and (2) the regulatory solution to this transparency issue has been to anonymize the data instead of requiring that individuals be explicitly informed of its collection.[69] In this regime, privacy decisions have been made well before an individual confronts a creepy use. In other words, anonymity, which enables disclosure while ostensibly protecting identity, has become the brute-force method of resolving the fact that the individual did not truly consent to the ultimate data grab. Anonymity thus functions as an increasingly permanent placeholder for privacy.

Much of the discussion about protecting privacy seems to embrace, at its core, the feeling that someone else simply knows too much about us, even though the results of that knowledge may never materially impact our lives.[70] The situation is complicated by the notion that the data collection which enables big data is both unavoidable and beneficial. For instance, Great Britain's decision to open every National Health Service patient's anonymized record to researchers and pharmaceutical companies reflects a belief that such access will lead to significant breakthroughs in medicine.[71] Thus, the

66. In *Minority Report*, three mutants were able to predict with extremely high accuracy murders before they were committed. As a result, law enforcement began arresting people before the crimes actually occurred. PHILIP K. DICK, MINORITY REPORT (1956).

67. The President's Council of Advisors on Science and Technology (PCAST) has discussed several areas in which big data may yield significant insights and possible harms, including health care, education, and the more private sphere of the home. PCAST, TECHNOLOGY, *supra* note 24, at 13–15. Britain's National Health Service attempted to implement a "care.data" project that would link all NHS data about patients into one database. The goal of this project was twofold: to use for lifesaving research, and to use commercially to create "billions for the UK economy." Ben Goldacre, *The NHS Plan To Share Our Medical Data Can Save Lives — But Must Be Done Right*, GUARDIAN (Feb. 21, 2014, 1:30 PM), http://www.theguardian.com/society/2014/feb/21/nhs-plan-share-medical-data-save-lives.

68. *See* Tene & Polonetsky, *supra* note 33.

69. *See supra* notes 35–40 and accompanying text.

70. Most discussion of the privacy harms resulting from big data remains relatively abstract or trivial. Some incidents have been highly publicized, such as the pregnant Target shopper. *See* Duhigg, *supra* note 30. However, most big data impacts are far more subtle, ranging from ads seen on Amazon to Google's search query completion tool. *See* Janna Anderson & Lee Rainie, *Main Findings: Influence of Big Data in 2020*, PEW RESEARCH CENTER (Jul. 20, 2012), http://www.pewinternet.org/files/old-media//Files/Reports/2012/PIP_Future_of_Internet_2012_Big_Data.pdf.

71. *See* Sarah Knapton, *Health Records of Every NHS Patient To Be Shared in Vast Database*, TELEGRAPH (Jan 10, 2014), http://www.telegraph.co.uk/news/10565160/Health-records-of-every-NHS-patient-to-be-shared-in-vast-database.html; *The Care.data Programme*, NAT'L HEALTH SERV. ENGLAND, http://www.england.nhs.uk/ourwork/tsd/care-data/ (last visited May 8, 2015).

solutions proposed throughout this debate seem half-hearted and reflect the dilemma between creepiness and innovation that big data has generated. We want privacy (or maybe just anonymity), but we also want the benefits of big data.

Focusing on the way data is used rather than how it is collected may provide a more effective means of implementing what is arguably the current core of privacy: controlling access to information about oneself.[72] Existing statutes implement access control through a regulated third party tasked with protecting individual anonymity, either by simply not releasing information that would reveal individual identities or by anonymizing any disclosed data.[73] The beauty of such a regime is that it facilitates data sharing between individuals and the parties who need their data for services like health care, while also permitting publicly beneficial secondary uses, such as research.[74] To ensure that an individual is properly protected when her information is shared, the third party only needs to obtain her consent or de-identify the data, because the data recipient simply cannot do much with the data — it remains siloed.

However, big data creates a strikingly different environment. Today, we have both the storage capacity to collect even the most inconsequential pieces of information and the analytic capability to derive meaningful insights from these trivial datasets. Further, the rapid evolution of this technological space means that the existing consent regime is effectively illusory; even if users read terms of service explaining what personal data will be gathered, they arguably cannot knowingly consent because the future uses of their data are not fully defined at the time of collection.[75] Finally, even the possibility of rendering individuals' data anonymous through de-identification is increasingly under fire with some researchers arguing that "there is no evidence that de-identification works either in theory or in practice."[76]

Anonymity is useful, but only as one aspect of a privacy policy. It can be augmented by regulating what correlative inferences are permissible on the part of the data recipient. Thus, even if individuals

72. *See* PCAST, TECHNOLOGY, *supra* note 24, at 38.

73. *See, e.g.*, Family Educational Rights and Privacy, 34 C.F.R. § 99.31(b) (2014) (stating that under FERPA, PII may be released if it meets the de-identified records exception).

74. *See* Knapton, *supra* note 71.

75. *See* Solove, *supra* note 34, at 1881.

76. Narayanan & Felten, *supra* note 44. Ann Cavoukian and Arvind Narayanan recently debated the efficacy of de-identification. Cavoukian focused on debunking various re-identification studies by pointing out flaws in their methodologies and emphasizing that re-identification risks are inflated dramatically. *See* Ann Cavoukian & Daniel Castro, *Big Data and Innovation, Setting the Record Straight: De-identification* Does *Work*, PRIVACY BY DESIGN 1–2 (June 16, 2014), https://www.privacybydesign.ca/content/uploads/2014/06/pbd-de-identification_ITIF1.pdf. In response, Narayanan argued that the danger of trying to show that de-identification works effectively "promote[s] a false sense of security" by unrealistically constraining what an adversary who is motivated to re-identify would do. Narayanan & Felten, *supra* note 44.

cannot control the collection of their personal information, they can control how collecting entities access and use it. This is not a traditional vision of privacy; it concedes that going online necessarily involves transferring significant personal information. But it does shift the focus from relatively harmless and largely unavoidable acts of collection to the harms stemming from data use.

The following case study of FERPA and MOOCs highlights the challenges of privacy and anonymity presented by big data in the online education context. More generally, this study investigates the tensions that arise when a pre-big-data privacy statute meets big data. The analysis shows just how poorly traditional privacy statutes deal with metadata, the shifting nature of PII, and de-identification as a privacy-protecting solution. Further, this case study indicates that significant improvements to traditional privacy statutes cannot simply add new protections or update a few definitions. Meaningful reforms must instead modify core statutory conceptions of privacy, thus impacting what qualifies as a privacy-protection measure or a privacy violation.

III. FERPA

FERPA is one of the few federal privacy statutes currently in effect.[77] It protects student education records.[78] Like most privacy laws, FERPA constrains access to records, protects the PII within them, and regulates how information from those records may be used, maintained, and shared. The use of PII as the privacy protection trigger operationalizes privacy as anonymity: If PII is removed, you are anonymous, and thus your data can be disclosed without unreasonable risk to your privacy. FERPA can also be viewed as a use-regulating statute, since statutory exceptions that permit PII disclosure focus on data uses, such as research.[79] However, the PII trigger still tethers FERPA to a largely data-focused approach. Because current definitions of PII focus on what can be gleaned from the data at a static point — the time of disclosure — FERPA assumes that if disclosed information is de-identified, there will be no *ex post* privacy harm.

FERPA was a floor amendment proposed in 1974 by New York Senator James Buckley[80] to address student privacy violations at a

77. *See Existing Federal Privacy Laws*, CTR. FOR DEMOCRACY & TECH. (Nov. 30, 2008), https://cdt.org/insight/existing-federal-privacy-laws.

78. *See Legislative History of Major FERPA Provisions*, US DEP'T OF EDUC. 2 (June 2002), http://www2.ed.gov/policy/gen/guid/fpco/pdf/ferpaleghistory.pdf.

79. *See* Family Educational Rights and Privacy, 34 C.F.R. § 99.31(a) (2014) (listing the various exceptions to the consent requirement for PII-disclosure, including for research and student safety reasons).

80. 121 CONG. REC. S13,990 (daily ed. May 13, 1975) (statement of Sen. James Buckley). Because FERPA was proposed as an amendment to the General Education Provisions

time when government privacy intrusion was a hot-button issue.[81] Senator Buckley noted that the most pressing reasons for the amendment were "growing evidence of the abuse of student records across the nation"[82] and the need for individual privacy protection:

> More fundamentally, my initiation of this legislation rests on my belief that the protection of individual privacy is essential to the continued existence of a free society. There has been clear evidence of frequent, even systematic violations of the privacy of students and parents by the schools through the unauthorized collection of sensitive personal information and the unauthorized, inappropriate release of personal data to various individuals and organizations.[83]

These violations included psychiatric tests, research experiments, and surveys that sometimes resulted in the negative labeling of students as "potential delinquent[s]" or "bad citizen[s]."[84] The surveys' intimate inquiries, coupled with a lack of oversight and transparency with respect to the information collected, angered parent groups and provided the momentum necessary to pass the amendment.[85]

FERPA applies to educational institutions that maintain records and receive federal funding, including financial aid.[86] To protect student records, FERPA imposes two major requirements. First, educational institutions must make a student's record available upon the student's request and provide the student an opportunity to contest any errors.[87] Second, educational institutions may not have a policy or practice of disclosing a student's education records or the PII contained therein without the student's consent, unless the disclosure is

---

Act on the Senate floor, it lacks a traditional legislative history. *Legislative History*, *supra* note 78 at 1.

81. Much of the privacy legislation that was passed in the mid-1970s cited the Watergate scandal as a driving factor. For example, the Privacy Act of 1974 that established the Code of Fair Information Practice was proposed and passed in part because of the post-Watergate conclusion that government surveillance of citizens must be limited. *See* LEGISLATIVE HISTORY OF THE PRIVACY ACT OF 1974 S. 3418 (PUBLIC LAW 93-579): SOURCE BOOK ON PRIVACY 4 (1976), *available at* http://www.loc.gov/rr/frd/Military_Law/pdf/LH_privacy_act-1974.pdf.

82. 121 CONG. REC. at S13,990 (1975).

83. 121 CONG. REC. at S13,991 (1975).

84. *Id.* Senator Buckley noted that these labels had actually harmed students, citing an example of a New York high school student who was prohibited from attending her graduation because she was labeled a "bad citizen" and was unable to examine her record. *See id.*

85. *See* Margaret L. O'Donnell, *FERPA: Only a Piece of the Privacy Puzzle*, 29 J.C. & U.L. 679, 681–83 (2003) (noting that privacy issues motivated the amendment's proposal).

86. Family Educational and Privacy Rights, 20 U.S.C. § 1232g(a) (2012).

87. Family Educational Rights and Privacy, 34 C.F.R. §§ 99.7, 10 (2014).

permitted by a statutorily defined exception.[88] If an educational institution fails to meet these requirements, the Department of Education ("DOE") is authorized to withhold federal funds, "[i]ssue a complaint to compel compliance," or "[t]erminate eligibility to receive [federal] funding."[89] There is no private right of action under FERPA for students who believe their rights have been violated.[90]

While the consequences of violating FERPA can be dire, both the statute's definitions and the DOE's enforcement policies limit its scope. Educational institutions are only those that provide services to students,[91] who in turn are individuals who have attended an institution providing educational services and about whom the institution has maintained education records.[92] Education records are broadly defined as those "[d]irectly related to a student; and [m]aintained by an educational . . . institution."[93] Many of the cases involving FERPA have focused on what documents or files qualify as education records.[94] Further, FERPA's disclosure limitations only apply to PII — absent personally identifying content, data is not FERPA-protected (with the exception of some limitations on directory information),[95] and the educational institution may share it freely.[96] Finally, DOE enforcement of FERPA largely focuses on guidance documents and training;[97] it has not imposed the severe sanctions at its disposal.[98]

IV. MOOCs

Cloud-based education technologies, aggregators of student records, and other services offering big data analytics are increasingly controversial innovations in the education space. InBloom, for example, was a prominent education data analytics company funded with

88. 20 U.S.C. § 1232g(b)(1); *see infra* Part V.B.1 for a discussion of the exceptions relevant to MOOCs.

89. 34 C.F.R. §§ 99.67(a)(1)–(3).

90. *See* Gonzaga Univ. v. John Doe, 536 U.S. 273, 276 (2002).

91. 34 C.F.R. § 99.1(a)(1).

92. 34 C.F.R. § 99.3.

93. *Id.*; *see also* Lynn M. Daggett, *FERPA in the Twenty-First Century: Failure To Effectively Regulate Privacy for All Students*, 58 CATH. U. L. REV. 59, 62 (2008).

94. *See infra* notes 171–176 and accompanying text.

95. *See* 34 C.F.R. § 99.37 (discussing the conditions that apply to disclosing directory information), § 99.3 ("Directory Information").

96. *See id.* at § 99.31(b) (de-identified records exception).

97. *See About PTAC*, PRIVACY TECHNICAL ASSISTANCE CENTER, http://ptac.ed.gov/about (last visited May 8, 2015) (noting that Privacy Technical Assistance Center (PTAC) is a "one-stop resource for education stakeholders to learn about data privacy, confidentiality, and security practices related to . . . student data").

98. Daniel Solove has commented that FERPA sanctions are like "using a nuclear bomb to kill a cockroach" and "ha[ve] never been used in [FERPA's] 40-year history." Daniel Solove, *Big Data and Our Children's Future: On Reforming FERPA*, SAFEGOV (May 6, 2014), http://safegov.org/2014/5/6/big-data-and-our-children%E2%80%99s-future-on-reforming-ferpa.

$100 million and backed by the Bill and Melinda Gates Foundation.[99] It closed in 2014 after its district and state partners withdrew their support, citing privacy concerns.[100] Indeed, concerns with third-party companies providing data analytics to schools led the President's Council of Advisors on Science and Technology ("PCAST") to call for FERPA reform[101] and caused Massachusetts Senator Ed Markey to propose legislation regulating third-party use of student data.[102] Thus far, however, these concerns have focused on impacts to younger students in the K–12 space instead of more general online education platforms, such as MOOCs.[103] Precisely because they do not focus on K–12,[104] MOOCs defy categorization under FERPA and provide an ideal illustration of the challenges and flaws that arise when existing privacy legislation attempts to regulate a technology that does not clearly fit into the statute's domain.

In 2011, higher education was confronted with a new teaching platform that promised to disrupt the centuries-old classroom model: the massive open online course. MOOCs have since generated controversy,[105] been dismissed,[106] and been championed;[107] their potential

99. Olga Kharif, *Privacy Fears over Student Data Tracking Lead to InBloom's Shutdown*, BLOOMBERG BUSINESSWEEK (May 1, 2014), http://www.bloomberg.com/bw/articles/2014-05-01/inbloom-shuts-down-amid-privacy-fears-over-student-data-tracking.

100. *See id.*

101. PCAST, VALUES, *supra* note 56, at 64.

102. Press Release, Senator Ed Markey, Markey, Hatch Introduce Legislation to Protect Student Privacy (July 30, 2014), http://www.markey.senate.gov/news/press-releases/markey-hatch-introduce-legislation-to-protect-student-privacy; Protecting Student Privacy Act of 2014, S. 2690, 113th Cong. § 2 (2014).

103. *See, e.g.*, PCAST, VALUES, *supra* note 56, at 25–26 (focusing on privacy concerns for "children and teenagers" and "especially those [students] in K–12 education").

104. *See, e.g.*, *MIT and Harvard Announce edX*, *supra* note 19.

105. *See, e.g.*, Steve Kolowich, *Why Professors at San Jose State Won't Use a Harvard Professor's MOOC*, THE CHRONICLE OF HIGHER EDUCATION (May 2, 2013), http://chronicle.com/article/Why-Professors-at-San-Jose/138941/ (noting that San Jose State philosophy professors refused to teach a course developed by edX, a MOOC provider, because it seemed like an attempt to "replace professors . . . and . . . diminish[] education for students in public universities."); JEFFREY R. YOUNG, BEYOND THE MOOC HYPE: A GUIDE TO THE HIGH-TECH DISRUPTION OF COLLEGE ch. 5 (2013).

106. *See, e.g.*, James Grimmelmann, *The Merchants of MOOCs,* 44 SETON HALL L. REV. 4, 11 (2013) (arguing that MOOCs are not groundbreaking, and "are often mediocre and occasionally terrible"); *Beware of MOOCs*, ASQ HIGHER EDUC. BRIEF (Mar. 2013), http://rube.asq.org/edu/2013/03/innovation/beware-of-moocs.pdf (stating that MOOCs "seem like a total nightmare" and that "this kind of unidirectional lecturing style will be extremely ineffective in teaching students").

107. MOOCs have received interest and support from major educational institutions, Silicon Valley investors, the Bill and Melinda Gates Foundation, and Google. *See, e.g.*, YOUNG, *supra* note 105, ch. 1 (noting the flow of Silicon Valley venture capital money into for-profit MOOC providers Coursera and Udacity, and that Harvard and MIT made substantial investments — a reported $30 million each — into the non-profit edX); MOOC RESEARCH, http://www.moocresearch.com/ (last visited May 8, 2015) (the MOOC Research Hub, whose goal is to "evaluat[e] MOOCs and how they impact teaching, learning, and education in general," is funded by the Bill and Melinda Gates foundation); Will Oremus, *Google and edX Are Building a "YouTube for MOOCs,"* SLATE (Sept. 10, 2013),

impact on education has been the subject of intense speculation. Thus far, however, the regulatory issues posed by massive aggregation of student data and its possible uses have largely escaped discussion.

One reason MOOCs gained traction is because many believed they represented a revolutionary shift in education technology.[108] While initial optimism may have been overwrought — many MOOC courses see a significant decline in registrant involvement as the course progresses, with the majority of students failing to complete the class[109] — proponents still believe MOOCs are an integral part of a re-imagining of education.[110] According to edX's Anant Agarwal, "we won't solve [the education system] just by tweaking one aspect of it . . . . [W]e need to change everything on campus . . . . [W]e have to rethink all aspects of education from the ground up."[111] Agarwal views MOOCs as integral to this evolution.[112]

MOOCs are online courses open to any person who decides to register. The three largest MOOC providers are Coursera, Udacity, and edX.[113] Many courses are created by universities with the logistical and technological support of a MOOC provider.[114] Three characteristics define a MOOC: (1) short videos made for the Internet, not the classroom; (2) an automated grading system that does not require any professor involvement; and (3) discussion forums.[115] MOOCs

---

http://www.slate.com/blogs/future_tense/2013/09/10/mooc_org_google_edx_online_classes_partnership_is_youtube_for_moocs.html.

108. *See, e.g.*, Steven Leckart, *The Stanford Education Experiment Could Change Higher Learning Forever*, WIRED (Mar. 20, 2012), http://www.wired.com/2012/03/ff_aiclass/; Tamar Lewin, *Instruction for Masses Knocks Down Campus Walls*, N.Y. TIMES (Mar. 4, 2012), http://www.nytimes.com/2012/03/05/education/moocs-large-courses-open-to-all-topple-campus-walls.html.

109. *See* Ho et al., *supra* note 6, at 2. Ho et al. argue that "certification rates are misleading and counterproductive indicators of the impact and potential of open online courses." *Id.*; *see also* Keith Devlin, *MOOCs and the Myths of Dropout Rates and Certification*, HUFFINGTON POST (Mar. 2, 2013), http://www.huffingtonpost.com/dr-keith-devlin/moocs-and-the-myths-of-dr_b_2785808.html.

110. *See* Tamar Lewin, *After Setbacks, Online Courses Are Rethought*, N.Y. TIMES (Dec. 10, 2013), http://www.nytimes.com/2013/12/11/us/after-setbacks-online-courses-are-rethought.html; Helen Walters, *We Need To Change Everything on Campus*, TED BLOG (Jan. 27, 2014), http://blog.ted.com/2014/01/27/we-need-to-change-everything-on-campus-anant-agarwal-of-edx-on-moocs-mit-and-new-models-of-higher-education/.

111. Walters, *supra* note 110.

112. *Id.*

113. Robert McGuire, *The Best MOOC Provider: A Review of Coursera, Udacity and edX* (June 19, 2014), http://www.skilledup.com/articles/the-best-mooc-provider-a-review-of-coursera-udacity-and-edx/.

114. Laura Pappano, *The Year of the MOOC*, N.Y. TIMES (Nov. 2, 2012), http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html. These providers host the content, have developed their own message boards and testing systems, and in the case of Udacity, create all the videos in-house. *Id.*

115. YOUNG, *supra* note 105 ch. 1. The video component is not a class or lecture recording, but rather a video created specifically to convey a concept. *Id.* They are typically short, and often interrupted by brief quizzes. *Id.* The most common grading mechanism is a multiple choice quiz, but some MOOCs have simulated lab environments (such as edX's Circuits

also facilitate more sustained and structured student involvement than other online educational platforms: usually about a semester's worth of time with suggested time allocations of ten or twelve hours per week.[116] Coursera and edX offer university-created courses on a wide variety of topics from art to molecular biology,[117] whereas Udacity develops courses in-house focused on technology education.[118] Both Coursera and Udacity are venture capital backed, for-profit corporations,[119] while edX is a non-profit supported by funding from Harvard and MIT.[120]

Although MOOCs have existed for several years, their business model is still very much in flux.[121] Failure to effectively monetize online education precipitated the failure of early course providers like Columbia University's Fathom,[122] and many question whether MOOCs will suffer the same fate.[123] The big three — Coursera, Udacity, and edX — have several business model options, but they seem to be following a common saying in Silicon Valley: "[I]f you build a

---

and Electronics course), and others — typically classes like English — have used peer grading systems. *Id.*

116. For instance, Coursera's collection of courses range from four to twelve weeks in duration. *Courses*, COURSERA, https://www.coursera.org/courses (last visited May 8, 2015). In contrast, Codeacademy and Kahn Academy are two non-MOOC platforms that provide brief modules that allow the user to pick and choose what they learn, in what order, and at what pace. *See* CODECADEMY, http://www.codecademy.com/about (last visited May 8, 2015); KAHN ACADEMY, http://www.khanacademy.org/about (last visited May 8, 2015). There is no structured course program, no homework assignments (beyond suggestions to practice), and no clear outside-university involvement. *See id.*; Courtney Boyd Myers, *Codecademy: Learning To Code Just Became Fun, Easy and Slightly Addictive,* THE NEXT WEB (Oct. 14, 2011), http://thenextweb.com/apps/2011/10/14/code-academy-learning-to-code-just-became-fun-easy-and-slightly-addicting/.

117. *See* HARVARDX, http://harvardx.harvard.edu/who-we-are (last visited May 8, 2015) (emphasizing that the platform is a tool for developing online courses driven or created by faculty); YOUNG, *supra* note 105 ch. 1 (noting how then-Stanford professor Sebastian Thrun personally created his first MOOC); Coursera, http://www.coursera.org/courses (last visited May 8, 2015).

118. UDACITY, http://www.udacity.com/faq (last visited May 8, 2015) (describing how courses are developed "to provide the most relevant and cutting-edge tech education").

119. YOUNG, *supra* note 105 ch. 1.

120. Each institution put in $30 million. *Id.*

121. *See* Chrysanthos Dellarocas & Marshall Van Alstyne, *Money Models for MOOCs*, COMMC'NS OF THE ACM, Aug. 2013, at 25; *see also The Attack of the MOOCs*, THE ECONOMIST (July 20, 2013), http://www.economist.com/news/business/21582001-army-new-online-courses-scaring-wits-out-traditional-universities-can-they (noting that there is uncertainty over the most revenue-producing business model and "disagreement over how big the market will be"); *What Campus Leaders Need To Know About MOOCs*, EDUCAUSE 2 (Dec. 20, 2012), http://net.educause.edu/ir/library/pdf/PUB4005.pdf (listing possible business models).

122. *See* TAYLOR WALSH, UNLOCKING THE GATES: HOW AND WHY LEADING UNIVERSITIES ARE OPENING UP ACCESS TO THEIR COURSES 33–39 (2011) (Columbia injected $25 million into Fathom, an online learning project, over three years, but ultimately closed the venture in 2003 because of low revenues.).

123. *See* Andrew Delbanco, *MOOCs of Hazard*, NEW REPUBLIC (Mar. 31, 2013), http://www.newrepublic.com/article/112731/moocs-will-online-education-ruin-university-experience.

[website] that is changing the lives of millions of people, then the money will follow."[124] The models explored by these providers include low-cost course certification,[125] course enrollment fees,[126] licensing courses to universities,[127] and matching students to jobs.[128] These options fall under two umbrellas: (1) charge participants for complements (value-adding activities) or (2) charge a third party invested in the user group.[129] Groups such as education researchers, recruiters, and data brokers are interested in obtaining the user data collected by MOOC providers, regardless of whether it becomes a component of the MOOC business model.[130] The existence of this data implicates FERPA.

V. FERPA AND MOOCS

Like most websites, a MOOC records practically every move its users make. In the MOOC context, this information can include the content of forum comments and time spent watching a video.[131] HarvardX and MITx,[132] the two entities that initially launched courses on

124. YOUNG, *supra* note 105 ch. 1 (statement from Daphne Koller — a Coursera cofounder — regarding their venture capital funding).

125. Both edX and Coursera offer identification-verified certificates to prove you actually took the course for a small fee. *Verified Certificates of Achievement*, EDX, https://www.edx.org/verified-certificate (last visited May 8, 2015); COURSERA, https://www.coursera.org/signature/ (last visited May 8, 2015).

126. Udacity offers all courseware for free, but registrants can enroll in a course for a per-month fee which includes "projects, code-review and feedback, a personal Coach, and verified certificates." *Frequently Asked Questions*, UDACITY, https://www.udacity.com/faq#section-0-1 (follow "How much does it cost to enroll in a Udacity course?" hyperlink) (last visited May 8, 2015).

127. EdX experimented with offering a course at Massachusetts Bay Community College on "Practical Python Programming" which was taken by seventeen paying students in a "blended" online/in-person style. YOUNG, *supra* note 105 ch. 3. Coursera has also explored the blended model, and entered into agreements with ten state university system's to incorporate "MOOC technology and content to improve completion, quality, and access to higher education . . . ." *Ten U.S. State University Systems and Public Institutions Join Coursera to Explore MOOC-Based Learning and Collaboration on Campus*, COURSERA BLOG (May 29, 2013) http://blog.coursera.org/post/51696469860/10-us-state-university-systems-and-public-institutions.

128. Udacity has partnered with more than 350 companies for its job-matching program that connects promising students and tech employers. YOUNG, *supra* note 105 ch. 3. Coursera offers a similar job-matching service. *Coursera and Your Career*, COURSERA BLOG (Dec. 4, 2012), http://blog.coursera.org/post/37200369286/coursera-and-your-career.

129. Dellarocas & Alstyne, *supra* note 121, at 25.

130. *See* YOUNG, *supra* note 105 ch. 5 (noting concerns that student data will be sold); Ed Finkel, *Data Mining the MOOCs*, UNIVERSITY BUSINESS (Oct. 2013), http://www.universitybusiness.com/article/data-mining-moocs.

131. Ho et al., *supra* note 6, at 5. Other data points include standard registration information like school attended, level of education, location, sex, and birthdate, along with data more specific to MOOCs, like motivations for enrolling in the class, video interaction, test results, and length of time spent in different modules. *Id.*

132. HarvardX is an independent entity overseen by a Faculty Committee and a Research Committee. HARVARDX, http://harvardx.harvard.edu/who-we-are (last visited May 8, 2015).

edX, highlight the opportunities this quantity of data provides for "more rigorous research on learning for on-campus courses."[133] Both have released non-PII data visualization tools to foster greater understanding of MOOC-user demographics.[134] They hope to publicly release as much MOOC data as FERPA allows in order to facilitate research that will improve understanding of how students learn.[135] To this effect, they released their first de-identified dataset in the summer of 2014.[136]

The HarvardX/MITx goal was recently echoed by PCAST, which suggested that data from MOOCs and traditional classrooms will enable significant improvements to education and allow researchers to "discover what skills, taught to which individuals at which points in childhood, lead to better adult performance in certain tasks."[137] MOOC providers are potentially big data players: They collect data straight from the source (the user) and can analyze the data for insights through in-house researchers or outsourcing.

Sharing MOOC data with researchers appears to be a natural consequence of improved accessibility to student data; it also seems beneficial. After all, FERPA itself has an exception for disclosing PII to education researchers.[138] But what about sharing this data with textbook companies, consumer data collectors, or advertisers?[139] This is not a hypothetical: EdX entered into an agreement with textbook publisher Elsevier to share anonymized data in exchange for free textbooks.[140] A recent consolidated multi-district litigation against Google included allegations that the search giant mined student emails for its

---

MITx is an organization under the umbrella of MIT's Office of Digital Learning. MITX, http://odl.mit.edu/mitx/ (last visited May 8, 2015).

133. Ho et al., *supra* note 6, at 5.

134. *HarvardX Insights*, HARVARDX, http://harvardx.harvard.edu/harvardx-insights (last visited May 8, 2015); *MITx Insights*, MIT OFFICE OF DIGITAL LEARNING, http://odl.mit.edu/insights/ (last visited May 8, 2015).

135. *See Developing Big Data Analysis Tools*, BERKMAN CTR. FOR INTERNET & SOC'Y, https://cyber.law.harvard.edu/research/dpsi/usecases#Developing (last visited May 8, 2015); Elise Young, *Big Data Team — Navigating Regulation and Data Sets*, DIGITAL PROBLEM-SOLVING INITIATIVE PILOT @ HARVARD (Nov. 17, 2013), http://dpsipilot.tumblr.com/post/67313629440/big-data-team-navigating-regulation-and-data-sets; Ho et al., *supra* note 6, at 33.

136. *MITx and HarvardX Dataverse*, HARVARD DATAVERSE NETWORK, http://dx.doi.org/10.7910/DVN/26147 (last visited May 8, 2015) (providing hyperlink to HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0).

137. PCAST, TECHNOLOGY, *supra* note 24, at 14.

138. Family Educational Rights and Privacy, 34 C.F.R. § 99.31(a)(6) (2014).

139. Audrey Watters discusses data mining in the context of online education resources and implications of extending the metaphor of student data as "new oil." Audrey Watters, *Student Data Is the New Oil: MOOCs, Metaphor, and Money*, HACK EDUCATION (Oct. 17, 2013), http://hackeducation.com/2013/10/17/student-data-is-the-new-oil/.

140. *Elsevier To Provide Textbooks for Five New edX MOOCs*, ELSEVIER (Oct. 23, 2013), http://www.elsevier.com/about/press-releases/science-and-technology/elsevier-to-provide-textbooks-for-five-new-edx-moocs.

advertising tools without student consent.[141] While the DOE has issued some guidance on using online resources in the K–12 context,[142] it has remained silent on the issue of MOOCs in higher education.

The convergence of education and big data thus raises several difficult questions. First, does FERPA even apply to MOOCs? Second, if it does, what are MOOC providers' and associated institutions' obligations and permissible ranges of action under FERPA? And third, is there some alternative regime that would better implement FERPA's privacy-protecting purposes?

A. Does FERPA Even Apply to MOOCs?

The broader issue of applying an existing privacy statute to big data innovation can be examined more closely by assessing FERPA's applicability to MOOCs. While some big data entities such as risk management utilities are directly regulated by existing legislation,[143] the rest touch on regulated spaces in a relatively ad hoc manner. MOOCs present a case somewhere in between: They are not directly regulated under FERPA, but they are sufficiently connected to educational institutions to run the risk of occasionally falling within FERPA. In other words, there is a meaningful area of intersection between "clearly FERPA-regulated" and not. At the heart of this intersection is the data. When FERPA applies to MOOCs, it applies to the data that MOOCs host or create. This same logic can extend to other big data players whose services are used in some capacity by educational institutions.

FERPA's application to an entity is significant because of the affirmative protections it may require an organization to make. When combined with the uncertainty over what data is covered under FERPA, this treatment can result in over- or under-protection. If an entity over-protects its data to comport with FERPA, the result may be increased administrative costs and stifling of innovation. If the en-

---

141. *See In re* Google Inc. Gmail Litig., No. 13-MD-02430, 2014 WL 1102660, at *1–*3 (N.D. Cal. Mar. 18, 2014) (describing Google Apps services for educational institutions, including student email services, and plaintiffs' allegations that Google processed student email content and metadata to create secret user profiles). Google's contracts with the educational institutions included an agreement that the institution would "obtain the necessary authorization from end users" so that Google could provide its services. *Id.* at *2. However, the educational institutions used different methods to obtain authorization, such as by linking to Google's Privacy Policy and Terms of Service on the email sign-in page or to online FAQs discussing Google's automatic scanning of email content. *Id.* at *5–*6. The court denied a motion for class certification of the educational users class because of the "substantial individual questions regarding express consent" in these cases. *Id.* at *15.

142. *See generally Protecting Student Privacy*, *supra* note 47 (discussing best practices for school districts).

143. *See* FTC, *supra* note 12, at 53. While risk mitigation products are covered by the Fair Credit Reporting Act (FCRA) in certain situations, the FTC outlined several scenarios whereby FCRA may not apply, such as confirming an individual's identity. *Id.*

tity under-protects, however, it risks losing deals with educational institutions invested in FERPA compliance, and threatening student privacy.

Answering whether FERPA applies to MOOCs is no easy task. The DOE has withheld comment,[144] and most industry leaders have avoided or only superficially touched on the topic:[145] EdX has stated that users' education records are protected by FERPA "to the extent FERPA applies,"[146] whereas Coursera and Udacity do not address FERPA in their privacy policies.[147] Industry groups and universities also hold wide-ranging views on FERPA's application to MOOCs. The University of Illinois at Urbana-Champaign has stated that it "does not consider participants in [its] Coursera courses to be students . . . and thus FERPA regulations do not apply."[148] The University of Virginia's associate general counsel has noted that university instructors may blur the lines between participants and students if they require student use of MOOC content.[149] The National Association of College and University Attorneys has suggested that university legal counsel ensure faculty understand the risk of invoking FERPA when incorporating MOOC modules in their own courses.[150]

FERPA could apply to MOOCs in two ways, each with significant consequences. First, a MOOC provider could be classified as an educational institution and thus subject to all the requirements of a school, such as providing access to records upon student request and limiting disclosure of PII.[151] Second, even if MOOC providers are not themselves regulated by FERPA, the data they compile from registrants taking MOOCs may be protected and thus subject to the statute's disclosure constraints.[152] In this category, MOOCs would be

144. Although the DOE recently issued guidance on student privacy and online educational services, its focus has been on the K–12 space with no discussion of MOOCs. *Protecting Student Privacy*, *supra* note 47.

145. An EDUCAUSE webinar on legal issues raised by MOOCs noted that there was no authoritative holding on whether FERPA applied to MOOCs, but noted edX's acknowledgement that FERPA might apply. Madelyn F. Wessel, *EDUCAUSE Live! Legal Issues in MOOCs*, EDUCAUSE (Sept. 26, 2013), https://net.educause.edu/ir/library/pdf/LIVE1319.pdf.

146. *EdX Privacy Policy*, EDX, https://www.edx.org/edx-privacy-policy (last visited May 8, 2015).

147. *Privacy Policy*, COURSERA, https://www.coursera.org/about/privacy (last visited May 8, 2015); *Privacy Policy*, UDACITY, https://www.udacity.com/legal/privacy (last visited May 8, 2015).

148. *FAQ for Faculty*, MOOCS @ ILLINOIS (Feb. 7, 2013), http://mooc.illinois.edu/resources/faqfaculty/.

149. *Legal Issues in MOOCs*, EDUCAUSE 3 (Nov. 2013), https://net.educause.edu/ir/library/pdf/LIVE1319S.pdf.

150. Megan W. Pierson et al., *Massive Open Online Courses (MOOCS): Intellectual Property and Related Issues*, HIGHER EDUC. COMPLIANCE ALL. 18–19 (June 2013), http://www.higheredcompliance.org/resources/publications/AC2013_5G_MOOCsPartI1.pdf.

151. *See supra* notes 87–88 and accompanying text.

152. *See infra* Part V.A.1.

handling FERPA-protected data and thereby required to abide by FERPA's terms for that particular data. This situation is complicated by contractual relationships with educational institutions and limitations on the DOE's ability to sanction the behavior of errant MOOC providers.[153]

Whether FERPA applies to a particular MOOC provider depends on a number of factors such as the entity that owns or controls the data and how the MOOC registrant is classified. The reason for this complexity lies in the arrangements between universities and MOOC providers and the triggering definitions for FERPA application: (1) educational institution receiving funds administered by the DOE and (2) students for whom records are maintained.[154] At first glance, the funding limitation would seem to rule out many MOOC providers that operate with private backing. However, FERPA specifies that funds are made available if they are provided through a subcontract or to students attending the institution and using the funds for educational purposes.[155]

One could also argue that MOOC registrants are not students, since a student must be "in attendance" at the educational institution,[156] and it seems a stretch to say that MOOC registrants attend a university simply because that university created the course. However, the legislative history suggests that FERPA's definition of "student" includes anyone who audits a class and for whom records are maintained.[157] Furthermore, although the DOE declined to change the definition of "educational agency or institution" to "include entities that do not necessarily have students in attendance but still receive DOE funding under a program administered by the [DOE]," the DOE's authority extends to these "other recipients of [DOE] funds," including nonprofit organizations and student loan lenders.[158] This regime could implicate MOOC providers if they receive federal fund-

153. The only applicable enforcement ability with respect to a MOOC provider who receives no federal funding is the "five-year rule" which prohibits an educational institution from providing access to PII to contractors who improperly disclose PII for a period "of at least five years." Family Educational Rights and Privacy, 34 C.F.R. § 99.67 (2014); *see infra* text accompanying notes 164–167.

154. *See* Family Educational Rights and Privacy, 34 C.F.R. §§ 99.1, 99.3.

155. *Id.* at § 99.1(c).

156. *Id.* at § 99.7(a)(1).

157. 120 CONG. REC. 39,863 (1974) ("a student who is only auditing a course, but on whom the institution maintains a personal file, would be included in [FERPA]'s coverage"). However, the Fifth Circuit has held that an auditor who was not fully admitted as a student may not be entitled to all of the same rights as a "student" under FERPA. *See* Tarka v. Franklin, 891 F.2d 102, 103 (5th Cir. 1989).

158. Family Educational Rights and Privacy, 76 Fed. Reg. 75,604, 75,631 (Dec. 2, 2011) (referring to amendments to FERPA's enforcement procedures at 34 C.F.R. §§ 99.61–99.67). In its current incarnation, FERPA refers to the responsibilities of "an educational agency or institution, a recipient of DOE funds, or *a third party outside of an educational agency or institution*." 34 C.F.R. § 99.61 (emphasis added); *see also, e.g.*, §§ 99.65(a), 99.66(a).

ing through universities using their services, if they are considered contractors of the educational institution, or if students receive course credit or are otherwise considered to be using funds for an educational purpose in their use of MOOCs. Since the "educational institution" definition requires students in attendance and receipt of federal funds,[159] however, most MOOC providers do not appear to fit the baseline definition of educational institutions.

While MOOCs, in their current form, may have avoided the educational institution label, they may still be regulated by FERPA. This wrinkle is created by the relationship between MOOCs and universities. There are several threads animating FERPA's application given this relationship. Universities are paradigmatic "educational institutions"; FERPA thus protects student records (data) maintained by the university. The easiest example of FERPA application would involve universities that use MOOC platforms internally — for example, hosting interactive modules or lectures for students enrolled in a "real-life" course at the university.[160] FERPA allows an educational institution to disclose PII to a third party only in select situations. The contractor must "perform[] an institutional service or function for which the agency or institution would otherwise use employees" and must be "under the direct control of the agency or institution with respect to the use and maintenance of education records."[161] A MOOC in this situation would meet the first requirement; as long as a university retained control over its students' data, it would be allowed to disclose PII to the MOOC. Once it did, the MOOC would become subject to FERPA through its relationship with the university.[162] This situation would cover the use of any PII shared from the school with the MOOC and any records created during the course of the relation-

159. While some traditional institutions may be more likely to receive federal funding through Free Application for Federal Student Aid ("FAFSA") loans granted to its students, MOOCs have explored other business models wherein very few MOOC registrants actually pay for the course (i.e., for a certificate) and thus any federal funding received by a MOOC provider through FAFSA grants given to a student would be negligible. *See supra* notes 121–130 and accompanying text.

160. While arguably a clear case, this situation is still complicated by the fact that many online educational services used in university classrooms involve a student's independent registration on the MOOC platform. This implicates the consent exception to disclosure of PII: if the student voluntarily agrees to whatever terms of service the service provider requires, then one could argue that FERPA would not apply to any hypothetical, consented-to data release. *See, e.g.*, *edX Privacy Policy, supra* note 146 (describing user's "consent to the collection, use, disclosure, and retention by edX of [user's] Personal Information").

161. Family Educational Rights and Privacy, 34 C.F.R. § 99.31(a) (2014). PTAC has noted that schools may share information with online educational service providers that may implicate FERPA. *Protecting Student Privacy*, *supra* note 47, at 2. This would be the case if, for example, the school provides students' names and contact information obtained from education records in order to create student accounts with the service provider. *Id.*

162. 34 C.F.R. § 99.31(a)(3).

ship.[163] At the other end of the spectrum would be a course where the MOOC provider owns the data collected, the course is not being used by any educational institution for credit or educational purposes, and the MOOC provider receives no federal funding. Here, FERPA should not be implicated because the MOOC provider is not an educational institution and has no third-party relationship with an educational institution.

Should FERPA be implicated, the "five-year rule" describes the applicable enforcement ability with respect to a MOOC provider that receives no federal funding.[164] Under this rule, "[i]f the Office finds that a third party, outside the educational agency or institution, violates [the PII disclosure provision], then the educational . . . institution from which the personally identifiable information originated may not allow the third party . . . access to [PII] . . . for at least five years."[165] The DOE has clarified that "third party" is broad and refers to "any entity outside of the educational agency or institution from which the PII from education records was originally disclosed . . . ."[166] The rule also applies to improper redisclosure of PII, as well as the failure to destroy PII from records when the data is used for research.[167] Thus, the five-year rule would apply to MOOC providers who received FERPA-protected PII from an educational institution, such as through a contract to provide MOOC modules to a university in which the MOOC provider also collected data. The rule would be triggered if the MOOC provider then improperly redisclosed that PII to, for example, another MOOC provider or a textbook publisher. As a result, the educational institution that originally shared the data could not share any such information with the MOOC provider for five years. This would limit the ability of that institution and that provider to continue contracting for MOOC modules that required any disclosure of FERPA-protected data.

1. Education Records

Complications arise when a school owns the data or when a professor requires student use of a MOOC course or module for a class. In each of these cases, the data shared with the provider or generated through a MOOC may be protected by FERPA. Possessing or creating FERPA-protected data would require compliance with FERPA's dis-

163. PTAC recommends that schools include data access provisions in their agreements with service providers, because the provider may create new education records that would be subject to various FERPA requirements. *Protecting Student Privacy*, *supra* note 47, at 8–9.

164. *See* 34 C.F.R. § 99.67.

165. *Id.*

166. Family Educational Rights and Privacy, 76 Fed. Reg. 75,604, 75,633 (Dec. 2, 2011).

167. *Id.* at 75,634.

closure provisions to avoid DOE sanctions. This situation raises questions over the impact of data ownership and maintenance, as well as what actually constitutes an education record.

Under a digitized regime, there may be questions about who "maintains" an education record. The Supreme Court has stated that the term "maintain" suggests files kept in a "filing cabinet" or "permanent secure database."[168] Arguably, data kept on a remote secure server with limited access is analogous to traditional paper records kept in a filing cabinet, creating confusion over who maintains the data when it is technically stored by third-party cloud storage providers. However, the DOE's Chief Privacy Officer, Kathleen Styles, has noted that "the [storage] provider never 'owns' the data," and thus outsourcing data to cloud storage should not create problems under FERPA.[169]

While ownership of data may be relatively clear when the data is created by the educational institution, some data generated outside the institution may still be considered an education record. Conversely, much of the data generated over the course of a MOOC may not be considered an education record at all and thus would not be FERPA-protected. It is vital to understand what qualifies as an education record in order to assess when a MOOC or institution would be required to follow FERPA.

There has also been much debate over the definition of an "education record,"[170] and this definition under FERPA has been "the subject of significant litigation."[171] The Supreme Court has held that peer-graded classroom assignments are not education records, in part because they are not kept by a central custodian, but rather, handled by students within a class.[172] Meanwhile, the Court of Appeals for the Sixth Circuit has stated that disciplinary records do qualify as education records because "Congress made no content-based judgments with regard to its 'education records' definition."[173] While the definition is not entirely clear, the Supreme Court has used language that conveys an idea of permanence and direct connection to the student.[174] Therefore, some lower courts have held that MAC address-

168. Owasso Indep. Sch. Dist. v. Falvo, 534 U.S. 426, 433 (2002).

169. Daniel Solove, *Interview with Kathleen Styles, Chief Privacy Officer, U.S. Department of Education,* SAFEGOV (Apr. 18, 2013), http://www.safegov.org/2013/4/18/interview-with-kathleen-styles,-chief-privacy-officer,-us-department-of-education.

170. *See* Louis N. Schulze, Jr., *Balancing Law Student Privacy Interests and Progressive Pedagogy: Dispelling the Myth that FERPA Prohibits Cutting-Edge Academic Support Methodologies*, 19 WIDENER L.J. 215, 224 (2009).

171. *Id.*

172. *Owasso*, 534 U.S. at 435–36.

173. United States v. Miami Univ., 294 F.3d 797, 812–15 (6th Cir. 2002).

174. *See Owasso*, 534 U.S. at 434–35 (noting that FERPA's language "suggests Congress contemplated that education records would be kept in one place with a single record of access . . . . FERPA implies that education records are institutional records kept by a single

es[175] and emails not saved to the student's permanent file[176] are not part of the student's education records.

The DOE's interpretation seems to conflict with these rulings. Recently, the DOE's Privacy Technical Assistance Center ("PTAC") issued guidance for use of online educational services, suggesting that some metadata elements could be FERPA-protected if they are not de-identified.[177] Metadata is ancillary information that provides "meaning and content to other data being collected," such as time spent on a task, desktop or mobile access, keystroke information, and location.[178] It is, perhaps, as far as one can get conceptually from a physical file kept for a student. Indeed, metadata's physical analog is more like a library record that contains information about a book and its circulation history: data about data. But PTAC was quite clear that metadata that is linked to FERPA-protected information must be de-identified before a provider can use it for other purposes.[179]

Practically speaking, PTAC's approach reflects the realities of outsourcing educational services to online providers: Metadata is unavoidable when managing information. However, metadata does not easily fit within the Court's conception of "education record" as a permanent, centrally held piece of information. A MOOC test score might be part of an education record[180] — but the data about how long the student took to complete the test, where the student was located while taking the test, and so on, would likely not be included in the record.[181] These pieces of information, however, are arguably the more direct identifiers. PTAC seems to be operating from this understanding, with a focus on steps an educational service provider would need to take to avoid data disclosures that could link back to FERPA-protected data.[182] In other words, the DOE is emphasizing the privacy

central custodian, such as a registrar, not individual assignments handled by many student graders").

175. *See, e.g.*, Arista Records LLC v. Does 1–4, 589 F. Supp. 2d 151, 153 (D. Conn. 2008) ("A student's MAC address is not part of his or her educational records, and so its disclosure is not restricted by FERPA."); Fonovisa, Inc. v. Does 1–9, No. 07-1515, 2008 WL 919701, at *8 (W.D. Pa. Apr. 3, 2008) ("The MAC address does not appear to fall within the purview of FERPA").

176. S.A. ex rel. L.A. v. Tulare Cnty. Office of Educ., No. 08-1215, 2009 WL 3126322, at *5, *7 (E.D. Cal. Sept. 24, 2009) (holding that emails would only constitute education records if they personally identified the student and were maintained — kept in the student's permanent file — by the educational institution).

177. *See Protecting Student Privacy*, *supra* note 47, at 3.

178. *Id.* at 2–3.

179. *Id.* at 3.

180. *See Owasso*, 534 U.S. at 436.

181. *See id.* at 435 (noting that FERPA did not imply that "individual assignments handled by many student graders" would be covered under the statute, and similarly, that the "elaborate procedural machinery" that afforded parents a right to a hearing to contest the child's record did not apply to "challenge the accuracy of the grade on every spelling test and art project the child completes").

182. *See Protecting Student Privacy*, *supra* note 47, at 2–3.

purposes of FERPA over the plain language of its definitions. While a MOOC provider could contest this interpretation in court — and may even prevail — the safer bet is to de-identify any metadata connected to FERPA-protected data.

Although some metadata may be FERPA-protected, several types of data fall outside of FERPA's application, such as data not qualifying as an education record and data about registrants who would not be considered students under FERPA. FERPA does not protect materials that course creators use for their own purposes, like internal notes or email chains about a course.[183] These materials would likely include communications between professors and teaching assistants about conducting the class. Similarly, FERPA would not protect peer-reviewed and graded assignments that are not actually "recorded" as a grade.[184] However, if — as is the case in some MOOCs — the grade given by the peer reviewer is the final grade, it may be protected.[185]

2. Students

Significantly, information is only an education record if it concerns a student. FERPA defines a student as someone who attends an educational institution and about whom the institution maintains education record.[186] While this definition is somewhat circular, it seems that MOOC registrants would not be students for FERPA purposes; they do not pay tuition or receive recognized credit for their participation, except in the case of paid-for certificates. Indeed, as previously noted, the University of Illinois at Urbana-Champaign concluded that FERPA did not apply to its MOOCs because it did not consider the participants to be students of the institution.[187] However, this perspective ignores legislative history suggesting that some student auditors "would be included in [FERPA]'s coverage"[188] and overlooks court rulings that have acknowledged that auditors may receive some rights under FERPA.[189] One could certainly analogize MOOC participants to auditors: They arguably get similar benefits to auditors, who sit in on courses at a university without receiving course credit. Yet qualifying MOOC participants as auditors only leads to more questions. How does one treat people who register for a MOOC course but never watch the videos, participate in discussions, or work on assignments?

183. Family Educational Rights and Privacy, 34 C.F.R. § 99.3 (2014) ("Education records").

184. *Id.* ("The term [education records] does not include . . . [g]rades on peer-graded papers before they are collected and recorded by a teacher.").

185. *Owasso*, 534 U.S. at 436 (declining to decide "whether the grades on individual student assignments, once they are turned in to teachers, are protected by [FERPA]").

186. 34 C.F.R. § 99.3 ("Student").

187. *FAQ for Faculty*, *supra* note 148.

188. 120 CONG. REC. 39,863 (1974).

189. *See, e.g.*, Tarka v. Franklin, 891 F.2d 102, 107 (5th Cir. 1989).

What about those participants who only watch a few videos? Is there an "auditor" threshold for MOOCs where FERPA might apply? Above all, why is it important to differentiate between participants in this fashion?

B. Impacts of FERPA Application to MOOCs

If FERPA applies to MOOCs, it is more likely to apply to the data, not the MOOC provider itself. Thus, data ownership becomes an important component of how FERPA relates to MOOCs. If data is owned by an actual educational institution, then use of that data must follow a fairly standard pattern: The institution can share the data with student consent or share the data absent consent through exceptions or de-identification.[190] However, these FERPA rules appear to apply only to educational institutions;[191] as discussed above, a MOOC provider is unlikely to fall within this category. Therefore, if FERPA applies to MOOC providers, it will likely be through the constraints levied on the MOOC providers through their relationships with educational institutions — both data shared with the MOOC provider by the educational institution and possibly also data generated by the MOOC provider during the course of a MOOC if the data is deemed FERPA-protected. In either situation, the MOOC provider would not be allowed to re-disclose the FERPA-protected data unless the provider has stripped the data of its PII.[192] These constraints are significant as MOOC providers explore different monetization models that may include use of data generated from MOOC registrants. This Part will briefly consider the exceptions to disclosing PII under FERPA with a focus on de-identifying protected data.

1. FERPA PII Disclosure Exceptions

Educational institutions may disclose PII in two situations: (1) with express consent of the student and (2) without consent if the institution uses one of several statutorily defined exceptions.[193] One exception includes de-identification of data, removing the need for FERPA protections. If the data does not have PII or does not contain FERPA-protected PII elements, then FERPA does not apply at all and

190. *See* 34 C.F.R. §§ 99.30–31.

191. Per the terms of FERPA and its regulations, any exceptions can be pursued by an "educational institution." *Id.*

192. PTAC recently noted that an online educational service provider could use metadata that was linked to PII so long as it has been properly de-identified. *Protecting Student Privacy*, *supra* note 47, at 2–3. De-identified data is no longer FERPA-protected, so the statutory constraints no longer apply. *See id.*

193. 34 C.F.R. §§ 99.30–31.

the data may be shared with the public.[194] Significantly, an institution may share some PII without de-identification under other exceptions such as for research purposes, thus providing substantially more data to the receiving party but placing more limitations on the scope of data or how the data is used.[195]

PII is information that would allow a "reasonable person in the school community . . . to identify the student with reasonable certainty."[196] This information may include the student's name, social security number, and — most relevant to MOOC purposes — other information that "in combination[] is linked or linkable to a specific student."[197] If a researcher wants access to PII and she is not considered a school official, she must use the more constrained research exception. Many researchers will likely want access to PII because it would enable more detailed analysis and assessment of student performance.[198]

a. Consent

An institution may disclose any records with the consent of the student (or parent, if the student is still in K–12).[199] This consent requirement is fairly exacting: The consent must be in writing, specify the records disclosed, state the purpose of the disclosure, and identify the party to whom the disclosure may be made.[200] Because of these limitations, FERPA consent is an ineffective tool for many uses of student data. For example, if an educational institution wants to share PII with MOOC providers to better tailor course offerings, FERPA would require the institution to acquire written consent for this specific purpose from each of its students.

b. Directory Information Exception

An institution may also make directory information public. This information could include the student's "name, address, telephone listing, date and place of birth, major field of study," and school-related information such as activities, honors and awards, and dates of attendance.[201] In order to disclose directory information, the institution must notify the student (or parent) of the categories of information designated as "directory information" and allow the student a

194. *Id.* at § 99.31(b).
195. *Id.* at § 99.31(a).
196. *Id.* at § 99.3 ("Personally Identifiable Information").
197. *Id.*
198. *See* Yakowitz, *supra* note 52, at 8–10.
199. Family Educational and Privacy Rights, 20 U.S.C. § 1232g(b)(1) (2012).
200. 34 C.F.R. § 99.30(b).
201. 20 U.S.C. § 1232g(a)(5)(A).

reasonable period of time to opt out of the disclosure.[202] Since directory information is already publicly disclosed, there should be no FERPA barrier to disclosure by a MOOC provider. However, since re-identification risks increase as more data is released, it would be a better policy to also remove directory information from any data disclosures.

c. Research Exception

The research exception is useful in promoting innovations in education, but it is likely only relevant for the MOOC space if the data generated in MOOCs may eventually be shared with researchers through the educational institution, and possibly by the MOOC provider directly if contractual agreements allow.

FERPA allows an institution to disclose PII to "organizations conducting studies for, or on behalf of, educational agencies or institutions to . . . improve instruction."[203] This exception is not a "general" research exception, but rather, limited to the purposes of the study outlined by the researcher.[204] Thus, the initial study parameters influence the extent to which the data may be released and how the disclosed data may be used. However, an educational researcher should qualify for this exception if she meets the FERPA requirements. The primary limiting factor to this exception is the "written agreement" requirement which governs restrictions on the research organization's use of PII as per its agreement with the educational institution.[205]

The study must be "for, or on behalf of" the institution for the development of tests, administration of aid programs, or improvement of instruction.[206] If the study benefits the institution, it would be "on behalf of" that institution.[207] In addition to these requirements, researchers must also ensure that third parties cannot identify the students covered by the study and researchers must destroy the information once it is no longer needed.[208]

d. School Official Exception

The school official exception is particularly relevant for MOOCs, because this exception is likely the safe harbor through which an educational institution can share PII with a MOOC provider. This excep-

202. *Id.* at § 1232g(a)(5)(B).
203. 34 C.F.R. § 99.31(a)(6).
204. *Id.*
205. *Id.* at § 99.31(a)(6)(iii).
206. *Id.* at § 99.31(a)(6)(i).
207. Family Educational Rights and Privacy, 76 Fed. Reg. 75,604, 75,627 (Dec. 2, 2011).
208. 34 C.F.R. § 99.31(a)(6)(iii).

tion may also provide insight as to how MOOCs can be conceptualized with respect to FERPA. The exception allows the institution to disclose PII to "a contractor, consultant, volunteer, or other party to whom an . . . institution has outsourced institutional services or functions."[209] The third party must fulfill a function for which the institution would normally use its own employees; moreover, the third party's use of education records must be subject to the institution's direct control,[210] and the third party must follow FERPA's re-disclosure requirements which specify that the third party must obtain a student's consent before re-disclosure is allowed.[211] The DOE has suggested that this exception is the most likely to apply to a school's use of online educational services.[212]

If an institution uses the school official exception to share PII with MOOC providers, there is a significant limitation: The MOOC provider "cannot use FERPA-protected information for any other purpose than the purpose for which it was disclosed."[213] Herein lies the utility of de-identification. If the PII is de-identified, the MOOC provider can use it for any purpose (subject to any non-FERPA limitations in its agreement with the educational institution).[214]

e. Data De-identification Exception

De-identified information may be shared with anyone, for any purpose, because once data has been de-identified, it is no longer considered FERPA-protected data.[215] According to the DOE, data is de-identified when the educational institution has removed all PII and the institution has made the "reasonable determination that the student's identity is not personally identifiable . . . taking into account other reasonably available information."[216] While some elements in education records are statutorily PII and must be removed,[217] others are PII only if, in combination, they create such a unique footprint that identification would be possible with reasonable certainty. These elements are known as "quasi-identifiers."[218]

The standard for de-identification is that a reasonable member of the institution's community would not be able to identify the student

209. *Id.* at § 99.31(a)(1).
210. *Id.* Direct control may be fulfilled through a contract with the third-party provider. *Protecting Student Privacy*, *supra* note 47, at 4.
211. 34 C.F.R § 99.33(a)(1).
212. *Protecting Student Privacy*, *supra* note 47, at 4.
213. *Id.* at 5.
214. *Id.* at 3, 5.
215. *Id.* at 3.
216. 34 C.F.R § 99.31(b).
217. *Id.* at § 99.3 ("Personally Identifiable Information").
218. Sweeney, *supra* note 43, at 563.

with "reasonable certainty."[219] According to PTAC, "de-identification is considered successful when there is no reasonable basis to believe that the remaining information in the records can be used to identify an individual."[220] In assessing the possibility of identification, the institution must take into account the risk of re-identifying an individual from previously released data and other "reasonably available information."[221] De-identification does not mean perfect anonymity, but rather, the minimization of unintentional disclosure of a student's identity.[222]

Methods for de-identification remain rather ad hoc across institutions, although the DOE has referred institutions generally to the Federal Committee on Statistical Methodology's *Statistical Policy Working Paper 22*.[223] Use of standard statistical methods should be sufficient to de-identify data for FERPA purposes as discussed below.

While the DOE currently finds removal of direct and indirect identifiers sufficient for de-identifying data, this view may change in the near future. True anonymization may be difficult or impossible to achieve; studies abound highlighting the ease by which re-identification is possible.[224] For instance, Harvard Professor Latanya Sweeney identified the Massachusetts governor's health records from a "de-identified" public data release she linked with public voting

219. For example, if the community was aware through media publicity that a student brought a gun to school, then any indication of this event in the student's record would qualify as PII because a community member could identify the student with reasonable certainty from that piece of information. Family Educational Rights and Privacy, 73 Fed. Reg. 74,806, 74,832 (Dec. 9, 2008).

220. *Data De-identification: An Overview of Basic Terms*, PRIVACY TECHNICAL ASSISTANCE CENTER 3 (May 2013), http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf.

221. *Id.*

222. *See id.* at 2–3 (noting that "it may not be possible to remove the disclosure risk completely").

223. Family Educational Rights and Privacy, 73 Fed. Reg. at 74,835. Although its primary goal is to assist federal agencies with confidentiality requirements, the working paper provides an overview of statistical procedures that can be used by any organization to minimize the risk of disclosing information about individuals when collecting and disseminating statistical data. *See Report on Statistical Disclosure Limitation Methodology* 1–2 (Fed. Comm. on Statistical Methodology, Statistical Policy Working Paper 22, 2005), *available at* https://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf.

224. Latanya Sweeney conducted several studies over a decade ago showing the inadequacies of stripping identifiers for de-identification; she found that 87% of the United States population "had reported characteristics that likely made them unique based only on {5-digit ZIP, gender, date of birth}." Sweeney, *supra* note 43, at 558. Arvind Narayanan and Vitaly Shmatikov were able to re-identify a number of Netflix users by linking IMDb reviews to the Netflix Prize dataset. *See* Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, *in* 2008 IEEE SYMPOSIUM ON SEC. & PRIVACY 111, 123 (2008). They found that "very little auxiliary information [from other sources such as IMDb] is needed [to] de-anonymize an average subscriber record from the Netflix Prize dataset." *Id.* at 124.

records.[225] Generally speaking, standard de-identification techniques are increasingly ineffective because of big data. In this case, big data represents the increased quantity of publicly available data (e.g., voter records, census data, and information collected by Facebook and Google) and the computing power that makes it possible to link these datasets and thus uncover identity, or "re-identify."[226] While there is some debate over the efficacy of de-identification, the consensus appears to be that for high-dimensional datasets (essentially, "big" big data sets), de-identification is a weak solution for protecting privacy.[227] PCAST adopts this position in its reports on big data, noting that anonymization techniques are "increasingly easily defeated" and that while somewhat useful, "approaches that deem [anonymization], by itself, a sufficient safeguard need updating."[228] This view may herald a shift in the enforcement of privacy statutes in the coming years, especially when combined with PCAST's recommendation to modernize FERPA to better ensure that student data is used appropriately[229] and shifting notions of what privacy means and where anonymity fits within that concept.[230]

As long as the DOE maintains the presumption that de-identification is possible, then the degree of anonymity required for FERPA compliance falls somewhere in the middle of an anonymity continuum: somewhat de-identified, but not truly anonymous.[231] Practically speaking, de-identification is probably sufficient to deter an average person who may want to discover a person's identity, and the kind of data currently generated by MOOCs is not so sensitive as to raise alarm. However, databases that lead to re-identification "can be built almost entirely with nonsensitive data," and thus even innocuous information can be used to build a profile that may have significant and harmful consequences to the re-identified individual.[232] Notwithstanding these concerns, FERPA as it stands permits de-identification as a method for protecting privacy.[233] Therefore, MOOC providers

225. Sweeney, *supra* note 43, at 558–59. The governor was one of only six people with his birthday. Three of those six were men, and he was the only one in his zip code. *Id.* at 559.

226. *See* Polonetsky & Tene, *supra* note 14, at 240.

227. *See supra* note 76 and accompanying text.

228. PCAST, TECHNOLOGY, *supra* note 24, at xi.

229. PCAST, VALUES, *supra* note 56, at 64.

230. Cohen, *supra* note 46, at 1906; Ohm, *supra* note 21, at 1740; Richards et al., *supra* note 62, at 407; Solove, *supra* note 34, at 1880. Arvind Narayanan, one of the researchers involved in the Netflix Prize dataset study, has proposed that datasets should not be release to third parties at all, and rather that analysts should go to the database. Kim Zetter, *Computer Scientist: Arvind Narayanan*, WIRED (June 18, 2012), http://www.wired.com/2012/06/wmw-arvind-narayanan/.

231. *Protecting Student Privacy*, *supra* note 47, at 3.

232. Ohm, *supra* note 21, at 1768.

233. Family Educational Rights and Privacy, 34 C.F.R § 99.31(b) (2014).

who possess FERPA-protected data may release it so long as the data has been de-identified.

2. Method for De-identification

MOOC providers may wish to use FERPA-protected data for many reasons — some financially motivated, others more explicitly aligned with improving education. Financially-beneficial releases could involve deals with a textbook publisher,[234] contracts with marketing or advertising agencies, or sales to a data broker.[235] To improve education, the provider may want to release the data to the public for research purposes or use it to improve their own MOOC offerings.[236] Regardless, the more data the provider releases, the more valuable the database becomes.[237] As a natural corollary, more complete releases increase the likelihood of privacy right violations and substantive harms. FERPA's de-identification regime is a rather brute-force approach to this tension: It removes most of the information to avoid the privacy harms, but in doing so it nullifies much of the data's value and potentially beneficial impact.[238] Since de-identification is still the rule, understanding it in broad terms is valuable.

The goal of any FERPA-sufficient method should be compliant with the de-identification standards released by the DOE: de-identification such that a reasonable person in the community could not identify the person behind the data.[239] Various government organizations including PTAC,[240] the National Center for Education Statistics ("NCES"),[241] and the National Research Council[242] have issued their own reports or guidelines suggesting methods for protecting stu-

234. *Elsevier to Provide Textbooks for Five New edX MOOCs*, ELSEVIER (Oct. 23, 2013), http://www.elsevier.com/about/press-releases/science-and-technology/elsevier-to-provide-textbooks-for-five-new-edx-moocs.

235. The FTC noted that in addition to using publicly available data sources, data brokers also purchase data from commercial sources. *See* FTC, *supra* note 12, at 13.

236. *See* Ho et al., *supra* note 6, at 4 (discussing HarvardX and MITx's shared goals of researching and improving digital education).

237. This reasoning is the logic behind data brokers' data-accumulation efforts: the more data you have about a person, the more complete your picture, and the better your product. *See* FTC, *supra* note 12, at iv (noting that one data broker "has information on 1.4 billion consumer transactions and over 700 billion aggregated data elements").

238. *See* Yakowitz, *supra* note 52, at 8.

239. *See* 34 C.F.R. § 99.3 ("Personally Identifiable Information").

240. *See generally Data De-identification*, *supra* note 220.

241. *See, e.g.*, Marilyn Seastrom, *Basic Concepts and Definitions for Privacy and Confidentiality in Student Education Records*, NAT'L CTR. FOR EDUC. STATISTICS (Nov. 2010), http://nces.ed.gov/pubs2011/2011601.pdf; Marilyn Seastrom, *Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting*, NAT'L CTR. FOR EDUC. STATISTICS (Dec. 2010), http://nces.ed.gov/pubs2011/2011603.pdf [hereinafter Seastrom, *Statistical Methods*].

242. *See generally* MARGARET HILTON, PROTECTING STUDENT RECORDS AND FACILITATING EDUCATION RESEARCH (2008), http://www.nap.edu/openbook.php?record_id=12514; McCallister et al., *supra* note 35.

dent data. Broadly speaking, a FERPA-sufficient de-identification method involves: (1) associating an individual student's data with a random code; (2) removing all statutorily required[243] PII elements; and (3) applying statistical techniques to obscure outliers and other instances where threshold removal was insufficient for de-identification.[244]

The proposed method, while FERPA-sufficient, may also be useful in other fields for de-identification purposes, with the caveat that true anonymity is impossible for any data release that is not a high-level aggregation of data.[245] While re-identification may be possible after a de-identification method is applied, it is important to remember that re-identification still requires an actor who (1) wants to identify the individuals behind the data and (2) has the skill to do so.[246]

As an initial step, the student's data should be associated with a code instead of the student's registered name.[247] A code serves several purposes: (1) it is an initial anonymizing step of the individual's data; (2) it allows for re-identification internally; and (3) it provides qualifying parties with the ability to obtain more data on the particular record.[248] Because MOOC content differs in the degree of public access, two codes should be used: one for internal data, such as student responses and grades, and another for external data, such as forum posts and "upvotes."[249] Coursera has adopted this method because forums are publicly viewable to all course registrants.[250] When the code that identifies the student is combined with the often unique content of forum posts, the code could be used to re-identify an individual within a coded dataset.[251]

---

243. *See* 34 C.F.R. § 99.3 ("Personally Identifiable Information").

244. *See generally, e.g.*, *Data De-identification*, *supra* note 220; HILTON, *supra* note 242; McCallister et al., *supra* note 35.

245. By high-level aggregation, I mean publishing broad conclusions or facts about a dataset. *See, e.g.*, *World Map of Enrollment*, HARVARDX, http://harvardx.harvard.edu/harvardx-insights/world-map-enrollment (last visited May 8, 2015).

246. *See* Narayanan & Felten, *supra* note 44.

247. For instance, the code could be a hexadecimal hash. McCallister et al., *supra* note 35, at 4-5.

248. For example, a permanent and unique code could be assigned to a student's record for the student's time in a school district so that a researcher could track her test performance over multiple years. 34 C.F.R. § 99.31(b)(2); *see also Data De-identification*, *supra* note 220, at 3.

249. MITx and HarvardX have adopted number of forum comments as a quasi-identifier that must be removed or obscured in their de-identification method. *See* Daries, *De-Identification*, *supra* note 5, at 9.

250. Coursera, *Data Export Procedures* 4–5 (June 10, 2013) (on file with author).

251. *Id.* The Coursera procedure discusses a scenario in which an actor, who has access to a data export on the course, searches the Coursera forums for posts by a particular student. Once the actor finds a post by the student with information contained in the data export (such as full-text forum comments), the actor can simply search the data export using this information and arrive at the student's entire record. *See id.* at 21.

In addition, the code should be unique on a per-student and per-course basis; the same code should not be used for the same registrant across different classes. If the code were tied to the registrant and used for each course she took, this would create a quasi-identifier problem.[252] For example, few people are likely to simultaneously take Introduction to Biology, CopyrightX, and a course on Greek mythology. If the code were static across courses, it would be relatively easy to re-identify that person because she took a unique combination of classes.

Next, statutorily defined PII elements should be removed. These elements are explicitly set forth by FERPA and include information such as social security number, birthdate, and mother's maiden name.[253] Removing these items is slightly complicated by the disclosure exception for directory information. Institutions may disclose directory information elements, such as an address, provided that they give notice and an opportunity to opt out.[254] For entities considering a data release, however, directory information can increase the chance that non-directory PII will be inadvertently disclosed.[255] Therefore, a MOOC provider should treat directory information as PII that must be removed or obscured before data is released.

The FERPA regulations also require removal or obfuscation of elements that "in combination" could identify the student[256] — in other words, quasi-identifiers.[257] Quasi-identifiers can be manually selected or they could be identified through a re-identification process. MITx and HarvardX have taken the manual approach.[258] They selected course ID number, number of forum posts, year of birth, country, and gender because these elements are either public or commonly-used descriptors and thus more likely to lead to inadvertent disclosure.[259] Quasi-identifiers may be removed or obscured through various statistical techniques. This may include adding noise by using differential privacy methods[260] or by perturbing the values slightly.[261]

---

252. *See id.* at 4–5.

253. Family Educational Rights and Privacy, 34 C.F.R. § 99.3 (2014) ("Personally Identifiable Information").

254. *See* 34 C.F.R. § 99.37.

255. Family Educational Rights and Privacy, 73 Fed. Reg. 74,806, 74,834–35 (Dec. 9, 2008).

256. 34 C.F.R. § 99.3 ("Personally Identifiable Information").

257. *See supra* notes 43–44 and accompanying text.

258. *See* Daries, *De-Identification*, *supra* note 5, at 3.

259. *Id.*

260. Differential privacy is a technique used for query-limited databases and is an increasingly favored privacy-protection method. Jane Bambauer et al., *Fool's Gold: An Illustrated Critique of Differential Privacy*, 16 VAND. J. ENT. & TECH. L. 701, 704–05 (2014). Differential privacy involves introducing a certain amount of noise into a query result such that an individual cannot be identified from the result of that query. Cynthia Dwork, *Differential Privacy*, MICROSOFT RESEARCH 8–10 (2006), http://research.microsoft.com/pubs/64346/dwork.pdf.

Significantly, many states have adopted minimum group size reporting rules; where the number of respondents is below a certain number — most states use ten — that field is simply not reported.[262] This comports with Sweeney's *k*-anonymity protocols, which mandate that any field below a certain *k* value should go unreported.[263]

Additional steps can be taken to mitigate disclosure risks, although with each additional step the utility of the dataset is reduced. One could simply not report enrollment data used to compute the percentage distribution across achievement levels, such that no "number of students" is even disclosed.[264] Similarly, achievement levels could be collapsed (i.e., "pass" or "fail" instead of grade breakdowns) or not reported.[265] Additionally, data at extremes can be consolidated to lower or higher than a certain percentage.[266]

After the dataset has been cleansed of possible identifying information and any particularly unique fields (those below the adopted *k* value) have been removed, the dataset could be released to the public. A more detailed and technical example of this procedure and accompanying dataset has been released by HarvardX and MITx as part of their commitment to providing data for education researchers.[267]

C. FERPA's Purpose and Flaws

MOOCs will likely not fall within FERPA simply because they are not educational institutions. However, the data they generate may sometimes be FERPA-protected. The difficulty in assessing when this situation may arise, and how to appropriately protect or disclose the data, emphasizes two main issues with FERPA specifically and existing privacy legislation more broadly. First, FERPA's definitions reflect non-digital presumptions and are thus ill-suited to confront the reality of today's records and data collection. Second, and more significantly, the issues presented by a MOOC's big data capabilities question FERPA's fundamental assumption: Privacy can be protected with anonymity.

The problem with FERPA's definitions can be summed up as an inability to adequately answer who is regulated, what is protected, and when protection is triggered. The core of our education system still falls squarely within the auspices of FERPA, but this core increasingly uses third-party applications to provide sophisticated and valuable

261. See HILTON, *supra* note 242, at 19.
262. *See* Seastrom, *Statistical Methods*, *supra* note 241, at 1.
263. *See* Sweeney, *supra* note 43, at 557.
264. *See* Seastrom, *Statistical Methods*, *supra* note 241, at 14.
265. *Id.*
266. *Id.*
267. The dataset was de-identified in accordance with a procedure outlined in a document accompanying the aggregated data. *See* Daries, *De-identification*, *supra* note 5, at 6–11.

educational tools. Outsourcing itself is allowed and regulated under FERPA's school official exception,[268] but it raises the next difficult question: What exactly should be protected by this third party? In other words, what is an education record? The answer is arguably even murkier than the "who." Indeed, what information constitutes an education record has the potential to snowball — and perhaps already has — precisely because so much information is now maintained about a student. Furthermore, the multiplicity of entities generating and collecting information on students makes determining which bytes of data are protected an even greater challenge. Finally, connected to these challenges is the difficulty of determining when protection attaches. This last point reflects the fact that the PII designation happens at a static point in time: upon contemplation of a data release. However, de-identified and released information may eventually *become* PII again or enable the generation of PII. We are simply unable to predict which information has the potential to become PII with accuracy.

Big data's impact on how records are maintained, what can function as an identifier, and the many benefits of its analytic insights have made these definition problems more pressing. It has also created the tension between privacy and unlocking the potential of vast amounts of data, calling into question not just FERPA's assumptions regarding privacy, but its very purpose. As discussed above, the PII trigger presumes that anonymized data protects a person's privacy by concealing her identity. When combined with access restrictions, these statutes are reasonably effective if judged by the technology contemporaneous to their passage: pre-digital environments when data could not be aggregated en masse and could, quite literally, be locked up. Access protection is still fairly straightforward to implement: educate holders of records (although this group has increased in size given uncertainty over what constitutes records) and develop data breach prevention measures. The bigger challenge nowadays is not preventing access, but rather, deciding what needs protection, when, and how.

VI. CONCLUSION

If the discussion of FERPA's applications to MOOCs has suggested nothing else, it is that fitting big data into existing privacy legislation is similar to fitting a round peg into a square hole. Instead of trying to place third-party educational providers in the FERPA framework, we should instead re-assess FERPA from the ground up: What do we want a privacy statute to accomplish in this space? What should be private?

268. Family Educational Rights and Privacy, 34 C.F.R. § 99.31(a)(1) (2014).

While various sources have suggested updates and amendments to FERPA,[269] they have left the core goals of the legislation largely undiscussed. Re-assessing the purpose of FERPA is useful in that it engages more meaningfully with what the statute should balance and how it should weigh innovation and privacy. A more holistic process creates the space necessary to develop an effective educational privacy statute. Senator Markey's FERPA amendment is an excellent example of the flaws in an approach that accepts FERPA's current framework and purpose.[270] The Protect Student Privacy Act is limited to controlling how schools outsource educational tools and therefore records, and it effectively codifies existing DOE regulations.[271] Its oversights do not include modifications to FERPA's "nuclear" enforcement provisions or FERPA's definitions, such as what constitutes an education record or PII. As a result, the Protect Student Privacy Act would simply bolster what the DOE has already told schools[272] while failing to provide the DOE with an improved mechanism for implementing the statute. The proposed amendment also has other downsides, including a new restraint on recordkeeping that requires schools to delete data on students who no longer attend the institution.[273] Rhetorically, this sounds like a win for privacy advocates, but the practical impact is to nullify alumni status. Perhaps the biggest harm of a flawed amendment is that it creates the illusion of greater privacy protections. This allows for continued avoidance of the real debate: what privacy means, and how we should protect it.

If one steps back and looks at FERPA's passage,[274] text, and genesis, it becomes increasingly clear that FERPA was a reactionary piece of legislation that seems to have done no more than create administrative headaches and provide an occasional shield to protect against Freedom of Information Act requests.[275] Perhaps no section better demonstrates the inefficacy of FERPA than its enforcement provisions, which authorize the DOE to remove federal funding and

269. *See, e.g.*, PCAST, VALUES, *supra* note 56, at 64; Senator Markey Press Release, *supra* note 102.

270. *See* Senator Markey Press Release, *supra* note 102.

271. For example, the amendment requires that third parties with access to education records with PII must meet or surpass those protections put in place by educational institutions. Protecting Student Privacy Act of 2014, S. 2690, 113th Cong. § 2(6) (2014).

272. For example, the DOE's 2008 regulations require that a school control a consultant or contractor's (e.g., an educational service provider) "maintenance, use, and redisclosure of education records" in order to share them in the first place. 73 Fed. Reg. 74,806, 74,816 (Dec. 9, 2008).

273. *See* S. 2690 § 2(7)(b).

274. *See supra* notes 80–85 and accompanying text.

275. Chicago Tribune v. Univ. of Ill., 680 F.3d 1001, 1002 (7th Cir. 2012) (noting that plaintiff requested information about student admissions under the Illinois Freedom of Information Act, which the school opposed, citing FERPA).

effectively nothing else.[276] This aspect of FERPA, coupled with the lack of a private cause of action, goes a long way toward explaining why so many school districts fail to properly follow DOE guidelines on student record and privacy protection.[277]

So why have FERPA, given today's education environment? Let us continue with the presumption that there is significant value in privacy as space in which one can innovate and make choices without undue influence.[278] With this perspective, we might reconceptualize FERPA's purpose as ensuring that students are not erroneously described through their records and, similarly, that their records do not define them for the world at large. As a result, FERPA functions as a means for protecting a student's ability to self-create and identify — at least in areas distanced from their education.

If we are to keep FERPA, legislators must change several key aspects. First, the DOE needs a more reasonable "stick" with which to punish violators. Alternatively, Congress could provide a private right of action, thus opening an additional enforcement mechanism. Second, Congress should update FERPA's definitions to reflect the now dominant nature of technology for recordkeeping and education generally. Third, Congress ought to reconsider FERPA's underlying purposes and other key definitions, such as PII, in light of current privacy debates.

From a self-regulation perspective, institutions could take several steps to protect privacy. These measures fall loosely under a framework of transparency and control: transparency in what is collected and how data is used and control over data. One stakeholder is implementing a similar framework in the data broker space: Acxiom's "About the Data" initiative,[279] which the FTC has lauded as an important step towards improving privacy protections.[280] The challenge with this approach still lies in how providers monetize online services

276. Family Educational Rights and Privacy, 34 C.F.R. § 99.67 (2014) (noting that the DOE has authority to "terminate eligibility to receive funding under any applicable program").

277. *See* Reidenberg et al., *Privacy and Cloud Computing in Public Schools*, CTR. ON LAW & INFO. POLICY (2013) (noting that "only 25% of districts inform parents of their use of cloud services, 20% of districts fail to have policies governing the use of online services" and that most of these services fail to address "parental notice, consent, or access to student information").

278. Cohen, *supra* note 46, at 1906–07.

279. Natasha Singer, *Acxiom Lets Consumers See Data It Collects*¸ N.Y. TIMES (Sept. 4, 2013), http://www.nytimes.com/2013/09/05/technology/acxiom-lets-consumers-see-data-it-collects.html; *see also About the Data*, ACXIOM, https://aboutthedata.com/portal/registration/step1 (last visited May 8, 2015).

280. Julie Brill, Comm'r, Fed. Trade Comm'n, A Call to Arms: The Role of Technologists in Protecting Privacy in the Age of Big Data, Sloan Cyber Security Lecture at Polytechnic Institute of NYU (Oct. 23, 2013), *available at* https://www.ftc.gov/sites/default/files/documents/public_statements/call-arms-role-technologists-protecting-privacy-age-big-data/131023nyupolysloanlecture.pdf.

(typically through data sales), and for MOOCs, this transparency/choice regime may not be a viable economic strategy given the current experimentation in business models and the remote likelihood of DOE involvement in the MOOC space.[281]

This framework is, however, appropriate where traditional educational institutions use MOOCs for educational purposes, and it would also apply to K–12 third-party providers. An effective way of implementing this approach could be to show users the data about themselves, in a quantified-self kind of format. Educational services already exist with this method in mind, and it could be a less "creepy" revelation when one sees the potential benefits of data collection through the ability to track one's learning progress. Such an approach would still inform the user of the data at the service provider's fingertips. The challenging part of the transparency step is informing the user precisely how the data will be used. This step is difficult simply because it is nearly impossible to predict how the provider might want to use the data and how third parties might use the data downstream (an argument against notice-consent regimes). Unfortunately, there do not seem to be satisfactory answers to this problem beyond recognizing that if providers share the data, the users will not be "anonymous." One nascent solution is to create a service where the user has complete control over data sharing[282] — but this may be difficult to implement in the education context. Ultimately, the best approaches may simply be for educational institutions or users to pay more to keep the data in-house, or for the industry to monitor and curb inappropriate downstream uses.

Control, unfortunately, is the more challenging aspect of this suggestion. It requires the provider to give users the meaningful (i.e., actual) ability to delete data about themselves or otherwise limit secondary and tertiary uses. Several services like this already exist, most of which have small user bases because, unsurprisingly, they are fee-based to make up for lost advertising revenue. Here, the degree and quality of transparency could be helpful: Users are likely to be more comfortable with trading their information as companies develop more useful or interesting data outputs and conclusions from the users' information. Therefore, users would be less likely to need the control element. While this proposal carries a somewhat paternalistic specter — certainly it does not fit the Brandeisian notion of priva-

281. Steve Kolowich, *Are MOOC-Takers 'Students'? Not When It Comes to the Feds Protecting Their Data*, THE CHRONICLE OF HIGHER EDUCATION (Dec. 3, 2014), http://chronicle.com/article/Are-MOOC-Takers-Students-/150325/ (noting that Kathleen Styles, the DOE's chief privacy officer, stated that "data in the higher-education context for MOOCs is seldom FERPA-protected" in a recent symposium on student privacy).

282. LaVonne Reimer, The Lumenous Project: How Open Data Systems Allow the Commercial Credit Market to Deliver Greater Transparency and Accountability at Cloud Speed and Scale 8–9 (Feb. 2, 2015) (unpublished manuscript) (on file with author).

cy — it does reflect other tradeoffs that society has made: public safety for reduced privacy, urban living for less solitude, and so on.

It is largely control that is lacking from the MOOC registrant who, at least with edX, has a fairly good idea of her rights to privacy (as edX sees it). Unfortunately, few websites provide a privacy policy as readable as edX's, which lists specific purposes for data disclosure in a bullet point format;[283] even fewer websites have a federal statute that might even tenuously apply to how they use data. Meaningful — that is, clear and relatable — transparency in data collection and use, coupled with the ability to control collection, deletion, and subsequent use of data, would go a long way towards restoring not simply a nebulous conception of privacy, but perhaps more significantly, our ability to be master over how we are perceived.

283. *See edX Privacy Policy, supra* note 146.