THE DE-IDENTIFICATION OF MOOC DATATSETS TO DEMONSTRATE THE POTENTIAL FERPA COMPLIANCE OF MOOC PROVIDERS

A thesis presented by

Michelle H. Lessly, M.Ed.

To

Doctor of Law and Policy Program

In partial fulfillment of the requirements for the degree of
Doctor of Law and Policy

College of Professional Studies
Northeastern University
Boston, Massachusetts

June 2016

ACKNOWLEDGEMENTS

Completing this thesis was not a solitary task. I want to thank the faculty and staff of the Doctorate of Law and Public Policy at Northeastern University. I want to extend a special extension of gratitude to Dr. Edward F. Kammerer, Jr., my primary advisor, and Dr. Neenah Estrella-Luna for her patience and support throughout this endeavor. I would like to thank William D. McCants, Esq., my second reader. Additionally, it is with deep appreciation that I want to recognize my peers and friends, Cohort VIII. I am forever grateful for your challenge, support, and friendship over the past few years, and the many years to come.

I also want to recognize my family and friends who supported me throughout this program. Specifically, I want to thank my parents who have been an unrelenting source of encouragement. Since I was young, you have provided me the resources and opportunities through which I could pursue my dream of earning a terminal degree. I am proud to be your daughter; I hope I have made you proud in return.

Additional thanks to: Clinton Blackburn, Todd Karr, John Daries, Rachel Meidl, Monqiue Cunningham Brijbasi, Keenan Davis, Ted Johnson, Bryan Coyne, Noradeen Farlekas, Jalisa Williams, Joni Beshansky, Michelle Puhlick, Jonathan Kramer, Melissa Feiser, Melody Spoziti, Dr. Anne McCants, Jon Daries, Julie Rothhaar-Sanders, Nivedita Chandrasekaran, Rebeca Kjaerbye, Kristen Covino, and the many friends and colleagues who supported me throughout this program.

ABSTRACT

The disruptive technology of massive open online courses (MOOCs) offers users access to college level courses and gives MOOC providers access to big data concerning how their users learn. This data, which is often used for educational research, also includes users' personally identifiable information (PII). The Family Educational Rights and Privacy Act of 1974 (FERPA) protects PII and the educational records of students who attend traditional educational institutions, but the protection of this legislation is not currently extended to MOOC providers or their users.

A legal analysis of FERPA demonstrates analogous relationships between key statutory definitions and MOOC users, providers, and their datasets. By imposing the *k*-anonymity and *l*-diversity standards, this replication study of Daries et al.'s (2014) work attempts to de-identify MOOC datasets in accordance with C.F.R. Title 34 Part 99 Subpart D §99.31 (b)(1) and (2) to exhibit how to redact these datasets to be FERPA compliant and still maintain their utility for research purposes. This study also seeks to determine if this de-identification method can be standardized for universal use across MOOC providers.

The replication study, coupled with the legal analysis, suggest FERPA may not be the proper statute to regulate the privacy protections MOOC providers afford their users. Rather, the U.S. Department of Education and Congress should promulgate policy that outlines the minimum privacy standards MOOC providers and other disruptive technologies afford their users. Future research will aid in determining best practices for de-identifying MOOC datasets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Massive open online courses (MOOCs) offer a promising 21st Century solution to the problem of access and affordability of higher education. Initially launched in the United States in 2011, MOOCs offer low-to-no cost college-level courses through partnerships with universities or corporations. This disruptive educational model differs from the traditional college model or online courses. MOOCs have no admission requirement and occur entirely online, allowing thousands of users from around the world to simultaneously take a class to learn from each other through interactions on discussion forums (Jones & Regner, 2015, Young, 2015). These courses are often offered on demand and deliver course content through videos, filmed lectures, discussion boards, forums, readings, and homework, all without the active intervention of a professor. MOOCs, operated by third party providers, can be affiliated with a post-secondary institution such as Harvard and MIT's edX, the only open source, nonprofit MOOC provider (edX, 2016). They can also operate as a private company such as Udacity and Coursera, both of which were co-founded by former Stanford professors.

MOOC enrollment continues to grow annually by 6% (Allen & Seaman, 2014), now reaching approximately 16 million users worldwide (Shah, 2014). The New York Times declared 2012 as the "year of the MOOC" (Pappano, 2012), but by 2014, skepticism regarding the MOOC revolution was at an all-time high (Friedman, 2014). This doubt may have been propelled by developmental setbacks such as San Jose State University's unsuccessful attempt to offer Udacity courses to its underprepared students (Rivard, 2013)[1]. Numerous reports reveal MOOC

[1] In January 2013, San Jose State University announced a pilot program, in partnership with Udacity, to offer three entry-level courses MOOC courses to matriculating students (Fain, 2013). However, due to poor student performance, the pilot was cancelled in June 2013 (Rivard, 2013).

course attrition rates consistently teeter between 90-96% (Pope, 2014). Still, the claims that MOOCs miss the mark overlook the innovations they contribute to the field of educational technology and research. The truly transformative nature of this non-formal education platform rests not in the method of knowledge delivery or course retention rates, but the in opportunities it creates for the analysis of knowledge acquisition, especially in the digital age.

MOOCs have a multi-pronged business model, for in addition to providing access to college courses, MOOCs function as education data warehouses. This information is known as metadata, or "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (National Information Standards Organization, 2004, p. 1). For MOOCs, this includes users' personally identifiable information (PII)[2] as described in the Federal Rights and Privacy Act of 1974 (20 U.S.C §1232g; Title 34 CFR Part 99), commonly known as FERPA, as well as data about the amount of time a user spends watching a video, mouse clicks on a page within the course's site, and the frequency in which a user logs onto the learning platform. With an average of 43,000 registrants per course (Ferenstein, 2014), one MOOC course can generate up to 20 terabytes of data (Hazlett, 2014). Such collections of metadata can accumulate to become big data, which are datasets that are not only massive, but are easily searchable and sortable (Boyd & Crawford, 2012) and are retained for the purposes of evaluating minute details to determine patterns or trends within representative sample populations (Young, 2015).

Big data creates privacy concerns for both users and data holders. As a wider cross-section of organizations and companies collect data on the different facets of a user's life which,

[2] FERPA defines PII as the student's name, the student's family member's names, the student's address, personal identification numbers, other indirect identifiers such as birthdate, and other information that may be linked to a specific student (Title 34 Part 99 Subpart D §99.3).

when compiled into a digital dossier, creates new privacy challenges. Big data has an exceptional ability to connect seemingly isolated pieces of information to create a holistic depiction of an individual's identity. These digital dossiers create a tension between the utility, or usability, of big data and expectations for consumer privacy grounded in law and ethics. The legal and ethical framework that guides data management must be broad enough in scope to address the potentially conflicting needs of both data holders and the individuals providing the content of the dataset. Dataset owners must respect those individuals by assuming the responsibility for protecting their privacy rights (Hoser & Nitschke, 2010).

Within the context of education, FERPA, a federal statute, and its attendant regulations with interpretative guidance detail the regulatory obligations schools have when safeguarding student data. This law protects student privacy by regulating the collection, retention, and distribution protocols educational institutions use to collect the information included in student educational records. Unfortunately, the protections afforded to traditional students have yet to be extended to MOOC users since the U.S. Department of Education has not yet determined if MOOCs providers are classified as an educational agency under FERPA (Young, 2015). This leaves some educators to speculate that the Department does not believe it has the authority to determine if FERPA is applicable to this new learning platform (Kolowich, 2014). This conjecture is further supported by the fact that MOOC providers do not currently receive federal funding, a prerequisite of the FERPA compliance structure. Moreover, to further complicate the question of applicability:

> 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 (Young, 2015, p. 578).

Thus, FERPA was not prepared for the reciprocal partnership between MOOC providers and postsecondary institutions when it was conceived over four decades ago, and neither the Dept. of Education nor Congress have made intentional steps to address the issue of MOOC user data privacy. MOOC providers have not been officially recognized by the Dept. of Education as educational agencies, and since MOOCs do not receive federal funding, which would require them to comply with FERPA, MOOC users are left without the same safeguards afforded to their university student counterparts who attend the same class in-person or on-line. edX is currently the only MOOC provider that voluntarily complies with FERPA (edX, 2014).

That said, MOOCs are becoming a more widely-accepted form of higher education, as demonstrated by the partnership between edX and Arizona State University (ASU). Their collaboration, known as the Global Freshman Academy, offers for-credit courses to matriculating students at a significantly reduced tuition rate. MIT’s MicroMaster’s admissions-free program provides would-be students the opportunity to take an entire semester’s worth of courses on the edX platform before taking a qualifying exam in order to earn admission to the on-campus, one-semester full master’s degree program. Since the MIT’s MicroMaster’s program requires taking edX courses as part of the degree, might the enrollees of the program be classified as MOOC users or students who should receive FERPA protection?

Therefore, the question of whether MOOCs should comply with FERPA warrants an urgent response from the Dept. of Education. This leaves the related question of whether MOOCs can be compliant with FERPA and still generate usable data for the purposes of research. Policy makers must address the conflict between the regulatory requirements of FERPA and the uniqueness of MOOCs. Examining this conflict through the lens of digital privacy theory, and Solove's taxonomy of privacy will provide a critical perspective and necessary understanding of how the Dept. of Education should address MOOCs' evolving impact on the American higher education system.

This study seeks to provide a solution to this burgeoning policy concern by asking: in what ways might MOOC provider datasets be de-identified to meet the requirements of C.F.R. Title 34 Part 99 Subpart D §99.31 (b)(1) and (2) of the Family Educational Rights and Privacy Act of 1974, and still maintain their utility for the purposes of research dissemination? To answer this question, an examination on the literature on MOOCs, FERPA, and digital privacy theory to will provide the context in which MOOC providers and policy makers must resolve this issue. A legal analysis on the legislative and judicial history of FERPA will inform a methodology of de-identifying MOOC datasets to be FERPA compliant. The results of this study will yield recommendations for MOOC providers, researchers and policy makers to resolve the concerns of user privacy, data utility, and the potential need for MOOC to comply with FERPA.

Chapter 2

Literature Review

MOOCs and Public Policy

MOOCs first focused on providing open access to courses at globally-recognized, highly ranked universities such as Harvard, Oxford, Stanford, and MIT. They have since evolved to offer courses ranging from Google-developed coding classes to public relations seminars and conversational English courses for non-native speakers. Though a general level of digital literacy is required for MOOC course navigation, users are not limited by course prerequisites or admissions requirements to enroll in their course of choice. MOOCs operate under an open learning model, requiring users to rely on self-motivation to progress through a course, rather than external motivators such as deadlines for homework assignments or attendance requirements. Moreover, by divorcing online learning from the matriculating enrollment model at a traditional university, MOOCs have developed into a new type of non-traditional educational program.

In light of the collaboration between ASU and edX to create the Global Freshman Academy, many MOOC providers and postsecondary institutions are exploring, and in some cases implementing, such hybrid educational models. The American Council on Education (ACE) recommends colleges and universities offer credit for up to five MOOC courses (ACE, 2013). By 2013, both California and Florida state legislators considered recommendations to make MOOCs part of the degree-granting curriculum for their public college systems. While Florida legislators did approve the use of MOOC classes in the K-12 system, concerns regarding course quality prevented expanding the bill to public postsecondary institutions (Inside Higher Ed, 2013). Faculty union fears prevented California lawmakers from making MOOCs a component of the state’s three public higher education systems (Kolowich, 2013). In Arizona, however, the Global Freshman Academy drew over 34,000 registrants in its first year by offering 6 credit-granting, transferable classes at $200 per credit hour (Straumsheim, 2015). This MOOC hybrid-model, if proven successful, challenges the traditional post-secondary education experience.

This new type of educational experience is at the center of the FERPA compliance problem for MOOCs as it presents many challenges for MOOC providers, their university partners, and legislators. The amorphous state of the MOOC provider does not match current legal constructs (Jones & Regner, 2015), nor do the privacy and safety needs of MOOC users equitably align with current legislation. For example, the Cleary Act requires colleges and universities to track crime data on and around their campuses, but is a MOOC required to report threats or an incident of sexual harassment between two students on a course's discussion board? Can a MOOC provider's course site be considered a campus? What if these two students reside in different countries?

The hybrid MOOC model presents even more of a challenge for FERPA in that its requirement for compliance requires a student's enrollment at a recognized educational agency that receives some form of federal funds. If a user signs up for a university-created, certificate granting course through edX's platform, is the user enrolled as student at that FERPA regulated university, or at edX, which is not currently an educational agency under FERPA rules? Or, is the user not entitled to any of the FERPA protections available to a student in a physical classroom?

The President's Council of Advisors on Science and Technology (PCAST) recognized the range of MOOC related privacy challenges in their 2014 report. The big data element of MOOCs makes protecting user privacy much more demanding than in the case of a traditional student whose FERPA-protected information is confined to PII, including their name, birthday, and email address, and their educational record which contains information such as graded coursework and transcripts. PII does not include the wide range of metadata collected by MOOCs such as a user's highest level of education, how many times they watched a course-related video, or the date of their last activity on a discussion board. Thus, the majority of the information held by MOOC providers likely may be unregulated, even if FERPA were to apply (Young, 2015). This does leave tort law as a potential safeguard for metadata, but an ideal privacy apparatus protects both PII and metadata. Thus, PCAST's recommendations for privacy protections include encryption[1] and de-identification by removing full and quasi-identifiable[2] variables from a dataset (Daries, Reich, Waldo, Young, Whittinghill, Ho, & Chuang, 2014). These recommendations surpass FERPA's current privacy regulations, demonstrating the revisions necessary to bring FERPA up-to-date with digital privacy needs. No longer is simply redacting PII sufficient to protect a student's identity. Lawmakers must contemplate the totality of the data collected on students when promulgating privacy legislation.

The Family Educational Rights and Privacy Act of 1974

First introduced as the Buckley Amendment and signed into law by President Ford in the summer of 1974, FERPA enables students to control both the access and content included in their educational record (Graham, Hall, & Gilmer, 2008). This statute regulates the privacy needs of students in the K-12 system by allowing both students and their parents to have the ability to review and correct their educational record. FERPA does revoke parental review rights for

[1] PCAST defines encryption as the process which converts data into cryptography-protected rendering it useless to those without the decryption key.

[2] Quasi-identifiers are pieces of data that, when combined with other data, can generate the ability to uniquely identify an individual. Examples include gender and birth date (Sweeney, 2002).

students once they turn 18 or are enrolled in a post-secondary institution, but it otherwise remains applicable to colleges and universities.

Compliance is required of all institutions that receive federal funds, including federal student aid and grant monies. Withholding these funds is the only statutorily authorized enforcement mechanism permitted. However, when a FERPA complaint is filed, the Dept. of Education prefers to resolve the matter through administrative actions such as required policy revisions or trainings (Family Policy Compliance Office, 2015), rather than revoking federal funds. The consequences of the latter not only penalize the academic institution, it can also have significant, negative repercussions that are passed on to the student. Revoking an institute's federal funds due to a FERPA violation potentially means the institution can no longer afford to educate students in the same way prior to the complaint. To date, the Dept. of Education has not withheld funds for a FERPA violation (Young, 2015).

Since 1974, FERPA has been amended eleven times (20 U.S.C §1232g). As a result, this statute is notoriously challenging to interpret and at times seems contradictory. Until 2008, the Dept. of Education actively abstained from providing clarity for colleges and universities on how to interpret and implement FERPA (Lomonte, 2010). In that same year, the Secretary of Education issued an amendment to FERPA in order to implement stricter written notification requirements for the release of student records to a third party, including parents, while simultaneously making notification exceptions when information is released for the purposes of research (Ramirez, 2009, Family Educational Rights and Privacy Act, 2008).

These recent amendments demonstrate the conflicting nature of the privacy expectations of students and their institution's need to share student information for the purposes of scholarship or safety. They highlight that FERPA was created in a time when its drafters were unable to conceive of a virtual learning environment in which the scope of personally identifiable data collected would be much more expansive than the current statutory definition of PII. FERPA permits disclosing student data when the PII is de-identified[3], but how might this process be accomplished to scale for a MOOC course?

The historical interpretation of FERPA's standard for de-identifying student PII may not be enough to prevent the re-identification of MOOC users. The removal of PII in compliance with FERPA will still leave behind additional quasi-identifying information, such as VPNs, gender, and online user-generated content, which can be used to re-identify MOOC users. Unfortunately, FERPA does not account for these quasi-identifiers. Therefore, even once a MOOC dataset is de-identified according to FERPA's regulations, the statute's safeguards will not be applied to the dataset's quasi-identifiers, leaving that information public and unprotected.

Theoretical Framework

Digital Privacy Theory. As MOOC providers continue to develop their ability to gather both PII and quasi-identifiers from their users, the need to ensure individual users' privacy grows. However, increasing privacy protections on this data may negatively impact the utility of the dataset. To combat this problem, MOOC providers might employ the *k*-anonymity algorithm (Sweeney, 2002), the *l*-diversity standard (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007), and Dwork's (2008) differential privacy model.

Sweeney's k-anonymity Algorithm. In an effort to better secure privacy within datasets while retaining research utility, Sweeney (2002) recommends employing the *k*-anonymity algorithm. Using *k*-anonymity on an individual-data point structured dataset, can "produce a release of the data with scientific guarantees that the individuals who are the subjects of the data

[3] See C.F.R. Title 34 Part 99 Subpart D §99.31 (b)(1), (2)

cannot be re-identified while the data remain practically useful" (p. 557). To be successful, a *k*-anonymous dataset maintains a value of *k*-1 between data points, or attributes, which reduces the ability to re-identify an individual based on the totality of the information provided in the dataset. By utilizing anonymization through the methods of generalization and suppression, *k*-anonymity introduces noise into a dataset to dilute the information to make it comprehensively secure and maintain utility.

The two redaction methods, generalization and suppression, alter data while retaining the type of attributes collected within the dataset. It is through these two methods that noise is injected into the dataset and generates the *k*-value between attributes. Through generalization, a specific attribute is removed but still captured through a generic, yet representative category. It replaces specific attributes, such as ages or other data that can be represented accordingly, with ranges. For example, using generalization, a 25-year old male who lives in Boston, MA could be represented in a *k*-anonymous dataset as a 25-29-year old male who lives in the region of New England. However, this method only works for certain types of data. Suppression is employed for data cannot be easily generalized. As in the previous example, the gender of the 25-year old male could be represented in a *k*-anonymous dataset as a symbol, most commonly an asterisk, indicating the data was collected but suppressed for the purposes of anonymization. It is important to note generalization and suppression may be used alone or in combination depending upon different types of data and different research questions.

L-diversity. Whereas *k*-anonymity is a fairly comprehensive data privacy theory, Machanavajjhala et al., (2007) argue it still provides contextual information in which individuals may be re-identified. *l*-diversity adds an additional level of protection for datasets that are sensitive to privacy breaches due to the totality of the data made available to the public, including not only the attributes represented in the data, but the background of the attributes. Therefore, even if the 25-year old male who lives in Boston, MA is represented in an *k*-anonymous dataset as * (25-29-year old) who lives in Massachusetts, if that information is published in an unaggregated manner that provides the context in which the data was collected, the *k*-anonymous data is still vulnerable to attack. These attacks fall into two category types: homogeneity attacks and background knowledge attacks.

Homogeneity attacks occur when the attributes in the dataset are not diverse enough to create true anonymity on an individual level. For example, an attacker may know a user enrolled in a MOOC who is a prolific poster on the course's discussion board. The attacker knowing that person's age, gender, zip code, and course may be able to determine how many posts that individual made, through the process of elimination, if given access to that class' discussion board. Homogeneity attackers do not need to know the user, but rather simply have access to that user's demographic information to make an identification.

Background attacks build on homogeneity attacks by using contextual information to make an identification. Background attacks are a result of an attacker having personal knowledge about a user and making connections between sensitive data and quasi-identifiers based on societal background knowledge or information on a specific population represented in the dataset. Continuing with the previous example, if the attacker also knew the user was struggling with the course content and sought assistance from others in the class, the attacker may be able to determine which posts were the user's. This example demonstrates a background attack using quasi-identifiers. Based upon the vulnerability presented by these attacks, the *l*-diversity algorithm increases the noise in a dataset by increasing the diversity of sensitive attributes. However, as sensitive attributes become more *l*-diverse, the utility of the data may be reduced.

The Differential Privacy Model. Dwork (2008) also challenges *k*-anonymity, claiming there is no such thing as an impenetrable privacy protection algorithm, and suggests the differential privacy model provides a more optimum anonymization solution. This algorithm uses noise by interjecting it on the release mechanism of the data, not the data itself. The layering of protection, that is encoding the data release rather than the data through methods such as generalization and suppression, interferes with an attacker's ability to accurately capture information or to trace back the information to re-identify individuals and retains the utility of the data for the purposes of analysis. The differential privacy model focuses on producing information about the data released in a published dataset.

This algorithm prevents an attacker from being "able to learn any information about any participant that they could not learn if the participant had opted out of the database" (Tockar, 2104, n.p.). By adding noise to the release mechanism, such as a chart or graph, an attacker is unable to determine seemingly random patterns in the data that may lead to re-identification. Thus, the differential privacy model redefines the concept of digital privacy, moving from a system that attempts to defend the entire dataset against attacks, to a tiered design that makes datasets systematically less vulnerable when an inevitable attack occurs.

Solove's Taxonomy of Privacy. In the context of MOOCs, user privacy should not simply be reduced to the application of a security algorithm or a debate about identity protection. A more satisfactory understanding of user privacy looks beyond anonymity and scrutinizes the rationales behind the collection of the data in order to determine if it should be collected in the first place. Solove's (2008) taxonomy of privacy provides a framework for MOOC providers to ethically develop and disseminate their user-populated datasets while maintaining the necessary type of privacy. His argument that a single concept of privacy is not constant and cannot be consistently applied reflects the complexities of the MOOC user privacy issue. By shifting the locus of privacy from the data owner to the data subject, Solove's taxonomy can explore the impact of the integration of six privacy concepts: the right to be left alone, limited access to the self, secrecy, control over personal information, personhood, and intimacy.

The concept of the right to be left alone is the underpinning for today's privacy torts and is similar to the notion that privacy is limited access to the self, a principle that insists an individual should be the gatekeeper of their own personal information (Warren & Brandeis, 1890). The concept of secrecy, as popularized by Posner (1978), is the "appropriation [of] social benefits to the entrepreneur who creates them while in private life it is more likely to conceal discreditable facts" (p. 404). The desire for secrecy leads individuals to limit access to information about themselves and leads to the concept of control over personal information, which recognizes information as one's personal property. The concept of personhood expands upon that of personal property by viewing one's information as a manifestation of one's identity and reputation. Finally, the concept of intimacy asserts the need to keep information private is not just for the protection of one's self, but to secure the information of those with whom the individual may be associated. Whereas Sweeney and Dwork consider privacy from the utilitarian perspective of the dataset owner, Solove recognizes that it is the individual who assumes more risk when a third party, such as a MOOC provider, collects and disseminates data.

This becomes especially problematic due to exclusion, or "the failure to provide individuals with notice and input about their records" (Solove, 2006, p. 521). Exclusion presents a harm different from that of data privacy and security in that rather than being concerned with re-identification, exclusion removes an individual's ability to control what happened to their data (Solove, 2006). FERPA's primary goal is to eliminate exclusion, but it is this goal that further complicates the application of FERPA to MOOCs. In order to register for a course, users are often required to agree to the MOOC provider’s terms of service, which can exclude them from the decision making process as to how and when their information is used, or to have the ability to review the data to ensure it is an accurate portrayal of their identity. This may become problematic if MOOCs are required to become FERPA compliant, as it requires educational agencies to grant students access to their educational record and the ability to correct it when necessary. That said, those terms of service agreements that do not align with FERPA may become void under the law, which easily resolves the policy concern, but still leaves MOOC providers with the responsibility to audit massive amounts of data to ensure compliance.

When examining digital privacy from the user’s perspective, Solove’s model highlights the porous nature of the relationship between data subjects and data holders. To rectify this, the taxonomy identifies four activities of the data collection process: information collection, information processing, information dissemination, and invasion. The taxonomy’s intentional design around the data subject, identified as “the individual whose life is most directly affected by the activities classified in the taxonomy” (Solove, 2008, p. 103), and not around a specific privacy conception, allows for the evolution of privacy needs in the digital age.

A MOOC provider’s act of collecting information includes user registration information and the surveillance of their subsequent activity online. This leads to the second action in the taxonomy, processing information, which may be aggregated and analyzed without user knowledge. Though the purpose of MOOC data research includes learning about the potential functionality of the platform and to expand the field of knowledge on education technology, sharing this information can violate user trust. Moreover, the third activity, information dissemination, reveals the vulnerability of MOOC users’ information. Poorly managed user information creates opportunities in which information may be inappropriately disclosed or privacy agreements may be violated, leading to the fourth activity of invasion. If a user's information is improperly disclosed, leading to an attack on their personhood, what impact might this have on the likelihood they will feel safe enough to enroll in another MOOC course?

Critical Review of the Literature

If the Dept. of Education is to evaluate the relationship between MOOC providers and user privacy concerns, so must it consider FERPA's definition of PII as it pertains to big data. The current statutory standard for de-identification is reducing or eliminating PII to create a reasonable determination of anonymization (C.F.R. Title 34 Part 99 Subpart D §99.31 (b)(1)). This binary conceptualization of privacy successfully operates in a traditional educational setting, but cannot be reasonably applied in an-online setting. Metadata, such as the course name, when the course started, and the user's VPN, are quasi-identifying data points that may be concatenated for the purposes of re-identification (Daries, 2014). The current assumption, that redacting what FERPA clearly considers to be PII provides sufficient user privacy protections, is antiquated and may not hinder MOOC providers from openly sharing quasi-identifiers.

However, an examination of the relevance of the current understanding of PII in a digital learning environment might be irrelevant as some critics suggest FERPA does not pertain to MOOCs. Since the Dept. of Education has remained silent on the matter, MOOC providers currently have the liberty to make their own determination as to whether or not their course users are protected by FERPA. Both Udacity and Coursera make no mention of their stance on FERPA on their websites, whereas edX, a provider owned and operated by Harvard and MIT, specifically states it complies with FERPA.

Still, the undetermined status of FERPA's applicability to MOOCs has the potential to diminish the future utility of different providers' data (Hollands & Tirthali, 2014). For if the Dept. of Education or Congress determine that MOOC providers are required to comply with FERPA or other privacy regulations, those MOOC providers that have decided to not create FERPA compliant datasets may be limited in their capacity to operate under their own business models when attempting to share data with researchers. Moreover, if MOOC providers have no clarity on what may legally or ethically be released, how then are researchers to take advantage of MOOC-sourced big data?

Yet, a determination of mandatory compliance will not immediately resolve the issue of user data privacy. Standardizing the privacy protection practice of traditional colleges and universities is seemingly impractical if not impossible in the MOOC classroom. Whereas redefining PII will aid in privatizing data, it does not remedy the problem of user exclusion (Solove, 2011). MOOC providers require users to agree with their terms of service when registering for a course, but the efficacy of these documents is dubious (Solove, 2013). Terms of service agreements often rely on the average user not being well versed in the language and structure of such documents, leading to common user misperceptions about the quality of privacy controls (Turow, Feldman, & Meltzer, 2005). Since less than ten percent of individuals actually read a terms of service agreement when registering for an online service (Smithers, 2011), trusting in such contracts as a form of user consent for metadata collection is questionable at best.

Fair Information Practice Principles (FIPPs) should be used to reduce users' confusion about their waived privacy. FIPPs insist that data holders act ethically with their data by maintaining transparency of the data management process, keeping users informed of what personal data is recorded, and to seek user consent when their data is repurposed (U.S. Department of Health, Education, & Welfare, 1973). Incorporating FIPPs into FERPA's regulatory structure will help to reduce user confusion over their privacy controls and increase MOOC provider accountability for data management practices. Or, MOOCs may use FIPPs and FERPA as guidelines to create their own data privacy protection standards.

Additionally, policy makers need also consider how the global scope of MOOCs will complicate statutory compliance. Whereas digital privacy theory can address the concerns regarding data protection, it cannot account for cultural privacy norms. Solove's taxonomy intentionally allows for applicability within a cross-cultural context, but it fails to anticipate how a culture's understanding of power dynamics ebb and flow through each activity of the data collection process (Sweeney, 2012). This can be especially problematic when determining how public policy applies to a MOOC dataset when the MOOC and the partner institution, or user, are not American. Notably, the European Union has very detailed requirements for protecting their citizens' privacy, even when their users are accessing education resources outside of the EU. Policy makers and MOOC researchers must pay additional attention to the issue of governance in an international educational setting.

The National Association of College and University Attorneys (NACUA) recognizes that the legal uncertainty surrounding FERPA and MOOCs may change at any point in time due to a number of factors. For example, in the instances when a user borrows federal funds to pay for a course, a professor incorporates MOOC course elements into their on-campus classroom instruction, or postsecondary institutions require students to enroll in a MOOC course to gain degree-seeking credit, MOOC providers will need to comply with FERPA (NACUA, 2013). It is an unreasonable expectation that MOOC providers, as they interact with hundreds of thousands of users and numerous institutional partners in a given day, to self-monitor for these factors that might change their compliance requirements. In order to optimize for both educational and research potential, policy makers should examine how MOOCs can be effectively regulated under FERPA.

Finally, the most prolific critics of MOOCs, university professors, claim this educational delivery platform jeopardizes their tradition of the academy and the American system of higher education. However, the data collected by MOOC providers may be advantageous in the classroom and when conducting research. Unfortunately, the vast majority of MOOC research is quantitative, and almost exclusively examines MOOCs from the perspective of user satisfaction. Shifting the focus of MOOC research from determining the efficacy of the delivery method to the utility of their user data will aid in the sustainability and mainstreaming of MOOCs in the education marketplace for the public, private, and online organizations.

Critiques and research on MOOCs can help MOOC providers and policy makers understand better the barriers to the platform's success. Rigorous studies of the San Jose State failure have led to vast improvements in course design and content delivery (Lewin, 2013). Investigations on open, self-directed learning indicate that user success may be contingent upon their perception of the security of the online learning environment (Fournier, Kop, & Durand, 2014). If users think their metadata is too readily accessible to MOOC provider personnel or believe that their privacy has been compromised, they are less likely to be retained (Hughes, Ventura, & Dando, 2007). There is a need for increased attention to metadata privacy and for regulatory oversight of MOOCs as a means of ensuring user retention.

Chapter 3

Method and Research Design

Objectives and Research Question

My study explored the feasibility of requiring MOOC providers to be FERPA compliant by asking in what ways might MOOC provider datasets be de-identified to meet the requirements of C.F.R. Title 34 Part 99 Subpart D §99.31 (b)(1) and (2) of the Family Educational Rights and Privacy Act of 1974, and still maintain their utility for the purposes of research dissemination. In addition to this question, my study also sought to determine a process to create standard, systematic method for de-identifying MOOC platform datasets.

My study was motivated by Daries et al.'s (2014) claim:

> It is possible to quantify the difference between replications from the de-identified data and original findings; however, it is difficult to fully anticipate whether findings from novel analyses will result in valid insights or artifacts of de-identification. Higher standards for de-identification can lead to lower-value de-identified data. . . If findings are likely to be biased by the de-identification process, why should researchers spend their scarce time on de-identified data? (p. 57)

To answer the research question, my study assumed a mixed methods approach by conducting a document review and legal analysis of FERPA, and attempting to replicate Daries et al.'s research on measuring the impact the *k*-anonymity standard has on a MOOC provider dataset while ensuring the potential for FERPA compliance. Daries and his team, comprised of MIT and Harvard researchers, examined the feasibility of generating "a policy-based solution that allows open access to possibly re-identifiable data while policing the uses of the data" (p. 58) according to the regulations promulgated in FERPA. Whereas Daries et al. approached the problem of de-identification for the purposes of finding an equilibrium between privacy and utility in advancing social research, my study examined the question of the application of C.F.R. Title 34 Part 99 Subpart D §99.31(b)(1) and (2) to a publishable MOOC dataset for the purpose of evaluating the feasibility of applying FERPA or other relevant public policies to MOOC providers in order to protect users and their data.

Table 3.1. Measures

Measures	**Definitions**
Can the de-identification process be successfully executed using the same protocol on sample MOOC datasets?	The de-identification process can be executed in the same manner on sample MOOC datasets and yield viable utility while maintaining FERPA compliance.
What is an acceptable level of utility?	Maintains a *k*-5 value of for quasi-identifying variables and l-diversity for sensitive variables while minimizing entropy of the dataset after the de-identification of explicit-identifying variables (Daries et al., 2014).

Daries et al.'s research focused on the first edX dataset to be made publicly available, known as the HarvardX-MITx Person-Course Dataset AY2013 (Person-Course). In an effort to validate and expand upon their work, my study employed the *k*-anonymity standard, a process in which data unique to a user are removed to reduce the risk of re-identification, on at least one dataset from two MOOC providers. Since FERPA does not require a precise value for *k*-anonymity, Daries et al. consulted the Department of Education's Privacy Technical Assistance Center standards and determined that a *k-value* of five (*k*-5) created a safely de-identified dataset and met MIT's standards for de-identification. My study used the same metric of de-identification.

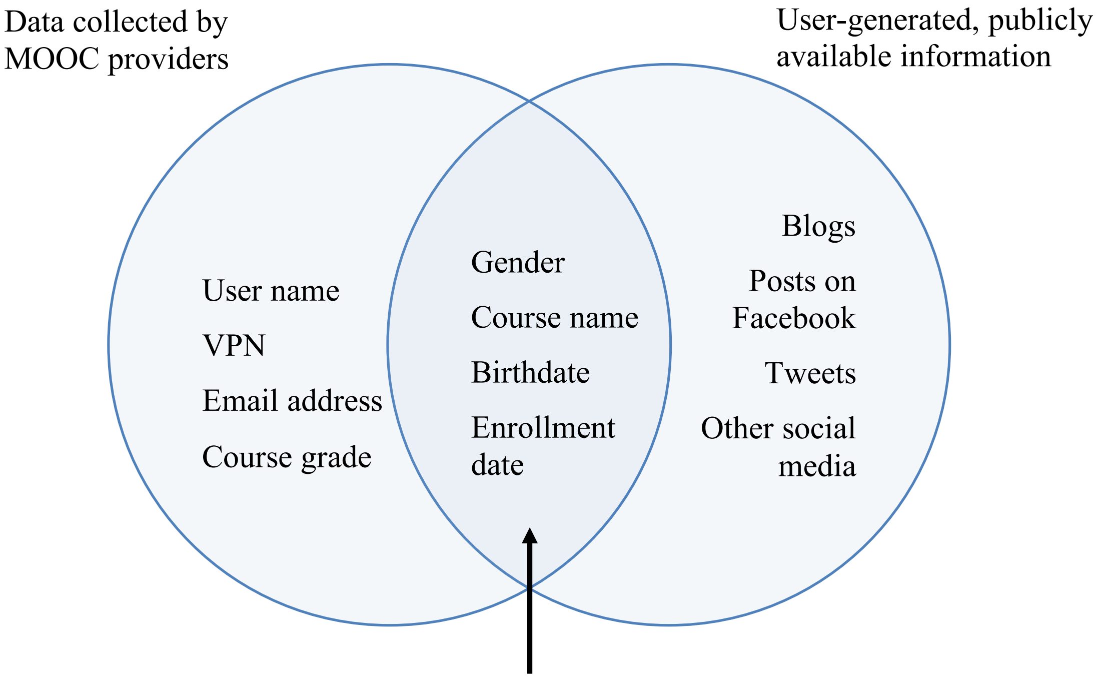
In keeping with the original research, I generated *k*-anonymous datasets through the methods of generalization emphasis and suppression emphasis. Daries et al. stressed the purpose of engaging both generalization and suppression emphases was to evaluate both methods' merits and challenges as it related to the utility impact of the data. Therefore, my study evaluated both generalization and suppression on their ability to better secure users' personally identifiable information and to meet the standards as promulgated in C.F.R. C.F.R. Title 34 Part 99 Subpart D §99.31(b)(1) and (2).

Understanding Daries et al.'s De-identification Process

Daries et al.'s method for de-identification included applying *k*-anonymity and *l*-diversity to MOOC datasets. Additionally, to quantifiably measure the shift in efficacy of the datasets, they employed a utility matrix as seen in Table 4.3. The authors' utility matrix was modeled after Dwork's (2006) utility vector, which combined descriptive and general statistics to assess the utility impact the de-identification had on the MOOC datasets.

***K*-anonymity.** To begin the de-identification process, Daries et al. determined which attributes, or quasi-identifiers, within the existing identified dataset should be removed to meet MIT's Institutional Research standards for both anonymization and report composition. The challenge in de-identifying Person-Course came with the amount of quasi-identifiers available within the data. One quasi-identifier may not be enough to distinguish a user, but as more unique attributes are made available, a holistic account becomes available making a user more vulnerable to attack. Additionally, if a user were actively posting about their MOOC experience on social media during the course, this increases the likelihood for re-identification based upon the information provided in the publicly available Person-Course dataset (see Figure 4.1).

Figure 3.1. Risk of Re-identification due to the Intersection of MOOC User Data, Quasi-identifiers, and User-generated, Publically Available Information



Potential data (quasi-identifiers) used to identify a MOOC user if not anonymized properly, Adapted from Sweeney, 2002.

Controlling for this potential variable was too challenging for Daries et al., but it was theorized that it could potentially be offset by using a higher standard for anonymization.

To do this, Daries et al. (2014) used Sweeney's (2002) *k*-anonymity model. In the case of a MOOC dataset, which can have quasi-identifiers ranging from username to the number of mouse clicks per page, a greater *k*-value is required to promote anonymity. For the purposes of the Person-Course dataset, the researchers assigned a value of *k*-5, meaning the "*k*-anonymized dataset has at least 5 records for each value combination" (Emam & Dankar, 2008, p. 628). In order for this de-identification approach to be successful, the researchers determined they needed to remove at least five quasi-identifiers from the dataset, which in turn served as a filtering mechanism in reducing the risk of re-identification. As *k*-value increases, the data's vulnerability to attack decreases. However, as Daries et al. noted, as the *k*-value increases, so does the likelihood that the utility of the data may be compromised.

To impose the *k*-anonymity model on the MOOC datasets, Daries et al. (2014) employed both the suppression and generalization emphases. The suppression emphasis removed identifiable attributes from the dataset and replaced it with a character to represent information that was collected and subsequently redacted. The generalization emphasis replaced attributes with corresponding or representative values. For example, in order to de-identify a dataset containing users' age, the suppression technique eliminated the cell value while maintaining the attribute category. The generalization technique replaced the cell value with an age range, as seen in Figure 4.2.

Figure 3.2. Example of Suppression and Generalization Emphases

	Suppression	**Generalization**
User_1 Age	*	20-24
User_2 Age	*	15-19
User_3 Age	<Null>	30-34

In the case of Person-Course, Daries et al. (2014) identified 20 attributes as variables that may be used to identify MOOC users (see Table 3.2). The attributes were categorized into two categories: administrative, meaning the data was generated by the MOOC provider or was generated by the researchers, and user-provided, which were data points generated by the user at the time of registration with the MOOC provider. Attributes that were altered as a result of the *k*-

anonymity process were tagged with the suffix *DI*. Null cells, or data that was not made available by either the MOOC provider or the user was indicated in the attribute *inconsistent_flag*.

Table 3.2. Variables

Attributes	**Code**	**Type**	**Description**
Course ID	*course_id*	Administrative	Course name, institution, and term
User ID	*userid_DI*	Administrative	Research assigned indiscriminate ID number that correlates to a given dataset
Registered for course	*registered*	Administrative	User register for a given course
Gender	*gender*	User-provided	Values include female, male, and other
Country of residence	*final_cc_cname_DI*	Administrative, user provided	IP address or user disclosed, was altered through generalization emphasis
Birth year	YoB	User provided	User's year of birth
Education	*LoE*	User provided	User's highest level of completed education
Registration	*start_time_DI*	Administrative	Date user registered for course
Forum posts	*nforums_posts*	Administrative	Number of user post to discussion forum
Activity	*ndays_act*	Administrative	Number of day user was active in the course
Class visits	*viewed*	Administrative	Users who viewed content in the course tab
Course interactions	*nevents*	Administrative	Number of user interactions with the course as determined by tracking logs
Video events	*nplay_video*	Administrative	Number of times user played course videos
Chapters accessed	*nchapters*	Administrative	Number of course chapters accessed by user
Chapters explored	*explored*	Administrative	Users who read at < half of chapters assigned
Seeking certificate	*certified*	Administrative	Users who earn a course certificate

Table 3.2. Variables, continued

Final grade	*grade*	Administrative, *l*-diversity sensitive	User’s final grade in the course
Activity end	*last_event_DI*	Administrative	Date of user’s final interaction with course
Non-user participant	*role*	Administrative	Classifies instructors or staff in the course
Null values	*inconsistent_flag*	Administrative	Classifies values that are not available due data inconsistencies

***L*-diversity.** Daries et al. (2014) also accounted for *l*-diversity in the de-identified Person-Course dataset. The researchers were able to create a *k*-anonymous dataset that was effective in reducing identification risks for individual MOOC users, but it still left the possibly for a "homogeneity attack" (Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2007, p. 3). This type of data breach capitalizes on an attacker's contextual knowledge of a given individual, perhaps learned through social media sites, and in employing deductive reasoning, as informed by the data provided in a k-anonymous data, can re-identify that individual. The initial *k*-anonymity process yielded individual-user population groups with sensitive variables that might be used for re-identification. In the case of Person-Course, by knowing how a user was classified in a few sensitive variable categories, such as date of enrollment, course name, and their IP address at the time of their involvement in the course, it might be possible to determine which specific user posted on a discussion board on a given date.

L-diversity could also be used to reduce statistical based reasoning data breaches known as "background knowledge attacks" (Machanavajjhala et al., 2007, p. 4). This type data breach allows an attacker to capitalize on the information they have about a specific demographic of user and might enable the attacker to use that information to reduce number of attributes to be examined when attempting to identify a specific user. However, for the purposes of their research, Daries et al. (2014) decided to focus only on their datasets' vulnerability based upon a homogeneity attack.

After the Person-Course dataset was de-identified for *k*-anonymity, Daries et al. (2014) assessed the data for *l*-diversity sensitive variables, or attributes that may be especially vulnerable if an attacker learned of their values. For example, a study about students in a traditional college course may provide the gender, age, and ethnicity of the learners, but in order for the data to be considered *l*-diverse, the sensitive variable of a student's GPA would need to be redacted in order to protect the privacy of those students. For the purposes of Person-Course, Daries et al.'s (2014) analysis determined that the only sensitive variable was final course grade (*grade*) and would be subject to removal from the dataset if believed to present homogeneity vulnerability. My research also ascribed the sensitive variable value to the *grade* attribute.

Replication of Daries et al.'s Method

To replicate Daries et al.'s (2104) study, I received approval from Northeastern University's Institutional Review Board and signed a data release with MIT's Office of Institutional Research. Correspondence with Daries provided access to a GitHub page featuring his study's de-identification process manual and the open-source Python code I used to de-identify my datasets. Daries also provided additional information regarding the background, theory, and process for his study via the MITx and HarvardX Dataverse which inclued the Person-Course Documentation (Daries, 2014) and Person-Course De-identification (Daries, 2014) files. I frequently consulted throughout the data collection, coding, and analysis processes.

Data Collection

The research process consisted of two distinct phases: the simultaneous document review and legal analysis of FERPA, and the coding of MOOC identified datasets. The document review and analysis included an evaluation of the case law that examines the application of the key terms found in C.F.R. Title 34 Part 99 Subpart A §99.3, and Subpart D §99.31(b)(1) and (2) which regulate the conditions in which an institution may disclose information without seeking a student's prior consent. The process of de-identifying the MOOC datasets included running the Python code-based program written by Daries.

FERPA Document Review and Legal Analysis. The document review and legal analysis was conducted in order to determine the statutory definition of key terms and regulations for the collection, retention, and dissemination of a student's education record. Subpart A §99.3 provided term definitions and Subpart D §99.31(b)(1) and (2) stipulated the regulations for releasing student information without that student's consent. The definitions and case law review provided the infrastructure for the analysis of both the de-identified datasets and the content included in the datasets that might be considered an educational record. The key terms reviewed included student, attendance, educational agency or institution, educational record, and personally identifiable information (PII). The review of Subpart D §99.31,9 (b)(1) and (2) provided the context in which the de-identification process would be necessary in order to permit the release of a dataset.

Sampling Populations for De-Identification Process. My study sought to expand the scope of Daries et al.'s (2014) study through purposive sampling which included datasets from the two most popular MOOC providers, edX and Coursera. These platforms were selected not only due to their prominence in the MOOC industry, but for their focus on accessibility to higher education, wide-range of course offerings, average amount of users per course, terms of service agreements, and user privacy policies. Udacity, the another popular MOOC provider, was not included in this study as it recently shifted its focus to providing courses solely on computer science and nanotechnologies through partnerships with corporate sponsors, not post-secondary institutions.

Datasets were requested from edX, Coursera, and Daries. edX was unable to provide datasets per their agreement with their partner institutions, but recommended requesting datasets directly from those partner institutions, which included MIT, Daries' home institution. Coursera did not respond to any inquires. My requests for datasets from 12 of Coursera's partner institutions were also denied. Daries responded by providing instructions for requesting access to the datasets he and his team used for his study, which were MITx courses hosted on the edX platform, as well as links to the de-identification Python code stored on GitHub, an open-source, project hosting website.

Through the MITx Data Request protocol, I received access to four MITx datasets: MITx 2.01x (2013), MITx 3.091x (2013), MITx 8.02x (2013), and MITx 8.MReV (2013). These datasets were selected from the collection of the original 16 datasets used in the Person-Course study and were chosen due to the size of the user population. Sampling from courses with smaller user populations allowed for easier data management and the reduced number of records to be deleted. Yet these datasets were still large enough to be well representative of a typical MOOC dataset with a mean user population of 20,586. The datasets were stored on a secure, encrypted external hard drive and transferred electronically using a Pretty Good Privacy (PGP) key. Once de-identified and assessed, the original datasets were deleted.

Using the data request method suggested by edX, datasets were solicited from Coursera's partner institutions. Using convenience sampling, 12 institutions located in the United States, and thus could be potentially subject to FERPA compliance were contacted via email to requests access to their Coursera-hosted course datasets. However, no institution was willing to participate in this study. Even though partner institutions have unique, individual contracts with Coursera, many of the universities I contacted declined my request for data citing their terms of use agreement with the provider. These agreements prohibited sharing their participants' identities without seeking permission from the users whose information was included in the datasets (Coursera, 2015). Providing me with their datasets would require the partner institutions to contact potentially thousands of domestic and international users. Resources were not available to accomplish this task.

Attempting to De-Identify MOOC Datasets

The original de-identification code was forked, or imported, from Daries' GitHub page onto my GitHub page and then imported into the software program PyCharm. The datasets were also imported into private directory in PyCharm, which allowed for the code to be run on the raw dataset in a protected virtual environment. The data was then converted from SQL to CSV files and ran through the de-identification code in Jupyter Notebook. The results were imported and saved in PyCharm.

De-identification Code and Process. I attempted to de-identify the MITx 2.01x and MITx 3.091x datasets. Due to programming errors, I was unable to perform the de-identification process on the MITx 8.02x and MITx 8.MReV datasets. The de-identification progam was run on the MITx 2.01x dataset six times and the MITx 3.091x dataset once.

In order to prepare the datasets for de-identification, and per Daries et al.'s (2014) original research design, each user was given a 16-digit identification number comprised of both a unique identifier and the course ID. The datasets were then evaluated by quasi-identifiable, user-specific attributes: IP address, gender, year of birth, enrollment date, last day active, days active, and number of forum posts. I selected these attributes to be consistent with the original study. Daries et al. report choosing these variables due to their increased probability to be publicly available.

I used the generalization and suppression emphases on these attributes to reduce re-identification risks and delete extreme outliers in the dataset, which allowed for the analysis of the truncated mean. Country names, derived from the users' IP addresses, were changed to their respective geographic regions, and, in order to reduce skew in the results, users with 60 or more forum posts were deleted. Then the data was concatenated by stringing the quasi-identifier variables into groups no smaller than 5 students. In order to minimize the impact on entropy, the code was applied systematically to each quasi-identifier represented in the utility matrix. This process attempted to yield a *k*-anonymous and *l*-diverse dataset ready for its utility assessment.

I then attempted to determine the utility of the *k*-anonymous and *l*-diverse datasets by completing the utility matrix. Comprised of a nine by three grid, this matrix measured the de-identified dataset's entropy, mean, and standard deviation of each quasi-identifier (see Table 3.3). Generated by the Python code, this matrix was run on the original identified dataset and once again each time a variable was coded for *k*-anonymity. The utility matrix was to be recorded for each iteration of the analysis for each dataset, but the program encountered an error, preventing the utility matrix from being completed.

Table 3.3. PreUtility Matrix for MITx 2.01x

Variables	Entropy	Mean (n)	Standard Deviation
viewed	0.893515	0.689704	0.462615
explored	1.38352	0.194336	0.395689
certified	0.345462	0.0646054	0.245828
grade	1.80109	0.0692774	0.211211
nevents	8.29603	799.21	2229.94
ndays_act	4.177	9.48864	17.5364
nplay_video	5.49129	78.207	239.401
nchapters	2.92928	3.90965	3.7522
nforum_posts	0.640539	5.8006	26.4147

Analysis

Document Review of FERPA. I determined if MOOC providers' datasets could meet the statutory requirements of FERPA by analyzing the regulatory definition of the terms educational record, PII, student, and educational agency or institution as found in Subpart A §99.3. I also assessed if MOOC users may be considered students according to §99.3 and the relevant case law. An in-depth analysis of the statute's applicability to MOOCs is provided in the subsequent chapter.

Measuring *K*-anonymous Utility. The de-identified datasets were then analyzed to determine their utility. In the original study, this process allowed Daries (2014) and his team to quantify the impact the deletion of variables had on the accuracy of the de-identified dataset.

The analysis was to measure the change between the raw datasets and the *k*-anonymous, *l*-diverse datasets by employing a utility matrix (see Table 3.3) modeled on Dwork's (2006) utility vector. This matrix was also designed with the intent to measure the shift in a common metric in information theory known as Shannon entropy, mean, and standard deviation of nine nominal variables from the pre and post-de-identified datasets. However, due to unresolved bugs in the code, I was unable to measure the utility of the any of the k-anonymous datasets.

Limitations

My inability to gain access to a Coursera dataset was a significant limitation of this study. Without a representative dataset from a second MOOC provider, I was unable to determine if this methodology can effectively de-identify non-edX data. Therefore, I was unable to answer my secondary research goal of determining a standardized methodology for MOOC data de-identification. Additionally, Daries et al.'s (2014) did not provide the standards by which they determined if a dataset has maintained its utility. This is problematic as the utility impact may very dependent upon how the attributes are grouped, categorized, or eliminated. Also, currently there are no industry standards for quantifying dataset utility.

With the additional goal of creating a systematic process for de-identifying datasets that may be used on any type of MOOC provider and still maintain the dataset's efficacy, my study necessitated defining utility as the "the validity and quality of a released dataset to be used as an analytical resource" (Woo, Reiter, Oganian, & Karr, 2009). The values for entropy, mean, and standard deviation will be discussed in Chapter 5. The broad scope of this term offered a baseline understanding of what should be the resulting usability of a de-identified dataset. However, it must be noted that though a general definition of utility is provided for my study, in practice, utility may be determined on a case-by-case basis, dependent upon the needs of the individual using the dataset.

Finally, I encountered a number of bugs in Daries' program, which will be disucssed more in depth in Chapter 5. Due to these problems with the code, I was unable to complete the method in its entirety as outlined in this chapter. This limitation of my study is reflective of the problem with Daries' code, not the method itself.

Chapter 4

Legal Analysis

In the aftermath of the Watergate scandal, when the public's desire for governmental transparency was at an all-time high (Stone & Stoner, 2002), Senator James Buckley proposed an amendment to the General Education Provisions Act (GEPA) that would become the Family Educational Rights and Privacy Act of 1974, more commonly known as FERPA (20 U.S. C. §1232g). The rationale for FERPA, as articulated in Senator Buckley's initial appeal to Congress, recognizes the need to curtail the "abuses of personal data by schools and Government [sic] agencies" (120 Congressional Record, 14580). Months later in the *Joint Statement in Explanation of Buckley/Pell Amendment* (120 Congressional Record, 39862-39866), Senator Buckley claimed the purpose of the law is to provide both parents and eligible students the ability to review their education records, as well as limit the sharing of those records without student or parental consent in an effort to promote student privacy. FERPA was authorized as an amendment to GEPA, therefore it did not undergo Congressional committee review, limiting its legislative history to the Joint Statement (Stone & Stoner, 2002). FERPA became law in the summer of 1974.

Over the past 40 years, FERPA has been amended eleven times and faced significant criticism. Many of these amendments were enacted in response to nationally publicized, critical incidents in higher education, such as the Campus Security Act in 1990, the USA PATRIOT Act of 2001, and the Amendments of 2008 (Ramirez, 2009). However, because these amendments were made in conjunction with other laws, such as the Jeanne Clery Act, or as an addendum to the Higher Education Act, the legislative history for these amendments is also limited.

Despite these modifications, the statute's language is indisputably imprecise, leaving institutions to interpret the statute's terminology of educational record to meet their own needs

(Graham, Hall, & Gilmer, 2008). Until 2008, the Dept. of Education actively abstained from providing clarity for colleges and universities on how to interpret and implement FERPA (Lomonte, 2010). This hesitation by the Dept. of Education to offer more guidance on FERPA compliance is a consequence of the statute's lack of detailed legislative history.

FERPA regulates K-12 and post-secondary education systems, but critics suggest it fails to take into account the distinctive needs of these two very different populations (Lomonte, 2010). The Dept. of Education first recognized the disparate privacy goals of higher education students and institutions through its 2011 proposal to strengthen protections around statewide longitudinal data systems (L'Orange, Blegen, & Garcia, 2011). However, the application of this law and the sharing of information is contingent on multiple factors including timing, the relationship of the parties in question to the student, and the purpose for disclosure (Meloy, 2012).

To date, MOOCs have not been litigated in any United States Course. Therefore, the following examination of the statutory definitions of key terms in FERPA, and the review of applicable case law, is intended to be persuasive only. The cases presented are not an authoritative assertion of the binding precedent to be enforced on MOOC providers or MOOC users.

Statutory Definitions of Key Terms as they Pertain to MOOCs

For MOOCs and the privacy needs of their users, examining how the definitions included in FERPA and how these regulations relate to this learning platform is essential in determining if MOOC datasets can and should be de-identified in a manner that is compliant with FERPA. In order to determine qualifications for compliance, as determined for the purposes of this study, the terms evaluated in this analysis include student, attendance, educational agency or institution, educational record, and personally identifiable information (PII). These definitions provided in FERPA in §99.3, as authorized by 20 U.S.C 1232g.

Who is a Student? The statute defines a student as “any individual who is or has been in attendance at an educational agency or institution and regarding whom the agency or institution maintains education records” (§99.3). However, determining who and under what conditions an individual meets the statutory definition of a student is a complicated process. The term student appears 208 times in the statute, often in correlation with other key terms such as educational record or PII. This is especially problematic considering these terms heavily rely on the designation of student in their own definition. For example, FERPA classifies many of types of information that may be considered a component of an educational record, but each relies on the qualifier that it relates to the student in some way. The definition of a student is not independent from the term educational record, and the meaning of educational record cannot be understood without including the term student. The same is true of PII and attendance.

FERPA is only authorized to regulate records and information that pertains to students, therefore it is reasonable to conclude that the reliance on term student is necessary for the success of the statute, but is problematic due to its circular nature (Young, 2015). FERPA’s definition of “student” is vague, creating difficulty in determining if a new type of learner may seek protection under FERPA, or if a new learning platform may be subject to regulation.

The term student has maintained its original meaning from FERPA’s enactment in 1974. Without any amendments that directly address the definition of the term student, one must turn to case law in assessing if a MOOC user can be considered a student under the statute. The application of the definition of student is examined in a number of cases, including *Klein*

Independent School District v. Mattox, 830 F.2d 756 (5th Cir. 1987), and *Tarka v. Franklin*, 891 F.2d 102 (5th Cir. 1989).

Klein Independent School District v. Mattox. Under the newly established Texas Open Records Act, a request to review the college transcripts of Rebecca Holt, a teacher in the Klein Independent School District, raised questions regarding the FERPA rights of employees. *Klein v. Mattox* (1987) examines if FERPA may be used to protect educational records that are included in a personnel record. The United States Court of Appeal for the 5th Circuit held that, because Holt's relationship with the Klein Independent School District was as an employee and never as a student who attended classes within the district, she could not seek relief under FERPA.

Klein's significance for MOOCs extends beyond the definition of student and raises the question of the value of personal privacy when contrasted against the public's best interest in the context of FERPA. The court did suggest the need to vet the competency and credentialing of the school district's educators outweighs Holt's desire to keep her transcripts private, thus the release of such information does not constitute an unjustifiable invasion of privacy. The court did not interpret FERPA as upholding Holt's request for privacy when weighed against the countervailing public interest that favored disclosure. This raises the question: if the metadata collected by MOOC providers for the purposes of educational research creates a public interest that supersedes users' need for privacy, is FERPA the appropriate statute to regulate MOOC providers?

Tarka v. Franklin. *Tarka v. Franklin* (1989) also tests FERPA's definition of student by determining if an individual who was not admitted to a graduate program may, under FERPA, review their application file. Mark Tarka was denied admittance to the University of Texas Graduate School but was subsequently granted permission to attend classes at the University as an auditor. In order to understand the admissions decision, Tarka requested to review the letters of recommendation included in his application file, but the University rejected his request. Tarka filed suit alleging violations under several laws, including FERPA. The 5th Circuit found Tarka did not have a private right of action under FERPA[1], but it still had to determine if a FERPA violation could be the basis for a civil rights claim under 42 USC §1983. During this analysis, the court conducted a thorough review of the meaning of the term student under FERPA and determined Tarka did not meet the statutory definition.

As specified by the Joint Statement (see Table 4.1), the court determined Congress did not intend FERPA protections to extend to would be students, but did afford individuals who audit courses the rights and benefits of the law, even though the extent of those protections was not explicitly enumerated. The court ultimately held the University of Texas did not have to release the contents of Tarka's application file his auditing of courses at the university did not give him rights of review of that file under FERPA.

A preliminary reading of *Tarka* is problematic when making an argument as to why MOOC users should be classified as students, but a more in-depth reading of the case proves to be beneficial. MOOC users' relationship with both MOOC providers and partnership institutions is established on the premise that they will not be attending classes in the conventional sense, but rather will be attending class by logging onto their course's portal. Thus, using *Tarka's* interpretation of the term "student," MOOC users are students from the first time they click on their class' page. *Tarka* may also be interpreted to make the distinction that an individual who registers with a MOOC provider but does not sign up for a course or never logs onto their course's portal may not be classified as a student. This is problematic in that even though

[1] See *Gonzaga University v. John Doe*, 536 U.S. 273 (2002).

metadata will not be collected on these users as a result of taking a course, their directory and demographic information is collected through the site registration process.

MOOC users are not admitted to a traditional institution for the purposes of taking a course for credit or to matriculate, but the argument can be made that they may be analogous to auditor, if not full students. MOOC users do not participate in courses in a conventional way, but they do engage with the material in the same manner as class auditors; and, just as a traditional academic agency collects and retains information on class auditors, so too do MOOC providers. If auditors are eligible to seek protections under FERPA akin to that of a traditional student, then a similar burden of regulatory compliance should fall upon MOOC providers.

Table 4.1. Definitions of a Student

34 C.F.R. §99.3	*Student,* except as otherwise specifically provided in this part, means any individual who is or has been in attendance at an educational agency or institution and regarding whom the agency or institution maintains education records.
20 U.S.C. §1232g(a)(6)	[T]he term "student" includes any person with respect to whom an educational agency or institution maintains education records or personally identifiable information, but does not include a person who has not been in attendance at such agency or institution.
Joint Statement in Explanation of Buckley/Pell Amendment	The "student" to whom the right of access belongs is defined as any person concerning whom the educational agency maintains education records of personal information, but does not include anyone who has not been in attendance at such agency or institution. This means that the rejected applicant for admission is not given the right under the Buckley Amendment to see and challenge his letters of recommendation, nor does the amendment give him the right to challenge the institution's decision not to admit him. Such a right accrues only to the individual who actually attends the institution. For the purpose of this definition, a student who is only auditing a course, but on whom the institution maintains a personal file, would be included in the Amendment's coverage (120 Cong. Rec. 39865).

How is Attendance Defined? For MOOC users, the determination of whether they qualify as "students" under FERPA may not rest squarely on the definition of that term, but rather on the meaning of the term "attendance," as also defined by this law. FERPA states:

> Attendance includes, but is not limited to—in person or by paper correspondence, videoconference, satellite, Internet, or other electronic information and telecommunications technologies for students who are not physically present in the classroom; and the period during which a person is working under a work-study program (§99.3).

As *Tarka* demonstrates, attendance is a key component in determining when an individual, or user in the case of MOOCs, becomes a student eligible for FERPA protection. The period of attendance also functions as the bookends for when educational records may be collected. Thus, this term is essential in answering any inquiring relating to FERPA compliance. However, there is little discussion about the meaning of this term in the federal courts beyond this case. Reviewing the Joint Statement provides more guidance as to what constitutes attendance.

Though the Joint Statement does include a number of references to the term attendance, it does not offer a definition or insight as to meaning of the term. The Statement merely clarifies that a parent does not need to have a child currently attending an educational agency in order to request access to their student's record, and that an applicant who was denied admissions to an institution is not entitled to access their letters of recommendation. This limited information or insight as to what the drafters of the legislation understood attendance to mean requires deferring to the language including in §99.3.

The statutory definition of attendance is inclusive of a number of ways in which an individual might access course material, and now may potentially permit the incorporation of MOOCs. A plain reading of this text indicates MOOC users might reasonably be considered in attendance when registering and logging onto a MOOC provider's course portal. However, MOOC users do not have a direct relationship with the institutions that provide courses on MOOC platforms, but with the providers themselves. Therefore, it is also necessary to examine if MOOC providers meet the statutory definition of educational institution or agency.

Are MOOC Providers Educational Institutions or Agencies? An educational institution or agency, as defined in FERPA includes "any public or private agency or institution to which this part applies under §99.1(a)" for which "funds have been made available under any program administered by the Secretary [of Education]" (§99.1(a)). The language in the Joint Statement regarding what qualifies as an educational agency or institution is similar in nature to that of the statutory regulation, and offers additional clarity in that educational programs that receive federal funds, including Headstart and the National Institute of Education, which are covered under FERPA. Critics of the idea that MOOCs should be required to comply with FERPA suggest that their private status, or more specifically, lack of federal funding, makes the entire discussion moot (Young, 2105). The key qualifier for FERPA compliance is the receiving of federal dollars. Thus, if a MOOC provider does not meet this minimum standard, the question of compliance is irrelevant. This perspective is limited as it does not contemplate the vast amount of metadata collected by MOOCs that warrants privacy protestations, but is legally correct.

As the use of MOOCs grows in both the general population and in the classroom, the line between MOOC provider and institutional partnership becomes less clear. Integrating MOOC modules into to a traditional classroom setting or syllabus may make MOOC providers educational institutions or vendors by proxy, thus potentially increasing the burden for compliance (Pierson, Terrell, & Wessel, 2013). For example, hybrid models such as the Global Freshman Academy and MIT's MicroMaster's program, reliance on the role of MOOC provider as an educational agency is vital to their success. Users enroll in these curriculum-based programs with the expectation that they will become students at ASU or MIT based upon their performance and willingness to pay for an edX course.

Moreover, in October 2015 the Dept. of Education announced the Educational Quality through Innovative Partnerships (EQUIP) pilot program. This initiative permits individuals to use federal student aid to assist with tuition costs at non-traditional educational programs such as coding boot camps and MOOCs. This action by the Dept. of Education, as supported by the Obama Administration (Office of the Press Secretary, 2013), suggests MOOCs might qualify as an educational program as discussed in the Joint Statement. Fortunately, a definition for education program is included in FERPA:

> Any program that is principally engaged in the provision of education, including, but not limited to, early childhood education, elementary and secondary education, postsecondary education, special education, job training, career and technical education, and adult education, and any program that is administered by an educational agency or institution (§99.3).

The inclusion of this term, and its reference in the Joint Statement, suggests the drafters of the statute intended for educational programs, which now could potentially include MOOCs, to be protected under FERPA. MOOCs, due to their relationship with institutional partners, may be classified as an educational program or even a third-party vendor. Prior to MOOCs, institutions that might meet the statutory definition of an educational agency were essentially limited to traditional institutions or educational programs, and therefore there is little precedent or case law from which to work in determining how a hybrid educational model might be categorized under FERPA.

What Constitutes an Educational Record? When determining if a MOOC provider is an educational agency covered by FERPA and if their users can be classified as students who attend courses online, it is also essential to establish what constitutes an educational record. As defined in FERPA, this information includes "records that are directly related to the student; and maintained by an educational agency or institution or by a party acting for the agency or institution" (§99.3), but is not inclusive of memory aids, employee files, information generated due to normal business operations, and medical or treatment records. A sensible interpretation of this definition indicates information collected by MOOC providers may fit the statutory understanding of educational record. However, a review of the case law regarding such records may deliver a more precise reading. *Owasso Independent School District v. Falvo,* 534 U.S. 426 (2002), *State ex rel. Miami Student v. Miami University,* 680 N.E.2d 956 (Ohio 1997), and *United States v. Miami University,* 294 F.3d 797 (6th Cir. 2002) offer such an examination.

Owasso Independent School District v. Falvo. Kristja Falvo, the parent of grade school children in an Owasso, Oklahoma school asked that the school district to end the practice of peer grading as she feared it not only embarrassed her children but violated their FERPA rights. In a unanimous opinion, the United States Supreme Court held students grading their peers' papers were not acting as agents of the school and therefore their act of grading could not be considered a FERPA violation, nor the graded papers an educational record under FERPA. Moreover, the Court determined that a teacher's gradebook is not a mechanism through which an educational institution maintains student records.

Justice Kennedy, writing for the majority, stated, “FERPA implies that education records are institutional records kept by a single central custodian, such as a registrar, not individual assignments handled by many student graders in their separate classrooms” (p. 435). Other interpretations of the statute would create an excessive burden on instructors to protect all types of information and interactions with their students.

However, the single central custodian concept may fail to consider the complex nature of educational institutions. The number of instances in which students and school officials interact and result in the collection and retention of educational records cannot be reasonably managed in a central filing system. Moreover, this theory fails to provide a mechanism through which school can ensure all records are appropriately given to the central custodian, thus increasing the burden of liability on the institution or agency.

Regardless of any criticism of the concept, MOOC datasets are the exemplar of *Falvo’s* single central custodian concept, especially since the Court’s opinion recognizes the use of an electronic filing system. The case highlights the retention and sharing of the records is more important than the records themselves. For MOOCs, controlling access to and auditing these records could be a simple as making a few changes in the software’s line of code. The built-in record maintenance system diminishes the burden traditional educational institutions may endure in order to meet the single central custodian standard. However, tension among state and federal statutory expectations for privacy, such as variations in state freedom of information laws, creates discrepancies between what might be considered protected information from state-to-state (Daggett, 2008) and contributes to the confusion surrounding the ambiguous language used in FERPA.

Miami Student v. Miami University. While working on a story on campus crime for *The Miami Student*, Miami University's student newspaper, the editor-in-chief made a request to access student disciplinary records. The University denied the petition, and the newspaper in turn filed another request under the Ohio Public Records Act, Ohio Rev.Code § 149.43. In order to remain compliant with FERPA, the University turned over the records, but only after redacting the involved students' PII and specific details about the incidents that caused the creation of the disciplinary records. The student editors found the records to be excessively redacted and filed an original mandamus request with the Ohio Supreme Court. The court awarded the writ of mandamus arguing student conduct records did not fit the statutory definition of educational record as described in FERPA per the opinion in *Red & Black Publishing Co. v. Board of Regents of University System of Georgia,* 427 S.E.2d 257 (Georgia 1993). The University requested a review from the United States Supreme Court, but the Court did not grant certiorari.

United States v. Miami University. In response the the Ohio Supreme Court's opinion in *Miami Student v. Miami University*, *The Chronicle of Higher Education*, also filed an open records request with the University for non-redacted student conduct records. Fearful it would no longer be in compliance with FERPA, the University fulfilled *The Chronicle's* request, and the University informed the Dept. of Education of the situation. The Department argued the Ohio Supreme Court was incorrect in that student disciplinary records are part of a student's FERPA protected educational record.

However, Miami University did comply with *The Chronicle's* request per the University's policy to release student disciplinary records to a third-party, even without the students' consent. Ohio State, which also received such a request, followed suit and released student conduct records to *The Chronicle*, and did so without prior student consent. The Dept. of

Education filed suit against both Miami University and Ohio State in the federal district court to prevent further disclosure of the disciplinary records without prior student consent. The Department argued that disciplinary records are educational records protected by FERPA (*United States v. Miami University,* 91 F.Supp.2d 1132 (S.D. Ohio 2000). *The Chronicle* intervened in the case and filed a motion to dismiss. The federal district court denied *The Chronicle's* motion and granted the Dept. of Education's motion for summary judgement holding that disciplinary records are covered by FERPA. *The Chronicle* appealed the decision to the Court of Appeals for the 6th Circuit and upheld the federal district court's decision.

United States v. Miami University interprets FERPA as placing value on student privacy above that of the public's need to be made aware of specific information with very limited exceptions. The court, relying on the text of FERPA, found that the definition of "educational record" is quite broad. This case presents a challenge for some institutions when enforcing FERPA. Neither the statute or the court offered content-based descriptions of a student record. Rather, the court determined student conduct records constitute an educational record because they are records retained by an institution and directly relate to a student. *United States v. Miami* suggests a case-by case determination as to what might be considered necessary to protect under the law. The Circuit Court's opinion invalidates this fear by arguing Congress expressly made disclosure exemptions, thus institutions should be able to successfully decipher how and when to share the contents of a student record.

United States v. Miami provides some insight for MOOC providers as to what information they collect may be considered part of a student record and what may be shared with the public. Though MOOC providers do not resolve student conduct issues or collect such information, they do gather a great deal of information on their users' activity while they engage in their courses. An analogous relationship be made between student conduct records as they pertain to disciplinary issues and student behavior, and the information user conduct records may contain on what a user does while online. For example, a MOOC user that harassed a fellow MOOC user while online may necessitate the keeping of a conduct record by the MOOC provider. Additionally, the Circuit Court highlighted Congress's expressed conditions for of the release of student records, none of which Miami University or Ohio State would be able to meet if they complied with *The Chronicle's* information request. MOOC providers, per their business model of using user data the purposes of educational research, meet the educational research disclosure exemption standards as described in §99.31(b)(1) and (2).

What is PII and how is it Protected? The most significant portion of an educational record is the student's PII. The statutory definition states PII:

> Includes, but is 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, 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 educational record relates (§99.3).

Osborn v. Board of Regents of the University of Wisconsin System, 634 N.W. 2d 536 (Wisconsin 2002), and *Press-Citizen Company, Inc. v. University of Iowa*, 817 N.W.2d 480 (Iowa 2012) explain, from the perspective of two state supreme courts interpreting a federal statute (FERPA), what steps educational agencies or institutions must take to protect PII when releasing records to a third party or the public.

Osborn v. Board of Regents of the University of Wisconsin System. The Center for Equal Opportunity filed an open records request, but the Board of the Regents University of Wisconsin prohibited the Center from accessing the application materials of prospective students, including non-matriculating students from a wide cross-section of campuses in the system. The Wisconsin Court of Appeals recognized the University's FERPA argument, *Osborn v. Board of Regents of the University of Wisconsin System*, 634 N.W.2d 563 (Wisc. Ct. App. Dist. IV 2001), but the Center for Equal Opportunity prevailed on appeal to the Wisconsin Supreme Court.

The University agreed the open records request issued by the Center of Equal Opportunity could be released after redacting the students' PII, but argued that the burden of doing so would be cost-prohibitive and essentially generate a new record, violating the Wisconsin Open Records Law, Wis. Stat. § 19.35. The University also claimed the public interest of preserving the records per FERPA's regulations served a greater public interested than would the sharing of the information with the Center of Equal Opportunity. The Wisconsin Supreme Court rejected both arguments stating that the University erred in denying the open records inquiry since no PII was specifically requested, and that the University could seek financial compensation for the task of redacting the records.

Osborn highlights the significance of §99.31(b)(1) and (2), the FERPA regulation which determines when educational records may be shared:

> An 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 §99.30 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 (b)(1). An educational agency or institution, or a party that has received education records or information from education records under this part, may release de-identified student level data from education records for the purpose of education research by attaching a code to each record that may allow the recipient to match information received from the same source (§99.31(b)(1) and (2)).

The Wisconsin Supreme Court's holding is in keeping with §99.31(b)(1) and (2) in that permitted the release of PII-redacted educational records, but presents a contrary understanding to what the Family Policy Compliance Office, the unit within the Dept. of Education responsible for the interpretation and enforcement of FERPA, considers an educational record (The Catholic University of America Office of General Counsel, 2008) and how that information might be released. Moreover, the Court specified "access is limited only to disclosure of information that is not personally identifiable, [and] that an institution may release personally identifiable information contained in a record, but only upon written consent" (*Osborn v. Board of Regents of the University of Wisconsin System*, 2002, p. 23-24). For MOOC providers, *Osborn* may be persuasive in that potentially creates a standard for the release of all user information, especially when sharing user-populated datasets for the purposes of educational research.

Osborn's interpretation of FERPA's scope, that "once personally identifiable information is deleted, by definition, a record is no longer an education record since it is no longer directly related to a student" (*Osborn v. Board of Regents of the University of Wisconsin System*, 2002, p. 19) expectations that an educational record can be sufficiently redacted differs from those in *Press-Citizen Company, Inc. v. University of Iowa* (2012). *Press-Citizen* held that if even after the PII has been removed from an educational record, but record is still identifiable, the record still warrants protection. This demonstrates the tension between the levels of governance as it relates to privacy regulations and open information laws. Though FERPA, as interpreted by *Osborn,* permits the release of redacted educational records, the point at which a record is sufficiently redacted is still unclear. In addition to issues of federalism, MOOC providers, due to their global nature, may find it necessary to balance the privacy laws of international users and partner institutions.

Press-Citizen Company, Inc. v. University of Iowa. After the sexual assault of a female student by two campus athletes at the University of Iowa, the *Iowa City Press-Citizen* requested access to the records containing information on the incident under the Iowa Open Records Act, Iowa Code §22.2, .7, .9 (2007). The University partially complied with the requests, but declined to turn over documents containing PII as such information was protected by Iowa State Code §22.7(1)[2], and in a later motion to an Iowa district court claimed the ability to retain the records under FERPA. The district court denied the University's motion and required the release of both redacted and unredacted student records.

[2] Iowa State Code §22.7(1) requires student records and PII to remain confidential unless ordered to release such information by the court order or by a requesting legal guardian, or accredited educational institution for the purpose of obtaining the records of a transferring student.

On appeal to the Supreme Court of Iowa, the University claimed the federal statute superseded the authority of the state law. The court determined the Iowa Open Records Act already gave priority to FERPA, and held the University could not release either redacted or original documents without the students' consent per *United States v. Miami University,* 294 F.3d 797 (6th Cir. 2002). With the provision that releasing such information would jeopardize the University's federal funding, the court was compelled to uphold the University's FERPA claim. The Supreme Court of Iowa reversed the district court's judgment.

The holding in *Press-Citizen* allows institutions in Iowa to withhold student records in the instance in which the requester would be able to identify the students, even with the PII redacted. In the digital era, in which the data collected on students expands far beyond FERPA's definition of PII, *Press-Citizen* persuades institutions to take a proactive stance on protecting their student's privacy. This is also beneficial for MOOC providers and users. Whereas the argument may be made that volume of data collected by MOOCs reduces the risk of identifying a user with the sharing of redacted datasets, but the prevalence of social media makes *Press-Citizens* specifically relevant to MOOC providers. If users share about their MOOC learning experience on their personal blog or social media, the ability to recognize a specific user through quasi-identifiers in a published MOOC dataset increases (Daries et al., 2014). *Press-Citizens* does not establish a legal standard for institutions or MOOC providers to monitor such behavior, but it does point to a need to further develop PII that consider external factors such as a student's behavior on social media. However, in doing so, MOOC providers may have an especially high burden to ensure their users' PII is expressly preserved in a manner that prevents re-identification when sharing datasets for the purposes of research.

FERPA’s Application to MOOCs

The statutory definitions included in FEPRA as discussed in the relevant case law provide insights as to how the statute might be interpreted to include MOOCs. However, problems created by the circular nature of these definitions cannot be resolved solely through a legal analysis. An in-depth analysis as to how the relevant case law and statutory definitions might be applied to MOOCs is provided in Chapter 5.

Chapter 5

Results

The study provides results that aid in answering my research question regarding the ability of MOOC provider datasets to be de-identified to meet the requirements of FERPA and still maintain their utility for the purposes of research dissemination. The study also shows that determining a standardized process through which MOOC datasets may be de-identified is more challenging than I originally anticipated. My legal analysis reveals there may not be a direct requirement for MOOCs to comply with FERPA, but parallel relationships may be established between MOOCs and educational programs and MOOC users and students.

Results of De-identification Process

I ran Daries' de-identification program on the MITx 2.01x dataset six times, and once on the MITx 3.091x dataset. The research design calls for the program to be applied to each of the four datasets, but due to issues with Daries' code, I was unable to accomplish this task. Technically, the program partially operates correctly in that it does run the de-identification code on the datasets, however it consistently returns an empty utility, or null, matrix table. Without a completed utility matrix table, I am unable to assess the success or effectiveness of the de-identification process.

Iteration I, MITx 2.01x. My first attempt at running Daries' program on the MITx 2.01x dataset was unsuccessful. At the program's prompting to choose variables to render an initial *k*-anonymity reading, I chose the variables *viewed*, *explored*, *certified*, *gender*, *nevents*, *ndays_act*, *nplay_videos*, *nchapters*, *nforum_posts*. I picked these quasi-identifiable attributes as they are the variables measured by the utility matrix. In response to the next prompt, to choose variables in order to "checking [sic] k-anonymity for records with some null values" (Daries, 2014, n.p.). I selected *course_id*, *registered*, *countryLabel*, *LoE*, *YoB*, *roles*, *role_isStaff*, *npause_video*, and *email_domain*. These attributes were chosen as they were a mix of quasi-identifiers and variables that would not be reflected in the utility matrix.

I was then asked to select the tails to be trimmed of both the *nforum_posts* and *YoB* variables. The trimming of these tails generates the interquartile range on which the standard deviation, entropy, and mean are calculated for the utility matrix. It was at this prompting that I encountered my first error in the program. The program provided the option to trim the high, low, or both high and low tails. I elected to trim both tails, which yielded an error and terminated the program. I reran the program and selected to trim only the high tail for the variable *nforum_posts.* For the MITx 2.01x dataset, this tail was 20:4, meaning four users posted to the course forum 20 times. I overwrote the data to reflect this new tail in the utility calculations and repeated the process when prompted to trim the tails for the variable *YoB*.

Trimming the tails for *YoB* mirrored the process for trimming the tails for *nforum_posts.* When prompted to trim the high, low, or both high and low tails, I opted to trim both tails, which also returned the same error. I reran the program and selected to trim only the high tail. For the MITx 2.01x dataset, this tail was 1998:39, meaning 39 users reported being born in 1998. I overwrote the data to reflect this new tail in the utility calculations.

The program then asked me to select the variables for *k*-anonymous wrapping, the "step where non-k-anonymous records are removed" (Daries, 2014, n.p.). I selected the same variables from the initial *k*-anonymous prompting: *viewed*, *explored*, *certified*, *gender*, *nevents*, *ndays_act*, *nplay_videos*, *nchapters*, *nforum_posts*. The final prompt request was for me to choose the variables to be exported, accompanied by the warning "to be careful to only export the columns you are okay with others seeing" (Daries, 2014, n.p.). I elected to export the variables that would be measured in the utility matrix. Choosing these variables would have allowed me to directly assess their utility as demonstrated in the matrix, which in turn should have allowed me to assess the effectiveness of Daries' program as a means by which to de-identify MOOC datasets.

The program exported the variables to the secured PyCharm database file. However, the post-utility matrix, which is calculated based upon the values selected in the tail trimming and k-anonymous variable nomination process, returned a table with zeros in each cell. This null utility matrix indicates an error occurred in the variable selecting process or the in program itself.

Iterations II-IV, MITx 2.01x. In order to determine if the return of the null utility matrix was an error on my part or a bug in the program, I ran the code a total of six times on the MITx 2.01x dataset. My original plan was to run the code up to ten times, choosing different variables each for each iteration. However, after my sixth attempt at running the program and encountering the same errors during each iteration, regardless of variable selection, I determined that there would be no need to continue to attempt to run the program on the MITx 2.01x dataset.

The attempts to run Daries' code yielded the same results: an empty utility matrix. I repeated the process an additional five times, using a mix of variables, some chosen at random and others selected in order to replicate the variables used in Iteration I (see Table 5.1). Each attempt returned a null utility matrix. After the sixth iteration, I decided to run the program on a different dataset to establish if the problem with the utility matrix is related to Daries' code or the MITx 2.01x dataset.

Iteration I, MITx 3.091x. I ran program on the MITx 3.091x dataset, choosing the *k*-anonymous variables and tails for the *nforum_posts* and *YoB* variables that mirrored Iteration I of the MITx 2.01x dataset. The second iteration also returned a null utility matrix. This result indicated that my inability to run the de-identification is not due to the variable selection or the datasets, but the code itself. Upon this assessment, I determined it was not necessary to run the program on the two additional datasets. However, I did to attempt to resolve my inability to successfully run Daries' code by troubleshooting the program.

Troubleshooting the Program. After identifying the potential source of the problem in the replication process, the program itself, I contacted Daries to solicit his help in pinpointing and debugging the code. Daries declined my request for assistance. I also enlisted the support of an expert software engineer who was unable to locate the source of the error in the code. Even after reviewing the coding documentation in GitHub provided by Daries and another researcher attempting to expand upon Daries' et al.'s 2014 study, Harvard University's Jim Waldo, it was too difficult to determine where the program bug was located within the code. In order to complete the full de-identification process, the software engineer recommended rewriting the entire program. Upon this recommendation, I determined Daries study was not able to be replicated without significant revisions.

Table 5.1. Variables Selected when Running Daries De-identification Program on MITx 2.01x, Iterations I-III

Variables	Iteration I	Iteration II	Iteration III
K-Anon Variable1	*viewed*	*YoB*	*city*
K-Anon Variable2	*explored*	*is_active*	*nforum_votes*
K-Anon Variable3	*certified*	*nvideos*	*LoE*
K-Anon Variable4	*gender*	*cc_by_ip*	*nplay_videos*
K-Anon Variable5	*nevents*	*start_time*	*nproblem_check*
K-Anon Variable6	*ndays_act*	*last_event*	*cert_status*
K-Anon Variable7	*nplay_videos*	*nforum_threads*	*course_combos*
K-Anon Variable8	*nchapters*	*registered*	*viewed*
K-Anon Variable9	*nforum_posts*	*user_id*	*gender*
Check Null Variable1	*course_id*	*viewed*	*YoB*
Check Null Variable2	*registered*	*explored*	*is_active*
Check Null Variable3	*countryLabel*	*certified*	*nvideos*
Check Null Variable4	*LoE*	*gender*	*cc_by_ip*
Check Null Variable5	*YoB*	*nevents*	*start_time*
Check Null Variable6	*roles*	*ndays_act*	*last_event*
Check Null Variable7	*role_isStaff*	*nplay_videos*	*nforum_threads*
Check Null Variable8	*npause_video*	*nchapters*	*registered*
Check Null Variable9	*email_domain*	*nforum_posts*	*user_id*
Trim Tails *nforum_posts*	High, 20 count 4	Low, 21 count 1	High, 29 count 3
Trim Tails *YoB*	High, 1998 count 39	High, 1998 count 39	Low, 1894 count 2
K-Anon Wrap1	*viewed_NF*	*kCheckFlag*	*registered_NF*
K-Anon Wrap2	*explored_NF*	*entropy*	*YoB_NF*
K-Anon Wrap3	*certified_NF*	*uniqUserFlag*	*YoB_DI*
K-Anon Wrap4	*gender_NF*	*nchapters_NF*	*nforum_posts*
K-Anon Wrap5	*nevents_NF*	*viewed_NF*	*start_time_NF*
K-Anon Wrap6	*ndays_act_NF*	*nforum_posts_NF*	*LoE_NF*
K-Anon Wrap7	*nplay_videos_NF*	*sdv_dt*	*is_active_NF*
K-Anon Wrap8	*nchapters_NF*	*sum_dt*	*course_combo*
K-Anon Wrap9	*nforum_posts_NF*	*kkey*	*nvideo_NF*
Exported Variable1	*viewed_NF*	*kCheckFlag*	*viewed*
Exported Variable2	*explored_NF*	*entropy*	*explored*
Exported Variable3	*certified_NF*	*uniqUserFlag*	*certified*
Exported Variable4	*gender_NF*	*nchapters_NF*	*gender*
Exported Variable5	*nevents_NF*	*viewed_NF*	*nevents*
Exported Variable6	*ndays_act_NF*	*nforum_posts_NF*	*ndays_act*
Exported Variable7	*nplay_videos_NF*	*sdv_dt*	*nplay_videos*
Exported Variable8	*nchapters_NF*	*sum_dt*	*nchapters*
Exported Variable9	*nforum_posts_NF*	*kkey*	*nforum_posts*
Utility Matrix Output	NULL	NULL	NULL

Table 5.1, *continued*: Variables selected when running Daries de-identification program on MITx 2.01x, Iterations IV-VI

Variables	**Iteration IV**	**Iteration V**	**Iteration III**
K-Anon Variable1	*course_id*	*YoB*	*city*
K-Anon Variable2	*registered*	*is_active*	*nforum_votes*
K-Anon Variable3	*countryLabel*	*nvideos*	*LoE*
K-Anon Variable4	*LoE*	*cc_by_ip*	*nplay_videos*
K-Anon Variable5	*YoB*	*start_time*	*nproblem_check*
K-Anon Variable6	*roles*	*last_event*	*cert_status*
K-Anon Variable7	*role_isStaff*	*nforum_threads*	*course_combos*
K-Anon Variable8	*npause_video*	*registered*	*viewed*
K-Anon Variable9	*email_domain*	*user_id*	*gender*
Check Null Variable1	*ip*	*LoE*	*city*
Check Null Variable2	*email_domain*	*postalCode*	*nforum_votes*
Check Null Variable3	*ndays_act*	*start_time*	*LoE*
Check Null Variable4	*nchapters*	*last_event*	*nplay_videos*
Check Null Variable5	*certified*	*explored*	*nproblem_check*
Check Null Variable6	*nforum_comments*	*nevetns*	*cert_status*
Check Null Variable7	*explored*	*ndays_act*	*course_combos*
Check Null Variable8	*nseek_video*	*nforum_endorsed*	*viewed*
Check Null Variable9	*grade*	*roles*	*gender*
Trim Tails *nforum_posts*	Low, 21 count 1	High, 14 count 4	High, 20 count 4
Trim Tails *YoB*	Low, 2001 count 3	High,1964 count 47	High, 1998 count 39
K-Anon Wrap1	*email_domain_NF*	*YoB*	*city*
K-Anon Wrap2	*nprogcheck_NF*	*is_active*	*nforum_votes*
K-Anon Wrap3	*LoE_DI*	*nvideos*	*LoE*
K-Anon Wrap4	*nforum_posts_DI_avg*	*cc_by_ip*	*nplay_videos*
K-Anon Wrap5	*grade_NF*	*start_time*	*nproblem_check*
K-Anon Wrap6	*nforum_posts_DI*	*last_events*	*cert_status*
K-Anon Wrap7	*nchapters_NF*	*nforum_threads*	*course_combos*
K-Anon Wrap8	*nforum_posts_NF*	*registered*	*viewed*
K-Anon Wrap9	*nseek_video_NF*	*user_id*	*gender*
Exported Variable1	*email_domain_NF*	*LoE*	*viewed_NF*
Exported Variable2	*nprogcheck_NF*	*postalCode*	*explored_NF*
Exported Variable3	*LoE_DI*	*start_time*	*certified_NF*
Exported Variable4	*nforum_posts_DI_avg*	*last_event*	*gender_NF*
Exported Variable5	*grade_NF*	*explored*	*nevents_NF*
Exported Variable6	*nforum_posts_DI*	*nevetns*	*ndays_act_NF*
Exported Variable7	*nchapters_NF*	*ndays_act*	*nplay_videos_NF*
Exported Variable8	*nforum_posts_NF*	*nforum_endorsed*	*nchapters_NF*
Exported Variable9	*nseek_video_NF*	*roles*	*nforum_posts_NF*
Utility Matrix Output	NULL	NULL	NULL

Assessing Replicability

Though the original purpose for attempting the de-identification process is to determine the replication potential of the method in order to establish a standardized process for anonymizing MOOC datasets, the complications with Daries' de-identification program reveals a number of insights regarding this research goal. Since MOOC providers do not have a consistent business model or research goals, it stands to reason the data they collect on their users is not all the same. Moreover, this research goal assumes that MOOC providers are willing to or are actively turning over their data for educational purposes or research. If it is determined that MOOC providers are required to comply with FERPA, a universal de-identification model or program may not be effective regulatory solution for all MOOC platforms.

Effectiveness of Daries' De-identification Program. The two limitations of this study, my inability to obtain datasets from a Coursera institutional partner and my failure to execute a successful run of Daries' program, actually demonstrate the limitation of my research goal. Per his original research design, Daries program is specifically written to work on datasets from the edX platform. His program assumes which MOOC user data is collected and retained, as well as what PII or quasi-identifiers might be released. Running the program on a different provider's dataset requires rewriting the Python code as the value names of the variables or attributes will be different. Thus, the edX custom tailoring of Daries' program lacks the universalness necessary to be established as standard methodology for de-identifying all MOOC datasets. I must note, however, that I was not able to obtain Coursera dataset, and therefore I cannot test the universality of Daries' de-identification program on a non-edX provider's dataset. Nonetheless, the theoretical framework of Daries' program, *k*-anonymity, *l*-diversity and the utility matrix, may be a sufficient model for creating a standardized process for de-identifying datasets. However, shifting the focus from standardizing the de-identification process to standardizing the outcomes of the de-identification process may prove to be more useful when determining which variables require redaction.

Role of Terms of Service Agreements and Privacy Policies on Data Releases for De-identification. In my attempt to gain access to MOOC provider datasets, I examined the terms of service agreements and privacy policies for both edX and Coursera. My review of these documents indicates the collection and release of user data is a more complicated process than originally anticipated and focuses primarily on the protection of user data rather than the providers' commercial use of the data.

edX. edX's terms of service agreement includes the expectations for users' online behavior, warranties and limitations of liability, indemnification policy, and the honor code. It also specifies the information users must provide in order register for a course or to be awarded a certificate for a verified course. When opening a user account in order to register for a course, users must provide their name, email and user password. For the purposes of user authentication for paid verified courses, edX users must submit a photo from a valid government or state ID and provide a current webcam headshot of themselves. edX's privacy policy states that this information, as well as other personal information defined as contact information, birthdate, employment, and gender, is protected but that this data may be shared with a third party for fourteen distinct purposes including processing payments, monitoring user participation, and educational or scientific research. edX also provides eight exemptions to their privacy policy for the sharing of user data, including personal information, to partner institutions. These exemptions include the development of individual user's educational goals, responding to subpoenas, and institutional research requests.

Nonetheless, edX's terms of service agreement states that users retain the rights to the content they publish to an edX's course's discussion board. However, the agreement also states that upon publishing such postings, users agree to:

> Grant to edX a worldwide, non-exclusive, transferable, assignable, sub licensable, fully paid-up, royalty free, perpetual, irrevocable right and license to host, transfer, display, perform, reproduce, modify, distribute, re-distribute, relicense and otherwise use, make available and exploit your User Postings, in whole or in part, in any form and in any media formats and through any media channels (now known or hereafter developed) [sic]. (edX, 2014, n.p.)

Though this language is fairly standard for the terms of service agreement, as seen in the service agreements for providers like Facebook, Apple, and Google (Bradshaw, Millard, Walden, 2011), it appears to be contradictory to the spirit of edX's privacy policy. This statement, juxtaposed with edX's privacy policy which states the personal information collected by the provider and the user's educational record is protected by FERPA, is even more perplexing. The terms of service agreement clarifies that users are not enrolled by proxy at an edX partner institution or entitled to the student benefits of those institutions. Still, the exemptions provided in the privacy policy directly align with the exceptions provided in C.F.R. Title 34 Part 99 Subpart D §99.31. Therefore, it may be reasonable to conclude that edX recognizes itself as an educational agency and that its users are students of the platform.

Coursera. The terms of use provided by Coursera are similar to that of edX in that it clarifies the guidelines for user conduct, includes disclaimers for liability and indemnification, and establishes enrollment in a Coursera course does not constitute a relationship between the user and the partner institution. It also states Coursera may use or share user content at its discretion. Coursera’s terms of use specifically addresses the issue of educational research in the statement, “records of [user] participation in courses may be used for education research. Research findings will typically be reported at the aggregate level. [User] personal identity will not be publically disclosed in any research findings without [user] express consent” (Coursera, 2015, n.p.).

Coursera’s privacy policy defines two types of collected data, non-personal and personal information. Non-personal information is collected through cookies and includes user metadata such as the number of Coursera page visits and the duration of time spent on those pages, IP addresses, browser software, and operating system information. Coursera’s classification of personally identifiable information, which is collected for the purpose of registration and identify verification, is similar to that of edX and includes a user’s name, address and birthday. For paid verified courses, users must also submit a photo, typing samples, and income level for those applying for financial aid. Coursera also discloses user information is stored on servers housed in the United States and therefore international users’ data is subject to United States law and may not be regulated or protected by their home countries’ laws.

Under Coursera’s safe harbor policy, seven principles guide the collection and release of users’ personal information: notice, choice, onward transfers (to third parties), data security, data integrity, access, and enforcement. As it pertains to the release user data, the principle of notice states Coursera (2015) “will provide [users] with timely and appropriate notice in [the] Terms of Service, describing what Personal Information [Coursera is] collecting, how [Coursera] will use that information, and the types of third parties with whom [Coursera] shares such information” (n.p.). The notice principle and the Coursera’s requirement per the terms of use to obtain user consent prior to data sharing was specifically referenced when I unsuccessfully solicited

Coursera institutional partners for access to their datasets. Additionally, through my correspondence with a representative from a Coursera partner institution, I learned the MOOC provider does not release identified data without a justifiable rationale from the requesting institutional partner (T. Karr, personal communication, February 11, 2016). For example, Coursera may release demographic information for a partnering institution's course per that institution's request, but it does not share user data on an individual record basis.

Protecting and Releasing User Data. The terms of service agreements and privacy policies of edX and Coursera reveal that the providers assume the significant burden of protecting user privacy, but have different positions on the applicability of FERPA to their datasets. edX, a non-profit provider, publishes a privacy policy stating the provider complies with FERPA (edX, 2014). Coursera, a for-profit company, does not expressly state that it considered itself subject to FERPA (Pierson, Terrell, & Wessel, 2013), but its principles and protocol to refrain from releasing identifiable datasets, even to partner institutions, indicates Coursera recognizes the need to protect users' information and privacy. Moreover, Coursera's (2015) use of the term personally identifiable information and the reference to users "apply[ing] for financial aid in connection to these services" in the privacy policy indicates an anticipation of FERPA regulation.

Results of Legal Analysis

The results of my document review of FERPA and the applicable case law as it relates to the statutory definitions of the terms student, attendance, educational agency or institution, educational record, and PII do not inform the legal question of FERPA's applicability to MOOCs as originally anticipated. Whereas a compelling argument may be made regarding the analogous relationship between MOOC providers, users, and datasets and FERPA's recognition of and what constitutes an educational agency, student, and an educational record which includes PII, the case law is insufficient for determining a legal requirement for MOOC to comply with FERPA. Moreover, while some case law may be persuasive in the argument for FERPA to be applicable to MOOCs, the ultimate authorities in determining FERPA compliance rest with Congress, the Dept. of Education, or the United State Supreme Court.

Are MOOC Users Students? FERPA states a student is “any individual who is or has been in attendance at an educational agency or institution and regarding whom the agency or institution maintains educational record” (C.F.R. Title 34 Part 99 Subpart A §99.3). This term is inclusive of students who attend classes in person, via correspondence, and online and does not differentiate between matriculating and non-matriculating students (C.F.R. Title 34 Part 99 Subpart A §99.3). *Klein Independent School District v. Mattox* (1987) and *Tarka v. Franklin* (1990) demonstrate that the statutory definition of student may be flexible enough to be inclusive of MOOC users.

A preliminary reading of §99.3 indicates MOOC users do not fit precisely within the definition of student, but it also does not explicitly exclude MOOC users. In fact, those users who take MOOC classes for the purpose of matriculating to a residential postsecondary program, such as MIT’s MicroMasters and the Global Freshman Academy, may already fall within the definition of student as set forth in FERPA. For users who are enrolled in a MOOC course without the goal of matriculation, on the other hand, the case law does not provide clarity regarding their student status according FERPA.

Both edX and Coursera explicitly state a user’s enrollment in a course does not constitute a relationship with the affiliated institution. This statement is consistent with the reasoning of the U.S. Court of Appeals for the Fifth Circuit in *Klein Independent School District v. Mattox*

(1987); having a relationship with an educational agency does not equate to that agency having an obligation to provide FERPA protection. The 5th Circuit court in *Tarka v. Franklin* (1990) further interprets the meaning of the term student as set forth in FERPA. That court held that those who are not admitted to an institution are not students and therefore do not have FERPA rights, but course auditors may garner some degree of FERPA protection per the *Joint Statement in Explanation of Buckley/Pell Amendment* (120 Congressional Record, 39862-39866). Though the Joint Statement is not binding, it provides a foundation for the argument that MOOC users, who essentially audit courses online, may qualify for some degree of FERPA protection from their MOOC provider.

MOOC providers are the medium through which a MOOC user attends a course. Though MOOC users access their courses through an online portal as opposed to a physical classroom, they do engage with the material in the same manner as class auditors. Both parties read the course materials as outlined in the course syllabus, participate in class discussions, and complete course assignments. Furthermore, just as traditional academic agency collects and retains information on class auditors, so do MOOC providers. If auditors are eligible to seek protections under FERPA similar to that of student, then perhaps MOOC users should be afforded the same benefit.

Does Enrolling in a MOOC Constitute Attendance? The definition of student, as outlined in FERPA, relies on the prerequisite of attendance. The statute's definition of attendance includes the ways in which an individual may attend an educational agency, such as "in person or by paper correspondence, videoconference, satellite, Internet, or other electronic information and telecommunications technologies for students who are not physically present in the classroom" (§99.3), but does not specify when attendance begins. Case law on the issue of attendance as it pertains to FERPA is limited, and is often discussed in conjunction with determining student status. Moreover, the Joint Statement does not provide clarity as to how attendance might be defined in the context of MOOCs.

Therefore, a plain reading of the statute indicates that MOOCs users do “attend” courses on a MOOC provider’s platform. However, this assessment does not equate to a requirement of FERPA compliance. As previously mentioned, in order to garner FERPA protection MOOC users must qualify as students and MOOC providers must be classified as an educational agency or institution. Determining that a MOOC user’s enrollment in MOOC courses does meet the statutory definition of attendance partially resolves the research question, and requires further evaluation of the term educational agency or institution.

Are MOOC Providers Educational Institutions or Agencies? FERPA defines an educational agency or institution as a private or public school that receives federal funds under the authority of the Secretary of the U.S. Department of Education (C.F.R. Title 34 Part 99 Subpart A §99.1, §99.3). At the time of my study, no MOOC provider met FERPA’s definition of an educational agency or institution, as they did not receive federal funds, even in the form of federal student aid. However, in October 2015 the Dept. of Education announced the Educational Quality through Innovative Partnerships (EQUIP) pilot program that enables MOOC users to apply for federal financial aid to cover the expense of their verified certificate-granting or credentialing courses (U.S. Department of Education, 2015). This initiative, coupled with partnerships such as edX and ASU’s Global Freshman Academy and MIT’s MicroMasters program, may change the status of MOOC providers into educational agencies subject to FERPA.

Are MOOC Datasets Classified as Educational Records and do they Include PII?

An educational record, as defined by FERPA, includes "records that are directly related to the student; and maintained by an educational agency or institution or by a party acting for the agency or institution" (§99.3). In that determining if MOOC providers' datasets meet the statutory definition of educational record also relies upon the definition of student and educational agency, the decision of the U.S. Supreme Court in *Owasso Independent School District* v. *Falvo* (2002) may inform the applicability of the term MOOC metadata.

Is Metadata an Educational Record? *Owasso Independent School District* v. *Falvo* (2002) raises questions, for the purposes of this analysis, regarding the parallel relationship between traditional educational records and MOOC datasets. The Court determined a teacher's gradebook, which includes any number of individual data points about a student's academic performance such as class participation and peer-graded papers, does not constitute a student record as defined by FERPA. Therefore, it stands to reason that a MOOC user's contribution on a course's discussion board or submission of a homework assignment would also not meet FERPA's legal standard for educational record.

However, Justice Kennedy's single central custodian theory, as articulated in *Owasso*, demonstrates how MOOC users may not benefit from a limited interpretation of educational record. MOOC providers have an exceptional nature to collect and indefinitely retain educational records, as opposed to traditional educational agencies, have multiple contact points with students that must be collated to generate an educational record. For example, a teacher's gradebook does not serve as a custodian for educational records, but a student's final grade on a transcript, which is reflective of the data housed in the gradebook, is submitted to the registrar, a single central custodian. MOOC providers have one access point to user information, and under Kennedy's theory, makes the provider a single central custodian by default.

Furthermore, MOOC datasets are arguably the epitome of *Falvo's* single central custodian concept, especially since the Court's opinion, while predating MOOCs, recognized the use of an electronic filing system, which in for MOOCs has become the exclusive means of data tracking. The case emphasizes that the retention and sharing of the records is just as important as the records themselves. For MOOCs, controlling access to and auditing these records could be as simple as making a few changes in the software's line of code. The built-in record maintenance system diminishes the burden traditional educational institutions may endure in order to meet the single central custodian standard, thus making FERPA compliance for MOOCs are more reasonably obtainable goal.

Still, the type of data collected by MOOC providers does not easily correlate to the information collected by a traditional educational agency. A student's college transcript may include completed courses and GPAs, but it does not provide information on how many times they attended those class or how long they stayed in the physical classroom. MOOC providers collect significantly more detailed information on their users, surpassing what a traditional educational agency is able to obtain, or of what the Joint Statement could even conceive. Even though a compelling argument may be made that MOOC datasets function as an educational records and therefore should be classified accordingly for the purposes of FERPA, and based upon the persuasiveness of Kennedy's theory as articulated in *Owasso*, a definitive conclusion cannot be reached without recommendations from the Dept. of Education or action by Congress.

Is Metadata PII? Moreover, once a decision is made regarding classification of MOOC datasets as they pertain to the definition of educational record, in order to determine which data should be removed from the datasets to yield FERPA compliance, the definition for PII as it pertains to the regulation promulgated in C.F.R. Title 34 Part 99 Subpart D §99.31(b)(1) and (2) requires evaluation. In addition to a student's name, address, social security number, student ID, and birth information, FERPA considers "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" (§99.3). Therefore, the data collected and concatenated by MOOC providers into a dataset for the purposes of research could be considered PII.

This definition is emphasized by the Iowa Supreme Court in *Press-Citizen Company, Inc. v. University of Iowa* (2012), which affirms the discretion of institutions to deny the release of a de-identified student's educational record if the totality of the information included in that record can easily lead to a re-identification. The court in *Press-Citizen* indicates that what is traditionally considered PII, such as a student's name, contact information, and parent's names, is not exhaustive. Rather, it reveals that any type of information, when examined in context, may be reasonably used to identify an individual. This distinction is important for MOOC providers as it reinforces the importance of ensuring that MOOC provider datasets are properly protected in order to safeguard MOOC users' privacy. However, it also indicates that the volume of data collected on MOOC users far surpasses that which is collected by traditional institutions, which may make FERPA compliance especially burdensome for MOOC providers.

Chapter 6

Recommendations and Conclusions

Despite the technical issues with the implementation of Daries' program, I conclude that MOOC datasets can be de-identified within the regulatory structure of FERPA. However, since I was not able to measure the utility of the de-identified dataset via the utility matrix, I am unable to determine if the de-identification yields a redacted dataset that may still be useful for research purposes. Additionally, the results of my study indicate the research goal of determining a standardized process for de-identifying MOOC datasets may not be feasible or even desirable. Instead, this goal should be realigned to establish standardized outcomes rather than processes. Future research, coupled with potential policy actions from the Dept. of Education and Congress, should guide the formation of these standards.

Conclusion

My study shows MOOC datasets can be de-identified to satisfy C.F.R. Title 34 Part 99 Subpart D §99.31 (b)(1) and (2) or any other regulation determined by the Dept. of Education or Congress. The results of the de-identification process indicate that though Daries' specific code was broken, the ability for MOOC datasets to be de-identified in accordance with FERPA is possible, but does require more testing to determine reliability. Reliability and replication is essential if this, or any, de-identification model is to be implemented as a policy solution. Still, a legal requirement to de-identify these datasets is contingent upon Congressional action to either amend FERPA to include of MOOCs or to draft new legislation that specifically addresses privacy in this specific context. MOOCs can, and should, ethically determine standards to address these privacy concerns in the absence of changes to FERPA.

The need to de-identify MOOC datasets prior to distribution or research currently exists as an ethical concern, not a legal requirement. As my legal analysis shows, MOOC providers are not currently under any obligation to provide FERPA protections to their users beyond their own terms of service agreements and privacy policies. MOOC providers do not currently receive federal aid, thus they are not legally required to comply with FERPA. Therefore, based upon the ethical values reflected in FERPA, I conclude that MOOCs should voluntarily comply with FERPA until further guidance is provided by Congress or the Dept. of Education.

Still, MOOCs have not yet been categorized by the Dept. of Education as an educational agency or institutions, thus even the application of federal student aid to cover the cost of a MOOC course may not be enough to require FERPA compliance. However, considering the rise of hybrid MOOC programs, including the Global Freshman Academy and MIT's MicroMaster's degree, as well as the Dept. of Education's EQUIP initiative which allows the use of federal student aid to cover the cost of MOOCs, inaction on the part of Congress of the Dept. of Education is less likely. As more traditional post-secondary educational agencies employ the use of MOOCs to achieve their educational and business goals, and if individuals are now able to use their federal aid to pay for MOOC certificates or courses that will be used to pursue a degree, Congress and the Dept. of Education must to act to make MOOC operations and current legislation align. Therefore, the question will soon change from can MOOCs comply with FERPA, to when will MOOC providers be required to comply with FERPA or other privacy regulations.

Moreover, my results indicate the research goal of determining a standardized process for de-identifying MOOC datasets should, instead, work to determine required outcomes as opposed to a specific processes. The de-identification of MOOC datasets should be a descriptive process, not a prescriptive one. As research needs change, so must the variables that are chosen to be de-identified. For example, a study examining the persistence of MOOC users are it correlates to age, de-identifying a MOOC dataset according to Daries' process immediately renders that dataset useless to the researchers as it redacts users' ages. In order to ensure privacy and utility, de-identification standards should be flexible or on a sliding scale. For as in the previous example, another quasi-identifier, such as gender or course name may be de-identified as a substitute for age. Therefore, and regulations governing the de-identification process should be focus on standardizing outcomes based upon the research question or intended goals of the study, not the process or specific quasi-identifiers.

Moreover, in order to meet the privacy needs of users and the business needs of providers, the redaction procedure should reflect the type and volume of data collected by MOOC providers. As demonstrated by this study's data collection and de-identification processes, too much variance exist between MOOC providers, therefore having a mandated or standardized de-identification process runs the risk of both overextending and underperforming privacy expectations. Considering the differences in data collected by MOOC providers, a standardized de-identification process may not accurately capture the necessary or correct information, missing some key data points while requiring other data that are extraneous.

Thus, shifting the focus of potential policy solutions from the means through which MOOC datasets are de-identified to the types of information that should be protected or redacted will lead to more nimble regulations, allowing for flexibility among MOOC providers, which results in better privacy outcomes for MOOC users and better data for educational researchers. These specific outcome goals should be developed by the Dept. of Education, but may be generated through partnerships with other administrative agencies and in collaboration with

MOOC providers and trade organizations. These outcomes may also be informed by the EU's standards.

Recommendations

Based upon the findings and conclusions of my study, I offer the following recommendations for future researchers and key stakeholders when attempting to resolve the question of imposing FERPA compliance for MOOC providers and their datasets:

For the Department of Education. To date, the Dept. of Education has yet to issue an official statement regarding its stance on the status of MOOCs. As MOOCs continue to persist and grow in both domestic and international markets, the Department will need devote time and attention to determine whether and how it will recognize or classify MOOCs. How it chooses to classify MOOCs will require further guidance from the agency about privacy concerns for user records, and may have trickle-down impacts on issues as accreditation, Title IX, and the Clery Act. This guidance, published via a Dear Colleague Letter, the Dept. of Education's standard method of communication with impacted parties, should include best practices for protecting user privacy, standardized outcomes for de-identified records that are made public for the purposes of research or otherwise, and the mechanisms through which these standards will be enforced.

For Congress. Though this research focuses on imposing FERPA's regulations on MOOC providers, the results of the study indicate this piece of legislation may not be well suited for digital learning platforms. Recent FERPA amendments indicate a trend towards acknowledging online learning environments, but it does not fully contemplate a global, online educational program that partners with traditional educational agencies, does not offer credit for its courses, but is experimenting with hybrid models in which matriculating students may substitute its courses for transferable credit at their home institution. Amending FERPA once more to accommodate MOOCs may make an already complex piece of legislation even more difficult to interpret and administer.

Therefore, Congress should consider drafting new legislation that addresses the unique regulatory challenges posed by disruptive educational technologies such as MOOCs and coding boot camps. Several bipartisan student privacy bills, primarily focused on K-12 education, have been proposed in both the House and the Senate, indicating members of Congress recognize the problem of student digital privacy and are already taking steps to address the issue. Though these bill have yet to pass in Congress, expanding upon or using one of these proposals as a framework for MOOCs may make regulatory compliance for MOOC providers both more feasible and swifter than if required to draft original legislation.

For Researchers. My study attempted to examine a redaction procedure on the datasets from one MOOC provider, as informed by the *k*-anonymity and *l*-diversity theories. My use of this procedure, edX datasets, and these theories was determined by my choice to replicate Daries research, and is not reflective of an assessment of the superiority of de-identification process based upon the *k*-anonymity and *l*-diversity theories.

Therefore, future researchers should explore other digital privacy theories and experiment with various de-identification processes on MOOC providers other than edX. Further research will lead to efficient redaction methods which both the Dept. of Education and MOOC providers can use to determine best practices for the de-identification of MOOC datasets to be compliant with FERPA or other relevant legislation.

For MOOC Providers. Even though MOOC providers are not currently legally required to comply with FERPA, the ethics of digital privacy as presented in Solove’s taxonomy and their terms of service agreements as demonstrated by edX and Coursera may compel MOOC providers to follow this law. Moreover, as MOOC providers continue to expand partnerships with institutions to create hybrid education models, and with the Dept. of Education's initiative to increase federal aid coverage to non-traditional educational platforms, MOOC providers should prepare for some form of regulatory oversight. Following the example of edX, which is currently experimenting with de-identification methods, MOOC providers should partner with researchers and trade organizations to begin to develop their own de-identification processes and best practices. Doing so may ease the inevitable transition from non-compliance to mandatory regulatory compliance. Ultimately, MOOCs may soon be subject to some form of federal rules and therefore should plan accordingly.

References

120 Cong. Rec. 14580 (1974).

120 Cong. Rec. 39862-39866 (1974).

Allen, I. E., & Seaman, J. (2014). Grade change: Tracking online education in the United States. Retrieved from http://www.onlinelearningsurvey.com/reports/gradechange.pdf

American Council on Education. (2012, November 13). ACE to assess potential of MOOCs, evaluate courses for credit-worthiness. Retrieved from http://www.acenet.edu/news-room/Pages/ACE-to-Assess-Potential-of-MOOCs-Evaluate-Courses-for-Credit-Worthiness.aspx

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662-279. doi: 10.1080/1369118X.2012.678878

Bradshaw, S., Millard, C., & Walden, I. (2011). Contracts for clouds: Comparison and analysis of the terms and conditions of cloud computing services. *International Journal of Law and Technology, 19*(3), 187-223. doi: 10.1093/ijlit/ear005

Cort v. Ash, 422 U.S. 66 (1975).

Coursera. (2015, April 3). Terms of Use. Retrieved from https://www.coursera.org/about/terms.

Daggett, L. M. (2008). FERPA in the twenty-first century: Failure to effectively regulate privacy for all students. *Catholic University Law Review, 85*(1), 59-114.

Daries, J. (2014, May 27). Person-course de-identification process. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147
Dixon v. Alabama, 294 F. 2d 150 (5th Cir., 1961).

Dwork, C. (2008). Differential privacy: A survey of results. In M. Agarwal, D. Du, D. Duan, & A. Li (Eds.), *Theory and applications of models of computation*, (pp. 1-19) Springer.

Dwork, C. (2006). Differential privacy. *Automata, languages and programming*, 1-12. Springer Berlin Heidelberg.

edX (2014, October 22). Privacy Policy. Retrieved from https://www.edx.org/edx-privacy-policy

Emam, K. E., & Dankar, F. K. (2008). Protecting privacy using k-anonymity. *Journal of the American Informatics Association*, 15(5), 627-637.

Fain, P. (2013, January 16). As California goes? *Inside Higher Ed.* Retrieved from https://www.insidehighered.com/news/2013/01/16/california-looks-moocs-online-push

Family Educational Rights and Privacy Act: Rules and Regulations, 73 Fed. Reg. 237 (Dec. 9, 2008) (to be codified at 34 C.F.R. pts. 99.5, 99.31, 99.31(a)(1), 99.31(a)(2), 99.331(a)(6), 99.35, & 99.36)

Family Policy Compliance Office. (2015). FERPA for school officials. Retrieved from http://familypolicy.ed.gov/ferpa-school-officials

Ferenstein, G. (2014, March 3). Study: Massive online courses enroll an average of 43,000 students, 10% complete. *TechCrunch.* Retrieved from http://techcrunch.com/2104/03/03/study-massive-online-courses-enroll-an-average-of-43000-students-10-completion/

Fournier, H., Kop, R., & Durand, G., (2014). Challenges to research in MOOCs. *Journal of Online Learning and Teaching, 10*(1), 1-15.

Franken, A. (2016, January 13). Sen. Franken presses Google on student data privacy concerns. Retrieved from http://www.franken.senate.gov/?p=press_release&id=3352

Friedman, D. (2014). The MOOC revolution that wasn't. *TechCrunch.* Retrieved from http://techcrunch.com/2014/09/11/the-mooc-revolution-that-wasnt/

Gardner, J. (2008). HIDE: An integrated system for health information de-identification, *21st IEEE International Symposium on Computer-Based Medical Systems*, 254-259. doi: 10.1109/CBMS.2008.129

Graham, R., Hall, R., & Gilmer, W.G. (2008). Connecting the dots. . . : Information sharing by post-secondary educational institutions under the family education rights and privacy act (FERPA). *Education & the Law, 20*(4). 301-316. doi: 10.1080/09539960903450548

Gonzaga University v. Doe, 536, U.S. 273 (2002)

Hazlett, C. (2014, January 21). Harvard and MIT release working papers on open online courses. Retrieved from: http://blog.edx.org/harvard-mit-release-working-papers-open/

Hollands, F. M., & Tirthali, D. (2014). *MOOCs: Expectations and reality*. Retrieved from http://files.eric.ed.gov/fulltext/ED547237.pdf

Hoser, B., & Nitschke, T. (2010). Questions on ethics for research in the virtually connected world. *Social Networks, 32*, 180-186.

Hughes, M., Ventura, S., & Dando, M. (2007). Assessing social presence in online discussion groups: A replication study. *Innovations in Education and Technology International, 44*(1), 17-29.

Jones, M. L. & Regner, L. (2015, August 19). Users or students? Privacy in university MOOCs. *Science and Engineering Ethics.* doi: 10.1007/s11948-015-9692-7

Kaplin, W.A., & Lee, B.A. (2007). *The law of higher education* (4th ed.). San Francisco, CA: John Wiley & Sons, Inc.

Klein Independent School District v. Mattox, 830 F.2d 576 (5th Cir. 1987).

Kolowich, S. (2014, December 3). Are MOOC-takers ‘students’? Not when it comes to the Feds protecting their data. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/j-MOOC-Takers-Students-/150325

L'Orange, H.P., Blegen, J., & Garcia, T.I. (2011). Improving student attainment requires more from higher education data. *Educause Review, 46*(5), 62-63.

Lomonte, F.D. (2010). Ferpa frustrations: It's time for reform. *Chronicle of Higher Education, 56*(35), A56-A56.

Lessig, L. (1999). *Code: And Other Laws of Cyberspace*, New York, NY: Basic Books.

Lewin, T. (2013, December 10). After setbacks, online courses are rethought. *The New York Times.* Retrieved from http://www.nytimes.com/2013/12/11/us/after-setbacks-online-courses-are-rethought.html

Machanavajjhala, A. Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data, 1*(1), 1-53. doi: 10.1145/1217200.1217302.

Markey, E. (2014). Markey to introduce legislation to protect student privacy: Press release. Retrieved from http://www.markey.senate.gov/news/press-releases/markey-to-introduce-legislation-to-protect-student-privacy

Melear, K. B. (2003). From in loco parentis to consumerism: A legal analysis of the contractual relationship between institution and student. *NASPA Journal, 40*(4), 124-148.

Meloy, A. (2012). Legal watch: Crisis on campus: What you need to know for compliance. *Presidency*, 1-3.

National Association of College and University Attorneys. (2013, July 25). *MOOCs: The key legal and policy issues for colleges and universities.* [PowerPoint slides]. Retrieved from https://net.educause.edu/ir/library/pdf/CSD6233.pdf

Norwood v. Slammons, 788 F.Supp. 1020 (W.D. Ark. 1991)

Office of the Press Secretary. (2013, August 22)/ FACT SHEET on the President’s plan to make college more affordable: A better bargain for the middle class. Retrieved from https://www.whitehouse.gov/the-press-office/2013/08/22/fact-sheet-president-s-plan-make-college-more-affordable-better-bargain-

Osborn v. Board of Regents of the University of Wisconsin System, 247 Wis. 2d 957, 634 N.W. 2d 536 (Wisconsin 2002)

Owasso Independent School District No. I-011 v. Falvo, 534 U.S. 426 (2002)

Pappano, L. (2012, November 2). The year of the MOOC. *The New York Times*. Retrieved from http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html?pagewanted=all&_r=1

Pierson, M. W., Terrell, R. R., & Wessel, M. F. (2013). *Massive open online courses (MOOCs): Intellectual property and related issues.* Retrieved from http://www.higheredcompliance.org/resources/publications/AC2013_5G_MOOCsPartI1.pdf

Pope, J. (2014, December 15). What are MOOCs good for? *MIT Technology Review.* Retrieved from http://www.technologyreview.com/review/533406/what-are-moocs-good-for/

Posner, R. A. (1978). The right of privacy. *Georgia Law Review, 12*(3), 393-422.

President's Council of Advisors on Science and Technology. 2014). Big data and privacy: A technological perspective. Retrieved from https://www.whitehouse.gov/sites/default/files microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

Press-Citizen Company, Inc. v. University of Iowa, 817 N.W.2d 480 (Iowa 2012)

Ramirez, C. A. (2009). *FERPA clear and simple: The college professionals guide to compliance.* San Francisco, CA: Jossey-Bass.

Red & Black Publishing Co. v. Board of Regents of University System of Georgia, 262 Ga. 848, 427 S.E.2d 257 (Georgia, 1993)

Rivard, R. (2013, July 18). Udacity project on pause. *Inside Higher Ed.* Retrieved from https://www.insidehighered.com/news/2013/07/18/citing-disappointing-student-outcomes-san-jose-state-pauses-work-udacity

State ex rel. Miami Student v. Miami University*,* 680 N.E.2d 956 (Ohio 1997)

Shah, D. (2014, December 26). MOOCs in 2014: Breaking down the numbers. Retrieved from https://www.edsurge.com/news/2014-12-26-moocs-in-2014-breaking-down-the-numbers

Shah, D. (2014, October 15). How does Coursera make money? *EdSurge*. Retrieved from https://www.edsurge.com/n/2014-10-15-how-dnnoes-coursera-make-money

Smithers, R. (2011, May 11). Terms and conditions: not reading the small print can mean big problems. The Guardian. Retrieved from http://www.theguardian.com/money/2011/may /11/terms-conditions-small-print-big-problems

Solove, D. (2008). *Understanding Privacy.* Cambridge, MA: Harvard University Press.

Solove, D. (2011). Why privacy matters even if you have 'nothing to hide'. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Why-Privacy-Matters-Even-if/127461/

Solove, D. (2013). Privacy self-management and the consent dilemma. *Harvard Law Review, 126,* 1880-1903.

Straumsheim, C. (2015, December 21). Less than 1%. *Inside Higher Ed.* Retrieved from https://www.insidehighered.com/news/2015/12/21/323-learners-eligible-credit-moocs-arizona-state-u

Stone, K. J. & Stoner, E. N. (2002). Proceedings from 23rd Annual National Conference on Law and Higher Education: *Revisiting the purpose and effect of FERPA.* Orlando, FL.

Sweeney, M. (2012). Understanding privacy. *The Information Society, 28*, 344-345. doi: 10.1080/01972243.2012.712488

Sweeney, L. (2002). Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems, 10*(5), 571-588.

Tarka v. Franklin, 891 F.2d 102 (5th Cir., 1989)

The Catholic University of America Office of General Counsel. (2008). Cases under FERPA. Retrieved from http://counsel.cua.edu/ferpa/fedlaw/cases.cfm

The White House. (2015). FACT SHEET: Safeguarding American consumers & families. Retrieved from https://www.whitehouse.gov/the-press-office/2015/01/12/fact-sheet-safeguarding-american-consumers-families

Tockar, A. (2014, September 8). Differential Privacy: The Basics. Neustar Research. Retrieved from https://research.neustar.biz/2014/09/08/differential-privacy-the-basics/

Turow, J., Feldman, L., & Meltzer, K. (2005). Open to exploitation: America's shoppers online and offline. *A Report from the Annenberg Public Policy Center of the University of Pennsylvania*. Retrieved from http://repository.upenn.edu/asc_papers/35

United States v. University of Miami, 294 F.3d 797 (6th Cir. 2002)

U.S. Department of Education. (2015, October 15). Notice inviting postsecondary educational institutions to participate in experiments under the experimental sites initiative; Federal student financial assistance programs under Title IV of the Higher Education Act of 1965, as amended. Retrieved from https://s3.amazonaws.com/public-inspection.federal register.gov/2015-26239.pdf

U.S. Department of Education. (2011). Family Educational Rights and Privacy Act of 1974. *Federal Register, 76(*232). Retrieved from: http://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title34/34cfr99_main_02.tpl

U.S. Department of Health, Education, & Welfare. (1973). Records, computers and the rights of citizens. *Secretary's Advisory Committee on Automated Personal Data Systems.* Retrieved from http://www.justice.gov/opcl/docs/rec-com-rights.pdf

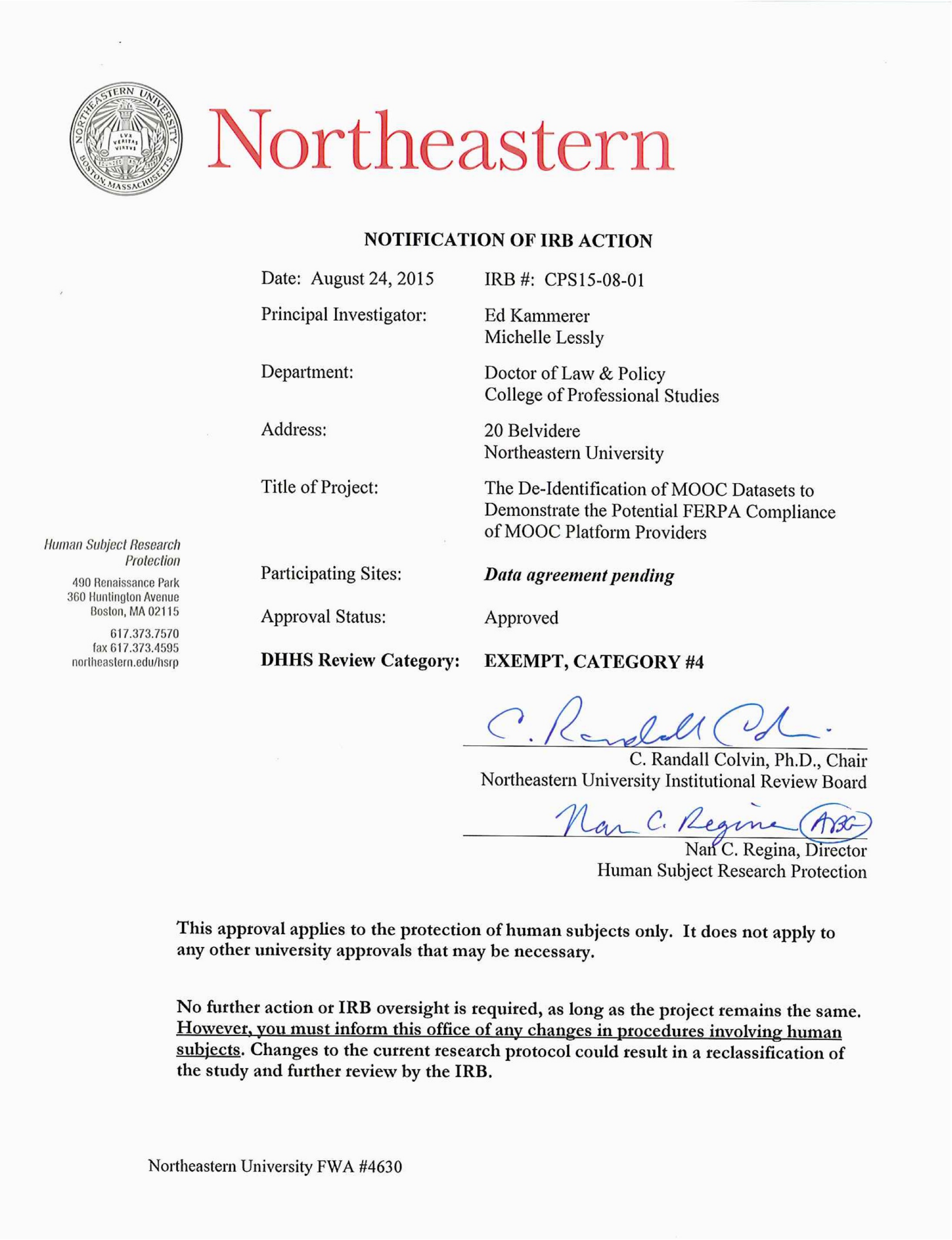
Ward, M.A. (2008). Reexamining student privacy laws in response to the Virginia Tech tragedy. *Journal of Health Care Law and Policy, 11*(2), 407-435.

Warren, S., & Brandeis, L. (1980). The right to privacy. *Harvard Law Review, 4*(5), 193-220.

Woo, M., Reiter, J.P., Oganian, A., & Karr, A.F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *The Journal of Privacy and Confidentiality, 1*(1), 111-124.

Young, E. (2015). Educational privacy in the online classroom: FERPA, MOOCs, and the big data conundrum. *Harvard Journal of Law & Technology, 28*(2), 549-592.

Appendix A: Notification of IRB Action



Northeastern

Human Subject Research Protection

490 Renaissance Park
360 Huntington Avenue
Boston, MA 02115

617.373.7570
fax 617.373.4595
northeastern.edu/hsrp

NOTIFICATION OF IRB ACTION

Date: August 24, 2015	IRB #: CPS15-08-01
Principal Investigator:	Ed Kammerer Michelle Lessly
Department:	Doctor of Law & Policy College of Professional Studies
Address:	20 Belvidere Northeastern University
Title of Project:	The De-Identification of MOOC Datasets to Demonstrate the Potential FERPA Compliance of MOOC Platform Providers
Participating Sites:	***Data agreement pending***
Approval Status:	Approved
DHHS Review Category:	**EXEMPT, CATEGORY #4**

C. Randall Colvin

C. Randall Colvin, Ph.D., Chair
Northeastern University Institutional Review Board

Nan C. Regina (ABG)

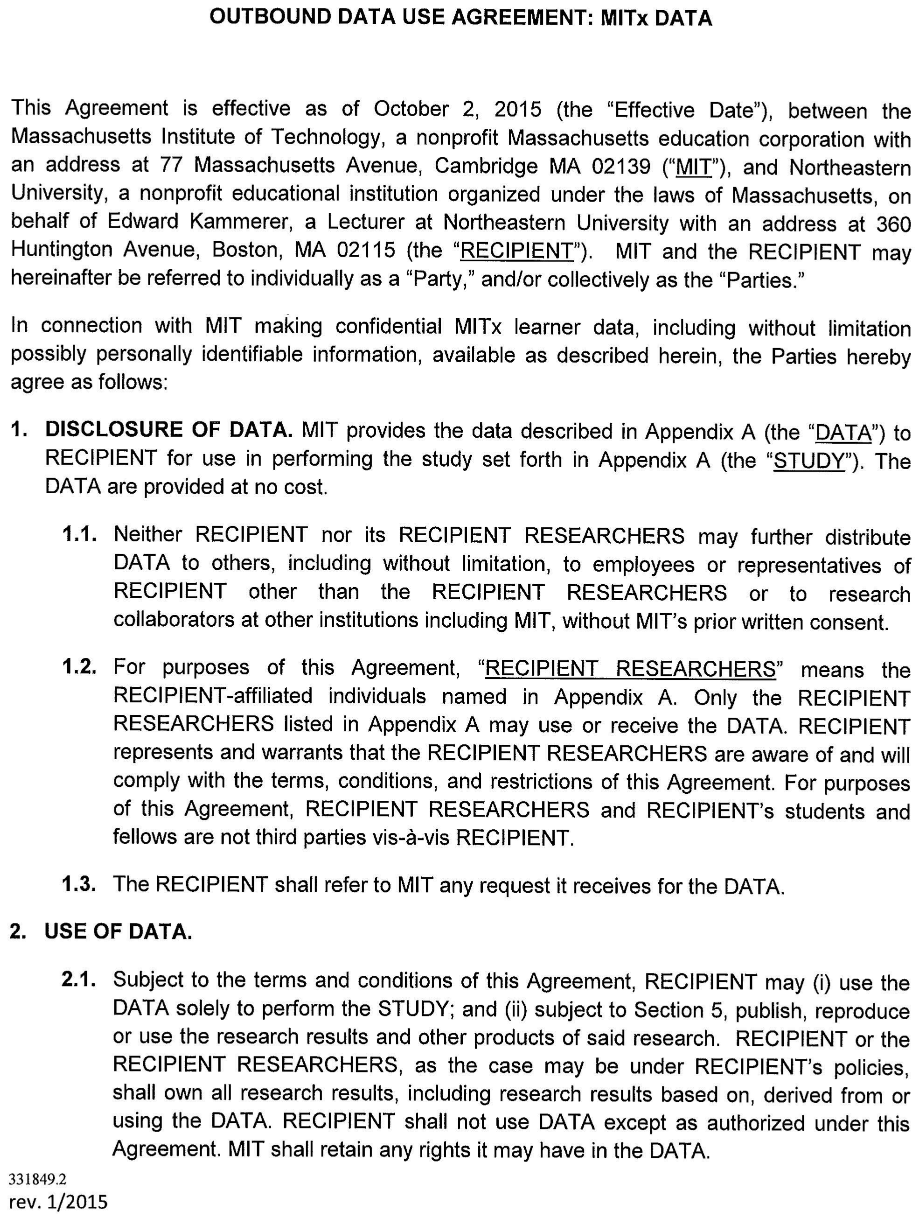
Nan C. Regina, Director
Human Subject Research Protection

This approval applies to the protection of human subjects only. It does not apply to any other university approvals that may be necessary.

No further action or IRB oversight is required, as long as the project remains the same. <u>However, you must inform this office of any changes in procedures involving human subjects</u>. Changes to the current research protocol could result in a reclassification of the study and further review by the IRB.

Northeastern University FWA #4630

Appendix B: Outbound Data Use Agreement: MITx Data



OUTBOUND DATA USE AGREEMENT: MITx DATA

This Agreement is effective as of October 2, 2015 (the "Effective Date"), between the Massachusetts Institute of Technology, a nonprofit Massachusetts education corporation with an address at 77 Massachusetts Avenue, Cambridge MA 02139 ("MIT"), and Northeastern University, a nonprofit educational institution organized under the laws of Massachusetts, on behalf of Edward Kammerer, a Lecturer at Northeastern University with an address at 360 Huntington Avenue, Boston, MA 02115 (the "RECIPIENT"). MIT and the RECIPIENT may hereinafter be referred to individually as a "Party," and/or collectively as the "Parties."

In connection with MIT making confidential MITx learner data, including without limitation possibly personally identifiable information, available as described herein, the Parties hereby agree as follows:

1. **DISCLOSURE OF DATA.** MIT provides the data described in Appendix A (the "DATA") to RECIPIENT for use in performing the study set forth in Appendix A (the "STUDY"). The DATA are provided at no cost.

   1.1. Neither RECIPIENT nor its RECIPIENT RESEARCHERS may further distribute DATA to others, including without limitation, to employees or representatives of RECIPIENT other than the RECIPIENT RESEARCHERS or to research collaborators at other institutions including MIT, without MIT's prior written consent.

   1.2. For purposes of this Agreement, "RECIPIENT RESEARCHERS" means the RECIPIENT-affiliated individuals named in Appendix A. Only the RECIPIENT RESEARCHERS listed in Appendix A may use or receive the DATA. RECIPIENT represents and warrants that the RECIPIENT RESEARCHERS are aware of and will comply with the terms, conditions, and restrictions of this Agreement. For purposes of this Agreement, RECIPIENT RESEARCHERS and RECIPIENT's students and fellows are not third parties vis-à-vis RECIPIENT.

   1.3. The RECIPIENT shall refer to MIT any request it receives for the DATA.

2. **USE OF DATA.**

   2.1. Subject to the terms and conditions of this Agreement, RECIPIENT may (i) use the DATA solely to perform the STUDY; and (ii) subject to Section 5, publish, reproduce or use the research results and other products of said research. RECIPIENT or the RECIPIENT RESEARCHERS, as the case may be under RECIPIENT's policies, shall own all research results, including research results based on, derived from or using the DATA. RECIPIENT shall not use DATA except as authorized under this Agreement. MIT shall retain any rights it may have in the DATA.

2.2. RECIPIENT may use the DATA in the form of raw data and in aggregated form. RECIPIENT may combine the DATA with other data sets.

2.3. RECIPIENT agrees to perform the STUDY and to use the DATA in compliance with all applicable laws and regulations.

2.4. RECIPIENT may not use the DATA in connection with the diagnosis or treatment of human subjects.

3. PURPOSE OF PROJECT. RECIPIENT agrees to use the DATA solely for the STUDY, which relates to instruction, learning, and related sciences.

4. TERM. The Term of this Agreement commences on the Effective Date and ends on the earlier of: (i) June 30, 2016 or (ii) termination under Section 9.

5. PUBLICATION. MIT acknowledges that RECIPIENT is receiving the DATA in anticipation of preparing and publishing scholarly papers ("SCHOLARLY WORKS"). Either prior to or concurrent with submission of any SCHOLARLY WORK for publication, RECIPIENT shall submit a copy of said SCHOLARLY WORK to MIT's Director of Institutional Research to review solely for any disclosure of DATA. MIT shall within fourteen days give RECIPIENT notice identifying specifically any DATA it believes is disclosed in the SCHOLARLY WORK. If MIT does not provide timely notice, it will be deemed to have waived any objection to disclosure of DATA in the SCHOLARLY WORK.

6. CONFIDENTIALITY AND SECURITY.

6.1. RECIPIENT (i) will use reasonable care, consistent with accepted standards in the academic community, to protect the security of the DATA but in all events will ensure that the DATA are encrypted at rest and in transit; (ii) will limit access to the DATA to the RECIPIENT RESEARCHERS; (iii) will not at any time during or after the term of this Agreement disclose DATA to any other person without first obtaining MIT's prior written consent (except as otherwise required by law in which case RECIPIENT shall, unless prohibited by law, notify MIT prior to such disclosure so that MIT may seek a protective order or similar remedy); (iv) will not present, submit for publication, publicly post or publish any personally identifiable information; (v) will not attempt to contact any individuals whose personally identifiable information is contained within the DATA without first obtaining MIT's prior written consent; (vi) will not attempt to re-identify any de-identified DATA by any means, including by combining them with other data sets; and (vii) will promptly notify MIT in the event of unauthorized access to or disclosure of the DATA.

6.2. Notwithstanding the foregoing, in no event is information considered DATA if it (a) was lawfully in the possession of RECIPIENT before receipt from MIT, other than under a prior agreement with MIT; (b) is or becomes publicly available through no fault of RECIPIENT; (c) is received by RECIPIENT, without restriction as to further disclosure, from a third party having an apparent bona fide right to disclose the information to RECIPIENT; or (d) can be demonstrated as being independently developed by RECIPIENT. The foregoing exceptions shall not apply to personally identifiable information contained within the DATA.

7. **USE OF NAMES.** Except as expressly authorized in this Agreement, neither MIT nor RECIPIENT may use (alone or as part of another name) any names, logos, seals, insignia or other words, symbols or devices that identify the other Party or any unit, division or affiliate of the other Party ("PROTECTED NAMES") for any purpose in connection with this Agreement except with the prior written approval of, and in accordance with restrictions required by, the other Party. In the case of MIT, such written approval must be obtained from MIT's Technology Licensing Office. Neither Party may seek to register any PROTECTED NAME of the other Party in any manner in any jurisdiction. Without limiting the foregoing, each Party shall cease all use of PROTECTED NAMES of the other authorized under this Agreement on the termination or expiration of this Agreement, except as otherwise expressly provided herein.

8. **NO REPRESENTATIONS OR WARRANTIES.**

8.1. All DATA are provided "as is." MIT MAKES NO REPRESENTATIONS AND EXTENDS NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED. THERE ARE NO EXPRESS OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, OR THAT THE USE OF THE DATA WILL NOT INFRINGE ANY PATENT, COPYRIGHT, TRADEMARK, OR OTHER PROPRIETARY RIGHTS. RECIPIENT assumes all liability for claims for damages against it by third parties that may arise from its use, storage or disposal of the DATA.

8.2. IN NO EVENT SHALL EITHER PARTY, MEMBERS OF ITS GOVERNING BOARDS, OR ITS OFFICERS, EMPLOYEES, FACULTY, FELLOWS, STUDENTS OR AFFILIATES BE LIABLE FOR INCIDENTAL OR CONSEQUENTIAL DAMAGES OF ANY KIND, INCLUDING ECONOMIC DAMAGES OR LOST PROFITS, REGARDLESS OF WHETHER THE PARTY WAS ADVISED, HAD OTHER REASON TO KNOW OR IN FACT KNEW OF THE POSSIBILITY OF THE FOREGOING.

9. **TERMINATION.**

9.1. This Agreement shall expire as of the date described in Section 3, unless extended by agreement of the parties or terminated earlier under this Section 9. All provisions of this Agreement that are intended to survive expiration or termination of this Agreement, including but not limited to Sections 5 ("Publication"), 6 ("Confidentiality"), 7 ("Use of Names"), 8 ("No Representations or Warranties"), 10 ("Notice"), and 11 ("Miscellaneous Provisions"), shall survive such termination or expiration.

9.2. RECIPIENT may terminate this Agreement without cause by providing thirty (30) days' prior written notice to the other Party. Either Party may terminate this

Agreement if the other Party breaches any material term or condition of the Agreement and fails to cure such breach within 30 days after receipt of written notice of such breach.

9.3. Upon the earlier to occur of completion of the STUDY or termination of this Agreement by MIT for material breach under Section 9.2, RECIPIENT shall destroy the DATA in its possession, and any notes, analyses, documents, or other materials in its possession containing personally identifiable information, and shall certify to MIT as to such destruction, within a reasonable time, not to exceed thirty (30) days from the earliest to occur of completion of the STUDY or termination or expiration of this Agreement, as the case may be ("Destruction Date"); provided, however, that RECIPIENT may retain a copy of the DATA to the extent necessary to comply with records retention obligations of a federal governmental agency or as set forth in any other sponsored research agreement. Following the Destruction Date, and to the extent permitted by law, MIT shall reasonably accommodate requests of RECIPIENT for access to the DATA for purposes of scientific validation of the research, upon the DATA requestor's execution of a data use agreement reasonably acceptable to MIT.

10. NOTICES. Any notices to be given under this Agreement (excluding the actual provision of DATA) shall be in writing and addressed to the parties in care of their respective primary contacts listed in Appendix A. Notices may be delivered in hand or given by certified mail, commercial courier, electronic mail or facsimile transmission.

11. MISCELLANEOUS PROVISIONS.

11.1. Independent Contractors; Non-Exclusive. The parties are independent contractors and do not intend that any agency, partnership, joint venture, or exclusive relationship is created between the parties by this Agreement. Neither Party is authorized to act on behalf of the other or to incur any obligations in the name of the other. Nothing in this Agreement shall be construed as obligating the parties to enter into any subsequent agreement or relationship.

11.2. Entire Agreement; Amendment. This Agreement contains the entire understanding of the parties with respect to the transactions that are the subject matter hereof and supersedes all prior agreements relating to the transactions. This Agreement may be amended or modified only by a written instrument signed by an authorized representative of each Party. The terms of this Agreement govern only the disclosure and use of the DATA for the STUDY as defined herein and do not apply to any other exchange of data between MIT and RECIPIENT.

11.3. Assignment. This Agreement and rights and obligations hereunder may not be assigned by either Party without the other Party's prior written consent.

11.4. Severability. The provisions of this Agreement are severable. In the event any provision of this Agreement is determined to be invalid or unenforceable, such invalidity or unenforceability shall not affect the validity or enforceability of the remaining provisions hereof.

11.5. Waiver. Any waiver of compliance with the terms of this Agreement must be in writing, and any waiver in one instance shall not be deemed a waiver in any future instance.

11.6. Counterparts. This Agreement may be executed in two or more counterparts, each of which will be deemed to be an original, but all of which together constitute one and the same instrument.

11.7. Governing Law and Language. This Agreement will be governed by, and construed in accordance with, the substantive laws of the Commonwealth of Massachusetts, without giving effect to any choice or conflict of law provision.

Executed as of the Effective Date:

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Signed: ______________________________ Date: ___________________

Name: ______________________________

Title: ______________________________

RECIPIENT

Signed: John F. Counts Date: 10/6/2015

Name: John F. Counts

Title: Acting Director, Office of Research Administration & Finance

Appendix A

DATA:

DATA shall mean: Identified MITx 2.01x 2013_Spring with Primary Instructors Simona Socrate, Alexie M. Kolpak and Kun Qian, Identified MITx 3.091x 2013_Spring with Primary Instructor Michael Cima, Identified MITx 8.02x 2013_Spring with Primary Instructors John Belcher and Walter Lewin, and Identified MITx 8.MReV 2013_Spring with Primary Instructor David Pritchard and Colin Fredericks.

STUDY:

The STUDY means the research study identified and described in the Application #[CPS15-08-01] ("[The De-Identification of MOOC Datasets to Demonstrate the Potential FERPA Compliance of MOOC Platform Provideers]") to RECIPIENT'S Institutional Review Board, attached hereto as Exhibit A].

Description of Study: The goals of this replication study Daries et al.'s (2014) study is determine a standardized methodology to that may be used to de-identify massive open online course (MOOC) datasets in order to better protect MOOC users' information when these datasets are made publically available. Moreover, this study aims to remove barriers for MOOC's FERPA compliance.

The operational research question for this study is in what ways might third-party platform provider datasets be de-identified to meet the requirements of C.F.R. Subpart D §99.31(9)(b)(1) of the Family Educational Rights and Privacy Act of 1974, and still maintain their utility for the purposes of research dissemination? This question will address the study's null hypothesis that MOOC user data, when de-identified to be in accordance with FERPA, will not better protect the privacy of MOOC users or maintain the utility of the data itself. Furthermore, this study will support the proposition that the information collected by third-party platform providers constitutes an educational record, making MOOC users students as defined by FERPA Subpart §99.3.

This research seeks to understand FERPA's capacity to meet the needs of MOOC users and providers, and how those parties view themselves in relationship to one another can influence the legal determination of a provider's obligation or duty of care to users. If Daries et al.'s (2014) study can be successfully replicated, a standardized methodology for de-identifying MOOC dataset may be establish and adopted by third-party platform providers. Moreover, this study may begin to remove barriers for MOOC's FERPA compliance and, therefore, result in recommendations to Congress, the Department of Education, and third-party platform providers as how to resolve the concerns of user data management as it relates to research.

Additional recommendations may also be offered to providers concerning best practices on how to protect users' data, transitioning into FERPA compliance if deemed necessary, and how to inform users of their safety entitlements.

Proposed Time Period of Study (if more than one year): 10/02/2015-06/30/2016

RECIPIENT RESEARCHERS (NAME + DATE OF COMPLETION OF CITI HUMAN SUBJECTS TRAINING):

Edward Kammerer, J.D., Ph.D., 08/2009, (Principal Investigator)

Michelle Lessly, 07/11/2014

____

____

____

____

PRIMARY CONTACTS:

For MIT:

Name: Jon Daries____________________

Title: Senior Research Analyst________

Address: 77 Massachusetts Avenue, Cambridge, MA 02139

Phone: 617-324-4810________________

Email: daries@mit.edu_______________

For RECIPIENT:

Name: Dr. Edward Kammerer________________

Title: Lecturer, Principal Investigator____

Address: 360 Huntington Avenue, Boston, MA 02115

Phone: 617-373-8900________________

Email: e.kammerer@neu.edu__________

Appendix C: De-identification Code

Jon Daries' project page, De-identification scripts from first year of MITx and HarvardX courses is located at https://github.com/jdaries/de_id.

His open-source di-identification code is located at:

https://github.com/jdaries/de_id/blob/master/De-identification.ipynb,

https://github.com/jdaries/de_id/blob/master/De-identification.py,

https://github.com/jdaries/de_id/commit/516666f92d5eae5ff12b4291a53df219ba9fb114,

https://github.com/jdaries/de_id/commit/516666f92d5eae5ff12b4291a53df219ba9fb114,

and

https://github.com/jdaries/de_id/commit/516666f92d5eae5ff12b4291a53df219ba9fb114.

Jim Waldo's project page, De-identification scripts from first year of MITx and HarvardX courses is located at https://github.com/harvard/de_id.