acmqueue Privacy, Anonymity, and Big Data in the Social Sciences

Quality social science research and the privacy of human subjects requires trust

Jon P. Daries

Open data has tremendous potential for science, but, in human subjects research, there is a tension between privacy and releasing high-quality open data. Federal law governing student privacy and the release of student records suggests that anonymizing student data protects student privacy. Guided by this standard, we de-identified and released a data set from 16 MOOCs (massive open online courses) from MITx and HarvardX on the edX platform. In this article, we show that these and other de-identification procedures necessitate changes to data sets that threaten replication and extension of baseline analyses. To balance student privacy and the benefits of open data, we suggest focusing on protecting privacy *without* anonymizing data by instead expanding policies that compel researchers to uphold the privacy of the subjects in open data sets. If we want to have high-quality social science research and also protect the privacy of human subjects, we must eventually have trust in researchers. Otherwise, we'll always have the strict tradeoff between anonymity and science illustrated here.

The *open* in massive open online course has many interpretations. Some MOOCs are hosted on open-source platforms, some use only openly licensed content, and most MOOCs are openly accessible to any learner without fee or prerequisites. We would like to add one more notion of openness: open access to data generated by MOOCs. We argue that this is part of the responsibility of MOOCs, and that fulfilling this responsibility threatens current conventions of anonymity in policy and public perception.

In this spirit of open data, on May 30, 2014, a team of researchers from Harvard and MIT (including this author team) announced the release of an open data set containing student records from 16 courses conducted in the first year of the edX platform. (In May 2012, MIT and Harvard launched edX, a nonprofit platform for hosting and marketing MOOCs. MITx and HarvardX are the two respective institutional organizations focused on MOOCs.)⁶ The data set is a de-identified version of that used to publish *HarvardX and MITx: The First Year of Open Online Courses*, a report revealing findings about student demographics, course-taking patterns, certification rates, and other measures of student behavior.⁶ The goal for this data release was twofold: first, to allow other researchers to replicate the results of the analysis; and second, to allow researchers to conduct novel analyses beyond the original work, adding to the body of literature about open online courses.

Within hours of the release, original analysis of the data began appearing on Twitter, with figures and source code. Two weeks after the release, the data journalism team at *The Chronicle of Higher Education* published "8 Things You Should Know about MOOCs," an article that explored new dimensions of the data set, including the gender balance of the courses.¹³ Within the first month of the release, the data had been downloaded more than 650 times. With surprising speed, the data set began fulfilling its purpose: to allow the research community to use open data from online learning platforms to advance scientific progress.

The rapid spread of new research from this data is exciting, but the excitement is tempered by a necessary limitation of the released data: it represents a subset of the complete data. To comply with federal regulations on student privacy, the released data set had to be de-identified. This article demonstrates tradeoffs between the need to meet the demands of federal regulations of student privacy, on the one hand, and our responsibility to release data for replication and downstream analyses, on the other. For example, the original analysis found that approximately 5 percent of course registrants earned certificates. Some methods of de-identification cut that percentage in half.

It is impossible to anonymize identifiable data without the possibility of affecting some future analysis in some way. 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. This could have a chilling effect on the motivations of social science researchers. If findings are likely to be biased by the de-identification process, why should researchers spend their scarce time on de-identified data?

At the launch of edX in May 2012, the presidents of MIT and Harvard spoke about the edX platform, and the data generated by it, as a public good. If academic and independent researchers alike have access to data from MOOCs, then the progress of research into online education will be faster and results can be furthered, refined, and tested. These ideals for open MOOC data are undermined, however, if protecting student privacy means that open data sets are markedly different from the original data. The tension between privacy and open data is in need of a better solution than anonymized data sets. Indeed, the fundamental problem in our current regulatory framework may be an unfortunate and unnecessary conflation of privacy and anonymity. Jeffrey Skopek¹⁷ of Harvard Law School outlines the difference between the two as follows:

under the condition of privacy, we have knowledge of a person's identity, but not of an associated personal fact, whereas under the condition of anonymity, we have knowledge of a personal fact, but not of the associated person's identity. In this sense, privacy and anonymity are flip sides of each other. And for this reason, they can often function in opposite ways: whereas privacy often hides facts about someone whose identity is known by removing information and other goods associated with the person from public circulation, anonymity often hides the identity of someone about whom facts are known for the purpose of putting such goods into public circulation (p. 1755).

Realizing the potential of open data in social science requires a new paradigm for the protection of student privacy: either a technological solution such as differential privacy,³ which separates analysis from possession of the data, or a policy-based solution that allows open access to possibly re-identifiable data while policing the uses of the data.

This article describes the motivations behind efforts to release learner data, the contemporary regulatory framework of student privacy, our efforts to comply with those regulations in creating an open data set from MOOCs, and some analytical consequences of de-identification. From this case study in de-identification, we conclude that the scientific ideals of open data and the current regulatory requirements concerning anonymizing data are incompatible. Resolving that

incompatibility will require new approaches that better balance the protection of privacy and the advancement of science in educational research and the social sciences more broadly.

BALANCING OPEN DATA AND STUDENT PRIVACY REGULATIONS

As with open-source code and openly licensed content, support for open data has been steadily building. In the United States, government agencies have increased their expectations for sharing research data.⁵ In 2003 the National Institutes of Health became the first federal agency to require research grant applicants to describe their plans for data sharing.¹² In 2013 the Office of Science and Technology Policy released a memorandum requiring the public storage of digital data from unclassified, federally funded research.⁷ These trends dovetailed with growing interest in data sharing in the learning sciences community. In 2006 researchers from Carnegie Mellon University opened DataShop, a repository of event logs from intelligent tutoring systems and one of the largest sources of open data in educational research outside the federal government.⁸

Open data has tremendous potential across the scientific disciplines to facilitate greater transparency through replication and faster innovation through novel analyses. It is particularly important in research into open, online learning such as MOOCs. A study released earlier this year¹ estimates that more than 7 million people in the United States alone have taken at least one online course, and that that number is growing by 6 percent each year. These students are taking online courses at a variety of institutions, from community colleges to research universities, and open MOOC data will facilitate research that could be helpful to all institutions with online offerings.

Open data can also facilitate cooperation between researchers with different domains of expertise. As George Siemens, president of the Society for Learning Analytics Research, has argued, learning research involving large and complex data sets requires interdisciplinary collaboration between data scientists and educational researchers.¹⁶ Open data sets make it easier for researchers in these two distinct domains to come together.

While open educational data has great promise for advancing science, it also raises important questions about student privacy. In higher education, the cornerstone of student privacy law is FERPA (Family Educational Rights and Privacy Act). FERPA is a federal privacy statute that regulates access to and disclosure of a student's educational records. In our de-identification procedures, we aimed to comply with FERPA, although not all institutions consider MOOC learners to be subject to FERPA.¹¹

FERPA offers protections for PII (personally identifiable information) within student records. Per FERPA, PII cannot be disclosed, but if PII is removed from a record, then the student becomes anonymous, privacy is protected, and the resulting de-identified data can be disclosed to anyone (20 U.S.C. § 1232g(b)(1) 2012; 34 C.F.R. § 99.31(b) 2013). FERPA thus equates anonymity the removal of PII—with privacy.

FERPA's PII definition includes some statutorily defined categories, such as name, address, social security number, and mother's maiden name, but also

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 (34 C.F.R. § 99.3, 2013).

In assessing the reasonable certainty of identification, the educational institution is supposed to take into account other data releases that might increase the chance of identification.²² Therefore, an adequate de-identification procedure must remove not only statutorily required elements, but also quasi-identifiers. These quasi-identifiers are pieces of information that can be uniquely identifying in combination with each other or with additional data sources from outside the student records. They are not defined by statute or regulatory guidance from the Department of Education but left up to the educational institution to define.²²

The potential for combining quasi-identifiers to uniquely identify individuals is well established. For example, Latanya Sweeney,²¹ from the School of Computer Science at Carnegie Mellon University, has demonstrated that 87 percent of the U.S. population can be uniquely identified with a reasonable degree of certainty by a combination of ZIP code, date of birth, and gender. These risks are further heightened in open, online learning environments because of the public nature of the activity. As another example, some MOOC students participate in course discussion forums—which, for many courses, remain available online beyond the course end date. Students' usernames are displayed beside their posts, allowing for linkages of information across courses, potentially revealing students who enroll for unique combinations of courses. A very common use of the discussion forums early in a course is a self-introduction thread where students state their age and location, among other PII.

Meanwhile, another source of identifying data is social media. It is conceivable that students could verbosely log their online education on Facebook or Twitter, tweeting as soon as they register for a new course or mentioning their course grade in a Facebook post. Given these external sources, an argument can be made that many columns in the person-course data set that would not typically be thought of as identifiers could qualify as quasi-identifiers.

The regulatory framework defined by FERPA guided our efforts to de-identify the person-course data set for an open release. Removing direct identifiers such as students' usernames and IP addresses was straightforward, but the challenge of dealing with quasi-identifiers was more complicated. We opted for a framework of *k*-anonymity.²⁰ A data set is *k*-anonymous if any one individual in the data set cannot be distinguished from at least *k*-1 other individuals in the same data set. This requires ensuring that no individual has a combination of quasi-identifiers different from *k*-1 others. If a data set cannot meet these requirements, then the data must be modified to meet *k*-anonymity, either by generalizing data within cases or suppressing entire cases. For example, if a single student in the data set is from Latvia, we can employ one of these remedies: generalize her location by reporting her as from Europe rather than Latvia, for example; suppress her location information; or suppress her case entirely.

This begins to illustrate the fundamental tension between generating data sets that meet the requirements of anonymity mandates and advancing the science of learning through public releases of data. Protecting student privacy under the current regulatory regime requires modifying data to ensure that individual students cannot be identified. These modifications can, however, change the data set considerably, raising serious questions about the utility of the open data for replication or novel analysis. The next sections describe our approach to generating a *k*-anonymous data set, and then examine the consequences of our modifications to the size and nature of the data set.

DE-IDENTIFICATION METHODS

The original, identified person-course data set contained the following information:

- Information about students (username, IP address, country, self-reported level of education, self-reported year of birth, and self-reported gender).
- The course ID (a string identifying the institution, semester, and course).
- Information about student activity in the course (date and time of first interaction, date and time of last interaction, number of days active, number of chapters viewed, number of events recorded by the edX platform, number of video play events, number of forum posts, and final course grade).
- Four variables computed to indicate level of course involvement (*registered*: enrolled in the course; *viewed*: interacted with the courseware at least once; *explored*: interacted with content from more than 50 percent of course chapters; and *certified*: earned a passing grade and received a certificate).

Transforming this person-course data set into a k -anonymous data set that we believed met FERPA guidelines required four steps: 1) defining identifiers and quasi-identifiers; 2) defining the value for k ; 3) removing identifiers; and 4) modifying or deleting values of quasi-identifiers from the data set in a way that ensures k -anonymity, while minimizing changes to the data set.

We defined two variables in the original data set as identifiers and six variables as quasi-identifiers. The username was considered identifying in and of itself, so we replaced it with a random ID. IP address was also removed. Four student demographic variables were defined as quasi-identifiers: country, gender, age, and level of education. Course ID was considered a quasi-identifier since students might take unique combinations of courses and because it provides a link between PII posted in forums and the person-course data set. The number of forum posts made by a student was also a quasi-identifier because a determined individual could scrape the content of the forums from the archived courses and then identify users with unique numbers of forum posts.

Once the quasi-identifiers were chosen, we had to determine a value of k to use for implementing k -anonymity. In general, larger values of k require greater changes to de-identify, and smaller values of k leave data sets more vulnerable to re-identification. The U.S. Department of Education offers guidance to the de-identification process in a variety of contexts, but it does not recommend or require specific values of k for specific contexts. In one FAQ, the department's Privacy Technical Assistance Center states that many "statisticians consider a cell size of 3 to be the absolute minimum" and goes on to say that values of 5 to 10 are even safer.¹⁵ We chose a k of 5 for our de-identification.

Since our data set contained registrations for 16 courses, registrations in multiple courses could be used for re-identification. The k -anonymity approach would ensure that no individual was uniquely identifiable using the quasi-identifiers *within* a course, but further care had to be taken to remove the possibility that a registrant could be uniquely identified based upon registering in a unique combination or number of courses. For example, if only three people registered for all 16 courses, then those three registrants would not be k -anonymous across courses, and some of their registration records would need to be suppressed in order to lower the risk of their re-identification.

The key part of the de-identification process was modifying the data such that no combination of quasi-identifiers described groups consisting of fewer than five students. The two tools employed for this task were *generalization*, the combining of more granular values into categories (e.g., 1, 2, 3, 4, and 5 become "1-5"); and *suppression*, the deletion of data that compromises k -anonymity.²¹ Many strategies for de-identification, including Sweeney's Datafly algorithm, implement both tools with

different amounts of emphasis on one technique or the other.¹⁸ More generalization would mean that fewer records are suppressed, but the remaining records would be less specific than the original data. A heavier reliance on suppression would remove more records from the data, but the remaining records would be less altered.

The following section illustrates differential tradeoffs between valid research inferences and de-identification methods by comparing two de-identification approaches: one that favors generalization over suppression (hereafter referred to as the generalization emphasis, or GE, method), and one that favors suppression over generalization (hereafter referred to as the suppression emphasis, or SE, method). There are other ways of approaching the problem of de-identification, but these were two that were easily implemented. Our intent is not to discern the dominance of one technique over the other in any general case but rather to show that tradeoffs between anonymity and valid research inferences a) are unavoidable and b) will depend on the method of de-identification.

The SE method used generalization for the names of countries (grouping them into continent/region names for countries with fewer than 5,000 rows) and for the first- and last-event time stamps (grouping them into dates by truncating the hour and minute portion of the time stamps). Suppression was then employed for rows that were not k -anonymous across the quasi-identifying variables. For more information on the specifics of the implementation, please refer to the documentation accompanying the data release.¹⁰

The GE method generalized year of birth into groups of two (e.g., 1980–1981), and number of forum posts into groups of five for values greater than 10 (e.g., 11–15). Suppression was then employed for rows that were not k -anonymous across the quasi-identifying variables. The generalizations resulted in a data set that needed less suppression than in the SE method, but also reduced the precision of the generalized variables.

Both de-identification processes are more likely to suppress registrants in smaller courses: the smaller a course, the higher the chances that any given combination of demographics would not be k -anonymous, and the more likely that this row would need to be suppressed. Furthermore, since an activity variable (number of forum posts) was included as a quasi-identifier, both methods were likely to remove users who were more active in the forums. Since only 8 percent of students had any posts in the forums at all, and since these students were typically active in other ways, the records of many of the most active students were suppressed.

THE CONSEQUENCES OF TWO APPROACHES TO DE-IDENTIFICATION

Both of the de-identified data sets differ from the original data set in substantial ways. We reproduced analyses conducted on the original data set and evaluated the magnitude of changes in the new data sets. This section highlights those differences.

Both de-identified data sets are substantially smaller than the original data set (see table 1), but de-identification did not affect enrollment numbers uniformly across courses. Table 1 shows the percentage decrease of enrollment in each de-identified data set compared with the original file. Only a small percentage of records from CS50x were removed because CS50x was hosted off the edX platform; thus, we have no data about forum usage (one of our quasi-identifying variables).

Table 2 shows that de-identification has a disproportionate impact on the most active students.

TABLE 1 Percent decrease in records by course and by de-identification method

Institution	Course Code	Baseline N	GE Reduction	SE Reduction	Average Reduction
HarvardX	CS50x	181,410	4%	6%	5%
MITx	6.002x	51,394	15%	21%	18%
MITx	6.00x	72,920	15%	21%	18%
MITx	6.00x	84,511	16%	21%	18%
MITx	6.002x	29,050	17%	23%	20%
MITx	8.02x	41,037	17%	24%	21%
HarvardX	PH278x	53,335	18%	26%	22%
HarvardX	ER22x	79,750	21%	28%	25%
MITx	14.73x	39,759	22%	30%	26%
HarvardX	CB22x	43,555	23%	31%	27%
HarvardX	PH207x	61,170	25%	32%	28%
MITx	3.091x	24,493	33%	42%	37%
MITx	8.MReV	16,787	33%	44%	38%
MITx	7.00x	37,997	35%	45%	40%
MITx	3.091x	12,276	39%	50%	44%
MITx	2.01x	12,243	44%	54%	49%
Total		841,687	18%	24%	21%

Andrew Dean Ho et al.⁶ identified four mutually exclusive categories of students: *Only Registered* enrolled in the course but did not interact with the courseware; *Only Viewed* interacted with at least one, and fewer than half, of the course chapters; *Only Explored* interacted with content from half or more of the course chapters but did not earn a certificate; and *Certified* earned a certificate in the course. Table 2 shows that the proportions of students in each category seem to change only slightly after de-identification; however, the percentage of certified students in the de-identified data set is nearly half the percentage in the original data set. Given the policy concerns around MOOC certification rates, this is a substantially important difference, even if only a small change in percentage points.

Demographic data from the de-identified data sets was similar to the original person-course data set. Table 3 shows the distributions of gender and bachelor's degree attainment, respectively, for each data set. The proportions of bachelor's degree holders in all three data sets are nearly identical. The

TABLE 2 Percent decrease in records by activity category and by de-identification method

Activity Category	Baseline N	Baseline Percentage	GE Percentage	SE Percentage	GE Change	SE Change
Only Registered	292,852	34.8%	37.3%	37.6%	+2.5%	+2.8%
Only Viewed	469,702	55.8%	56.2%	56.1%	+0.4%	+0.3%
Only Explored	35,937	4.3%	3.6%	3.5%	-0.7%	-0.7%
Certified	43,196	5.1%	2.9%	2.8%	-2.2%	-2.4%
Total	841,687	100%	100%	100%		

TABLE 3 Changes in demographics and activity by de-identification method

Statistic	Baseline	GE	SE	GE % Change	SE % Change
Percent Bachelor's or Higher	63%	63%	63%	0.1%	-0.2%
Percent Female	29%	26%	26%	-2.2%	-2.9%
Median Number of Events (explored + certified)	3645	2194	2052	-40%	-44%

de-identified data sets report slightly lower percentages of female students than the original data set. The gender bias of MOOCs is a sensitive policy issue, so this difference raises concerns about analyses conducted with the de-identified data sets.

The suppression of highly active users substantially reduces the median number of total events in the courseware. Table 3 shows the median events for all three data sets, and the de-identified data sets have median event values that are two-thirds of the value reported by the original data set.

Finally, we analyzed the correlations among variables in all three of the data sets. We use correlations to illustrate possible changes in predictive models that rely on correlation and covariance matrices, from the regression-based prediction of grades to principal components analyses and other multivariate methods. Although straight changes in correlations are dependent on base rates, and averages of correlations are not well formed, we present these simple statistics here for ease of interpretation. No correlation changed direction, and all remained significant at the 0.05 level. For all registrants, the SE data set reported correlations marginally closer to the original data set than the GE method, while for explored and certified students only, the GE data set was slightly closer to the original (see table 4).

It is possible to use the results from the previous tables to formulate a multivariate model that has population parameters in these tables. By generating data from such a model in proportion to the numbers we have in the baseline data set, we would enable researchers to replicate the correlations and mean values above. Such a model, however, would lead to distorted results for any analysis that is not implied by the multivariate model selected. In addition, the unusual distributions seen in MOOC data² would be difficult to model using conventional distributional forms.

The comparisons presented here between the de-identified data sets and the original data set provide evidence for the tension between protecting anonymity and releasing useful data. We emphasize that the differences identified here are not those that may be most concerning. The above analyses characterize the differences that researchers conducting replication studies might expect to see. For novel analyses that have yet to be performed on the data, it is difficult to formulate an a priori estimate of the impact of de-identification. For researchers hoping to use de-identified, public data sets to advance research, this means that any given finding might be the result of perturbations from de-identification.

BETTER OPTIONS FOR SCIENCE AND PRIVACY WITH RESPECT TO MOOC DATA

As illustrated in the previous section, the differences between the de-identified data set and the original data range from small changes in the proportion of various demographic categories to large

TABLE 4 Changes in Pearson Correlations by de-identification method and activity category

Variable 1	Variable 2	Registrants	Baseline Correlation	GE Correlation	SE Correlation	GE Change (+/-)	SE Change (+/-)
Grade	Number of days active	All	0.800	0.750	0.745	-0.050	-0.055
Grade	Number of days active	Explored + Certified	0.553	0.558	0.564	+0.005	+0.011
Grade	Number of events	All	0.722	0.701	0.697	-0.021	-0.025
Grade	Number of events	Explored + Certified	0.458	0.495	0.501	+0.037	+0.043
Grade	Number of forum posts	All	0.146	0.064	0.156	-0.082	+0.010
Grade	Number of forum posts	Explored + Certified	0.074	0.036	0.108	-0.038	+0.034
Grade	Number of video plays	All	0.396	0.397	0.403	+0.001	+0.007
Grade	Number of video plays	Explored + Certified	0.159	0.194	0.189	+0.035	+0.030
Number of events	Number of days active	All	0.844	0.837	0.835	-0.007	-0.009
Number of events	Number of days active	Explored + Certified	0.736	0.773	0.776	+0.037	+0.040
Number of events	Number of video plays	All	0.665	0.698	0.714	+0.033	+0.049
Number of events	Number of video plays	Explored + Certified	0.587	0.628	0.634	+0.041	+0.047
Number of forum posts	Number of days active	All	0.207	0.104	0.207	-0.103	+0.000
Number of forum posts	Number of days active	Explored + Certified	0.180	0.103	0.200	-0.077	+0.020
Number of forum posts	Number of events	All	0.287	0.117	0.194	-0.170	-0.093
Number of forum posts	Number of events	Explored + Certified	0.279	0.113	0.176	-0.166	-0.103
Number of forum posts	Number of video plays	All	0.091	0.035	0.100	-0.056	+0.009
Number of forum posts	Number of video plays	Explored + Certified	0.051	0.014	0.050	-0.037	-0.001
Number of video plays	Number of days active	All	0.474	0.492	0.505	+0.018	+0.031
Number of video plays	Number of days active	Explored + Certified	0.311	0.404	0.407	+0.093	+0.096
Average		All	0.463	0.420	0.456	-0.044	-0.008
Average		Explored + Certified	0.339	0.332	0.361	-0.007	+0.022

decreases in activity variables and certification rates. It is quite possible that analyses not yet thought of would yield even more dramatic differences between the two data sets. Even if a de-identification method is found that maintains many of the observed research results from the original data set, there can be no guarantee that other analyses will not have been corrupted by de-identification.

At this point it may be possible to take for granted that any standard for de-identification will increase over time. Information is becoming more accessible, and researchers are increasingly sophisticated and creative about possible re-identification strategies. Cynthia Dwork of Microsoft Research, in a presentation at the Big Data Privacy Workshop sponsored by MIT and the White House in early 2014, pointed out that de-identification efforts have been progressing as a sort of arms race, similar to advances in the field of cryptography.⁴ Although *k*-anonymity is a useful heuristic, researchers have challenged that it alone is not sufficient. Ashwin Machanavajjhala et al.⁹ point out that a *k*-anonymous data set is still vulnerable to a “homogeneity attack.” If, after undergoing a process that ensures *k*-anonymity, there exists a group of size *k* or larger for whom the value of a sensitive variable is homogenous (i.e., all members of the group have the same value), then the value of that sensitive variable is effectively disclosed even if the attacker does not know exactly

which record belongs to the target. Machanavajjhala et al. define this principle as *l*-diversity. Other researchers have advanced an alphabet soup of critiques to *k*-anonymity such as *m*-invariance and *t*-similarity.⁴ Even if it were possible to devise a de-identification method that did not impact statistical analysis, it could quickly become outmoded by advances in re-identification techniques.

This example of our efforts to de-identify a simple set of student data—a tiny fraction of the granular event logs available from the edX platform—reveals a conflict between open data, the replicability of results, and the potential for novel analyses on one hand, and the anonymity of research subjects on the other. This tension extends beyond MOOC data to much of social science data, but the challenge is acute in educational research because FERPA conflates anonymity and therefore de-identification—with privacy. One conclusion could be that the data is too sensitive to share; so if de-identification has too large an impact on the integrity of a data set, then the data should not be shared. We believe that this is an undesirable position, because the few researchers privileged enough to have access to the data would then be working in a bubble where few of their peers have the ability to challenge or augment their findings. Such limits would, at best, slow down the advancement of knowledge. At worst, these limits would prevent groundbreaking research from ever being conducted.

Neither abandoning open data nor loosening student privacy protections is a wise option. Rather, the research community should vigorously pursue technology and policy solutions to the tension between open data and privacy. A promising technological solution is differential privacy.³ Under the framework of differential privacy, the original data is maintained, but raw PII is not accessed by the researcher. Instead, it resides in a secure database that has the ability to answer questions about the data. A researcher can submit a model—a regression equation, for example—to the database, and the regression coefficients and R-squared are returned. Differential privacy has challenges of its own, and remains an open research question because implementing such a system would require carefully crafting limits around the number and specificity of questions that can be asked in order to prevent identification of subjects. For example, no answer could be returned if it drew upon fewer than *k* rows, where *k* is the same minimum cell size used in *k*-anonymity.

Policy changes may be more feasible in the short term. An approach suggested by the U.S. PCAST (President's Council of Advisors on Science and Technology) is to accept that anonymization is an obsolete tactic made increasingly difficult by advances in data mining and big data.¹⁴ PCAST recommends that privacy policy emphasize that the use of data should not compromise privacy and should focus on the what rather than the how.¹⁴ One can imagine a system whereby researchers accessing an open data set would agree to use the data only to pursue particular ends, such as research, and not to contact subjects for commercial purposes or to rerelease the data. Such a policy would need to be accompanied by provisions for enforcement and audits, and the creation of practicable systems for enforcement is, admittedly, no small feat.

We propose that privacy can be upheld by researchers bound to an ethical and legal framework, *even if* these researchers can identify individuals and all of their actions. If we want to have high-quality social science research and privacy of human subjects, we must eventually have trust in researchers. Otherwise, we'll always have a strict tradeoff between anonymity and science.

REFERENCES

1. Allen, I. E., Seaman, J. 2014. Grade change: tracking online education in the United States; <http://sloanconsortium.org/publications/survey/grade-change-2013>.

2. DeBoer, J., Ho, A. D., Stump, G. S., Breslow, L. 2013. Changing course: reconceptualizing educational variables for massive open online courses. *Educational Researcher*. Published online before print February 7, 2014.
3. Dwork, C. 2006. Differential privacy. *Automata, Languages and Programming*. Springer Berlin Heidelberg: 1–12.
4. Dwork, C. 2014. State of the art of privacy protection; video <http://web.mit.edu/bigdata-priv/agenda.html>.
5. Goben, A. Salo, D. 2013. Federal research: data requirements set to change. *College & Research Libraries News* 74(8): 421425; <http://crln.acrl.org/content/74/8/421.full>.
6. Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J., Chuang, I. 2014. HarvardX and MITx: the first year of open online courses, fall 2012-summer 2013; <http://ssrn.com/abstract=2381263>.
7. Holdren, J. P. 2013. Increasing access to the results of federally funded scientific research; http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
8. Koedinger, K. R., Baker, R. S. J. d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. 2010. A data repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*, ed. C. Romero, S. Ventura, M. Pechenizkiy, R. S. J. d. Baker. Boca Raton, FL: CRC Press.
9. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M. 2007. L-diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1(1): 3.
10. MITx and HarvardX. 2014. HarvardX-MITx person-course academic year 2013 de-identified dataset, version 2.0; <http://dx.doi.org/10.7910/DVN/26147>.
11. MOOCs @ Illinois. 2013. *FAQ for Faculty*; <http://mooc.illinois.edu/resources/faqfaculty/>.
12. National Institutes of Health. 2003. Final NIH statement on sharing research data; <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>.
13. Newman, J. Oh, S. 2014. 8 things you should know about MOOCs. *The Chronicle of Higher Education* (June 13); <http://chronicle.com/article/8-Things-You-Should-Know-About/146901/>.
14. President's Council of Advisors on Science and Technology. 2014. Big data and privacy: a technological perspective; http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.
15. Privacy Technical Assistance Center. 2012. Frequently asked questions—disclosure avoidance; http://ptac.ed.gov/sites/default/files/FAQs_disclosure_avoidance.pdf.
16. Siemens, G. 2014. The Journal of Learning Analytics: supporting and promoting learning analytics research. *Journal of Learning Analytics* 1(1): 3–5; <http://epress.lib.uts.edu.au/journals/index.php/JLA/article/view/3908/4010>.
17. Skopek, J. M. 2014. Anonymity, the production of goods, and institutional design. *Fordham Law Review* 82(4): 1751–1809; <http://ir.lawnet.fordham.edu/flr/vol82/iss4/4/>.
18. Sweeney, L. 1998. Datafly: a system for providing anonymity in medical data. In *Database Security, XI: Status and Prospects*, ed. T. Lin, and S. Qian. Amsterdam: Elsevier Science.
19. Sweeney, L. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* 671: 1–34.
20. Sweeney, L. 2002a. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5): 557–570.
21. Sweeney, L. (2002b). Achieving k-anonymity privacy protection using generalization and

suppression. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5): 571–588.

22. United States Department of Education. 2008. Family educational rights and privacy. *Federal Register* 73(237). Washington, DC: U.S. Government Printing Office; <http://www.gpo.gov/fdsys/pkg/FR-2008-12-09/pdf/E8-28864.pdf>.

LOVE IT, HATE IT? LET US KNOW

feedback@queue.acm.org

The authors are a group of researchers and administrators from MIT and Harvard who have been working with the data, and policies related to the data, from the MITx and HarvardX MOOCs on the edX platform:

JON P. DARIES, Massachusetts Institute of Technology

JUSTIN REICH, Harvard University

JIM WALDO, Harvard University

ELISE M. YOUNG, Harvard University

JONATHAN WHITTINGHILL, Harvard University

DANIEL THOMAS SEATON, Massachusetts Institute of Technology

ANDREW DEAN HO, Harvard University

ISAAC CHUANG, Massachusetts Institute of Technology

© 2014 ACM 1542-7730/14/0700 \$10.00