Person-Course De-identification Process

Document Date: May 27, 2014

I. Overview

The HarvardX-MITx Person-Course dataset AY2013 (filename "HMXPC13_DI_v2_5-14-14.csv ") contains data from the first year of HarvardX and MITx courses on edX. The raw learner data originating from these courses contains personally identifiable information, and is considered by MIT and Harvard to be protected by the Family Educational Rights and Privacy Act (FERPA) (10 USC. § 1232g; 34 CFR Part 99). This Person-Course dataset was produced from the original learner data by de-identification, removing personally identifiable information using best-practices and expert determination methods, including aggregation, anonymization via random identifiers, and blurring, among other techniques. The resulting dataset is still insightful, and contains data from which results in the "*HarvardX and MITx: The first year of open online courses*" report[1] can be largely reproduced.

The de-identification process is challenging, in particular because edX courses contain significant components that generate publicly accessible data, such as the online discussion forum. Thus, de-identification requires more than simply removing names and email addresses. Even without explicit identifiers, quasi-identifying variables—variables that can, when combined, identify a person—allow someone to identify a record in the data. For example, since many course participants post comments in the course discussion forums, and those forums may persist online long after the end of the course, some of the activity variables in the Person-Course dataset can serve as quasi-identifiers.

This document outlines the process applied by Harvard and MIT to generate the de-identified AY2013 Person-Course dataset, starting from an aggregated (but not further de-identified) Person-Course dataset. The intent of this document is to document the strategy employed and the policy choices made. This process results in a dataset that satisfies FERPA guidelines for release without learner consent, in the view of institutional experts who reviewed the process and the datasets, at both MIT and at Harvard. It is hoped that this process will be of use in the de-identification of future datasets from edX, and will also open doors to research with MITx and HarvardX data.

[1] HarvardX and MITx working paper "*HarvardX and MITx: The first year of open online courses"* by Andrew Ho, Justin Reich, Sergiy Nesterko, Daniel Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2381263)

We begin by providing (Section II) background about the dataset, and the general issues and approaches to de-identification. We then detail the technical implementation and associated policy decisions and their rationale (Section III), before concluding (Section IV) and providing references (Section VI).

II. Background

Researchers from MITx and HarvardX have been collaboratively analyzing the data from the institutions' respective courses released on the edX platform. This joint research effort resulted in the release of course reports (Ho et al., 2014), detailing course-by-course, patterns of student behavior and interactions with the edX courses from Harvard and MIT during the first year of edX. The data behind these reports are in the person-course dataset[2]. This dataset is at the level of one row per-person, per-course. So, for example, if one individual enrolled in three MITx or HarvardX courses during the period covered by the dataset (for this release, Fall 2012, Spring 2013, and Summer 2013), that person would have three rows associated with their userID. The dataset includes both administrative variables and variables generated from user-provided data.

Both institutions are motivated to make as much data freely available as possible. The key constraint to the release of data is that both Harvard and MIT are considering their X-students as protected under the Family Educational Rights and Privacy Act (FERPA) (10 USC. § 1232g; 34 CFR Part 99), which protects the privacy of student records. In order to strike a balance between these competing interests, we have decided to release a version of the person-course dataset that is de-identified in such a way that it would require significant effort to re-identify the students.

The four subsections below describe the approach we have taken for de-identification, which is based on two complementary approaches. On the one hand, we use k-anonymity (Section IIa) to measure degree of de-identification, achieved using generalization of quasi-identifiers. We also check for k-anonymity of users across courses (Section IIb). Matching this, on the other hand, we use L-diversity to ensure that the data is still meaningful and protected against "background knowledge" attacks (Section IIc). A utility matrix is also employed (Section IId) to quantify the degree of change to the pre- de-identified data.

[2] For more details, and definitions for the variables in the dataset, see "The HarvardX-MITx Person-Course Dataset AY2013" reference document, which accompanies the dataset.

IIa. *k*-Anonymity, Quasi-Identifiers, and Generalization

The principle of *k*-anonymity, as described by Sweeney (2002a), was chosen as the goal for the de-identification process. Briefly, k-anonymity is the idea that each person in the dataset cannot be distinguished from at least k-1 other individuals who appear in the same dataset (Sweeney, 2002a, p. 557). For the de-identification of person-course, a k of 5 was chosen, which is consonant with the minimum cell size for survey data reports generated by MIT's Institutional Research group.

Sweeney (2002b) defines the techniques of generalization (whereby "a value is replaced by a less specific, more general value that is faithful to the original") and suppression (whereby the value is removed entirely from the record) as the two primary tools in de-identifying a dataset (p. 575). Along with deleting entire records, these were also the primary tools used to identify the person-course dataset. For example, each country name was associated with a continent/region, and then countries were generalized into their continent/region value based on the number of members in each country group.

We also want to mention Sweeney's (2002a) definition of the quasi-identifier as a variable which alone is not enough to identify an individual, but two or more quasi-identifying variables in conjunction can form unique combinations that could identify an individual. It is also important that the quasi-identifying variables be something that may exist in another available dataset. So, in the person-course dataset, there may be a unique value for *nplay_video* (the number of times the student watched a video in the course), but it is reasonable to assume that the number of video views is not publicly available, so *nplay_video* is not considered a quasi-identifier. The quasi-identifiers we used in de-identifying the person-course dataset were: course name, gender, year of birth, country name, and number of forum posts. The last one was chosen as a quasi-identifier because the edX forums are somewhat publicly accessible and someone wishing to re-identify the dataset could, with some effort, compile the count of posts to the forum by username. We would like to point out that we are using a flexible standard of what can "reasonably" be assumed to be publicly available. In other words, if a student chooses to compile their own statistics and grade along with their edX username and post all of this information on Facebook, this would present a threat to the de-identification of the dataset. However, this is a threat we cannot guard against without significant information loss, and so we determine such circumstances to be outside of the realm of reasonable threats.

IIb. User *k*-anonymity

In addition to checking for k-anonymity across the quasi-identifying variables, we had the added challenge of students who took a unique combination of courses in edX. We wanted the userIDs to be consistent across records, so that researchers

using the dataset could make analyses of how the same individuals behave in different classes. As a trivial example, it could be that only one student registered for both MITx 6.002x and HarvardX JusticeX, and so we would want to delete one or the other row in order to make that *user* k-anonymous, even if their quasi-identifying variables are already k-anonymous.

To implement the user k-anonymity, each user was assigned a 16-digit binary string where each digit (0/1) represented registration in one of the 16 courses in the HarvardX MITx person-course dataset. The number of unique users associated with each of these 16-digit strings was computed, and the strings were divided into two groups: those with k or more members, and those with fewer than k members.

Within the group of strings with fewer than k members, the goal was to selectively delete rows in order to form larger groups from the members of the small groups. Further, the goal was to delete the rows in a manner that wasn't biased against large courses. So, to extend on the earlier example: in addition to our hypothetical student who uniquely registered for 6.002x and JusticeX, suppose we have two students who are the only two students who registered for both JusticeX and 8.02x (See Fig. 1). The simplest way to make these users k-anonymous is to delete all three rows associated with JusticeX, and since no course has fewer than k enrollees, these users would be k-anonymous after one iteration of deletion. However, this approach would be biased against courses that have much higher enrollment than other courses because deleting those rows would have the biggest impact on the dataset.

Before Deletion	After Deleting JusticeX
6.002x, JusticeX	6.002x
8.02x, JusticeX	8.02x
8.02x, JusticeX	8.02x

Figure 1: Deleting user records with bias against classes with higher enrollments.

Before Deletion	After Deleting 6.002x	After Deleting 8.02x
6.002x, JusticeX	JusticeX	JusticeX
8.02x, JusticeX	8.02x, JusticeX	JusticeX
8.02x, JusticeX	8.02x, JusticeX	JusticeX

Figure 2: Deleting user records with more iterations, but less bias against classes with higher enrollments.

In order to minimize information loss and bias, we instead took the approach of abstracting the binary course-combination strings into a decision space and speculatively deleting the rows associated with each of the classes, one at a time, from the set of non-user-k-anonymous groups. The Shannon entropy (a measure of the amount of information in a dataset) of the set of non-user-k-anonymous groups was measured before and after each deletion. Then, we proceeded with the deletion

that had the lowest impact on entropy. Furthermore, rows were only deleted from users who would then be members of a group of size >= k after the deletion. This process was repeated, over many iterations, until the process converged and no more course records could be deleted (See Fig. 2 for a simple example). At the end of this process, there remained users who were not k-anonymous, and all of the records associated with those rows were deleted. In the future, we would like to implement the ability to speculatively delete 2+ courses at a time in order to further reconcile those records.

IIc. L-Diversity and Sensitive Variables

Machanavajjhala et al. (2007) point out that k-anonymity can result in datasets that are 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. For example, if you know the gender, age, and country of residence of an edX student (perhaps from reading their comments online), then if you wanted to learn if they got a certificate or not, you may download the de-identified person-course dataset. By k-anonymity, there are at least k people in the dataset who have the attributes you already know; however, all k of them did not get a certificate, so you now know that your target student did not get a certificate.

Machanavajjhala et al. also discuss vulnerability to "background knowledge attacks (ibid.), whereby an attacker uses statistics about a population to make inferential disclosures. Their example involves the attacker using the fact that people of Japanese ethnicity have a lower-than-average incidence of heart disease to draw conclusions from k-anonymous healthcare records (ibid.). However, we are considering these types of attacks to be outside of the domain of reasonable threats to de-identification.

To ensure l-diversity in the person-course dataset, once the data were k-anonymous, each k-anonymous group (i.e. each group with a unique set of quasi-identifying variables) was evaluated for l-diversity along sensitive variables. Final course grade was considered a sensitive variable. If this value was homogenous within a k-anonymous group, then that homogenous value was redacted.

IId. Utility Matrix

In order to quantify the changes made to the data through the process of de-identification, we employed a concept from Dwork (2006): the utility vector. A utility vector is an object that contains aggregate statistics of key variables in the dataset so that changes in the dataset can be represented in a low-dimensional

framework that is easily interpreted; Dwork describes it as a list of "answers to questions about the data".

For this de-identification process, a 9 by 3 utility matrix was constructed that contained the entropy, mean, and standard deviation of nine numeric variables (see the Person-Course dataset description document for definitions): viewed, explored, certified, grade, nevents, ndays_act, nplay_video, nchapters, nforum_posts. Ideally, the dataset could be edited in such a way that the statistics in the utility matrix would change minimally, while the data would be k-anonymous and l-diverse. Since we are considering some of the utility variables to be quasi-identifiers, the statistics likely change since extreme values are deleted.

III. Implementation

The HarvardX-MITx Person Course dataset contains aggregate data about students' interactions with MITx and HarvardX courses on the edX platform. In order to demonstrate their institutional commitments to open access to the data generated by edX courses, Harvard and MIT have decided to release this dataset to the public. However, as an effort to comply with FERPA, we have made a thorough effort to de-identify the dataset in a way that enforces k-anonymity of quasi-identifying variables, enforces l-diversity of sensitive variables, and minimizes the loss of data. K was chosen to be 5, generally agreed to be a sufficient value for such datasets.

IIIa. Technical Implementation in Detail

The de-identification process, in detail, was as follows.

First, records were analyzed user-wise, to delete records that identify users who have taken a unique combination of courses. Next, courseID, gender, year of birth, country, start date, last date active, number of days active, and number of forum posts were chosen as quasi-identifying variables. These were chosen as most likely to be available in public sources of data (including through the edX platform) and therefore targets for re-identification efforts.

The general strategy was to delete edge cases with the goal of maintaining the statistics of key columns. Country names were replaced with continent/region name for countries with fewer than 5,000 records.

Number of forum posts were similarly trimmed: rows with 60 or more forum posts were deleted.

K-anonymity was evaluated case-wise by concatenating the quasi-identifying variables into a string, and then counting members of unique groups. Next, the

process was to look at the impact further deletions would have on the statistics (mean, standard deviation, and entropy) of key numeric columns, and to find a balance between deletions and impact on the key statistics.

L-diversity was evaluated by checking that no k-anonymous group had a uniform value for any sensitive column. Values of non-diverse sensitive variables were redacted for groups failing the test for l-diversity.

User_IDs were de-identified by taking the original user_ID, adding a random string at the end, hashing the result, sorting the hashed id alphanumerically, and then assigning a sequential ID to the resulting list. The new user_IDs are persistent across courses, but will not persist from dataset to dataset.

The end result does not contain the original user_ID or username, nor any of the IP addresses from the original file.

For the de-identification of the HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 1.0, created on May 14, 2014 (File name: "HMXPC13_DI_v2_5-14-14.csv" md5sum 2b09c674af772d45dae429045cf7acfc) the following apply. The entropy (a measure of information in a dataset) changed by approximately 2% overall, and the entropy changes of each column in the utility matrix (see section IId) ranged from 2% for the viewed variable to 77% for the number of events. Figure 3 (below) details the change in number of records per course as a result of de-identification.

INSTITUTION	COURSE	TERM/YEAR	BEFORE	AFTER
HarvardX	CB22x	2013_Spring	43555	30002
HarvardX	CS50x	2012	181410	169621
HarvardX	ER22x	2013_Spring	79750	57406
HarvardX	PH207x	2012_Fall	61170	41592
HarvardX	PH278x	2013_Spring	53335	39602
MITx	14.73x	2013_Spring	39759	27870
MITx	2.01x	2013_Spring	12243	5665
MITx	3.091x	2012_Fall	24493	14215
MITx	3.091x	2013_Spring	12276	6139
MITx	6.002x	2012_Fall	51394	40811
MITx	6.002x	2013_Spring	29050	22235
MITx	6.00x	2012_Fall	84511	66731
MITx	6.00x	2013_Spring	72920	57715
MITx	7.00x	2013_Spring	37997	21009
MITx	8.02x	2013_Spring	41037	31048
MITx	8.MReV	2013_Summer	16787	9477
		TOTAL	**841687**	**641138**

Figure 3. Change in number of rows by course during de-identification

The de-identification process removes outliers and highly active users because these users are more likely to be unique and therefore easy to re-identify. The resulting dataset will have different demographics and different activity statistics. See Figure 4 for examples.

		ORIGINAL	DE-IDENTIFIED
VIEWED	Yes	65%	62%
	No	35%	38%
EXPLORED	Yes	9%	6%
	No	91%	94%
CERTIFIED	Yes	5%	3%
	No	95%	97%
GENDER	Male	71%	74%
	Female	29%	26%
AGE	Mean	29.3	27.7

Figure 4. Change in engagement variables and demographics during de-identification.

III.b Policy Decisions

As part of the technical implementation, parameters were chosen that constitute policy choices for de-identification. Below is an outline of decisions made, along with the rationale for these decisions and their implications.

Value for K

A value of 5 was chosen because it matches current practice in MIT's Institutional Research group of not releasing survey results for fewer than 5 respondents. This is consistent with advice provided by the Chief Technical Officer of Harvard (Jim Waldo) and his Red Team.

Implications

A higher value of k decreases the risk of re-identification, but increases the number of records that must be removed from the dataset.

Quasi-identifiers

The quasi-identifiers chosen are: course ID, gender, year of birth, country, and number of forum comments. For more detail on quasi-identifiers, please see the section "k-Anonymity, Quasi-identifiers, and Generalization."

Chosen Quasi-identifiers:

- Course ID and number of forum posts: These variables are the **required** quasi-identifiers because they are publicly available through the edX platform.
- Year of birth, country, gender: These variables are **strongly advised** quasi-identifiers because they are the most-commonly-used descriptors that people use to describe themselves online. These can often be inferred from someone's comments on the forums as well.

Not chosen possible quasi-identifiers:

- Level of education: not included because it is less-commonly disclosed in forums. Inclusion would not significantly increase the number of records deleted.
- Date of first activity: suggested because students could announce their registration on social media (Facebook, Twitter) thereby giving a key with which to re-identify their records. Not included because it would have an extreme impact on the number of records to be deleted (from ~9% of dataset with core quasi-identifiers to 81% with this variable). It would be better to remove it from the file entirely than to use as a quasi-identifier.
- Date of last activity: rationale same as date of first activity.
- Number of days active: suggested because it could possibly be guessed at from social media activity and forum activity. Not included because reasonable that few students post to forums *every* time they log in, or few announce *every* active day on social media. Large but not detrimental impact on number of records to delete (from ~9% without to ~30% with).

Implications

The more variables that are considered quasi-identifiers, the lower the risk of re-identification; however more quasi-identifiers also decrease the quality of the resulting dataset.

Sensitive variables

The sensitive variables are the variables that are being protected by de-identification. These are the private data protected by FERPA. For this file, the final grade in a course was the only sensitive variable. Whether or not a student earned a certificate was also considered as a sensitive variable, but as so few students receive certificates (<10% on average), it would be prohibitive to use certification as a sensitive variable. For more information on sensitive variables, please read the section "L-diversity and Sensitive Variables."

Implications

Any variable that is not a sensitive variable and is also not a quasi-identifier, is at minimal risk of disclosure under a homogeneity attack (detailed in the section "L-diversity and Sensitive Variables").

Implementation Strategy

The general strategy chosen was to prioritize integrity of the original values over the number of records preserved. As a result, 10-20% of the dataset will be deleted, but the remaining records will be virtually unaltered. The one exception is that many records will have the location listed by continent/region instead of specific countries. It is presumed that this will not have a large impact on many of the research questions that this dataset can answer.

The most unique users are also the most active users, so another impact of this strategy is that the means and standard deviations on activity variables decrease significantly during the de-identification process. However, it is unavoidable that activity outliers would have been removed, regardless of the de-identification strategy chosen.

Risk of re-identification

It is difficult to quantify the risk of re-identification, but the implementation points to a low probability that the dataset will be re-identified. The risk is low for a variety of reasons:

1. One common re-identification strategy involves using social media or other external sources of data to determine one user_ID, and then use that to determine the algorithm by which the user_IDs were anonymized, and work backwards to determine the rest of the user_IDs. This attack is *not possible* under this implementation because the original userids are "salted" before they are hashed—random text is included in the hash. For re-identification to be possible, students would have to be re-identified more-or-less one-by-one.

2. Location is not 100% reliable. Due to VPNs and anonymous proxies, not all users will end up in the de-identified dataset with a location that matches their actual location.
3. Not all students are heavy social media users. Without a verbose external log of activity on the edX platform (e.g. multiple tweets describing courses in which a student is enrolled), re-identification would be very difficult.
4. Many of the courses in this release are already archived. The risk that an "attacker" could use the edX forums to re-identify the dataset is much lower if said attacker cannot access the forums for those courses. However, if edX opens these to indexing by Google, this would change.
5. Many students are in a group much larger than k. Although worst-case scenarios must be imagined for the groups of size k, many students are even more difficult to identify than that because they are part of a large group of students with the same values for their quasi-identifiers.

However low the risk, it will never be zero. As researchers publish papers based on identified datasets, it is possible that their charts, figures, or results will include details that allow for incremental re-identification of the dataset. It is a goal of this de-identification to ensure that all such re-identification efforts are necessarily incremental.

IV. Conclusion

This de-identification process was designed with the goal of not compromising student privacy, but there is always the risk that these data will be re-identified. The primary threats to our de-identification arise from data that students disclose about themselves on the forums, data that students disclose about themselves elsewhere online (Facebook, Twitter, etc.), and subsequent data releases. As more research is conducted using identified edX data sources, it is increasingly likely that multiple releases of data could become available which, in conjunction, could also open additional avenues to re-identify the dataset.

Moving forward, we would like to improve the de-identification process by reducing the amount of supervision and abstracting more of the generalizing decisions into a decision space where the optimal generalization path could be chosen (c.f. El Emam et al, 2009). Another goal is to automate the de-identification process, making it more accessible to a broader range of users.

V. References

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

El Emam, K et al, A Globally Optimal k-Anonymity Method for the De-Identification of Health Data. *Journal of the American Medical Informatics Association,* 16 (5), 2009; 670-682.

Ho, A. D., Reich, J., Nesterko, S., Seaton, D. T., Mullaney, T., Waldo, J. & Chuang, I. *HarvardX and MITx: The first year of open online courses* HarvardX and MITx Working Paper (1), 2014.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. l-diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3.

Samarati, P. and Sweeney, L. Protecting Privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report, Carnegie Mellon University, SRI, 1998.

Sweeney, L. Datafly: a system for providing anonymity in medical data. *Database Security, XI: Status and Prospects,* T. Lin and S. Qian (eds.), Elsevier Sience, Amsterdam, 1998.

Sweeney, L. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,* 10 (5), 2002; 557-570.

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