Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Dataset

Created February 19, 2016

De-Identification

This document explains some of the thinking and methodology behind the Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Dataset as relates to the preservation of participant privacy through de-identification techniques.

Opening up access to learning data creates more opportunities for researchers and technologists to solve educational challenges, especially in the realm of online education. However, sharing learning data always risks the privacy of the participants. The preparation of this dataset aimed for participant anonymity through the de-identification processes. De-identification of data scrubs identifying information from data using computer algorithms (to remove, replace, or transform) so that the chance of re-identifying an individual is negligible.

The structure and variables of the Canvas Network Person-Course (1/2014 - 9/2015) De-Identified Dataset is based on the HarvardX-MITx Person-Course data release of 2014, and we used similar methods of de-identification. The Harvard and MIT researchers wrote about the legal and ethical considerations around privacy, as well as the technical challenges of thoroughly de-identifying the data while preserving utility. One of the most prominent characteristics of the Harvard and MIT data set is how necessarily limited the resulting data set was, reflecting the degree to which they sought to anonymize the data to protect participant privacy. In fact, the only prominent criticism of the Harvard and MIT data release was that it didn't include enough useful data for researchers.

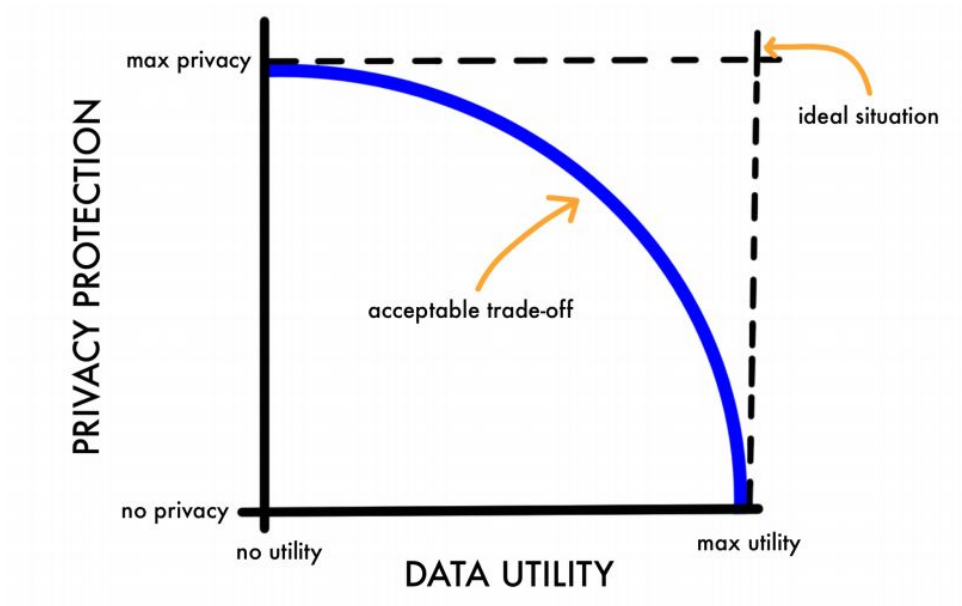

Figure 1. Privacy protection vs data utility. From Gregory Nelson[1].

It may not be possible to fully de-identify such a dataset and retain significant usefulness. It has been posited that the more thorough the de-identification, the less valuable the learning data is (Figure 1). Despite these challenges, the Instructure Research and Education team, working with the Canvas Network team, believes there is value in offering a similar dataset from the Canvas Network open course project to the public research community:

First, there may be value in the data set itself, despite the challenges discussed elsewhere. Comparisons between the different offerings on the different platforms that can be of high-level interest to researchers of the MOOC phenomenon. This is a larger dataset (200+ courses), and we hope to add value by including variables that researchers and commentators have asked for, such as self-reported data on learner intention.

Second, the inherently limited utility of the resulting dataset may reinforce discussions on the feasibility of de-identifying these kinds of learning data. Since this dataset is not dramatically different in its variable scope may encourage work on new methods of protecting user privacy without sacrificing the richness of or access to data. This innate challenge has already been acknowledged by Harvard and MIT researchers, and suggestions for future work have been proposed or are underway. We have made another, richer dataset available to researchers under restricted terms of use as an alternative means of safeguarding privacy. This discussion should also continue to address the question of user privacy and protection of data in open online courses and the open web in general.

[1] Nelson, Gregory S. "Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification."

Third, even if this dataset does not yield impactful insights for education, it may still be useful to student researchers just getting into data analysis.

Methodology (High-Level)

Canvas Network data from January 2014 - September 2015 was queried, organized, and de-identified based on promising known practices in data anonymization, informed and influenced by the work of Harvard and MIT. Harvard and MIT's anonymization methodology included de-identifying both direct and indirect identifiers. Their methods of de-identification involved a technique called *K*-anonymity, which programmatically aims to make any random record from the data set indistinguishable from *K*-1 other records.

In the planning phase of this dataset, we began by studying secondary research and talking with other researchers familiar with the subject. We then manually reviewed all relevant variables case-by-case, explored additional variables beyond what has been released elsewhere, and considered alternatives variables when the breadth or depth of Canvas Network data diverged from published results on MOOC person-course data. Even though Canvas Network courses offer a wide variety of design choices and activities (e.g. groups, peer review, wiki pages, etc), the diversity of course designs and the inconsistency of participation by users discouraged us from including many of those variables in a person-course dataset.

After identifying target variables and querying for the data, we first excluded all users who reported their age as < 18. We identified and removed all direct identifiers, and then used the ARX Data Anonymization Tool to further de-identify the data. The ARX tooling allowed us to both evaluate the risks of releasing different versions of the dataset as well as limit the information loss. Data that is quasi-identifying was generalized using the tooling to produce *K*-anonymity. We followed Harvard and MIT's lead and adopted a *K* value of 5, which means that any random person in the dataset should not be distinguishable from at least 4 other individuals.

In addition, we generalized the names of the courses up to discipline or subject.

Differences in the Canvas Network Person-Course Dataset

Even though this dataset's structure and methodology is based on the HarvardX-MITx data release, there are some important differences:

Some organizations or instructors who had designed Canvas Network courses opted-out of this project, and we have excluded those courses.

Some variables yielded no useful values after de-identification, so we omitted those entirely (e.g. country of origin).

The nature of the Canvas Network data made certain fields irrelevant or too distinct to include as a parallel to HarvardX-MITx data, so we omitted them or introduced alternatives. For example, Canvas Network courses are each uniquely designed for the varying goals of the instructors or sponsoring organizations, and may or may not have sufficient required checkpoints to "certify" a user, let alone certificates per se. This makes any calculated "certified_score" impossible to rely upon across all courses and users. We therefore created alternative fields that aim for similar measures, such as "completed_%" which relays the amount of explicitly required items in the Canvas Network course that were completed by the user, provided that the total amount of required items in the course design met a minimum threshold.

This de-identification process of this dataset resulted in the removal of potentially important demographic data (gender, country of origin); we are investigating methods of restoring this data in a future release.

Because Canvas allows course designers to both create scored activities (for grades) and require specific course activities -- scored or unscored -- there will be person-course records that show one or the other, or both, or neither. This should not be seen as problematic, but rather as reflective of course design choices.

Finally, we added self-reported (survey) data, which suggests individual user's intentions in registering for the open course.