How to De-identify Your Data

OLIVIA ANGIULI, JOE BLITZSTEIN, AND JIM WALDO

HARVARD UNIVERSITY

**BALANCING
STATISTICAL
ACCURACY AND
SUBJECT PRIVACY IN
LARGE SOCIAL-SCIENCE
DATA SETS**

Big data is all the rage; using large data sets promises to give us new insights into questions that have been difficult or impossible to answer in the past. This is especially true in fields such as medicine and the social sciences, where large amounts of data can be gathered and mined to find insightful relationships among variables. Data in such fields involves humans, however, and thus raises issues of privacy that are not faced by fields such as physics or astronomy.

Such privacy issues become more pronounced when researchers try to share their data with others. Data sharing is a core feature of big-data science, allowing others to verify research that has been done and to pursue other lines of inquiry that the original researchers may not have attempted. But sharing data about human subjects triggers a number of regulatory regimes designed to protect the privacy of those subjects. Sharing medical data, for example, requires adherence to HIPAA (Health Insurance Portability and Accountability Act); sharing educational data triggers the requirements of FERPA (Family Educational Rights to Privacy Act). These laws require that, to share data generally, the data be de-identified or anonymized (note that, for the purposes of this article, these terms are

interchangeable). While FERPA and HIPAA define the notion of de-identification slightly differently, the core idea is that if a data set has certain values removed, the individuals whose data is in the set cannot be identified, and their privacy will be preserved.

Previous research has looked at how well these requirements protect the identities of those whose data is in a data set.² Violations of privacy, like re-identification, generally work by linking data from a de-identified data set with outside data sources. It is often surprising how little information is needed to re-identify a subject.

More recent research has shown a different, and perhaps more troubling, aspect of de-identification. These studies have shown that the conclusions one can draw from a de-identified data set are significantly different from those that would be drawn when the original data set is used.¹ Indeed, it appears that the process of de-identification makes it difficult or impossible to use a de-identified (and therefore easily sharable) version of a data set either to verify conclusions drawn from the original data set or to do new science that will be meaningful. This would seem to put big-data social science in the uncomfortable position of having either to reject notions of privacy or to accept that data cannot be easily shared, neither of which are tenable positions.

This article looks at a particular data set, generated by the MOOCs (massive open online courses) offered through the edX platform by Harvard University and the Massachusetts Institute of Technology during the first year of those offerings. It examines which aspects of the de-

identification process for that data set caused it to change significantly, and it presents a different approach to de-identification that shows promise to allow both sharing and privacy.

DEFINING ANONYMIZATION

The first step in de-identifying a data set is determining the anonymization requirements for that set. The notion of privacy that was used throughout the de-identification of this particular data set was guided by FERPA, which requires that personally identifiable information be removed, such as name, address, Social Security number, and mother's maiden name. FERPA also requires that other information, alone or in combination, must not enable identification of any student with "reasonable certainty."

To meet these privacy specifications, the HarvardX and MITx research team [guided by the general counsel, for the two institutions] opted for a k -anonymization framework, which requires that every individual in the data set have the same combination of identity-revealing traits as at least $k-1$ other individuals in the data set. Identity-revealing traits, termed *quasi-identifiers*, are those that allow linking to other data sets; information that is meaningful within only a single data set is not of concern.

Anonymizing a data set with regard to quasi-identifiers is important in order to prevent the re-identification of individuals that would be made possible if these traits were linked with external data that share the same traits. The example in figure 1 illustrates how two data sets could be

combined in such a way that allows re-identification.²

In the edX data set, the quasi-identifiers were course ID, level of education, year of birth, gender, country, and number of forum posts. The number of forum posts is considered to be a quasi-identifier because the forum was a publicly accessible Web site that could be scraped in order to link user IDs with their number of forum posts. Course ID is considered a quasi-identifier because unique combinations of courses could conceivably enable linking personally identifiable information that a student posts in a forum with the edX data set.

The required value of k within k -anonymization was set to 5 in this context, based on the U.S. Department of Education’s Privacy Technical Assistance Center’s claim that “statisticians consider a cell size of 3 to be the absolute minimum” and that values of 5 to 10 are even safer. A higher

FIGURE 1: **COMBINATION OF TWO DATA SETS THAT ALLOW RE-IDENTIFICATION**

value of k corresponds to a stricter privacy standard, because more individuals are required to have a given combination of identity-revealing traits.³

Note that this is not a claim that de-identifying the data set to a privacy standard of $k = 5$ assures that no one in the data set can be re-identified. Rather, this privacy standard was chosen to allow legal sharing of the data.

WHAT METHODS ALLOW ANONYMIZATION?

There are two techniques to achieve a k -anonymous data set: generalization and suppression. Generalization occurs when granular values are combined to create a broader category that will contain more records. This can be achieved both for numerical variables [e.g., combining ages 20, 21, and 22 into a broader category of 20-22] and for categorical variables [e.g., generalizing location data from “Boston” to “Massachusetts”]. Suppression occurs when a record that violates anonymity standards is deleted from the data set entirely.

Generalization and suppression techniques introduce differing kinds and degrees of distortion during the anonymization process. Relying on suppression can mean that a large number of records in the data set will be removed. Suppression-only de-identification also skews the integrity of a data set when values are eliminated disproportionately to the original distribution of the data, causing distortion in resulting analyses.

On the other hand, generalized values are often less powerful than granular values—it may be difficult, for

example, to fit a linear regression line on generalized numeric attributes. Further, while generalization-only de-identification leaves non-quasi-identifier fields intact, quasi-identifiers may become generalized to a point where few conclusions can be drawn about their relationship with other fields. Finally, since generalization is applied to whole columns, it decreases the quality of the entire data set, whereas suppression decreases the quality of the data set on a record-by-record basis.

The anonymization process used to de-identify edX data for public release in 2014 employed a “suppression-emphasis” approach toward k -anonymization. In this approach, the names of the countries were first generalized to region or continent names, then date-time stamps were transformed into date stamps, and finally any existing records that were not k -anonymous after these generalizations were suppressed. In the process, records that claimed a birth date before 1931 (which seemed unlikely to be correct) were automatically suppressed.

Daries et al.’s 2014 study of edX data confirmed that a suppression-emphasis approach tended to distort mean values of de-identified columns, whereas a generalization-emphasis approach tended to distort correlations between de-identified columns.¹

UNDERSTANDING THE MECHANISMS OF DISTORTION

Daries et al. showed that de-identification distorted measures of class participation by suppressing records of rare (generally higher) levels of participation. We pursued

investigating where distortion of summary statistics was being introduced into the data set. Intuitively, distortion is introduced whenever a row becomes generalized or suppressed. Under k -anonymity, this occurs only when a row's combination of quasi-identifier values occurs fewer than k times. If rare quasi-identifier values tend to be associated with high grades or participation levels, then the de-identified data set would be expected to have a lower mean grade or participation level than the original data set.

We did, in fact, find that a quasi-identifier characteristic whose *frequency of occurrence* is correlated with a numeric attribute is most likely to create distortion in that numeric attribute. Specifically, we confirmed this hypothesis in three ways, using the edX data:

- As privacy requirements increase (i.e., k is increased), distortion increases in such numeric attributes as mean grade, shown in figure 2. The fact that more distortion is introduced as more rows are suppressed is consistent with the hypothesis that the association of rare quasi-identifier values with high grades will cause more distortion of the data set as the privacy standard is increased.

- The deletion of quasi-identifier columns whose values' frequency of occurrence is highly correlated with numeric attributes results in a decreased amount of distortion in numeric attributes. This supports the hypothesis that the presence of a correlation between the frequency of quasi-identifier values and numeric attributes introduces distortion of the data set by de-identification.

- As the correlation between the frequency of

FIGURE 2: **DISTORTION OF MEAN GRADE INCREASING WITH K**

occurrence of quasi-identifier values and other numeric attributes is manually increased, more distortion is introduced into those attributes. This, too, supports the hypothesis that the magnitude of a correlation between the frequency of quasi-identifier values and numeric attributes increases distortion of those attributes by de-identification.

What methods may alleviate distortion introduced by de-identification?

The above analyses indicate that associations between quasi-identifier traits and numeric attributes may introduce distortion of means by suppression during de-identification. We therefore consider a prospective role for generalization in alleviating distortion during de-identification.

Since the number of forum posts is the quasi-identifier whose frequency of values is most correlated to grade, we first explore the effect of generalizing this attribute. As the bin size increases [e.g., from 0,1,2,3 to values of 0-1,2-

3,etc.], the number of rows requiring suppression decreases, as shown in figure 3. Further, the mean grade approaches the true value (of 0.045) as bin size increases, suggesting that generalization may alleviate distortion by preventing records associated with rarer quasi-identifier values from becoming suppressed.

Generalization, however, can make it difficult to draw statistical conclusions from a data set. Certain statistical properties of a column, like its mean, can be maintained after generalization by computing a weighted mean of the pregeneralized values within each bin. The average of these bin averages will be equal to the true mean of the pregeneralized values.

Such a solution, however, cannot easily preserve two-dimensional relationships among generalized values. Table 1 illustrates that the correlation of the number of forum posts with various numeric attributes becomes increasingly

FIGURE 3: **DISTORTION OF MEAN GRADE DECREASING WITH BIN SIZE**

TABLE 1: **INCREASING DISTORTION OF CORRELATION WITH INCREASING BIN SIZE**

CORRELATIONS OF FORUM POSTS WITH NUMERIC ATTRIBUTES							
	Original	BIN SIZE					
		1	2	3	4	5	6
Grade	0.159	0.105	0.0980	0.0919	0.0833	0.0732	0.0533
Viewed	0.0444	0.0683	0.0582	0.0462	0.0372	0.0294	0.0228
Explored	0.127	0.0744	0.0710	0.0661	0.0620	0.0554	0.0418
Certified	0.152	0.0868	0.0810	0.0758	0.0699	0.0598	0.0482
# Active Days	0.236	0.117	0.111	0.106	0.0940	0.0855	0.0649
# Chapters	0.154	0.143	0.127	0.115	0.100	0.0858	0.0715
# Events	0.283	0.103	0.103	0.0964	0.0986	0.0913	0.0597
# Video Plays	0.0929	0.0943	0.105	0.103	0.125	0.110	0.0683

distorted with increasing forum post bin size.

Thus is encountered the fundamental tradeoff between generalization and suppression as discussed earlier: although an approach emphasizing suppression may introduce bias into an attribute where a correlation exists between quasi-identifier frequency and numeric attributes, generalization may also distort correlational and other multidimensional relationships inherent within data sets.

Decreasing distortion introduced by generalization

One potential improvement to generalization may be to distribute the number of records more evenly within each bin, using small bucket sizes for values that are well represented and larger bucket sizes for less-well-represented values.

When the number of forum posts is generalized into

groups of five for values greater than 10 (e.g., 1,2,3,...,11-15, 16-20, etc.), the correlations between the number of forum posts and other characteristics become less distorted than with generalization schemes that use constant bin widths. This suggests that optimizing for equal numbers of records within each bin may enable a compromise between the loss of utility and the distortions caused in numeric analysis, such as correlations between different variables. Using this framework for generalization, let's now explore its relationship to suppression in more detail.

A TRADEOFF BETWEEN GENERALIZATION AND SUPPRESSION

To reach a compromise between the distortions introduced by suppression and by generalization, we first want to quantify the relationship between suppression and generalization. As generalization is increased, how much suppression is prevented, and does this change at a constant rate as generalization is increased?

Each of the quasi-identifiers was individually binned to ensure a minimum number of records in each bin, termed *bin capacity*. An increase in bin capacity from 1,000 to 5,000 drastically decreases the number of records that have to be suppressed, but this improvement drops off as bin capacity continues to increase. Furthermore, in figure 4, the decreasing slope of the lines as the bin size increases suggests that the larger the chosen bin capacities, the smaller the marginal cost of a greater degree of anonymity.

We then quantify the distortion that was introduced under each choice of bin capacity. Concentrating on sets

FIGURE 4: **NUMBER OF ROWS SUPPRESSED VS. BIN CAPACITY**

that were 5-anonymous with bin capacities of 3k, 5k, and 10k, we compare the resulting de-identified data sets with the original set on the percentage of students who simply registered for the course; those who registered and viewed (defined as looking at less than half of the material); those who explored (defined as looking at more than half of the material but not completing the course); and those who were certified (completed the material). This comparison shows the greatest disparity in the de-identification scheme

FIGURE 5: ORIGINAL AND DE-IDENTIFIED DATA, 5-ANONYMOUS, 3K BINS

that favors suppression; the results are skewed by as much as 20 percent with the suppression-emphasis de-identification approach.

A generalization scheme using bin capacities of 3,000 entries, as shown in figure 5, produces a distribution of participation that is somewhat closer to the original distribution than the suppression-only approach. While

FIGURE 6: ORIGINAL AND DE-IDENTIFIED DATA, 5-ANONYMOUS, 5K BINS

in some categories the distortion is large (such as the certification rates for MITx|7.00x during the spring semester), others are much closer to the original values.

The situation gets considerably better by using bins with a minimum of 5,000 entries, as shown in figure 6. The distribution of participation is nearly the same in the de-identified set as in the original data set. The maximum

FIGURE 7: ORIGINAL AND DE-IDENTIFIED DATA, 5-ANONYMOUS, 10K BINS

difference between the measures is less than three percentage points; most are within one percent.

Moving to a bin capacity of 10,000 gives even better results, as shown in figure 7. While there are one or two cases of results differing by almost three percentage points, in most cases the difference is a fractional percentage.

As expected, the decrease in the distortion of the mean

FIGURE 8: CORRELATION BETWEEN NUMBER OF FORUM POSTS WITH VARIOUS ATTRIBUTES

of certain attributes is accompanied by an increase in the distortion of the correlation between quasi-identifier fields with numeric attributes as bin capacity increases. The table in figure 8 shows the correlations between the number of forum posts and numeric attributes under various bin capacities. The column corresponding to a bin capacity of 1 represents a suppression-only approach.

Encouraged, we observe that a bin capacity of 3,000 produces a data set whose correlations are close to those of the original, non-de-identified data set, as shown in figure 8.

Even though a bin capacity of 3,000 did not produce optimal results in terms of minimization of class participation distortion, these results may signal the existence of a bin capacity that produces an acceptable balance of distortion between single- and multidimensional relationships.

FURTHER OPPORTUNITIES FOR OPTIMIZATION

Given these results, the question naturally arises whether bin capacities may be chosen differently for each quasi-identifier in order to minimize distortion further.

The edX data set contains two numeric, generalizable quasi-identifier values: year of birth and number of forum posts. Experimentation with different bin capacity combinations yielded the results shown in table 2. This table illustrates the number of records that must be suppressed with the respective amounts of generalization. It is particularly noteworthy that generalization of each quasi-identifier has uneven effects: the required number of suppressed values drops off much more quickly as the bin capacity for number of forum posts increases, as compared with the bin capacity for year of birth.

Such an analysis of the tradeoffs between generalization versus suppression becomes exponentially harder as the number of quasi-identifier values increases. A brute-force method of calculating the number of suppressed records would demand excessive computation time with data sets like edX's that contain six quasi-identifier fields. The development of approximation algorithms for these calculations would enable researchers quickly to determine a near-optimal generalization scheme that strikes an ideal balance between

TABLE 2: **NUMBER OF ROWS SUPPRESSED: NUMBER OF FORUM POSTS BIN SIZE VS. YEAR OF BIRTH BIN SIZE**

		YEAR OF BIRTH: BIN CAPACITY				
		100	200	300	400	500
NUMBER OF FORUM POSTS: BIN CAPACITY	100	1119	1074	1057	1017	1009
	200	577	530	508	468	456
	300	400	356	337	291	280
	400	355	303	281	238	221
	500	283	236	210	169	167
	600	261	208	187	146	131
	700	220	161	140	107	100
	800	224	164	148	112	101
	900	208	138	129	88	76
	1000	186	122	111	72	61
	1100	183	120	109	70	61
	1200	164	94	78	53	46
	1300	145	81	70	48	41
	1400	145	81	70	48	41
	1500	136	78	63	42	35
	1600	134	76	61	45	35

distortions introduced by generalization versus suppression. This is an area where further research is needed.

CONCLUSION

De-identification techniques will continue to be important as long as the regulations around big-data sets involving human subjects require a level of anonymity before those sets can be shared. While there is some indication that regulators may be rethinking the tie between de-identification and

ensuring privacy, there is no indication that the regulations will be changed any time soon. For now, sharing will require de-identification.

But de-identification is hard. We have known for some time that it is difficult to ensure that the data set does not allow subsequent re-identification of individuals, but we now find that it is also difficult to de-identify data sets without introducing bias into those sets that can lead to spurious results.

A combination of record suppression and data generalization offers a promising path to solving the second of these problems, but there seems to be no magic bullet here; our best results were obtained by trying a number of different combinations of generalization, sizing, and record suppression. There is further work to be done, such as investigating the possibility of choosing different bin capacities for different quasi-identifiers, which may mitigate some of the distortions introduced by anonymity. We are more confident than we were a year ago that some form of de-identification may allow sharing of data sets without distorting the analyses done on those shared sets beyond the point of usefulness, but there is much left to investigate.

References

1. Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A.D., Seaton, D. T., Chuang, I. 2014. Privacy, anonymity, and big data in the social sciences. *Communications of the ACM* 57(9): 56-63.
2. Sweeney, L. 2002. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and*

Knowledge-Based Systems 10[5]: 557–570.

3. 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.

LOVE IT, HATE IT? LET US KNOW feedback@queue.acm.org

Olivia Angiuli received her Bachelor's degree in Statistics and Computer Science in 2015 from Harvard College. She began working at Quora as a Data Scientist in July 2015. She is ultimately interested in harnessing big data for social good. She can be reached at oangiuli@post.harvard.edu.

Joe Blitzstein is a professor of the practice of statistics at Harvard University, whose research is a mixture of statistics, probability, and combinatorics. He is especially interested in graphical models, complex networks, and Monte Carlo algorithms. He received his Ph.D. from Stanford University. He can be reached at blitz@fas.harvard.edu.

Jim Waldo is a Gordon McKay Professor of the Practice in Computer Science, a member of the faculty of the Kennedy School, and the Chief Technology Officer at Harvard University. His research centers around distributed systems and topics in technology and policy, especially around privacy and cyber security. Jim was a Distinguished Engineer at Sun Microsystems, where he worked on the Java programming language and various projects in Sun's research lab. He can be reached at waldo@seas.harvard.edu.