Statistical Tradeoffs between Generalization and Suppression in the De-Identification of Large-Scale Data Sets

Olivia Angiuli[1], Jim Waldo[2]
Harvard University
Cambridge, Massachusetts
[1] oangiuli@post.harvard.edu
[2] waldo@seas.harvard.edu

***Abstract*—Data sets containing private information about individuals must satisfy privacy standards before being publicly released. One such standard, k-anonymity, reduces the probability of the re-identification of individuals by requiring that rare combinations of personally-identifiable information be represented by at least *k* distinct individuals. Records that violate this standard must be altered, which can lead to significant distortion of the statistical properties of the data set. In this paper, we discuss improvements to two techniques used to achieve k-anonymity, generalization and suppression, that confer k-anonymity while better preserving the statistical properties of an educational data set taken from a massive online open course platform, edX.**

Keywords—anonymity; de-identification; data privacy

I. Introduction

Though large-scale data sets have the potential to fuel advances in medicine, education, the social sciences, and technology, their utility is limited by the extent to which they can legally be publicly shared due to concerns over subject privacy. Many such data sets include personally identifiable information that must be excluded or amended in order to preserve the privacy of their subjects. This paper examines one of the anonymization standards, k-anonymity, to which subject privacy can be held. Specifically, it explores the tradeoffs that two techniques used to achieve k-anonymity -- generalization and suppression -- involve for the statistical representativeness of de-identified data sets with respect to their original counterparts. We find that improvements to the generalization and suppression algorithms can mitigate bias introduced during the k-anonymization process, enabling the generation of de-identified data sets that are significantly more representative of their original counterparts than had been generated with past versions of the algorithms.

II. Anonymity Standards

Current legal standards have equated privacy to a notion of anonymity, allowing disclosure of data as long as there exists a minimum level of uncertainty about the identity of the subjects. Required levels of anonymity for data sets are often legally enforced and may differ across fields. Medical data, for example, is protected by the Health Insurance Portability and Accountability Act (HIPAA), which requires that highly sensitive fields like names, addresses, social security numbers and telephone numbers, generally referred to as *directory information*, be removed and that any subject's chance of being re-identified is "statistically small" [1]. Educational data is protected by the Family Educational Rights and Privacy Act (FERPA), which similarly requires the removal of highly sensitive fields and that remaining information, alone or in combination, must not enable identification of any student with "reasonable certainty " (34 C.F.R. § 99.3,2013).

Re-identification can occur from just the information within a data set if directory information is retained. These fields are obvious mechanisms for re-identification and must be removed if anonymity is to be maintained. A more subtle vector for re-identification can occur whenever two data sets share fields in common, as shown in Fig. 1 below. By joining together rows that share the same identifying values, additional information can be determined about a person that is potentially identity-revealing. Traits that could conceivably be linked with external data sets to reveal identity -- in this case, zip, birth date, and sex -- are termed *quasi-identifiers*.

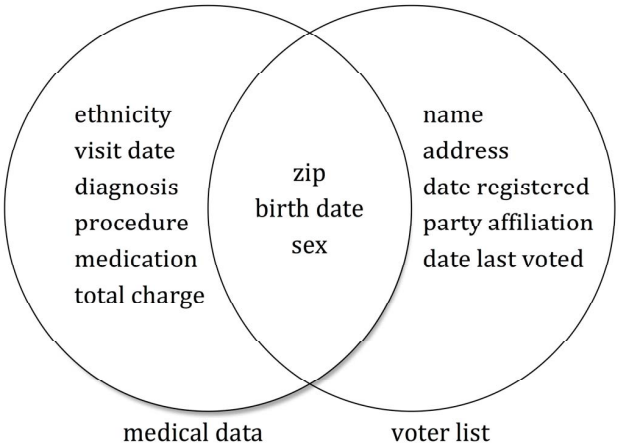

Fig. 1. Combination of two data sets that allow re-identification [2].

A. k-Anonymity

This article examines an educational data set obtained from the Massive, On-line Open Courses (MOOCs) offered by Harvard University and the Massachusetts Institute of Technology through the edX platform over the years 2013-2015. Our prior analysis was of the data set for academic year 2013-2014, and this new analysis is of the data set for academic year 2014-2015. Unlike previous studies of de-identifying data sets, we are not concerned with the possibility

0730-3157/16 $31.00 © 2016 IEEE
DOI 10.1109/COMPSAC.2016.198

of re-identification in these sets. Instead, we focus on the statistical changes that occur in such sets as a result of the de-identification mechanisms chosen to reach desired levels of k-anonymity, and their possible repercussions for conclusions reached in scientific research performed on those data sets.

We opted to use a *k-anonymity* framework to satisfy FERPA requirements for minimizing the ability to re-identify subjects by the linking of data sets through quasi-identifiers. Our choice was dictated by the legal analysis of the FERPA requirements contained in Young [3]. This standard requires that there be in a data set at least *k* individuals with any unique combination of quasi-identifier characteristics, ensuring that, even if the data set shares quasi-identifier traits with an external data set, there is at most only a 1/k chance of an individual's being correctly tied to a given row in that external data set. For the example in Fig. 1 above, this would mean that k different records in the medical data set share the quasi-identifier values of zip, birth date, and sex, and linking an individual from medical data to the voter list cannot be accomplished.

For the purposes of releasing the edX data set, a value of k=5 has been chosen in compliance with the statement by the U.S. Department of Education's Privacy Technical Assistance that "statisticians consider a cell size of 3 to be the absolute minimum needed to prevent disclosure, though larger minimums (e.g., 5 or 10) may be used to further mitigate disclosure risk" [4].

As a more concrete example of k-anonymity, consider the tables below, one with the incomes of five different females, and one with the sex, age, and name of a voter registrant. The table of incomes satisfies a k-anonymity of k=5, since at least 5 different rows contain any given quasi-identifier combination (sex=F, age=37). The probability that an adversary could correctly link the name=Taylor Smith column from the second data set to a corresponding income is therefore 1/5. If the second table were itself k-anonymous at a level of k=5, there would be four other voter registrant females for whom age=37, reducing the probability of linking to 1/25.

TABLE I. INCOME AND VOTER REGISTRATION TABLES

Income Table		
Sex	***Age***	***Income***
F	37	$52,000
F	37	$37,000
F	37	$119,000
F	37	$12,000
F	37	$88,000

Voter Registration Table		
Sex	***Age***	***Name***
F	37	Taylor Smith

III. TWO METHODS OF ACHIEVING *K*-ANONYMITY: GENERALIZATION AND SUPPRESSION

We explore two main methods, *generalization* and *suppression,* for de-identifying a data set to a desired level of k-anonymity prior to its statistical analysis.

Generalization occurs when multiple values in a column are collapsed into more-encompassing entries. Numeric values, for example, can be generalized from single values to ranges, e.g., from age 37 to age 35-40, and non-numeric values can be generalized to broader granularity, e.g., from Atlanta to Georgia. Generalization helps achieve k-anonymity by favoring more rows with any given quasi-identifier value and fewer with unique quasi-identifier combinations.

Suppression occurs when all of a row's values are deleted from the data set. Simply deleting a row may be desirable to preserve the integrity of a data set where the presence of a single outlier quasi-identifier value would otherwise require a large degree of generalization.

There are tradeoffs between generalization and suppression. Generalization mitigates the number of rows that have to be suppressed but, since generalization affects all values, lessens the precision of results across the entire data set. Suppression, on the other hand, affects only one row at a time but, in removing the values of deleted rows, distorts the true means of columns, lessening the quality of and introducing bias into the data set.

A. *k-Anonymity Algorithm*

The edX data set that is being analyzed has six quasi-identifiers: course ID, level of education, year of birth, gender, country, and number of forum posts. The number of posts to publicly-accessible online forums is considered a quasi-identifier because these forums are public, and a scraper could potentially link post counts with user IDs to re-identify students. Similarly, course ID is considered a quasi-identifier because publicly-accessible class lists or personally-identifiable information from online posts could conceivably link unique combinations of courses with user IDs.

Each record in the data set represents a particular student-course pair, specifying both student information (e.g., user ID, level of education, year of birth, gender, country) and course performance data (e.g., level of participation, grade, start time). Data concerning course performance is not linkable to other data sets, and therefore are not considered quasi-identifiers.

We anonymize the data set to a desired level of k-anonymity by the following algorithm.

1) Generate a list of records for suppression due to rare course combinations. Identify each user-course combination representing fewer than k students, then sequentially suppress records with the lowest user-course participation levels until all user-course combinations represent k or more students.

2) Generalize numeric quasi-identifiers by converting single values to ranges of values. Here the numeric quasi-identifiers are year of birth and number of forum posts. Using a user-inputted bin capacity c, sequentially merge quasi-

identifier values until there are at least c records for each merged value.

3) Generalize user countries to regions or continents where necessary. Identify each user country representing fewer than 5,000 students, then broaden those values to regions or continents to preserve k-anonymity.

4) Generalize user level of education. Generalize this field to reflect whether the student had finished a college-level education or not. (The educational researchers who were using the data set indicated that this was the granularity of interest.)

5) Suppress all remaining records that violate k-anonymity. Suppress from the data set any record surviving generalization and suppression from steps 1-4, but whose combination of quasi-identifier values are shared with fewer than k-1 other students.

Steps 1 and 3 are dependent on one-dimensional distributions of, respectively, course and country quasi-identifiers that result, for a given data set, in fairly constant and usually minimal degrees of generalization and suppression. Most records alteration is thus accounted for by generalization and suppression occurring in steps 2 and 5, with step 2's chosen bin capacity c determining the balance in step 5: high bin capacity increases generalization, reducing the need to suppress rows, while low bin capacity decreases generalization, requiring that more rows be suppressed.

IV. Minimizing Distortion of a Data Set

Daries et al. showed that de-identification of the 2013-2014 edX data set led to significant changes in the data set, including shifts in the underlying demographics and grade distributions of students who completed courses [5].

Angiuli et al.'s follow-up study of the 2015 edX data set found that suppression-heavy anonymization approaches tend to bias column means, while generalization-heavy approaches tend to bias correlations among columns. Their work further demonstrated as a contributive factor that rare quasi-identifier values tended to be associated with high-performing students, making records with high grades more likely to require suppression [6].

We now pursue understanding how distortion can be mitigated at multiple stages of the anonymization process. Specifically, we investigate how improving suppression, generalization, and balancing of the two in achieving k-anonymity affects the statistical properties of a data set, especially toward preserving its means and correlations – statistical properties that are among its most basic, most important to interpretation, and most crucial to the accuracy of linear and other simple models.

A. Improving Suppression

Recall that suppression occurs when records exist whose quasi-identifier combinations occur fewer than k times in a data set. We previously found that suppression most drastically distorts the means of data sets. This effect is especially pronounced in data sets where rare quasi-identifier values, most likely to require suppression, are associated with outlier values of numeric traits. In the edX data set, since students with high numbers of forum posts tend to achieve higher grades, suppression is liable to lower significantly the mean grade.

In order to mitigate this effect, we investigate how adding simulated rows, constructed so as to maintain important statistical properties of the data set, can reduce the need to suppress rows.

Accordingly, for every record whose quasi-identifiers occur in $n < 5$ total records, we generate 5-n "fake" records with non-quasi-identifier values drawn from one of the following distributions:

1) Marginal. From the column's values in the original data set, independently draw each column value randomly.

2) Marginal mean. From the column's values in the original data set, independently draw each column value as the mean of 5 random choices.

3) Joint. From the column's values for records with identical quasi-identifier values, randomly draw each column value, with additive noise to mask which rows are fake.

4) Joint mean. From the column's values for records with identical quasi-identifier values, equate column value to their mean, with additive noise to mask which rows are fake.

We generated 50 datasets from each method, and compare resultant bias of the mean and correlations.

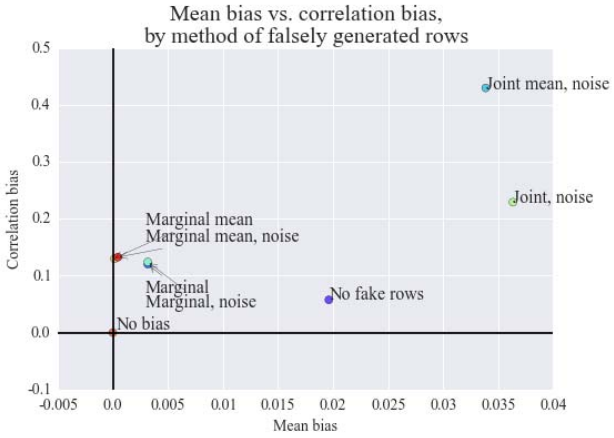

Fig. 2. Bias in mean and correlation resulting from anonymization by generating fake rows.

We reach three major conclusions from Fig. 2.

First, generating fake values drawn from the marginal distributions of columns can anonymize data sets with low distortion to means. This follows logically from the fact that random records from this distribution should, on average, have a mean close to the true mean of the data set. The marginal mean method, which sets the fake value as the average of five randomly drawn values, offers even more protection against drawing outlier values, thereby showing even lower bias of means. Interestingly, the addition of noise appears to have a negligible effect on overall bias, suggesting that it provides a

good way to mask which values are fake without significantly affecting results.

Second, generating fake values from the marginal distributions of columns substantially mitigates bias of means (versus the "No fake rows" datapoint), but imposes a cost in bias of correlations. In a setting where a researcher cares about preserving correlations, as when performing linear regressions between variables, this tradeoff may be undesirable.

Third, generating fake values from the joint distributions of anonymity-violating rows produces anonymized data sets with high bias of both means and correlations. This is attributed to the fact that anonymity-violating rows tend to be for students with high numbers of forum posts, and in turn high grades. Fake grades drawn from this distribution thus highly bias, by inflating the mean and the correlation of the anonymized data set.

B. Improving Generalization

The original generalization algorithm merged sequential values of quasi-identifiers until each generalized value has at least c records with that value, where *c* corresponded to a user-inputted bin capacity. For example, consider a data set with the following frequencies of quasi-identifier values for year of birth:

TABLE II. GENERALIZATION EXAMPLE.

Year of Birth	1950	1951	1952	1990	1991
Frequency	10	4	4	6	11

For a bin capacity of 10, the algorithm would scan from left to right, joining year of birth values together until there was a combined frequency of 10 in each bucket. In this case, the algorithm would produce the bins: 1950, 1951-1990, 1991.

However, this bucketing algorithm could be improved by optimizing for the minimization of bin width. A bucketing schema of 1950-1952 and 1990-1991 would enable a greater resolution into a person's true year of birth, while still satisfying a bin capacity of 10. Furthermore, certain statistical properties of data sets require a placeholder value to be used in place of the binned values: the mean of the pre-discretized values, which is a natural choice for this placeholder, would be preserved more closely under this improved generalization algorithm.

To implement this, we run a greedy algorithm in which values whose bins have fewer than the bin capacity c records get merged with the next smaller or larger value. The direction of merging is chosen to minimize the change in mean of the pre-discretized values: at every step, the least costly merge is performed, and this repeats until every bin has at least c records, at which point the algorithm stops.

The following graphs compare the resulting bias in the de-identified data sets using the original versus the greedy generalization algorithm. A diagonal line is included in both graphs to represent where points would lie if the two generalization schemes led to equally-biased results; points lie above the line when the greedy scheme is better, and below the line when the original scheme is better.

Fig. 3 below indicates that the greedy algorithm leads to de-identified data sets that have less bias in means for all bin sizes. This suggests that this algorithm mitigates the bias in means that is caused when privacy-violating quasi-identifiers are systematically associated with high values of a non-quasi-identifier column, as is the case in the edX data set.

Interestingly, however, the greedy binning algorithm worsens bias in correlations as compared with the original binning algorithm for all bin sizes except 5k (Fig. 4). This may occur if locally-optimal greedy decisions do not lead to globally optimal results.

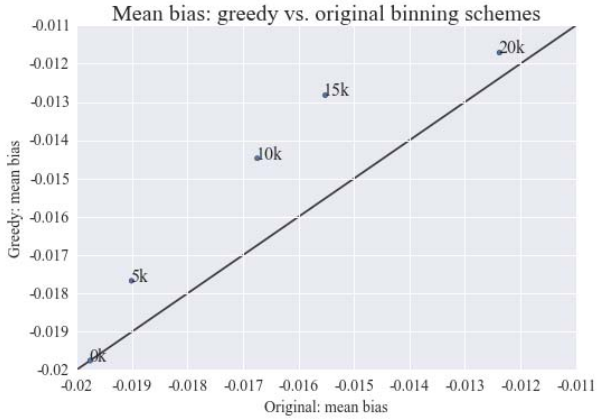

Fig. 3. Bias in mean grade resulting from generalization by original versus greedy algorithm.

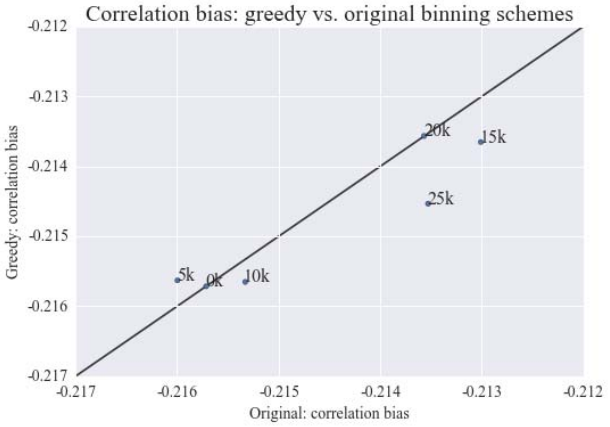

Fig. 4. Bias in correlation between grade and number of forum posts resulting from generalization by original versus greedy algorithm.

C. Combining Suppression and Generalization Improvements

The power of the demonstrated improvements in generalization and suppression compounds significantly when the two algorithms are combined. Fig. 5 shows the biases in mean (x-value) and the biases in correlation (y-value) for the improved algorithms. In this plot, "Bin" connotes the use of

the improved generalization algorithm, "Marginal mean" connotes the use of the improved suppression algorithm, and "Combined" connotes the use of both. The horizontal and vertical black lines indicate where data sets with no bias in the correlation or mean would be located. All of the "Combined" approaches lead to significantly lower biases in both means and correlations than the separate approaches.

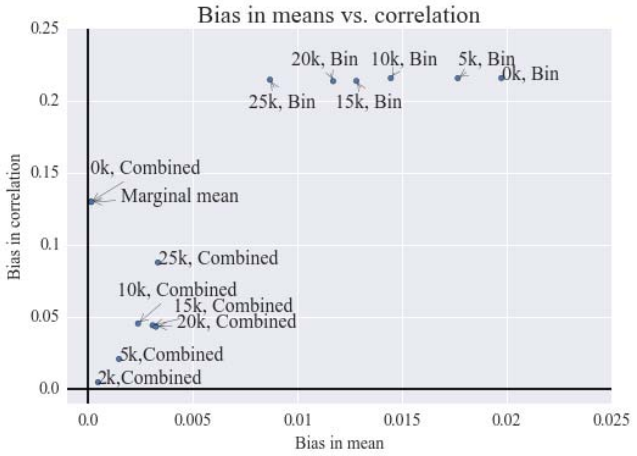

Fig. 5. Bias in means and correlations resulting from generalization, suppression, and combined generalization and suppression improvements, with specified bin sizes.

V. Conclusion

This research points to an exciting new direction that k-anonymization can take to improve the statistical representativeness of data sets by improving algorithms for generalization and suppression. It further suggests that a combined approach, improving both, can mitigate the separately-distorting effects on data sets of suppression and generalization, minimizing bias while preserving privacy. In this paper, we have demonstrated the feasibility of such algorithmic improvements for the edX data set.

Interesting questions remain. The current combined approach minimizes bias when bin sizes are reduced down to 2k, but shows a large increase in bias for the 0k bin size. Understanding the precise relationship between bin size and bias reduction is yet to be done, as is determining whether different bin sizes for different categories of quasi-identifiers have an impact on the bias. Applying these techniques to other data sets to find if the approach is generalizable also remains to be done. However, these early results are promising.

Developing an algorithm that programmatically balances generalization and suppression schemes to minimize bias in the means, correlations, and other statistical properties of data sets would be groundbreaking. Such a globally-optimizing algorithm would allow sharing information-sensitive data sets without compromising the privacy of the participant subjects, opening powerful new fronts for harnessing the analytical powers of big data.

Acknowledgment

This research extends previous work completed in collaboration with Joe Blitzstein of Harvard University, who continues to provide valuable technical feedback and who originally suggested the idea of exploring anonymity from a statistical point of view.

References

[1] US Department of Health and Human Services. Guidance regarding methods for de- identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule. http://www.hhs.gov/, May 2014.

[2] Latanya Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557–570, 2002.

[3] Young, Elise; 2015. Educational Privacy in the Online Classroom: FERPA, MOOCs, and the Big Data Conundrum. Harvard Journal of Law and Technology, Volume 28, Number 2. http://jolt.law.harvard.edu/articles/pdf/v28/28HarvJLTech549.pdf

[4] Privacy Technical Assistance Center. 2012. Frequently asked questions — disclosure avoidance; http://ptac.ed.gov/sites/default/files/FAQs_disclosure_avoidance.pdf.

[5] 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.

[6] Angiuli, O., Blitzstein, J., and Waldo, J. How to de-identify your data. Queue 13, 8 (Sept. 2015).