Big Data and the Social Sciences:
Can Accuracy and Privacy Co-Exist?

Jim Waldo
Harvard University

Abstract

Big Data Science, which combines large data sets with techniques from statistics and machine learning, is beginning to reach the social sciences. The promise of this approach to investigation are considerable, allowing researchers to establish correlations between variables over huge numbers of participants using data that has been gathered in a non-invasive fashion and in natural settings.

Unlike large-data projects in the physical sciences, however, use of these data sets in the social sciences require that the subjects generating the data be treated in a fair an ethical fashion. This is often taken as requiring either compliance with the common rule, or that the data be de-identified to insure the privacy of the subjects.

But de-identification turns out to be far more difficult than one might think. In particular, the ability to re-identify subjects from a set of attributes that can be linked to other data sets has led to a number of mechanisms, such as k-anonymity or l-diversity, that attempt to define technical solutions to the de-identification problem.

But these mechanisms are not without their cost. Recent work has shown that de-identification of a data set can introduce statistical bias into that data, making the results extracted by analysis of the de-identified set vary significantly from those same analyses applied to the original set.

In this paper, we will look at how this bias is introduced when a particular form of de-identification, k-anonymity, is applied to a particular large data set generated by the Massive Open On-line Courses (MOOCs) offered by Harvard and MIT. We will discuss some of the tensions that arise between privacy and big-data science as a result of this bias, and look at some of the ways that have been proposed to avoid the trade-off between accurate science and privacy. Finally, we will outline a promising new approach to de-identification which appears to avoid much of the bias introduction, at least on the data set in question.

Keywords – big data; privacy; k-anonymity;

1 Introduction

Studies based on large-scale data sets and techniques from statistics, jointly labeled “big data science”, are beginning to make their appearance in the social sciences, medicine, and education. While these techniques have been used in the physical sciences for some time, their application in these new areas raise concerns about the privacy of the subjects of the studies. De-identifying the data sets so that those sets can be shared for further investigation or for verification of results has been the goal for these fields, but the goal has been particularly difficult to achieve in practice.

The first worry raised by de-identification attempts centered around the ease with which naively de-identified sets could be linked to other, easily accessible, datasets to re-identify the participants in the study. A number of studies have shown how surprisingly little data can be linked to outside data sets to re-identify individuals [1, 2]. These concerns led to a number of enhanced technical definitions of de-identification, including k-anonymity [3], l-diversity [4] and differential privacy [5]. Each of these frameworks attempts to provide a technical solution that allows sharing of data about human subjects without allowing (or at least making it very difficult to) re-identification of the individuals whose data is shared.

More recent work has raised a new concern about de-identified data sets. An early study of de-identifying the data sets for students of MOOCs offered by Harvard and MIT through the edX platform showed that de-identifying those sets so that they were 5-anonymous introduced significant statistical bias into the de-identified set [6]. Similar results have called into question the accuracy of data sets protected by differential privacy [7].

While these results are preliminary, they pose a dilemma for the researcher wanting to use big data techniques in the social sciences. Privacy requirements mean that the raw data used cannot be openly shared. But science requires both the ability for others to reproduce and check your results, and the ability for others to extend and enhance an analysis using the data that was used by others. But if de-identification introduces significant statistical bias into a data set, sharing of that de-identified set is a disservice to a field, as researchers using that data set will come to erroneous conclusions.

In what follows, we will look at some of the possible ways of slipping between the horns of this dilemma. We will look at the technical background of the problem, but also entertain the kinds of policy solutions that might be possible. We will end by describing a technical solution that offers some hope.

2 Context

In what follows, we will frame out discussion around the de-identification of a particular data set, the person-course dataset generated from the MOOC courses offered by Harvard and MIT through the edX platform during the years 2014 and 2015. The dataset contains 3,040,773 records. Each record records information about a single student's interaction with a single course. Information includes basic demographic information such as age, level of education, and gender, as well as information about the interaction of the student with the course including completion status, number of forum posts, performance on quizzes, and the like. Each record contains 132 fields for each student, although some of the fields may be empty.

We chose our de-identification standard based on the legal requirements as best we could understand them. We took these records to be educational records associated with Harvard and MIT; in the United States such records are governed by the Family Educational Rights to Privacy Act (FERPA). Like many laws, this one is less than clear on the privacy requirements, but the best interpretation we could find [8] required that prior to openly sharing the information, the data set needs to be de-identified using to a standard of k-anonymity, where k = 5. Note that we are not claiming that de-identifying the data set in this fashion will guarantee that no subject in the data set cannot be re-identified. We are simply saying that we have followed the legal rules set up to protect the privacy of the subjects.

We begin by removing all direct identifiers from the data set; these include such fields as name, address, and obvious directory information. Once this is done, the first step towards de-identification is to determine the set of quasi-identifiers in the data set. Quasi-identifiers are those entries in the data set that could be used to connect this data set with other data sources. Only a surprisingly small number of such quasi-identifiers are needed to re-identify a subject; in the classic case the combination of zip, birth date, and gender can be used to link a medical record in which directory information has been removed to a voter list that re-identifies the subject, as illustrated in Figure 1.

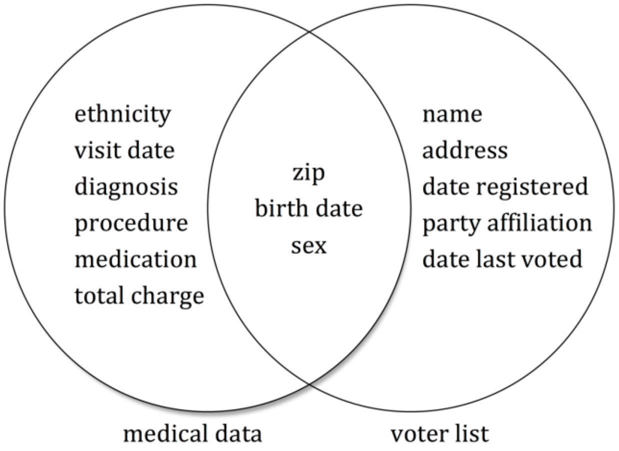

Figure 1. Combination of two data sets that allow re-identification (from [3]).

Of the 132 fields in the data set, only six fields were judged to be quasi-identifiers: the course itself, gender, country, year of birth, level of education, and number of forum posts (because the forums were public, and anyone could scrape the forums for any class). It would seem that it would be easy to find a mechanism to insure that, for any combination of these six identifiers, there would be at least 5 entries in the set that had identical values.

Two mechanisms are used to de-identify such a data set. The first, generalization, joins distinct values into a more general range to increase the number of records that report the range of values. For example, we could generalize the year of birth by giving a range of years, joining all the records for the multiple years into a single reported value.

The second mechanism used to achieve k-anonymity is suppression of particular records. If a record is difficult to generalize in such a way that it can be made k-anonymous, the record can be suppressed from the set.

Given the large number of records in the set and the small number of quasi-identifiers, the first mechanism used to produce a 5-anonymous data set favored suppression over generalization. The intuition was that, with such a small number of quasi-identifiers, the number of suppressed records would be small. We were surprised to find that this was not at all the case; in analysis of the resulting data set we discovered that approximately 20% of the records in the original set were suppressed. Moreover, the records that were suppressed tended to be the ones that the researchers found most interesting. There were lots of students who would sign up for a MOOC course and then do little or nothing else; from a k-anonymity point of view all of these students look the same and thus were retained in the de-identified set. Students who completed the course, on the other hand, had far more variation in their data (especially with respect to forum posts) and were thus more likely to be suppressed by the simple approach.

These anomalies, and some anecdotal evidence that this sort of thing had occurred in other, unrelated, data sets motivated a more detailed and directed study in the ways that data sets could be k-anonymized, and what the effects of those k-anonymization techniques were on the statistical properties of the set; the results of that work can be found in [9] and [10]. In summary, these works found that the mechanisms for k-anonymization, suppression and generalization, each introduced a different form of statistical bias. Suppression of individual records tended to bias the means of individual quasi-identifier values, while generalization tended to introduce bias in the correlation between the quasi-identifier values. The higher the level of generalization, the less suppression needed to be performed (although we found no level of generalization that completely eliminated the need for suppression). Our preliminary conclusion, at the end of this study, was that there was no way of reaching a level of 5-anonymity that did not distort the data to such a degree that sharing the data would be of scientific use.

3 The Tensions

If there is no way to de-identify data sets without introducing significant bias into those sets, we are presented with a dilemma when attempting to do big data science on human subjects. It is difficult to advocate that this sort of science should not rest on the ability to share the data on which conclusions are based, both to allow checking of results and to allow new questions to be addressed. But if sharing requires either degrading the utility of the data set or exposing the people represented in the set, we seem to be forced into a choice between good science and privacy.

One possible way out of this dilemma is to change the way in which we guarantee the privacy of the subjects whose data is contained in the set. Current notions of k-anonymity, l-diversity, or differential privacy all rest on the notion that privacy is preserved when the data does not allow the re-identification of the subjects. In effect, these approaches all assume that anonymity is the guarantor of privacy.

This connection between privacy and anonymity is, on investigation, not obvious *a priori*. One discussion of this can be found in the recent report from the President's Council of Advisors on Science and Technology report on big data and privacy [11]. This report notes that there are activities that are private but not anonymous (for example, voting) and also activities that are anonymous but not private (such as some political pamphlets). This report suggests that a more fruitful approach to privacy in the era of big data would rely on restrictions on how the data is *used* rather than on the form of the data itself.

Such an approach would change the interaction between big data and privacy in a fundamental way. Rather than trying to insure that no privacy violations could occur by any analysis of the data, this approach would try to audit, find, and punish those who violated the privacy of individuals by misusing the data. Identification or re-identification would be such a misuse. Rather than trying to avoid privacy violations before the fact, this approach would find privacy violations once they had occurred.

The emphasis on use has its own set of technical challenges. In particular, data sets would need to be marked or their provenance tracked in such a way that after-the-fact privacy violations could be detected. On the other hand, this approach would allow researchers to share the original data sets, with all of the information intact, meaning that there would be no statistical bias introduced.

But the real problem with this approach is not technical, but legal and political. Making such a change would require re-writing many of the regulations and laws that are used to protect human subjects in research. This may dismay many in the field of data science, as technical problems can be worked on at the speed of technology change, while policy changes occur at the speed of bureaucracy.

4 Some Technology Hope

As a technologist, I would be remiss if I did not end on a note of technological hope. While the current literature seems to indicate that de-identifying a data set will neither insure the impossibility of leaking private information nor allow accurate science to be done on that data set, there is some research that is showing some signs that a more careful approach to de-identification may avoid some of these pitfalls. Recent work by the author and his collaborators has re-visited the techniques used to achieve k-anonymity on the MOOC data set with some encouraging initial results [12].

On this approach, rather than simply generalizing numeric values into bins of a particular size to aid in k-anonymity, the binning has been done in a fashion that will adjust the bins to minimize the variation of each of the members in the bin to the mean value of all the members. To do this optimally is computationally infeasible, so we have instead relied on a greedy approximation that terminates in a reasonable amount of time.

As with more common forms of generalization, this led to a data set that was not fully 5-anonymous. Rather than suppressing the records that violated 5-anonymity, we added *chaff*. These are synthetic records having the same set of quasi-identifiers as those that needed to be made 5-anonymous, with the value of other fields being picked randomly from the values of those fields in the entire set.

We found that this combination of mechanisms gave us a 5-anonymous data set with almost no added statistical bias; those interested in the details should consult the original paper. Note also that these results are preliminary; they need to be generalized to see if they work as well with other, unrelated data sets.

Even if these results are generally useful, they do not guarantee privacy, only k-anonymity. But doing this without introducing statistical bias would itself be an advance in the state of the art, and we are hopeful that the future results will confirm our initial findings.

Acknowledgements

Much of the work referenced in this paper was undertaken in collaboration with Olivia Angiuli. The original paper, and many of the ideas, was shaped by Joe Blitzstein. Recent developments have been funded by a grant from the Harvard Initiative on Learning and Teaching.

References

1. Sweeney, L, Re-identification of De-identified Survey Data, Carnegie Mellon University, School of Computer Science, Data Privacy Laboratory, Technical Report. Pittsburgh: 2000.
2. Malin, Bradley, and Latanya Sweeney. "Re-identification of DNA through an automated linkage process." *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001.
3. Sweeney, L, Achieving K-Anonymity Privacy Protections Using Generalization and Suppression, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2000.
4. Machanavajjhala, Ashwin, et al. "l-diversity: Privacy beyond k-anonymity."*ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007): 3.
5. Dwork, Cynthia. "Differential privacy: A survey of results." *International Conference on Theory and Applications of Models of Computation*. Springer Berlin Heidelberg, 2008.
6. Daries, Jon P., et al. "Privacy, anonymity, and big data in the social sciences."*Communications of the ACM* 57.9 (2014): 56-63.
7. Bambauer, Jane, Krishnamurty Muralidhar, and Rathindra Sarathy. "Fool's gold: an illustrated critique of differential privacy." *Vand. J. Ent. & Tech. L.* 16 (2013): 701.
8. Young, Elise. "Educational privacy in the online classroom: FERPA, MOOCs, and the big data conundrum." *Harv. J. Law & Tec* 28 (2015): 549-593.
9. Angiuli, Olivia, Joe Blitzstein, and Jim Waldo. "How to de-identify your data. "*Communications of the ACM* 58.12 (2015): 48-55.
10. Angiuli, Olivia Marie. *The effect of quasi-identifier characteristics on statistical bias introduced by k-anonymization*. Diss. 2015.
11. President's Council of Advisors on Science and Technology, Big Data and Privacy: A technological Perspective, Executive Office of the President, May 2014 (https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf)
12. Angiuli, O. and Waldo, J, Statistical Tradeoffs between Generalization and Suppression in the De-Identification of Large-Scale Data Sets, 2016 IEEE 40th Annual Computer Software and Applications Conference, June, 2016.