PSI Tools: Building Replications Project

Clara Wang
Mentors: James Honaker and Marco Gaboardi

Summer 2016

1 Introduction

The Building Replications Project falls under the umbrella of a larger project, the Private data Sharing Interface (PSI) Tools Project, which seeks to create tools that allow social scientists to more easily access and analyze sensitive data. The PSI Tools Project achieves this goal by applying differential privacy algorithms to datasets, adding noise to the data so that individuals cannot be re-identified. This summer, my work consisted of two primary tasks: (1) creating a corpus of datasets for testing differential privacy algorithms, and (2) using differential privacy mechanisms to replicate existing social science studies and evaluate the results. In addition to this work, I also conducted research for the Berkman-Klein Center for Internet and Society at Harvard University. For this secondary project, I analyzed the privacy harms and controls of a research dataset using a framework developed by members of the PSI Tools Project.

2 Corpus of Datasets

2.1 Objectives

We collected studies and replicated their analyses in order to verify the findings, as well as establish a baseline of comparison for differentially private replications. Utilizing datasets from actual studies allowed us to consider the practical issues that may arise for social scientists when using the PSI tools. We were also able to better understand the types of sensitive data most common in social science research, such as survey responses and demographic information.

In order to test the existing differential privacy algorithms, we required datasets that met a number of criteria: (1) over 2000 observations, (2) consisted of near-private or sensitive data, and (3) were used in studies that include simple models of analysis. Specifically, univariate statistics (i.e. means, medians, histograms, etc.) or regression models with few independent variables.

2.2 Results

To find datasets that met the requisite criteria, I first looked at journals that had replication policies such as the American Journal of Political Science and Political Analysis. After sifting through a number of different political science journals, I then looked through the Harvard Dataverse website as well as the data repository for Yale's Institution for Social and Policy Studies (ISPS). [1]

I identified 24 datasets and studies that fulfilled our requirements and documented metadata about each one (e.g. the number of variables in the dataset, types of analysis used in the study, number of observations). [2] While many of the datasets I collected had replication code provided by the author, some lacked replication

[1] The Harvard Dataverse website URL is `https://dataverse.harvard.edu/`, and the Yale ISPS website URL is `http://isps.yale.edu/research/data`.

[2] The documentation is available in a Google Drive folder titled "Privacy Tools 2016 Replication".

code or the provided code was written in Stata. For these studies, I selected the ones that were the most promising for testing our differential privacy algorithms (e.g. studies that mainly used univariate statistics), and I wrote code in R to replicate the analyses. I produced or translated replication code in R for six studies:

1. Berkman and Plutzer. "Defeating Creationism in the Courtroom, But Not in the Classroom." *Science* 331(6016). January 2011.
2. Nyhan et al. "One Vote Out of Step? The Effects of Salient Roll Call Votes in the 2010 Election." *American Politics Research* 40(5). 2012.
3. Plutzer et al. "Climate confusion among U.S. teachers." *Science* 351(6274). February 2016.
4. Saenz and Barrera. "Findings from the 2005 College Student Survey (CSS): National Aggregates." *Higher Education Research Institute.* February 2007.
5. Pew Research Center. "Many in Emerging and Developing Nations Disconnected from Politics." *Pew Research.* December 2014.
6. Furia and Lucas. "Arab Muslim Attitudes Toward the West: Cultural, Social, and Political Explanations." *International Interactions* 34(2). June 2008.

Unfortunately I was unable to exactly replicate the last three studies listed above, as I was unsure of how the authors recoded some of the variables and how they chose which observations to exclude from their analyses. For the Plutzer et al. (2016) study, the replication code provided by the authors had a line of code to recode one observation, but it failed to make any changes. Thus, one of the observations is different between the dataset the authors used and the one I used for replication purposes. I also did not take the analytic weights into account when I replicated this study.

3 Differentially Private Replications

I successfully completed non-private and differentially private replications of three studies: Berkman and Plutzer (2011), Nyhan et al. (2012), and Plutzer et al. (2016). In addition to writing replication code in R for these studies, I also wrote functions in R to run multiple, differentially private replications using different epsilon values and evaluate the results against the original data using three metrics: mean squared error, root mean squared error, and mean absolute error. [3]

3.1 Berkman and Plutzer (2011)

The Berkman and Plutzer (2011) study examines the beliefs that American high school teachers have about evolution, and how they approach the topic of evolution versus creationism in the classroom. The authors found that a majority of teachers advocate neither evolution nor creationism in their lessons, and that teachers who have taken a course on evolution are more likely to advocate evolutionary biology to their students.

3.1.1 Histograms

As part of their analysis, Berkman and Plutzer examined how many teachers advocated evolution, creationism, or neither. They found that of 926 surveyed teachers, 257 advocated evolution, 118 advocated creationism, and 551 advocated neither of the two. I used the differential privacy algorithms for histograms to replicate this distribution, and I ran 500 iterations of the algorithm for each epsilon value from 0.01 to 1. The results of these differentially private replications are displayed in Figure 1.

As epsilon increases, the differentially private results converge around the true value, but smaller epsilons preserve more privacy. Thus, there is a trade-off between privacy and utility if a research wants to know the

[3] The replication code for these studies and the code for the functions are available on GitHub.

exact counts for each bin in a histogram. However in Berkman and Plutzer's study, the important question is what the values of the three counts are in relation to one another, and this relationship is preserved at lower epsilon values (i.e. 0.1). Hence, the utility of differentially private results is highly dependent on the question being asked of the data.

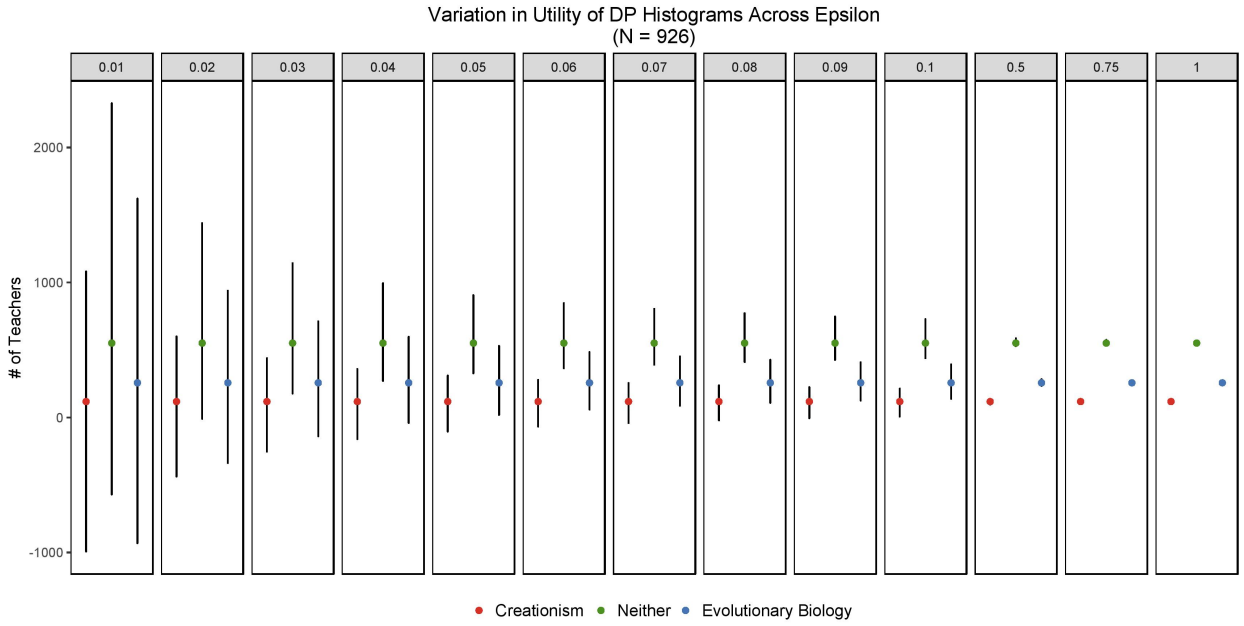


Figure 1: The colored dots in this figure show the true values for the number of teachers who advocate one of the three approaches for teaching evolution. The black lines show the range of values for the differentially private replications.

3.1.2 Subsetting

The authors also looked at the number of teachers who completed a course on evolution within each of the three "types" of teacher (i.e. advocate of evolution, creationism, or neither). I replicated a histogram showing the number of teachers advocating creationism who did and did not complete a course on evolutionary biology. Since only 118 teachers completed a course on evolutionary, the small N may have significantly impacted the accuracy of the differentially private replications. So, I tested four different methods for subsetting and compared their performance using RMSE and variance as metrics (See Figure 2).

The first subsetting method I tested was the standard method, where I used a dataframe that only contained the teachers who advocated creationism (N = 118) to produce differentially private histograms.

For my second subsetting method, I created an interaction term that equaled one if a teacher was both an advocate of creationism and took a course on evolution, and zero otherwise. This interaction term returned a count for the number of teachers advocating creationism who took a course on evolution. To get the number of teachers advocating creationism who did *not* take a course on evolution, I calculated the differentially private N for the total number of teachers advocating creationism, and subtracted the number of teachers advocating creationism who took an evolution course. The resulting value was the number of teachers who advocate creationism and did not take an evolutionary biology course. This method allowed me to preserve the full N of the dataset in the differentially private calculations, but it required two queries of the data and used more of the "privacy budget."

For my third subsetting method, I created a new variable that indicated whether a teacher did not ad-

vocate creationism and did not take a course on evolution (value = 0), advocates creationism and did not take a course (value = 0.5), does not advocate creationism and did take a course (value = 1), and advocates creationism and did take a course (value = 1.5). This method allowed me to preserve the full N of the dataset, and it only required one query of the data. It also returned extra information about the data (i.e. the number of teachers who do not advocate creationism and did not take a course on evolution, and the number of teachers who do not advocate creationism and took a course on evolution). [4]

For my fourth subsetting method, I created two new variables: one indicated whether a teacher was both an advocate of creationism and took a course on evolution, and another indicated whether a teacher was both an advocate of creationism and did *not* take a course on evolution. I used the counts from these two variables to create a differentially private histogram of course completion rates for teachers who advocate creationism. This method of subsetting used the entire N, but it required more of the privacy budget as I had to query the data twice.

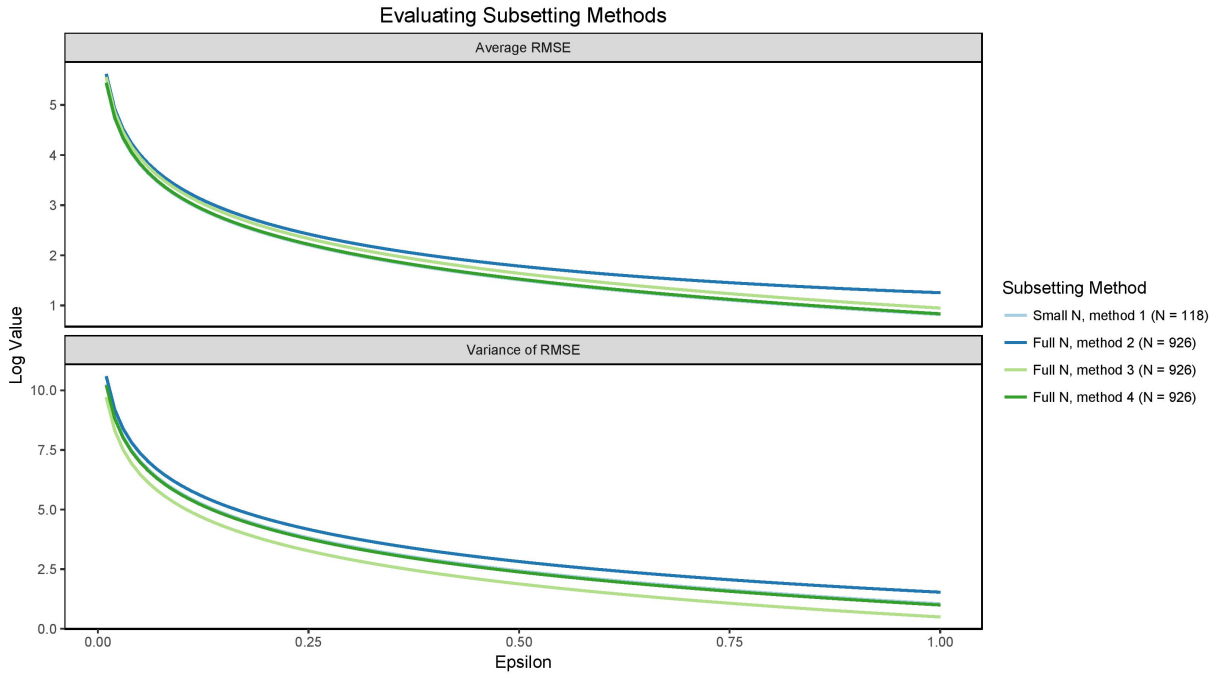


Figure 2: This figure shows the differences in average RMSE and variance of RMSE for the different subsetting methods. Method 1 seems to have similar performance to method 4. It should be noted that at lower epsilon values, the difference between the four methods is greater in magnitude than at the higher epsilon values. In this figure, the y-axis is on a log scale to better show the discrepancies at higher epsilons.

3.2 Nyhan et al. (2012) Replication

In this study, the authors examined whether Democratic incumbents in the 2010 midterm elections lost at the polls due to their support for health care reform. They found that voters perceived these Democratic incumbents to be more ideologically liberal than themselves, which led them to support other candidates in the 2010 elections.

[4] I created this variable by adding together 1/2 of a binary variable (values = 0 or 1) indicating whether a teacher advocated creationism, and another binary variable (values = 0 or 1) indicating whether a teacher took a course on evolution.

3.2.1 Means

To get an idea of how constituents perceived Democratic incumbents, the authors calculated the difference between the political ideology of a constituent and a constituent's perceived political ideology of a Democratic incumbent. They calculated the average difference in political ideology for self-identifying Democrats, independents, and Republicans. The results suggested that most Republicans and independents perceived Democratic incumbents to be more ideologically liberal than themselves.

I replicated the authors' calculations, and then I replicated the three averages using differential privacy mechanisms. Figure 3 shows the accuracy of the differentially private means using RMSE as a measure of error. The figure clearly demonstrates the effect that N has on the accuracy of a differentially private value, as on average, the differentially private means for independents (N = 2043) are less accurate than those of Democrats (N = 10692) and Republicans (8841). Thus, differential privacy may have limited utility for calculating means in datasets with few observations, and researchers should note that subsetting a dataset may yield less accurate differentially private values.

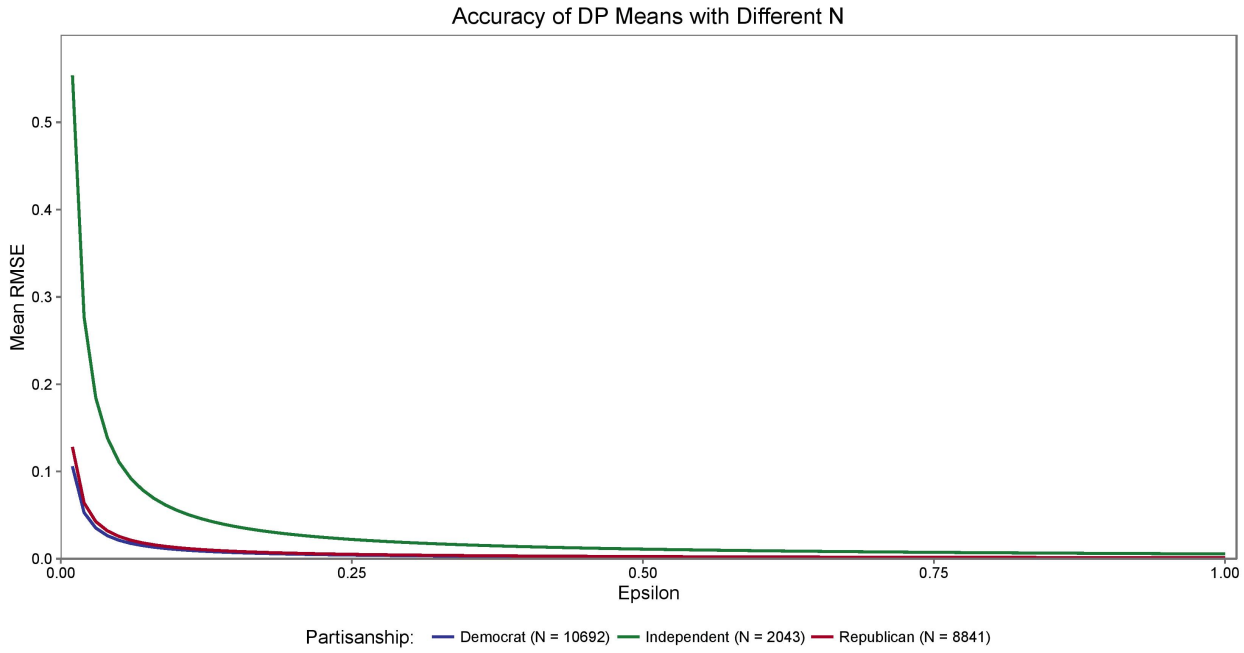


Figure 3: The three colored lines show the average root mean squared error (RMSE) for the mean difference in political ideology between Democratic incumbents and constituents, as grouped by political affiliation. The average error for independents is much greater than the other two populations, suggesting that the number of observations (N) influences the accuracy of differentially private values.

3.2.2 Histograms: Subsetting

The authors also examined the perceived ideologies of Democratic incumbents who voted in favor of health care reform in comparison to those who opposed it. They found that those who supported health care reform were perceived to be more liberal by their constituents.

I replicated one of the histograms that showed the distribution of perceived ideologies for Democratic incumbents who opposed health care reform. I then generated 500 differentially private versions of the histogram at different epsilon values ranging from 0.01 to 1. To generate the histograms, I subset the data to only include Democratic incumbents who opposed health care reform (N = 3285).

I also tried a different method of subsetting where I created an interaction term using a binary variable that indicated whether an incumbent voted for health care reform (0 = supported, 1 = opposed), and a categorical variable indicating the perceived ideology of an incumbent (values = 1 to 7). So, the interaction term only took on a value other than zero if the incumbent *opposed* health care reform - which was the population of interest. I then produced differentially private histograms of this interaction term, and I excluded the bin with zeroes from the final histogram so it only included Democratic incumbents who voted against health care reform.

To evaluate which subsetting method yielded the most accurate results, I calculated the differences in (1) average RMSE and (2) variance of RMSE between the differentially private histograms produced by each method (See Figure 4). Although the N was much smaller for the typical method for subsetting, the average RMSE and average variance of RMSE was approximately the same for differentially private histograms from both subsetting methods. However, the average RMSE was slightly greater for the smaller subsetting method, and the variance of RMSE was slightly less for the smaller subsetting method.

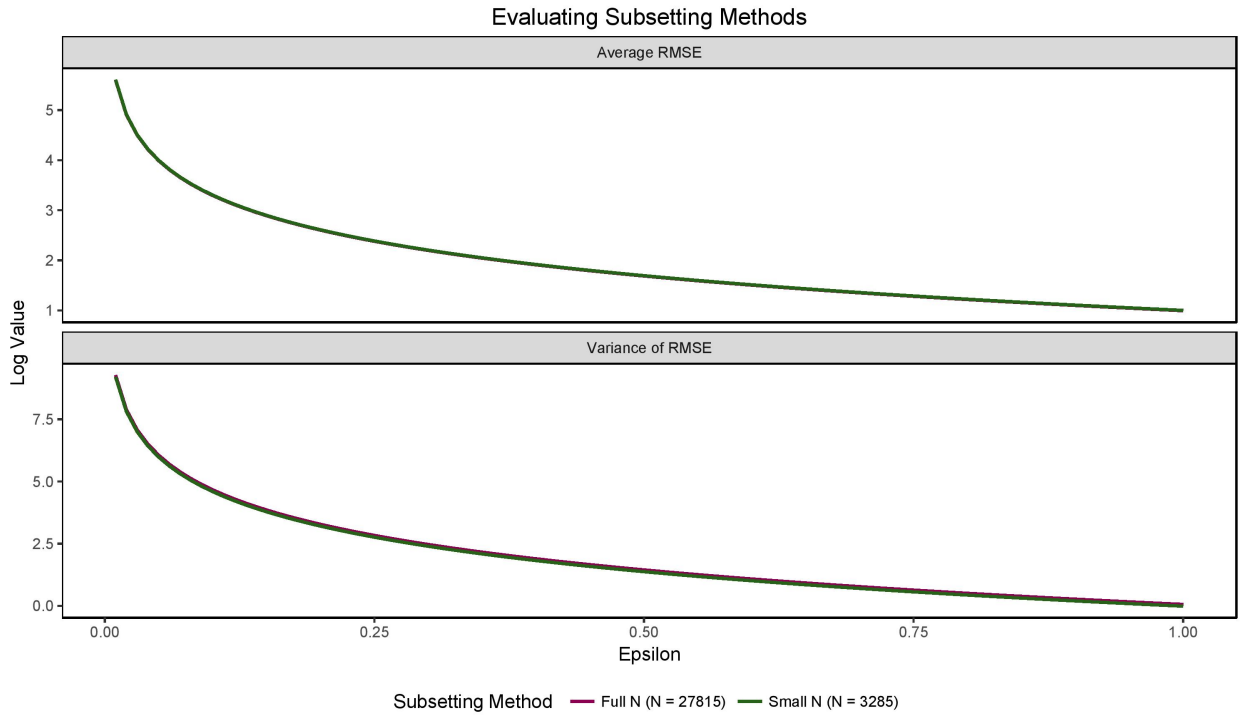


Figure 4: This figure compares the performance of two methods of subsetting: one that involves a smaller N and another that includes the full N of the dataset.

3.3 Plutzer et al. (2016) Replication

The authors of this study surveyed American public middle-school and high-school science teachers about their opinions on climate change and their approaches for teaching the subject in the classroom. They found that while many teachers spend time educating students about climate change, some teachers relay explicitly contradictory messages about the topic. These teachers tell their students that scientists believe climate change stems from both natural and human causes in equal parts, when in fact 81-100% of scientists think that global warming is primarily caused by humans.

3.3.1 Standard Deviation

One question in the survey of American public school teachers asked whether they were mostly pressured by fellow teachers to include global warming as part of their teaching curriculum. When coding the question, the authors used a binary variable to indicate whether a teacher reported "yes" (value = 1), or "no" (value = 0) to the question. The standard deviation of this variable was 0.1755259, indicating that the values were generally concentrated around the mean (0.0317997).

Figure 5 is a violin plot that demonstrates the relationship between epsilon and the accuracy of a differentially private value. Notably, some of the differentially private standard deviations are negative values, which is unrealistic as standard deviation cannot be negative. Thus, the differentially private values that take on negative values use all utility, as they cannot be true. However the current mechanism of adding LaPlace noise to preserve privacy allows for differentially private values to be negative, [5] and the low value of the standard deviation (0.17553) increases the likelihood that the differentially private value is negative, even at higher values of epsilon.

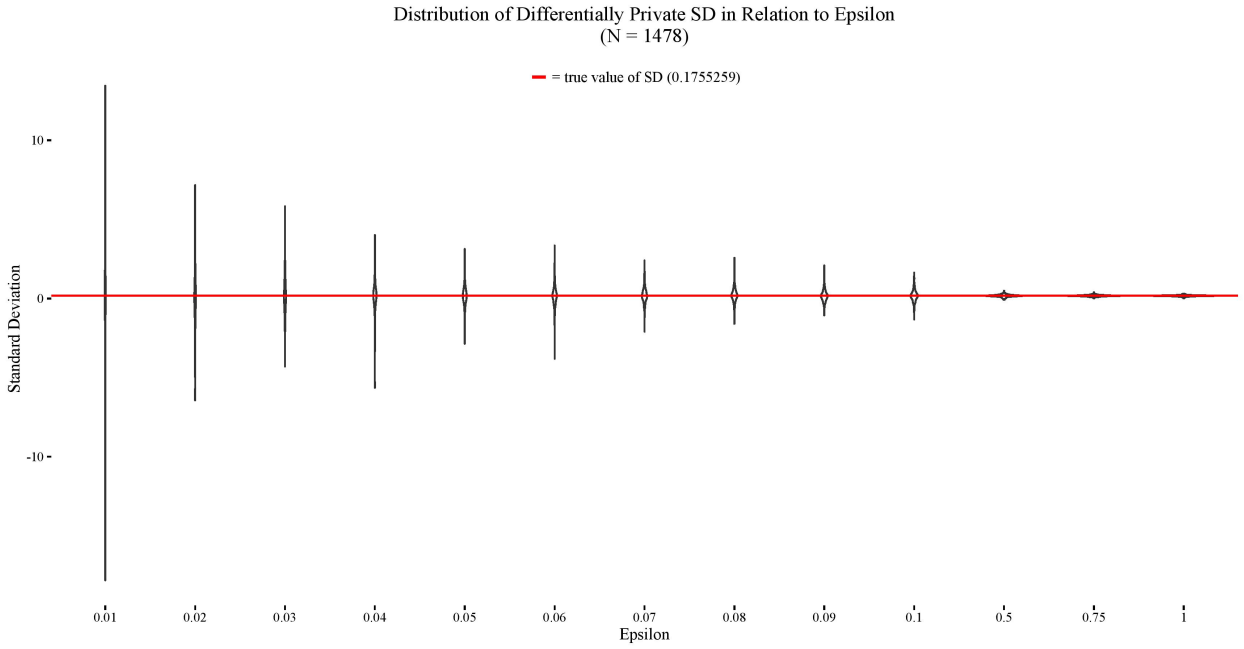


Figure 5: This violin plot shows the distribution of 500 differentially private standard deviations at each value of epsilon. As epsilon increases, the differentially private values converge around the value of the original dataset, as indicated by the red line.

3.3.2 Two-way Table

The authors also examined the relationship between teachers' personal beliefs about climate change and teachers' beliefs about scientific consensus on the topic. They visualized the information in a two-way table as shown in Table 1.

Table 2 shows a differentially private version of the original table. The differentially private values were calculated by taking the observations in each cell of the original table, adding LaPlace noise with a sensitivity of two, and rounding to the nearest whole number. [6]

[5] Code for differentially private standard deviation is in its first stages. See Appendix A.
[6] Code in Appendix A.

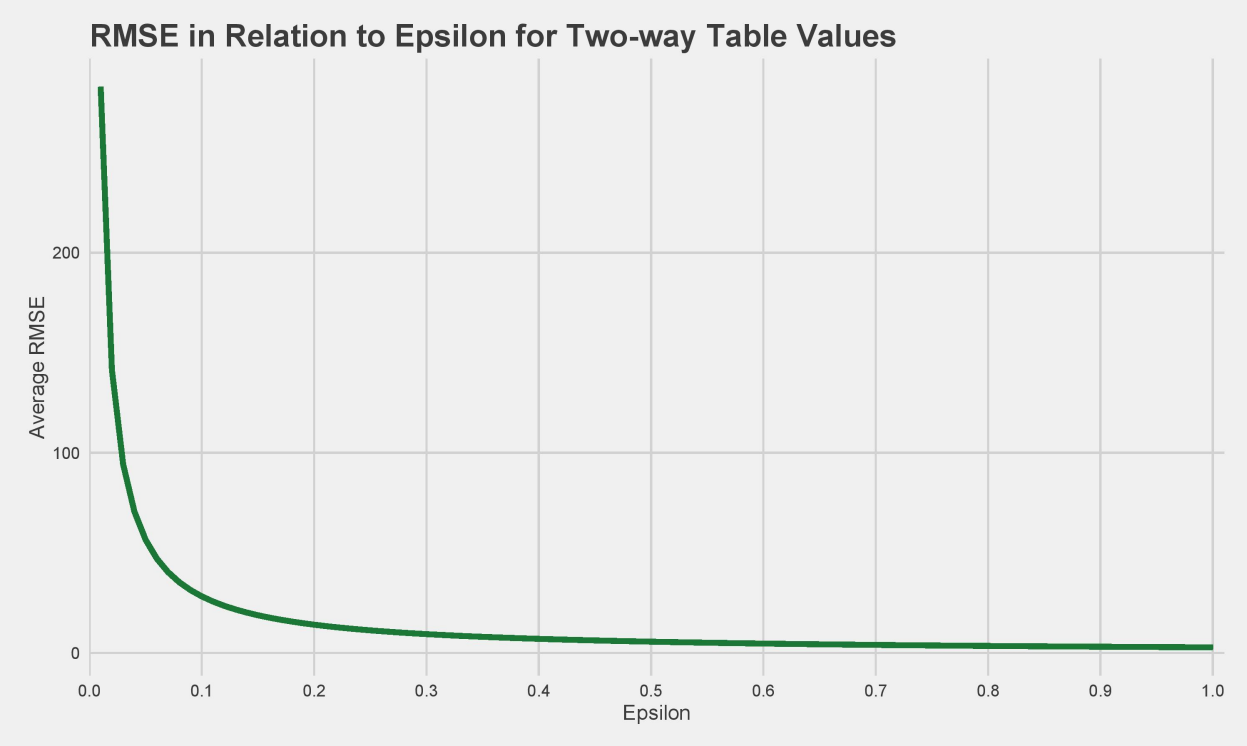
Table 1: Two-way table of perceived causes of climate change vs. estimated percentage of scientists who believe that climate change is primarily caused by humans.

	0-20%	21-40%	41-60%	61-80%	81-100%	Don't Know
Humans	5	23	44	199	506	139
Mix of Humans/Nature	2	19	23	44	24	51
Nature	16	38	56	38	20	71
Not Occurring	2	5	5	8	0	9
Other	0	0	0	0	0	1
NA	0	5	2	4	23	7

Table 2: A differentially private replication of Table 1.

	0-20%	21-40%	41-60%	61-80%	81-100%	Don't Know
Humans	5	19	93	201	498	138
Mix of Humans/Nature	2	15	30	46	16	50
Nature	16	34	63	40	12	70
Not Occurring	2	1	12	10	-8	8
Other	0	-4	7	2	-8	0
NA	0	1	9	6	15	6

Figure 6: This figure shows the relationship between RMSE and epsilon for the differential privacy algorithm used to produce two-way tables. 500 iterations of the algorithm were run at each epsilon value.



4 Discussion

My work this summer has resulted in two main outcomes. First, the corpus of datasets will allow other researchers on the project to test the PSI tools software on the types of datasets used by social scientists, who are the intended users of the software. Also, many of the studies include regression analyses that will be useful for testing differentially private regression code once it is finished.

The replications offer insight into the practical applications of the PSI tools software and some of the potential issues to address. First, some of the studies included weights in their analysis, which have yet to be integrated or addressed in the PSI tools differential privacy algorithms. Social scientists will likely be interested in incorporating such statistical methods in their analysis, so allowing for weighting in the PSI tools software will increase its usability. Additionally, the number of observations in a study influences the accuracy of differential private values. However, many social scientists choose to subset their data to examine trends within specific populations in the dataset. To allow for such subsetting queries in the PSI tools software, it would be valuable to identify the best methods for subsetting that preserve the entire N of the dataset when applying differential privacy algorithms.

Finally, this replications project helped highlight issues related to the utility of differential privacy. While I used metrics such as mean squared error, root mean squared error, mean absolute error, and variance of error to evaluate my differentially private replications, these metrics are not perfect measures of utility. For example, if a researcher's question is about the proportional relationship between three variables, the exact counts become less important and mean squared error is no longer the best measure of utility. However, if the researcher's question is about the exact count of a variable, then mean squared error is a great measure of utility. Subtle nuances in what a researcher is asking of the data can change whether or not a differentially private value has greater or less utility, and finding a quantitative metric for this qualitative issue may be a central question for future research in the PSI Tools Project.

5 Privacy Controls Case Study: "Long-term effect of September 11 on the political behavior of victims' families and neighbors" (Hersh 2013)

5.1 Introduction

In order to prevent researchers from using fraudulent data and ensure that a study is replicable, many scholars in the social sciences have called for the mandatory sharing of data.[7] However, a key concern is that much of the data collected and analyzed contains private information about individuals, such as their opinions on controversial topics (e.g. LGBTQ+ issues, beliefs about creationism, etc.). Thus, privacy concerns are important barriers to the sharing of data in social science research, as releasing datasets that contain sensitive information about individuals could result in significant harms for the individuals in that dataset. Assessing the potential harms that come from releasing social science data can help identify privacy issues that need to be addressed before researchers make their data available for others to access.

Eitan Hersh's study, "Long-term effect of September 11 on the political behavior of victims' families and neighbors" (2013) offers an interesting case study for analyzing privacy concerns for social science data.[8] He collected information about the political behavior of 9/11 victims' family members and friends by aggregating data from New York State voter files and the Federal Election Commission's (FEC) election donation database. Since this information is publicly available, his study qualified for exemption from Institutional Review Board (IRB) review. The replication data and code for the study is accessible on Yale University's Institution for Social and Policy Studies (ISPS) page.[9]

His study examined a somewhat sensitive topic – the political behavior of the families and close neighbors of 9/11 victims before and after the 9/11 attacks. By comparing the behavior of 9/11 victims' family and friends to people who lived in the same area, Hersh found that those close to the victims of 9/11 became more politically active after the attacks. They were more likely to become involved in politics by voting and donating to campaigns, and they were more likely to change their political affiliation to Republican. Hence, his study offered insight into the effects of 9/11 on Americans, and more generally, his study also provided information about the influence of terrorist attacks on the American people. As the author notes in his study: "Political changes among associates of victims are important, because system shocks like 9/11 can lead to rapid policy shifts, and relatives of victims often become leaders advocating for such shifts."[10]

To evaluate the potential privacy concerns in the Hersh (2013) study, I will utilize the framework proposed in Altman et al.'s paper, "Towards a Modern Approach to Privacy-Aware Government Data Releases."[11] This framework organizes the "data life cycle" into five stages: (1) Collection/Acceptance, (2) Transformation, (3) Retention, (4) Access/Release, and (5) Post-Access. In order to better capture Hersh's process, I adapted the framework by combining the first two stages.

5.2 Utility

As Hersh explicitly states in his paper, this study offers valuable insight into how 9/11 "catalyzed long-term changes in the political behaviors of victims' families and neighbors." He notes that "political changes among associates of victims are important because system shocks like 9/11 can lead to rapid policy shifts,

[7] Gary King, "Replication, Replication," Political Science & Politics 28, no. 3 (September 1995): 446, DOI: http://dx.doi.org/10.2307/420301.

[8] Eitan Hersh, "Long-term effect of September 11 on the political behavior of victims' families and neighbors." Proceedings of the National Academy of Sciences 110, no. 52 (2013): 20959-20963, DOI: 10.1073/pnas.1315043110.

[9] "Long-Term Effect of September 11 on the Political Behavior of Victims' Families and Neighbors," Yale ISPS, accessed June 29, 2016, `http://isps.yale.edu/research/data/d107#.V5-gqqLzPMs`.

[10] Hersh, 1.

[11] Micah Altman, Alexandra Wood, David R. O'Brien, Salil Vadhan, and Urs Gasser, "Towards a Modern Approach to Privacy-Aware Government Data Releases," Berkeley Tech. L.J. 30, no. 3 (2015), http://dx.doi.org/10.2139/ssrn.2779266.

and relatives of victims often become leaders advocating for such shifts."[12] Thus, this study adds to the base of scholarship on the effects of traumatic events on individuals, which can lead to major changes in public opinion and government policy.

5.3 Harms

Since the dataset focuses on a very specific population (i.e. 9/11 victims and their families and close neighbors), and the author offers a very detailed description of his methodology in his "Supporting Information" document, others may very easily re-identify subjects in the study. Connecting the political behavior information in the dataset with individuals could result in substantial harms. For example, a study by Iyengar and Westwood found that biases against members of the opposite political party are now stronger than racial biases.[13] Researchers conducted a survey experiment where they manipulated resume information to signal both race and political affiliation. Individual's names were used to signal race, while extracurricular activities such as "President of the Young Democrats" signaled political affiliation. They found biases against partisans to be even stronger than biases against African-Americans. This partisan animosity has steadily grown over time,[14] suggesting that the harms related to political behavior and partisanship may increase. Hence, it is possible that the individuals in this dataset may experience biases related to their political affiliation.

In a more abstract sense, other potential harms from this dataset arise from the aggregation of data to reveal new information about individuals. Daniel J. Solove, a Professor of Law at George Washington University, highlights the harms of aggregation, stating: "When analyzed, aggregated information can reveal new facts about a person that she did not expect would be known about her when the original, isolated data was collected." He notes that in the Information Age the harms of aggregation have become more pronounced, as "the data gathered about people is significantly more extensive, the process of combining it is much easier, and the computer technologies to analyze it are more sophisticated and powerful."[15]

Additionally, the release of the dataset in this study causes harms by increasing accessibility to data on individuals' voting behavior, and by exposing information about these individuals and drawing attention to them. Solove notes that increased accessibility to a dataset can increase the possibility of disclosure, or be exploited and used for purposes other than intended by the dataset's release. Although all the data used in Hersh's study is publicly available, by uploading the data as a cleaned data file on Yale's ISPS page, it is easily accessible and could be used for purposes other than research. The study also poses harms to the individuals in the dataset, as it exposes their political behavior, which they may consider to be embarrassing or humiliating. Exposure is different from disclosure in that it "rarely reveals any significant new information that can be used in the assessment of a person's character or personality. Exposure creates injury because we have developed social practices to conceal aspects of life that we find animal-like or disgusting. Further, in certain activities, we are vulnerable and weak, such as when we are nude or going to the bathroom."[16] In the case of Hersh's study, exposing a specific populations' political behavior in response to an event may cause significant social harms to the individuals in the dataset, as they are all residents or former residents of New York City – a generally liberal, Democratic area. Since the study reveals that the individuals in the dataset were more likely to be registered Republicans after the 9/11 attacks, they may experience social discomfort or bias from neighbors and people on the opposite end of the political spectrum.

[12] Hersh, 1.

[13] Shanto Iyengar and Sean J. Westwood, "Fear and Loathing across Party Lines: New Evidence on Group Polarization," American Journal of Political Science 59, no. 3 (2015): 690, https://pcl.stanford.edu/research/2015/iyengar-ajps-group-polarization.pdf.

[14] Shanto Iyengar, Gaurav Sood, and Yphtach Lelkes, "Affect, Not Ideology: A Social Identity Perspective on Polarization," Public Opinion Quarterly (2012): 1, DOI: 10.1093/poq/nfs038.

[15] Daniel J. Solove, "A Taxonomy of Privacy," University of Pennsylvania Law Review 154, no. 3 (January 2006): 506, http://ssrn.com/abstract=667622.

[16] Solove, 534.

5.4 Privacy Controls at the Collection/Acceptance and Transformation Stage

Hersh provides a thorough account of how he collected the data for his study, identifying the different sources of information and actors that received access to the data.[17] He first gained access to a database from Labels & Lists (now L2), which is a political data firm that aggregates information about individuals. The database Hersh obtained consisted of 9,995,513 records for all registered voters in New York State as of summer 2001, and these records included information about voters' registered political party, which primaries and elections they had voted in, their addresses, their names, their gender, and other information requested on voter registration forms.[18] Hersh then identified 9/11 victims who were registered voters and residents of New York. He began by finding the obituaries of 9/11 victims published by media outlets, and he collected information on the names, ages at times of death, employers, and city of residence from these sources.[19] With this information, he identified 9/11 victims in his New York State voter file by looking at first name, last name, city, and birth year. If an individual matched on all of these characteristics, he confirmed them as a match for a 9/11 victim. If Hersh was unable to match an individual from the obituaries to an individual in the voter file, he looked for the victims' family members, whose names were often listed in the obituaries as well. After identifying victims' family members by using their names, he was often able to find the victim by name as well. In other cases, Hersh accessed grave information from websites that allow individuals to upload such information.[20] From these websites, he obtained the exact birth dates of some of the 9/11 victims, and he used this information to distinguish between multiple potential matches for victims in the voter file. Through these matching methods, Hersh was able to identify 1,181 of 1,729 victims of the 9/11 attacks who were registered voters in New York state.

To identify the family members of 9/11 victims, he used a variable in the Labels & Lists dataset that grouped individual together who lived in the same household and were likely to be family members. He then identified close neighbors of 9/11 victims by using the address list in the voter database from Labels & Lists to narrow down the dataset to only include registered voters who lived in the same Census block and on the same street as a 9/11 victim in 2001. To find a victim's closest neighbors, Hersh restricted close neighbors to only include those who lived in the same apartment building, or in one of the 10 closest house numbers to a victim. So, he identified close neighbors by using geographic data and addresses in the voter file to pinpoint those living closest to a 9/11 victim.

To create a comparison group to evaluate the change in behavior for the family and close neighbors of 9/11 victims, Hersh used a combination of exact and distance matching to find up to five "control victims" who matched the 9/11 victims on a number of traits. He used the 2001 New York voter file from Labels & Lists to select the variables for exact matching: party affiliation, vote history in the 1998 and 2000 federal general elections, and indicator variable that represented whether the individual voted in a party primary between 1992 and 2000, sex, and state legislative district. For distance matching, Hersh used a statistical method known as Mahalanobis distance matching to find other "control victims." He used the same methodology to find "control neighbors" and "control family members" for the 9/11 victims as well.

He then shared identifying information about the individuals of interest (i.e. families, close neighbors, neighbors of 9/11 victims, and "control" individuals) with Catalist, another data vendor that collects and curates voter files that are often used by political campaigns. He contracted with Catalist to collect the voting information for the individuals of interest post-9/11.

Hersh then reached out to another political scientist, Adam Bonica, to access his cleaned database of campaign contribution data released by the Federal Election Commission (FEC). He matched the individuals in his dataset to those in the FEC data based on first name, last name, and zip code.

[17] Eitan Hersh, "Supporting Information," Proceedings of the National Academy of Sciences 110, no. 52 (2013): 1-7, `http://www.pnas.org/content/110/52/20959.abstract?tab=ds`.

[18] "L2," L2, accessed on July 23, 2016, `http://www.l2political.com/`.

[19] Hersh cites the New York Times and CNN in the supporting information for his journal article. See Hersh, "Supporting Information," 1.

[20] Hersh used two websites: (1) "Find A Grave," Newspapers.com, accessed on July 11, 2016, `http://www.findagrave.com/`, and (2) "Remember September 11, 2001," Legacy.com, accessed on July 11, 2016, `http://www.legacy.com/Sept11/Home.aspx`.

In summary, during this stage of the data life cycle Hersh collected information from a number of different sources: (1) Labels & Lists, (2) published obituaries from media outlets, (3) websites with information about deceased individuals, (4) Catalist, and (5) the FEC database of campaign contribution data. To collect data for his research, Hersh shared identifying information about specific individuals with Catalist.

Since all of the information that Hersh collected was public information, his study was exempt from IRB review.[21] IRBs are the main form of privacy control for the data collection/ acceptance stage of the life cycle, and they play a valuable role in protecting human research subjects. As stated in Yale's IRB policy document, the purpose of an IRB is to "review human subject research or activities regulated by the FDA to ensure all such research or activities conducted under the auspices of the University meets rigorous ethical standards and all applicable state, federal, and University requirements for the protection of human participants."[22] This study's exemption from IRB review suggests that the dataset was perceived to pose little to no harm towards the human subjects involved in the study.

5.5 Privacy Controls at the Retention Stage

At this stage of the data life cycle, the compiled data was retained by the author. Since Hersh conducted his research under Yale University's Human Research Protection Policy (HRPP), he was required by HRPP 1360.4 to ensure "adequate subject protection in the course of [his] interactions with subjects and/or [the] data."[23] It is likely that he stored the data on a computer at his academic institution (Yale University), meaning that Hersh had access to data encryption measures provided by Yale's IT services.[24] However, it is unknown whether or not Hersh utilized these services, as HRPP 1360.4 does not provide a concrete definition for what constitutes "adequate subject protection" in terms of data protection. It is possible that Hersh was not under very strict data protection requirements. The data used in his study came from publicly available sources, and his study was exempt from IRB review, suggesting that his study and data collection posed little to no harm for human subjects.

Copies of the data were also retained by the sources of each dataset (e.g. Labels & Lists, Catalist, etc.) and likely secured by their own data protection mechanisms.

[21] In Yale's "Policy 100 – IRB Review of Research Protocols," section 100.4 details the criteria for exemption from IRB review. It is probable that Hersh's study qualified for exemption based on the fourth criteria category, which exempts "Research not regulated by the FDA involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that participants cannot be identified, directly or through identifiers linked to the participants. 45 C.F.R. § 46.101(b)(4). Such research may not involve data from persons who are known to be currently imprisoned as participants or have been collected from incarcerated individuals under the exemption." For Yale's entire IRB policy document see: "Yale University Institutional Review Boards," Yale University, December 18, 2015, `http://www.yale.edu/hrpp/policies/documents/IRBPolicy100IRBreviewRevised-Final12-18-15.pdf`.

[22] "Yale University Institutional Review Boards," 1.

[23] "1360 Human Research Protection," Yale University, last modified on June 17, 2015, accessed on August 11, 2016, `http://your.yale.edu/policies-procedures/policies/1360-human-research-protection`.

[24] "Secure Computing," Yale Information Technology Services, accessed on August 11, 2016, `http://its.yale.edu/secure-computing/security-standards-and-guidance/data-and-application-security/protecting-yales-data/data-encryption`.

5.6 Privacy Controls at the Release and Access Stage

Prior to releasing the data to other parties, the author de-identified the data by removing the identification number from the dataset provided by Labels & Lists, and substituting it with a new variable that served the same purpose but included different values. It is unclear whether the author instituted any further measures to de-identify the data, but it is highly likely that he adhered to Yale ISPS's data archiving instructions and removed direct identifiers from his dataset for public release. This list of direct identifiers includes variables such as names, addresses and ZIP codes, social security numbers, etc.[25]

As the data was collected for a study that was later published in the PNAS journal, visualizations and analysis of the data are available in the online, published article. When Hersh submitted his study to PNAS for publication, the journal had policies in place that required authors to make their "materials, data, and associated protocols available to readers," and authors were "encouraged to deposit as much of their data as possible in publicly accessible databases."[26] Thus, Hersh was required to share his dataset in order to be published in PNAS.

Currently, Yale's ISPS has a copy of Hersh's data, replication code, and metadata stored and available for download on its website. The data is de-identified and was archived in 2014.[27] Prior to that date it is unclear how Hersh made his data publicly available (as required per PNAS editorial policy requirements). Privacy controls include de-identification of the data, recoding of variables, a terms of use agreement, and a number of other measures that trained data archivists implement.[28] The terms of use prohibit data requesters from using the data to identify individuals and collect information about them, and data requesters who want to redistribute the data or use them in a new data product or service must receive permission from ISPS.[29] A number of different online media outlets, such as *FiveThirtyEight* and the *Yale News*,[30] also published articles about the study that share the findings of the study, as well as the analysis of the data conducted by the author (e.g. regression results that model the effects of 9/11 on victims' families and close neighbors).

[25] "About the ISPS Data Archive," Yale ISPS, accessed on July 27, 2016, `http://isps.yale.edu/research/data/about#Access%20and%20Confidentiality`.

[26] "Editorial Policies," PNAS, accessed on August 11, 2016, `http://web.archive.org/web/20121223064228/http://www.pnas.org/site/authors/journal.xhtml`. NOTE: This editorial policy page was accessed using the wayback machine at https://archive.org/web/, which is a website that takes "snapshots" of web pages over time to create an archive. This webpage for PNAS's editorial policies was captured on December 23, 2012, which is a year prior to when Eitan Hersh's study was published by PNAS. Thus, these were the policies that Hersh had to operate under.

[27] "Long-Term Effect of September 11," Yale ISPS.

[28] "About the ISPS Data Archive."

[29] "Terms of Use," Yale ISPS, May 2015, `http://isps.yale.edu/research/data/terms-of-use#.V6VzG6LzPMs`.

[30] A number of media outlets published articles about the study, such as (1) Dan Hopkins, "The Enduring Political Impact of 9/11 For Those Who Were Closest," FiveThirtyEight, September 11, 2014, `http://fivethirtyeight.com/features/the-enduring-political-impact-of-911-for-those-who-were-closest/`, (2) Helen Dodson, "Families of 9/11 victims more politically active – and more Republican," Yale News, December 9, 2013, `http://news.yale.edu/2013/12/09/families-911-victims-more-politically-active-and-more-republican`, and (3) Jim Fitzgerald, "Yale study finds 9/11 kin in NY more political, a bit more Republican," Deseret News, December 9, 2013, `http://beta.deseretnews.com/article/765643137/Yale-study-finds-911-kin-in-NY-more-political.html`.

5.7 Privacy Controls at the Post-Access Stage

At this stage of the data life cycle, few privacy controls are in place. As the paper containing data visualizations and analysis are disseminated among a number of easily accessible, online sources (e.g. news sites, the author's website,[31] etc.), many individuals can collect information about the data.

The actual dataset and replication code is protected by Yale ISPS's terms of use, which prohibits data requesters from utilizing the data in an unlawful manner, or from using the data for re-identification purposes. However these privacy controls are quite weak, as the terms of use fail to specify concrete punishments that will deter data requesters from misusing the dataset. Furthermore, much of the terms of use seems concerned with protecting Yale ISPS from legal and financial harms (e.g. points 4, 7, 9).[32]

5.8 Discussion

At each stage of the data life cycle, the data is constantly vulnerable to access by other parties, as all the data is publicly available and can be aggregated by any individual in the same manner as Eitan Hersh. Thus, even though privacy controls are present throughout the data life cycle, they only protect the specific dataset that Hersh has on file. They do not protect the actual data.

Hence, the most notable feature of this case study is how easily the researcher was able to access, analyze, and publish information about the data without obtaining the consent of the individuals in the dataset or even notifying them. The primary purpose of IRBs is to protect human subjects from harms and to preserve certain ethical standards. The Federal Policy for the Protection of Human Subjects (AKA the "Common Rule") is the federal law that requires research on human beings to be subject to IRB oversight. As stated on the webpage for the U.S. Department of Health and Human Services, this policy system for protecting human research subjects "is heavily influenced by the Belmont Report,"[33] which states: "Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied."[34] However, it seems that none of the human subjects in Hersh's study were asked for their consent to participate in the study.

As more and more services move to electronic platforms and connect to social media, our information is increasingly at risk of being used for purposes we may not expect. This case study is just one example of how online, publicly available information may be used, and it raises a number of concerns about IRBs and whether or not they offer adequate protections for human research subjects. It also highlights the potential harms of information that is released as a matter of public record, like voter registration data and election campaign donation information. Putting such data in the public record serves a valuable purpose, as understanding whose finances may influence a politician is important for preserving the integrity of America's democracy. However, when this data is aggregated and analyzed, there may be greater harms to individuals in the dataset. Unfortunately, privacy protection measures such as differential privacy have limited efficacy in such cases, as the utility of public records such as campaign donation information comes from the ability to identify specific individuals in the dataset. Differential privacy would make it difficult or impossible to re-identify individuals.

Thus, no easy solutions to the issues posed by this case study exist. Nevertheless, the potential harms and risks to privacy presented by this study suggest that the issues require greater attention from researchers, lawyers, and government officials.

[31] "Research," Eitan D. Hersh, accessed on July 23, 2016, `http://www.eitanhersh.com/research.html`.

[32] "Terms of Use."

[33] "Federal Policy for the Protection of Human Subjects ('Common Rule')," U.S. Department of Health and Human Services, accessed on August 12, 2016, `http://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html`.

[34] "The Belmont Report," U.S. Department of Health and Human Services, accessed on August 12, 2016, `http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html`.

6 Works Cited

"About the ISPS Data Archive." *Yale ISPS.* Accessed on July 27, 2016. http://isps.yale.edu/research/data/about.

Altman, Micah, Alexandra Wood, David R. O'Brien, Salil Vadhan, and Urs Gasser. "Towards a Modern Approach to Privacy-Aware Government Data Releases." *Berkeley Tech. L.J.* 30, no. 3 (2015). http://dx.doi.org/10.2139/ssrn.2779266.

"The Belmont Report." *U.S. Department of Health and Human Services.* Accessed on August 12, 2016. http://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html.

Dodson, Helen. "Families of 9/11 victims more politically active – and more Republican." *Yale News.* December 9, 2013. http://news.yale.edu/2013/12/09/families-911-victims-more-politically-active-and-more-

"Federal Policy for the Protection of Human Subjects ('Common Rule')." *U.S. Department of Health and Human Services.* Accessed on August 12, 2016. http://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.

"Find A Grave." *Newspapers.com.* Accessed on July 11, 2016. http://www.findagrave.com/.

Fitzgerald, Jim. "Yale study finds 9/11 kin in NY more political, a bit more Republican." *Deseret News.* December 9, 2013. http://beta.deseretnews.com/article/765643137/Yale-study-finds-911-kin-in-NY-more-p
html.

Hersh, Eitan. "Long-term effect of September 11 on the political behavior of victims' families and neighbors." *Proceedings of the National Academy of Sciences* 110, no. 52 (2013): 20959-20963. DOI: 10.1073/pnas.1315043110.

Hersh, Eitan. "Supporting Information." *Proceedings of the National Academy of Sciences* 110, no. 52 (2013): 1-7. http://www.pnas.org/content/110/52/20959.abstract?tab=ds.

Hopkins, Dan. "The Enduring Political Impact of 9/11 For Those Who Were Closest." *FiveThirtyEight.* September 11, 2014. http://fivethirtyeight.com/features/the-enduring-political-impact-of-911-for-th

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. "Affect, Not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* (2012): 1-27. DOI: 10.1093/poq/nfs038.

Iyengar, Shanto and Sean J. Westwood. "Fear and Loathing across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59, no. 3 (2015): 690-707. https://pcl.stanford.edu/research/2015/iyengar-ajps-group-polarization.pdf.

King, Gary. "Replication, Replication." *Political Science & Politics* 28, no. 3 (September 1995): 444-452. DOI: http://dx.doi.org/10.2307/420301.

"L2." *L2.* Accessed on July 23, 2016. http://www.l2political.com/.

"Long-Term Effect of September 11 on the Political Behavior of Victims' Families and Neighbors." *Yale ISPS.* Accessed June 29, 2016. http://isps.yale.edu/research/data/d107#.V5-gqqLzPMs.

"1360 Human Research Protection." *Yale University.* Last modified on June 17, 2015. Accessed on August 11, 2016. http://your.yale.edu/policies-procedures/policies/1360-human-research-protection.

"Remember September 11, 2001." *Legacy.com.* Accessed on July 11, 2016. http://www.legacy.com/Sept11/Home.aspx.

"Research." *Eitan D. Hersh.* Accessed on July 23, 2016. http://www.eitanhersh.com/research.html.

"Secure Computing." *Yale Information Technology Services.* Accessed on August 11, 2016. http://its.yale.edu/secure-computing/security-standards-and-guidance/data-and-application-security/protecting-yales-data/data-encryption.

Solove, Daniel J. "A Taxonomy of Privacy." *University of Pennsylvania Law Review* 154, no. 3 (January 2006): 477-560. http://ssrn.com/abstract=667622.

"Terms of Use." *Yale ISPS.* May 2015. http://isps.yale.edu/research/data/terms-of-use#.V6VzG6LzPMs.

"Yale University Institutional Review Boards." *Yale University.* December 18, 2015. http://www.yale.edu/hrpp/policies/documents/IRBPolicy100IRBreviewRevised-Final12-18-15.pdf.

7 Appendix A

Code in multipleDP_functions.R

```
## ------------------------< DP STANDARD DEVIATION FUNCTIONS >------------------------
## Simple implementations of differentially private variances and standard deviation
## Honaker, July 7, 2016

# function to calculate LaPlace noise
rlaplacejames<-function(sensitivity, eps){
  flip <- rbinom(n=1,size=1,prob=0.5)
  rlaplace <- rexp(n=1, rate=eps/sensitivity) * (2*flip -1)
  return(rlaplace)
}


## x - variable as vector
## lb - lower bound
## ub - upper bound
## epsilon - epsilon for differential privacy

## function to calculate DP variance
dp.var<-function(x, lb, ub, eps){
  n <- length(x)
  var.sens <- ( (n -1)/ n^2 ) * ((ub - lb)^2)
  var.sample <- var(x)
  var.release <- var.sample + rlaplacejames(sensitivity=var.sens, epsilon=epsilon)

  return(var.release)
}


## function to calculate DP standard deviation
dp.sd <-function(x, lb, ub, eps, na.rm = FALSE){
  if(na.rm == TRUE){
    x <- x[!is.na(x)]
  }
  n <- length(x)
  sd.sens <- (sqrt(n - 1) / n) * (ub - lb)
  sd.sample <- sd(x)
  sd.release <- sd.sample + rlaplacejames(sensitivity = sd.sens, eps = eps)
  return(sd.release)
}



## ------------------------< DP TWO-WAY TABLES >------------------------
## add LaPlace noise to an observation
dp.val <- function(x, sensitivity = 2, eps){
  val.release <- x + rlaplacejames(sensitivity = sensitivity, eps = eps)
  return(val.release)
}
```