Causal Inference & Differential Privacy

REU 2016 Final Paper

Grace Rehaut

August 12, 2016

I. Introduction

Causal inference is a large and immersive field within the realm of statistics, concerning itself with making statements and conclusions about things, events, stimuli, etc. that have not actually happened. It therefore seeks to understand counterfactual occurrences. What would happen to a patient if he or she did not take his medication? Would unemployment rise or fall if that conservative fiscal bill was not passed? Unfortunately, researchers may never obtain exact answers to these counterfactual questions, thus giving rise to what is often referred to as the **fundamental problem of causal inference**: we will never be able to observe one observation or situation experience treatment and not-treatment at the same time.

Exploring these counterfactual outcomes, however, is important, providing a way to remove bias from causal estimates; in light of this, causal inference has come to incorporate various techniques and methodologies that allow researchers to incorporate useful logic about counterfactual possibilities and thereby answer questions about real-life trends and situations. Using statistical techniques such as difference-in-means and difference-in-differences, researchers have effectively come to bypass the fundamental problem of causal inference in order to make valuable conclusions about the world that we live in.

My project this summer has been to ensure that the Privacy Tools for Sharing Research Data group is doing all that it can to promote and employ causal inference, from creating new differentially private causal algorithms to testing the efficiency of the private causal inference methodologies that the Privacy Tools group has already developed. These goals, important and worthwhile, have provided me with the foundation for a full summer of fruitful and educational statistical research, the substance of which I will explain in the following report.

II. Overview of Replication Work

My primary task this summer was replication work, a process of perusing databases for interesting quantitative studies, replicating those studies' findings using non-privacy protecting tools, replicating those studies' findings using differentially private tools, and then analyzing the differences and similarities between the two. By browsing the Harvard Dataverse and other online data repositories, including the website of the Institute of Social and Policy Studies (ISPS) at Yale University and the website for the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan, I was able to find several useful studies which satisfied all my critiera, i.e. having accessible datasets containing over 2,000 observations, simple models, and replicable data analysis. In particular, I sought out studies employing causal inference methodologies so that, in replicating those studies using the Privacy Tools algorithms, I could make a study of how well the algorithms were performing.

Replication work is crucial to the Privacy Tools project at this stage as it provides a valuable opportunity to test the differentially private algorithms that the Privacy Tools researchers have constructed, analyze how exact and efficient those algorithms are, and find ways to improve the substance of the algorithms. Successful replications may also serve as an effective advertisement for the work that Privacy Tools does, a major motivating factor behind my project. By showing that other researchers' results can indeed be closely replicated using differentially private tools, I have sought to demonstrate the validity of Privacy Tools' differentially private approach, hoping to prove that seeking privacy need not entail a loss of time or data utility.

i. Replication Strategy

In conducting my replication work, I followed a well-defined strategy that allowed me to maximize the number of useful studies I could find in a minimal amount of time. I began my search by perusing journals with mandatory replication policies, those that require authors to make their datasets available for replication work such as mine. An excellent example is the American Journal of Political Science (AJPS), which I consulted extensively in my search for appropriate replication studies. AJPS requires that all datasets used in the studies that it publishes be made available in the Harvard Dataverse, rendering them easy for me to access and analyze. I ultimately found success by reading articles in journals such as AJPS and others like it; bookmarking those that employed simple, causal models in their analysis; locating the relevant datasets for the articles within the Harvard Dataverse and on other sites, such as those mentioned above; and checking that those datasets contained appropriate numbers of observations, easily coded variables, near-sensitive data, etc. Each time that I located an article fitting these criteria, I began the process of making sure that I could replicate the researchers' results using basic, non-privacy protecting approaches. After confirming my ability to reproduce those original results, I moved on to the next step: replication with the privacy tools algorithms. The results of this step were the most pertinent to my analysis and overall project.

Over the course of the summer, I located around twenty studies that satisfied the replication criteria I have been looking for. Of those, I focused on four studies in particular, all of which I was able to successfully replicate using standard non-privacy protecting analysis. I went on to successfully replicate each of those four studies using the differentially private algorithms as well. These four studies will form the main analytic component of this paper, and they are listed in **Figure 1** below.

Author	Title	N	V	Analysis
McClendon	Social Esteem and Participation in Contentious Politics: A Field Experiment at an LGBT Pride Rally	3647	17	ATE
Butler	Black Politicans Are More Intrinsically Motivated to Advance Blacks' Interests: A Field Experiment Manipulating Political Incentives	5593	45	DID
Gerber	Can Incarcerated Felons Be (Re)integrated into the Political System? Results from a Field Experiment	6441	14	DID
Winters	Can Citizens Discern? Information Credibility, Political Sophistication, and the Punishment of Corruption in Brazil	2001	10	ATE

Figure 1: Studies that I have succeeded in replicating this summer

The studies all have fairly large numbers of observations (all over 2000) and employ simple causal analysis, which I was able to replicate both privately and in a non-privacy protecting manner. The results of those replications will be detailed at length in this report.

ii. Replication Dataset Corpus

It bears mentioning that a useful byproduct of my summer's research is a neat and extensive corpus of replication datasets, which, though an unintended outcome of my project, may actually prove useful for other Privacy Tools team members in the future. The corpus incorporates information about over 50 datasets, including title, author, number of variables, number of observations, and type of analysis. An adjunct dataset documentation sheet has been created as well, including, for each dataset, a short summary of the study in which the data was employed, a BibTex citation for that study, information about where to locate relevant files, etc. Images of both the datasheet corpus and its associated documentation may be found in **Figures 2 and 3** below.

My hope is that this replication dataset corpus will be able to find use in the future projects of other members of the Privacy Tools group, whether for purposes of replication specifically or for related tests and examples.

Author	Title	Journal	N	V
Goldin	The Historical Evolution of Female Earnings and Occupations	NBER	936	27
Christenson	The Factors Shaping Public Support for Unilateral Action	AJPS	1000	14
Winters et al	Can Citizens Discern? Information Credibility, Political Sophistication, and the Punishment of Corruption in Brazil	UVA Journal of Politics	2001	56
Pew	Indians Want Political Change: Modi Viewed More Favorably than Ghandi	Pew Research Center	2464	138
Hanmer et al	Experiments to Reduce the Over-reporting of Voting: A Pipeline to the Truth	Political Analysis	2517	39
Pew	Asian Americans: A Mosaic of Faiths	Pew Research Center	3511	268
Broockman et	Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators	AJPS	4859	15
Broockman	Black Politicians Are More Intrinsically Motivated To Advance Blacks' Interests: A Field Experiment Manipulating Political Incentives	AJPS	5593	45
Pew	A Wider Ideological Gap Between More and Less Educated Adults	Pew Research Center	6004	208
Gerber et al.	Can Incarcerated Felons Be (Re)integrated into the Political System? Results from a Field Experiment	AJPS	6441	14
Barnes et al	Making Space for Women: Explaining Citizen Support for Legislative Gender Quotas in Latin America	Journal of Politics	17083	25
Fraga	Candidates or Districts? Reevaluating the Role of Race in Voter Turnout	AJPS	17,140	14
Panagopoulos	Timing Is Everything? Primacy and Recency Effects in Voter Mobilization Campaigns	Political Behavior	25000	20
Persson	Does Survey Participation Increase Voter Turnout? Re-examining the Hawthorne Effect in the Swedish National Election Studies	AJPS	31588	18
Gerber et al.	Do Robotic Calls from Credible Sources Influence Voter Turnout or Vote Choice? Evidence from a Randomized Field Experiment	JPM	55230	2469

Figure 2: Subset of the replication dataset corpus

XXVII. Keele & Minozzi (2013)
How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data

JOURNAL: *Political Analysis*
SUMMARY: Considering the states of Minnesota and Wisconsin, where election-day registration has been recently implemented, the authors seek to understand if such registration has an effect on election turnout overall. Employing an average treatment effect design, the authors find that election-day registration does not in fact increase turnout, producing no significant change in the turnout rates in these two states.
BIBTEX CITATION:

```
@article{keele2013much,
  title={How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data},
  author={Keele, Luke and Minozzi, William},
  journal={Political Analysis},
  pages={mps041},
  year={2013},
  publisher={SPM-PMSAPSA}
}
```

DATA:
m) Stored within Replication Files.zip (available on Dataverse)
CODE:
l) Stored within Replication Files.zip (available on Dataverse)

Figure 3: Sample documentation for the replication dataset corpus

III. Replication 1: Butler & Broockman (2011)

As my first demonstration of the utility and functionality of differentially private causal inference algorithms, we here replicate a section of analysis from Daniel Butler and David Broockman's 2011 field study on political behaviors among state legislators, using publicly available data drawn from the Harvard Dataverse. The goal of the study is to investigate the motivations behind legislators' political behaviors; to measure these motivations, the researchers conducted a field experiment wherein they sent advice-seeking emails to 4,859 United States legislators from either the putatively black alias of "DeShawn Jackson" or the putatively white alias of "Jake Mueller." The researchers further differentiated the purported political affiliation of the sender by manipulating whether the sender sought information about voting in the Democratic or Republican primary. Ultimately, by employing a difference-in-means design between emails from DeShawn Jackson and emails from Jake Mueller, the researchers found that only minority Democratic legislators were more likely to respond to emails from DeShawn Jackson than from Jake Mueller; all other categories (white Democratic legislators, white Republican legislators, and minority Republicans legislators) were more likely to respond to emails from Jake Mueller. Based on these conclusions, the researchers suggest that racial biases and backgrounds may play a larger role than formerly realized in shaping legislators' behaviors.

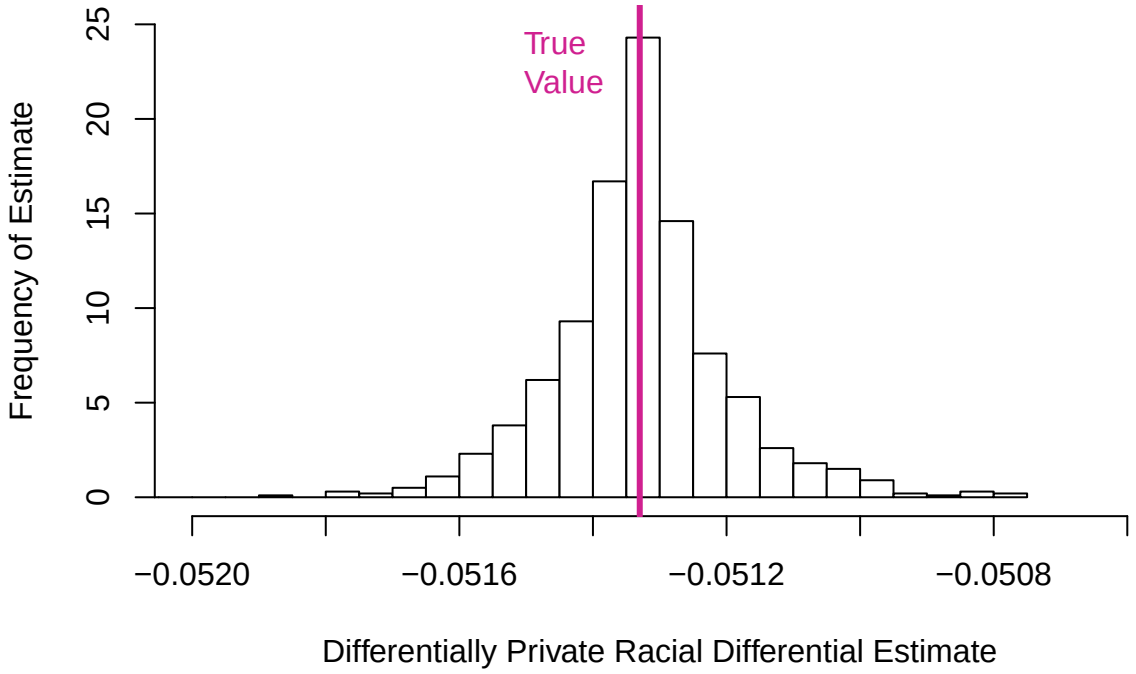
The dependent variable in the study is whether legislators responded to the emails they received in a meaningful way, and the independent variables are the putative race and political affiliation of each email's sender. Despite the fact that some of this information is public, as the individuals in the study are indeed public officials, the legislators' private behaviors and interactions with constituents may be considered sensitive materials, and as such, they represent just one variety of the kind of private, personal data that Privacy Tools' Private data-Sharing Interface (PSI) software might aim to protect. Consequently, we here imagine what kinds of valuable conclusions could be drawn from this study using PSI alone, employing only differentially private algorithms to produce the original conclusions that the researchers obtained through non-privacy protecting means.

We begin by replicating one of the most important and essential results of Butler and Broockman's original study: the result that legislators overall are more likely to respond to emails from the putatively white alias than emails from the putatively black alias (among emails with no partisanship signal). The researchers found this to be a strong conclusion, learning that legislators were 0.05133 or 5.1% more likely to respond to an email from "Jake Mueller" than to an email from "DeShawn Jackson"; with a p-value of 0.04, this analysis is significant at the 95% significance level.

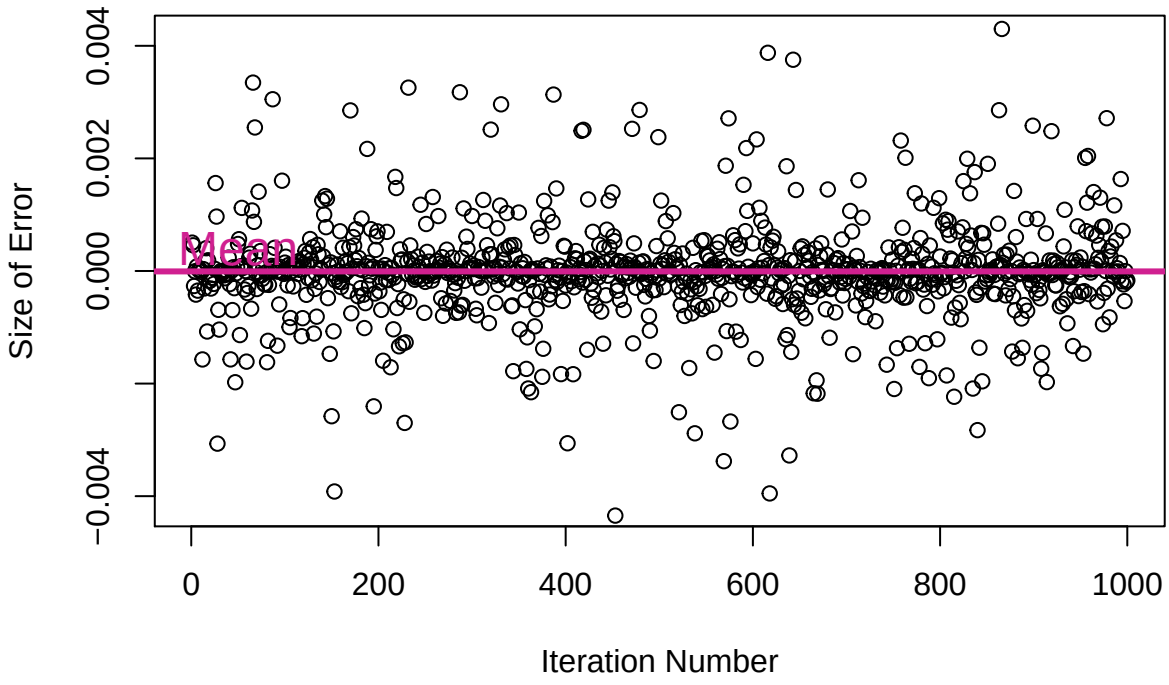
In **Plot 1** below, we replicate this difference-in-means analysis between rates of response to DeShawn Jackson and Jake Mueller (among all legislators) by running the researchers' original analysis with the Privacy Tools differentially private algorithms. For purposes of avoiding bias and inaccuracy, we run this analysis with the private algorithms 1,000 times, using a loop constructed in R. Again, the true value of this analysis, which was reported by the original researchers and which I was able to replicate myself as well, is 0.05133, derived by subtracting the rate of response to DeShawn Jackson from the rate of response to Jake Mueller. As may be seen in the histogram of the results below, across 1000 runs of the differentially private algorithms, the values that are returned cluster closely around this true value, designated by the pink line in the figure. The shape of the histogram, as may be noted, also neatly approximates a LaPlacian distribution. This result is both logical and supportive of the success of the differentially private algorithms; since they function by adding a small amount of noise derived from a LaPlace distribution to numerical releases, the fact that the histogram looks LaPlacian in nature is a positive sign that the differentially private algorithms are functioning as they were intended to do.

Another useful product of the analysis demonstrated in **Plot 1** is the 90% range provided by the differentially private algorithms over 1,000 simulations. In the same 1,000 runs of these algorithms shown in the figure, 90% of the results fall within the range of [0.0512,0.0515]; with the true value in this case being 0.0513, this analysis points to the conclusion that the differentially private algorithms are mostly producing substantively accurate results, an excellent indicator of the prowess of the algorithms.

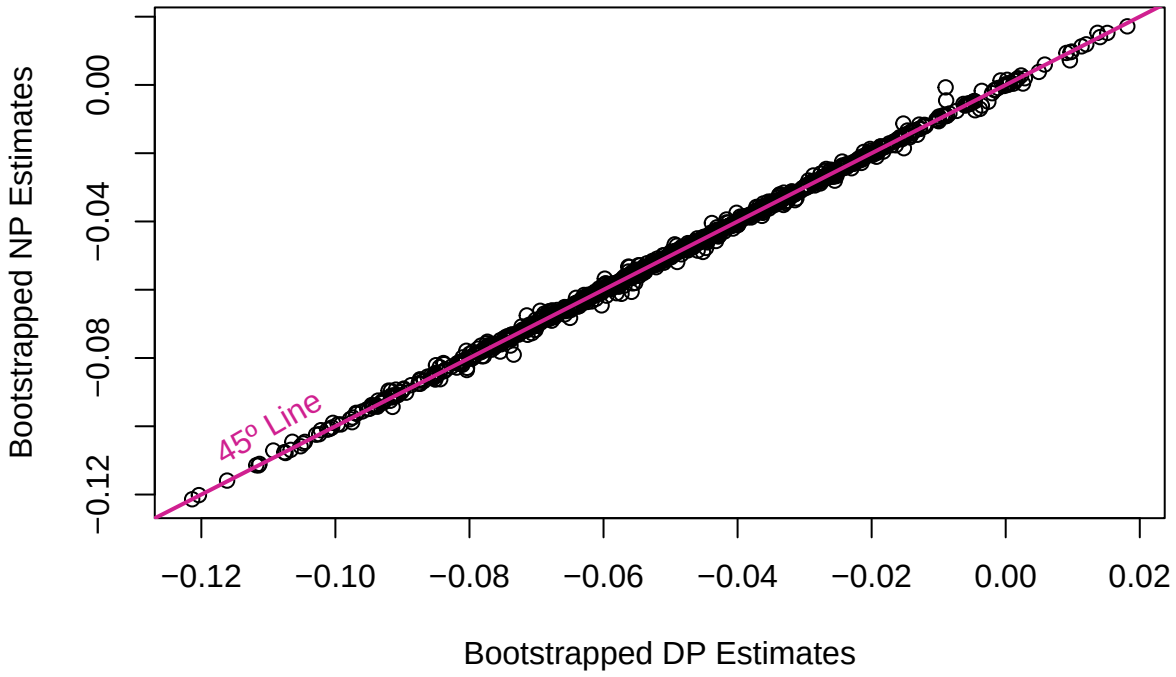
Plot 1: Distribution of DP Estimates of Race Differential



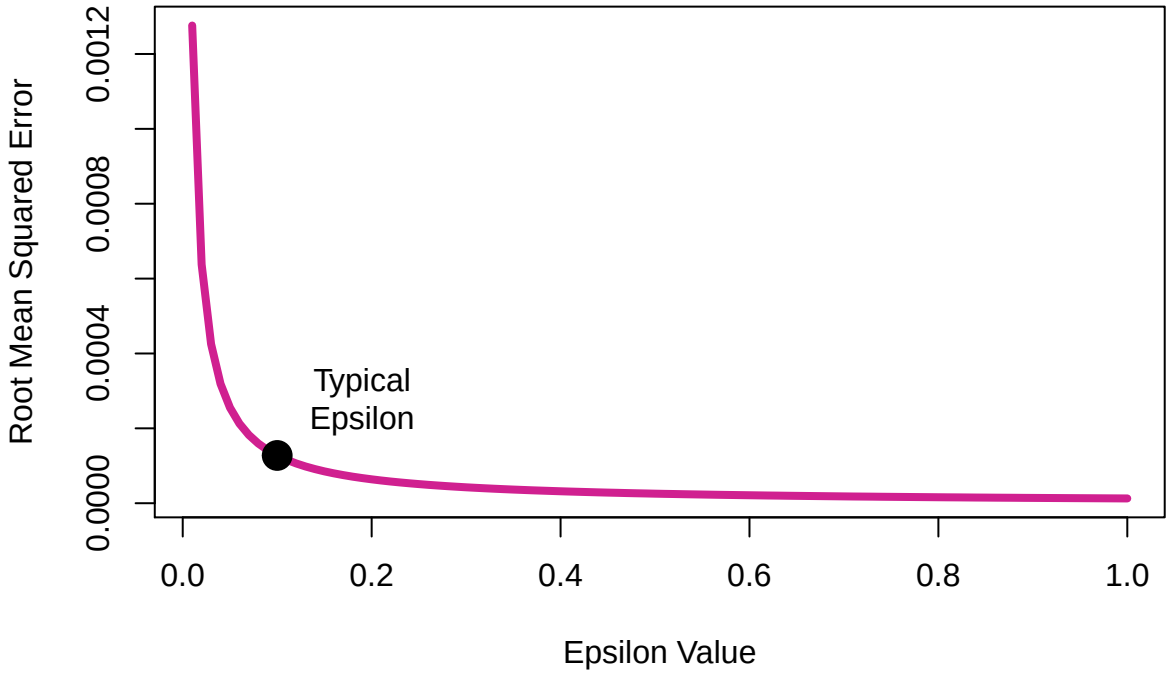
Plot 2: Race Differential Error in Bootstrapped Data



Plot 3: Bootstrapped DP Estimates vs Boostrapped NP Estimates



Plot 4: RMSE for Race Differential Calculation Across Epsilon Values



As another test of the algorithms, we next perform 1,000 bootstraps of Butler and Broockman's original dataset, creating simulated versions of the data in order to test the robustness of our analysis in the previous figure. For each of these 1,000 bootstraps, the same difference-in-means value (rate of response to Jake minus rate of response to Shawn) is calculated using both generic, non-privacy protecting tools and the Privacy Tools differentially private tools. We then compute the difference between these two results to develop an idea of the error generated across multiple simulations of the bootstrapped data. In **Plot 2** above, we graph those 1,000 error values and find that they cluster nicely around 0, demonstrating that even among a variety of simulated versions of the original study's dataset, the Privacy Tools algorithms continue to perform effectively. In **Plot 3** above as well, we also plot the differentially private bootstrapped estimates against the non-privacy protecting bootstrapped estimates; the approximately 45 degree line that the results create is a good sign, indicating that the sampling error due to the bootstrapping is larger and more dominant than the error due to the differentially private algorithms. This is a sign that the error generated by the algorithms is minimal.

Next, we test the epsilon level of the differentially private calculations. Epsilon is a way of quantifying the level of privacy that is guaranteed by differential privacy; as epsilon increases from 0.01 to 1, the privacy guarantee offered by the differential private algorithms decreases in strength. In **Plot 4** above, we run the same difference-in-means analysis that we have been working with at epsilon values ranging from 0.01 to 1, computing root mean squared error (RMSE) at each of those values. As expected, we see that RMSE, while high at first, drops dramatically as the epsilon values increase and begins to approach zero at the higher values of epsilon. Note that at a commonly chosen epsilon value of 0.1, we observe a small amount of RMSE, but this error is quite a bit smaller than the error generated at even smaller epsilon values.

As a final analysis of the success of this replication, we here reproduce an entire table from Butler and Broockman's original analysis. In **Figure 4** below, the table is reproduced using both non-privacy protecting and differentially private analysis. From the reported 90% range covered by the algorithms, it may be concluded that the values provided by the differentially private algorithms are consistent with and substantively similar in meaning to the true values provided by the non-privacy protecting analysis. This is another indicator that the algorithms are working effectively to produce results with utility.

Overall Effect Sizes -- Does Jake Receive More Replies than DeShawn?			
	No Partisanship Signal	**Republican Signal**	**Democratic Signal**
DeShawn Jackson	55.3% [52.4% , 58.4%] N=806	54.3% [51.4% , 57.4%] N=810	57.3% [54.4% , 60.3%] N=812
Jake Mueller	60.5% [57.6% , 63.5%] N=812	56.4% [53.6% , 59.5%] N=820	55.3% [52.4% , 58.4%] N=799
Race Differential	5.1% [5.11% , 5.15%] (p=0.04)	2.1% [2.10% , 2.18%] (p=0.39)	−1.9% [-2.00% , -1.89%] (p=0.39)

Figure 4: Table 1 from the original study. Numbers without brackets represent the non-privacy-protecting true values. Bracketed numbers represent the range of values covered by 90% of simulations of the algorithms.

IV. Replication 2: Gerber et al. (2014)

In this next replication, we reproduce the results of a 2014 study by Alan Gerber and others, entitled "Can Incarcerated Felons Be (Re)integrated into the Political System? Results from a Field Experiment." The study analyzes voter registration in the United States among one specific community, formerly incarcerated people, 93% of whom actually become eligible to vote after completing their incarceration period and satisfying certain components of their initial sentencing requirements. The authors of the study worked in tandem with government administrators in Connecticut to collect data about incarcerated people within the state and, using that data, configured a field experiment in the form of an informational outreach campaign which involved sending letters about voter registration to over 6,000 former felons in the state. Each letter included an official voter registration card for Connecticut; letters in the treatment condition also contained a letter written on the Secretary of State's letterhead which included specific information about the recipient's eligibility to vote as a former felon. The treatment condition Secretary of State letters came in two forms: one shorter and less informative (referred to as the "assurance" condition) and the other longer and more detailed (referred to as the "expanded assurance" condition). The dependent variables in the study tracked whether the letter recipients registered to vote following the letters' arrival and whether the recipients subsequently voted in the 2012 presidential election.

Although much of the information included in this study's dataset is technically public, as it involves crime-related info that is typically made publicy available, the data concerns human subjects and their personal decisions and in that respect represents the kind of sensitive data that the differential privacy tools and PSI software seek to protect. We invoke these tools now in replicating the analysis performed on this data as a test of how effective those tools may prove in analyzing real-world sensitive data.

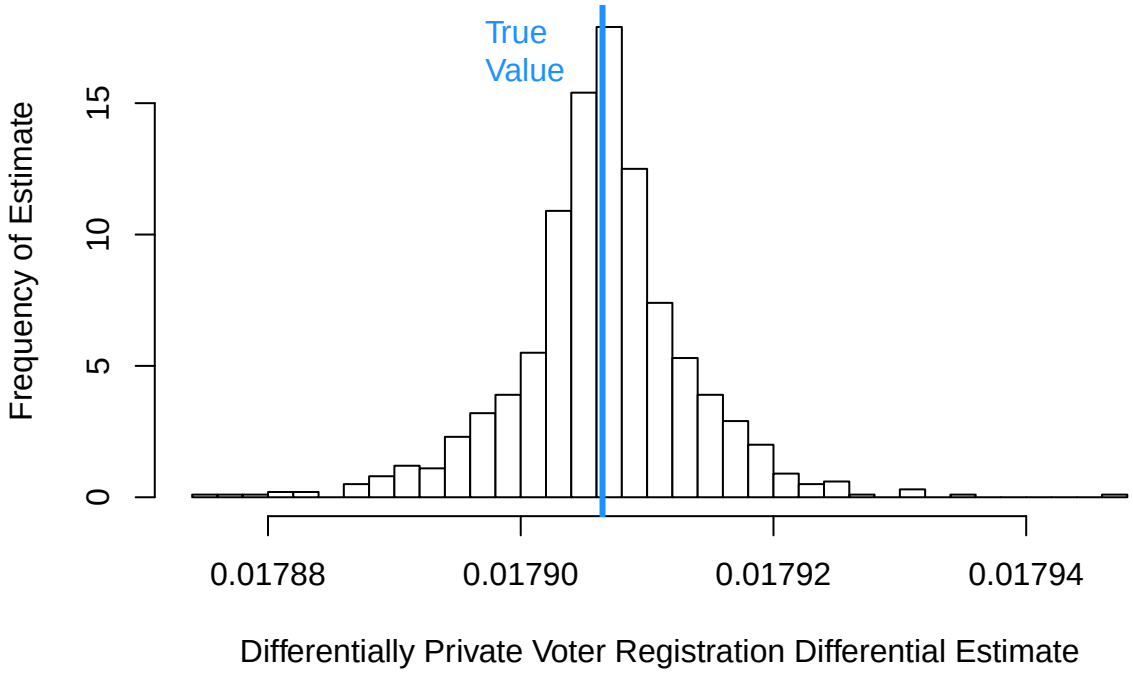
We will begin by replicating one of the more interesting and prominent results of this study, the finding that sending former felons both registration forms and a specific informational letter about registration for felons was more effective in increasing registration rates than sending the registration forms alone. This finding essentially compares both of the treatment conditions in the study combined (i.e., receiving one of two possible informational letters AND a registration form) compared to the control condition (i.e., receiving a registration form only). The value for this difference-in-means analysis is 0.0179 or 1.79%, indicating that the treatment did succeed in urging more former felons to register to vote. With a p-value of 0.002, this result is significant at the 95% significance level.

We first replicate this result using the Privacy Tools differentially private algorithms. In **Plot 5** below, we replicate these results with the algorithms 1,000 times in order to reduce bias and inaccuracy; those results are produced in a histogram format in order to show the density of the returned results. As may be seen, the results cluster closely around the true value produced by the researchers, denoted by the blue line on the graph. As in the last replication, the shape of the resultant histogram is resemblant of a LaPlacian distribution, an expected result as the differentically private algorithms do function by adding LaPlacian noise to data analysis releases. Regardless, this plot is a good indicator that the differentially private algorithms are working well in the context of this data and adding noise as they were designed to do.

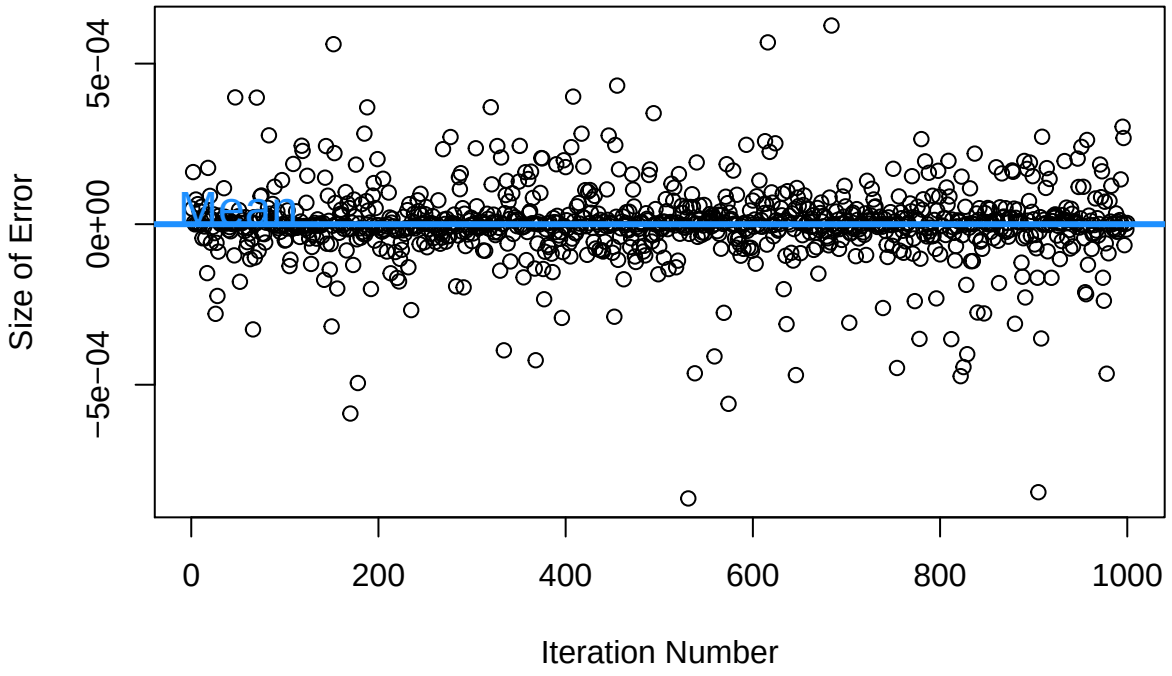
Another useful result from this plot is the 90% range provided by the differentially private algorithms when computing this significant difference-in-means result. 90% of the 1,000 results produced wih the algorithm fall within the range of [0.0178,0.0179]. Again, with the true value being 0.0179 or 1.79%, we see that the private releases fall very near the true value of the analysis, both in terms of numerical size and direction. More importantly, these results are substantively similar in meaning to the true value, indicating that the differentially private algorithms are indeed capable of performing useful and effective data analysis.

Differentially private T-test algorithms that are in development will soon be yet another useful test of how well the differentially private algorithms work. A differentially private T-test comparing the rate of registration among the control group and the rate of registration among the treatment group, for example, would be able to produce a p-value that would indicatei how significant our analysis is, providing another way of comparing the usefulness of the private. However, until those tools become available, this simulation provides a useful picture of how effectively the algorithms are working, particularly for a dataset of this size.

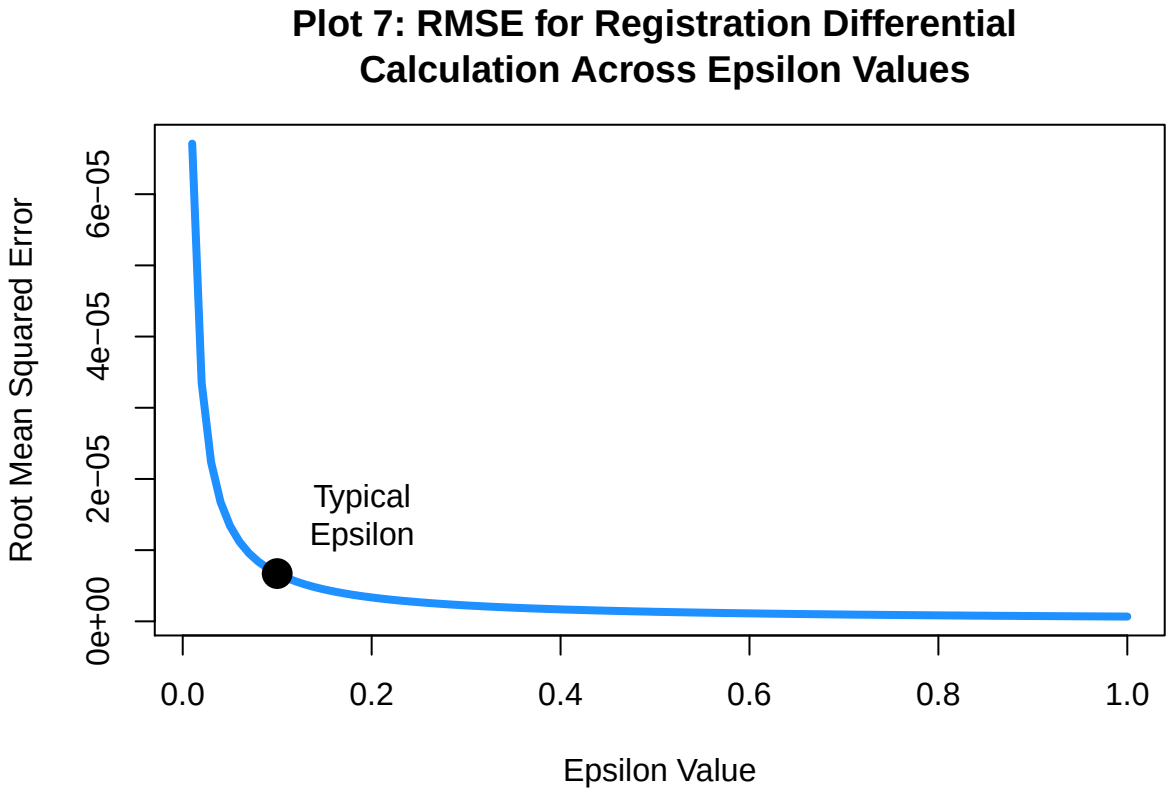
Plot 5: Distribution of DP Estimates of Voter Registration Differential



Plot 6: Registration Differential Error in Bootstrapped Data



In another test of the Privacy Tools algorithms, we create bootstraps of the original study dataset in order to further test the functionality of the algorithms. Bootstrapping involves producing simulated versions of data that are structurally identical to the initial data, providing supplemental datasets which may be used to test the robustness, stability, and uncertainty of some type of analysis. We do so here to test the functionality of our differentially private difference-in-means analysis on this same interesting finding about voter registration from Gerber et al's study. We begin here by producing 1,000 bootstraps of the dataset and then calculating the difference-in-means value for each, using both non-privacy protecting statistical tools and the Privacy Tools differentially private tools. As before, we find the difference between these two results by subtracting so as to visualize the error generated by the differentially private algorithms in each version of the bootstrapped data. Those 1,000 error calculations are graphed in **Plot 6** above; as anticipated, they all fall on or near zero, indicating that the algorithms continue to work effectively even on the bootstrapped datasets. It may be noted that these errors are also particularly small, smaller than in other replications produced in this paper – this may be a result of the fact that this dataset is large in terms of number of observations. With an N of 6,441, the dataset is large enough for the differentially private algorithms to be able to produce fine results with relatively little error.



In our next test, we consider the epsilon values associated with the differentially private calculations. As mentioned previously, epsilon measures the privacy guarantee offered by differentially privacy. When epsilon is low, the privacy guarantee is relatively weak; conversely, when epsilon is high, the privacy guarantee is relatively strong. In **Plot 7** above, we run 500 simulations of the same difference-in-means analysis at 100 values of epsilon, ranging from 0 to 1. We then compute RMSE for each of those analyses at each of those epsilon values. The results show that the RMSE begins high at the low epsilon values but drops dramatically as the epsilon values increase. We see as well that at a typical epsilon value of 0.1, RMSE is fairly small and is certainly minute compared to RMSE at lower epsilon values. However, the ploy shows overall that RMSE

remains fairly limited, even at the very low epsilon values, and this is a good sign that the differentially private algorithms can still produce useful results even without being assigned a large epsilon value.

As a final analysis of this replication, we here reproduce an entire table from the researchers' original analysis, shown in **Figure 5** below. As in other replications, the reported 90% ranges in this table show that the values produced by the private algorithms are similar both in terms of numerical size and substantive meaning to the original, non-privacy protecting results – another good sign of functionality of the differentially private algorithms. On the whole, the algorithms are working well in the context of this data and are producing results that are substantively and numerically useful.

2012 Registration and Turnout among Treatment Groups				
	Control	**Assurance**	**Expanded Assurance**	**Pooled Assurance**
Mean Registration	5.9% [5.0% , 6.5%]	7.6% [6.1% , 9.1%]	7.7% [6.1% , 9.1%]	7.7% [6.8% , 8.3%]
Difference-in-Means Relative to Control		1.8% [1.0% , 2.5%] (p=0.010)	1.8% [1.1% , 2.6%] (p=0.008)	1.8% [1.8% , 1.8%] (p=0.002)
Mean Turnout	3.0% [2.2% , 3.7%]	4.1% [2.5% , 5.5%]	3.8% [2.2% , 5.2%]	3.9% [3.1% , 4.6%]
Difference-in-Means Relative to Control		1.1% [0.3% , 1.8%] (p=0.028)	0.8% [0.0% , 1.5%] (p=0.085)	0.9% [0.9% , 0.9%] (p=0.024)
Num. of Observations	3,134	1,574	1,572	3,146

Figure 5: Table 1 from Gerber et al's original study. Numbers without brackets represent the non-privacy-protecting true values. Bracketed numbers represent the range of values covered by 90% of simulations of the differentially private algorithms.

V. Replication 3: McClendon (2013)

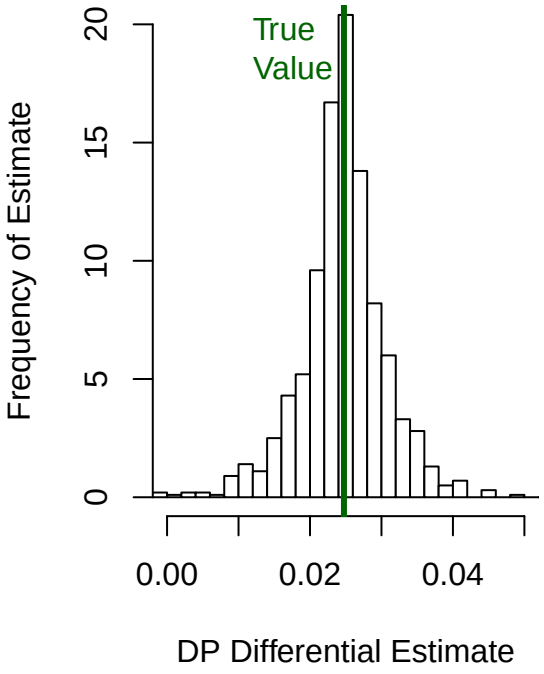
In the next replication, we reproduce a portion of the analysis from Gwyneth McClendon's 2013 study entitled "Social Esteem and Participation in Contentious Politics: A Field Experiment at an LGBT Pride Rally." The study seeks to analyze the motivations that move individuals to join in "contentious politics," controversial political activities that people may otherwise be deterred from partaking in. In her study, McClendon proposes that individuals may participate due to reasons relating to social esteem, and she presents the results of a field study which tests that hypothesis. In this experiment, prior to an LGBT pride event, the researche sent possible attendees (i.e., members of the hosting LGBT group) invitations to the event, varying the kind of email each received. In the control condition, invitees were sent a basic informational email that notified them about the event and its key details. In the "newsletter condition," invitees received the same basic email and were further told that if they chose to attend, their name would be included in a list sent out in the host group's next monthly newsletter. In the "Facebook condition," invitees received the same basic email and were encouraged to post pictures from the event on the organization's Facebook members so that other members could look at and "like" those photos. The two treatment conditions were intended to introduce two types of social esteem incentives in order to encourage individuals to attend the event. The dependent variables, accordingly, tracked whether invitees RSVPed to attend the event (measured with an online Evite website), whether they actually attended the event (measured by physically checking in members at the event), and whether they reported having attended the event one week after it occurred (measured by sending out a survey about the event to invitees). The study determines ultimately that invitees do seem to be motivated by social esteem in participating in contentious politics, finding that individuals in either of the two treatment conditions were in fact more likely to attend the rally.

The data included in McClendon's dataset is an excellent example of data that would benefit from some sort of privacy guarantee. Though the researcher received contact information for group members from administrators of the group itself, the individual group members did not actually consent to being included in the study; furthermore, they may not want their membership in the group made public, but this is a possible consequence of some sort of re-identification attack that could be directed at the data. After all, the sexual orientation of individuals, even if only implied by membership in an LGBT group, is sensitive information that should not be subject to privacy threats. This being the case, this dataset could definitely profit from the strong privacy protections that differentially private releases allow for.

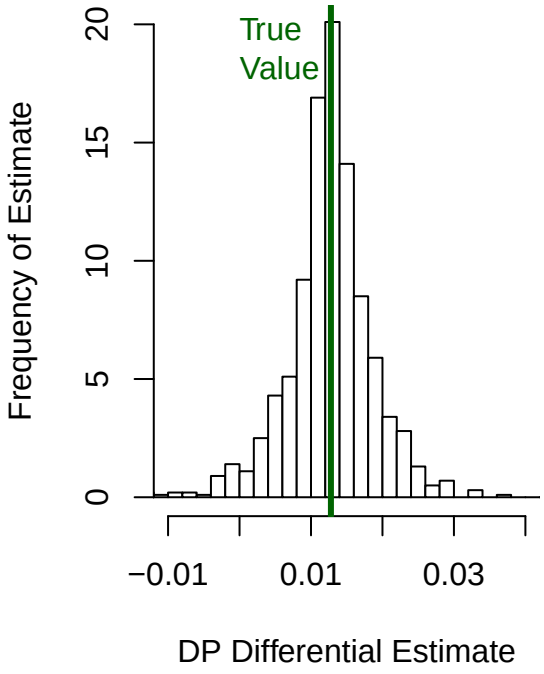
We begin our analysis by replicating a key result from the study, the finding that invitees in one of the two treatment conditions (the "newsletter condition" or the "Facebook condition") were 2.48% more likely to register to attend the rally and 1.28% more likely to actually attend the rally than invitees in the control condition. With respective p-values of 0.001 and 0.012, these results are significant at the 95% level, and represent two of the more interesting findings from the study, evidencing that McClendon's hypothesis may be correct: invitees with social esteem incentive to attend the event are more likely to do so.

In **Plots 8 and 9** below, we replicate these two difference-in-means values, derived by subtracting the rate of intended participation [rate of actual attendance] among control condition invitees from the rate of intended participation [rate of actual attendance] among treatment condition invitees. The 90% range of the values over 1,000 runs of the algorithms for the intended participation differential is [0.015,0.032], which is large, and the histogram for this analysis, with its high variance, also indicates that the private results that were generated are not highly precise. The 90% range for the attended participation differential is [0.003,0.020] and the histogram for this analysishas an even wider spread, extending so far as to dip into negative values; with a true, non-privacy protecting value of 0.0128, these private results are also not well-tuned. Though both of these graphs show that the differentially private algorithms are working adequately to inject LaPlace noise into the private releases (indicated by the approximate LaPlacian shape of both graphs), there appears to be a lot of noise in these estimates, which is is diluting their utility. It is also particularly concerning that a differentially private release for the actual participation differential could be returned as negative, since the true value for that analysis is positive. This is a sign that the differentially private algorithms, while seeming to work moderately well, are still not at their strongest here. This may be due to the small number of observations in this study, however – with just under 4,000 observations, this dataset is relatively small and may be generating particularly noisy private releases.

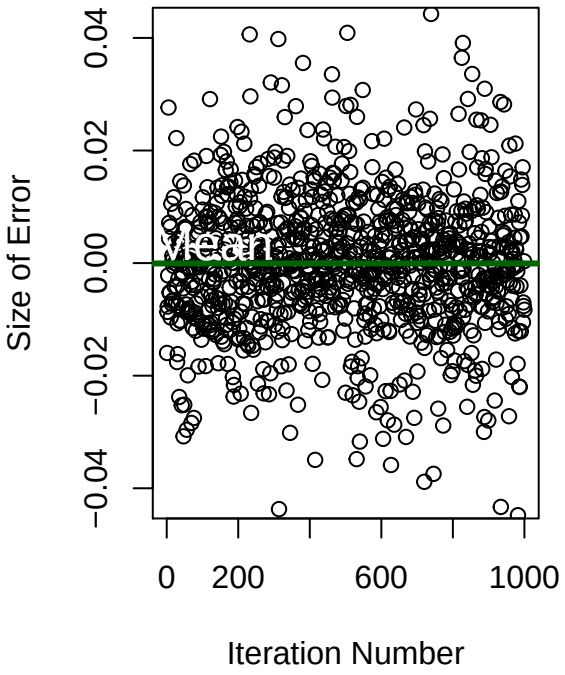
Plot 8: Distribution of DP Intended Participation Differentia



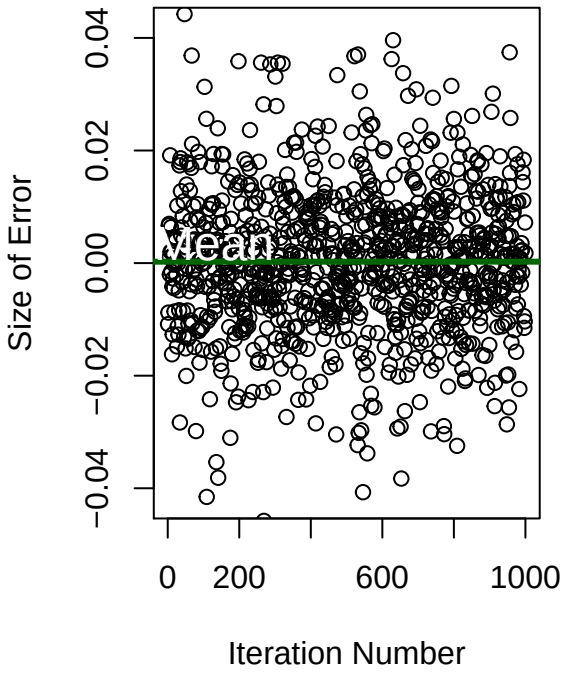
Plot 9: Distribution of DP Actual Participation Differential



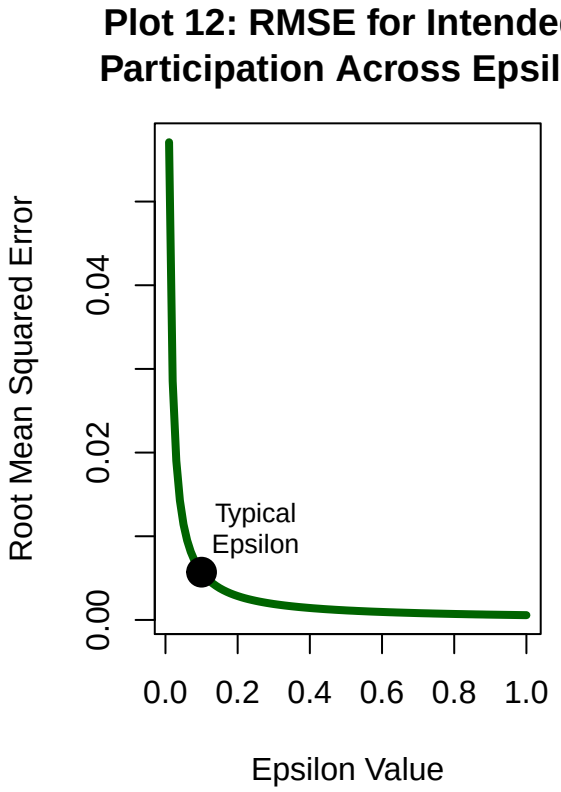
Plot 10: Intended Participation Error in Bootstrapped Data

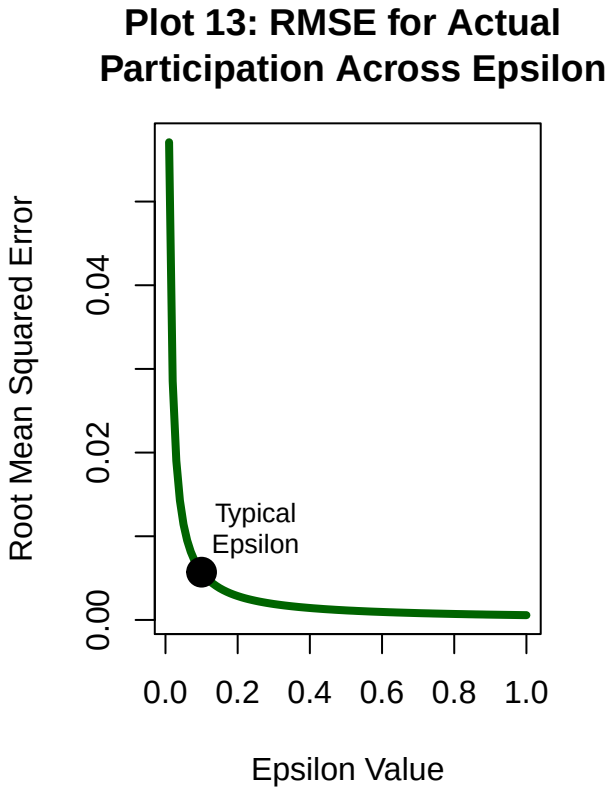


Plot 11: Actual Participation Error in Bootstrapped Data



In another test of this data, we bootstrap McClendon's dataset 1,000 times in order to produce simulated versions of the initial data. We then run both the differentially private difference-in-means algorithms and standard non-privacy protecting analysis on the data to form the same intended participation and actual participation differentials graphed in the previous plot. We then compute the difference between the private releases and the non-privacy protecting results to grasp a basic picture of the error generated by the differentially private algorithms over multiple bootstrapped datasets. This analysis is intended to provide stability and robustness to our research, and **Plots 10 and 11** above, which show that these errors cluster around zero, evidence this by demonstrating that the differentially private algorithms are generating closely-tuned answers. However, in both plots, the many plot points showing high error values, such as 0.04 or -0.04, show that the algorithms are producing a lot of noise, probably due to this dataset's small N.





Next, we consider how epsilon value affects the error generated by the differentially private algorithms. Epsilon, our privacy budget, determines how closely the privacy of some dataset or some release is protected, and it therefore greatly affects the accuracy of differentially private outputs. In **Plots 12 and 13** above, we calculate RMSE of the same difference-in-means analysis we have been considering 500 separate times at 100 different values of epsilon ranging from 0.01 to 1. As in the past replications, RMSE is high (around 0.06) when epsilon is equal to zero, but it decreases drastically and begins to approach zero as epsilon increases. Unlike past replications, however, RMSE is particularly high at the lower epsilon values here, with the ability to influence the outcome by up to, in this case, six percentage points. As mentioned previously, this tendency for the algorithms to introduce so much extra noise is likely due to the small N of this dataset.

Finally, as one last test of this replication, in **Figure 6** we reproduce an entire table from McClendon's study using our differentially private algorithms, reporting 90% ranges of the values across 1,000 simulations of the algorithms rather than the basic difference-in-means values. Ultimately, while the various tests run on this replication have indicated lapses in the algorithms' performance, this last test shows us that overall, the differentially private algorithms are still returning ranges that are similar in substantive meaning to the original researchers' values. This is an exciting and positive result.

Differences in Rates of Participation by Treatment		
	Intent to Participate	**Actual Participation**
Newsletter	6.08% [4.19% , 7.50%] N=1217	3.04% [1.11% , 4.45%] N=1217
Control	3.53% [1.61% , 4.94%] N=1217	1.72% [-0.20% , 3.14%] N=1217
Difference (p-value)	2.55% [0.00% , 5.28%] (p=0.003)	1.32% [-1.22% , 4.05%] (p=0.034)
Facebook	5.93% [4.01% , 8.00%] N=1213	2.96% [1.04% , 5.03%] N=1213
Control	3.53% [1.61% , 4.94%] N=1217	1.72% [-0.20% , 3.14%] N=1217
Difference (p-value)	2.41% [-0.15% , 5.14%] (p=0.005)	1.24% [-1.31% , 3.98%] (p=0.043)
Newsletter & Facebook	6.01% [5.05% , 7.04%] N=2430	3% [2.04% , 4.03%] N=2430
Control	3.53% [1.61% , 4.94%] N=1217	1.72% [-0.20% , 3.14%] N=1217
Difference (p-value)	2.48% [0.55% , 4.65%] (p=0.720)	1.28% [-0.64% , 3.45%] (p=0.509)

Figure 6: Table 1 from McClendon's original study. Numbers without brackets represent the non-privacy-protecting true values. Bracketed numbers represent the range of values covered by 90% of simulations of the differentially private algorithms.

VI. Replication 4: Weitz-Shapiro & Winters (2015)

For our final replication, we here reproduce a portion of the analysis from Rebecca Weitz-Shapiro and Matthew Winters' study entitled "Can Citizens Discern? Information Credibility, Political Sophistication, and the Punishment of Corruption in Brazil." The study concerns Brazilian citizens' perceptions of corruption, in particular their ability to distinguish corruption amidst sources of varying credibility. The researchers hypothesize that citizens should be able to discern between credible and fraudulent sources, and, using survey data about political corruption, test that hypothesis in the study.

In the survey data, participants were read a prompt that described a political figure and were asked, given the description, how likely they would be to vote for that political figure on a scale of 1 to 4. In the control condition, participants were given a simple, generally flattering description of the political figure; in the treatment conditions, participants were given that same simple description of the political figure, plus a warning that the politician is accused of being corrupt and easily bribed. In the first treatment condition, known as the "less credible source" condition, participants were told that the source of the politician's corruption accusations was the opposition party, objectively a less credible source. In the second treatment condition, known as the "more credible source" condition, participants were told that the source of the accusations was a federal audit, objectively a more truthful source. The dependent variable in the study was participants' professed likeliness to vote for the candidate, again rated on a scale of 1 to 4.

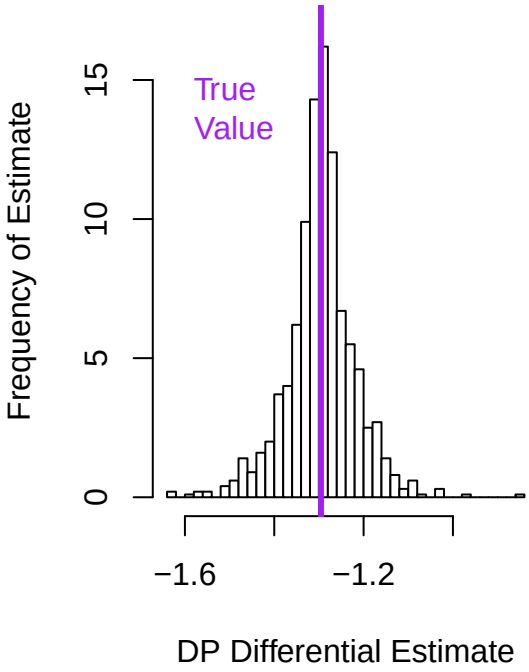
This information is actually quite sensitive! As Brazil is a country in which corruption is somewhat common, if information about study participants were to be leaked, thereby revealing participants' personal opinions on corruption, participants could be severely discriminated against or punished. Because this data should be protected, it is an excellent example of the kind of data that the PSI software seeks to protect.

An issue with this study, however, is its small number of observations. The dataset includes only 2002 observations, a marked decrease from the other dataset replications included in this report, and furthermore, certain queries on the data involve subsetting, which serves to reduce the N even more. This presents problems because, as previously mentioned, the Privacy Tools algorithms are best suited to large datasets with many thousands of observations. As we will see in the subsequent analysis, the algorithms are still capable of producing useful analysis from a dataset with this few observations, but it is worth considering the merit of the analysis overall and whether differentially privacy is compatible with a small dataset such as this one. We will ponder this thought in several tests to follow.

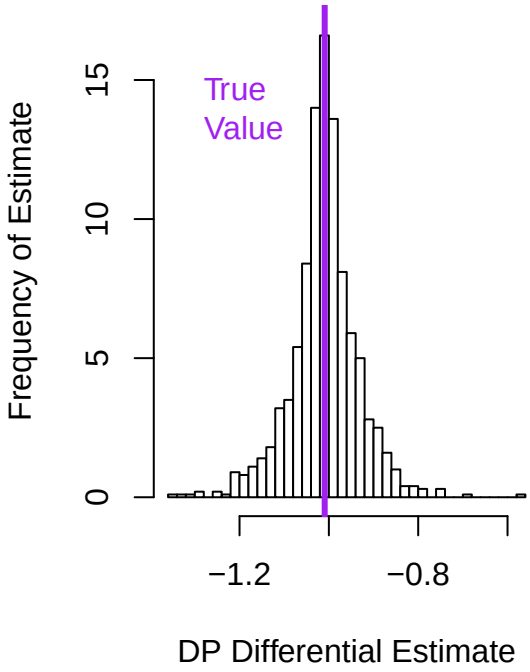
We begin by replicating a key result from the study, the finding that participants do in fact seem to be skilled at discerning between more credible and less credible sources. Compared to the control condition in which participants were told nothing about the political figure's possible corruption, participants in the treatment condition who were told about the politician's corruption ranked themselves as less likely to vote for the figure; furthermore, participants who were given a legitimate source (a federal audit) for the corruption accusations were the most likely to decrease their likelihood of voting for the candidate, as compared to participants given a weak source (the opposition party) for the accusations. Again, participants ranked their likeliness to vote for the political figure on a scale of 1 to 4, so whereas control condition participants ranked their likeliness on average as 3.38, the less credible accusation participants ranked their likeliness as 2.36 (a -1.02 difference) and the more credible accusation participants ranked their likeliness as 2.08 (a -1.30 difference).

In **Plots 14 and 15** below, we recreate the two difference-in-means analyses discussed above, using the Privacy Tools differentially private algorithms. The left graph shows the difference between the control and more credible accusation conditions, with a true value of -1.30; the right graph shows the difference between the control and less credible accusation conditions, with a true value of -1.02. The 90% range defined over 1,000 simulations of the left analysis is [-1.43%,-1.17%]; the 90% range defined over 1,000 simulations of the right analysis is [-1.14%,-0.89%]. Both of these ranges are honed in around the respective true values, but they are fairly broad, and this is reflected in the visuals provided by the histograms below. While the differentially private algorithms appear to be working functionally, again seeming to approximate LaPlacian distributions, they are not numerically exact, and their utility is likely better measured in terms of their substantive interpretations, which would be true to the results in the original study.

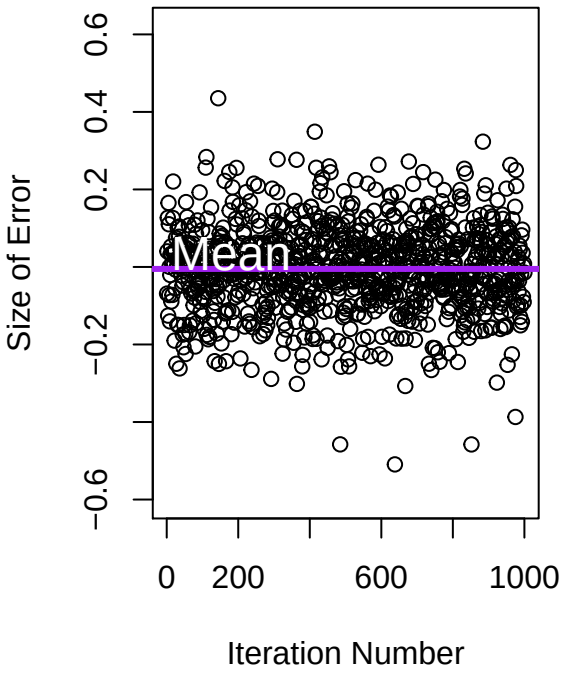
Plot 14: Distribution of DP Diff. for Credible Accusations



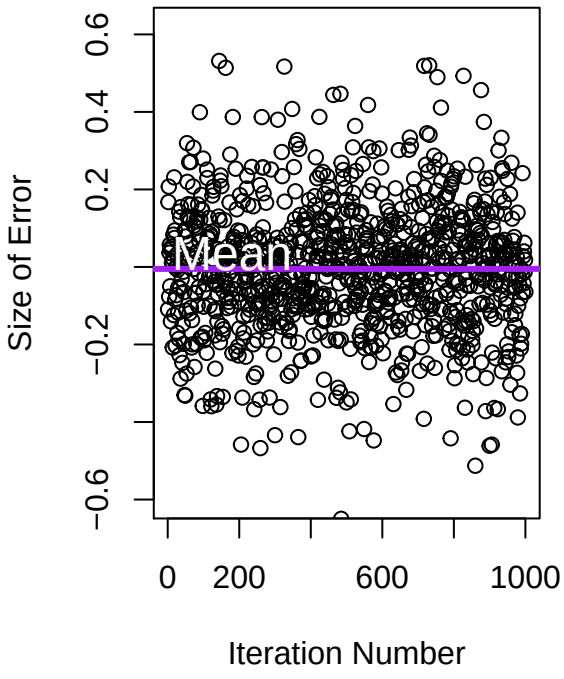
Plot 15: Distribution of DP Diff. for Less Credible Accusatio



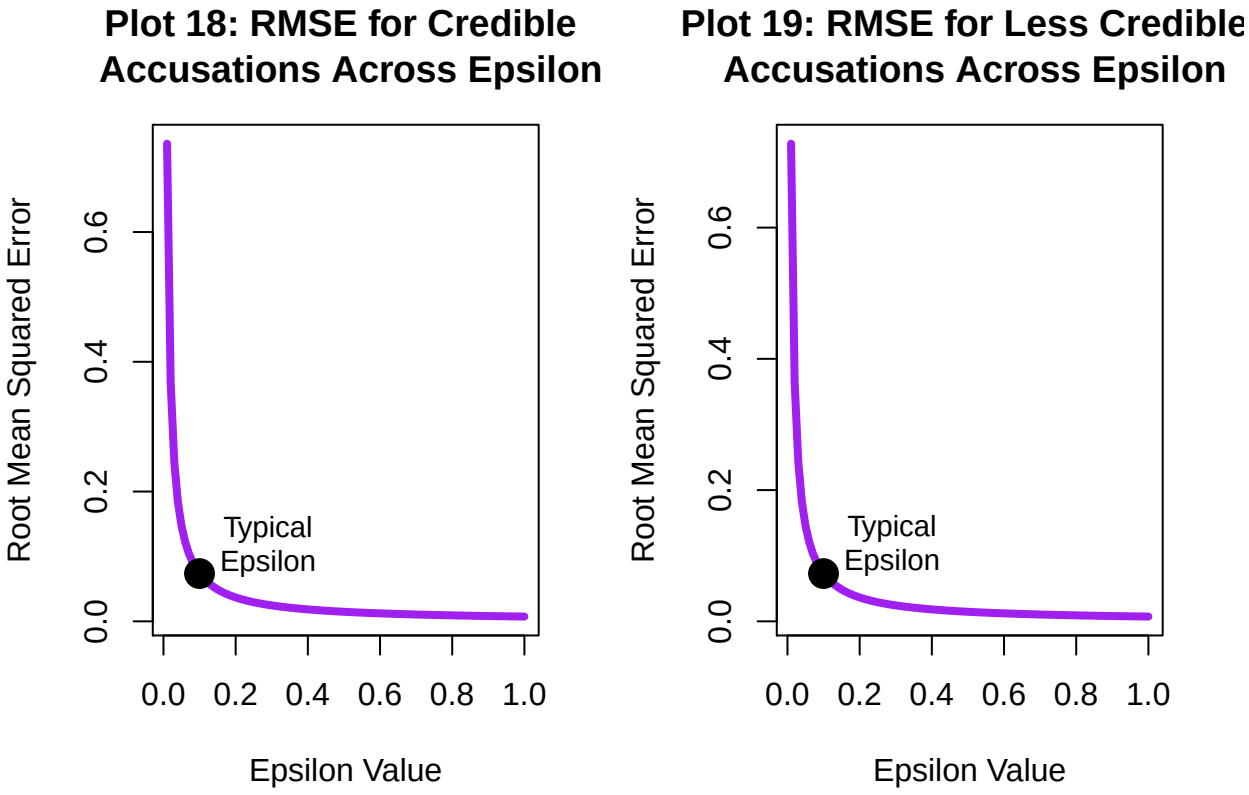
Plot 16: DP Diff for Credible Acc. in Bootstrapped Data



Plot 17: DP Diff for Less Credibl Acc. in Bootstrapped Data



In our next test of this data, we create 1,000 bootstraps of Weitz-Shapiro and Winters' dataset in order to produce new versions of the data that may be tested for accuracy and stability. With this goal in mind, we calculate the differentially private and non-privacy protecting difference-in-means values for each bootstrapped dataset and take the simple difference between the two values in order to produce a picture of the kind of error generated by the algorithms across multiple trials. Graphs of the error for each difference-in-means result are printed above, with the credible accusation analysis on the left in **Plot 16** and the less credible accusation analysis on the right in **Plot 17**. The credible accusation analysis graph shows 1,000 plot points closely honed in around zero, with some variance; the lowest error value and the highest error value extend to -0.45 and 0.44, respectively. The less credible accusation analysis shows a wide variety of error values with range extending to -0.62 on the low end and 1.35 on the high end, respectively; this may be because this analysis requires subsetting of the data, reducing its already limited N to an even smaller number. Regardless, for both analyses, there is a lot of error. An error value of 0.4 or 0.6, when applied to a four-point scale, can greatly tip the meaning of the result that is produced! These errors, thus imagined, may hinder both the numerical size and substantive interpretation of differentially private releases about this data.



In another test, we analyze the stability of the error generated by the differentially private algorithms across a wide range of epsilon values. To reiterate, epsilon is a quantification of the privacy guarantee that differential privacy offers, with lower epsilon values indicating a relatively weak privacy guarantee and higher epsilon values indicating a relatively strong privacy guarantee. In **Plots 18 and 19** above, we compute the differentially private difference-in-means values 500 times each at 100 unique values of epsilon ranging from 0.01 to 1. We see that epsilon value can greatly shape the amount of error, in this case RMSE, that the differentially private algorithms produce. This is a valuable result to note, as it indicates that although the algorithms are not producing very exact or finely-tuned releases for this dataset, likely because of its

small N, the algorithms are still capable of performing at a high level when the epsilon value allocated to the analysis is high. For example, a data analyst who is willing to expend much of her epsilon budget (meaning that she can use a high epsilon) on generating this differentially private release would still be able to get a highly accurate number, limited size of the dataset notwithstanding. That being said, if an analyst did not have much available epsilon to devote to this release, the result would be imprecise and likely not very valuable. This is an interesting and useful finding, indicating that while not all datasets are perfectly suited for use with the differentially privacy algorithms, the algorithms can still generate worthwhile results for most datasets as long there is a high enough epsilon value!

Finally, as a final test of the success of this replication, in **Figure 7** below, we recreate a table from Weitz-Shapiro and Winters' analysis using both non-privacy protecting and differentially private tools. We observe that these ranges, while somewhat wide, are generally consistent with the true values from the study, likely to provide results that are in the spirit of the original data analysis and interpretation. Ultimately, this is a final sign of the strong functionality of these algorithms; even in such a limited dataset, they are still capable of producing releases that, while imprecise numerically, are substantively in line with the true values they seek to replicate. This is yet another achievement for the differentially private algorithms.

How Likely Are You to Vote for the Political Figure? (On a Scale of 1 to 4)

	Average Response	Dif. from Control Conditions	Diff. from Unsourced Accusations	Diff. from Less Credible Accusations
Credible Accusations	2.07 [1.95 , 2.22] N=553	- 1.3 [-1.43 , -1.17] N=553	- 0.1 [-0.24 , 0.02] N=553	- 0.29 [-0.29 , -0.28] N=553
Less Credible Accusations	2.37 [2.24 , 2.50] N=547	- 1.02 [-1.14 , -0.89] N=547	0.18 [0.05 , 0.31] N=547	~~
Unsourced Accusations	2.18 [1.93 , 2.45] N=278	- 1.2 [-1.19 , -1.19] N=278	~~	~~
No Accusations	3.38 [3.12 , 3.64] N=560	~~	~~	~~

Figure 7: Table 2 from the authors' original study. Numbers without brackets represent the non-privacy-protecting true values. Bracketed numbers represent the range of values covered by 90% of simulations of the differentially private algorithms.

VII. Legal Research

As a supplementary project this summer I undertook legal research under the supervision of Privacy Tools legal fellow Alexandra Wood. This research provided me with a valued opportunity to engage with the legal/policy arm of the Privacy Tools project and to obtain a fuller picture of the importance of privacy protections for many research subjects. As part of this project, I selected a 2013 Pew Research Center dataset/study about LGBT Americans, conducted research about legal approaches to privacy and re-identification, and finally wrote a short paper on the nuanced risks, harms, and possible controls associated with the study I chose. My goal in choosing a study published by Pew, a large American think tank, was to participate in useful thought exercises about big data and data privacy that could be applied more broadly to data compiled by think tanks in general. As a public policy aficionado with a strong interest in the work of such policy-minded organizations, I found this inquiry fascinating.

A full copy of my legal research paper is included below. I envision this work as a useful complement to my replication project, as I believe it allows insight into the importance of the PSI software and thereby provides useful motivation for the replication experiments that I conducted.

A Survey of LGBT Americans by Pew Research Center

As part of its series entitled LGBT in Changing Times, nonpartisan think tank Pew Research Center released in fall 2014 the results of a large-scale public opinion poll conducted among LGBT Americans. The dataset includes 1,197 observations, each representing individual respondents' responses to a lengthy survey, and covers topics ranging from respondents' perspectives on sexual orientation to their support of public figures known for promoting LGBT rights. The data is de-identified to omit names, demographic information, and all identifying information, referring to respondents only by unique respondent identifiers that seem to be randomly generated. With these limited privacy protections in place, the dataset and its corresponding analysis have been made publicly available online, safeguarded only by a short fill-in form requesting personal information about the data user. Although Pew took steps to protect the privacy of respondents in the data through de-identification, privacy attacks likely remain viable for the respondents in the dataset. The following analysis will consider these possible vulnerabilities, data uses, and attacks at each stage of the data lifecycle, followed by a detailing of available controls that may function well in the unique context of this dataset.

It is important that data such as this is kept more secure, both for the benefit of the survey participants and for the good of the researchers as well. In particular, survey participants whose data is included in this dataset are vulnerable to harm if their data were to be leaked or accessed by an adversary. This is due to the fact that this dataset contains a wealth of sensitive and confidential information, ranging from participants' identification as gay, lesbian, bisexual, or transgender to their occasionally controversial opinions about their families, about religion, and about prominent political leaders. This information is sensitive and must be protected for the sake of the participants, who could be socially or economically harmed if it were to be released. For example, discrimination against LGBT individuals is well documented within the workplace, where issues of pay or sexual harassment can result from the sharing of such private information. Discrimination against LGBT individuals can also occur as well in the housing market, in the classroom, and even on simple streets; these could be products of privacy leaks on this data, Additionally, many LGBT individuals, especially those of a younger generation, may not communicate their sexual orientation to friends or family members, and reputational harm and harm to relationships could therefore also be consequences of leaks of the private information in this dataset.

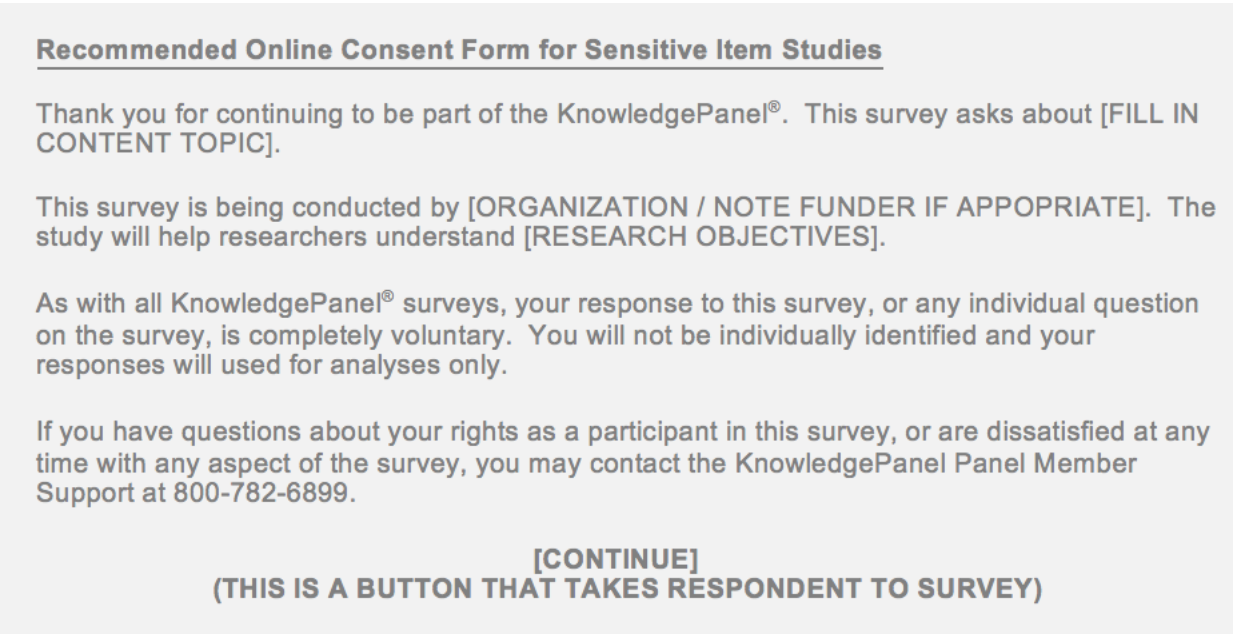
Protecting this information is crucial for researchers as well, whose credibility as authors and analysts would be damaged if their data fell prey to deleterious adversaries or motives. If it were to be publicized that a researcher had not protected his or her subjects' privacy adequately, he or she could be subject to such harms as loss of funding, loss of employment, reputational damage, criminal punishments, etc. It is thus mutually beneficial for both researcher and research subject that the privacy of this data be protected!

A. Collection and Acceptance Stage

The dataset for the survey of LGBT Americans was compiled by the GfK Group, a German market research group, using their online research software, KnowledgePanel. KnowledgePanel aims to conduct representative sampling of Americans by randomly selecting households to participate in their online surveys in exchange for small rewards and cash payments. Survey codes are distributed by mail, and households are given full choice over whether or not to participate in the research, which is entirely completed online through the group's private website. The research panel provides the information and data collected from its surveys to various companies, universities, news organizations, and think tanks, including Pew Research Center. There are no apparent restrictions on the data usage, as it is provided to institutions ranging from private companies to government organizations to NGOs.

Although it collects much identifying information about participants, including their name, location, payment information, billing address, and demographic information, GfK Group stores only minimal personally identifying information about its survey participants, which it uses for payment and other purposes. In a statement on the "Frequently Asked Questions" page of the company's website, GfK states, "We never give or re-sell your name, address or any information that could be used to identify you to anyone else outside GfK without your prior permission. But even more than that, we only let a select few people inside our company know who you are." Although it is likely necessary that the company retains personal information about participants for these uses, the availability of this personal, identifying information online, even stored deep within GfK Group's private online servers, poses some threat to participants' privacy. An online attacker may be able to use hacking techniques to penetrate these servers and obtain access to personal files. Furthermore, by linking external personal data with confidential survey data, an attacker may very well succeed in retrieving sensitive information about KnowledgePanel participants. The GfK Group can attempt to protect itself against these attacks through encryption of website data and more secure password controls, but ultimately there is not much the GfK group can do to truly secure itself. All in all, these attacks remain viable threats to data privacy and should be kept in mind as such. It may be good practice for the GfK Group to warn survey participants about these possibilities in their online privacy policy, which currently does not address the prospect of online hackers or the risk of re-identification.

Interestingly, the Gfk Group only uses consent forms for survey participants aged 13-17 and for surveys containing objectively sensitive information. In the latter case, the consent form looks as follows:



Recommended Online Consent Form for Sensitive Item Studies

Thank you for continuing to be part of the KnowledgePanel®. This survey asks about [FILL IN CONTENT TOPIC].

This survey is being conducted by [ORGANIZATION / NOTE FUNDER IF APPOPRIATE]. The study will help researchers understand [RESEARCH OBJECTIVES].

As with all KnowledgePanel® surveys, your response to this survey, or any individual question on the survey, is completely voluntary. You will not be individually identified and your responses will used for analyses only.

If you have questions about your rights as a participant in this survey, or are dissatisfied at any time with any aspect of the survey, you may contact the KnowledgePanel Panel Member Support at 800-782-6899.

[CONTINUE]
(THIS IS A BUTTON THAT TAKES RESPONDENT TO SURVEY)

Figure 8: Consent form employed by the GfK Group

As may be noted, this consent form is very limited and does not fully explain how de-identification of the data occurs, explaining only that individuals in the study will not be individually identified. Participants who do not understand the dangers of re-identification would not be fully prepared by this consent form if there were to be an information leak in this data, which could present problems. The vague nature of this consent form could certainly lead to potential mismatches between participants' expectations about their privacy protections and actual practice.

One main concern with the data collected by KnowledgePanel is that, due to the long-term relationship between research participants and the research panel, participants may be more likely to share data that they otherwise might not feel comfortable sharing. Participants who are comfortable working with the panel and have built up a level of trust and mutual respect over time may feel abnormally eager to share data that they might not otherwise. Put into other words in the survey methodology information released by Pew, "This is not an anonymous survey, but the level of trust established between respondents and the survey organization is likely to be high... Considerable research on sensitive issues (such as drug use, sexual behavior and even attendance at religious services) indicates that the online mode of survey administration is likely to elicit more honest answers from respondents on a range of topics." An example of sensitive data that survey participants may not otherwise have shared is their sexual orientation; in other cases, participants may not have felt comfortable indicating their sexual preferences, but under the conditions of trust established between participants and KnowledgePanel, they may have shared that information nonetheless. This possibility necessitates that KnowledgePanel data be treated with extreme carefulness and caution, as the possible harms of survey participants are greatened by their demonstrated likelihood to share more sensitive information in this context than in others.

Of note within the collection stage as well is the payment allotted to survey participants within KnowledgePanel. For those participants with Internet access at home, KnowledgePanel provides payment in the form of 1,000 points for each survey completed, which is equivalent to one dollar. For participants without Internet access, KnowledgePanel will provide a computer and dial-up Internet access as payment for survey completion. Though this latter payment for participants without Internet access is considerable, payment for other participants seems disproportional to the amount of time that is required for full survey completion. A Survey of LGBT Americans is a survey containing over 100 questions, which, if answered with appropriate thought and consideration, should take at least an hour to complete. Survey participants within KnowledgePanel would only be paid $1 for this hour of work! Paying participants more appropriately for their work is necessary, both in order to ensure that participants are incentivized to provide thoughtful answers to all parts of the survey and in order to reward participants for sharing private, personal information that, though unlikely, may expose them to possible privacy risks. As will be discussed, there do exist multiple threats to the privacy of data, and providing appropriate payment to survey participants would likely lessen participants' feelings of abuse or mistreatment if those threats were somehow realized.

One important control that KnowledgePanel has implemented is the use of multiple Institutional Review Boards (IRBs) which are responsible for evaluating their research practices and ensuring that the rights of their survey participants are fully observed and respected. Formally, an Institutional Review Board's goal is to "review and approve research involving human subjects... and to ensure that all human subject research be conducted in accordance with all federal, institutional, and ethical guidelines." IRBs are in charge of weighing the benefits and risks associated with each study and making decisions about whether the study is ultimately appropriate. IRBs also proofread all relevant consent forms and make rules about when and when not consent forms should be required, as we note occurs in this case.

According to the GfK Group's site, 56 IRBS at various colleges and universities have previously reviewed the practices of KnowledgePanel, checking that all of its procedures are compliant with legal and ethical protocols. An important product of this IRB oversight is an established relationship between Principal Investigators (PIs) on a study and their study participants. KnowledgePanel participants are able to contact their PIs if they have questions or concerns about the study or how their information is being used. This valuable connection exemplifies the importance of having IRBs involved and overseeing research, in a position to make useful recommendations about how to improve the process of working with human subjects in research.

B. Transformation Stage

During the transformation stage of the LGBT American survey data, prior to releasing it to the Pew Research Center, researchers at the GfK group de-identified the data, removing personal information such as names, ages, and addresses from the more pertinent survey response data. De-identification is a privacy-preserving technique that researchers have employed for many years in the hopes of protecting the confidentiality of survey participants, for whom the release of identifying information such as names, attached to sensitive survey information, could be disastrous. Unfortunately, research in the field of "re-identification" has proved that de-identifying data does not in fact provide a meaningful privacy guarantee to research participants. According to a 2000 paper by Latanya Sweeney, "About half of the U.S. population... are likely to be uniquely identified by only {place, gender, date of birth}, where place is basically the city, town, or municipality in which the person resides. And even at the county level, {county, gender, date of birth} are likely to uniquely identify 18% of the U.S. population." The implication of Sweeney's research is that even de-identified data is not secure: basic, seemingly non-identifying demographic information such as ZIP code and age can still uniquely identify participants, and by linking that information with other available data, much sensitive information can be uncovered. It is evident that de-identification is no longer a secure technique for data privacy. Various, more secure privacy controls exist in order to more effectively protect data at this crucial transformation stage; these controls will be further discussed in a later section of this analysis.

C. Retention Stage

The GfK Group claims to only store personally identifying information about its survey participants for purposes of payment and other uses. According to the "Frequently Asked Questions" page on the company's website, GfK "only [lets] a select few people inside our company know who [participants] are (because they need to help [participants] on the phone or send [them] checks)." Although it is likely necessary that the company retains personal information about participants for these uses, the availability of this personal, identifying information online, even stored deep within GfK Group's private online servers, poses some threat to participants' privacy. An online attacker may be able to use hacking techniques to penetrate these servers and obtain access to personal files; furthermore, by linking personal data with confidential survey data, an attacker may very well succeed in retrieving sensitive information about KnowledgePanel participants. Several options are available to the GfK Group in order to protect against these attacks, like encryption, two-step verification, Captcha verification, etc. However, none of these controls can guarantee an attack will not succeed, and these attacks therefore remain a viable threat to data privacy and should be kept in mind as such. It may be good practice for the GfK Group to warn survey participants about these possibilities in their online privacy policy, which currently does not address the prospect of online hackers, let alone the other possible privacy risks that have been mentioned in this report.

The data is also presumably stored by the Pew Research Center as well – Pew's website should also be better secured in order to protect privacy leaks at this storage site also.

D. Access and Release Stage

After the data is compiled by the GfK Group, it is passed on to the Pew Research Center, where data analysts look for trends in the data and publish reports about their findings. In the report based on the LGBT data, basic means and counts of the data are reported, along with corresponding figures which display the findings in a visually simple way. In addition to this report, which is publicly available on Pew's website, the entire contents of the LGBT dataset (minus the demographic and identifying information removed from the dataset by the GfK Group) are also made available, free for download from Pew's website as well. Those who wish to download the dataset must supply basic identifying and contact information in order to obtain access, such as name, email address, telephone number, and organization/institution – this may be so that Pew can later follow up with users of the data and see how they have used and analyzed it. The person requesting the data must also accept the terms of a short agreement about how the data may be used. The agreement makes specific reference to the fact that users should not attempt to re-identify the data they download,

stating, "In the event that you discover any... Personally Identifying Information in the Data, you shall immediately notify the Center and refrain from using any such Personally Identifying Information." After the user confirms their acceptance of this agreement, they are given complete access to the data, including a codebook, an informational document about the data, and a SPSS file containing the dataset.

There are several problems with the process of requesting and granting access to Pew datasets such as this one. Firstly, the form required to request access to the datasets is quite short and limited. Although it may be useful to collect contact information from people gaining access to this data in order to be able to contact and question data users later on, the form does not appear to check for accuracy, allowing those seeking access to enter false or inaccurate information that would not in fact prove useful. The form additionally does not request information about the user's intended uses for the data, helpful information that would allow better insight into how the data is being employed, if those uses are legitimate, etc. Though it is unlikely that all users would specify their honest intentions for data usage, a simple space soliciting that information from potential users might work to deter users who are seeking the data for inappropriate purposes. Another possible solution might be to require potential users to register for formal accounts through which they may access data – these accounts could furthermore provide a system for limiting data usage by allowing only registered accounts with certain confirmed credentials to gain access.

The short agreement which users must accept in order to obtain access to data is also in need of potential improvements. The language the agreement uses with respect to re-identification – ("refrain from using any such Personally Identifying Information") is far too weak for the important message that it conveys. Users should not be advised to merely "refrain" from re-identifying data – they should be specifically forbidden from doing so. The flimsy language currently employed by the data usage agreement creates legal ambiguity where there need be none, a quick fix that could be easily rectified.

Of note, however, is the fact that the conditions specified in Pew's terms of use policy are enforceable by law. The agreement specifically states that the "agreement shall be governed by, construed and interpreted in accordance with the laws of the District of Columbia" and confirms that the data user has agreed to "submit to the jurisdiction and venue of the courts of the District of Columbia for any dispute relating to this Agreement."

A threat may also be present in the full version of the LGBT Americans dataset that is available for download. The dataset has been stripped of almost demographic information, other than some relevant characteristics like gender, sexual orientation, marital status, and religion. This is beneficial for privacy purposes, but it does include respondent identifiers for each individual in the dataset. While including a respondent identifier in the data is certainly useful for purposes of organized data analysis, it may pose a privacy risk, for example the respondent identifiers that are included are identical to those used in other surveys. Retaining any unique identifier for survey participants poses some risk, and should be avoided if at all possible.

It is worth noting the considerable utility loss that researchers accept by removing so much useful demographic information. As interesting results could certainly generated by parsing survey results for age-related differences, geographic differences, etc., analysts at the GfK Group accepted a loss of utility by removing that information, likely for the sake of the privacy of survey participants, and this is a behavior that should be applauded and encouraged among social scientists.

E. Post-Access Stage

In its post-access stage, this dataset is made fully available to all those users who have filled out the request form and accepted the data use agreement. As mentioned, the dataset comes in the form of a .zip file which includes a complete codebook, a full SPSS file of the data, and an informational document about how the data was collected. In this way, the user gains full access to the de-identified dataset and to all materials necessary for data processing, allowing him or her to work with the data at his or her skill level. This level of access is associated with privacy threats on several levels.

Firstly, as previously discussed, the growing threat of re-identification renders public data, even de-identified public data, susceptible to release or to leaks at the hands of skilled online attackers. For example, an

attacker with advanced computer science knowledge may be capable of combining the data with other publicly available data in order to cross-reference unique personally identifying information and uncover the identities of individuals included in the dataset. Additionally, an attacker may be able to use the unique identifying numbers in the dataset to again cross-reference other datasets and connect the unique identifiers to named individuals, thereby uncovering sensitive information about them as well.

One particular concern about this dataset is the transparency with which the origin of the data is discussed. Pew makes no secret of the fact that the data was obtained by the Gfk Group through their survey panel KnowledgePanel. Though small, this information is powerful, as it could provide an adversary with an idea of where exactly the survey participants in the data were drawn from. Combined with other information, potentially be available online, that details the kinds of people and households that are represented in the KnowledgePanel participant community, this information could be extremely dangerous. Again, though this risk is small, with so much information and data currently available online, a skilled adversary could use information about the LGBT Americans dataset to learn sensitive information about individuals in the data.

Finally, although Pew does request contact information of users of its data, it does not seem to monitor how data is used in this post-access stage. Pew also does not have a set policy about how to provide redress to respondents harmed by possible privacy releases. The center could be doing a better job of monitoring data usage by following up with data users and establishing a system for how to deal with and address privacy issues that may arise.

F. Aligning Uses, Threats, and Vulnerabilities with Controls

As has been evidenced, there are a number of threats that could harm the survey participants whose information is recorded in the LGBT Americans dataset. In light of this, it is important that a good relationship is established between participants and researchers so that, if any issues were to arise in terms of information leaks or releases, participants would feel more secure, having already built a trusting and respectful connection with the researchers.

One control that should be implemented to begin building this stronger relationship is higher payment for survey participants; those who are putting so much time into furthering research succeed should be more appropriately rewarded for their efforts, and doing so would help to ensure that participants feel better compensated and respected if any privacy leaks were to occur. Participants should also be better advised about these privacy risks – more information about the types of leaks that could occur should be included in the privacy policy and the consent forms used by the GfK Group, so that participants understand that there are risks to their involvement in the research. A full understanding of the privacy threats intrinsic to social science research might discourage some participants from participating, but it is far better to inform participants about possible risks in advance than to face the repercussions if a problem were to arise and participants were not forewarned about that possibility. Furthermore, these controls would work well in tandem: if participants were offered higher payment for their participation, they might be more likely to partake in the research, even if they did understand the risks involved in doing so.

In terms of actually lessening the possibility of threats to data privacy, however, there are several controls that may be implemented as well. One option is to remove all identifying information from the dataset entirely. This means that the unique identifiers assigned to each observation within the dataset would be removed and the data republished without that variable present. As evidenced, unique identifiers can be dangerous to privacy, even if they may seem innocuous, and they should be deleted as such; doing so would also probably cause no utility loss, as the identifiers do not seem to be particularly useful for the analysis of this data.

Another option, however, that would provide a strong guarantee of privacy would be to release only a synthetic version of the data, derived through a differentially private technique known as synthetic data generation. According to the methodological paper explaining "On The Map," a Census software that employs synthetic data generation on commuter data, "The main idea behind synthetic data generation is to build a statistical model from the data and then to sample points from the model. These sampled points form the synthetic data, which is then released instead of the original data." In this way, re-identification of individuals, or learning about individuals, in the data is significantly reduced, as the observations in a synthetic dataset represent

simulated individuals, not real ones. Synthetic data generation offers a lesser privacy guarantee than other privacy techniques, such as epsilon differential privacy, but it does assuage concerns about re-identification, a powerful assurance.

Lastly, both institutions which work with this data, namely the GfK Group and the Pew Research Center, should better protect their websites in order to prevent wrongful uses of their data. In particular, Pew Research Center, the institution which makes this data available, might consider adding a two-step verification or a Captcha verification process to its data access form, as this would likely prevent misuse of the data.

The controls that have been suggested above and more generally throughout this paper may provide some assurance against privacy attacks on this data, and they may also help to better address and rectify privacy problems if they were to arise. Though these controls may not perfectly protect the data in the event of an attack, they likely would not entail any utility losses for either the GfK Group or the Pew Research Center, and therefore they are certainly worthy of consideration and implementation. It would behoove both researchers and their research subjects to do so!

VIII. Weighting and Matching Algorithms

Another assignment I worked on this summer was the construction of differentially private algorithms for employing weights/matching techniques within causal inference, a project I worked on with my mentors James Honaker and Vishesh Karwa. As part of this project, I assisted in researching and devising the algorithms, as well as conducted sensitivity analysis on my own regarding weighting techniques. I also hand-picked two studies, both using weights in their analyses, which can be replicated for the purposes of testing weighting algorithms, once created. One of these studies uses weights produced through coarsened exact matching (CEM); the other uses its own weights based on population demographics. The results of this research are currently being developed within a separate paper entitled "Weights and Matching within Differential Privacy."

VIX. Conclusion

A. Future Work

Although I am proud of the progres I made with this project, there is still much work to be done, and the avenues that I have opened with my research have revealed interesting roadblocks and questions that could be addressed in the future. Certainly there is much more work to be done still within the realm of replication alone; though the four replications that I completed are a great step in the right direction, more replications should be completed and should cover a wider range of statistical techniques that are being developed within Privacy Tools. Replications are an excellent advertisement for the prowess of the differentially private algorithms and more could definitely be used in providing evidence of the utility of differentially privacy in general.

The strides that I took this summer towards developing weighting/matching algorithms within differential privacy were quite productive, but much work remains to be completed here as well to ensure that those algorithms are working functionally and efficiently. Additionally, while I believe that the two replication examples I have chosen for weighting will be useful, more replications could also be used in this realm as well in order to test and streamline the algorithms once they are created.

As part of implementing more replications in the future, more differentially private algorithms may still need to be developed within the Privacy Tools project. An excellent example is regression, for which the differentially private algorithm is still not operational. Future efforts should focus on developing these algorithms and getting them running so that they can be tested using replications and eventually be made public!

B. Acknowledgements

I would like to thank my mentors James Honaker and Vishesh Karwa for their insight and advice as I progressed through this summer and this project. I would not be able to have made the great strides that I did were it not for their help and frequent feedback!

I would also like to thank my fellow Privacy Tools interns for their companionship and assistance this summer, especially Clara Wang and Marcelo Novaes, who were instrumental in helping me get comfortable working with the differentially private algorithms.

X. Bibliography

Altman, Micah, et al. "Towards a Modern Approach to Privacy-Aware Government Data Releases." Berkman Center Research Publication 2016-9 (2016): 30.

Broockman, David E. "Black politicians are more intrinsically motivated to advance Blacks' interests: A field experiment manipulating political incentives." American Journal of Political Science 57.3 (2013): 521-536.

Butler, Daniel M., and David E. Broockman. "Do politicians racially discriminate against constituents? A field experiment on state legislators." American Journal of Political Science 55.3 (2011): 463-477.

D'Orazio, Vito, James Honaker, and Gary King. "Differential Privacy for Social Science Inference." (2015).

Gerber, Alan S., et al. "Can incarcerated felons be (Re) integrated into the political system? Results from a field experiment." American Journal of Political Science 59.4 (2015): 912-926.

Iacus, Stefano M., Gary King, and Giuseppe Porro. "Causal inference without balance checking: Coarsened exact matching." Political analysis (2011).

Machanavajjhala, Ashwin, and Jerome P. Reiter. "Big privacy: protecting confidentiality in big data." XRDS: Crossroads, The ACM Magazine for Students 19.1 (2012): 20-23.

McClendon, Gwyneth H. "Social esteem and participation in contentious politics: A field experiment at an LGBT pride rally." American Journal of Political Science 58.2 (2014): 279-290.

Weitz-Shapiro, Rebecca, and Matthew S. Winters. "Can Citizens Discern? Information Credibility, Political Sophistication, and the Punishment of Corruption in Brazil." (2015).