# Data Privacy in the Digital World

Tianen (Benjamin) Liu

Department of Mathematics and Statistics, Wake Forest University

## Objective

In 2018, Facebook released data of over 87 million users to Cambridge Analytics for political use. Now in the big data-driven world, can we trust the companies to protect our privacy? In addition, years ago, the medical information of the former Massachusetts governor was leaked when hackers identified him by combining anonymous medical data set with anonymous voter list, which share zip code and date of birth information. Now it does not matter whether we trust the companies or not. Even when they are ethical enough to conceal people's sensitive information, it still can be revealed by linkage with another seemingly secure data set.

We as data providers want our sensitive information to be unidentifiable by any means. However, companies depend on the data sets that contain as much information as possible to ensure accurate analysis. Differential privacy sets a privacy level for a public data set that benefits both sides of the trade-off between accuracy and our privacy. It requires techniques to modify a data set and ensures its usability for companies who analyze it without the need to directly access the sensitive information in the original raw data. At the same time, data providers do not have to worry about data breach because, based on the differential privacy, it is unknown if information about them is the actual raw data or not. Moreover, since privacy concerns are reduced, people will be encouraged to contribute to the data set on sensitive topics, such as a drug consumption survey, which requires a large population to ensure better analysis.

## Definition

### • Definition in Practice

Say I am filling out a survey that contains my sensitive information. The sensitive information does not necessarily have to be my name or SSN. It could be anything that seems inconsequential. The company that does the survey will publish the survey results online for people to use. It is unsafe to publish the exact raw data set. Thus we need to modify it so that it is a useful data set that captures the characteristics of the raw data set while also making my information unidentifiable. Then this modified data set could be released. The way we ensure unidentifiability is to make one single answer have no difference on the probability of getting the released data set. If a hacker obtained the released data set, he/she would not be able to know if my information is in there or not, since whether I'm in it does not make a difference in how the released data set looks like.

### • Statistical Definition

Let $I$ be the population whose data are collected, $d_i$ be the information given by person $i$, $D_I = d_i | i \in I$ be the data set collected from all people in $I$, $Q$ be the privatized query run on a data set, and $R = Q(D_I)$ be the resultant modified data set released to the public.

Ideally, since whether one person is in the data set does not impact the released data set, we have $Q(D_{I-me}) = Q(D_I)$ [3]. This should hold whenever, meaning the probability of $Q(D_{I-me})$ being equal to $Q(D_I)$ should be similar. Thus $\epsilon$-differential privacy is defined as [3]:

$$\frac{Prob(Q(D_I) = R)}{Prob(Q(D_{I\pm i}) = R)} \leq e^{\epsilon} \text{ for small } \epsilon \geq 0.$$

### • Properties of Differential Privacy

Even if you did not fill out the survey, a hacker could still learn things about you, based on the analyses of the data in the survey. Also, individual information is protected but group information is not necessarily. A hacker can guess with a high probability whether or not you are in the survey if you are in a known cohesive group. What differential privacy can do is that it ensures the released data set $R$ does not provide additional information of you. Mathematically, $P(secrete(me)|R) = P(secrete(me))$ [3].

## Methods

### • Laplacian Noise

Adding Laplacian noise to a result ensures $\epsilon$-differential privacy in the case of a counting query, i.e., when one is only interested in the count of a certain value. For example, if we were to release the number of smokers in a medical data set, we can add Laplacian noise to this true number of smokers before releasing the count.

Let Global Sensitivity [3] of $F$, a function that modifies data sets, $= \Delta F = max_{(D_1,D_2)}|F(D_1) - F(D_2)|$, which means max difference in answers that adding or removing any individual from the data set can cause. [3]. If we take many noise samples from a Laplace distribution centered at 0 with scale $b = \frac{\Delta F}{\epsilon}$ to be added to true count, the released counts will have a Laplace distribution $Prob(R = x|D) = \frac{\epsilon}{2\Delta F} e^{-\frac{|x-F(D)|\epsilon}{\Delta F}}$ with $\epsilon$-differential privacy [3].

Graphically, as shown in Fig. 1, since 2 neighboring data set that differ only by one row has Laplace distributions that overlap greatly, they have similar probability in generating the same $R$.

### • Randomized Response

Similar to Laplacian noise, randomized response ensures epsilon differential privacy also in the case of a counting query.

Flip a biased coin with probability of heads $\alpha$. If heads, then answer truthfully with $d$. If tails, flip a coin with probability of heads $\beta$ and then answer with one response if heads and the other response if tails. [4]

We control the parameters $\alpha$ and $\beta$ that satisfy differential privacy using an extreme case of definition [4]: $\frac{P[Q(d_{yes},\alpha,\beta)=1]}{P[Q(d_{no},\alpha,\beta)=1]} \leq e^{\epsilon}$ from which we can get $ln(\frac{\alpha+(1-\alpha)\beta}{1-(\alpha+(1-\alpha)\beta)}) \leq \epsilon$.

One disadvantage of Laplacian noise and randomized response is that the relationship is not very well captured, as shown in Fig. 2.

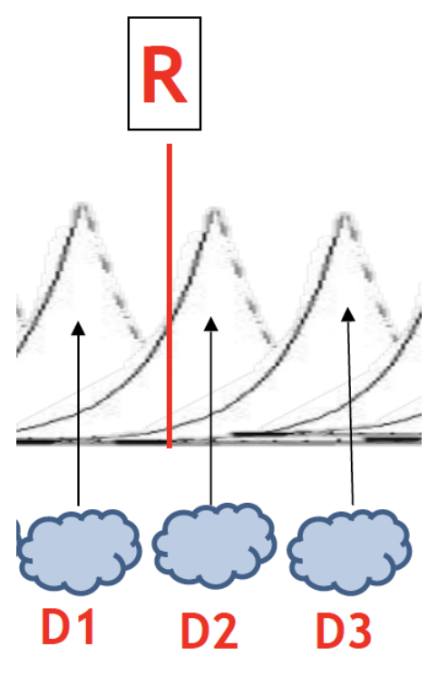

Fig. 1: Laplace Distribution [3]

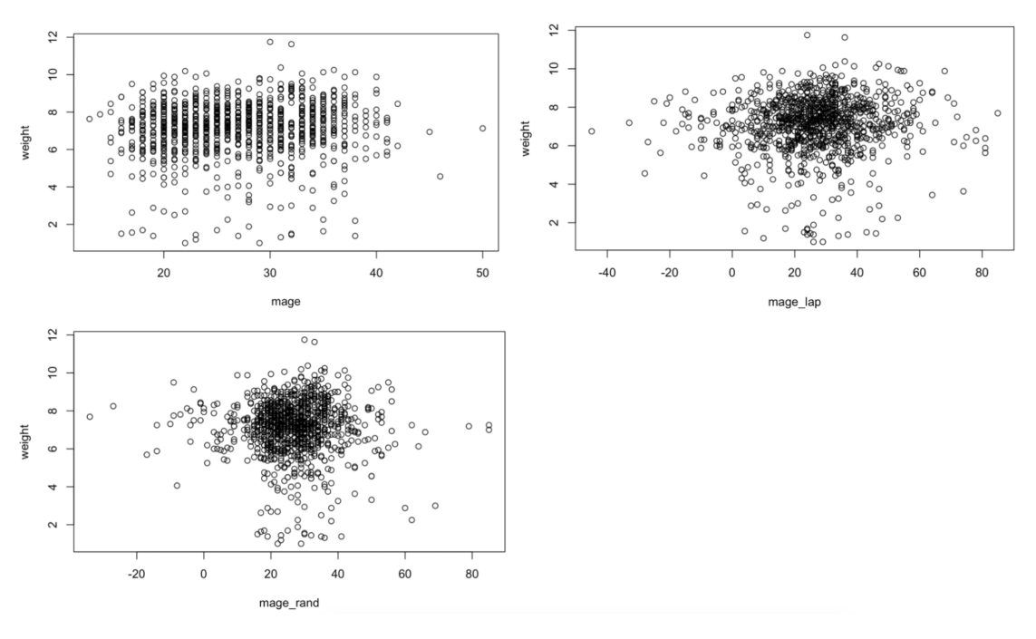

Fig. 2: A Simple Linear Regression Model Fit by Original, Laplacian Noised, and Randomized Data

- binomial logistic (`Discussed_MH_w_Empr ~ .`)

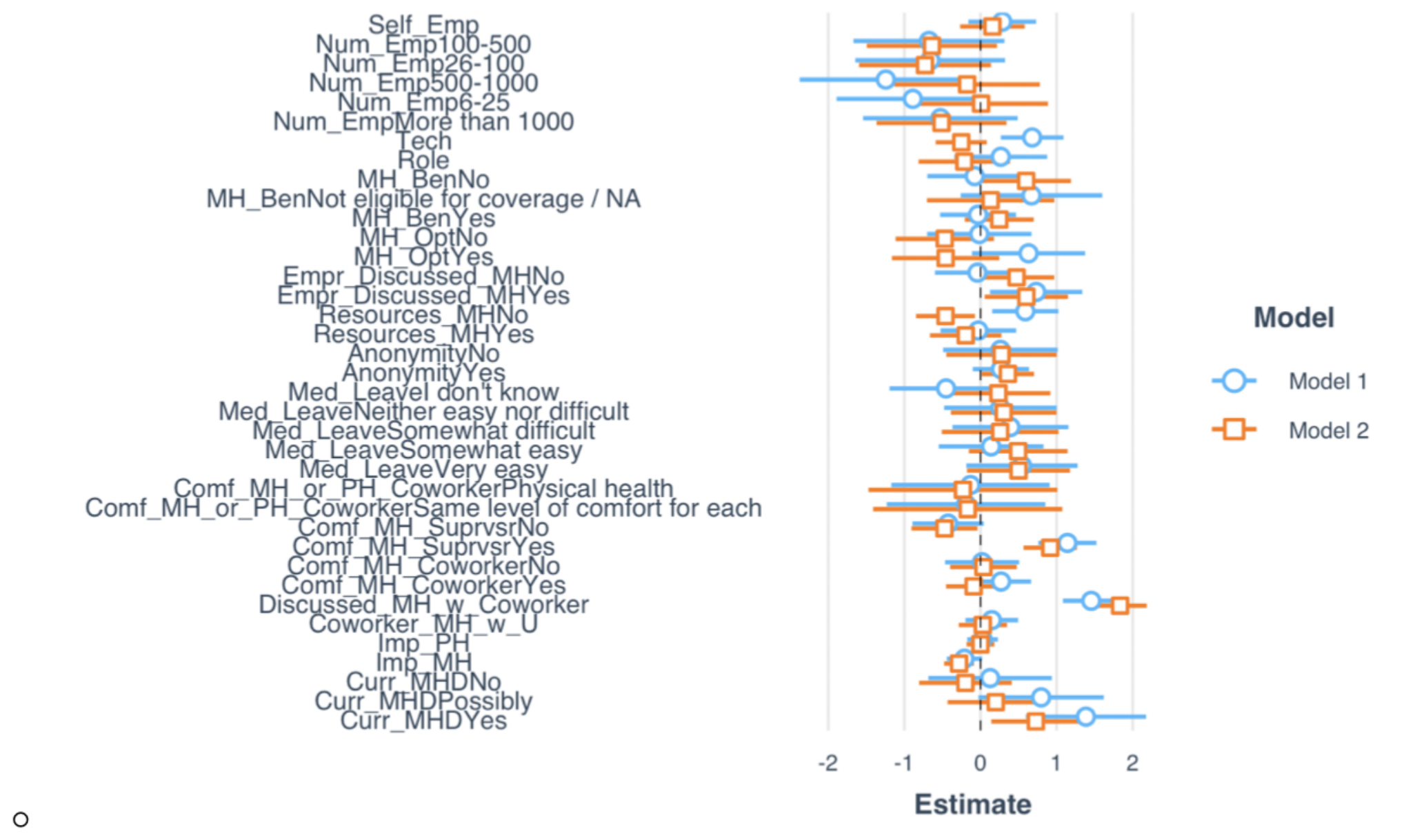

Fig. 3: Comparison of Confidence Intervals of Parameters Fit with Original vs. Synthetic Data Set

### • Bernoulli Posterior Predictive Distribution

Drawing from Bernoulli posterior predictive distribution works for binary data. Let $A$ to be information known by an intruder about records in $D$, and let $S$ to be info known by intruder about process of generating $R$. The probability density function of intruder's probability of knowing $Y_j$ given $(R, A, S)$ is $p(Y_j|R, A, S) \propto p(R|Y_j, A, S)p(Y_j|A, S)$.

The way we generate synthetic binary data is by drawing its value from a Bernoulli posterior predictive distribution based on a $Beta(\alpha_\epsilon, \beta_\epsilon)$ prior distribution whose parameters are turned to satisfy $\epsilon$-differential privacy. This Bernoulli posterior predictive distribution has form

$$p(Y_j^*|R, \alpha_\epsilon, \beta_\epsilon) = \text{Bernoulli}(\frac{\alpha_\epsilon + \sum_{j=1}^{n} Y_j}{\alpha_\epsilon + \beta_\epsilon + n})$$

with $p(Y_i^* = 1|D, \alpha_\epsilon, \beta_\epsilon) = \frac{\alpha_\epsilon + \sum_{j=1}^{n} Y_j}{\alpha_\epsilon + \beta_\epsilon + n}$. We tune $\alpha_\epsilon = \beta_\epsilon = \frac{1}{e^{\epsilon/n_s - 1}}$ to ensure $\epsilon$-differential privacy [2].

## Experiment on OSMI Data Set Using synthpop

The R package `synthpop` creates an entire synthetic data set rather than one variable. The features of the `synthpop` package is that it is possible to choose parametric or non-parametric methods for synthesis, which variables to synthesize first, and whether null data are considered in synthesis, as well as set certain rules of synthesis, such as not synthesizing negative numbers for age.

According to the sequence of synthesis, latter variables are synthesized using previous ones as predictors, which helps capture relationships. By comparing original and synthetic OSMI data sets [1] in logistic regression modeling and inference, we find that the confidence intervals for each parameter have overlap, which explains that the synthetic data set captures similar variability, shown in Fig. 3.

When fitting a multinomial logistic regression, to make the synthetic data set capture the trend, it is necessary to make sure that 1) the sequence of synthesis is ranked from the variables with less missing values to those with more; 2) for the first 3 variables to synthesize, use random sample method. CART does not work well when there are few less correlated predictors; 3) for the variables we are interested in modeling, we synthesize them later so that they would have more predictors when being synthesized; 4) we do not use variables that are less correlated with or do not make sense in predicting the target variable as its predictors.

## Acknowledgements

I would like to express my gratitude to Dr. Nicole Dalzell who helped and guided me through the research process this summer as my research advisor, and to the URECA Center of Wake Forest University for funding it.

## References

[1] Open Sourcing Mental Illness Ltd. *OSMI Mental Health in Tech Survey.* 2017, 2018. URL: `https://osmihelp.org/research`.

[2] David McClure and Jerome P. Reiter. "Differential Privacy and Statistical Disclosure Risk Measures: An Investigation with Binary Synthetic Data". In: *Trans. Data Privacy* 5.3 (Dec. 2012), pp. 535–552. ISSN: 1888-5063. URL: `http://dl.acm.org/citation.cfm?id=2423656.2423658`.

[3] Christine Task. *Privacy-preserving Datamining: Differential Privacy And Applications.* June 2014.

[4] Roberto Agostino Vitillo. *Differential Privacy for Dummies.* July 2016. URL: `https://robertovitillo.com/2016/07/29/differential-privacy-for-dummies/`.

Wake Forest
University