Short Course on Data Privacy

Bei Jiang

University of Alberta

Alberta Machine Intelligence Institute

International Undergraduate Summer Enrichment Program 2025

Why Should We Protect Data?

- Removing directly identifying information is not sufficient.
  - many legendary disclosure incidents due to failed de-identification
- For example:
  - Phd student Sweeney's experiment to identify the then Massachusetts governor's medical data using publicly accessible and insurance data
  - 87% of U.S. citizens are expected to be unique using zip code, gender, and birth date

Why Should We Protect Data?

- Most famously, 2006 Netflix $ 1 million prize competition, inviting researchers to improve the Netflix baseline recommendation algorithm by 10%.
  - The training data: 100 million user data and their ratings on movies (1 to 5 stars)
  - The testing data: 1.5 million user data without any ratings.
  - As required by the Video Privacy Protection Act of 1988, there was no information that could identify a user, like zip-code, birthdate, and of course name, etc.

Why Should We Protect Data?

- Narayanan and Shmatikov were able to connect the individuals in the Netflix dataset to real people, by cross referencing movie ratings in IMDA site (many users post publicly with their own names).
- It end up with privacy breaches, a big lawsuit, and cancelling of Netflix Challenge II.

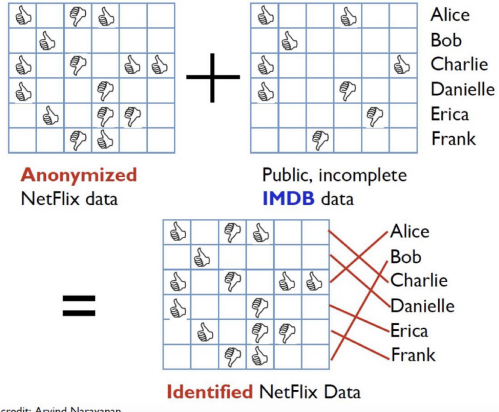


Image credit: Arvind Narayanan

Motivating principles for sharing data

- Transparency/reproducibility → trust in research
- Public good → collective knowledge sharing
- Fairness → reducing barriers to data access
- However, one needs to protect privacy.

How might we protect privacy?

- Removing direct identifiers is not enough
- In a typical dataset, especially those for research purposes, we have:
  - Demographics (e.g., age, sex, race)
  - Sensitive variables (e.g., income, medical status, political opinion)

How might we protect privacy?

Topics covered include:

- Sample new values based on the existing data? [Synthetic Data]
- Add some random noise to final released results? [Differential Privacy]

Other frameworks exist that we will not cover:

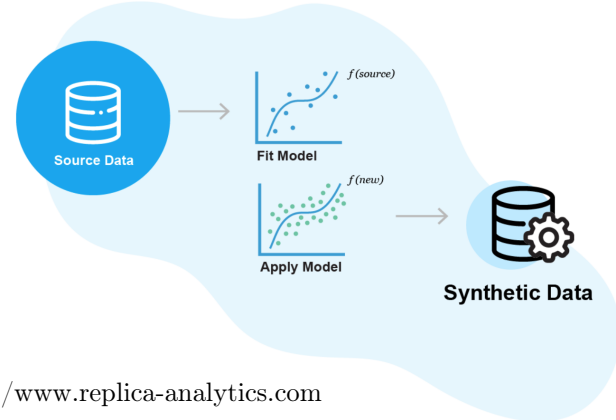
- Distributed learning/federated learning
- k-anonymity: generalization and suppression
- Classical statistical disclosure control methods: data swapping, top and bottom coding

Understanding the Utility-Risk Trade-off is Crucial

- Privacy is never free.
- Fundamentally all methods contain a trade-off:
  - Utility is the accuracy of the released data relative to the confidential data
    - global measures: overall similarity or resemblance of the released data to the confidential data
    - specific measures: analysis-specific measures for certain estimand of interest
  - Risk is the likelihood of learning confidential information: identity disclosure and attribute disclosure
- Generally, high privacy, low utility; low privacy, high utility.
- Goals are to maintain the same level of privacy while maximizing data utility.

What is synthetic data?

- The idea is not new: Rubin (1993) proposed the first model-based synthetic method using multiple imputation (MI):
  - replace the original data by *predictive* values drawn from the posterior predictive distributions
  - use combining rules to analyze the released multiple synthetic datasets.



source: https://www.replica-analytics.com

Challenges

- However, the quality of the resulting synthetic datasets depends on how well the assumed synthesis model would fit the original data. Information loss due to incorrect specification of such a model can weaken or invalidate the inferences obtained using synthetic datasets.
- As discussed by Winkler (1993) that “in producing confidential public-use data files, statistical agencies should first assure that the files are analytically valid", creating synthetic datasets with low quality/utility is not beneficial or worthwhile especially if they lead to incorrect scientific conclusions.

The setup

- All directly identifying variables (e.g., names, addresses) are removed from the confidential dataset.
- The dataset $\mathcal{D} = (X, Z)$ consists of two types of variables:
  - The quasi-identifying variables X (demographic variables, such as age, gender) that provide indirectly identifying information that intruders can use to identify a particular subject.
  - The non-identifying variables Z.
- Here we focus on synthesizing/perturbing X to prevent identity disclosure, while keeping Z unchanged.
  - An *identity disclosure/re-identification* occurs when it is possible to learn that a particular data record belongs to a particular subject.

Rubin's original MI-based synthetic methods

- The imputation model $\mathcal{M}_{imput}$: $p(X|Z,\gamma)$
- The synthetic values of X_i, $i = 1, \cdots, n$ are then drawn from the posterior predictive distribution:

$$P(\tilde{X}_i|\tilde{Z}_i, \mathcal{D}, \mathcal{M}_{imput}) = \int P(\tilde{X}_i|\tilde{Z}_i, \gamma, \mathcal{M}_{imput}) P(\gamma|\mathcal{D}, \mathcal{M}_{imput}) d\gamma,$$

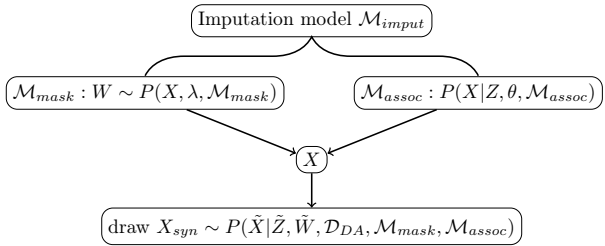
  where $\tilde{Z}_i$ represents an independent copy of Z_i, and $\tilde{X}_i$ represents the posterior predictive replicate of X_i.
- If $\mathcal{M}_{imput}$ does accurately reflect the true association between X and Z, the information loss in the synthetic data will be minimal, leading to a high level of data utility but also minimal privacy protection.

Limitations

- In practice, the assumed $\mathcal{M}_{imput}$ will rarely provide perfect predictions and hence will not perfectly preserve the information in the original dataset.
- The information about X that is preserved in the synthetic data is determined by the assumed association specified by the imputation model $\mathcal{M}_{imput}$.
- The data utility and disclosure risk in the synthetic datasets are determined once $\mathcal{M}_{imput}$ is chosen (more illustration in simulation studies).

The new data-augmented MI synthetic framework by J)

- Our new framework generalizes Rubin's MI framework:



  - We create pseudo-variable W based on X through $\mathcal{M}_{mask}$ and the tuning parameter λ (Details will be given later).
  - We build the joint model $\mathcal{M}_{mask}$ and $\mathcal{M}_{assoc}$ for the augmented $\mathcal{D}_{DA} = (\mathcal{D}, W)$.

$$p(X, W|Z, \lambda, \theta) = \underbrace{p(X|Z, \theta)}_{\mathcal{M}_{assoc}} \underbrace{p(W|X, \lambda)}_{\mathcal{M}_{mask}},$$

The new data-augmented MI synthetic framework

- The synthetic values of X_i, $i = 1, \cdots, n$ are then drawn from the posterior predictive distribution:

$$P(\tilde{X}_i|\tilde{Z}_i, \tilde{W}_i, \mathcal{D}_{DA}, \mathcal{M}_{imput})$$
$$\propto \int P(\tilde{X}_i|\tilde{Z}_i, \theta, \mathcal{M}_{assoc}) P(\tilde{W}_i|\tilde{X}_i, \lambda, \mathcal{M}_{mask}) P(\lambda, \theta|\mathcal{D}_{DA}, \mathcal{M}_{imput}) d\lambda d\theta,$$

  - $\tilde{Z}_i$ and $\tilde{W}_i$ represent independent copies of Z_i and W_i, respectively and $\tilde{X}_i$ represents the posterior predictive replicate of X_i.
  - $\mathcal{M}_{imput}$ consists of $\mathcal{M}_{mask}$ (always correctly specified) and $\mathcal{M}_{assoc}$.

The new data-augmented MI synthetic framework

- Without W, our DA-MI reduces to Rubin's original method.
- With W and $\mathcal{M}_{mask}$, the generated W will provide additional information about X and hence mitigate information loss.
- Further, one could control how much information in X to be transferred to W by tuning through λ under $\mathcal{M}_{mask}$.
- Thus, when $\mathcal{M}_{assoc}$ is misspecified or X is only weakly correlated with Z, $\mathcal{M}_{mask}$ offers protection against information loss and in turn helps balance quality of inference and disclosure risk.
- No need for new inference procedures; the combining rules of Reiter (2003) for inference using MI synthetic data sets are still applicable (regardless of the input data being $\mathcal{D}$ or augmented $\mathcal{D}_{DA}$).

How to choose $\mathcal{M}_{mask}$ to generate W?

- If the confidential X_i is continuous
  - for each X_i, we add random noise with the variance level chosen by the data provider, to produce K pseudo-variable, denoted by $W_i = (W_{ik}, i = 1, \cdots, n; k = 1, \cdots, K)$, according to the masking model $\mathcal{M}_{mask}$,

$$\mathcal{M}_{mask} : W_{ik} = X_i + e_{ik}, \text{ where } e_{ik} \sim_{i.i.d.} N(0, \sigma_e^2), \ k = 1, 2, \cdots, K.$$

  - By varying the values of K and σ_e^2, one can control the amount of information in X_i to be transferred to W_i.

How to choose $\mathcal{M}_{mask}$ to generate W?

- If the sensitive X_i is categorical taking values in $\{0, \cdots, C-1\}$
  - for each X_i, we create pseudo-variable W_i from a mixture distribution of C components according to the masking model $\mathcal{M}_{mask}$,

$$\mathcal{M}_{mask} : W_i = \alpha X_i + u_i, \text{ where } u_i \sim N(0, 1),$$

  - The value of α controls how close the C components are located to one and another; in other words, how accurate one can recover X_i from W_i.
  - Alternatively, W_i could be generated from

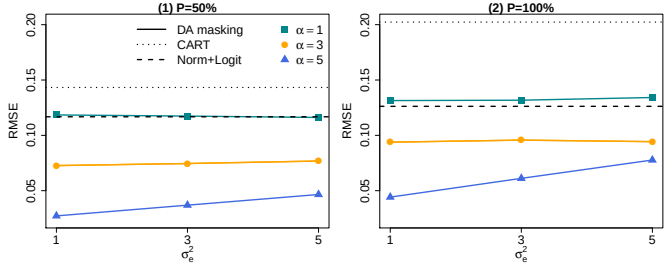
$$\mathcal{M}_{mask} : P(W_i = v | X_i = u) = \pi_{vu}$$

where higher π_{vu} means more information transfer from X_i to W_i.

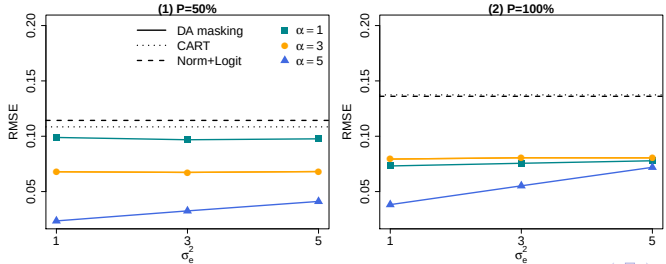
How does the tuning mechanism work?

RMSEs of the estimated coefficients using M synthetic datasets (Utility)

$M = 3$



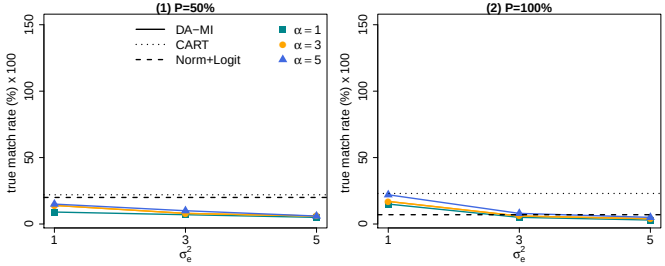
$M = 20$



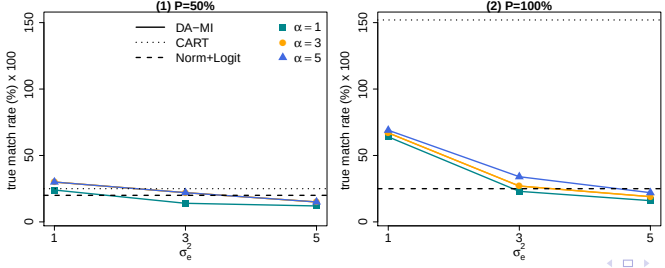
How does the tuning mechanism work?

Disclosure/Re-identification risk

$M = 3$



$M = 20$



Global utitlity measures

- aim to reflect the overall similarities between the distributions of the synthetic data and the confidential data.
- stack up the original dataset and the synthetic dataset resulting in a merged dataset of size $2n$
- and use a classification algorithm (e.g. logistic regression or CART) to predict whether an observation belongs to the original dataset or the synthetic dataset
- return a summary statistic U_p, which measures overall how close the predicted probability of each observation $\hat{p}_i$ is to $\frac{1}{2}$:

$$U_p = \frac{1}{2n} \sum_{i=1}^{2n} \left(\hat{p}_i - \frac{1}{2} \right)^2 .$$

- High level of similarity between the original and the synthetic datasets results in $U_p \approx 0$; low level of similarity results in $U_p \approx \frac{1}{4}$

Specific utility measures

Analysis-specific measures for the estimand of interest:

- The 95% confidence interval (CI) overlap measure, i.e., the percentage overlap of CIs calculated using the original and synthetic data sets, defined as,

$$0.5\left[\frac{\min(U_{\text{orig}}, U_{\text{syn}}) - \max(L_{\text{orig}}, L_{\text{syn}})}{U_{\text{orig}} - L_{\text{orig}}} + \frac{\min(U_{\text{orig}}, U_{\text{syn}}) - \max(L_{\text{orig}}, L_{\text{syn}})}{U_{\text{syn}} - L_{\text{syn}}}\right],$$

  where $U.$ and $L.$ denote the upper and lower bounds of the CI and subscripts, and "orig" and "syn" denote the CI bounds calculated using the original and synthetic data sets respectively. A higher positive value of this interval overlap measure corresponds to higher data utility. A negative value indicates no overlap.

Specific utility measures

Analysis-specific measures for the estimand of interest:

- The standardized difference between the estimates using the original and synthetic data sets, defined as $\left|\hat{\beta}_{\text{orig}} - \hat{\beta}_{\text{syn}}\right| / \text{SE}(\hat{\beta}_{\text{orig}})$, where $\hat{\beta}_{\text{orig}}$ and $\hat{\beta}_{\text{syn}}$ are the estimated coefficients obtained using the original and synthetic data sets respectively, and $\text{SE}(\hat{\beta}_{\text{orig}})$ is the estimated standard error of the coefficient using the original dataset.

Disclosure risk

- Let $\boldsymbol{t}_i$ represents the vector of the quasi-identifiers for individual i, for $i = 1, \cdots, n$. Recall that $\mathcal{D}_{pub} = \{\mathcal{D}_{pub}^{(1)}, \cdots, \mathcal{D}_{pub}^{(M)}\}$ denotes the M released synthetic datasets.
- For each target record $\boldsymbol{t}_i$ known by the intruder, we seek candidate records in each $\mathcal{D}_{pub}^{(m)}$ with the same values for each categorical identifier and with each continuous identifier located within a small range of the target value. When any such record within $\mathcal{D}_{pub}^{(m)}$ is found, we consider it to be a match of $\boldsymbol{t}_i$.
- Then we estimate the probability of matching on the i^{th} individual in each $\mathcal{D}_{pub}^{(m)}$, denoted by $\Pr(J^{\mathbf{t_i}} = i | \boldsymbol{t}_i, \mathcal{D}_{pub}^{(m)})$, by letting t_i be the record of the i^{th} individual and assuming that each of the matched records has equal probability of being picked by the intruder.

Disclosure risk

- If s records are matched on $\boldsymbol{t_i}$, this probability is estimated to be $1/s$; and when there are no defined matches for $\boldsymbol{t}_i$, this probability is estimated to be zero.
- Once $\Pr(J^{\mathbf{t}_i} = i|\boldsymbol{t}_i, \mathcal{D}_{pub}^{(m)})$ is determined for each released $\mathcal{D}_{pub}^{(m)}$, $m = 1, \cdots, M$, we average them across the M released data sets $\mathcal{D}_{pub}$ to obtain the estimate of $\Pr(J^{\mathbf{t}_i} = i|\boldsymbol{t}_i, \mathcal{D}_{pub})$ for each target record $\boldsymbol{t}_i$. That is, the individual-level re-identification risk, $I_i = \frac{1}{M}\sum_{m=1}^{M} \Pr(J^{\mathbf{t}_i} = i|\boldsymbol{t}_i, \mathcal{D}_{pub}^{(m)})$.

Disclosure risk

- Under this framework, we mimic the behaviour of an ill-intentioned intruder who has access to the true values of certain background and demographic information for certain individuals in the dataset, and seeks to identify the records that belong to these individuals in the released synthetic datasets.
- Specifically, we assume that the intruder knows that the target individuals are in the dataset and the values of the quasi-identifiers for all individuals in the dataset. In other words, those variables that are not considered as quasi-identifiers are not used in the calculation of disclosure risks.

Inference using the synthetic datasets

- Let $Q = Q(\mathcal{D})$ be a target estimand defined by the data user, which is a function of the unmasked data $\mathcal{D}$.
- For each synthetic dataset $\mathcal{D}_{pub}^{(m)}$, $m = 1, \cdots, M$, the data user could fit his/her choice of model, $\mathcal{M}_{analy}$, as if $\mathcal{D}_{pub}^{(m)}$ was the actual dataset.
- The target estimand Q is estimated by some estimator $q^{(m)} = q(X_{pub}^{(m)}, Z)$ with an associated variance, estimated via $v^{(m)} = v(X_{pub}^{(m)}, Z)$.
  - Let $\bar{q}_M = \sum_{m=1}^{M} q^{(m)}/M$, $\bar{v}_M = \sum_{m=1}^{M} v^{(m)}/M$ (i.e., average of the within-copy variance) and $b_M = \sum_{m=1}^{M}(q^{(m)} - \bar{q}_M)^2/(M-1)$ (the between-copy variance). Then Q is estimated by $\bar{q}_M$ and the variance of $\bar{q}_M$ is estimated by $T_M = M^{-1}b_M + \bar{v}_M$.
  - When the sample size n is large, the confidence interval for Q is constructed using a $t-$distribution with degrees of freedom $\nu = (M-1)(1 + r_M^{-1})^2$, where $r_M = M^{-1}b_M/\bar{v}_M$. As pointed out by Reiter (2003), in the case of a large enough ν, the normal distribution provides an adequate approximation to such a $t-$distribution.

Focusing on Complete Case

Table: Point estimates and 95% confidence intervals for **only 4** coefficients in a Probit regression model to predict **work disability** when perturbing binary Race, Sex and Education and continuous Age variables.

		M=3			M=20		
	Unperturbed	**CART**	**norm+logit**	**DA-MI**	**CART**	**norm+logit**	**DA-MI**
Age							
Estimate	-0.183	**-0.094**	**-0.08**	-0.195	**-0.084**	**-0.119**	-0.189
95% CI	(-0.312, -0.053)	**(-0.235, 0.048)**	**(-0.222, 0.063)**	(-0.324, -0.066)	**(-0.219, 0.051)**	**(-0.261, 0.022)**	(-0.319, -0.059)
Race							
Estimate	0.518	0.051	0.85	0.514	0.088	0.554	0.514
95% CI	(-0.018, 1.053)	(-0.491, 0.593)	(0.13, 1.57)	(-0.021, 1.049)	(-0.424, 0.601)	(-0.04, 1.147)	(-0.02, 1.049)
Sex							
Estimate	0.082	-0.132	-0.031	0.084	-0.058	0.017	0.016
95% CI	(-0.364, 0.527)	(-0.577, 0.312)	(-0.53, 0.468)	(-0.354, 0.522)	(-0.538, 0.423)	(-0.429, 0.464)	(-0.398, 0.431)
Education							
Estimate	-0.149	-0.041	-0.256	-0.153	-0.052	-0.216	-0.113
95% CI	(-0.435, 0.137)	(-0.332, 0.249)	(-0.717, 0.204)	(-0.444, 0.138)	(-0.349, 0.244)	(-0.522, 0.09)	(-0.401, 0.176)

Our MI-DA method reduces the risk of re-identification to 0% while still preserving 97.6% confidence interval overlap on average.

Focusing on Complete Case

Table: Point estimates and 95% confidence intervals for **only 4** coefficients in a Probit regression model to predict **one-year incidence of insterstitial lung disease (ILD)** when perturbing binary Race, Sex and Education and continuous Age variables.

		M=3			M=20		
	Unperturbed	**CART**	**norm+logit**	**DA-MI**	**CART**	**norm+logit**	**DA-MI**
Age							
Estimate	0.253	0.22	**0.029**	0.233	**0.100**	**0.013**	0.244
95% CI	(0.082, 0.424)	(0.019, 0.421)	**(-0.185, 0.244)**	(0.062, 0.403)	**(-0.072, 0.272)**	**(-0.158, 0.183)**	(0.074, 0.414)
Race							
Estimate	-0.347	0.164	0.215	-0.35	0.178	-0.032	-0.35
95% CI	(-0.857, 0.164)	(-0.526, 0.853)	(-0.602, 1.032)	(-0.858, 0.159)	(-61.472, 61.829)	(-0.611, 0.548)	(-0.861, 0.161)
Sex							
Estimate	0.167	0.4	0.083	0.062	0.138	-0.067	0.16
95% CI	(-0.422, 0.756)	(-0.304, 1.104)	(-0.447, 0.614)	(-0.467, 0.592)	(-0.502, 0.777)	(-0.59, 0.457)	(-0.385, 0.704)
Education							
Estimate	0.002	0.117	0.011	0.015	-0.072	0.046	0.045
95% CI	(-0.359, 0.364)	(-0.306, 0.541)	(-0.419, 0.441)	(-0.35, 0.381)	(-0.45, 0.306)	(-0.329, 0.422)	(-0.32, 0.41)

Note, one-year ILD is a new collected variable not used in the original data synthetic process.

CSRG Study with Missing Data

Table: Point estimates and 95% confidence intervals for **only 4** coefficients in both Work Disability and ILD when perturbing binary Race, Sex and Education and continuous Age variables.

	Predict Work Disability				Predict Onset of ILD			
	Unperturbed MI	**DA-MI**	**CART**	**Norm+Logit**	**Unperturbed MI**	**DA-MI**	**CART**	**Norm+Logit**
Age								
Estimate	-0.065	-0.082	-0.044	0.002	0.301	0.293	**0.138**	**-0.077**
95% CI	(-0.174, 0.043)	(-0.19, 0.026)	(-0.164, 0.076)	(-0.107, 0.11)	(0.042, 0.56)	(0.06, 0.526)	**(-0.189, 0.466)**	**(-0.343, 0.189)**
Gender								
Estimate	-0.017	-0.012	-0.072	-0.043	0.751	0.599	0.187	-0.238
95% CI	(-0.306, 0.272)	(-0.3, 0.277)	(-0.322, 0.179)	(-0.327, 0.24)	(-0.092, 1.595)	(-0.238, 1.437)	(-0.594, 0.969)	(-0.778, 0.302)
Race								
Estimate	0.19	0.209	0.086	0.002	-0.286	-0.277	0.186	0.147
95% CI	(-0.156, 0.537)	(-0.138, 0.556)	(-0.282, 0.454)	(-0.323, 0.326)	(-0.875, 0.303)	(-0.89, 0.336)	(-0.722, 1.093)	(-0.673, 0.966)
Education								
Estimate	-0.306	-0.304	**-0.184**	**-0.189**	-0.11	-0.117	-0.073	0.121
95 %CI	(-0.519, -0.093)	(-0.517, -0.091)	**(-0.422, 0.055)**	**(-0.404, 0.027)**	(-0.525, 0.306)	(-0.507, 0.273)	(-0.632, 0.486)	(-0.302, 0.544)

- DA-MI method can be combined with existing missing data methods to impute missing values.

Other challenges

- Most existing methods treat each individual equally in the synthesis model, resulting in no special treatment of high or low risk records.
- Individuals at higher risk of disclosure may not get sufficient protection, while those individuals at relatively lower risk of disclosure may get excessive protection and hence unnecessary information loss.
- Some authors propose to take into account the disclosure risks when designing their synthetic methods, although this is still under-explored in the literature.

Other challenges

- Little et al. (2004) developed an algorithm to synthesize the identifying variables for only a group of selected high risk individuals, which can be impractical or inefficient when the size of this selected subset is small.
- Hu and Williams (2021) proposed a risk-weighted synthetic method that down weighs the likelihood contributions from high-risk records.
- However, these methods may still suffer from potential imputation model mis-specification.

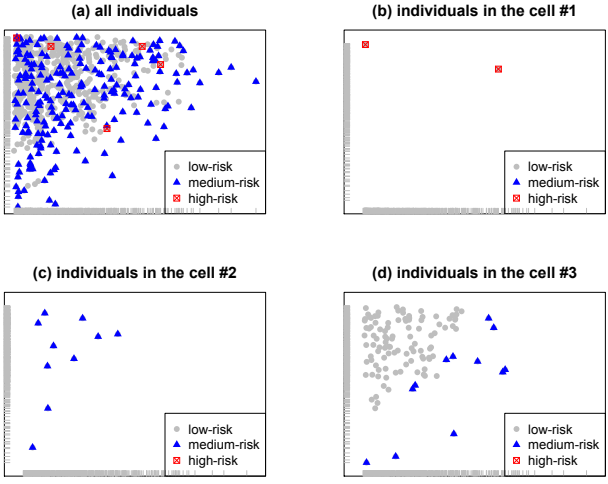
Revisit CSRG patient registry data

We tackle the more challenging task of perturbing the following quasi-identifying (QI) variables:

- Disease duration (in years)
- Age (in years)
- Work disability status (yes, no)
- Sex (female, male)
- Race (white, non-white)
- Education ($<$ high school, high school, $>$ high school)

Risk subgroups (informal definition)

Scatter plots of the age and disease duration variables (with axis labels removed)



- Low risk: have many neighbouring/matching individuals (i.e., sharing similar values of QIs).
- Medium risk: share the same values of categorical QIs with at least k_0 others, but less than k_0 of them have continuous QI values close enough
- High risk: otherwise

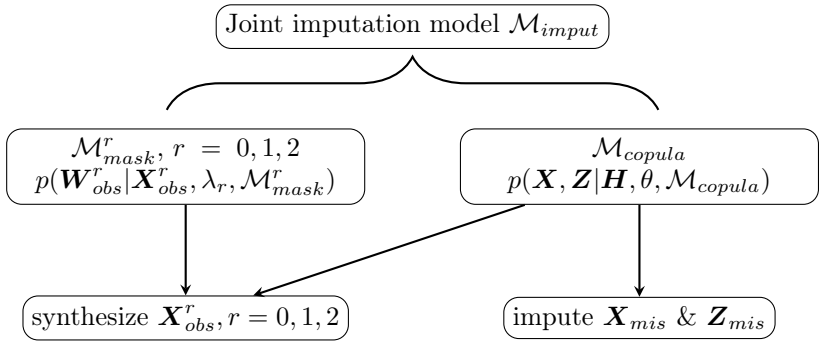
The joint subgroup-targeted imputation model

With the generated pseudo-observations $\boldsymbol{W}_{obs} = (\boldsymbol{W}^0_{obs}, \boldsymbol{W}^1_{obs}, \boldsymbol{W}^2_{obs})$ and the originally observed $\mathcal{D}_{obs} = (\boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs}, \boldsymbol{H})$, we build the following joint imputation model, denoted by $\mathcal{M}_{imput}$ using $\mathcal{D}_{DA} = (\mathcal{D}_{obs}, \boldsymbol{W}_{obs})$:

$$
\begin{aligned}
& p(\boldsymbol{X}_{obs}, \boldsymbol{Z}_{obs}, \boldsymbol{W}_{obs} | \boldsymbol{H}) \\
= & \int p(\boldsymbol{X}_{obs}, \boldsymbol{X}_{mis}, \boldsymbol{Z}_{obs}, \boldsymbol{Z}_{mis}, \boldsymbol{W}_{obs} | \boldsymbol{H}) d\boldsymbol{Z}_{mis} d\boldsymbol{X}_{mis} \\
= & \int \underbrace{p(\boldsymbol{X}_{obs}, \boldsymbol{X}_{mis}, \boldsymbol{Z}_{obs}, \boldsymbol{Z}_{mis} | \boldsymbol{H}, \theta)}_{\mathcal{M}_{copula}} \underbrace{p(\boldsymbol{W}^0_{obs} | \boldsymbol{X}^0_{obs}, \lambda_0)}_{\mathcal{M}^0_{mask}} \times \cdots \times \underbrace{p(\boldsymbol{W}^2_{obs} | \boldsymbol{X}^2_{obs}, \lambda_2)}_{\mathcal{M}^2_{mask}} d\boldsymbol{Z}_{mis} d\boldsymbol{X}_{mis}
\end{aligned}
$$

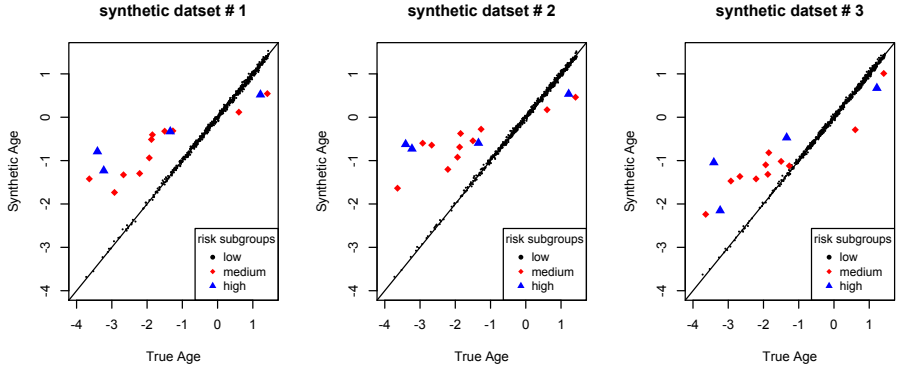
A risk-subgroup targeted synthesis scheme

Figure: A graphical representation of our joint imputation model for $\mathcal{D}_{DA}$ in the presence of re-identification risk subgroups.



How does this framework work?

Tuning leads to different degrees of perturbation to Age for different risk subgroups:



Utility of synthetic datasets: predict work disability

Partial table:

	Estimates (95% CI)			Std. Diff		95% CI Overlap	
	MI-complete	Risk-based DA-MI	CART	Risk-based DA-MI	CART	Risk-based DA-MI	CART
Diffuse							
Estimate	0.318	0.311	0.096	0.058	1.762	0.986	0.539
95% CI	(0.071, 0.564)	(0.065, 0.556)	(-0.138, 0.33)				
Centro Positive							
Estimate	-0.022	-0.035	0.064	0.09	0.581	0.978	0.845
95% CI	(-0.312, 0.267)	(-0.325, 0.254)	(-0.198, 0.326)				
Disease Duration							
Estimate	0.178	0.12	0.073	0.97	1.75	0.775	0.545
95% CI	(0.061, 0.296)	(-0.018, 0.257)	(-0.041, 0.187)				
FVC							
Estimate	-1.298	-1.3	0.174	0.003	2.491	0.996	0.34
95% CI	(-2.457, -0.138)	(-2.449, -0.15)	(-0.895, 1.244)				
DLCO							
Estimate	0.142	0.146	-0.246	0.01	1.054	0.983	0.722
95% CI	(-0.587, 0.87)	(-0.558, 0.849)	(-0.895, 0.404)				
TLC							
Estimate	0.499	0.456	0.001	0.07	0.81	0.982	0.787
95% CI	(-0.711, 1.708)	(-0.731, 1.643)	(-1.107, 1.11)				
Vitality							
Estimate	-0.548	-0.561	-0.114	0.045	1.507	0.988	0.605
95% CI	(-1.111, 0.016)	(-1.123, 0.001)	(-0.646, 0.418)				
HAQ							
Estimate	1.325	1.29	0.029	0.191	7.043	0.952	-1.026
95% CI	(0.965, 1.684)	(0.931, 1.649)	(-0.255, 0.313)				

Summary

- Our new risk-subgroup targeted synthesis framework preserves all strengths of the DA-MI synthetic framework in Jiang et al. (2022):
  - Protects against imputation model misspecification by introducing the data-augmentation masking step.
  - Flexibility in balancing disclosure risk and data utility through tuning
  - Simplicity: with valid inferences for the MI synthetic datasets can be obtained using simple combing rule.
- It adds additional features:
  - simultaneously imputing missing data for mixed categorical and continuous variables
  - provides subgroup specific perturbation schemes to suit the specific privacy protection needs of different risk subgroups.

Motivation for a New Privacy Framework

- Assumptions on attacker knowledge on participants
- Assumptions on publicly accessible variables
- Lack of consensus on risk definitions for synthetic data

Differential Privacy with provable guarantee

- A mechanism $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is ε-differentially private (DP) for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}$ such that $\delta(\mathbf{x}, \mathbf{y}) = 1$:

$$\left| \ln \left(\frac{\Pr[\mathcal{M}(\mathbf{x}) \in \mathcal{S}]}{\Pr[\mathcal{M}(\mathbf{y}) \in \mathcal{S}]} \right) \right| \leq \varepsilon$$

  Equivalently,

$$\Pr[\mathcal{M}(\mathbf{x}) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{M}(\mathbf{y}) \in \mathcal{S}]$$

  where $\delta(\mathbf{x}, \mathbf{y})$ is the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$ (that is, any two neighbouring datasets $\mathbf{x}$ and $\mathbf{y}$ that differ by one record or remove a record from $\mathbf{x}$ to get $\mathbf{y}$).
- How to achieve DP?
  - Add noise to the output of queries (i.e., summary statistics or any quantities derived from the data) made to databases
  - Added noise is random, which depends on a pre-determined privacy budget ϵ and the queries.

The global sensitivity

- Let f be some function $\mathcal{X}^n \to \mathbb{R}^k$. The ℓ_1-sensitivity of f is

$$\Delta f = \max_{X,X'} \|f(X) - f(X')\|_1$$

  where X and X' are two neighboring databases.
- The concept of ℓ_1 sensitivity of query (function) f is to quantify the maximum potential change in the ℓ_1 norm of the query f caused by the data of a single individual in the worst-case scenario.
- For example, f can be
  - Sample mean
  - Sample median
  - Coefficient of linear regression

The Laplace mechanism is ϵ-DP.

- Let f be a query (function) $\mathcal{X}^n \to \mathbb{R}^k$. The Laplace mechanism is defined as

$$\mathcal{M}(\mathbf{x}, f(\cdot), \epsilon) = f(\mathbf{x}) + (Y_1, \ldots, Y_k)$$

  where Y_i are independent Laplace$(\Delta f/\varepsilon)$ random variables.

The Laplace distribution

A random variable has a Laplace (μ, s) distribution if its probability density function is

$$
\begin{aligned}
f(x \mid \mu, s) &= \frac{1}{2s} \exp\left(-\frac{|x-\mu|}{s}\right) \\
&= \frac{1}{2s} \begin{cases} \exp\left(-\frac{\mu - x}{s}\right) & \text{if } x < \mu; \\ \exp\left(-\frac{x-\mu}{s}\right) & \text{if } x \geq \mu, \end{cases}
\end{aligned}
$$

- μ is a location parameter
- $s > 0$ is a scale parameter
- when $\mu = 0, b = 1$, the positive half-line is an exponential distribution scaled by $\frac{1}{2}$.

Proof

Proof: Let $\mathbf{x}, \mathbf{y} \in \mathbb{N}^{|\mathcal{X}|}$ and $\delta(\mathbf{x}, \mathbf{y}) = 1$, and let $f(\cdot)$ be some function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$. Let $p_{\mathbf{x}}$ and $p_{\mathbf{y}}$ denote the probability density functions of $\mathcal{M}(\mathbf{x}, f(\cdot), \epsilon)$ and $\mathcal{M}(\mathbf{y}, f(\cdot), \epsilon)$. We compare the two at some arbitrary output point $z \in \mathbb{R}^k$:

$$\begin{aligned}
\frac{p_{\mathbf{x}}(z)}{p_{\mathbf{y}}(z)} &= \frac{\prod_{i=1}^{k} \exp\left(-\frac{\varepsilon |f(\mathbf{x})_i - z_i|}{\Delta f}\right)}{\prod_{i=1}^{k} \exp\left(-\frac{\varepsilon |f(\mathbf{y})_i - z_i|}{\Delta f}\right)} = \prod_{i=1}^{k} \exp\left(-\frac{\varepsilon \left(|f(\mathbf{x})_i - z_i| - |f(\mathbf{y})_i - z_i|\right)}{\Delta f}\right) \\
&\leq \prod_{i=1}^{k} \exp\left(\frac{\varepsilon \, |f(\mathbf{y})_i - f(\mathbf{x})_i|}{\Delta f}\right) \\
&= \exp\left(\frac{\varepsilon \sum_{i=1}^{k} |f(\mathbf{x})_i - f(\mathbf{y})_i|}{\Delta f}\right) \\
&= \exp\left(\frac{\varepsilon \|f(\mathbf{x}) - f(\mathbf{y})\|_1}{\Delta f}\right) \\
&\leq \exp(\varepsilon)
\end{aligned}$$

Example: mean under DP guarantee

- $f(\mathcal{D})$ is sample mean of a dataset $\mathcal{D}$, where each record is a scalar in $[0, 1]$.
- So global sensitivity of f is $1/n$, where n is the sample size.
- **Laplace mechanism:** output sample mean + Z, where $Z \sim \frac{1}{n\epsilon}\,\mathrm{Lap}(0, 1)$, ϵ is the privacy budget.
- How about each record contains a continuous value (unbounded)?

Some properties of Differential Privacy

- Any function of an output that satisfies $\epsilon-$DP is also $\epsilon-$DP (post-processing property).
- The privacy loss as defined is guaranteed, which does not require assumptions about the attacker.
- The amount of privacy loss is quantifiable, in the form of privacy budget ϵ. It needs to be added across multiple releases (next: composition theorem).

Composition theorem

- Formally, let $\mathcal{M}_1 : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_1$ be an ϵ_1-differentially private algorithm, and let $\mathcal{M}_2 : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_2$ be an ϵ_2-differentially private algorithm. Then their combination, defined to be $\mathcal{M}_{1,2} : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_1 \times \mathcal{R}_2$ by the mapping: $\mathcal{M}_{1,2}(\mathbf{x}) = (\mathcal{M}_1(\mathbf{x}), \mathcal{M}_2(\mathbf{x}))$ is $(\epsilon_1 + \epsilon_2)$-differentially private
- A generalization: let $\mathcal{M}_i : \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}_i$ be an ϵ_i-differentially private algorithm for $i \in [k]$. Then if $\mathcal{M}_{[k]} : \mathbb{N}^{|\mathcal{X}|} \to \prod_{i=1}^{k} \mathcal{R}_i$ is defined to be $\mathcal{M}_{[k]}(\mathbf{x}) = (\mathcal{M}_1(\mathbf{x}), \cdots, \mathcal{M}_k(\mathbf{x}))$, then $\mathcal{M}_{|k|}$ is $\left(\sum_{i=1}^{k} \epsilon_i\right)$-differentially private.

What are Drawbacks to Differential Privacy?

- How about complex statistics and DP synthetic data?
  - Counts/histograms are the most developed and deployed in practice.
- The meaning of the privacy loss is less intuitive: $\epsilon = 0.1$ vs. $\epsilon = 1$.
  - Compared with measures such as risk of re-identification
- No consensus on how to set the privacy parameter/budget ϵ

Randomized response $\epsilon-$differentially private

Input: Data set of n bits: $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n, \varphi : \mathcal{X} \to \{0, 1\}$, and a parameter $\varepsilon > 0$

Output: Bits $Y_1, \ldots, Y_n$

1 for $i = 1$ to n do

2 $Y_i = \begin{cases} \varphi(x_i) & \text{w.p. } \frac{e^\varepsilon}{e^\varepsilon + 1} \\ \varphi(1 - x_i) & \text{w.p. } \frac{1}{e^\varepsilon + 1} \end{cases}$

3 return $(Y_1, \ldots, Y_n)$;

Proof: $\mathcal{RR}_\epsilon$ differentially private

Fix two neighboring data sets $\mathbf{x}$ and $\mathbf{x}'$, and let i be the position in which they differ (so that $x_i \neq x_i'$ but $x_j = x_j'$ for all $j \neq i$). First, consider a particular outcome $\mathbf{y} = (y_1, \ldots, y_n)$. Because we make selections independently for each i, we have

$$\mathbb{P}\left(RR_\varepsilon(\mathbf{x}) = \mathbf{y}\right) = \mathbb{P}\left(Y_1 = y_1 \mid x_1\right) \cdot \mathbb{P}\left(Y_2 = y_2 \mid x_2\right) \cdots \mathbb{P}\left(Y_n = y_n \mid x_n\right)$$

When we compare this to the probability that $RR_\varepsilon\left(\mathbf{x}'\right) = \mathrm{y}$, only one of the terms in the product will change. We thus get that

$$\frac{\mathbb{P}\left(RR_\varepsilon\left(\mathbf{x}'\right) = \mathbf{y}\right)}{\mathbb{P}\left(RR_\varepsilon(\mathbf{x}) = \mathbf{y}\right)} = \frac{\mathbb{P}\left(Y_i = y_i \mid x_i'\right)}{\mathbb{P}\left(Y_i = y_i \mid x_i\right)}$$

This ratio is at most $\frac{e^\varepsilon}{e^\varepsilon+1} / \frac{1}{e^\varepsilon+1} = e^\varepsilon$. Now let's take any subset $E \subseteq \mathcal{Y} = \{0,1\}^n$. The probability that $RR_\varepsilon(\mathbf{x})$ lies in E is just the sum over $\mathbf{y} \in \mathbf{E}$ of the probability that $RR_\varepsilon(\mathbf{x}) = \mathbf{y}$. We thus get

$$\mathbb{P}\left(RR_\varepsilon(\mathbf{x}) \in E\right) = \sum_{\mathbf{y} \in E} \mathbb{P}\left(RR_\varepsilon(\mathbf{x}) = \mathbf{y}\right) \leq \sum_{\mathbf{y} \in E} e^\varepsilon \cdot \mathbb{P}\left(RR_\varepsilon\left(\mathbf{x}'\right) = \mathbf{y}\right) = e^\varepsilon \cdot \mathbb{P}\left(RR_\varepsilon\left(\mathbf{x}'\right) \in E\right).$$

This completes the proof.

Composition

let $A_1 : \mathcal{X}^n \to \mathcal{Y}_1$ be ε_1-DP, and let $A_2 : \mathcal{Y}_1 \times \mathcal{X}^n \to \mathcal{Y}_2$ be ε_2-DP for all values of its first input (that is $A_2(a_1, \cdot)$ is ε_2-DP for every value of a_1). Let $A : X^n \to \mathcal{Y}_1 \times \mathcal{Y}_2$ be the randomized algorithm that outputs $A(\mathbf{x}) = (a_1, a_2)$ where $a_1 = A_1(\mathbf{x})$ and $a_2 = A_2(a_1, \mathbf{x})$. Then A is $(\varepsilon_1 + \varepsilon_2) - DP$.

Proof

We prove the discrete case here, for simplicity. Let $\mathbf{x}, \mathbf{x}'$ be neighboring data sets in $\mathcal{X}^n$, and let $a = (a_1, a_2)$ be an outcome in $\mathcal{Y}_1 \times \mathcal{Y}_2$.

$$\mathbb{P}\left(A(\mathbf{x}) = (a_1, a_2)\right) = \mathbb{P}\left(A_1(\mathbf{x}) = a_1\right) \cdot \mathbb{P}\left(A_2\left(\mathbf{x}, a_1\right) = a_2\right)$$

Since A_1 is ε_1-DP, and $A_2\left(a_1, \cdot\right)$ is ε_2-DP for every choice of a_1, we can bound the probability above.

$$\begin{aligned}\mathbb{P}\left(A(\mathbf{x}) = (a_1, a_2)\right) &\leq e^{\varepsilon_1}\mathbb{P}\left(A_1\left(\mathbf{x}'\right) = a_1\right) \cdot e^{\varepsilon_2}\mathbb{P}\left(A_2\left(\mathbf{x}', a_1\right) = a_2\right) \\ &= e^{\varepsilon_1+\varepsilon_2} \cdot \mathbb{P}\left(A\left(\mathbf{x}'\right) = (a_1, a_2)\right)\end{aligned}$$

Post-processing

Let $A : \mathcal{X}^n \to \mathcal{Y}$ and $B : \mathcal{Y} \to \mathcal{Z}$ be randomized algorithms, where $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are arbitrary sets. If A is ε-differentially private, then so is the composed algorithm $B(A(\cdot))$.

- When B is deterministic. In that case, the event $B(A(\mathbf{x})) = b$ is the same as the event $A(\mathbf{x}) \in B^{-1}(b)$ where $B^{-1}(b)$ is the preimage of b under B. So we can just apply the A's DP guarantee to $B^{-1}(b)$:

$$\mathbb{P}(B(A(\mathbf{x})) = b) = \mathbb{P}\left(A(\mathbf{x}) \in B^{-1}(b)\right) \leq e^{\varepsilon}\mathbb{P}\left(A\left(\mathbf{x}'\right) \in B^{-1}(b)\right) = e^{\varepsilon}\mathbb{P}(B(A(\mathbf{x}')) = b)$$

Post-processing

- To handle the case where B is randomized, we can write the $B(a)$ as the application of a deterministic function f applied to the pair (a, R) where R is a random variable independent of a that represents B 's random choices. Thus, $B(A(\cdot))$ is the application of a deterministic function to $A'(\mathbf{x}) = (A(\mathbf{x}), R)$. The algorithm A' is ε-DP (since R is independent of A). Thus $B(A(\cdot))$ is also ε-DP.

Example: $k-$means Clustering

Let's use some of the tools we now have—the Laplace mechanism and basic composition—to design a more complex algorithm. First, let's review the original algorithm given below.

Algorithm 1: Lloyd's algorithm with random initialization

Input: Data set $\mathbf{x} \in \mathcal{X}^n$ where $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_1 = 1\right\}$, parameter k

1 Initialize $c_1^{(0)}, c_2^{(0)}, ..., c_k^{(k)}$ randomly in $\mathcal{X}$;

2 **for** $t = 1$ *to* T **do**

3 **for** $j = 1$ *to* k **do**

4 $S_j = \left\{i : c_j^{(t-1)} \text{ is the closest current center to } x_i\right\}$;

5 $c_j^{(t)} = \frac{1}{|S_j|} \sum_{i \in S_j} x_i$;

6 **return** $c_1^{(T)}, c_2^{(T)}, ..., c_k^{(T)}$;

Idea to achieve DP mechanism

- We apply the Laplace mechanism to get noisy versions of the each of the algorithm's intermediate steps. We can divide our privacy budget ϵ into T parts, and assign ϵ/T to each intermediate step.
- Suppose we have already released centers $c_1^{(t-1)}, \ldots, c_k^{(t-1)}$ from the previous step. Then we can divide the universe $\mathcal{X}$ into k regions $B_1, \ldots, B_k$, where B_j consists of points closest to center $c_j^{(t-1)}$. To compute the next set of centers, we approximate two quantities for each B_j :
  - n_j (integer): the number of data records in B_j, and
  - a_j (vector in $\mathbb{R}^d$): the sum of the data records in B_j (as vectors)

Idea to achieve DP mechanism

- The counts $(n_1, \ldots, n_k)$ form a histogram. Their global sensitivity is thus 2 . Releasing $\hat{n}_1, \ldots, \hat{n}_k$ by adding noise Lap $\left(\frac{4T}{\varepsilon}\right)$ to each histogram entry thus consumes at most $\frac{\varepsilon}{2T}$ of our privacy budget.
- Similarly, we can view the sums $a_1, \ldots, a_k$ as one long vector of length kd. If we change one record in the data set, only two of the sums a_j can change, since the record either stays in the same bin or moves from one bin to another. These two sums gain or lose one term each, of ℓ_1 norm at most 1. The change in the long vector is thus at most 2 .
- Again, the algorithm adds noise Lap $\left(\frac{4T}{\varepsilon}\right)$ to each entry, consuming another $\frac{\varepsilon}{2T}$ of our privacy budget. The computation of the next cluster center is just postprocessing of the $\hat{n}_j$'s and $\hat{a}_j$'s, so it consumes no further budget.
- The total expenditure for the T step is thus $\frac{\varepsilon}{T}$. By Basic Composition, the algorithm as a whole is ε-DP.

Example: $k-$means Clustering

Algorithm 2: A differentially private version of Lloyd's algorithm

Input: Data set $\mathbf{x} \in \mathcal{X}^n$ where $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_1 = 1\right\}$, parameter k and privacy parameter $\varepsilon > 0$

$\varepsilon' = \frac{\varepsilon}{2T}$ (since we will compose $2T$ executions of the Laplace mechanism);

Initialize $c_1^{(0)}, c_2^{(0)}, ..., c_k^{(k)}$ randomly in $\mathcal{X}$;

for $t = 1$ *to* T **do**

  for $j = 1$ *to* k **do**

    $S_j = \left\{i : c_j^{(t-1)} \text{ is the closest current center to } x_i\right\}$;

    $n_j = |S_j|$ (this has global sensitivity 2 across all j);

    $a_j = \sum_{i \in S_j} x_i$ (this has global sensitivity 2 across all j);

    Release $\hat{n}_j = n_j + Y$ where $Y \sim \text{Lap}\left(\frac{2}{\varepsilon'}\right)$;

    Release $\hat{a}_j = a_j + (Z_1, ..., Z_d)$ where $Z_t \sim \text{Lap}\left(\frac{2}{\varepsilon'}\right)$ i.i.d.;

    $$c_j^{(t)} = \begin{cases} \frac{\hat{a}_j}{\hat{n}_j} & \text{if } \hat{n}_j \geq 1 \\ \text{uniform in } \mathcal{X} & \text{if } \hat{n}_j < 1 \end{cases};$$

return $c_1^{(T)}, c_2^{(T)}, ..., c_k^{(T)}$ (*NB:* One can also average over the last few iterations to reduce variance);

The exponential mechanism

- A set $\mathcal{Y}$ of possible outputs;
- A score function $q : \mathcal{Y} \times \mathcal{U}^n \to \mathbb{R}$ which measures the "goodness" of each output for a data set. Given $\mathbf{x} \in \mathcal{U}^n$, our goal is to find $y \in \mathcal{Y}$ which approximately maximizes $q(y; \mathbf{x})$. (When $\mathcal{Y}$ is finite, we can also think of q as a collection of $\mathcal{Y}$ separate low-sensitivity queries.)
- A sensitivity bound $\Delta > 0$ such that $q(y; \cdot)$ is Δ-sensitive for every y. That is, $\sup_{y \in \mathcal{Y}} \sup_{\substack{\mathbf{x}, \mathbf{x}' \in \mathcal{U}^n \\ \text{adjacent}}} |q(y; \mathbf{x}) - q(y; \mathbf{x}')| \leq \Delta$

The exponential mechanism

- The idea is that given the score function $q(\cdot; \mathbf{x})$ that assigns a number to each element $y \in \mathcal{Y}$, we define a probability distribution which generates each element in y in $\mathcal{Y}$ with probability proportional to $\exp\left(\frac{\varepsilon}{2\Delta} q(y; \mathbf{x})\right)$; that is, we sample elements with a probability that grows exponentially with their score. The symbol " $\propto$ " in Algorithm 2 means "proportional to".

Algorithm 1: Exponential Mechanism $A_{EM}(\mathbf{x}, q(\cdot;\cdot), \Delta, \varepsilon)$

Input: Assume that $q(y;\cdot)$ is Δ-sensitive for every $y \in \mathcal{Y}$.

1 Select Y from the distribution with $\Pr(Y = y) \propto \exp\left(\frac{\varepsilon}{2\Delta} q(y; \mathbf{x})\right)$;

2 **return** Y;

The exponential mechanism

- When is this algorithm even well defined? When $\boldsymbol{Y}$ is finite the algorithm is well-defined since we can set

$$P(Y = y) = \frac{e^{\frac{\varepsilon}{2\Delta} q(y;\mathbf{x})}}{\sum_{y' \in \mathcal{Y}} e^{\frac{\varepsilon}{2\Delta} q(y';\mathbf{x})}}.$$

- In fact, the mechanism makes sense over infinite domains, and even continuous ones. For inifinite discrete domains like the integers $\mathbb{Z}$, it must be that $\sum_{y \in \mathcal{Y}} e^{\frac{\varepsilon}{2\Delta} q(y;\mathbf{x})}$ is finite for every $\mathbf{x}$. Over continuous spaces like the real line, it must be that $\int_{y \in \mathcal{Y}} \exp\left(\frac{\varepsilon}{2\Delta} q(y;\mathbf{x})\right) dy$ is finite for every possible data set $\mathbf{x}$.
- Now that we have a well-defined algorithm, we'll try to understand why it is differentially private, and why it is useful.

The exponential mechanism

Proof:

- Assume for simplicity that $\boldsymbol{y}$ is finite. For any output y and data set $\mathbf{x}$ we have $P(y \mid \mathbf{x}) = \frac{e^{\frac{\varepsilon}{2\Delta} q(y;x)}}{\sum_{e^{\frac{\varepsilon}{2\Delta} q(y';\mathbf{x})}}}$. Let $\mathbf{x}'$ be a data set adjacent to $\mathbf{x}$. Since the sensitivity of $q(y;\cdot)$ is at most Δ, we have

$$\frac{e^{\frac{\varepsilon}{2\Delta} q(y;\mathbf{x})}}{e^{\frac{\varepsilon}{2\Delta} q(y;\mathbf{x}')}} = \exp\left(\frac{\varepsilon}{2\Delta}\left(q(y;\mathbf{x}) - q\left(y;\mathbf{x}'\right)\right)\right) \leq \exp\left(\frac{\varepsilon}{2\Delta} \cdot \Delta\right) = e^{\varepsilon/2}$$

- Similarly, for the normalizing constants,

$$\frac{\sum_{y' \in \mathcal{Y}} e^{\frac{\varepsilon}{2\Delta} q\left(y';\mathbf{x}'\right)}}{\sum_{y' \in \mathcal{Y}} e^{\frac{\varepsilon}{2\Delta} q(y';\mathbf{x})}} \leq \sup_{y'}\left(\exp\left(\frac{\varepsilon}{2\Delta}\left(q\left(y';\mathbf{x}'\right) - q\left(y';\mathbf{x}\right)\right)\right)\right) \leq e^{\varepsilon/2}.$$

Thus the ratio $\frac{Pr(y|\mathbf{x})}{P(y|\mathbf{x}')}$ is at most $e^{\varepsilon/2} \cdot e^{\varepsilon/2} = e^{\varepsilon}$. The case of an infinite domain is similar, with integrals over to the base measure replacing sums.

Thank You!

Any Questions?