Bayesian Pseudo Posterior Synthesis for Data Privacy Protection

Jingchen (Monika) Hu

Vassar College & ASA/NSF/BLS Fellow

Joint work with Terrance D. Savitsky (BLS)

March 12, 2019

Outline

The CE data at the BLS

- Conducted by the U.S. Census Bureau for the BLS.
- Contains data on expenditures, income, and tax statistics about consumer units (CU) across the country.
- Provides information on the buying habits of U.S. consumers.

The CE data at the BLS

- Conducted by the U.S. Census Bureau for the BLS.
- Contains data on expenditures, income, and tax statistics about consumer units (CU) across the country.
- Provides information on the buying habits of U.S. consumers.
- The public-use microdata (PUMD) files provide information for individual respondents, without any information that could identify respondents.
- Data users would like to access to more detailed versions of the CU's family income, however BLS is not releasing the original values of family income due to confidentiality concerns (Title 13).

The CE data at the BLS

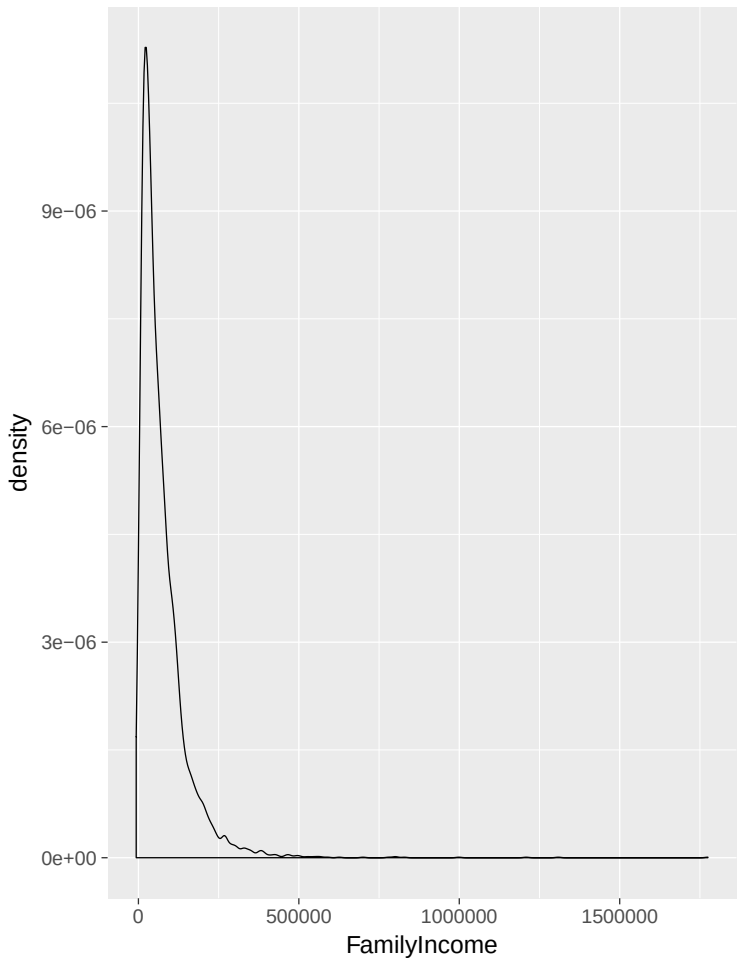
- Conducted by the U.S. Census Bureau for the BLS.
- Contains data on expenditures, income, and tax statistics about consumer units (CU) across the country.
- Provides information on the buying habits of U.S. consumers.
- The public-use microdata (PUMD) files provide information for individual respondents, without any information that could identify respondents.
- Data users would like to access to more detailed versions of the CU's family income, however BLS is not releasing the original values of family income due to confidentiality concerns (Title 13).
- How to make it happen?

The Consumer Expenditure Surveys data

Variable	Description
Gender	Gender of the reference person; 2 categories
Age	Age of the reference person; 5 categories
Education Level	Education level of the reference person; 8 categories
Region	Region of the CU; 4 categories
Urban	Urban status of the CU; 2 categories
Marital Status	Marital status of the reference person; 5 categories
Urban Type	Urban area type of the CU; 3 categories
CBSA	2010 core-based statistical area (CBSA) status; 3 categories
Family Size	Size of the CU; 11 categories
Earner	Earner status of the reference person; 2 categories
Family Income	Imputed and reported income before tax of the CU; approximate range: (-7K, 1,800K)

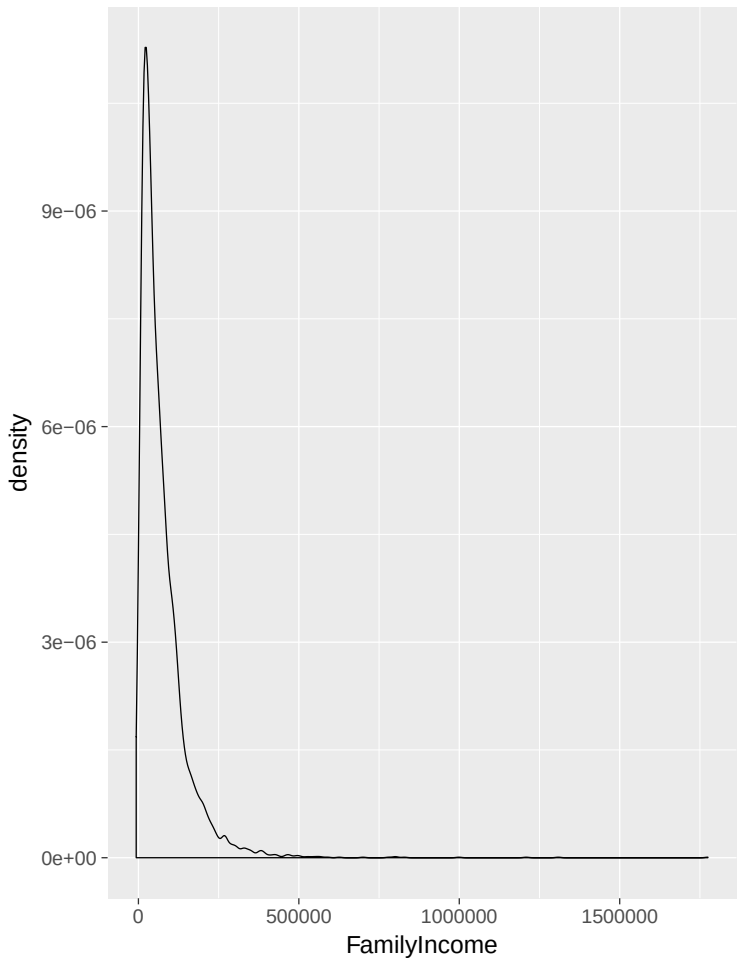
Table: Variables used in the CE sample. Data taken from the 2017 Q1 Consumer Expenditure Survey.

The CE data: highly skewed family income



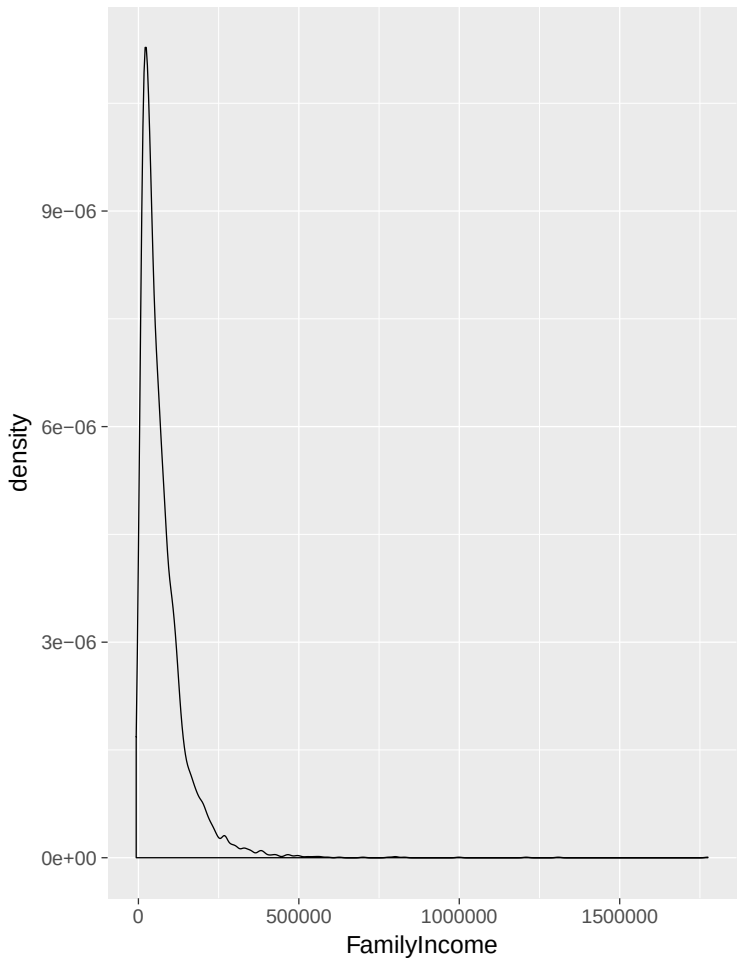
- Family income: imputed and reported income before tax of the CU.

The CE data: highly skewed family income



- Family income: imputed and reported income before tax of the CU.
- Record with high risks: assumed to be in the tail, i.e. CUs with extremely high family income.
- What to do?

The CE data: highly skewed family income



- Family income: imputed and reported income before tax of the CU.
- Record with high risks: assumed to be in the tail, i.e. CUs with extremely high family income.
- What to do? Topcoding (statistical disclosure control).

Protection for highly skewed family income: topcoding

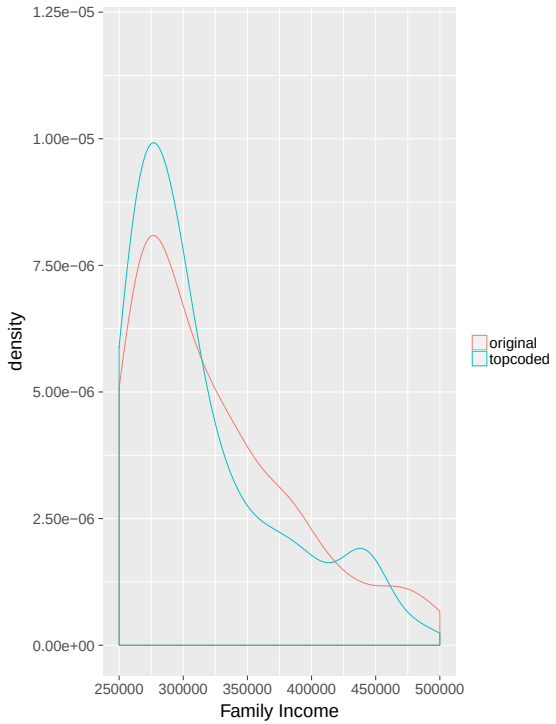


Figure: Range (250K, 500K)

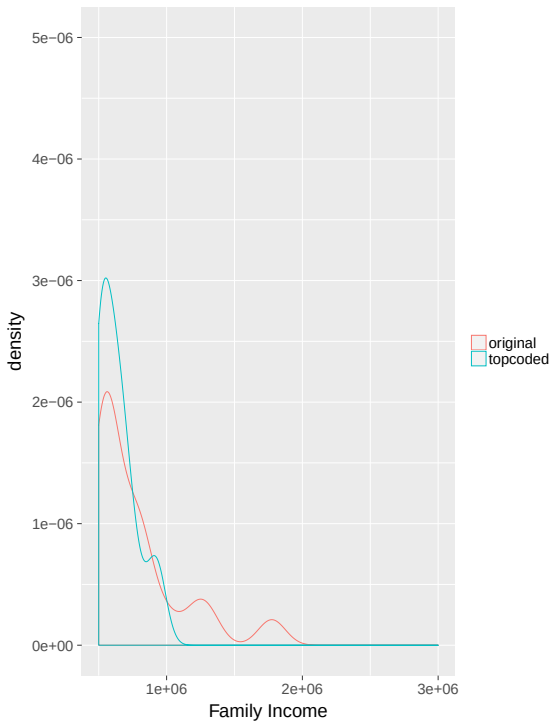


Figure: Range (500K, 3,000K)

- Topcoding affects observations in the tail.
- Reduction in data utility (An and Little, 2007).
- Reduction in disclosure risks?

Protection for highly skewed data: synthetic data

Rubin (1993) and Little (1993) proposed the synthetic data.

- Simulate records from statistical models that are estimated from the original confidential data.
- Balance of data utility and disclosure risks
  - preserve relationships of variables
  - low disclosure risks
- Allow data analysts to make valid inference for a wide class of analyses.

Protection for highly skewed data: synthetic data

- Census Bureau's synthetic data products:
  - Survey of Income and Program Participation (SIPP)
  - Longitudinal Business Databases (LBD)
  - OnTheMap

Protection for highly skewed data: synthetic data

- Census Bureau's synthetic data products:
  - Survey of Income and Program Participation (SIPP)
  - Longitudinal Business Databases (LBD)
  - OnTheMap
- Ongoing research at the Census
  - Economic Census microdata synthesis
  - Joint collaboration between the Census Bureau and Dr. Hang Kim (University of Cincinnati and ASA/NSF/Census Fellow)
  - Contact: Jenny Thompson (katherine.j.thompson@census.gov)

Protection for highly skewed data: proposed approach

- Typically, statistical agencies will
  - Develop a synthesizer, and simulate synthetic data.
  - Evaluate the data utility and disclosure risks of the synthetic data.
  - Make decision on the synthetic data release based on utility and risks profiles.

Protection for highly skewed data: proposed approach

- Typically, statistical agencies will
  - Develop a synthesizer, and simulate synthetic data.
  - Evaluate the data utility and disclosure risks of the synthetic data.
  - Make decision on the synthetic data release based on utility and risks profiles.
- What if the disclosure risks are deemed too high?
  - Option 1: develop a new synthesizer, or many new synthesizers if necessary.

Protection for highly skewed data: proposed approach

- Typically, statistical agencies will
  - Develop a synthesizer, and simulate synthetic data.
  - Evaluate the data utility and disclosure risks of the synthetic data.
  - Make decision on the synthetic data release based on utility and risks profiles.
- What if the disclosure risks are deemed too high?
  - Option 1: develop a new synthesizer, or many new synthesizers if necessary.
  - Option 2: use the evaluated disclosure risks to create a risk-adjusted synthesizer, which induces further disclosure protection on high risk records.

Protection for highly skewed data: proposed approach

- Typically, statistical agencies will
  - Develop a synthesizer, and simulate synthetic data.
  - Evaluate the data utility and disclosure risks of the synthetic data.
  - Make decision on the synthetic data release based on utility and risks profiles.
- What if the disclosure risks are deemed too high?
  - Option 1: develop a new synthesizer, or many new synthesizers if necessary.
  - Option 2: use the evaluated disclosure risks to create a risk-adjusted synthesizer, which induces further disclosure protection on high risk records.
- We will illustrate proposed Option 2 for the CE family income data synthesis.

Protection for highly skewed data: proposed approach

Our CE family income data synthesis methods involve:

- Develop a nonparametric mixture synthesizer, and simulate synthetic family income.
- Evaluate the identification disclosure risks of the synthetic family income.
- Create a new risk-adjusted synthesizer.

Protection for highly skewed data: proposed approach

Our CE family income data synthesis methods involve:

- Develop a nonparametric mixture synthesizer, and simulate synthetic family income.
- Evaluate the identification disclosure risks of the synthetic family income.
- Create a new risk-adjusted synthesizer.
- Synthesize family income from the new risk-adjusted synthesizer, and evaluate the identification disclosure risks.

Protection for highly skewed data: proposed approach

Our CE family income data synthesis methods involve:

- Develop a nonparametric mixture synthesizer, and simulate synthetic family income.
- Evaluate the identification disclosure risks of the synthetic family income.
- Create a new risk-adjusted synthesizer.
- Synthesize family income from the new risk-adjusted synthesizer, and evaluate the identification disclosure risks.
- Evaluate and compare risk profiles of:
  - Synthetic data from original synthesizer.
  - Synthetic data from risk-adjusted synthesizer.
  - Topcoded microdata.

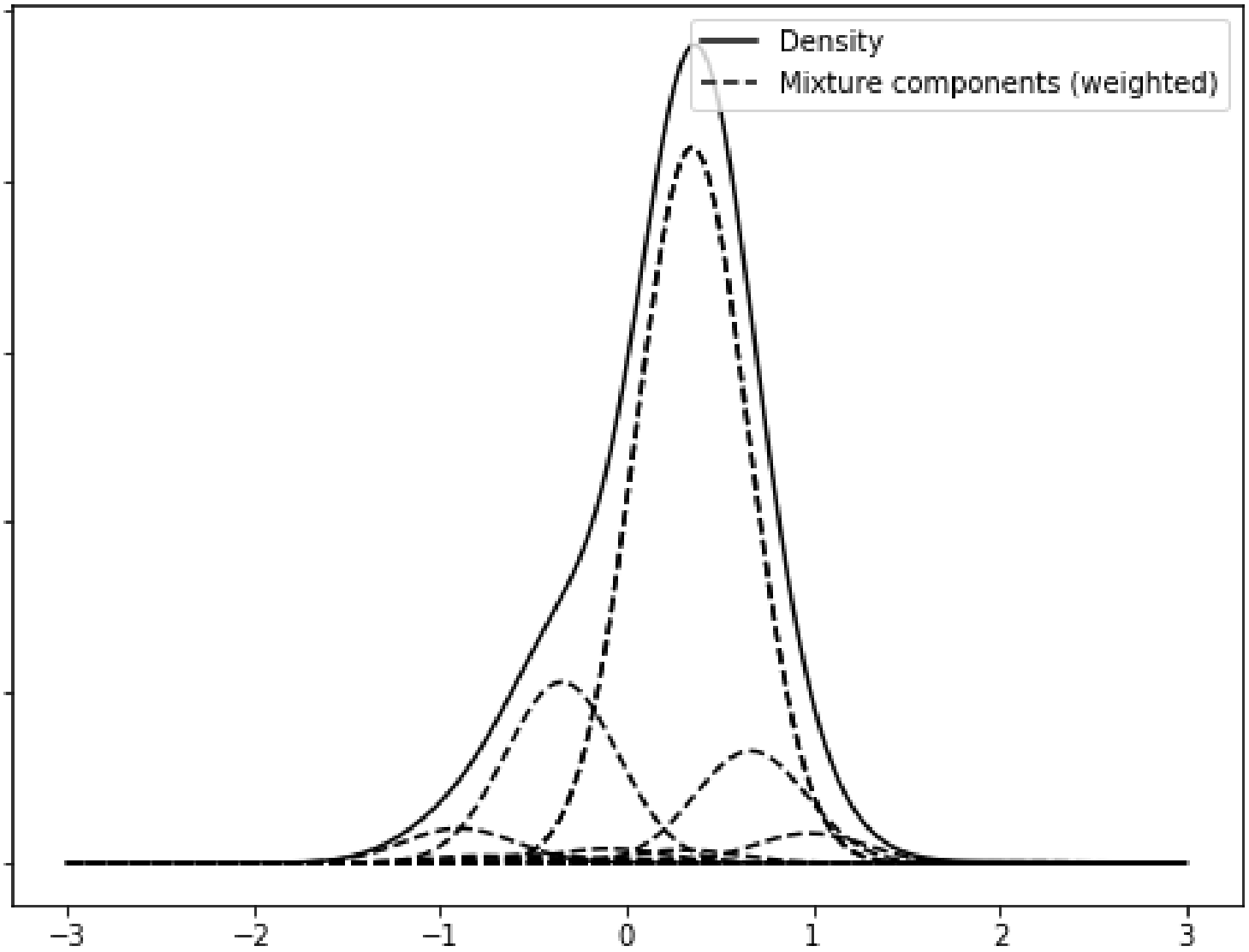
A nonparametric mixture synthesizer

$$y_i \mid \mathbf{X}_i, z_i, \beta, \sigma \sim \text{Normal}(y_i \mid \mathbf{x}_i^{'}\boldsymbol{\beta}_{z_i}^*, \sigma_{z_i}^*) \quad (1)$$

$$z_i \mid \pi \sim \text{Multinomial}(1; \pi_1, \cdots, \pi_K) \quad (2)$$

- y_i is the family income of CU i.
- $\mathbf{x}_i^{'}$ includes 10 predictors of CU i.
- $y_i \mid \mathbf{X}_i, z_i, \beta_i, \sigma_i \sim \text{Normal}(y_i \mid \mathbf{x}_i^{'}\boldsymbol{\beta}_i, \sigma_i)$, where $\boldsymbol{\beta}_i = \boldsymbol{\beta}_{z_i}^*$ and $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}_{z_i}^*$ given z_i.
- The nonparametric mixture synthesizer could model the long tail better with several mixture components.

A nonparametric mixture synthesizer



A nonparametric mixture synthesizer

$$y_i \mid \mathbf{X}_i, z_i, \beta, \sigma \sim \text{Normal}(y_i \mid \mathbf{x}_i^{'}\boldsymbol{\beta}^*_{z_i}, \sigma^*_{z_i}) \quad (3)$$

$$z_i \mid \pi \sim \text{Multinomial}(1; \pi_1, \cdots, \pi_K) \quad (4)$$

- To generate synthetic family income, y_i^*, for CU i:
  - Generate z_i from Equation (4).
  - Generate y_i^* given $\mathbf{x}_i^{'}$ and estimated $\boldsymbol{\beta}^*_{z_i}, \sigma^*_{z_i})$ from Equation (3).
  - Do this for every CU, and obtain one partially synthetic dataset $\mathbf{Z}^{(l)}$.
  - Repeat the above steps for m times, and obtain m independent partially synthetic datasets $\mathbf{Z} = (\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(m)})$.

Record-level identification disclosure

- Focus on record-level identification disclosure risks.
- As opposed to file-level identification disclosure risk summary.
- Want to surgically target high risk records.

Assumptions about intruder's knowledge

- Available information known by the intruder about CU i:
  - A known pattern of the un-synthesized categorical variables, $\mathbf{X}_i^p \subseteq \mathbf{X}_i$, e.g. (Gender, Age, Education Level, Marital, Earner).
  - The true value of synthesized family income y_i.
  - A name or identity of interest.
- Successful identification allows the intruder to learn other information of CU i in the released microdata.

Identification risks based on notion of isolation

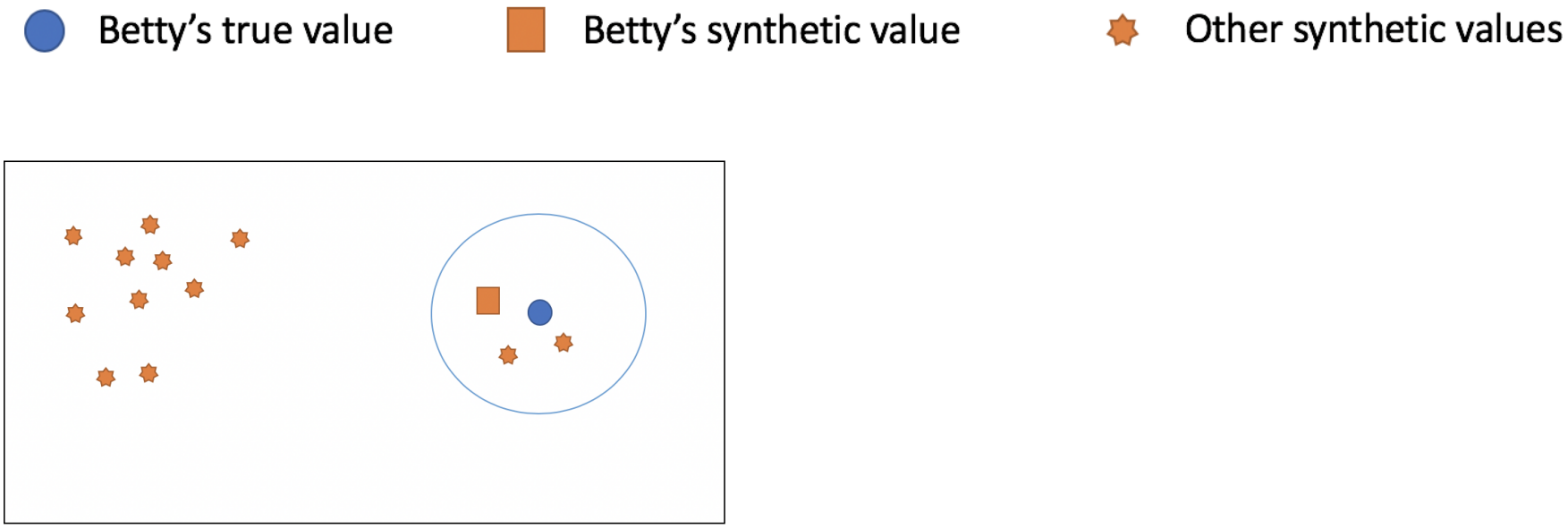
- Define radius r of synthetic data y^* in pattern p around the truth y.
- Use percentage radius, e.g. $r = 20\%$.
  - e.g. For a CU i with \$50,000 family income, the interval/ball is: [\$40,000, \$60,000].
- Outside of radius $\rightarrow$ isolation.
- Do this for each record y_i: all y_j^*'s in pattern p.
- Identification risk (IR) for each record is a probability.

Evaluation of identification disclosure risks

- Fewer synthetic values inside the interval/ball $\rightarrow$ the intruder has a higher probability of guessing the record of the name they seek.

Evaluation of identification disclosure risks

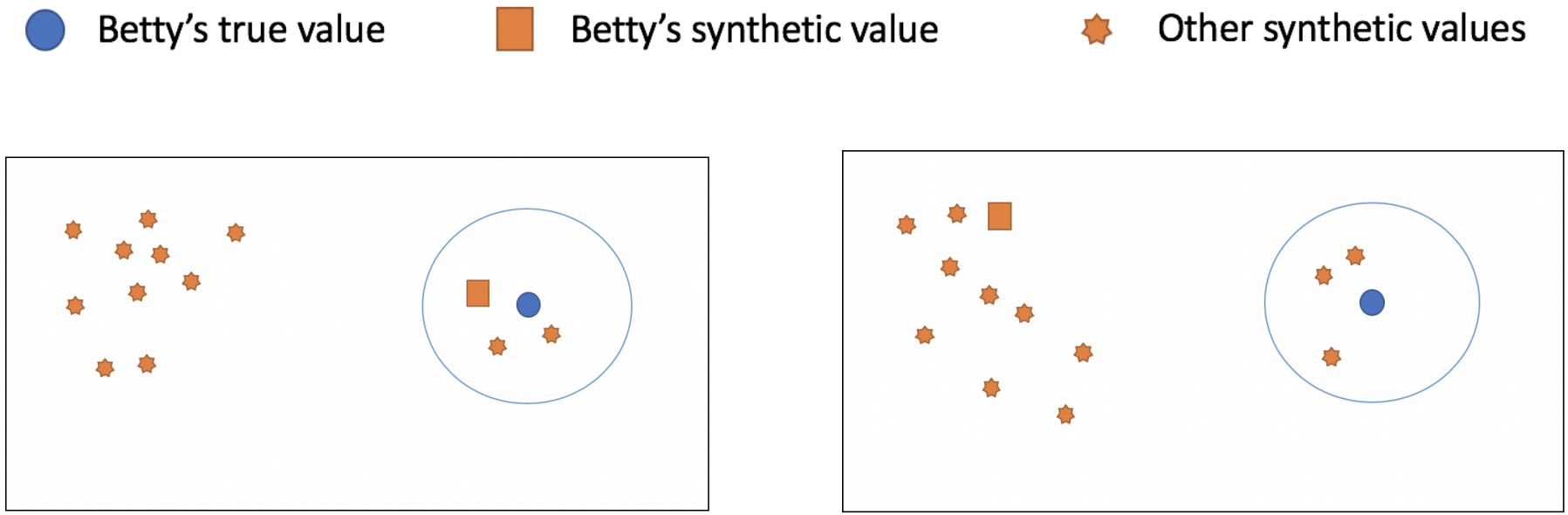
- Fewer synthetic values inside the interval/ball $\rightarrow$ the intruder has a higher probability of guessing the record of the name they seek.



Scenario 1:
$IR_i = \frac{10}{13} \times 1 = \frac{10}{13}$.

Evaluation of identification disclosure risks

- Fewer synthetic values inside the interval/ball $\rightarrow$ the intruder has a higher probability of guessing the record of the name they seek.

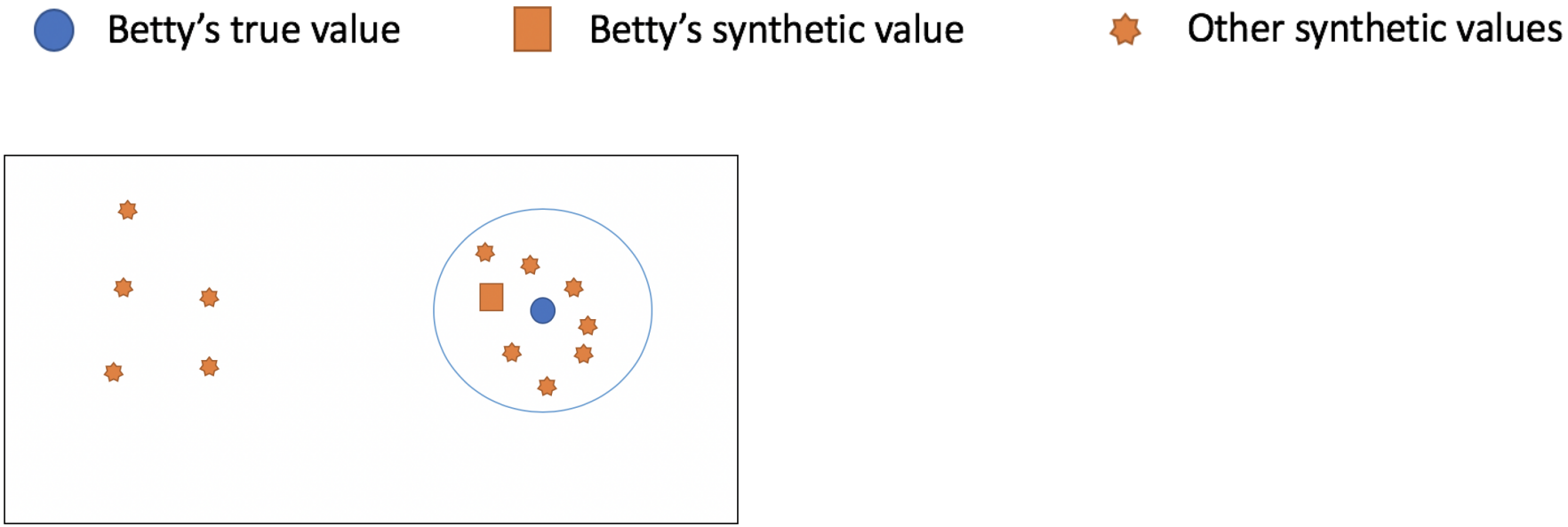


Scenario 1:
$IR_i = \frac{10}{13} \times 1 = \frac{10}{13}$.

Scenario 2:
$IR_i = \frac{10}{13} \times 0 = 0$.

Evaluation of identification disclosure risks

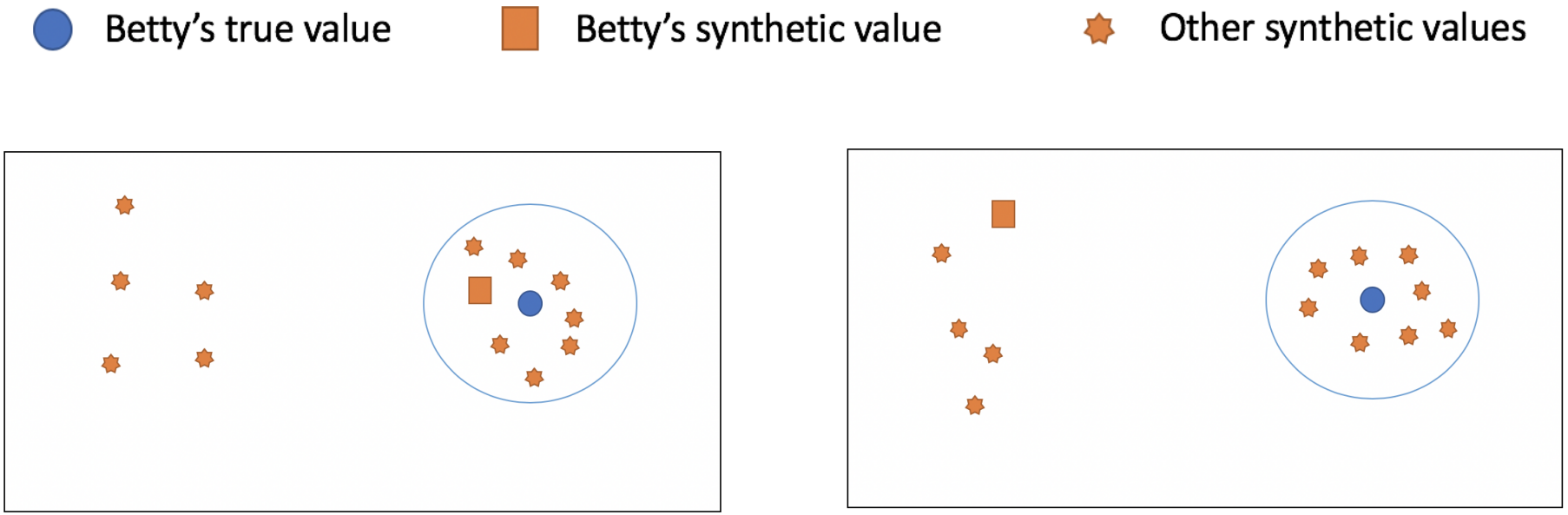
- More synthetic values inside the interval/ball $\rightarrow$ the intruder has a lower probability of guessing the record of the name they seek.



Scenario 3:
$IR_i = \frac{5}{13} \times 1 = \frac{5}{13}$.

Evaluation of identification disclosure risks

- More synthetic values inside the interval/ball $\rightarrow$ the intruder has a lower probability of guessing the record of the name they seek.



Scenario 3:
$IR_i = \frac{5}{13} \times 1 = \frac{5}{13}$.

Scenario 4:
$IR_i = \frac{5}{13} \times 0 = 0$.

Evaluation of identification disclosure risks

More formally, for CU i with pattern p:

$$\begin{aligned} IR_i &:= \text{Pr}\left(\text{identification disclosure of } i\right) \\ &= \frac{\sum_{j \in M_i} \mathbb{I}\left(y_j^* \notin B(y_i, r)\right)}{|M_i|} \times T_i. \end{aligned} \tag{5}$$

- $B(y_i, r)$: a ball of radius r around true value, y_i.
- $T_i = 1$ if the true value, y_i is among those records, $j \in M_i$ whose $y_j^* \in B(y_i, r)$; $T_i = 0$ otherwise.

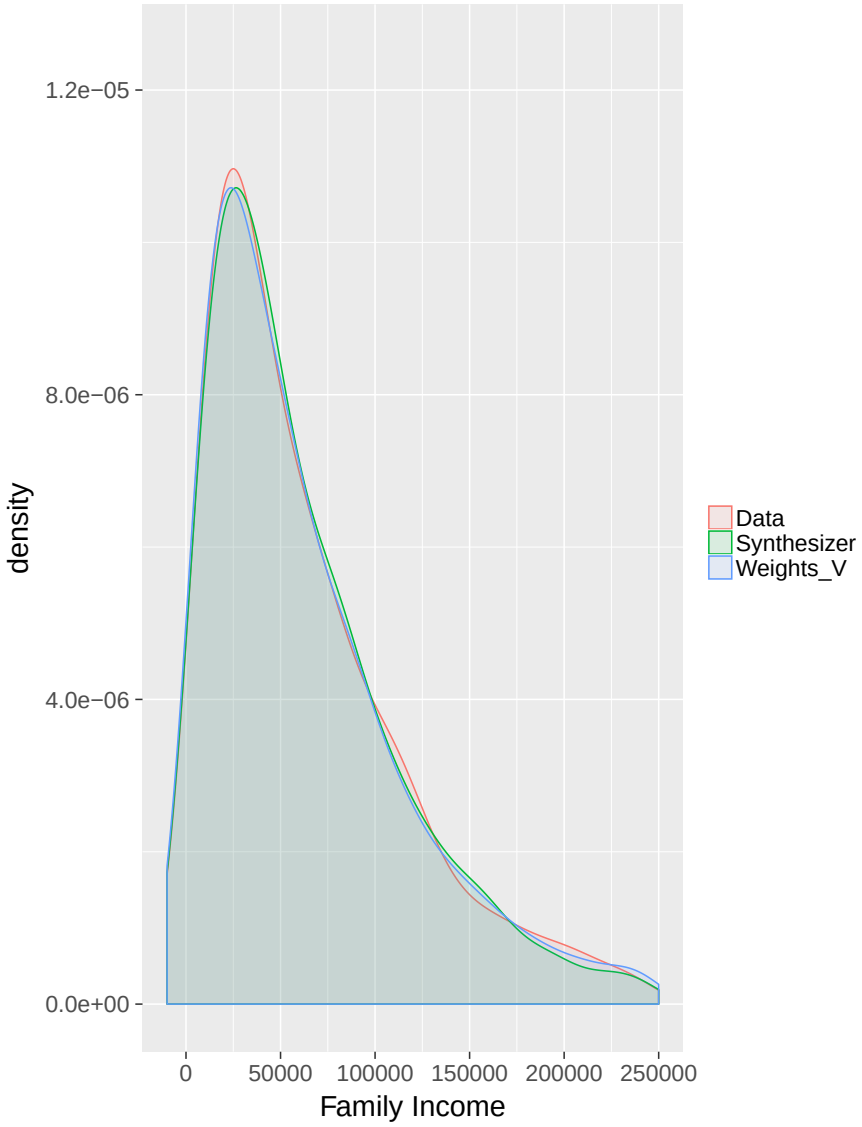
A new risk-adjusted synthesizer

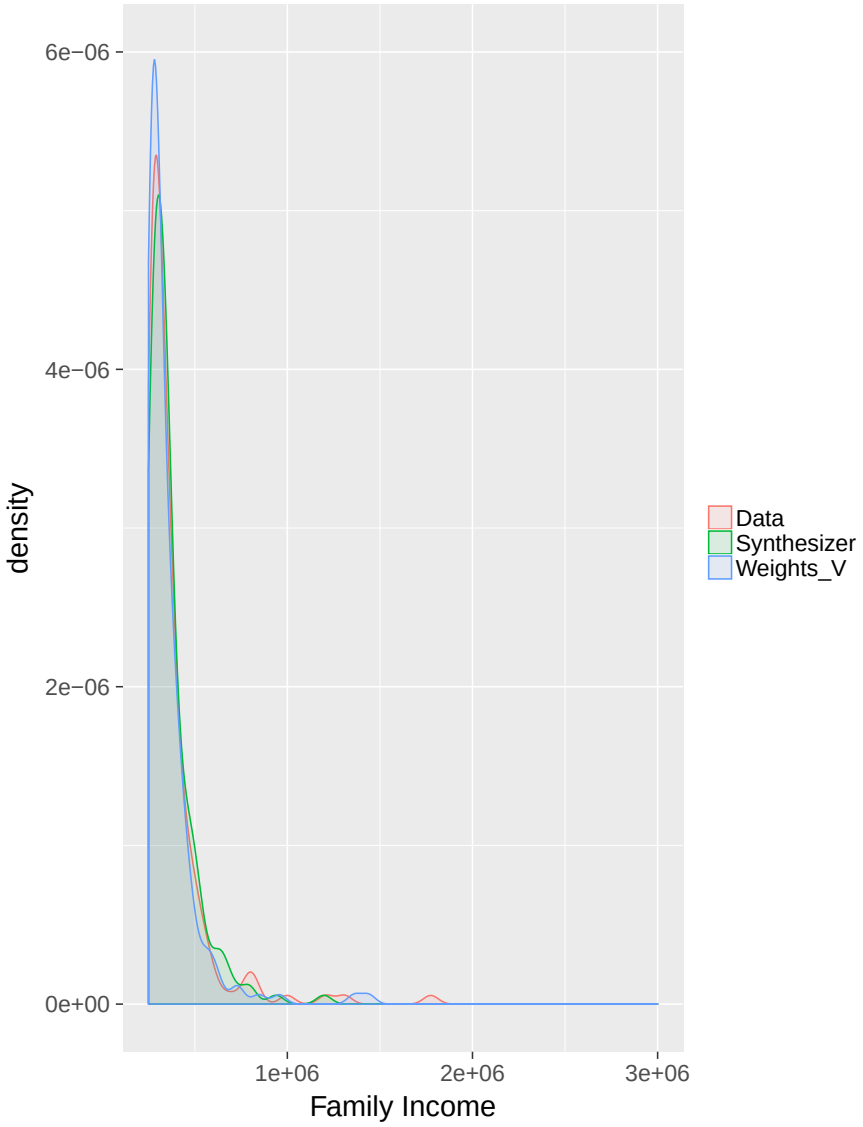
- Use weight $\alpha_i \in (0, 1)$ for CU i.
- $\alpha_i \propto \frac{1}{IR_i}$.
- Selectively downweight to defeat the likelihood principle:

$$\left[\prod_{i=1}^{n} p\left(y_i \mid (\pi_k, \boldsymbol{\beta}_k^*, \sigma_k^*)_{k=1}^{K}\right)^{\alpha_i}\right] \prod_{k=1}^{K} p\left(\pi_k, \boldsymbol{\beta}_k^*, \sigma_k^* \mid \theta\right). \tag{6}$$

- Surgical distortion: scalar α vs vector α_i.

Data utility results





Data utility results

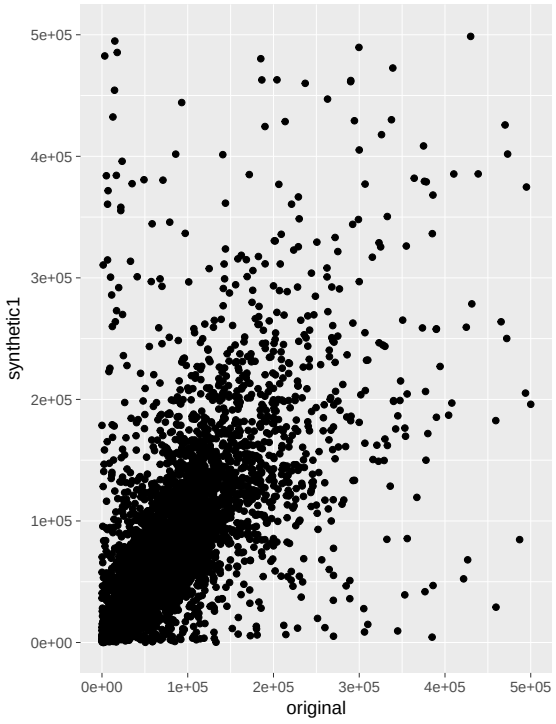


Figure: Synthesizer

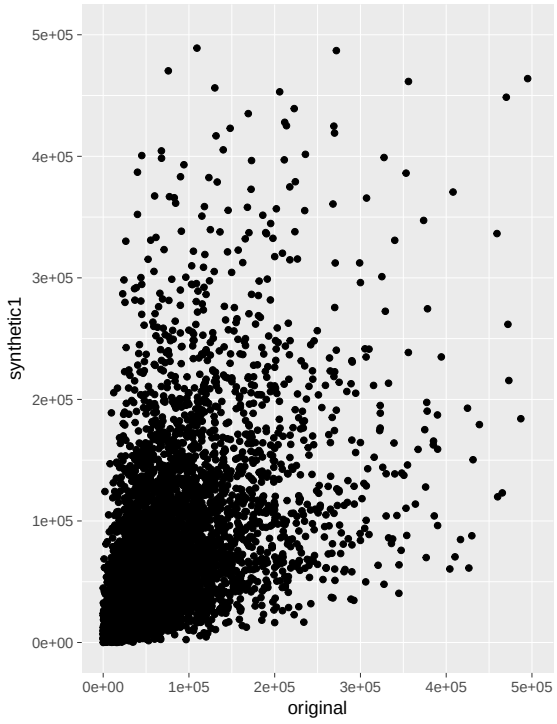


Figure: Vector Weights

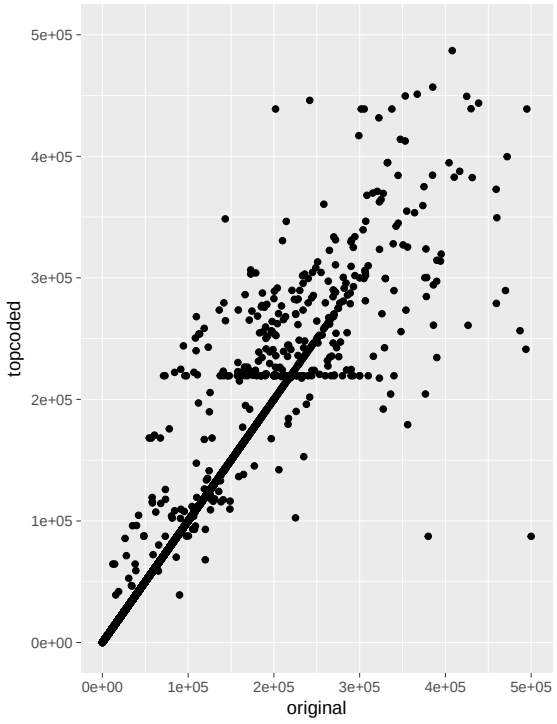


Figure: Topcoding

Violin plots of identification risks

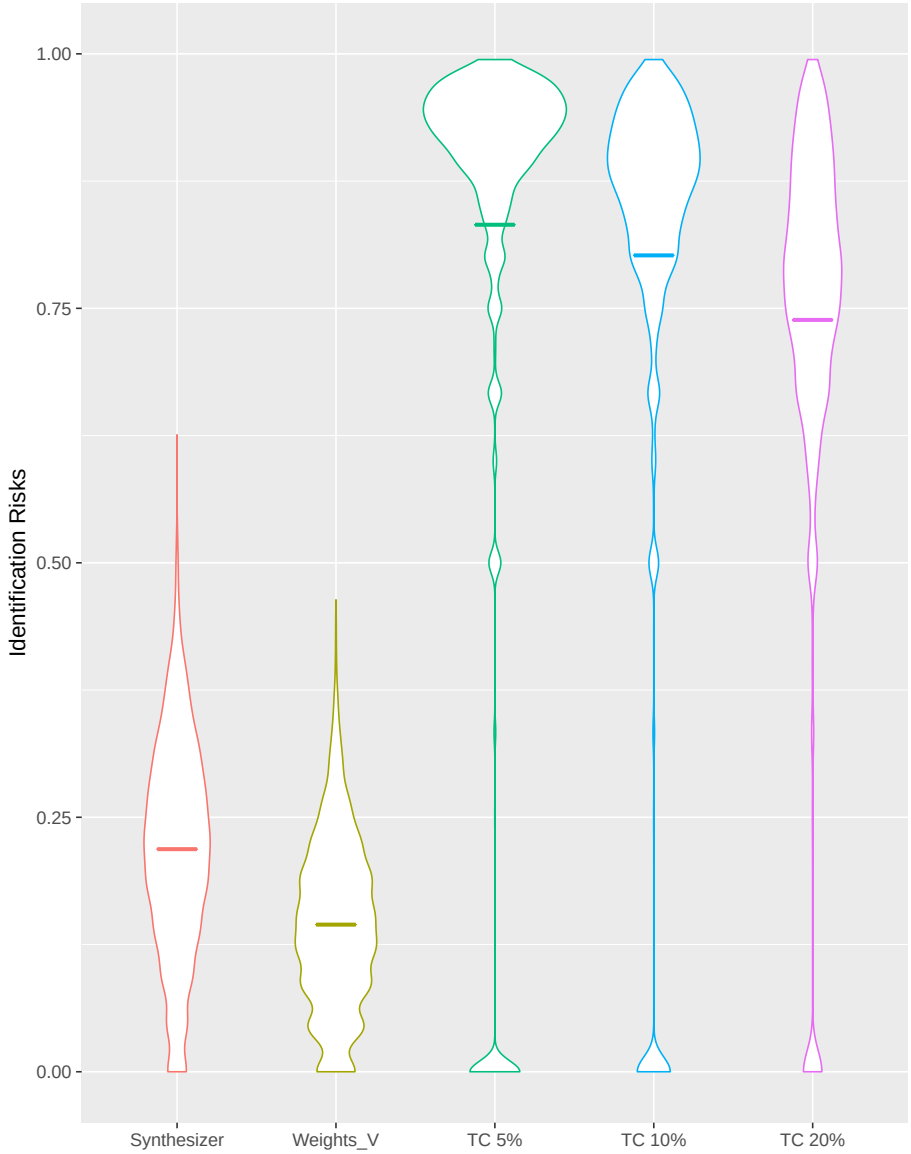


Table of identification risks for top 10 size/magnitude records

Data Value	Synthesizer	Weights_V	TC 20%
(250K, 3,000K)	0.0000	0.0486	0.0000
(250K, 3,000K)	0.0482	0.1446	0.0000
(250K, 3,000K)	0.0000	0.2473	0.0000
(250K, 3,000K)	0.0000	0.1440	0.0000
(250K, 3,000K)	0.0496	0.0987	0.0000
(250K, 3,000K)	0.0967	0.0000	0.9565
(250K, 3,000K)	0.0986	0.0989	0.0000
(250K, 3,000K)	0.0000	0.0986	0.0000
(250K, 3,000K)	0.0987	0.0000	0.0000
(250K, 3,000K)	0.0000	0.0000	0.0000

Table of identification risks for top 10 risky records

Data Value	Synthesizer	Weights_V	TC 20%
(-10K, 50K)	0.6258	0.0925	0.9762
(50K, 100K)	0.5318	0.1717	0.8312
(100K, 250K)	0.5725	0.1635	0.8108
(100K, 250K)	0.5274	0.2261	0.8889
(100K, 250K)	0.5369	0.1560	0.7738
(100K, 250K)	0.5258	0.1348	0.8788
(-10K, 50K)	0.5270	0.3250	0.9400
(100K, 250K)	0.5310	0.1708	0.8214
(100K, 250K)	0.5925	0.0460	0.9080
(50K, 100K)	0.5351	0.2629	0.7113

Risky records are unique

Variable	Record A	Record B
Gender	Male	Male
Age	20 - 40	60 - 80
Education Level	Bachelor's degree	Some college, no degree
Region	South	South
Urban	Urban	Urban
Marital	Married	Married
Uatype	Urbanized area	Urbanized area
Cbsastat	In a CBSA not in the Principal City	In a CBSA in the Principal City
Family Size	1	1
Earner	Member Earns Income	Member Earns Income
Family Income	(-10K, 50K)	(-10K, 50K)
IR: Synthesizer	0.6258	0.5270
IR: Weights_V	0.0925	0.3250
IR: TC 20%	0.9762	0.9400

Implications

- Our proposal is for agencies to release respondent-level synthetic data.
  - Users never see real data.
- We may directly interrogate the disclosure risks of the synthetic records.
  - As contrasted with dynamic queries of the real data.
- We should directly measure the disclosure risks of the synthetic data.

Implications

- Our proposal is for agencies to release respondent-level synthetic data.
  - Users never see real data.
- We may directly interrogate the disclosure risks of the synthetic records.
  - As contrasted with dynamic queries of the real data.
- We should directly measure the disclosure risks of the synthetic data.
- We propose respondent-level weights to surgically downweight the likelihood contribution of high risk records, to achieve a better utility-risk tradeoff.

Implications

- Our proposal is for agencies to release respondent-level synthetic data.
  - Users never see real data.
- We may directly interrogate the disclosure risks of the synthetic records.
  - As contrasted with dynamic queries of the real data.
- We should directly measure the disclosure risks of the synthetic data.
- We propose respondent-level weights to surgically downweight the likelihood contribution of high risk records, to achieve a better utility-risk tradeoff.
- The use of topcoding incorrectly assumes which records express high risks.

References

- An, D. and Little, R. J. A. (2007), Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A*, 170, 923-940.
- Hu, J., Savitsky, T. D., and Williams, M. R. (2018+), Bayesian pseudo posterior synthesis for data privacy protection, *arXiv pre-prints*.
- Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* 9, 407-426.
- Rubin, D. B. (1993). Discussion statistical disclosure limitation. *Journal of Official Statistics* 9, 461-468.