Synthetic Data and Randomized Sanitizers

John M. Abowd

Cornell University

May 2, 2008

NISS /NCHS Workshop

Data Confidentiality: The Next Five Years

Acknowledgements and Disclaimer

- This work was partially supported by the National Science Foundation (grants SES-0339191, SES-0427889 and CNS-0627680) and the Census Bureau .

- The opinions are those of the author and not the National Science Foundation or the Census Bureau.

Outline

- Definition of synthetic data
- Criteria for “good” synthetic data
- Relationship of synthetic data to privacy models
- Detailed example of how a synthesizer works
- Research goals for the next five years

Definition of Synthetic Data

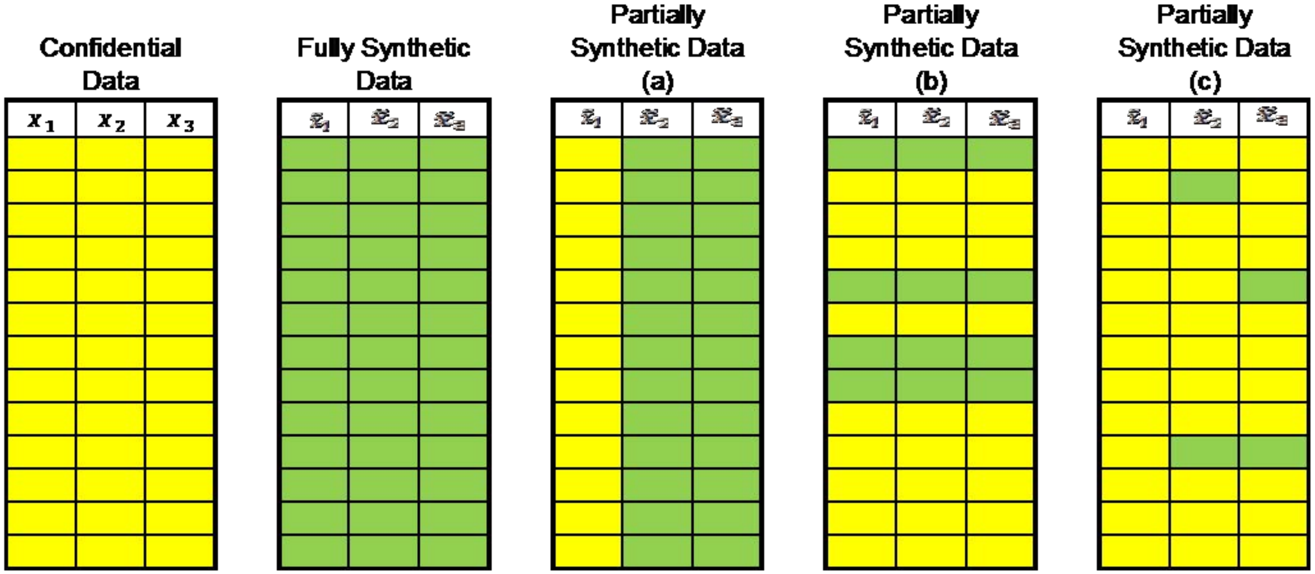
$$X \equiv \text{confidential data}$$

$$\Pr\left[\tilde{X} \middle| X\right] \equiv \text{PPD of } \tilde{X} \text{ given } X$$

$$\text{Release data are samples of } \tilde{X}$$

- Synthetic data are created by estimating the posterior predictive distribution (PPD) of the release data given the confidential data; then sampling release data from the PPD conditioning on the actual confidential values.
- The PPD is a parameter-free forecasting model for new values of the complete data matrix that conditions on all values of the underlying confidential data.

Types of Synthetic Data


Confidential Data
x_1 x_2 x_3
Fully Synthetic Data
$\tilde{x}_1$ $\tilde{x}_2$ $\tilde{x}_3$
Partially Synthetic Data (a)
$\tilde{x}_1$ $\tilde{x}_2$ $\tilde{x}_3$
Partially Synthetic Data (b)
$\tilde{x}_1$ $\tilde{x}_2$ $\tilde{x}_3$
Partially Synthetic Data (c)
$\tilde{x}_1$ $\tilde{x}_2$ $\tilde{x}_3$

- Fully synthetic data have all values of all variables replaced by draws from the PPD.
- Partially synthetic data have only some values replaced
  - (a) two variables synthesized
  - (b) four observations synthesized
  - (c) only “sensitive” values of some variables synthesized

Goals of Synthetic Data

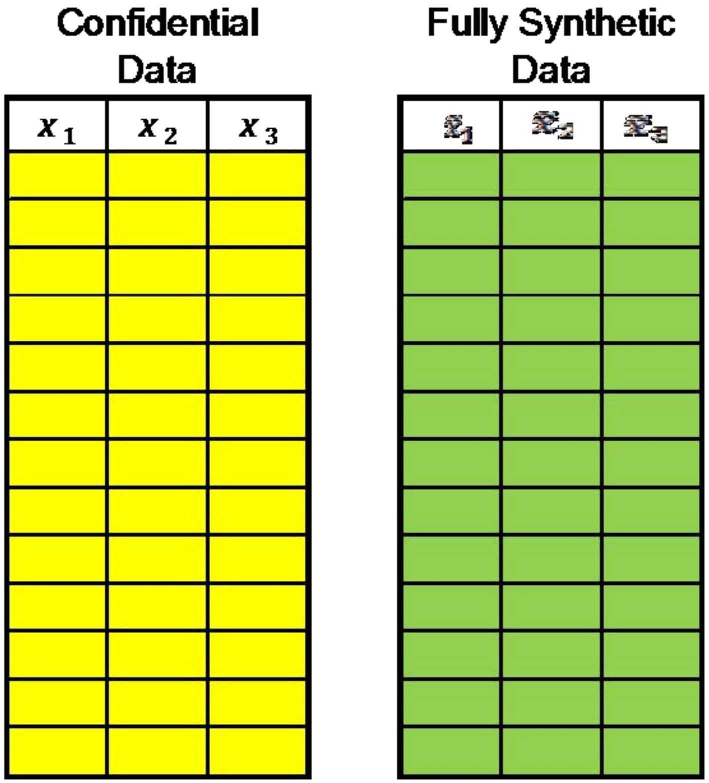
- Analytical validity
  - Statistical inferences based on the synthetic data should be “similar” to those based on the underlying confidential data.
- Confidentiality protection
  - The perturbation of the confidential data induced by replacing some, or all, of the values with draws from the PPD should be adequate to “protect” the confidential data.

Formal Models of Analytical Validity

- Unconditional analytical validity
  - The synthetic data process delivers the same inferences as the process that generated the confidential data. This property depends on both the synthesizer and the design of the confidential data
- Conditional analytical validity
  - The synthetic data process delivers the same inferences as the realized confidential data.
- The Rubin (1993) inference validity was based on using multiple samples (implicates) from the PPD. It is an unconditional analytical validity model.

Formal Models of Confidentiality Protection

- It was initially argued that the confidentiality protection embodied in fully synthetic data was complete since no elements of the confidential data are present in the release data.



Taking Account of Formal Privacy Models

- It is now known from a variety of papers in the computer science literature (Dwork, Nissim and their many collaborators, Gehrke and his collaborators, and others) that the confidentiality protection afforded by synthetic data depends upon properties of the transition probabilities that relate the confidential data to the release data.
- Not surprising since

$$\Pr[\tilde{X}|X] = I$$

implies that the PPD leaves the confidential data unchanged by the synthesizer.

Connection to Randomized Sanitizers

$$X \equiv \text{confidential data}$$

$$U \equiv \text{random noise}$$

$$\text{San}(X,U):(X,U) \rightarrow \tilde{X}$$

$$\Pr\left[\tilde{X} \middle| X\right] \equiv \text{probability of } \tilde{X} \text{ given } X$$

- A randomized sanitizer creates a conditional probability distribution for the release data given the conditional data.
- The randomness in a sanitizer is induced by the properties of the distribution of *U*.
- The PPD is just a particular randomized sanitizer.

Disclosure Limitation Definitions

$$X = x^{(1)} \text{ and } X = x^{(2)}$$

$$\tilde{X} = \tilde{x}$$

- Consider two confidential data matrices that differ in only a single row, $x^{(1)}$ and $x^{(2)}$.
- Use the PPD to evaluate the probability of a particular release data set given the two different confidential data sets.

Synthetic Data Can Leak Information about a Single Entity

$$\Pr\left[\tilde{X} = \tilde{x} \middle| X = x^{(1)}\right] \neq \Pr\left[\tilde{X} = \tilde{x} \middle| X = x^{(2)}\right]$$

- Changing a single row of the confidential data matrix changes the PPD.
- The PPD defines the transition probabilities from the confidential data to the release data.

Connection Between Synthetic Data and Differential Privacy

$$\frac{\dfrac{\Pr\left[X = x^{(1)} \middle| \tilde{X} = \tilde{x}\right]}{\Pr\left[X = x^{(2)} \middle| \tilde{X} = \tilde{x}\right]}}{\dfrac{\Pr[X = x^{(1)}]}{\Pr[X = x^{(2)}]}} = \frac{\Pr[\tilde{X} = \tilde{x} | X = x^{(1)}]}{\Pr[\tilde{X} = \tilde{x} | X = x^{(2)}]}$$

The posterior odds ratio for the gain in information about a single row of X is equal to the differential privacy from the randomized sanitizer that creates release data by sampling from the posterior predictive distribution.

A DETAILED EXAMPLE

The Multinomial-Dirichlet Model

- The data matrix X consists of categorical variables that can by summarized by a contingency table with *k* categories.
- n_i are counts.
- π_i are probabilities

$$\mathbf{n} = \left(n_1, \ldots, n_k\right), n = \sum n_i$$

$$\boldsymbol{\alpha} = \left(\alpha_1, \ldots, \alpha_k\right), \alpha_0 = \sum \alpha_i$$

$$\boldsymbol{\pi} = \left(\pi_1, \ldots, \pi_k\right)$$

$$\mathbf{n} \sim \mathrm{M}\left(n, \boldsymbol{\pi}\right)$$

$$\boldsymbol{\pi} \sim \mathrm{D}\left(\boldsymbol{\alpha}\right), \text{a priori}$$

$$\boldsymbol{\pi} \sim \mathrm{D}\left(\boldsymbol{\alpha} + \mathbf{n}\right), \text{a posteriori}$$

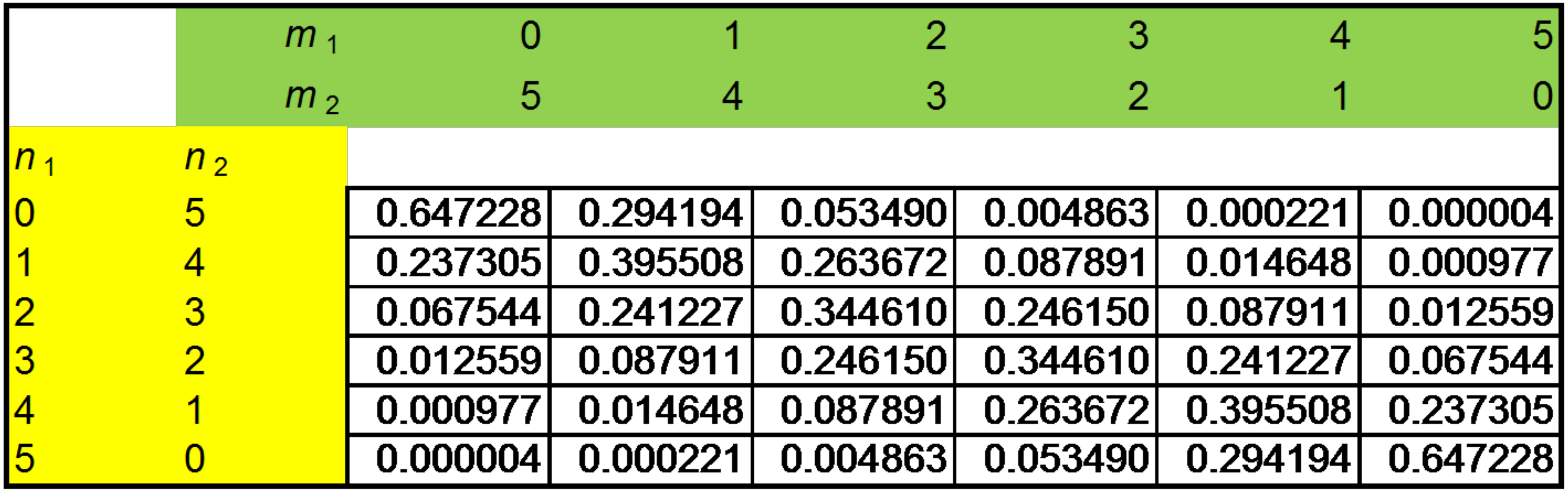
$$\mathbf{m} = \left(m_1, \ldots, m_k\right), m = \sum m_i$$

$$\mathbf{m} \sim \mathrm{M}\left(m, \boldsymbol{\pi}\right)$$

The Multinomial-Dirichlet Synthesizer

$$\Pr[\mathbf{m}|\mathbf{n}] = \mathrm{E}_{\boldsymbol{\pi}|\mathrm{n}}[\mathrm{M}(m, \boldsymbol{\pi})]$$

- The synthetic data are samples from the synthesizer that can be summarized by their counts, **m**.
- Since all the random variables are discrete, the synthesizer can be expressed as a simple transition probability matrix.



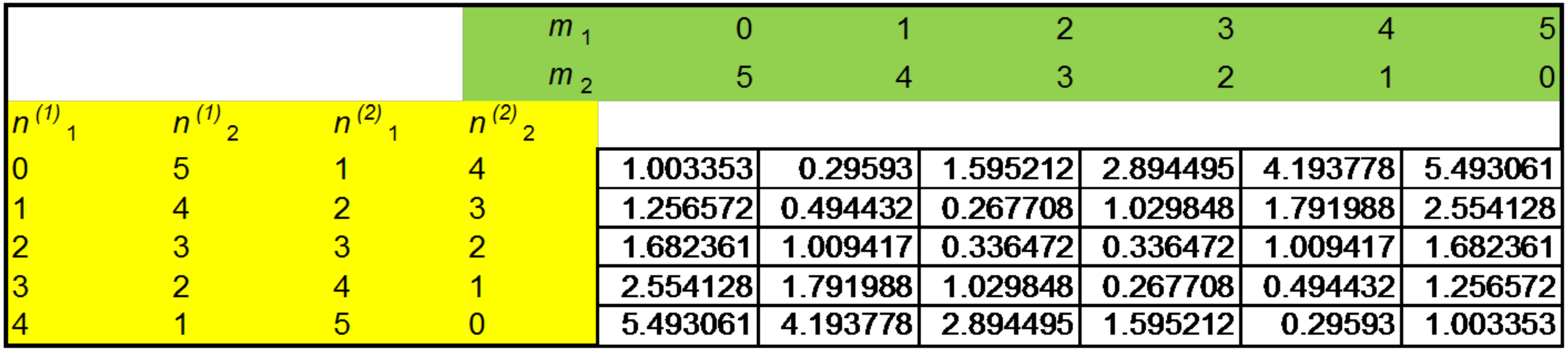
	m_1	0	1	2	3	4	5
	m_2	5	4	3	2	1	0
n_1	n_2						
0	5	0.647228	0.294194	0.053490	0.004863	0.000221	0.000004
1	4	0.237305	0.395508	0.263672	0.087891	0.014648	0.000977
2	3	0.067544	0.241227	0.344610	0.246150	0.087911	0.012559
3	2	0.012559	0.087911	0.246150	0.344610	0.241227	0.067544
4	1	0.000977	0.014648	0.087891	0.263672	0.395508	0.237305
5	0	0.000004	0.000221	0.004863	0.053490	0.294194	0.647228

- $k = 2$
- $\alpha_i = ½; \alpha_0 = 1$
- $n = m = 5$
- The table displays the transition probabilities that map **n** into **m**.

ε-Differential Privacy

$$\left|\ln \frac{\Pr\left[\mathbf{m}\middle|\mathbf{n}^{(\mathbf{1})}\right]}{\Pr\left[\mathbf{m}\middle|\mathbf{n}^{(\mathbf{2})}\right]}\right| < \varepsilon$$

- The two confidential data matrices, $\mathbf{n}^{(1)}$ and $\mathbf{n}^{(2)}$ differ by changing exactly one entity’s data.

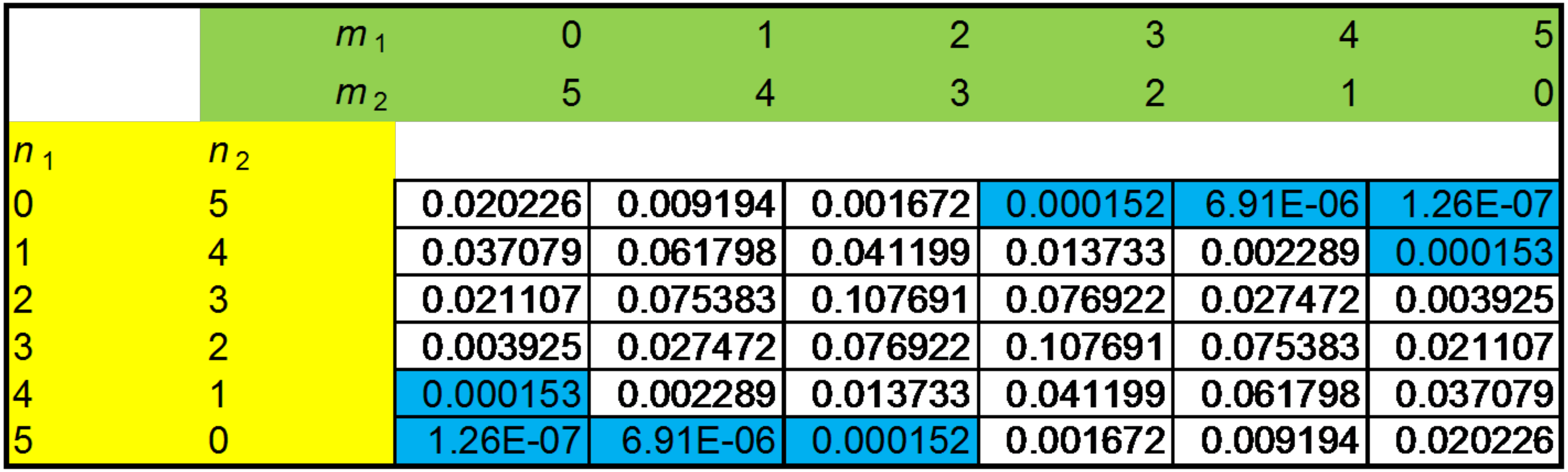


			m_1	0	1	2	3	4	5
			m_2	5	4	3	2	1	0
$n^{(1)}_1$	$n^{(1)}_2$	$n^{(2)}_1$	$n^{(2)}_2$						
0	5	1	4	1.003353	0.29593	1.595212	2.894495	4.193778	5.493061
1	4	2	3	1.256572	0.494432	0.267708	1.029848	1.791988	2.554128
2	3	3	2	1.682361	1.009417	0.336472	0.336472	1.009417	1.682361
3	2	4	1	2.554128	1.791988	1.029848	0.267708	0.494432	1.256572
4	1	5	0	5.493061	4.193778	2.894495	1.595212	0.29593	1.003353

- The table shows all of the differential privacy ratios for the example problem.
- The ε-differential privacy of this synthesizer is the maximum element in this table, 5.493061.
- The differential privacy limit is attained when the synthesizer delivers (0,5) and the underlying data are either (5,0) or (4,1).

Probabilistic Differential Privacy

- This definition of differential privacy allows the ε-differential privacy limit to fail with probability δ.
- To compute the PDP, the joint distribution of m and n must be examined for outcomes that occur with probability less than δ.

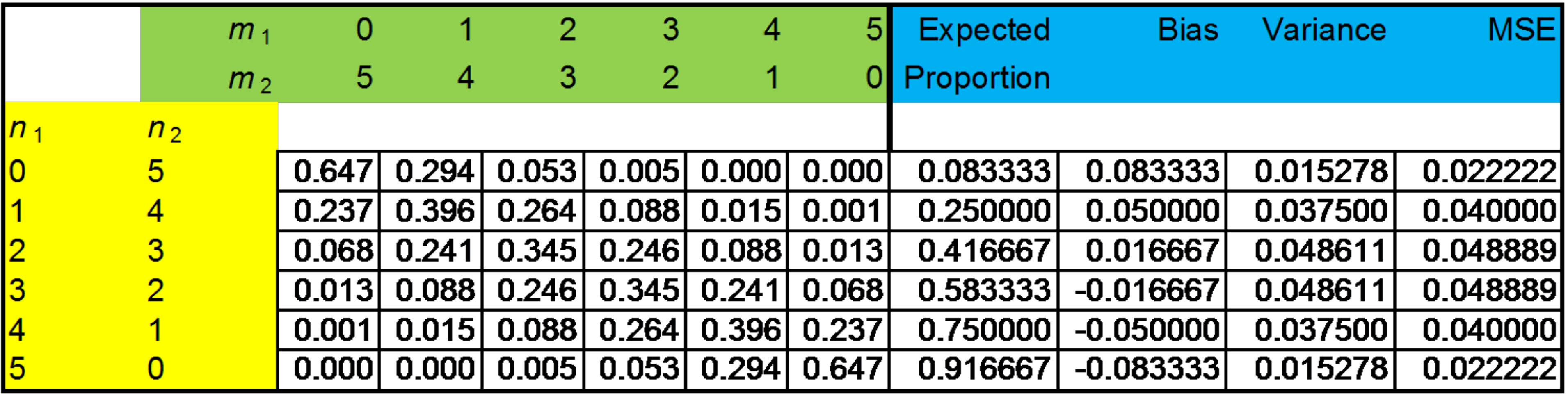


	m_1	0	1	2	3	4	5
	m_2	5	4	3	2	1	0
n_1	n_2						
0	5	0.020226	0.009194	0.001672	0.000152	6.91E-06	1.26E-07
1	4	0.037079	0.061798	0.041199	0.013733	0.002289	0.000153
2	3	0.021107	0.075383	0.107691	0.076922	0.027472	0.003925
3	2	0.003925	0.027472	0.076922	0.107691	0.075383	0.021107
4	1	0.000153	0.002289	0.013733	0.041199	0.061798	0.037079
5	0	1.26E-07	6.91E-06	0.000152	0.001672	0.009194	0.020226

- If we want to have ε-differential privacy of 2, then the synthesizer fails in the highlighted cells.
- The probabilistic differential privacy has δ = 0.000623, which is just the sum of the blue cells.

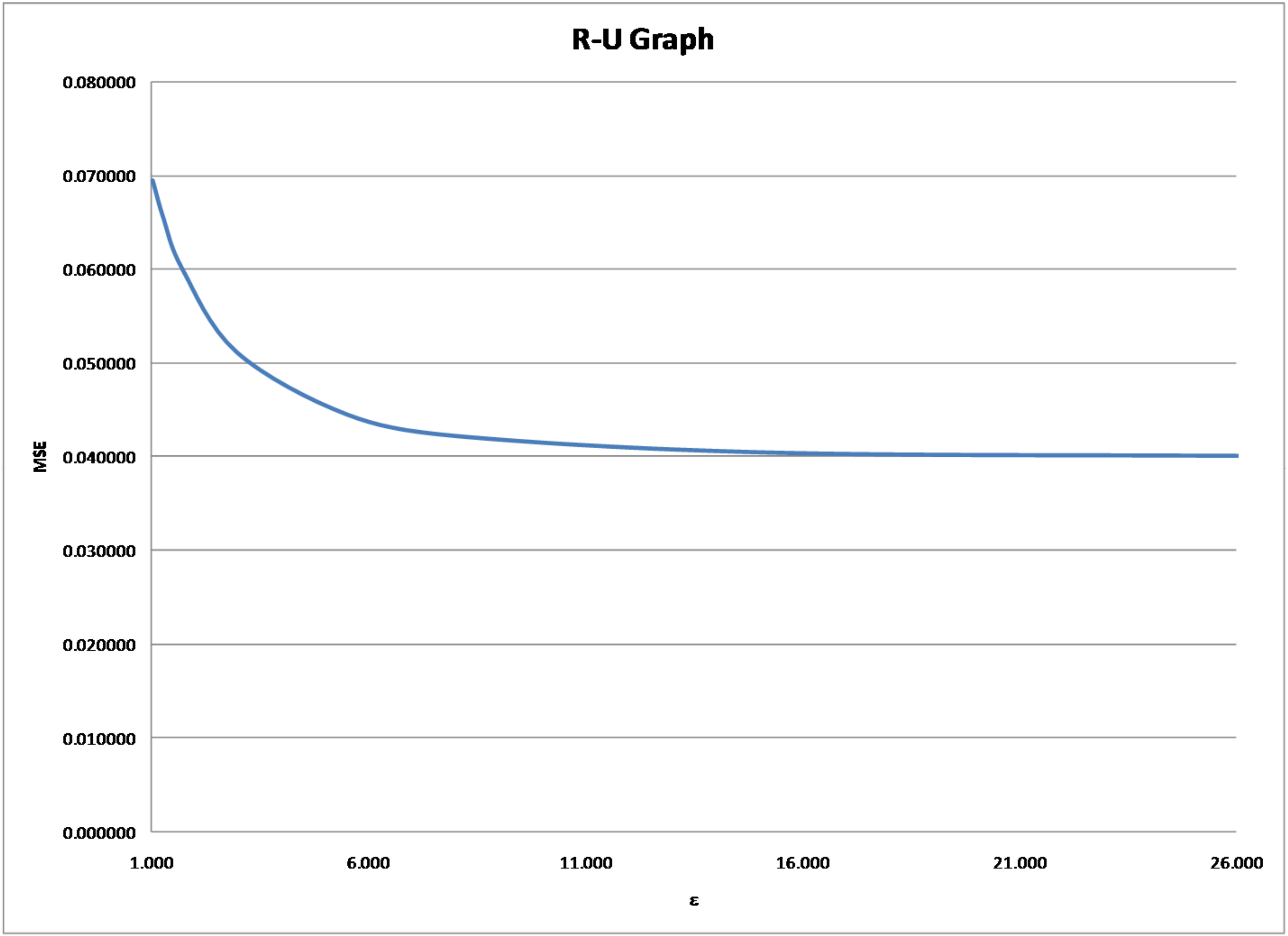
Analytical Validity

- The synthesizer’s analytical validity can be assessed with respect to estimates of the proportion in category 1.
- I use Mean Squared Error = Bias^2 + Variance as the analytical validity (data utility) criterion.
- Conditional validity depends on the bias and variance (MSE) in each row of the synthesizer.
- Unconditional validity depends on the expected value of the MSE over all possible confidential inputs.



	m_1	0	1	2	3	4	5	Expected	Bias	Variance	MSE
	m_2	5	4	3	2	1	0	Proportion			
n_1	n_2										
0	5	0.647	0.294	0.053	0.005	0.000	0.000	0.083333	0.083333	0.015278	0.022222
1	4	0.237	0.396	0.264	0.088	0.015	0.001	0.250000	0.050000	0.037500	0.040000
2	3	0.068	0.241	0.345	0.246	0.088	0.013	0.416667	0.016667	0.048611	0.048889
3	2	0.013	0.088	0.246	0.345	0.241	0.068	0.583333	-0.016667	0.048611	0.048889
4	1	0.001	0.015	0.088	0.264	0.396	0.237	0.750000	-0.050000	0.037500	0.040000
5	0	0.000	0.000	0.005	0.053	0.294	0.647	0.916667	-0.083333	0.015278	0.022222

- Table shows the conditional validity.
  - For example if the confidential data are (1,4), then the conditional analytical validity is 0.040000.
- The unconditional validity is MSE = 0.044444, which is the average of the six entries in the column labeled MSE with respect to the probability space induced on **n** by the Dirichlet prior.
- The R-U graph is shown on the next slide.


R-U Graph
0.080000
0.070000
0.060000
0.050000
0.040000
0.030000
0.020000
0.010000
0.000000
MSE
1.000
6.000
11.000
16.000
21.000
26.000
ε

Challenges and Research Program

- Realistic problems are all very sparse.
- Probabilistic differential privacy partially solves the sparsity problem.
  - But, it requires coarsening and domain shrinking to deliver acceptable analytical validity.
- Conditional analytical validity must be assessed by simulation, which requires very efficient synthesizers.