INFO 7470/ILRLE 7400 Methods of Confidentiality Protection

John M. Abowd and Lars Vilhuber

April 19, 2011

Outline

- What is Statistical Disclosure Limitation?
- What is Privacy-preserving Data Mining?
- Basic Methods for Disclosure Avoidance (SDL)
- Rules and Methods for Model-based SDL
- Noise Infusion Methods

Statistical Disclosure Limitation

- Protection of the confidentiality of the underlying micro-data
  - Avoiding identity disclosure
  - Avoiding attribute disclosure
  - Avoiding inferential disclosure
- Identity disclosure: who (or what entity) is in the confidential micro-data
- Attribute disclosure: value of a characteristic for that entity or individual
- Inferential disclosure: improvement of the posterior odds of a particular event

Privacy-preserving Datamining

- Formally define the properties of “privacy”
- Introduce algorithmic uncertainty as part of the statistical process
- Prove that the algorithmic uncertainty meets the formal definition of privacy

General Methods for Statistical Disclosure Limitation

- At the Census Bureau SDL is called Disclosure Avoidance Review
- Traditional methods
  - Suppression
  - Coarsening
  - Adding noise via swapping
  - Adding noise via sampling
- Newer methods
  - Explicit noise infusion
  - Synthetic data
  - Formal privacy-preserving sanitizers

Suppression

- This is by far the most common technique
- Model the sensitivity of a particular data item or observation (“disclosure risk”)
- Do not allow the release of data items that have excessive disclosure risk (primary suppression)
- Do not allow the release of other data from which the sensitive item can be calculated (complementary suppression)

Suppression in Model-base Releases

- Most data analysis done in the RDCs is model-based
- The released data consist of summary statistics, model coefficients, standard errors, some diagnostic statistics
- The SDL technique used for these releases is usually suppression: the suppression rules are contained (up to confidential parameters) in the RDC Researcher's Handbook

Coarsening

- Coarsening is the creation of a smaller number of categories from the variable in order to increase the number of cases in each cell
- Computer scientists call this “generalizing”
- Geographic coarsening: block-block group-tract-place-county-state-region
- Top coding of income is a form of coarsening
- All continuous variables in a micro-data file can be considered coarsened to the level of precision released
- This method is often applied to model-based data releases by restricting the number of significant digits that can be released

Swapping

- Estimate the disclosure risk of certain attributes or individuals
- If the risk is too great, attributes of one data record are (randomly) swapped with the same attributes of another record
- If geographic attributes are swapped this has the effect of placing the risky attributes in a different location from the truth
- Commonly used in household censuses and surveys
- Rarely used with establishment data

Sampling

- Sampling is the original SDL technique
- By only selecting certain entities from the population on which to collect additional data (data not on the frame), uncertainty about which entity was sampled provides some protection
- In modern, detailed surveys, sampling is of limited use for SDL

Rules and Methods for Model-based SDL

- Refer to Chapter 3 of the RDC Researcher's Handbook
- Suppress: coefficients on detailed indicator variables, on cells with too few entities
- Smooth: density estimation and quantiles, use a kernel density estimator to produce quantiles
- Coarsen: variables with heavy tails (earnings, payroll), residuals (truncate range, suppress labels of range)

Noise Infusion

- Introduction
- Application to the Quarterly Workforce Indicators
- Measures of Protection
- Measures of Analytical Validity

Explicit Noise Infusion

- Adding noise to the published item or to the underlying micro data to disguise the true value
- Example: QWIs and work place data in OTM
- Original method developed by Evans, T., Zayatz, L., and Slanta, J. (1998). “Using Noise for Disclosure Limitation of Establishment Tabular Data,” Journal of Official Statistics Vol. 14 (December): 537-51.

The Quarterly Workforce Indicator System

- Multiplicative noise infusion system
- Establishment level micro data are distorted according to a permanent distortion factor
- Distortion factor always moves the fuzzed item away from the actual item by a minimum and maximum percentage
- All release data are computed from the fuzzed items

References for QWI

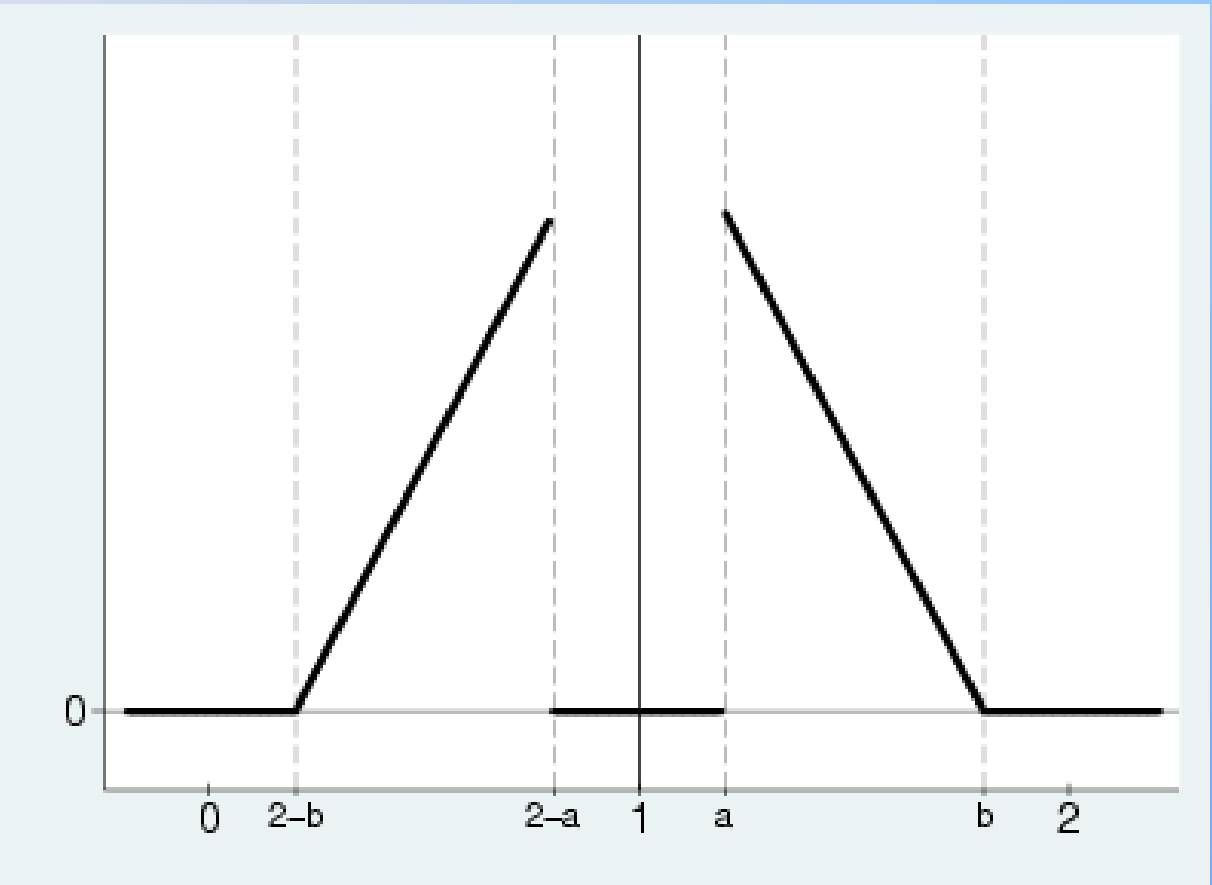
- Abowd, J. M., Stephens, B. E., Vilhuber L. (2005). Confidentiality Protection in the Census Bureau's Quarterly Workforce Indicators, mimeo, U.S. Census Bureau, LEHD and Cornell University (LEHD TP-2006-02)
- Abowd, J.M., Gittings, R.K., Stephens, B. E., and Vilhuber, L. "Combining Synthetic Data and Noise Infusion for Confidentiality Protection of the Quarterly Workforce Indicators," chapter 2 of Gittings Ph.D. thesis.

Noise Factor Distribution

$$p(\delta_j) = \begin{cases} (b-\delta)/(b-a)^2, & \delta \in [a,b] \\ (b+\delta-2)/(b-a)^2, & \delta \in [2-b,2-a] \end{cases}$$

$$F(\delta_j) = \begin{cases} 0.5 + \left[(b-a)^2 - (b-\delta)^2\right] / \left[2(b-a)^2\right], & \delta \in [a,b] \\ \left[(\delta+b-2)^2\right] / \left[2(b-a)^2\right], & \delta \in [2-b,2-a] \end{cases}$$

Graph of Noise Distribution



Implementation of Noise Infusion

- Counts: *B, E, M, F, A, S, H, R, FA, FS, W*1*, W*2*, W*3*, NA, NH, NR,* and *NS*
- Ratios: *Z_W*2*, Z_W*3*, Z_WA, Z_WS, Z_NA, Z_NH, Z_NR,* and *Z_NS*
- Differences: *JF, JC, JD, FJF, FJC, FJD, DWA, DWFA, DWS, DWFS*
- In *OTM*, *B* is distorted as a count

Multiplicative Noise Infusion

- *B* is beginning of quarter employment; *E* is end of period; *E-bar* is the average.
- *Z_W2* is end of quarter employee earnings, *W2* is total payroll for end of quarter employees.
- *JF* is net job flows
- Asterisk indicates distorted values.

$$B^{*}_{jt} = \delta_{j} \times B_{jt}$$

$$Z\_W^{*}_{2\,jt} = \frac{W^{*}_{2\,jt}}{E_{jt}} = \frac{\delta_{j} \times W_{2\,jt}}{E_{jt}}$$

$$JF^{*}_{kt} = G_{kt} \times \overline{E}^{*}_{kt} = JF_{kt} \times \frac{\overline{E}^{*}_{kt}}{\overline{E}_{kt}}$$

$$Z\_DWA^{*}_{kt} = \frac{\Delta WA_{kt}}{A_{kt}} \times \frac{A^{*}_{kt}}{A_{kt}}$$

Weighting

- Each fuzzed micro-data item is weighted by the QWI final weight before aggregation
- This means that all input data are real numbers (not integers)
- Final disclosure control formulas must reflect rounding of the counts

Interpreting the Algorithm

- Based on the public use version of the spreadsheet

Public_use_variable	Variable_stem	Employee_disclosure_stem_1	Employee_disclosure_stem_2	Employee_disclosure_stem_3	Employer_disclosure_stem_1	Employer_disclosure_stem_2	Employer_disclosure_stem_3	Disclosure_type	Fuzz_type	Fuzz_value_numerator	Fuzz_value_denominator	Fuzz_change_ratio_numerator	Fuzz_change_ratio_denominator
BeginEmp	B	B	B	B	n_B	n_B	n_B	count	count	delta_B	unity	unity	unity
EndEmp	E	E	E	E	n_E	n_E	n_E	count	count	delta_E	unity	unity	unity
JobFlowNet	JF	EBAR	EBAR	EBAR	n_EBAR	n_EBAR	n_EBAR	count	change	JF	unity	delta_EBAR	EBAR
JobCreate	JC	EBAR	EBAR	EBAR	n_EBAR	n_EBAR	n_EBAR	count	change	JC	unity	delta_EBAR	EBAR
JobDest	JD	EBAR	EBAR	EBAR	n_EBAR	n_EBAR	n_EBAR	count	change	JD	unity	delta_EBAR	EBAR
Acc	A	A	A	A	n_A	n_A	n_A	count	count	delta_A	unity	unity	unity
Hire	H	H	H	H	n_H	n_H	n_H	count	count	delta_H	unity	unity	unity
Recl	R	R	R	R	n_R	n_R	n_R	count	count	delta_R	unity	unity	unity
Sep	S	S	S	S	n_S	n_S	n_S	count	count	delta_S	unity	unity	unity
FulEmp	F	F	F	F	n_F	n_F	n_F	count	count	delta_F	unity	unity	unity
FulJobFlw	FJF	FBAR	FBAR	FBAR	n_FBAR	n_FBAR	n_FBAR	count	change	FJF	unity	delta_FBAR	FBAR
FulJobCre	FJC	FBAR	FBAR	FBAR	n_FBAR	n_FBAR	n_FBAR	count	change	FJC	unity	delta_FBAR	FBAR
FulJobDes	FJD	FBAR	FBAR	FBAR	n_FBAR	n_FBAR	n_FBAR	count	change	FJD	unity	delta_FBAR	FBAR
FulQrtTurn	FT	F	FA	FS	n_F	n_FA	n_FS	ratio	aggregate	delta_FAFS	delta_F	unity	unity
FulEmpFlw	FA	FA	FA	FA	n_FA	n_FA	n_FA	count	count	delta_FA	unity	unity	unity
FulHire	H3	H3	H3	H3	n_H3	n_H3	n_H3	count	count	delta_H3	unity	unity	unity
FulSep	FS	FS	FS	FS	n_FS	n_FS	n_FS	count	count	delta_FS	unity	unity	unity
EarnEnd	Z_W2	E	E	E	n_E	n_E	n_E	ratio	ratio	delta_W2	E	unity	unity
EarnFul	Z_W3	F	F	F	n_F	n_F	n_F	ratio	ratio	delta_W3	F	unity	unity
EarnFulAcc	Z_WFA	FA	FA	FA	n_FA	n_FA	n_FA	ratio	ratio	delta_WFA	FA	unity	unity
ChgEarnAcc	Z_dWA	A	A	A	n_A	n_A	n_A	ratio	change_ratio	dWA	A	delta_A	A
NonEmpAcc	Z_NA	A	A	A	n_A	n_A	n_A	ratio	ratio	delta_NA	A	unity	unity
NonEmpHire	Z_NH	H	R	R	n_H	n_R	n_R	ratio	ratio	delta_NH	H	unity	unity
NonEmpRecl	Z_NR	R	H	H	n_R	n_H	n_H	ratio	ratio	delta_NR	R	unity	unity
EarnFulSep	Z_WFS	FS	FS	FS	n_FS	n_FS	n_FS	ratio	ratio	delta_WFS	FS	unity	unity
ChgEarnSep	Z_dWS	S	S	S	n_S	n_S	n_S	ratio	change_ratio	dWS	S	delta_S	S
NonEmpSep	Z_NS	S	S	S	n_S	n_S	n_S	ratio	ratio	delta_NS	S	unity	unity
EarnFulHire	Z_WH3	H3	H3	H3	n_H3	n_H3	n_H3	ratio	ratio	delta_WH3	H3	unity	unity
Payroll	W1	M	M	M	n_M	n_M	n_M	sum	sum	delta_W1	unity	unity	unity
TotalEmp	M	M	M	M	n_M	n_M	n_M	count	count	delta_M	unity	unity	unity

Protection Properties of the QWI Algorithm

Table 1A

Variable: B

Unweighted/Undistorted vs. Unweighted/Distorted

	0	1	2	3	4	5+
0	99.61	0.39	0.00	0.00	0.00	0.00
1	0.00	98.57	1.43	0.00	0.00	0.00
2	0.00	1.04	96.10	2.85	0.00	0.00
3	0.00	0.00	2.19	93.21	4.60	0.00
4	0.00	0.00	0.00	7.30	82.52	10.18
5+	0.00	0.00	0.00	0.00	1.60	98.40

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent row percentages and sum to 100.

Table 1B

Variable: B

Unweighted/Undistorted vs. Weighted/Distorted

	0	1	2	3	4	5+
0	99.19	0.81	0.00	0.00	0.00	0.00
1	0.14	89.29	10.56	0.01	0.00	0.00
2	0.04	1.39	67.45	30.70	0.42	0.00
3	0.03	0.04	2.19	50.99	42.76	3.99
4	0.03	0.02	0.03	3.07	41.04	55.81
5+	0.01	0.00	0.00	0.02	0.33	99.64

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent row percentages and sum to 100.

Table 1C

Variable: B

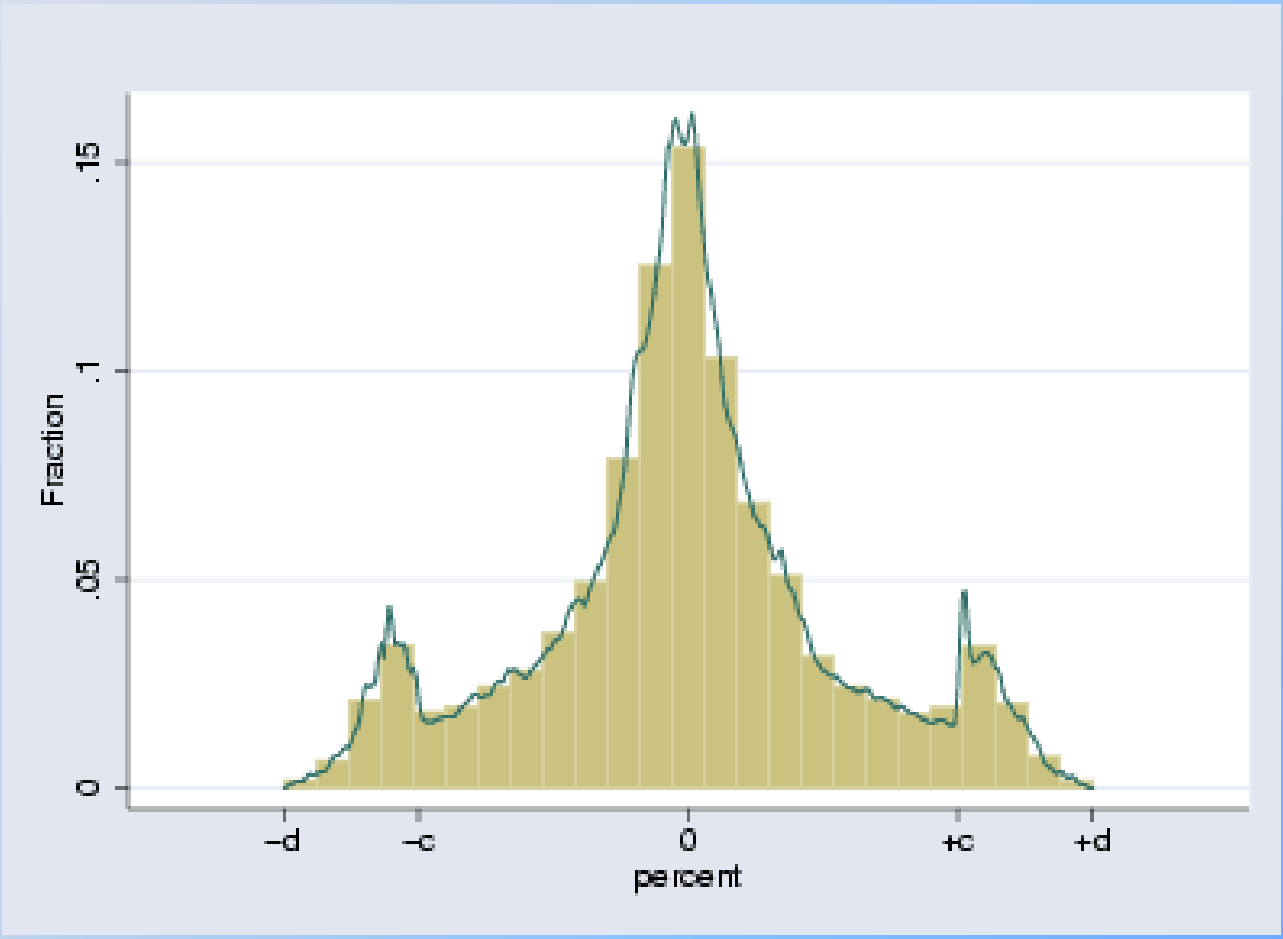
Unweighted/Undistorted vs. Synthesized

	0	1	2	3	4	5+
0	99.17	0.82	0.01	0.00	0.00	0.00
1	7.85	84.74	6.62	0.78	0.01	0.00
2	0.51	11.93	61.06	24.14	2.24	0.12
3	0.06	0.76	7.53	47.50	39.13	5.02
4	0.03	0.11	0.93	7.40	38.84	52.69
5+	0.01	0.01	0.01	0.11	0.71	99.16

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent row percentages and sum to 100.

Analytical Validity Properties

Error Distribution for B (Micro-data)



Error Distribution for W (Micro-data)

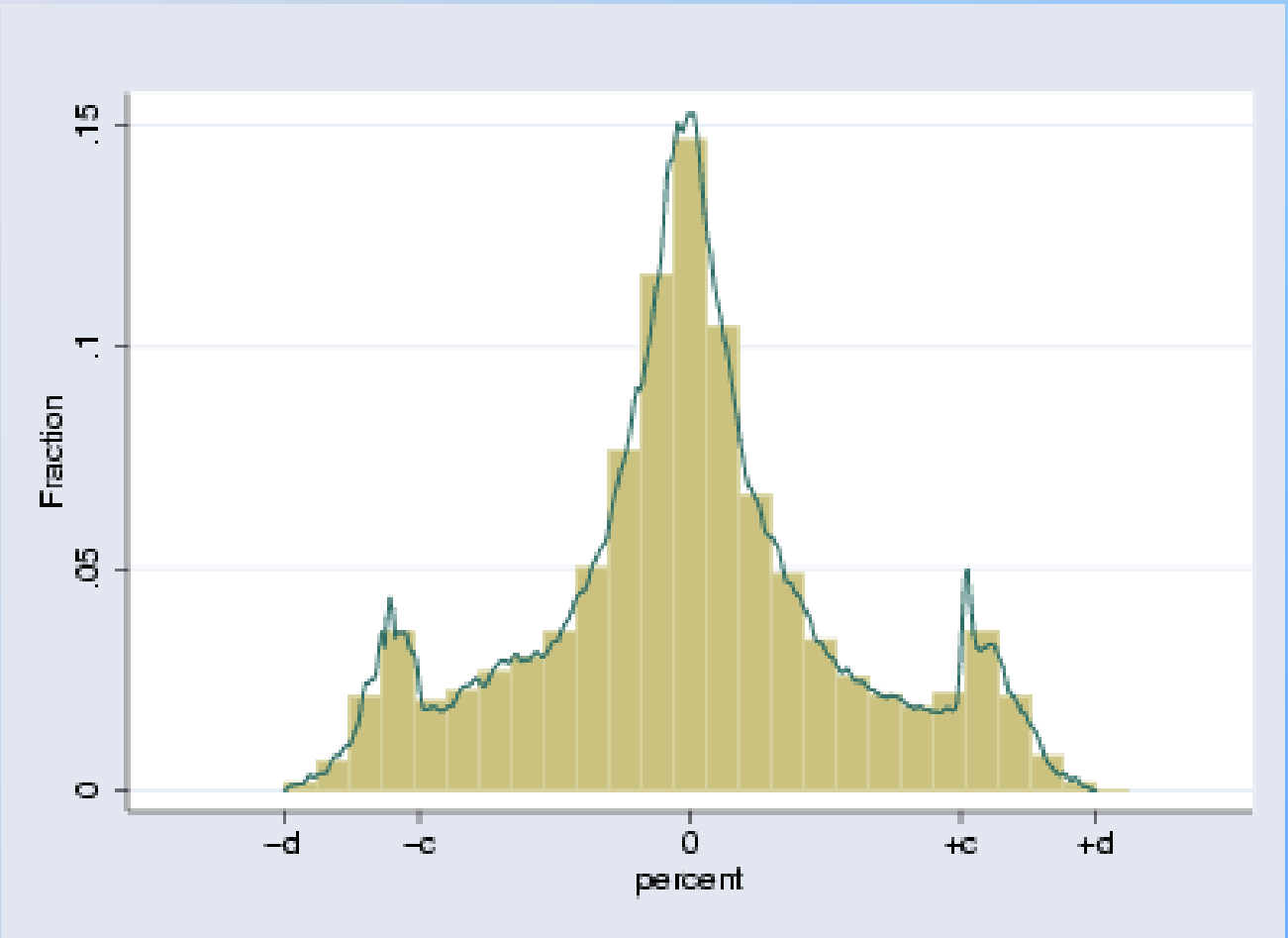


Table 7

Distribution of the Difference Between Autocorrelation Coefficients
Unweighted/Undistorted vs. Unweighted/Distorted

Percentile	B	H	R	E	A	S	M	F	FA	FS	H3
99	0.067	0.058	0.042	0.066	0.061	0.061	0.079	0.059	0.056	0.054	0.056
95	0.036	0.035	0.025	0.036	0.036	0.036	0.042	0.032	0.034	0.033	0.035
90	0.025	0.025	0.015	0.024	0.025	0.025	0.028	0.021	0.024	0.023	0.024
75	0.009	0.009	0.005	0.009	0.010	0.010	0.011	0.008	0.009	0.008	0.008
50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25	-0.008	-0.011	-0.005	-0.009	-0.010	-0.010	-0.011	-0.008	-0.009	-0.008	-0.009
10	-0.023	-0.028	-0.016	-0.023	-0.026	-0.026	-0.027	-0.022	-0.026	-0.024	-0.026
5	-0.035	-0.038	-0.025	-0.037	-0.036	-0.038	-0.041	-0.033	-0.037	-0.035	-0.038
1	-0.061	-0.063	-0.045	-0.065	-0.059	-0.065	-0.074	-0.061	-0.061	-0.059	-0.062

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.

Table 8

Distribution of the Difference Between Autocorrelation Coefficients
Unweighted/Undistorted vs. Weighted/Distorted w/ Suppressions

Percentile	B	H	R	E	A	S	M	F	FA	FS	H3
99	0.318	0.592	0.538	0.325	0.602	0.652	0.330	0.366	0.606	0.656	0.648
95	0.153	0.326	0.221	0.158	0.294	0.304	0.134	0.181	0.317	0.352	0.359
90	0.086	0.196	0.136	0.084	0.173	0.181	0.063	0.108	0.196	0.218	0.229
75	0.021	0.060	0.049	0.019	0.053	0.053	0.017	0.025	0.068	0.070	0.087
50	-0.002	-0.006	0.000	-0.003	-0.006	-0.008	-0.006	-0.001	-0.004	-0.004	-0.001
25	-0.029	-0.075	-0.063	-0.031	-0.071	-0.078	-0.037	-0.026	-0.070	-0.084	-0.082
10	-0.085	-0.188	-0.162	-0.091	-0.170	-0.190	-0.099	-0.077	-0.169	-0.200	-0.190
5	-0.150	-0.289	-0.242	-0.154	-0.258	-0.288	-0.176	-0.141	-0.262	-0.300	-0.286
1	-0.393	-0.551	-0.474	-0.426	-0.505	-0.531	-0.489	-0.419	-0.499	-0.538	-0.527

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.

Table 9

Distribution of the Difference Between Autocorrelation Coefficients
Unweighted/Undistorted vs. Weighted/Distorted w/ Synthetic Replacements

Percentile	B	H	R	E	A	S	M	F	FA	FS	H3
99	0.107	0.246	0.223	0.096	0.238	0.263	0.210	0.227	0.260	0.157	0.294
95	0.056	0.144	0.126	0.051	0.134	0.166	0.123	0.112	0.145	0.076	0.185
90	0.036	0.104	0.092	0.033	0.099	0.120	0.084	0.075	0.108	0.049	0.133
75	0.011	0.050	0.043	0.010	0.047	0.058	0.037	0.031	0.051	0.018	0.066
50	-0.004	0.003	0.007	-0.005	0.005	0.007	0.006	0.002	0.007	-0.003	0.009
25	-0.030	-0.043	-0.013	-0.030	-0.032	-0.032	-0.017	-0.018	-0.030	-0.034	-0.037
10	-0.065	-0.109	-0.067	-0.064	-0.089	-0.095	-0.049	-0.056	-0.095	-0.082	-0.128
5	-0.092	-0.167	-0.120	-0.093	-0.142	-0.145	-0.081	-0.088	-0.144	-0.116	-0.194
1	-0.166	-0.295	-0.257	-0.176	-0.257	-0.268	-0.168	-0.187	-0.277	-0.233	-0.343

Note: The data represent county data for Maryland/NAICS Industry Group. Cells represent the difference between the autocorrelation coefficients the percentile designated by the rows.

QWIPU: Aggregating Formulas

- Counts and magnitudes:
  - Add
- Ratios and differences
  - Multiply by released base
  - Aggregate numerator and denominator separately
  - Add
- Job creations and destructions
  - Handle like counts but understand that there is an inherent loss of information

QWIPU: Handling the Suppressions

- Status flag 5 suppressions
  - Must be treated as missing data and estimated
- Status flag 9 data
  - Not suppressed. Should be used as published.

Research Uses of the Micro-data

- QWI Micro-data (UFF_B)
  - Use the undistorted confidential data (establishment level) in models
  - Use conventional model-based disclosure avoidance rules
  - No fuzzed data should be used
- OTM Micro-data (WHAT_B)
  - All micro-data and all imputations of missing data used to build OTM
  - Contain both fuzzed and unfuzzed values with separate weights
  - Unfuzzed data should be used in models, weighted as appropriate

INFO 7470/ILRLE 7400 Cryptographic and Statistical Advances in Confidentiality Protection

John M. Abowd and Lars Vilhuber

April 19, 2011

Outline

- Motivation: formal privacy models and statistical disclosure limitation
- The basic OnTheMap application
- The statistical structure of the OnTheMap data
- Applying probabilistic differential privacy to OnTheMap
- The trade-off between analytical validity and confidentiality protection

Formal Privacy Models and Statistical Disclosure Limitation

- Formal privacy protection methods are based on open algorithms with provable properties
- The standard in privacy-preserving datamining is based on cryptography:
  - Only the private key (password, encryption key) is confidential; all algorithms and parameters are public
  - Attacker (= user) can have massive amounts of prior information

The Cryptographic Critique of SDL

- Standard SDL techniques fail because:
  - They and do not have provably protective properties when the attacker (= user) is allowed full access to the algorithm
  - They depend upon the realized data and not the algorithm
- Many standard SDL techniques are viewed as very risky when the cryptographic critique is applied

Point of Common Ground

- Federal Committee on Statistical Methodology working paper 22 offers the desirable disclosure avoidance property:

  Disclosure relates to inappropriate attribution of information to a data subject, whether an individual or an organization. Disclosure occurs when a data subject is identified from a released file (identity disclosure), sensitive information about a data subject is revealed through the released file (attribute disclosure), or the released data make it possible to determine the value of some characteristic of an individual more accurately than otherwise would have been possible (inferential disclosure). (page 4)

- Evfimievski, Gehrke and Srikant (2003), Dwork (2006) show that disclosure avoidance in this sense is impossible to achieve in general.

Focus on Synthetic Data and Randomized Sanitizers

- The SDL technique known as synthetic data most closely resembles the cryptographic data protection techniques
- The cryptographic techniques are known as privacy-preserving datamining, randomized sanitizers, differential privacy, and e-privacy.

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.

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 confidential data
- The randomness in a sanitizer is induced by the properties of the distribution of *U*
- The PPD is just a particular randomized sanitizer

ε-Differential Privacy

Definition (ε - Differential Privacy): Let A be a randomized algorithm, let S be the set of all possible outputs of the algorithm, and let $\varepsilon > 0$. The algorithm A satisfies ε - differential privacy if for all pairs of data sets (D_1, D_2) that differ in exactly one row,

$$\forall S \in \mathrm{S}, \frac{P(\mathrm{A}(D_1)) = S}{P(\mathrm{A}(D_2)) = S} \leq e^{\varepsilon} \text{ or } \left| \ln \frac{P(\mathrm{A}(D_1)) = S}{P(\mathrm{A}(D_2)) = S} \right| < \varepsilon.$$

- Differential privacy (Dwork, and many co-authors) is difficult to maintain in sparse applications when geographically near blocks have very different posterior probabilities

Disclosure Set

Definition (Disclosure Set) : Let D be a table and D be the set of tables that differ from D in at most one row. Let A be a randomized algorithm and S be the space of outputs of the algorithm A. The disclosure set of D, denoted $\mathrm{Disc}(D,\varepsilon)$, is

$$\left\{ S \in \mathsf{S} \middle| \exists X_1, X_2 \in \mathsf{D}(D), \left| X_1 \setminus X_2 \right| = 1 \wedge \left| \ln \frac{P(\mathsf{A}(X_1) = S)}{P(\mathsf{A}(X_2) = S)} \right| > \varepsilon \right\}.$$

- This set describes the outcomes where differential privacy fails

Probabilistic Differential Privacy

Definition (Probabilistic Differential Privacy): Let A be a randomized algorithm and S be the space of outputs of A. Let $\varepsilon > 0$ and $0 < \delta < 1$ be constants. Then A satisfies (ε, δ)-probabilistic differential privacy (or (ε, δ)-pdp) if for all tables D, $P(\mathsf{A}(\mathsf{D}) \in \mathrm{Disc}(D, \varepsilon)) \leq \delta$.

- PDP allows us to control the probability that differential privacy fails
- The analytical validity of sparse applications can be controlled with PDP because the restrictions on the prior used in the synthesizer are reasonable for use with sparse tables

Disclosure Limitation Definitions

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

$\tilde{X} = \tilde{x}$, realization of the synthesizer

- 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 or the random sanitizer
- The PPD or the random sanitizer define the transition probabilities from the confidential data to the release data
- True for all SDL procedures that infuse noise

Connection Between Synthetic Data and Differential Privacy

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

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 specified conditional distribution.

Connection Between Differential Privacy and Inferential Disclosure

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

The posterior odds ratio for the gain in information about a single row of X is the Dalenius (1977) definition of an inferential disclosure. Bounding the differential privacy therefore bounds the inferential disclosure.

Taking Account of Formal Privacy Models

- A variety of papers in the cryptographic data privacy literature (Dwork, Nissim and their many collaborators, Gehrke and his collaborators, and others) show that the confidentiality protection afforded by synthetic data or a randomized sanitizer depends upon properties of the transition probabilities that relate the confidential data to the release data.
- Exact data releases are not safe. Not surprising since

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

  implies that the sanitizer leaves the confidential data unchanged .
- Off-diagonal elements that are zero imply infinite differential privacy: exact disclosure in some cases with probability 1.
- For a full explanation of the relation between the transition matrix and differential privacy measures see Abowd and Vilhuber (2008).

Relationship to Post-randomization

- Post-randomization (Kooiman et al. 1997) focuses on the diagonal elements of

$$\Pr\left[\tilde{X} \middle| X\right]$$

- When off-diagonal elements of this transition matrix are zero, infinite differential privacy usually results
- Swapping, shuffling, stratified sampling, and most noise-infusion methods result in off-diagonal elements that are zero

A DETAILED EXAMPLE: SYNTHETIC DATA

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

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

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

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

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

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

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

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

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

The Multinomial-Dirichlet Synthesizer

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

- The synthetic data are samples from the synthesizer, and 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

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

- The two confidential data matrices, $\mathbf{n}^{(1)}$ and $\mathbf{n}^{(2)}$ differ by changing exactly one entity's data
- Bounding by ε the log inferential disclosure odds ratio in the M-D synthesizer amounts to controlling the probabilities in $\Pr[\mathbf{m}|\mathbf{n}]$ appropriately

				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) (or (0,5) with original data (1,4) or (5,0))
- If I release (5,0) and you know 4 people are in category 2, then the odds are 243:1 (= exp(5.493061)) that the unknown person is in category 1

Probabilistic Differential Privacy

- This definition of differential privacy allows the ε-differential privacy limit to fail with probability δ (Machanavajjhala *et al.* 2008)
- To compute the PDP, the joint distribution of **m** and **n** must be examined for outcomes with differential privacy that exceed the limit to ensure that they occur with total 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

- The table is Pr[**m**,**n**], where the marginal Pr[**n**] is based on the prior D(**α**)
- If we want to have ε-differential privacy of 2, then the synthesizer fails in the highlighted cells
- With prior D(**α**), probabilistic differential privacy has $\varepsilon = 2$ and $\delta = 0.000623$, which is just the sum of the highlighted cells

A DETAILED EXAMPLE: RANDOM SANITIZER

Laplace Sanitizer

- Dwork *et al.* (2006) show that ε-differential privacy can be achieved in the Multinomial model with a sanitizer using independent double exponential noise (Laplace noise) with mean zero and variance $2/\varepsilon$
- Note that in our application the total n is released without noise

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

$$u \sim i.i.d\ \mathrm{Lap}\left(0, \frac{2}{\varepsilon}\right)$$

	m_1	0	1	2	3	4	5
	m_2	5	4	3	2	1	0
n_1	n_2						
0	5	0.816060	0.159046	0.021525	0.002913	0.000394	0.000062
1	4	0.183940	0.632121	0.159046	0.021525	0.002913	0.000456
2	3	0.024894	0.159046	0.632121	0.159046	0.021525	0.003369
3	2	0.003369	0.021525	0.159046	0.632121	0.159046	0.024894
4	1	0.000456	0.002913	0.021525	0.159046	0.632121	0.183940
5	0	0.000062	0.000394	0.002913	0.021525	0.159046	0.816060

- $k = 2$
- $n = m = 5$
- $\varepsilon = 2$
- The table displays the transition probabilities that map **n** into **m**
- Note that the diagonals are larger than the M-D model and the extreme outcomes have greater probability

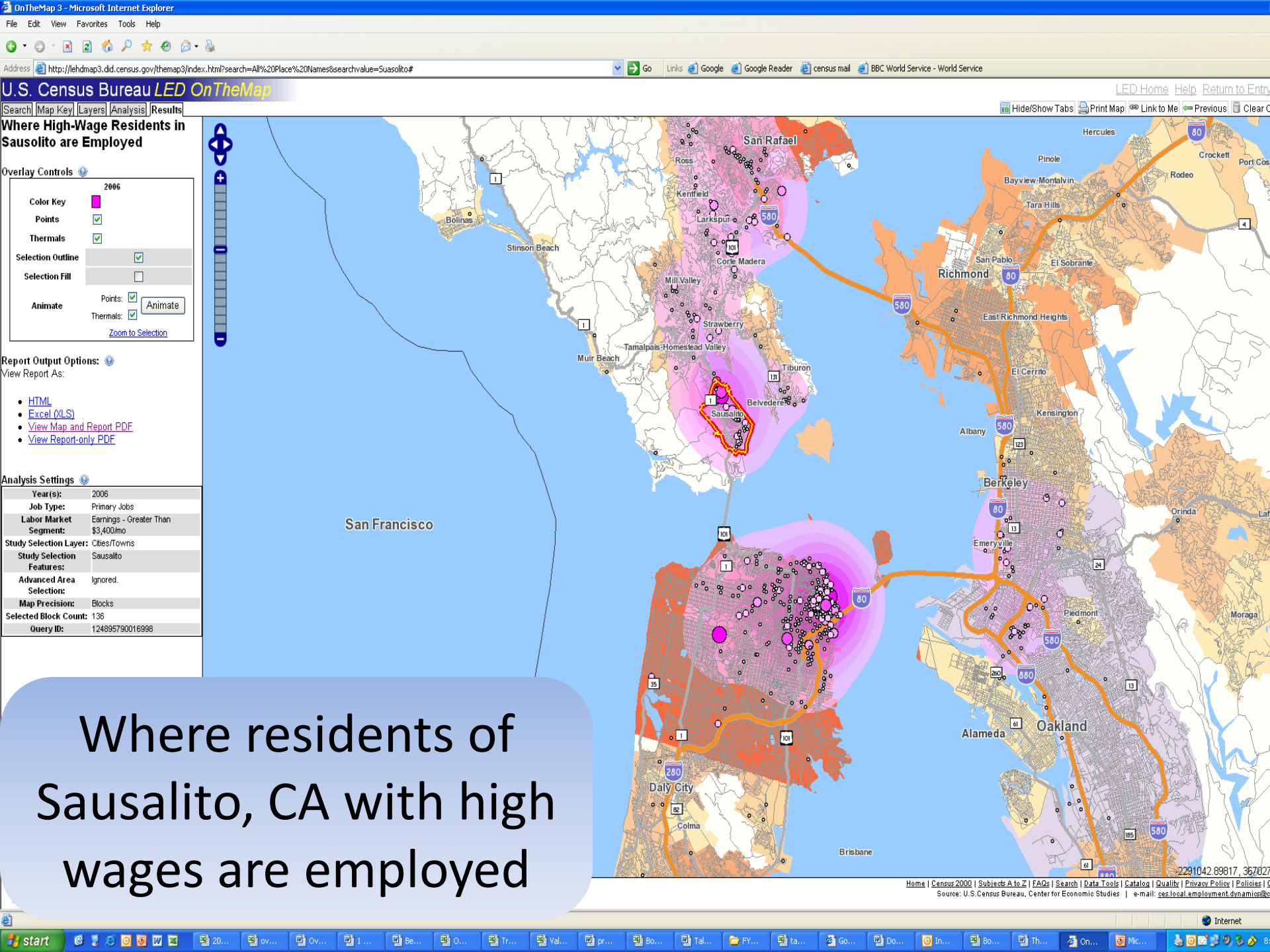
				m_1	0	1	2	3	4	5
				m_2	5	4	3	2	1	0
$n^{(1)}{}_1$	$n^{(1)}{}_2$	$n^{(2)}{}_1$	$n^{(2)}{}_2$							
0	5	1	4		1.489880	1.379885	2.000000	2.000000	2.000000	2.000000
1	4	2	3		2.000000	1.379885	1.379885	2.000000	2.000000	2.000000
2	3	3	2		2.000000	2.000000	1.379885	1.379885	2.000000	2.000000
3	2	4	1		2.000000	2.000000	2.000000	1.379885	1.379885	2.000000
4	1	5	0		2.000000	2.000000	2.000000	2.000000	1.379885	1.489880

- The table confirms that the transition matrix on the previous page has $\varepsilon = 2$

Challenges and Applications

- Realistic problems are all very sparse
- Probabilistic differential privacy can solve the sparseness problem
  - But, it requires coarsening and domain shrinking to deliver acceptable analytical validity.
- The Laplace synthesizer can solve the sparseness problem by adaptive histogram coarsening
  - But the user cannot directly control the coarsening hence analytical validity for some hypotheses is low
- OnTheMap uses probabilistic differential privacy

A REAL APPLICATION: US CENSUS BUREAU'S ONTHEMAP



Where residents of Sausalito, CA with high wages are employed

The OnTheMap Data Structure

- Set of linked data tables with a relational database schema
- Main tables (micro-data)
  - Job: [Person_ID, Employer_ID]
  - Residence: [Person_ID, Origin_Block, …]
  - Workplace: [Employer_ID, Destination_Block, …]
  - Geo-code: [Block, Tract, Latitude, Longitude, …]

Detailed Geo-spatial Data in OTM

- Workplace and residence geographies are defined using Census blocks
- Statistical analysis to estimate the PPD is based on Census tract-to-tract relations
- There are 8.2 million blocks and 65,000 tracts in the U.S.
- Every workplace block with positive employment has its own synthesizer

Dirichlet-Multinomial Synthesizer

- *I* origins
- Model each destination *d* separately for each demographic segment (age, earnings, industry)
- Sample data **X** tabulated into **n**
- Synthetic data tabulated into **m**
- Usually *m* = *n,* but not in the OTM application

$$\mathbf{n} = (n_1, \ldots, n_I), n = \sum n_i$$

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I), \alpha_0 = \sum \alpha_i = |\boldsymbol{\alpha}|$$

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_I)$$

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

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

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

$$\mathbf{m} = (m_1, \ldots, m_I), m = \sum m_i$$

$$\mathbf{m} \sim \mathrm{M}(\boldsymbol{\pi}, m), \text{ a posteriori}$$

Synthetic Data Model

- Likelihood of place of residence (index *i*) conditional on place of work (index *j*) and characteristics (index k):

$$p(n_{ijk} \mid \pi_{i|jk}) \propto \prod_{i=1}^{I} \pi_{i|jk}^{n_{ijk}}$$

- The resulting posterior for π is Dirichlet with parameter $\mathbf{n}_{jk} + \boldsymbol{\alpha}_{jk}$ for each unique workplace and characteristic combination (age, earnings, industry).
- Synthesize residence counts by sampling from the posterior predictive distributions conditional on already protected (and published) destination employment counts, m_{jk}

Search Algorithm Implements PDP

- We rely on the concept of (ε, δ)-probabilistic differential privacy, which guarantees ε-differential privacy with 1-δ confidence (Machanavajjhala *et al.* (2008)).
- Search algorithm finds the minimum prior sample size to guarantee ε-differential privacy with failure probability δ.
- This minimum prior sample size is then used as the lower bound of the prior sample sizes (α) .
- The privacy-preserving algorithm implemented in *OnTheMap* guarantees ε-differential privacy protection of 8.99 with 99.999999% confidence (δ = 0.000001).

Measures to Improve Validity

- Coarsening of the domain
  - Reducing the number of support points in the domain of the prior
- Editing the prior domain
  - Eliminating the most unlikely commute patterns (from prior and likelihood) based on previously published data
- Use of informative priors
  - Impose likely shape based on previously published data subject to minimum prior sample size that ensures (ε,δ)-PDP
- Pruning the prior
  - Randomly eliminating a fraction support points with no likelihood support.
  - Pruning comes with a penalty in terms of privacy protection

Refinement: Coarsening the Domain

- Blocks are collected into larger geographic areas- SuperPUMAs, PUMAs, Tracts
- Reduces the dimensionality of the domain of each destination's synthesizer
- Theorem 5.1 in Machanavajjhala et al. shows that ε-differential privacy, and (ε, δ)-probabilistic differential privacy both survive coarsening with unchanged parameters

Coarsening Steps

- If origin block very far away from destination block (distance > 90th percentile of CTTP commute distribution) coarsened to Super-PUMA (400,000 population in Census 2000)
- Else if origin block far away from destination block (distance > 50th percentile of CTTP commute distribution) coarsened to PUMA (100,000 population in Census 2000)
- Else if origin block close to destination block (distance < 50th percentile of CTTP commute distribution) coarsened to Census Tract (4,000 population on average).
- Idea: “marginal differences in commute distances between candidate locations have less predictive power in allocating workers the farther away the locations are”

Effects of Coarsening

- Coarsening in formal privacy models is effectively the same as coarsening in traditional methods
- After coarsening, an entity (in this case a block) is chosen randomly to represent the coarsened unit (one block per SuperPUMA, PUMA, or tract, as appropriate)
- This ensures that the transition matrix has no zero elements at the block level
- Ratios of the elements of this transition matrix determine the differential privacy

Refinement: Editing the Prior Domain

- For each work tract:
  - if point in domain has zero probability in prior data then do:
    - eliminate point with p=0.98 if distance > 500 miles
    - eliminate point with p=0.9 if distance > 200 miles
    - eliminate point with p=0.5 if distance > 100 miles
    - do not eliminate if distance < 100 miles
  - else retain point
- Note: contribution of any likelihood data in eliminated points also eliminated

Fraction of Points in the Prior Domain with Positive Counts in Census Transportation Planning Package Data

	State A		State B		State C	
Distance (in miles)	Mean	SD	Mean	SD	Mean	SD
- low-10	0.47	0.37	0.40	0.32	0.92	0.18
- 10-25	0.30	0.26	0.19	0.19	0.63	0.29
- 25-100	0.01	0.13	0.09	0.10	0.15	0.16
- 100-500	0.01	0.03	0.01	0.02	0.02	0.04
- 500-high	0.00	0.01	0.00	0.01	0.00	0.00
All	0.18	0.28	0.14	0.23	0.34	0.40

Fraction of Points in the Domain with Positive Counts in CTPP after Eliminating Extremely Unlikely Commute Patterns

	Large State		Medium State		Small State	
distance (in miles)	Mean	SD	Mean	SD	Mean	SD
- low-10	0.47	0.37	0.40	0.32	0.92	0.18
- 10-25	0.30	0.26	0.19	0.19	0.63	0.29
- 25-100	0.13	0.13	0.09	0.10	0.15	0.16
- 100-500	0.06	0.09	0.03	0.06	0.08	0.14
- 500-high	0.07	0.13	0.06	0.12	0.03	0.08
All	0.21	0.27	0.15	0.23	0.36	0.39

Fraction of likelihood data eliminated by eliminating unlikely commute patterns is about 3-7% depending on state and year

Support Points in Prior Domain (before pruning)

	Large State (A)			Medium State (B)			Small State (C)		
Support points:	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Total	1,005	583	2,067	1,027	619	1,560	672	602	818
By level of coarsening									
- Super-PUMA	526	519	538	526	518	539	537	535	539
- PUMA	39	9	73	47	7	79	10	4	19
- Census Tract	438	32	1,506	453	72	998	125	56	272
By distance (in miles) between centroids									
- low-10	265	1	878	188	1	438	15	1	49
- 10-25	127	8	794	195	13	612	16	1	60
- 25-100	85	23	289	121	45	296	54	15	169
- 100-500	139	119	206	181	151	238	80	29	233
- 500-high	389	361	412	343	300	373	508	486	519

Refinement: Informative Priors

- In year 2002: Public-use CTTP data
- In year 2003-2008: Public-use previous year *OnTheMap* data (not posterior)
- α = max[min_alpha, f(prior density)] minimum prior sample size is the larger of the PDP value (min_alpha) or the informative prior value
- Priors unique to each employment tract
- Not strictly Bayesian because the posterior is not published, and published data are required for prior by PDP

Refinement: Domain Pruning

- Domain may still have too many blocks for good analytical validity
- Algorithm 2 prunes the domain for a given destination *j*:
  - Keep all origins in the likelihood support (confidential data)
  - For all other origins, add to domain with probability f_i; (generates min_p below)
  - From Machanavajjhala et al. 2008:

Theorem 5.2 (summary): Applying the domain pruning algorithm 2 changes the (ε, δ)-pdp to

$$\varepsilon' = \varepsilon + \max_{i \in \{i | n_i = 0\}} \left(\ln\left(1/f_i\right)\right) + \max_{i \in \{i | n_i = 0\}} \lceil \alpha_i \rceil \ln 2$$

Effects of Domain Pruning

- Domain pruning leaves all of the support points that appear in the likelihood function in the posterior
- Domain pruning removes some of the prior support points that have no likelihood
- Domain pruning improves analytical validity, but because it depends upon the confidential data, it increases the effective differential privacy limit

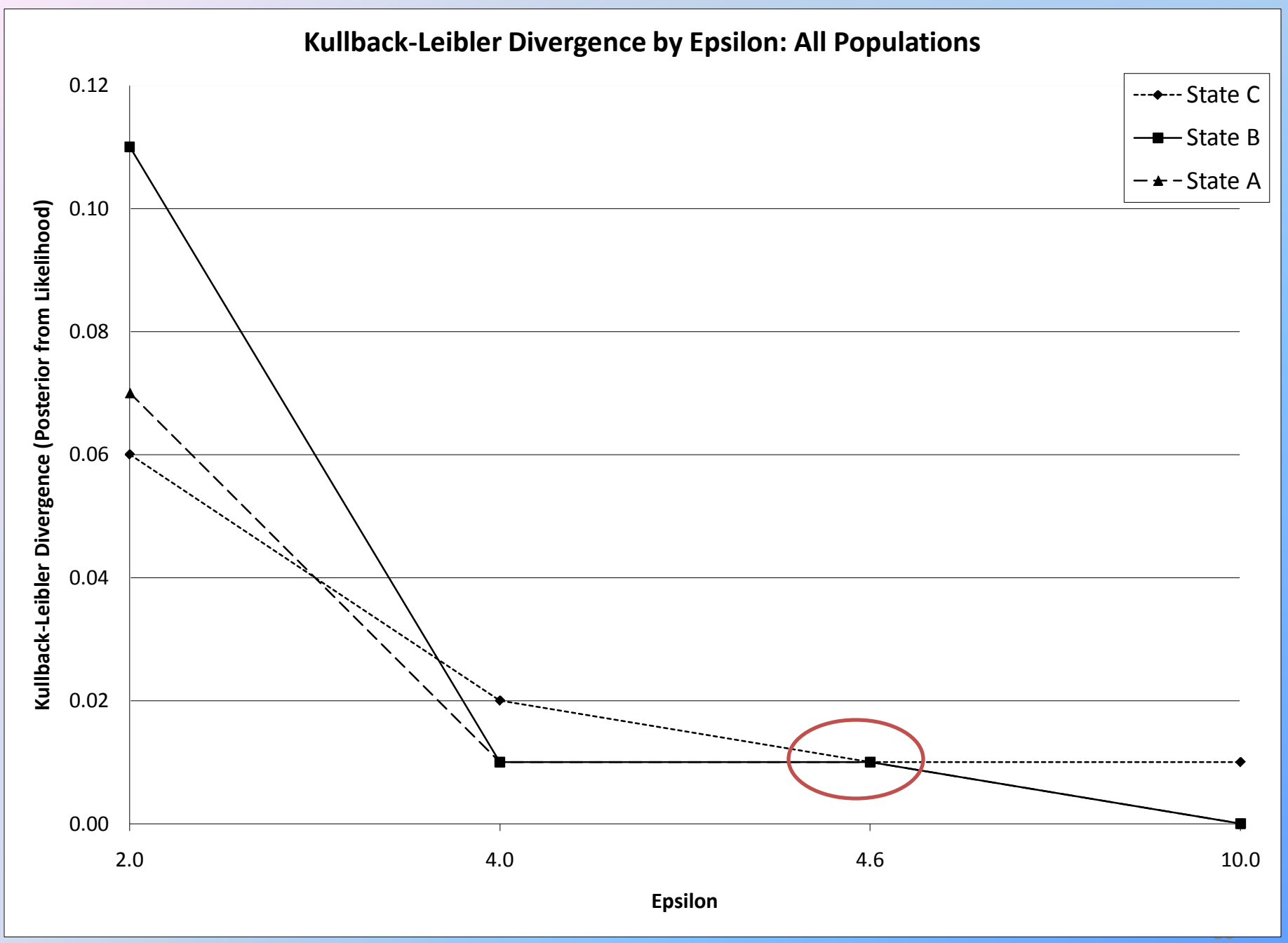
Final Privacy Settings for OnTheMap V3

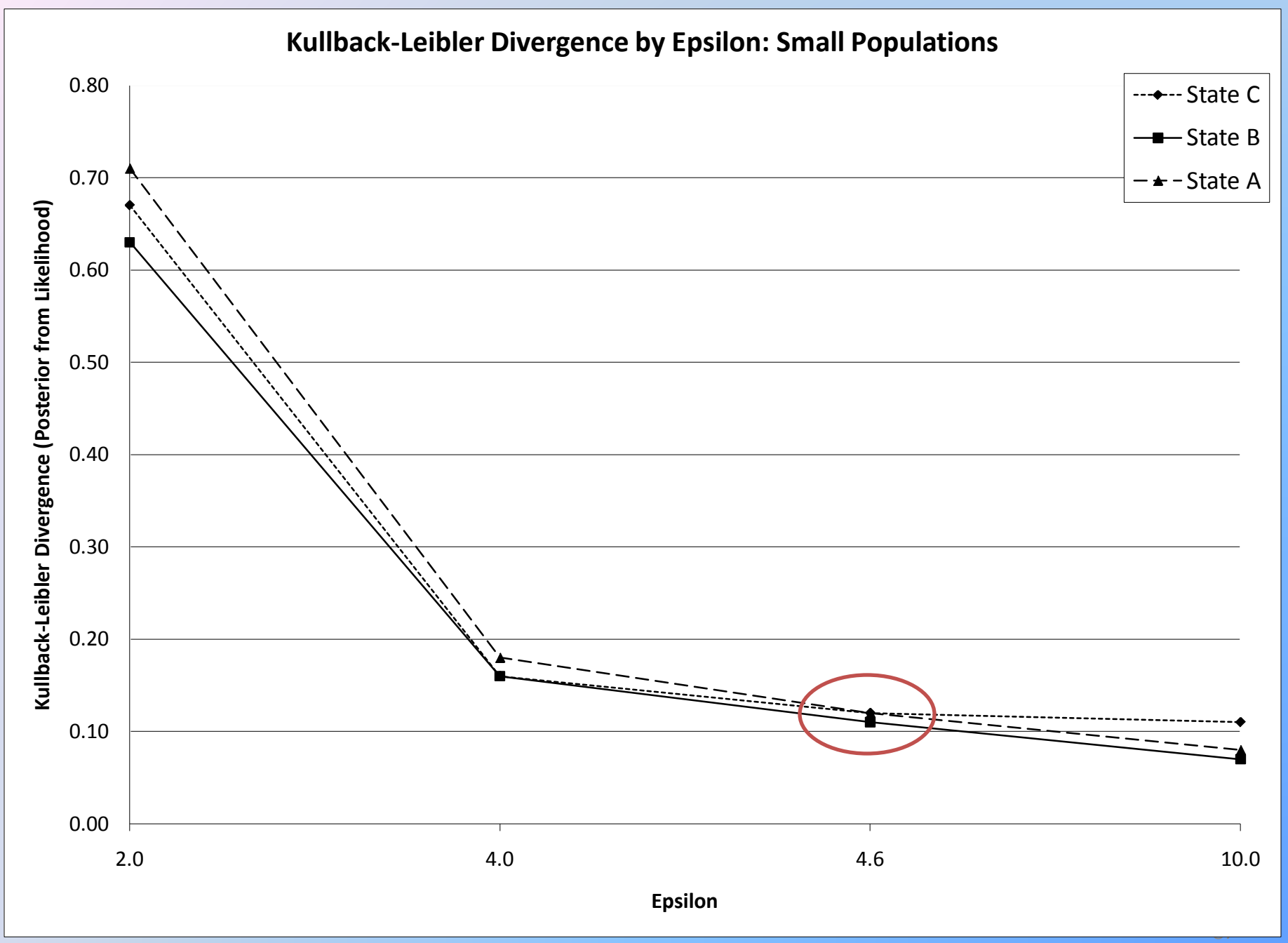
- Unadjusted $\varepsilon = 4.6$
- Probability of failure $\delta = 0.000001$
- Minimum retention probability *min_p*= 0.025
- Adjusted $\varepsilon = 8.9$
- Kullback-Leibler and Integrated Mean Squared Error loss functions used to set parameters of prior
- Multinomial-Dirichlet Posterior sampled for every workplace block in the U.S. (about 1.4 million)

Analytical Validity Measures

- The divergence between posterior and likelihood for a population is measured by the Kullback-Leibler Divergence index (KL) and the Integrated Mean Square Error (IMSE) over a 29 point grid defined by the cross product of:
  - 8 commute distance categories (in miles: 0, (0-1), [1-4), [4-10), [10-25), [25-100), [100,500), [500+]
  - 5 commute direction categories (NW, NE, SW, SE, "N/A")
- D_{KL} = 0 if identical; D_{KL} = ∞ if no overlap

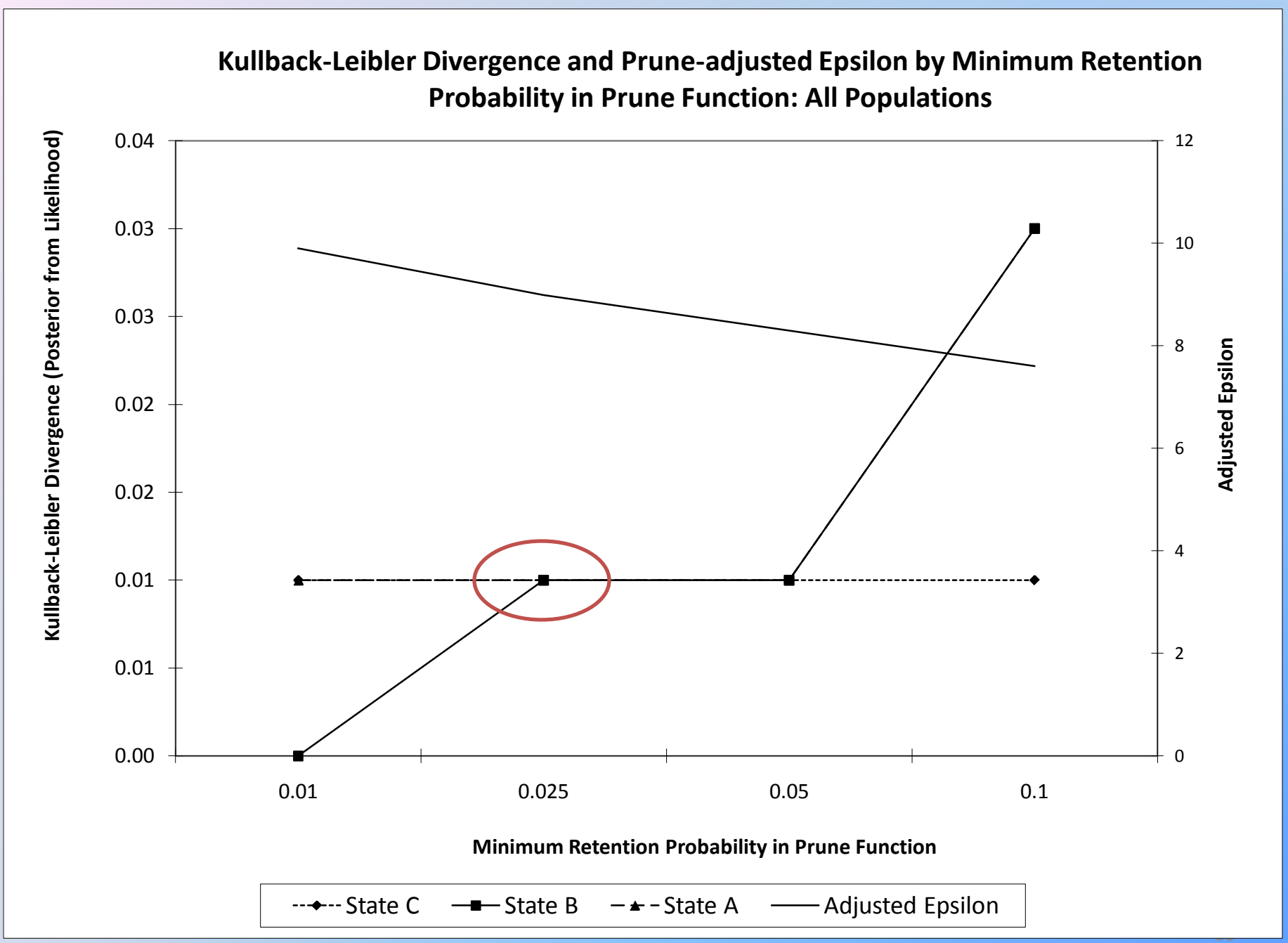
$$D_{KL}(P \parallel L) = \sum_i L(i) \ln \frac{P(i)}{L(i)}$$

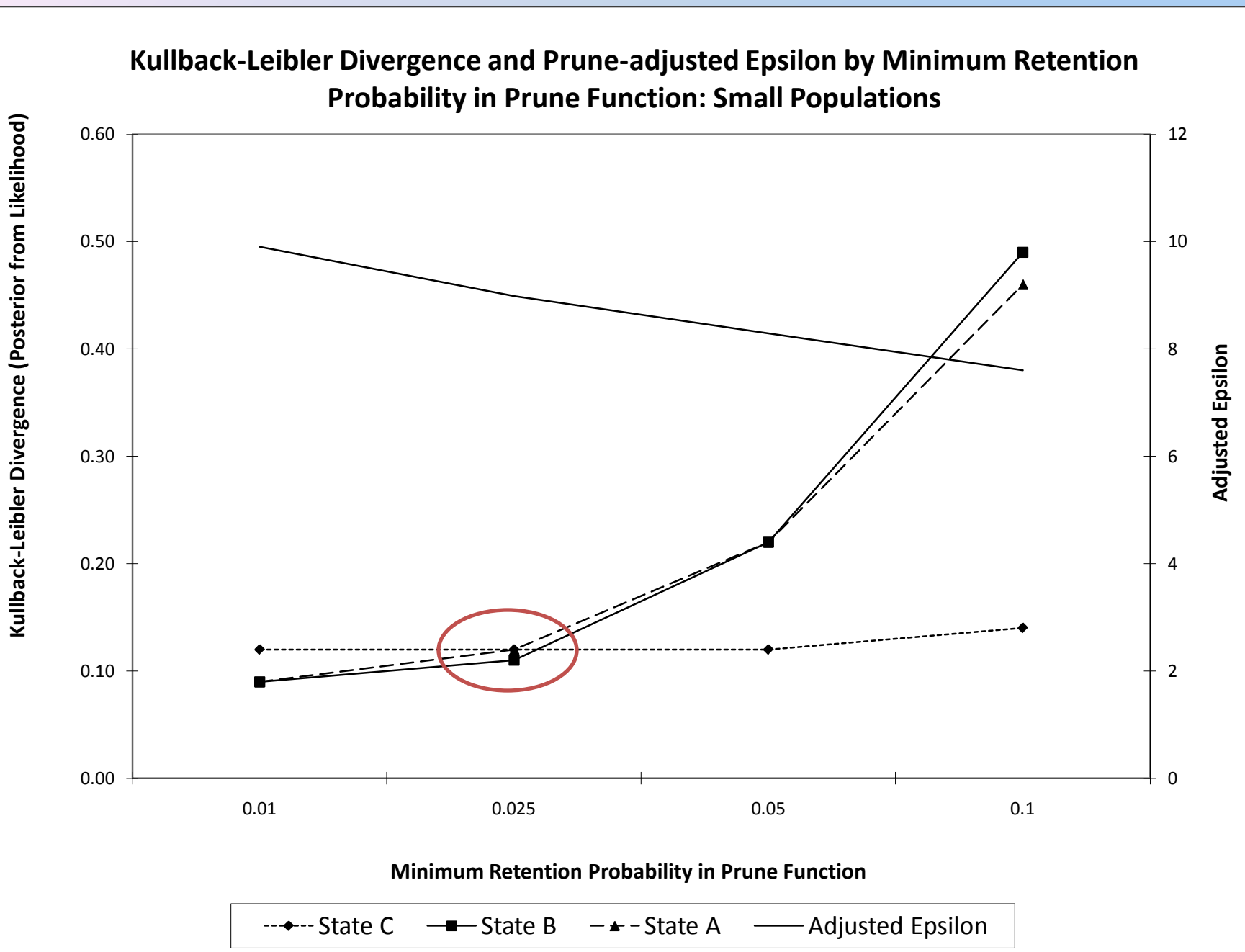

Kullback-Leibler Divergence by Epsilon: All Populations
Kullback-Leibler Divergence (Posterior from Likelihood)
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2.0
4.0
4.6
10.0
Epsilon
State C
State B
State A


Kullback-Leibler Divergence by Epsilon: Small Populations
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Kullback-Leibler Divergence (Posterior from Likelihood)
2.0
4.0
4.6
10.0
Epsilon
State C
State B
State A

Summary: Varying ε

- Figures show the population-weighted D_{KL} for all and small (1 to 9) workforce populations for ε = 2, 4, **4.6**, 10 and 25
- Overall, D_{KL} close to zero for values of ε > 4
- Significant gains in analytical validity for small populations as we increase ε further to 4.6
- The marginal improvements in analytical validity from even higher values of ε hard to justify in terms the costs in privacy protection loss


Kullback-Leibler Divergence and Prune-adjusted Epsilon by Minimum Retention Probability in Prune Function: All Populations
Kullback-Leibler Divergence (Posterior from Likelihood)
0.04
0.03
0.03
0.02
0.02
0.01
0.01
0.00
12
10
8
6
4
2
0
Adjusted Epsilon
0.01
0.025
0.05
0.1
Minimum Retention Probability in Prune Function
State C
State B
State A
Adjusted Epsilon


Kullback-Leibler Divergence and Prune-adjusted Epsilon by Minimum Retention Probability in Prune Function: Small Populations
Kullback-Leibler Divergence (Posterior from Likelihood)
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Adjusted Epsilon
12
10
8
6
4
2
0
0.01
0.025
0.05
0.1
Minimum Retention Probability in Prune Function
State C
State B
State A
Adjusted Epsilon

Summary: Varying *min_p*

- Figures show the population-weighted D_{KL} for all and small (1 to 9) workforce populations and ε for *min_p* = 0.1, 0.05, **0.025** and 0.001
- Large gains in analytical validity as *min_p* is decreased from 0.1 to 0.05 for all populations and further large gains for small populations as *min_p* is decreased to 0.025
- The marginal improvements in analytical validity from even lower values of *min_p*; hard to justify in terms the costs in privacy protection loss

Summary: Varying δ

- We evaluate δ = 0.001, 0.0001, 0.00001 and 0.000001
- Only very marginal improvements in analytical validity as we decrease confidence from 1 in a million to 1 in a 100
- No reason to consider values of δ > 0.000001

Posterior, Likelihood and Prior Mass across Commute Ranges for All and for Small Populations

Large State A						
	All			Small (min-10)		
Distance	Post.	Lik.	Prior	Post.	Lik.	Prior
0	0.07	0.07	0.01	0.30	0.32	0.18
(0-1)	0.15	0.15	0.03	0.13	0.16	0.03
[1-4)	0.23	0.23	0.07	0.25	0.27	0.17
[4-10)	0.26	0.26	0.24	0.28	0.27	0.31
[10-25)	0.28	0.28	0.39	0.21	0.22	0.17
[25-100)	0.14	0.13	0.19	0.18	0.16	0.31
[100-500)	0.03	0.03	0.07	0.04	0.03	0.11
[500-high]	0.02	0.02	0.05	0.01	0.00	0.08

Overall Summary

- Synthetic data as an privacy protection algorithm is a promising alternative to traditional disclosure avoidance methods, especially when data representation is sparse
- Hard to quantify degree of disclosure protection – synthetic data methods may leak more information than intended
- OnTheMap version 3 demonstrates the successful implementation of formal privacy guarantees based on the concept of probabilistic ε-differential privacy
- To achieve acceptable analytical validity results with privacy guarantees requires experimentation

References

- Theorems refer to A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. Privacy: From theory to practice on the map, ICDE, 2008.
- Federal Committee on Statistical Methodology, "Report on Statistical Disclosure Limitation Methodology," Working paper 22 (revised December 2005).
- Evfimievski, A., J. Gehrke, and R. Srikant. "Limiting privacy breaches in privacy-preserving data mining," PODS 2003.
- Kooiman, P., Willenborg, L.C.R.J. and Gouweleeuw, J.M., "PRAM: a method for disclosure limitation of microdata," Research paper no. 9705, Statistics Netherlands, (1997).
- Dalenius, T. "Towards a methodology for statistical disclosure control," *Statistik Tidskrift* (Statistical Review) (1977): 429-44.
- Chawla, S., C. Dwork F. McSherry, A. Smith, and H. Wee, "Towards privacy in public databases," in Proceedings of the 2nd Theory of Cryptography Conference (2005).
- Abowd, J. and L. Vilhuber, "How Protective are Synthetic Data," in J. Domingo-Ferrer and Y. Saygun, eds., Privacy in Statistical Databases, 2008" (Berlin: Springer-Verlag, 2008), pp. 239-246.
- Dwork, C "Differential Privacy," 33rd International Colloquium on Automata, Languages, and Programming—ICALP (2006): Part II, 1-12.