Formal Privacy Guarantees and Analytical Validity of *OnTheMap* Public-Use Data

Presented at NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection, July 31, 2009

Motivation - Problem

- *OnTheMap* provides conditional residence distributions across approximately 8 million blocks - most counts are too low to be released based on any typical cell suppression rules
- Traditional disclosure avoidance methods are not applicable to data that are sparsely distributed across sensitive attributes
- *Synthetic data* is a promising confidentiality protection approach
- Need to quantify degree of confidentiality protection, i.e. how much information leaks from the anonymization algorithms?

This presentation

- Version 3 of *OnTheMap* is based on a *synthetic data* anonymization algorithm that offers formal privacy guarantees in terms of ε-*differential privacy*
- Data developed by staff at the U.S. Census Bureau's *Longitudinal Employer-Household Dynamics Program* (LEHD)

Outline

1. Motivation
2. Overview of *OnTheMap*
3. The Synthetic Data Anonymization algorithm
4. Formal Privacy Guarantees
5. Implementation
6. Tradeoffs between analytical validity and privacy protection
7. Summary

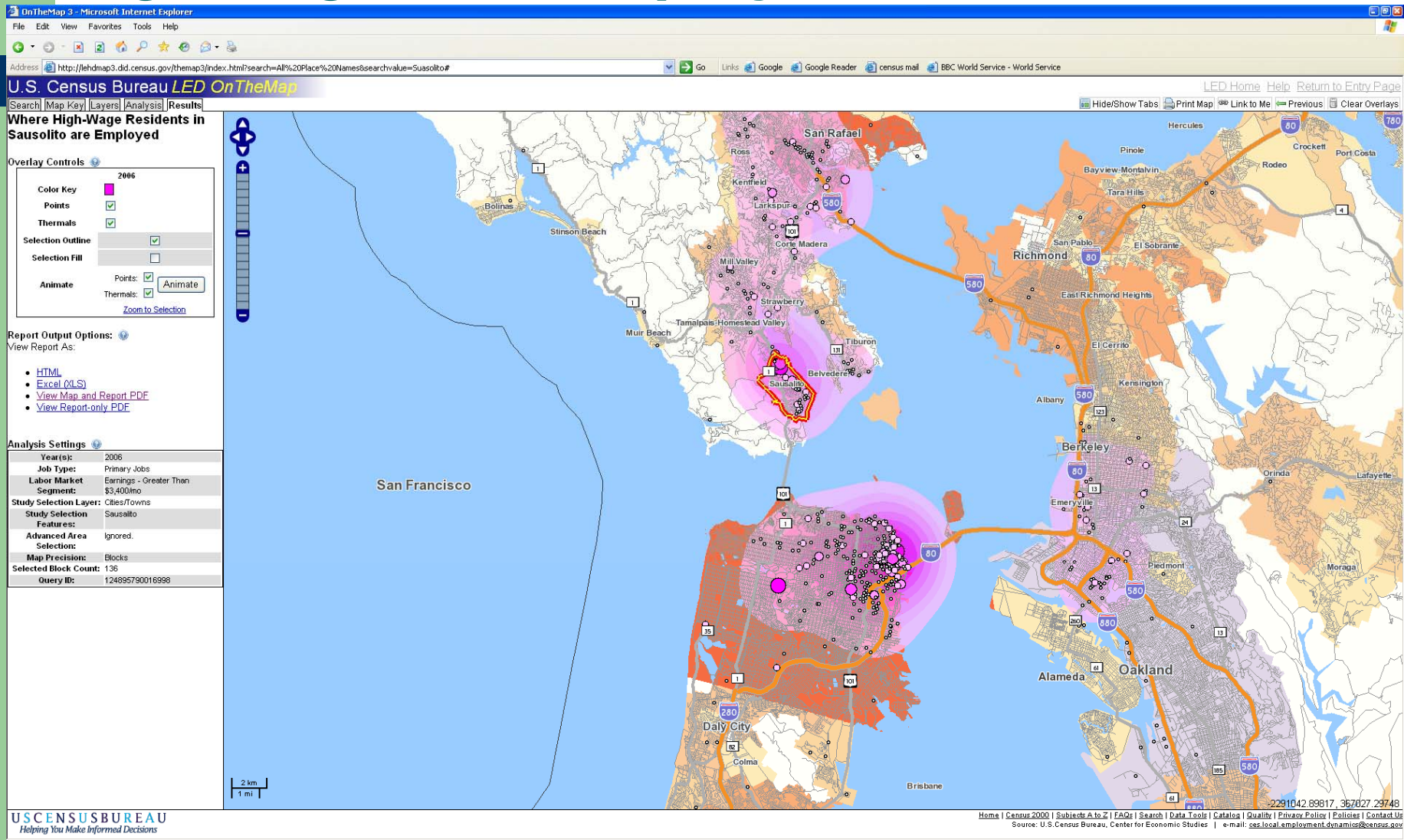
2. Overview of *OnTheMap*

OnTheMap – Overview

- *OnTheMap* is an interactive mapping application that shows in high geographic resolution where people reside and work along with characteristics of home and work areas → Valuable tool for transportation planning, emergency planning, and economic development purposes
- Version 3
  - Released in 2008
  - More data: all 48 states in production, additional years of data (2002-06)
  - Additional features in online application
  - Segmented data: O/D by age, earnings and industry
  - Refined disclosure avoidance methodology with formal privacy guarantee

http://lehdmap3.did.census.gov/themap3

Where residents of Sausolito, CA, with high wages are employed



OnTheMap – Features

- Analysis capabilities include:
  - Selection of work or home area by geographical layers or by freehand
  - Selection of year (2002-2006), of 4 job types (primary jobs vs. all jobs in the private vs. all sectors) and segmentation possibilities by earnings, age or industry groups
- Commute and Labor Shed Maps/Reports
- Area Characteristics Reports
- Block-level QWIs

OnTheMap – Public Use Micro Data

- The micro data that feed the application are available for download (unrestricted access)*
- An observation is a unique Origin Block-Characteristic**-Destination Block combination with information on the job counts
- Origin counts need protection (destination counts are public-use information***)

---

* See http://lehd.dsd.gov for more information about application and access

** A characteristic is defined by a combination of 3 industry groups, 3 earnings and 3 age categories

*** Destination counts that are subject to item suppression and replaced by synthetic values

3. The Synthetic Data Anonymization Algorithm

Modeling Objective

- To maximize analytical validity in terms of:
  1. Completeness of estimates
  2. Preservation of key properties of micro data
- Subject to confidentiality restrictions in terms of:
  1. Conditional (and unconditional) origin block counts that need protection
  2. Public-Use data on destination counts and characteristics

Synthetic Data Model

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

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

- The resulting posteriors of θ is Dirichlet with parameter $y + \alpha$
- Synthetic place of residence counts by sampling from the posterior predictive distributions conditional on already disclosure protected destination population counts, *Yjk*

4. Formal Privacy Guarantees

Epsilon-Differential Privacy

- The anonymization algorithm is said to provide ε-differential privacy if the amount of additional information the attacker can gain concerning an unknown data point is bounded by ε
- ε is the log odds favoring any data set over another generated from the anonymization algorithm that differ in exactly one row
- Note privacy audit is based on posterior transition matrix, not actual randomized data

Differential Privacy - Example

- Population: 10 workers distributed across 3 residence locations
- Consider an attacker that has complete information about:
  - all the data except one observation
  - all aspects of the anonymization algorithm, except for the seeds used in the randomization process

Privacy Audit

	A	B	C	All
Attacker's information	9	?	?	10
Data 1 (true)	9	1	0	10
Data 2 (not true)	9	0	1	10
Prior (known to attacker)	0.1	0.1	0.1	0.3
P[y|Data 1]	0.883	0.107	0.010	1.000
P[y|Data 2]	0.883	0.010	0.107	1.000

$$\max\left[\ln\left[\frac{P[Y_\bullet \mid \text{Data } 1 = \text{True}]}{P[Y_\bullet \mid \text{Data } 2 = \text{True}]}\right]\right] = \ln\left[\frac{0.107}{0.010}\right] \approx 2.4$$

Example of Infinite Differential Privacy

	A	B	C	All
Attacker's Information	9	?	?	10
Data 1	9	1	0	10
Data 2	9	0	1	10
Prior	0.1	0.1	0	0.2
P[y|Data 1= true]	0.892	0.108	0.000	1.000
P[y|Data 2= true]	0.892	0.010	0.008	1.000

$$\max\left[\ln\left[\frac{P[Y_\bullet \mid \text{Data } 2 = \text{True}]}{P[Y_\bullet \mid \text{Data } 1 = \text{True}]}\right]\right] = \ln\left[\frac{0.008}{0.000}\right] = \infty$$

Search algorithm

- Search algorithm to find minimum prior support to guarantee ε-differential privacy developed in Machavajjhala et al. (2008)
- We rely on the concept of (δ,ε)-differential privacy, where the search algorithm guarantees ε-differential privacy with 1- δ confidence
- The anonymization algorithm implemented in *OnTheMap* guarantees ε -differential privacy protection of 8.99 with 99.999999% confidence (δ = 0.000001)

5. Implementation

Main Complication

- Outcome domain has support across approximately 8 million blocks
- To avoid infinite differential privacy each point in the domain has to have minimum prior support

➔ For any model with acceptable formal privacy guarantees this will adversely impact the analytical validity of data

Measures to improve analytical validity

1. Coarsening of the outcome domain
   - Reducing the number of support points in the domain
2. Restricting the outcome space
   - Eliminating the most unlikely commute patterns (from prior and likelihood)
3. Use of informative priors
   - Impose likely shape based on published data (s.t. minimum prior support that will ensure epsilon differential privacy)
4. Pruning the prior
   - Randomly eliminating a fraction support points with no likelihood support.
   - Pruning comes with a penalty in terms of privacy protection

Coarsening of the outcome space

1. If origin block very far away from destination block (distance > 90th pctl of CTTP commute distribution) coarsened to Super-PUMA
2. Else if origin block far away from destination block (distance > 50th pctl of CTTP commute distribution) coarsened to PUMA
3. Else if origin block close to destination block (distance < 50th pctile of CTTP commute distribution) coarsened to Census Tract

Idea: “marginal differences in commute distances between candidate locations have less predictive power in allocating workers the farther away the locations are”

Note: The synthesizer samples from the coarsened domain. Conditional on draw, block is sampled based on Decennial 2000 population estimates

Support points in domain

	State A			State B			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

Fraction of points in the domain with support in CTPP 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

Restricting the outcome space

- For each work tract:
  - if point in domain has zero support 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 do not eliminate
- Note: contribution of any likelihood data in eliminated points also eliminated

5. Fraction of points in the domain with support in CTPP data after eliminating extremely unlikely commute patterns

	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	0.13	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

Informative priors

- In year 2002: Public use CTTP data
- In year 2003-2006: Public use year-1 *OnTheMap* data
- alfa = max[min_alfa, f(prior density)]
- Priors unique to each employment tract

Pruning the prior

- Unpruned prior is defined by alfa = max[alfa_min, f(prior density)]
- Pruned prior defined by alfa*
  - For points with likelihood support alfa* = alfa
  - For points with no likelihood support alfa* = 0 with probability 1-p and alfa* = alfa with probability p, where p=max[alfa,min_p]
- Benchmark min_p = 0.025
- Pruning comes with a cost in terms of privacy protection ➔ ε=g(ε*,min_p), where ε* is the “Nominal epsilon”

6. Tradeoffs between Analytical Validity and Privacy Protection

Benchmark case

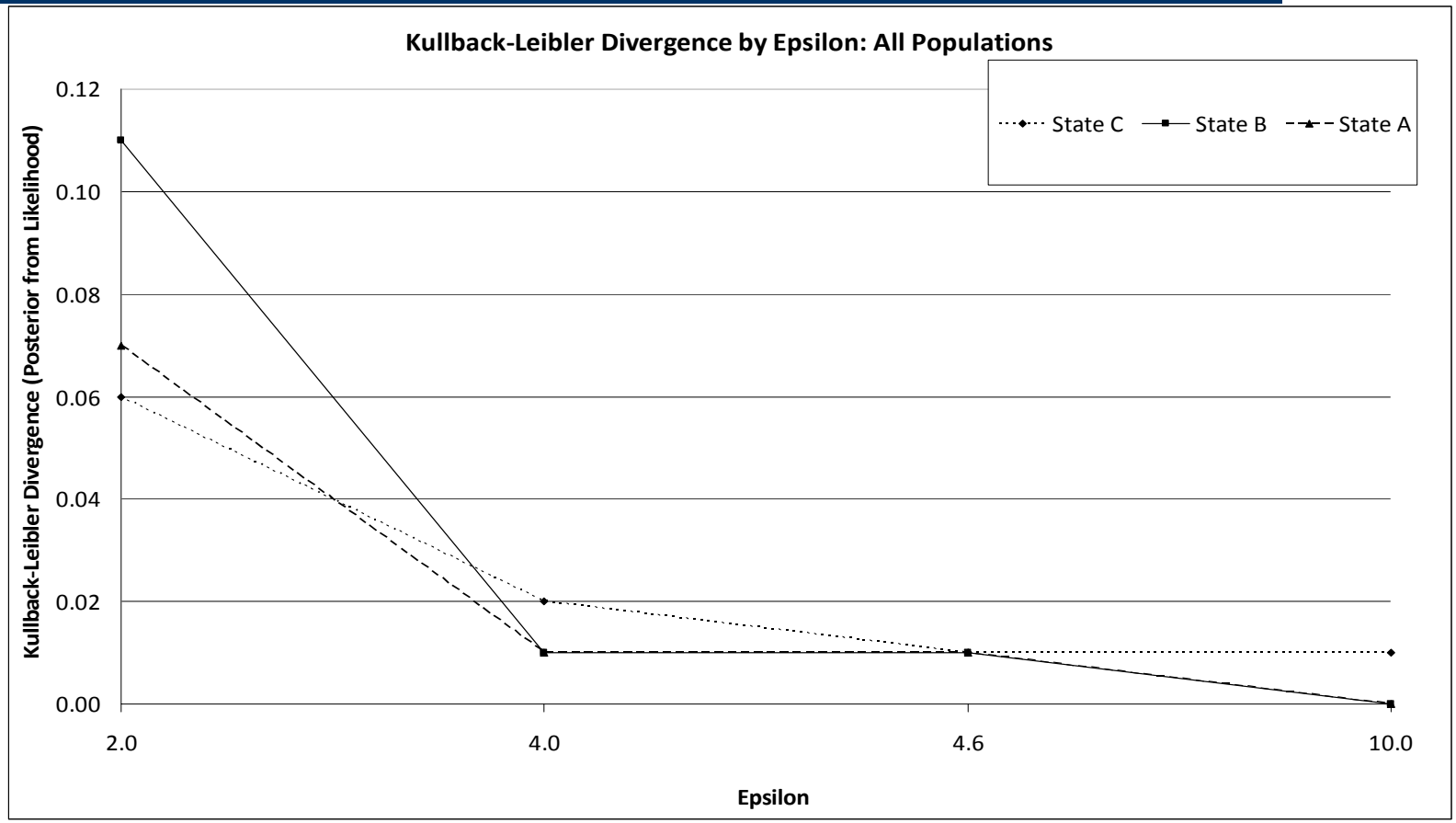
- Choice parameters in the model:
  - Parameters in domain coarsening algorithm
  - Parameters in domain restriction algorithm
  - *Nominal epsilon, delta and the pruning function*
- [ε *,δ, min_p] = [4.6,0.000001,0.025] ➔ ε < 9 in all cases with 99.99999% confidence
- We evaluate effects by changing one parameter at the time around benchmark case

Analytic validity metric

- As metrics for divergence between posterior and likelihood for a population we calculate 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

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

D_{KL} by nominal value of ε*



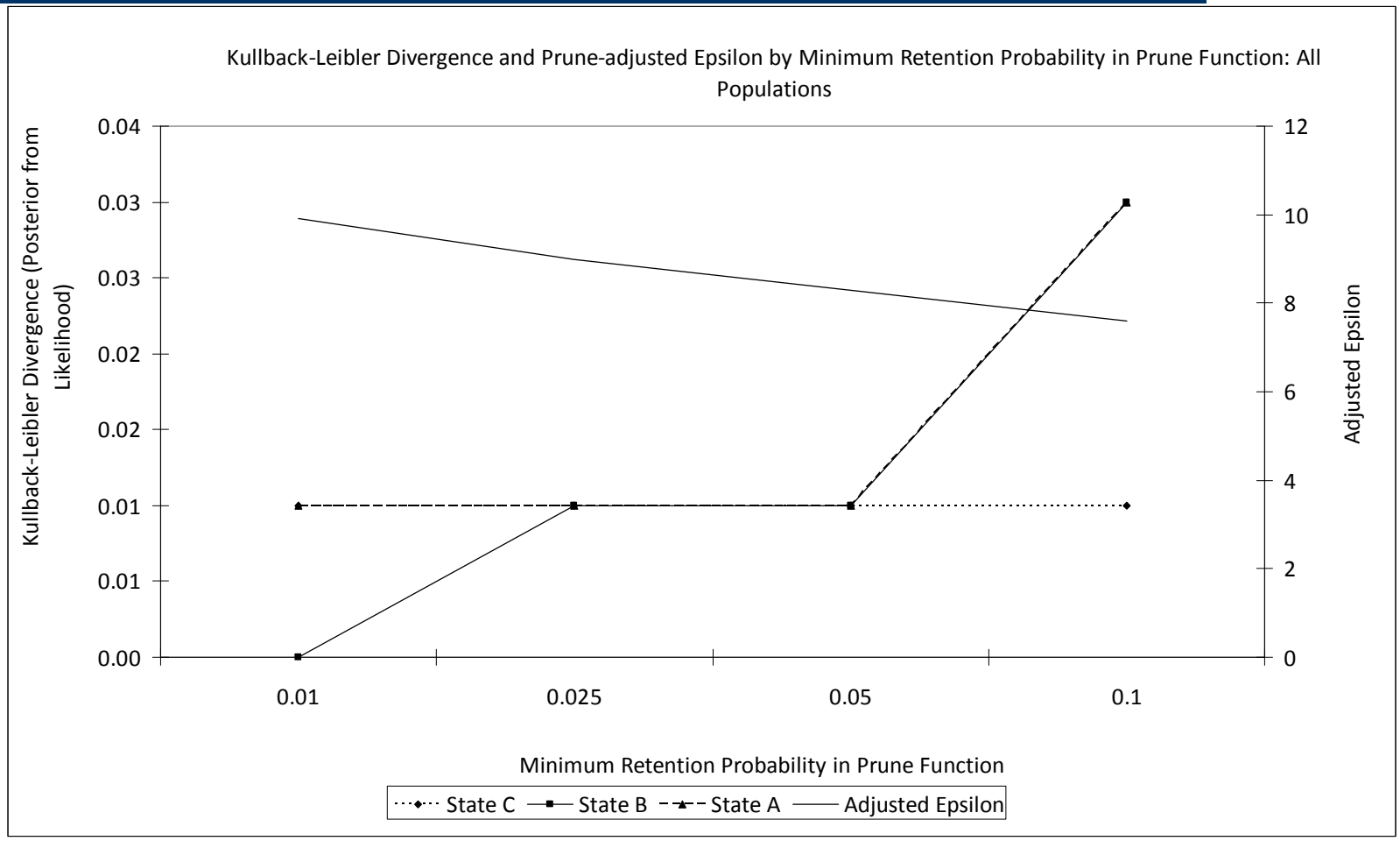
Varying ε* - Summary

- Figures show the population-weighted D_{KL} for all and small (<10) populations for ε* = 2, 4, **4.6**, 10 and 25
- 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

Varying δ - Summary

- 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

D_{KL} and ε vs. paramater, min_p, in pruning function



Varying min_p - Summary

- Figures show the population-weighted D_{KL} for all and small (<10) 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

Posterior, likelihood and prior mass across commute ranges for all and for small populations

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

7. Summary

- Synthetic data as an anyonymization algorithm 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 ε-differential privacy
- To achieve acceptable analytical validity results s.t. privacy guarantees requires experimentation