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

Presented at 3rd IAB Workshop on Confidentiality and Disclosure, November 20-21, 2008

1. This presentation

- *OnTheMap* provides residence distributions across roughly 8 million blocks conditional on workplace and other characteristics
- Traditional disclosure avoidance methods are not applicable to data that are sparsely distributed across sensitive attributes
- Version 3 of *OnTheMap* is based on a *Bayesian synthetic data* anonymization algorithm that offers formal privacy guarantees in terms of ε-*differential privacy*

1. 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

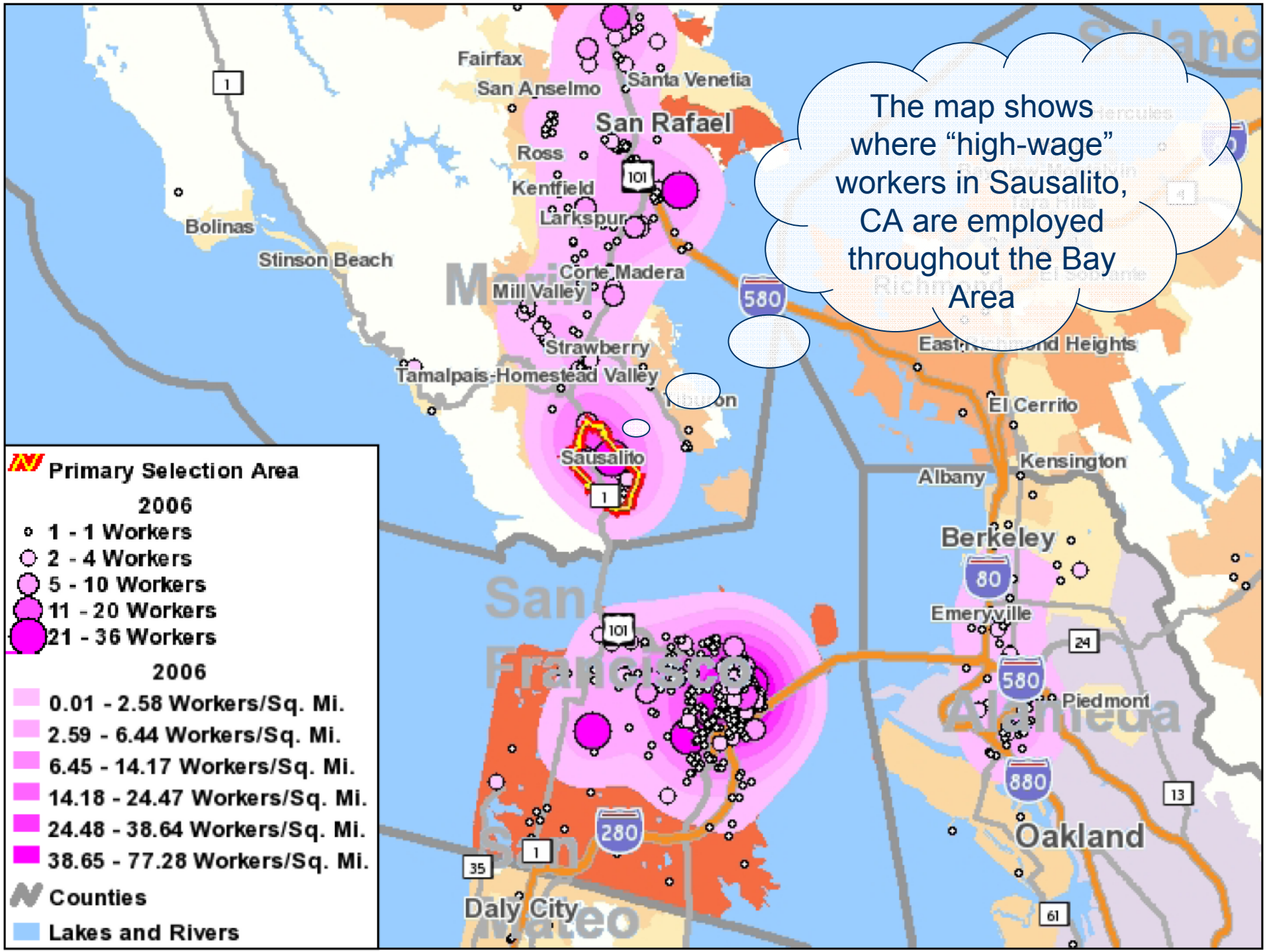
2. Overview of *OnTheMap*

2. *OnTheMap* – Overview of Online Application

- *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

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


The map shows where “high-wage” workers in Sausalito, CA are employed throughout the Bay Area
Primary Selection Area
2006
1 - 1 Workers
2 - 4 Workers
5 - 10 Workers
11 - 20 Workers
21 - 36 Workers
2006
0.01 - 2.58 Workers/Sq. Mi.
2.59 - 6.44 Workers/Sq. Mi.
6.45 - 14.17 Workers/Sq. Mi.
14.18 - 24.47 Workers/Sq. Mi.
24.48 - 38.64 Workers/Sq. Mi.
38.65 - 77.28 Workers/Sq. Mi.
Counties
Lakes and Rivers
Fairfax
San Anselmo
Santa Venetia
San Rafael
Ross
Kentfield
Larkspur
Bolinas
Stinson Beach
Corte Madera
Mill Valley
Strawberry
Tamalpais-Homestead Valley
Sausalito
El Cerrito
Kensington
Albany
Berkeley
Emeryville
Piedmont
Oakland
Daly City
1
101
580
80
24
880
280
35
13
61

2. *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
- Area Characteristics Reports

2. *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 4 different job counts,

  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

2. *OnTheMap* – A Brief History

- Beta/proof of concept version (~2004)
  - Limited scope: one state and one year of data (2004)
  - Limited geographical precision: block group
  - Limited usefulness: cell suppression used to protect confidentiality
- Version 1 (2006, First official release)
  - 11 states, additional years of data (2002-2003)
  - Online application
  - First ever synthetic data product released by a statistical agency

2. *OnTheMap* – A Brief History (cont.)

- Version 2 (2007)
  - 42 states, additional years of data (2002-04)
  - Increased geographical precision: block
  - Additional features in online application
  - Replicates of synthetic data available from Cornell virtual research data center
- Version 3 (2008)
  - 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, including formal privacy guarantees

3. The Synthetic Data Anonymization Algorithm

3. Modeling Objective

- Goal is to create residence distributions across blocks for each population defined by employment block and other characteristics
- To maximize analytical validity in terms of:
  1. Completeness of estimates, i.e. no suppression
  2. Preservation of key properties of micro data, e.g. commute patterns
- Subject to confidentiality restrictions in terms of residence block information that needs protection

3. Synthetic Data Model

- Bayesian approach: Each population has a likelihood and prior that describes the conditional residence distribution
- Synthetic data sampled from the posterior predictive distribution that combines information from likelihood and prior
- Size of synthetic population based on disclosable estimates
- Likelihood from confidential micro data
- Design of prior is the result of combining information from public use data with restrictions to ensure formal privacy definition

4. Formal Privacy Guarantees

4. Epsilon-Differential Privacy

- Privacy audit is based on posterior transition matrix
- For any two potential confidential data sets generated from the anonymization algorithm that differ in exactly one row, if the probability that the log odds favoring one data set over another is less than ε the algorithm is said to provide ε-differential privacy

4. 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

4. Privacy Audit

	A	B	C	All
Attacker's information	9	?	?	10
Data 1	9	1	0	10
Data 2	9	0	1	10
Prior (known to attacker)	0.1	0.1	0.1	0.3
P[y|Data 1= true]	0.883	0.107	0.010	1.000
P[y|Data 2= true]	0.883	0.010	0.107	1.000

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

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

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

4. Search algorithm

- 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
- In particular, our anonymization algorithm guarantees ε-differential privacy protection of 8.99 with 99.999999% confidence (δ = 0.000001)

5. Implementation

5. Main Complication

Dimensionality of problem

- Outcome domain has support across approximately 8 million blocks
- Recall each point in the domain has to have minimum support in prior

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

5. Measures to improve analytical validity

- To maximize analytical validity s.t. privacy guarantee:
  - Use of informative priors
  - Coarsening of the outcome domain
  - Restricting the outcome space
  - Pruning the prior

5. Informative priors

- In year 2002: Public use CTPP data
- In year 2003-2006: Public use year-1 *OnTheMap* data
- alfa = max[min_alfa, f(prior density)]

5. Coarsening of the outcome space

"The marginal difference in commute distances of two candidate locations has less predictive power in allocating workers the farther away the locations are"

- For each work tract:
  - if distance > 90th pctile of the CTPP commmute distribution, then collapse into Super-PUMAs
  - else if distance >50th pctile, then PUMAs
  - else Census Tracts

5. **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

5. 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

5. Restricting the outcome space

- For each work tract:
  - if point in domain has zero support in prior data the 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

5. Pruning the prior

- For each likelihood:
  - Do for each support point in domain:
    - Let alfa=max[alfa_min, prior density]
    - Set alfa*=0 with probability 1-p, where p=max[alfa,min_p], e.g. a high value of min_p will result in little pruning and vice versa
    - else alfa*=alfa
    - Calculate posterior data as y+alfa*
- Pruning comes with a cost in terms of privacy protection ➔ ε=f(ε*,min_p), where ε* is the “Nominal epsilon”
- Note: pruning does not depend on likelihood data

5. Creation of synthetic data

1. Sample support point in outcome domain from posterior
2. Conditional on coarsened geographical area, sample block-level residence locations based on 2000 Decennial block-level population estimates

6. Tradeoffs between Analytical Validity and Privacy Protection

6. Benchmark case

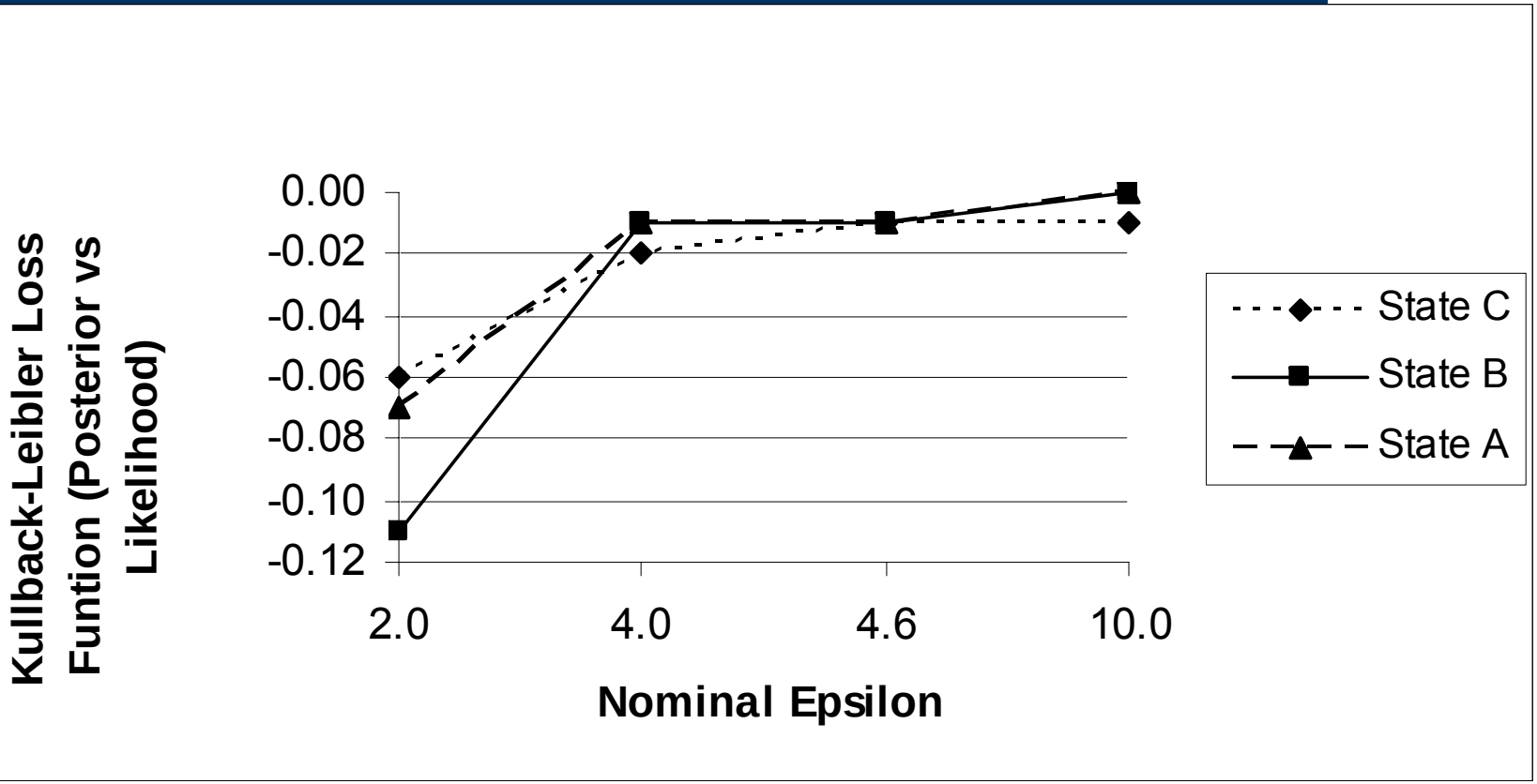
- Choice parameters in the model:
  - Parameters in domain coarsening algorithm
  - Parameters in domain restriction algorithm
  - *Nominal epsilon, delta and the pruning function*
- [ε *,δ, min_p] = [4.6,0.000001,0.025] ➔ ε < 9 in all cases with 99.99999% confidence
- We evaluate effects by changing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

6. Analytic validity metric

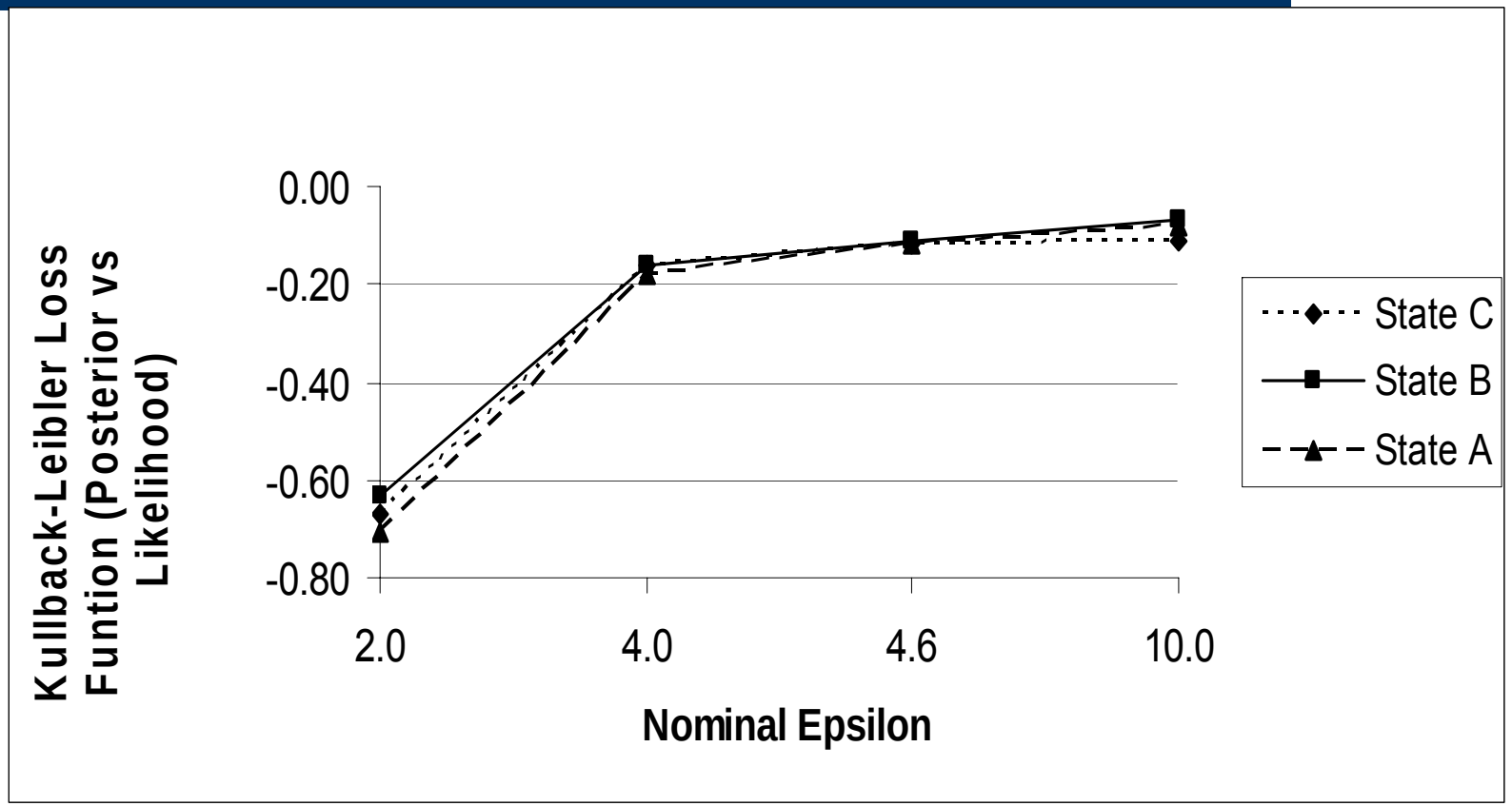
- As a metric for divergence between posterior and likelihood for a population we calculate the *Kullback-Leibler* Divergence index over a 29 point grid defined by the cross product of:
  - 8 commute distance categories (in miles: 0, (0-1), [1-4), [4-10), [10-25), [25-100), [100,500), [500+]
  - 5 commute direction categories (NW, NE, SW, SE, “N/A”)
- D_{KL}= 0 if identical; D_{KL}=-∞ if no overlap

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

6. D_{KL} by ε*: All Populations



6. D_{KL} by ε*: Small Populations



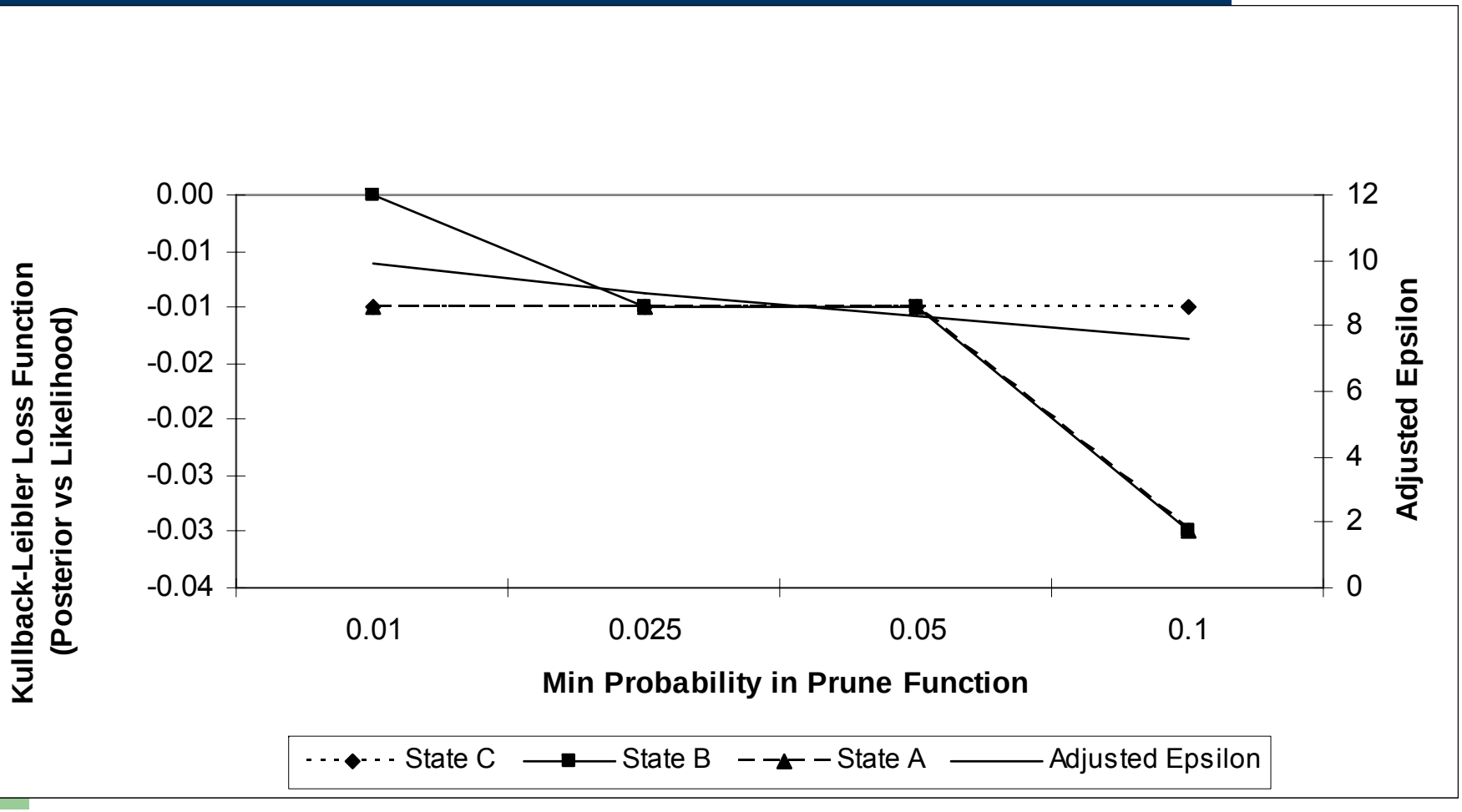
6. Varying ε* - Summary

- 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

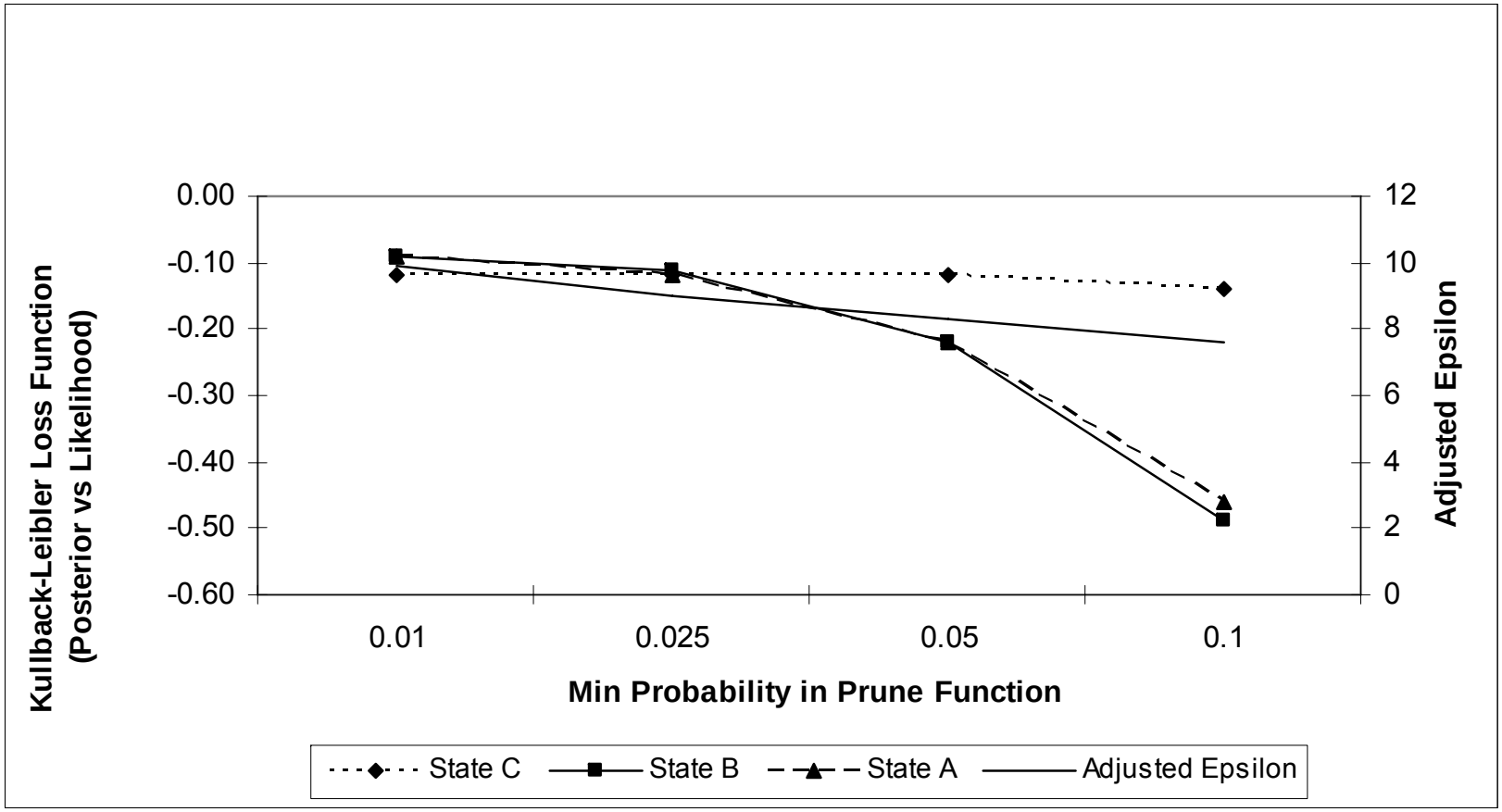
6. Varying δ - Summary

- 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 1000.
- No reason to consider values of δ > 0.000001

6. D_{KL} and ε vs. min_p: All populations



6. D_{KL} and ε vs. min_p: Small populations



6. 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

7. Summary

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