Synthetic data in US Census

CompSci 590.03
Instructor: Ashwin Machanavajjhala

Synthetic Data

- Rather than have a system for performing query answering, release a synthetic dataset.
- Analyst can now perform arbitrary analysis on this synthetic dataset.
- Very popular amongst statisticians
  - For ensuring privacy
  - For imputing (filling in) missing values

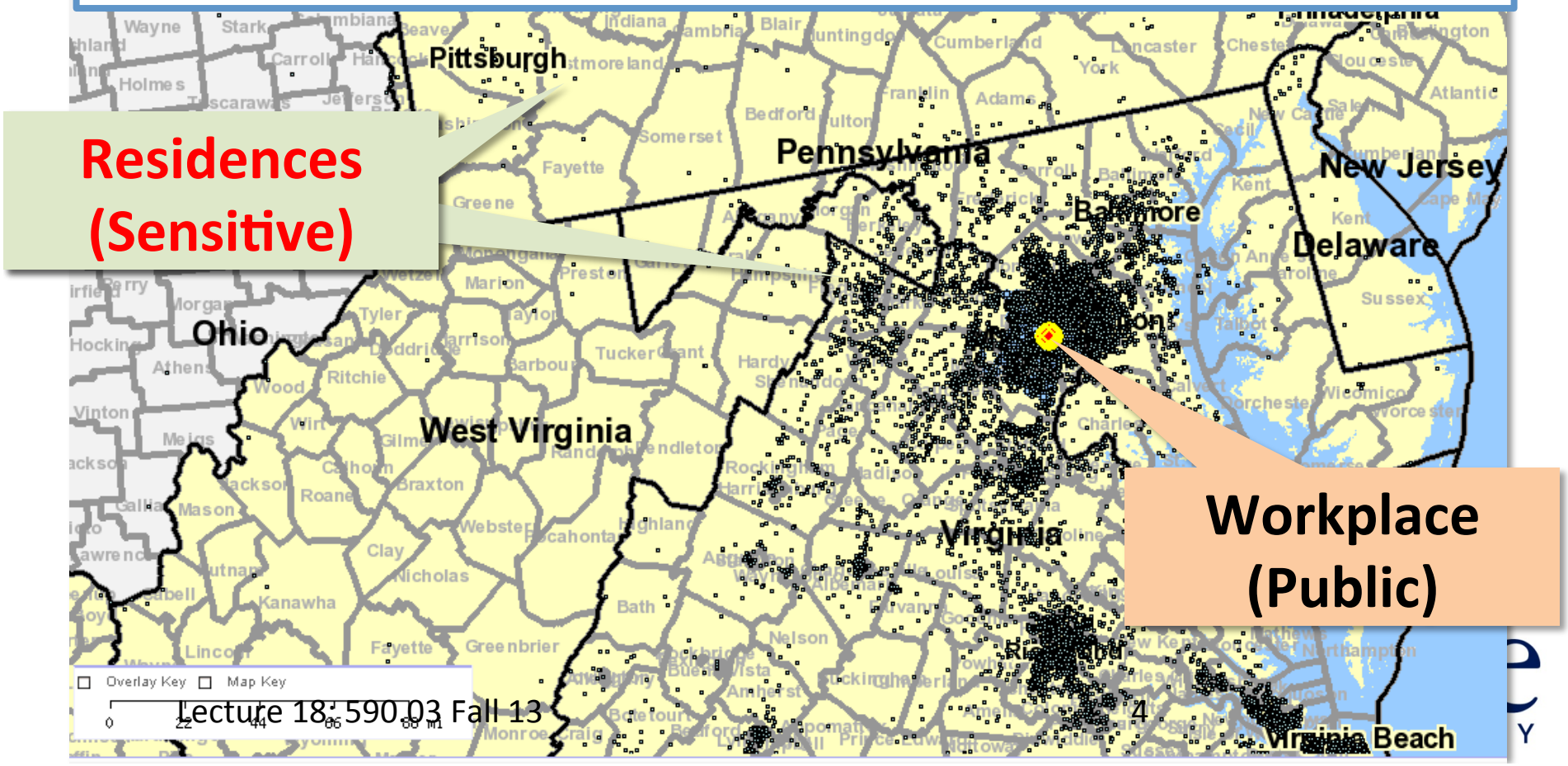
Duke UNIVERSITY

This Class

- Synthetic Data in the US Census
- Synthetic Data: What can we achieve?

OnTheMap: A Census application that plots commuting patterns of workers

http://onthemap.ces.census.gov/



OnTheMap: A Census application that plots commuting patterns of workers

Residence (Sensitive)

Workplace (Quasi-identifier)

Worker ID	Origin	Destination
1223	MD11511	DC22122
1332	MD2123	DC22122
1432	VA11211	DC22122
2345	PA12121	DC24132
1432	PA11122	DC24132
1665	MD1121	DC24132
1244	DC22122	DC22122

Census Blocks

Duke UNIVERSITY

Why publish commute patterns?

- To compute Quarterly Workforce Indicators
  - Total employment
  - Average Earnings
  - New Hires & Separations
  - Unemployment Statistics

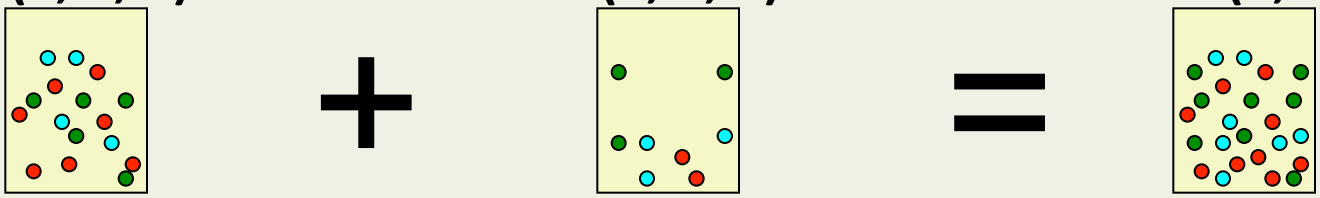
E.g., Missouri state used this data to formulate a method allowing **QWI to suggest industrial sectors where transitional training might be most effective** … to proactively reduce time spent on unemployment insurance …

Duke UNIVERSITY

A Synthetic Data Generator (Dirichlet resampling)

Step 1: **Noise Addition** *(for each destination)*

D (7, 5, 4) + **A (2, 3, 3)** = **D+A (9, 8, 7)**



Multi-set of Origins for workers in Washington DC.

Noise (fake workers)

Noise infused data

Washington DC **Somerset** **Fuller**

Noise added to an origin with at least 1 worker is > 0

Duke UNIVERSITY

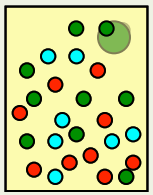
A Synthetic Data Generator (Dirichlet resampling)

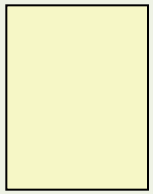
Step 2: **Dirichlet Resampling** *(for each destination)*

Replace two of the same kind.

Draw a point at random

(9, 8, 7)





S : Synthetic Data

frequency of block *b* in D+A = 0 ➔ frequency of *b* in S = 0
i.e., block *b* is ignored by the algorithm.

How should we add noise?

- Intuitively, more noise yields more privacy …

- How much noise should we add ?
- To which blocks should we add noise?

- Currently this is poorly understood.
  - **Total amount of noise added is a state secret**
  - **Only 3-4 people in the US know this value in the current implementation of OnTheMap.**

Privacy of Synthetic Data

Theorem 1:

The Dirichlet resampling algorithm preserves privacy if and only if for every destination *d*, the noise added to each block is at least

$$\frac{m(d)}{\varepsilon - 1}$$

where *m(d)* is the size of the synthetic population for destination *d* and *ε* is the privacy parameter.

Duke UNIVERSITY

1. How much noise should we add?

Noise required per block: (differential privacy)

1 million original and synthetic workers.

lesser privacy →

Privacy (e^{ε} =)	5	10	20	50
Noise per block (x 10^6)	0.25	0.11	0.05	0.02

2. To which blocks should we add noise?

Add noise to every block on the map.

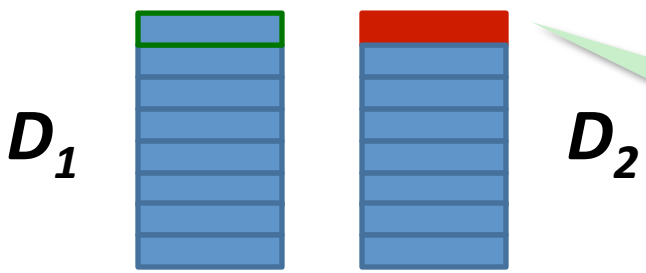
There are 8 million Census blocks on the map!

1 million original workers and 16 billion fake workers!!!

Duke UNIVERSITY

Intuition behind Theorem 1.

Two possible inputs



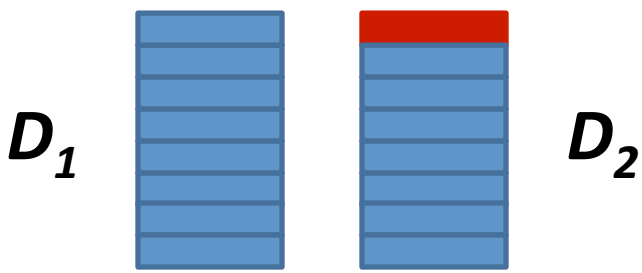
Adversary knows individual 1 is **Either blue or red**.

Adversary knows individuals [2..n] are **blue**.

blue and red are two different origin blocks.

Intuition behind Theorem 1.

Two possible inputs



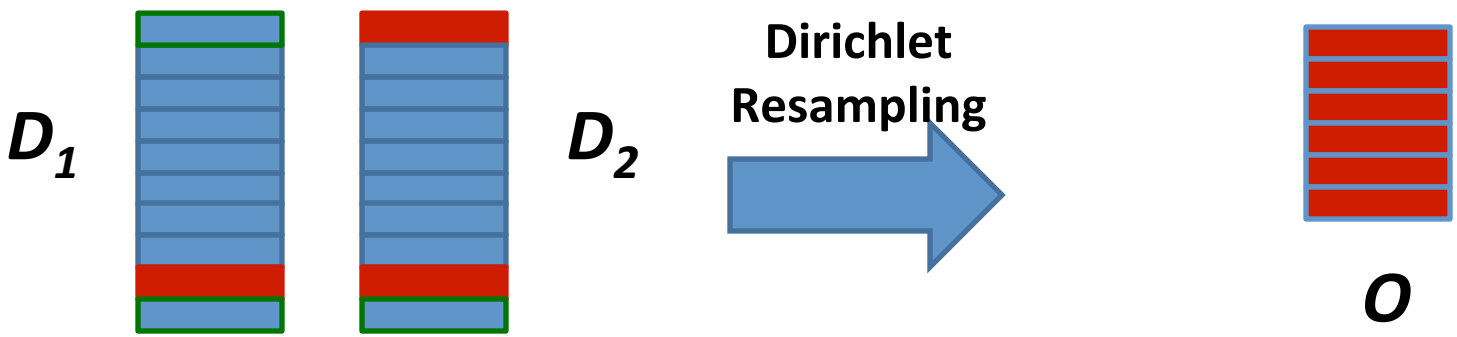
Noise Addition

blue and red are two different origin blocks.

Intuition behind Theorem 1.

Noise infused inputs

For every output …



$\Pr[D_1 \rightarrow O] = 1/10 * 2/11 * 3/12 * 4/13 * 5/14 * 6/15$

$\Pr[D_2 \rightarrow O] = 2/10 * 3/11 * 4/12 * 5/13 * 6/14 * 7/15$

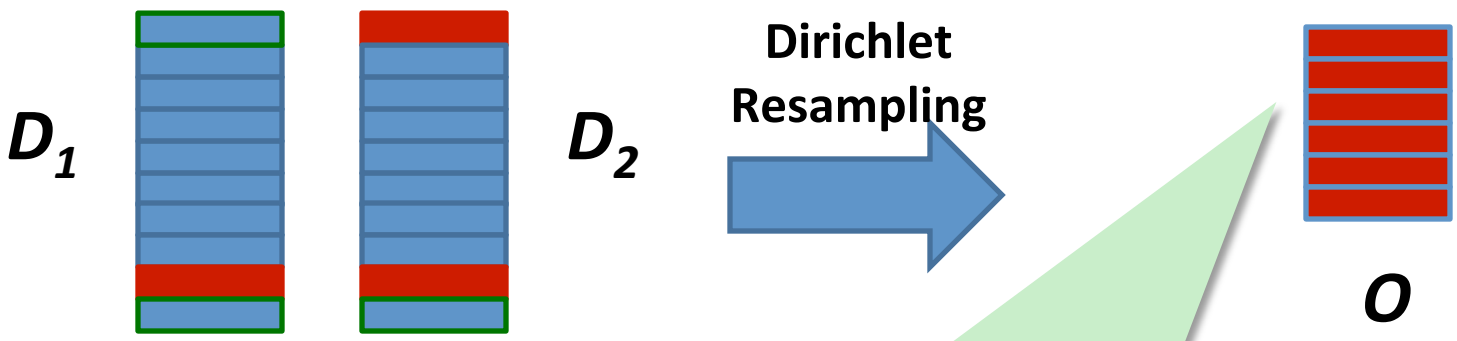
$$\frac{\Pr[D_1 \rightarrow O]}{\Pr[D_2 \rightarrow O]} = 7$$

blue and red are two different origin blocks.

Intuition behind Theorem 1.

Noise infused inputs

For every output …



Adversary infers that it is very likely individual 1 is red …
… unless noise added is very large.

blue and red are two different origin blocks.

Privacy Analysis: Summary

- Chose differential privacy.
  - Guards against powerful adversaries.
  - Measures privacy as a distance between prior and posterior.

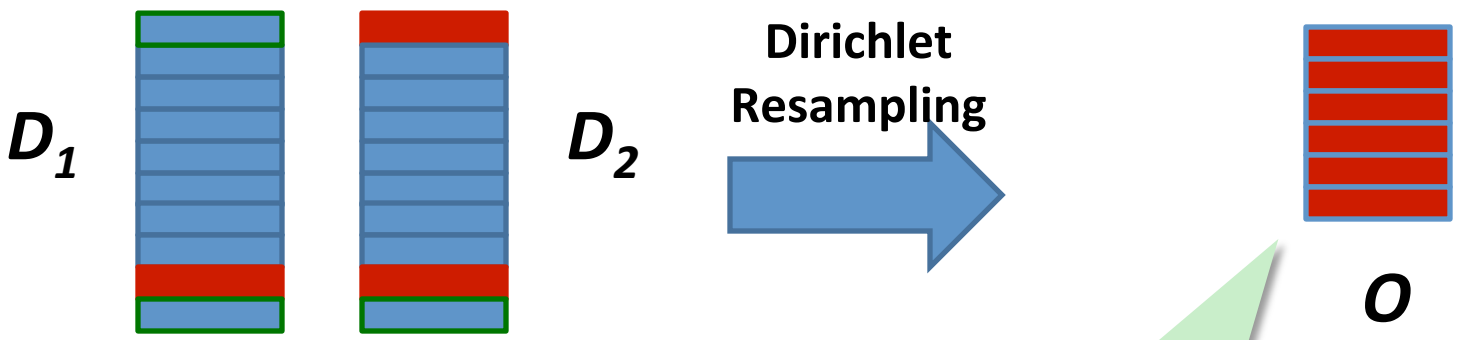
- Derived necessary and sufficient conditions when OnTheMap preserves privacy.

- The above conditions make the data published by OnTheMap useless.

But, breach occurs with very low probability.

Noise infused inputs

For every output …



Probability of O ≈ 10^{-4}

blue and red are two different origin blocks.

Negligible function

Definition:

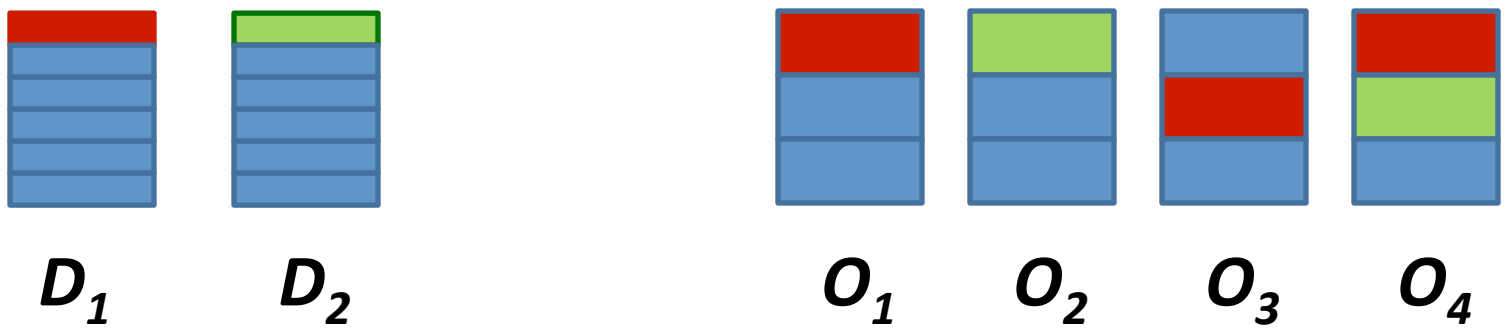
f(x) is *negligible* if it goes to 0 faster than the inverse of any polynomial.

e.g., 2^{-x} *and* e^{-x} *are negligible functions.*

(ε,δ)-Indistinguishability

For every pair of inputs that differ in one value

For any subset of outputs *T*



$$\Pr[D_1 \rightarrow T] \leq e^{\varepsilon} \Pr[D_2 \rightarrow T] + \delta(|D_2|)$$

If *T* occurs with negligible probability, the adversary is allowed to distinguish between D1 and D2 by a factor > ε using O_i in *T*.

Conditions for (ε,δ)-Indistinguishability

Theorem 2:

The Dirichlet resampling algorithm preserves (ε,δ)-indistinguishability if for every destination *d*, the noise added to each block is at least

$$\log n(d)$$

where
n(d) is the number of workers commuting to *d*
and $m(d) \leq n(d)$.

Probabilistic Differential Privacy

- (ε,δ)-Indistinguishability is an asymptotic measure
  - May not guarantee privacy when number of workers at a destination is small.

Definition (Disclosure Set *Disc(D, ε)*):
The set of output tables that breach ε-differential privacy for D and some other table D' that differs from D in one value.

Duke UNIVERSITY

Probabilistic Differential Privacy

For every pair of inputs that differ in one value

For every ***probable*** output

D_1 D_2 O

Adversary may distinguish between D_1 and D_2 based on a set of unlikely outputs with probability at most δ

$$\Pr\left[O \mid \frac{\Pr[D_1 \rightarrow O]}{\Pr[D_2 \rightarrow O]} < e^{\varepsilon}\right] > 1 - \delta$$

1. How much noise should we add?

Noise required per block:

lesser privacy →

Privacy (e^{ε} =)	5	10	20	50
Noise	25×10^4	11×10^4	5×10^4	2×10^4
Noise	17.5	5.5	2.16	0.74

← **Differential Privacy**

← **Probabilistic Differential Privacy ($\delta = 10^{-5}$)**

1 million original and synthetic workers.

Duke UNIVERSITY

Prob. Differential Privacy: Summary

- Ignoring privacy breaches that occur due to low probability outputs drastically reduces noise.

- Two ways to bound low probability outputs
  - (ε,δ)-Indistinguishability and Negligible functions.
    Noise required for privacy ≥ (log n(d)) per block
  - (ε,δ)-Probabilistic differential privacy and Disclosure sets.
    Efficient algorithm to calculate noise per block (see paper).

- Does probabilistic differential privacy allow useful information to be published?

Duke UNIVERSITY

1. How much noise should we add?

Noise required per block:

lesser privacy →

Privacy (e^{ε} =)	5	10	20	50
Noise	25×10^4	11×10^4	5×10^4	2×10^4
Noise	17.5	5.5	2.16	0.74

← **Differential Privacy**

← **Probabilistic Differential Privacy ($\delta = 10^{-5}$)**

1 million original and synthetic workers.

2. To which blocks should we add noise?

Why not add noise to every block?

Duke UNIVERSITY

Why not add noise to every block?

Noise required per block: (probabilistic differential privacy)

1 million original and synthetic workers.

lesser privacy →

Privacy (ε =)	5	10	20	50
Noise per block	17.5	5.5	2.16	0.74

- There are about **8 million** blocks on the map!
  - Total noise added is about **6 million.**
- Causes non-trivial spurious commute patterns.
  - Roughly 1 million fake workers from West Coast (out of a total 7 million points in D+A).
  - Hence, 1/7 of the synthetic data have residences in West Coast and work in Washington DC.

Duke UNIVERSITY

2. To which blocks should we add noise?

Noise required per block: (probabilistic differential privacy)

1 million original and synthetic workers.

lesser privacy →

Privacy (ε =)	5	10	20	50
Noise per block	17.5	5.5	2.16	0.74

Adding noise to all blocks creates spurious commute patterns.

Why not add noise only to blocks that appear in the original data?

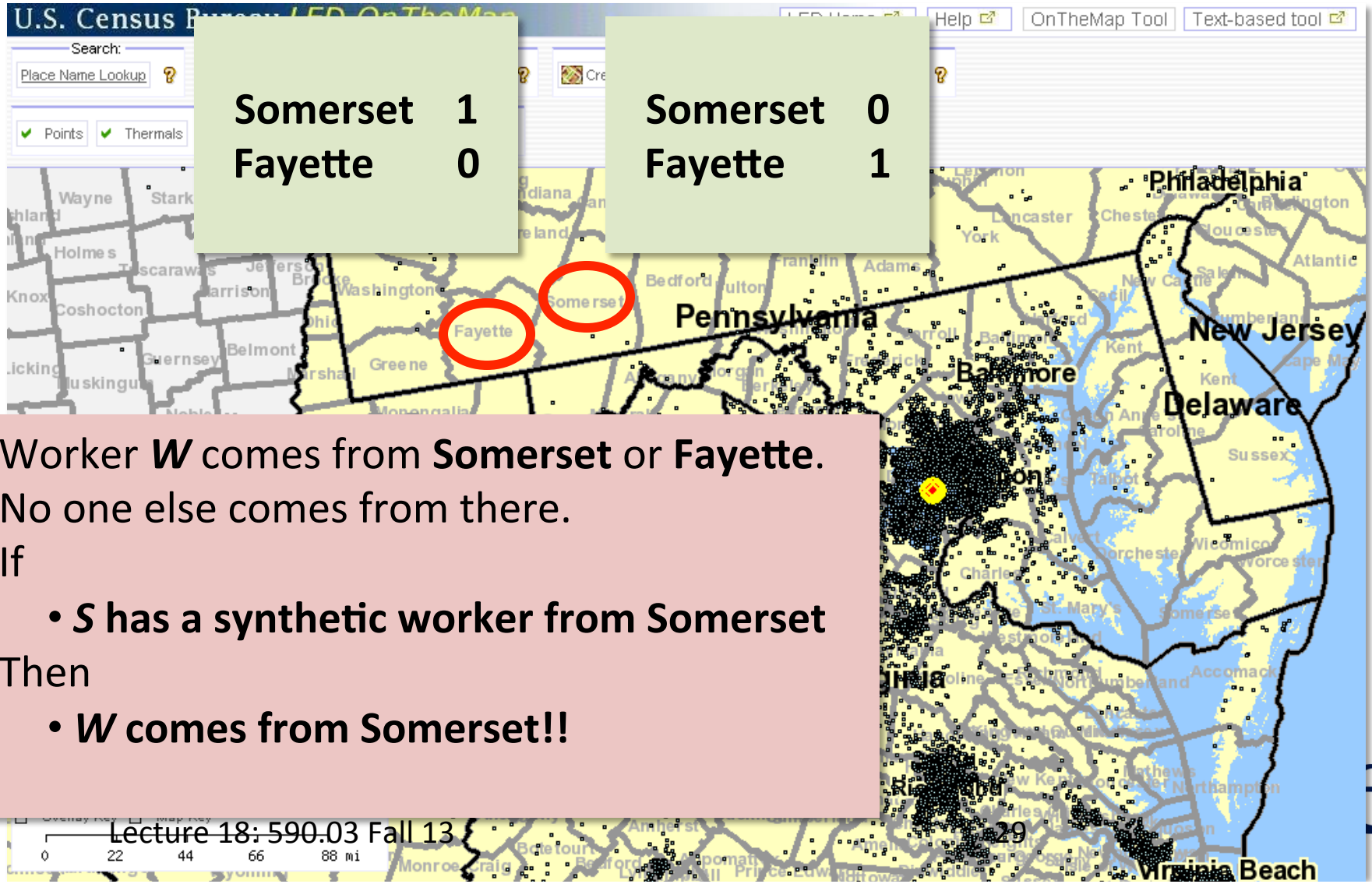
Duke UNIVERSITY

Theorem 3: Adding noise only to blocks that appear in the data breaches privacy.

If a block *b* does not appear in the original data and no noise is added to *b*

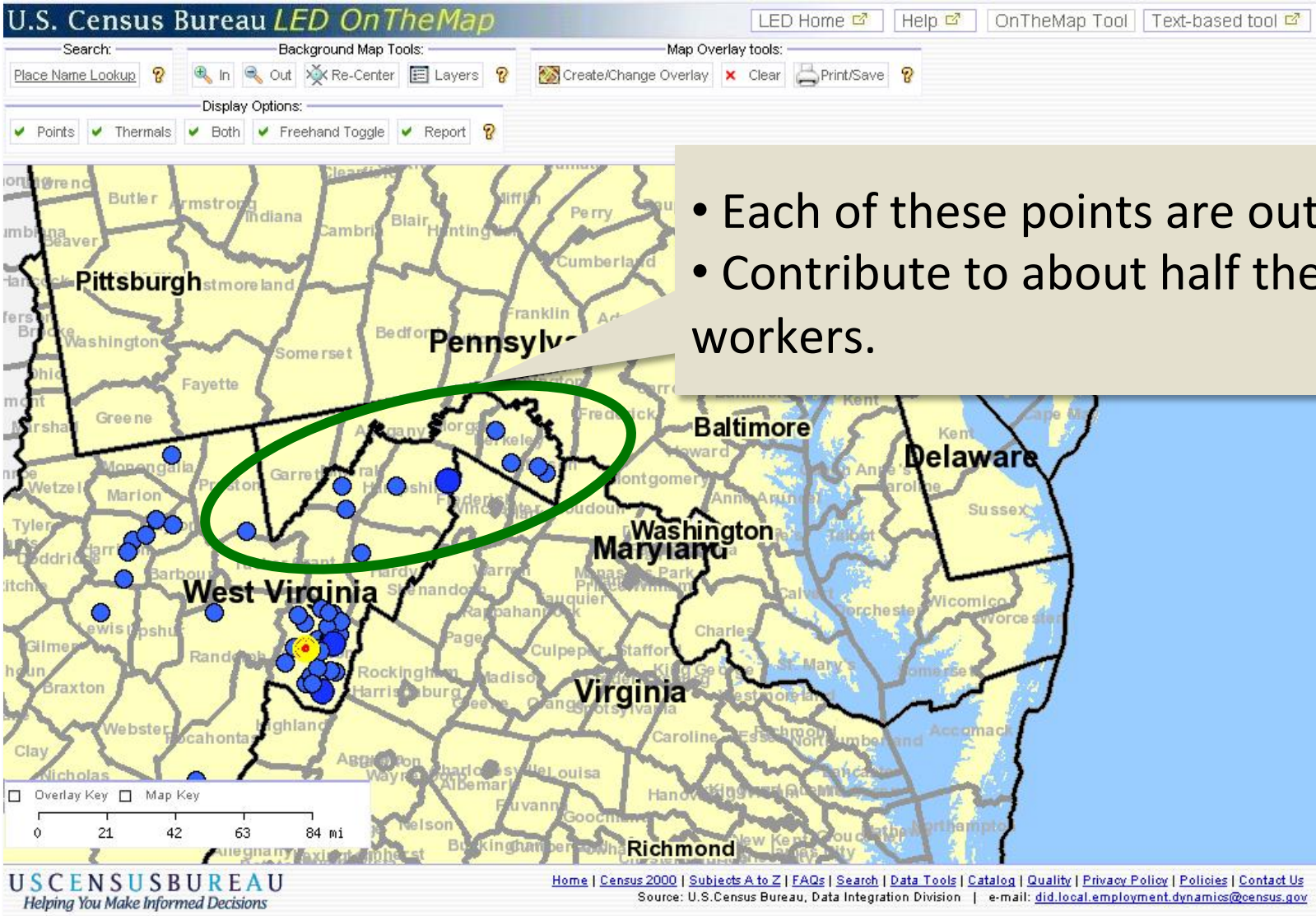
then *b* cannot appear in the synthetic data.

Theorem 3: Adding noise only to blocks that appear in the data breaches privacy.



- Worker ***W*** comes from **Somerset** or **Fayette**.
- No one else comes from there.
- If
  - ***S* has a synthetic worker from Somerset**
- Then
  - ***W* comes from Somerset!!**

Ignoring outliers degrades utility



- Each of these points are outliers.
- Contribute to about half the workers.

Our solution to "Where to add noise?"

Step 1 : **Coarsen the domain**

- Based on an existing public dataset (Census Transportation Planning Package, CTPP).

Our solution to "Where to add noise?"

Step 1 : **Coarsen the domain**

Step 2: **Probabilistically drop blocks with 0 support**

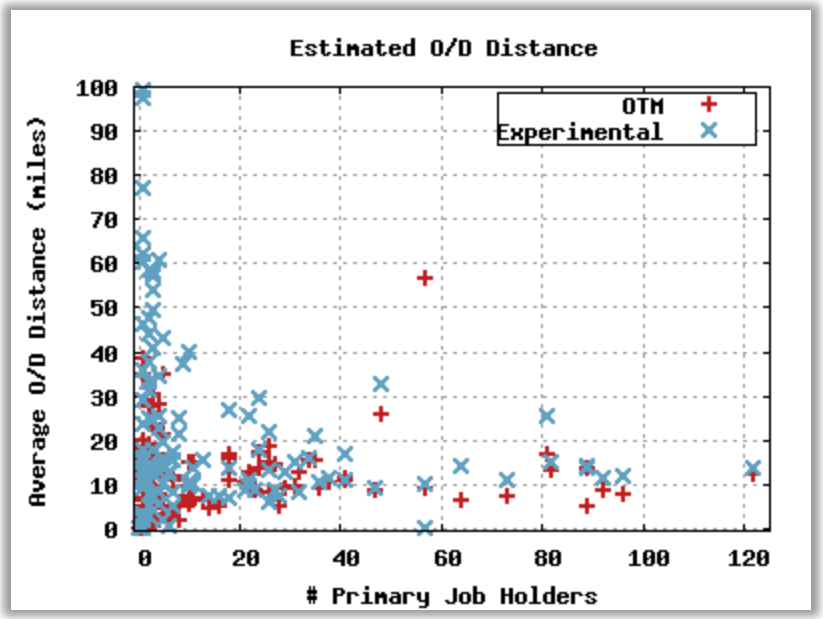
- Pick a function $f: \{b_1, \ldots, b_k\} \rightarrow (0,1]$ (based on external data)
- For every block b with 0 support,
  ignore b with probability f(b)

Theorem 4:

Parameter ε increases by $\max_b \left(\max \left(2^{\text{noise per block}}, f(b) \right) \right)$

Utility of the provably private algorithm

Utility measured by average commute distance for each destination block.

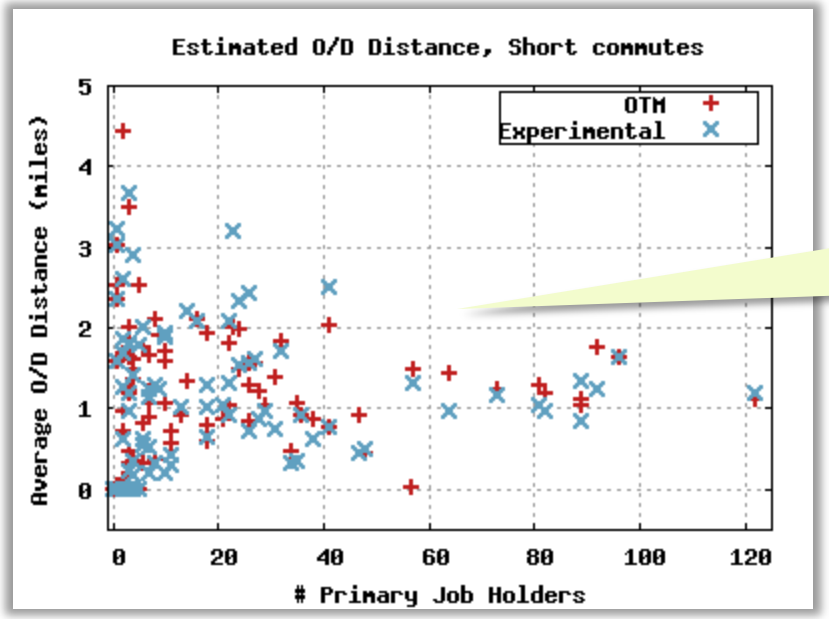


Experimental Setup:

- **OTM**: Currently published OnTheMap data used as original data.
- All destinations in Minnesota.
- 120, 690 origins per destination.
  - chosen by pruning out blocks that are > 100 miles from the destination.
- $\varepsilon = 100$, $\delta = 10^{-5}$
- Additional leakage due to probabilistic pruning = 4 *(min f(b) = 0.0378)*

Utility of the provably private algorithm

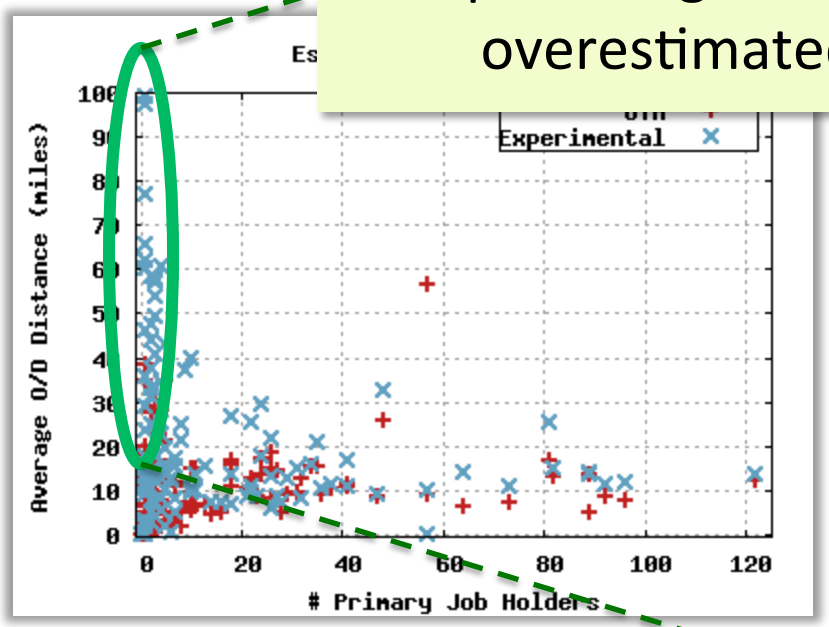
Utility measured by average commute distance for each destination block.

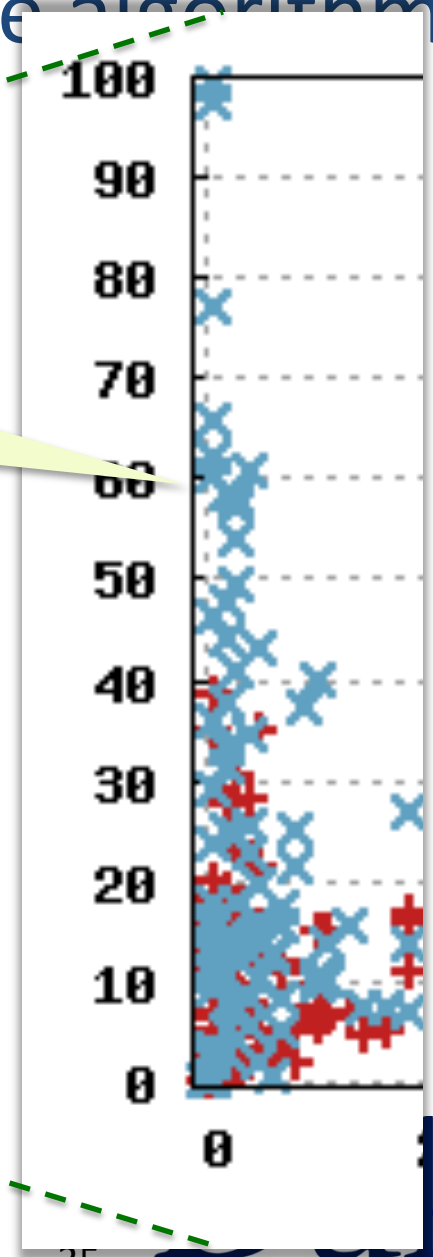


Short commutes have low error in both sparse and dense regions.

Utility of the provably private algorithm

Long commutes in sparse regions are overestimated.





OnTheMap: Summary

- OnTheMap: A real census application.
  - Synthetically generated data published for economic research.
  - Currently, privacy implications are poorly understood.
    - Parameters to the algorithm are **state secret**.

- First formal privacy analysis of this application.
  - Analyzed the privacy of OnTheMap using variants of Differential Privacy.
  - First solutions to publish useful information despite sparse data.

References

[**M** et al ICDE '08]
A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, L. Vilhuber, "*Privacy: From Theory to Practice on the Map",* ICDE 2008