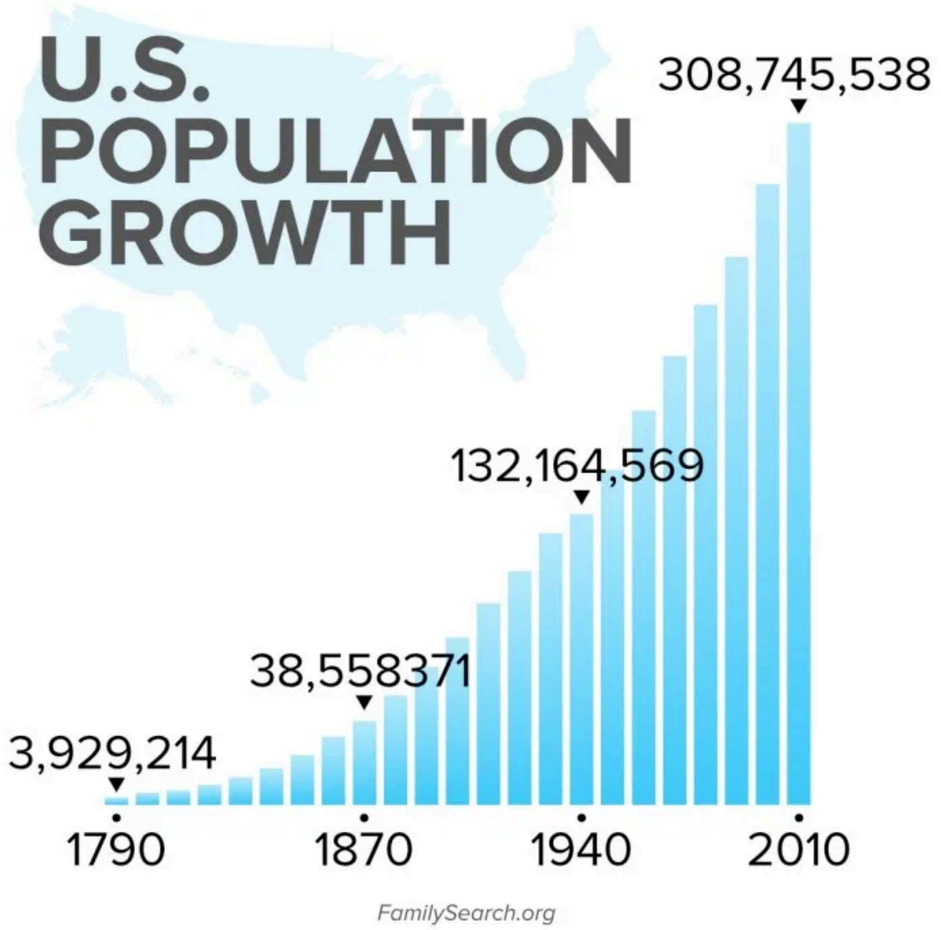
Differential Privacy of Hierarchical Census Data: An Optimization Approach

Ferdinando Fioretto
Syracuse University
Georgia Institute of Technology

Pascal Van Hentenryck
Georgia Institute of Technology

Oct 1, 2019

US Census Data Collection



Georgia Tech

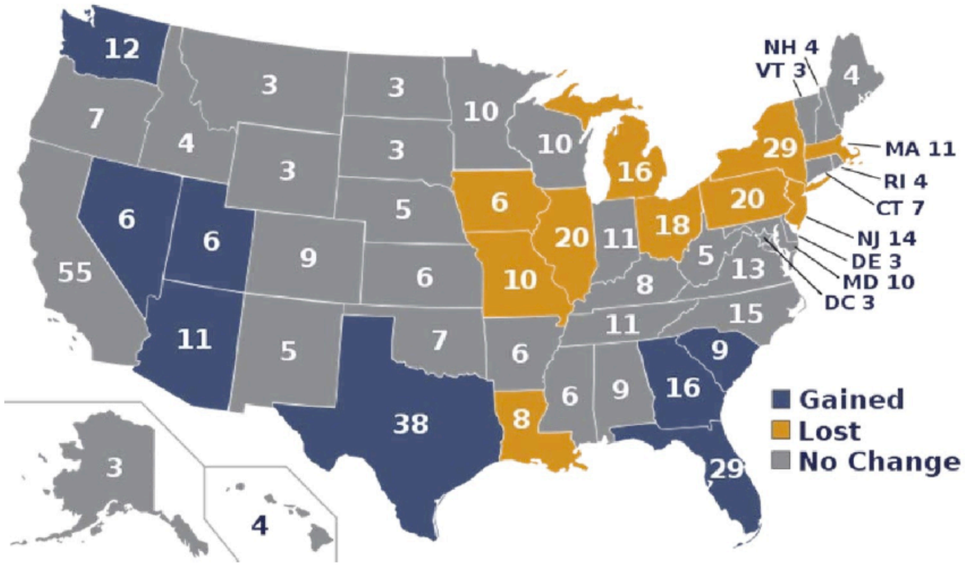
Accurate Count is Important

- Used to apportion multiple federal funding streams
  - $675 billions allocated to 132 programs (2015)

Medicaid

Highway Planning and Construction

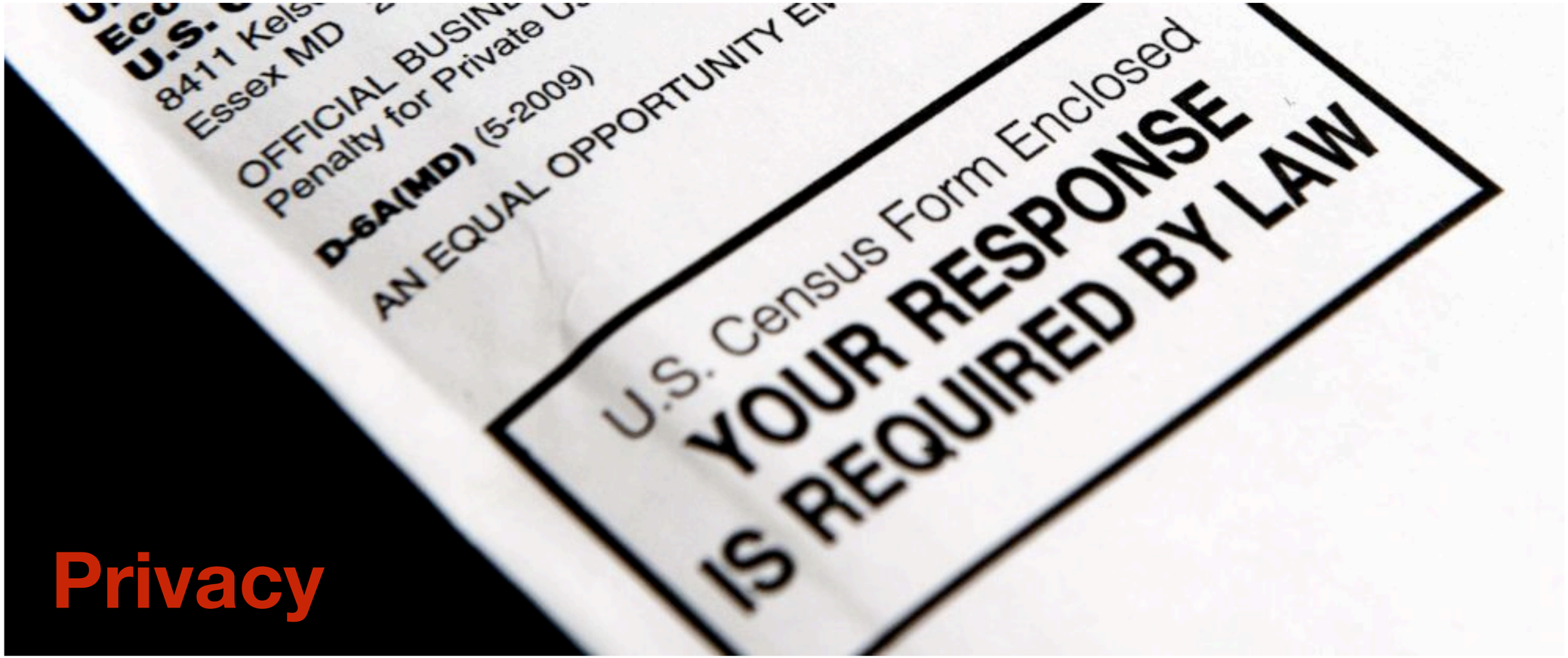
U.S. DEPARTMENT OF EDUCATION



- Determine the number of seats that states get in the US House of Representatives

Georgia Tech

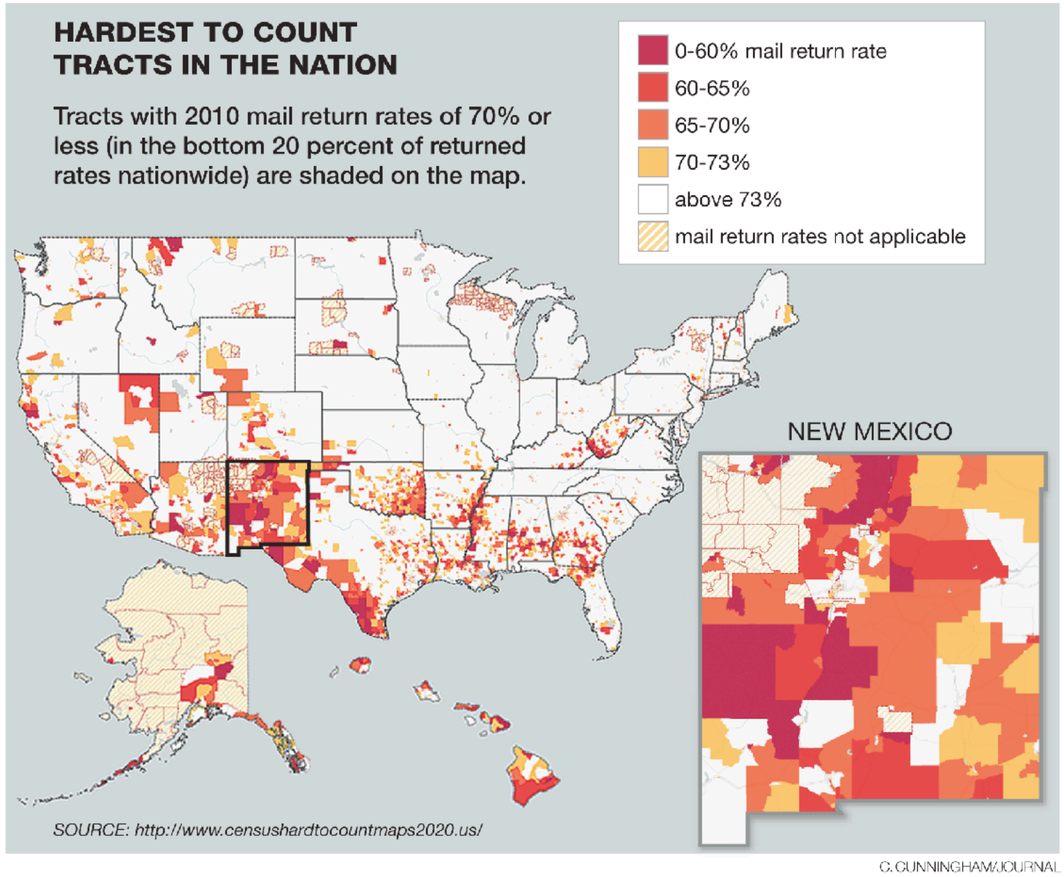
US Census Data Collection



Census is required to keep information confidential

Georgia Tech

US Census Data Collection



"Is this person a citizen of the United States?"

California has approved $100 million since 2017 to hire workers with the aim of reaching people who are hard to count

How do you keep survey data private?

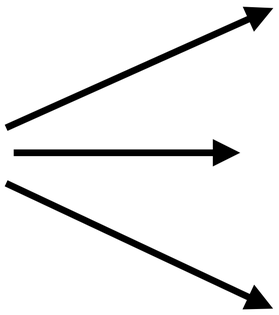
- The census bureau is supposed to keep the collected information confidential.
- How can you retain privacy while publishing results about the survey?

user	age	gender
Margaret	31	F
Luis	49	M
Maria	26	F
Carl	19	M
Isabelle	27	F

How do you keep survey data private?

- The census bureau is supposed to keep the collected information confidential.
- How can you retain privacy while publishing results about the survey?
- Database Reconstruction Theorem: Every information released contributes to violate the privacy of an individual. The more information it is publicly released the more privacy is violated.

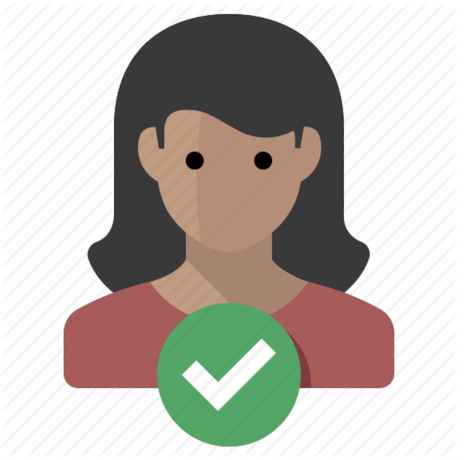
user	age	gender
~~Margaret~~	31	F
~~Luis~~	49	M
~~Maria~~	26	F
~~Carl~~	19	M
~~Isabelle~~	27	F



Avg.: 30.4

Avg.: 34 Male

Avg.: 28 Female

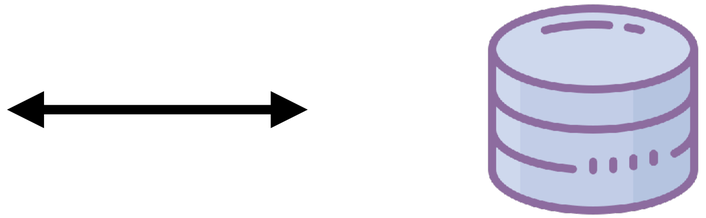


Irit:03

Reconstruction Attacks

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov



308,745,548 people in 2010 release which implements some "protection"

Commercial databases

Linkage Attacks — Results from UC Census

- Census blocks correctly reconstructed in all 6,207,027, inhabited blocks
- Block, sex, age, race, ethnicity reconstructed:
  - Exactly: 46% of population (142M)
  - Allowing age +/- 1 year: 71% of population (219M)
- Name, block sex, age, race, ethnicity:
  - Confirmed re-identification: 38% of population

Ramachandran:12

McKenna:18

Georgia Tech

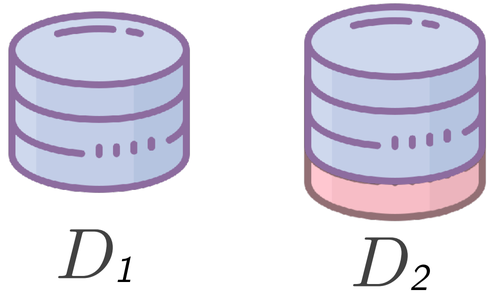
Differential Privacy

Differential Privacy

For every pair of inputs that differs in one row

For every output O

A randomized algorithm $\mathcal{A}$ is ε-differentially private if:



$$\frac{\Pr[\mathcal{A}(D_1) = O]}{\Pr[\mathcal{A}(D_2) = O]} \leq \exp(\epsilon)$$

Intuition: An adversary should not be able to use output O to distinguish between any D_1 and D_2

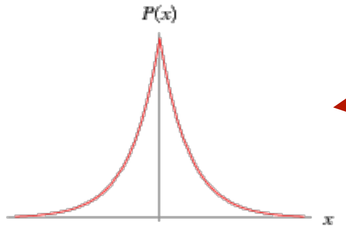
Dwork:06

How to Publish Differentially Private Data

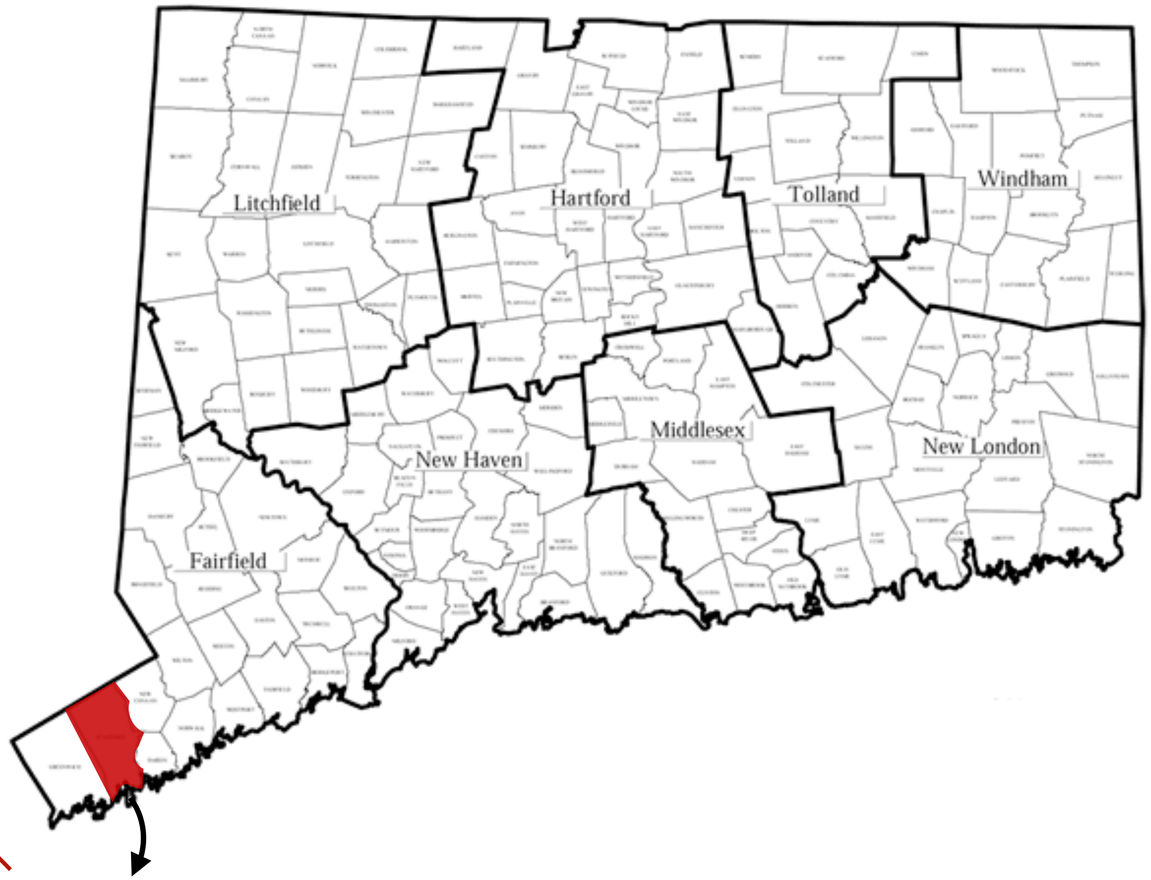
Q: count the number of people in the Stamford census block

Laplace Mechanism

$$\xi \sim Lap(\frac{1}{\epsilon})$$



P $f(x \mid \mu = 0, b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$



130,824+ξ_S

Differential Privacy

Notable Properties

- *No linkage attack*: *Adversary knows arbitrary auxiliary information*

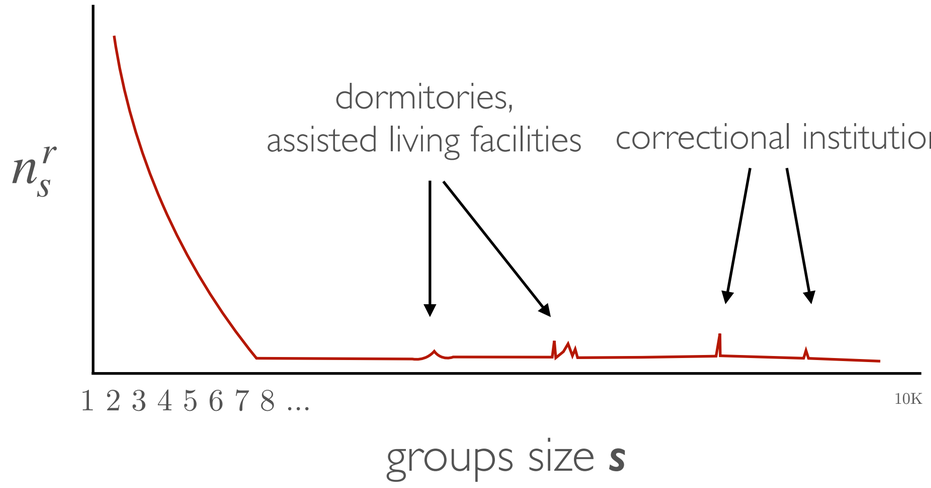
- *Post-processing immunity*: *If* A *enjoys* ε*-differential privacy and* g *is an arbitrary mapping,* $g \circ A$ *is* ε*-differential private*

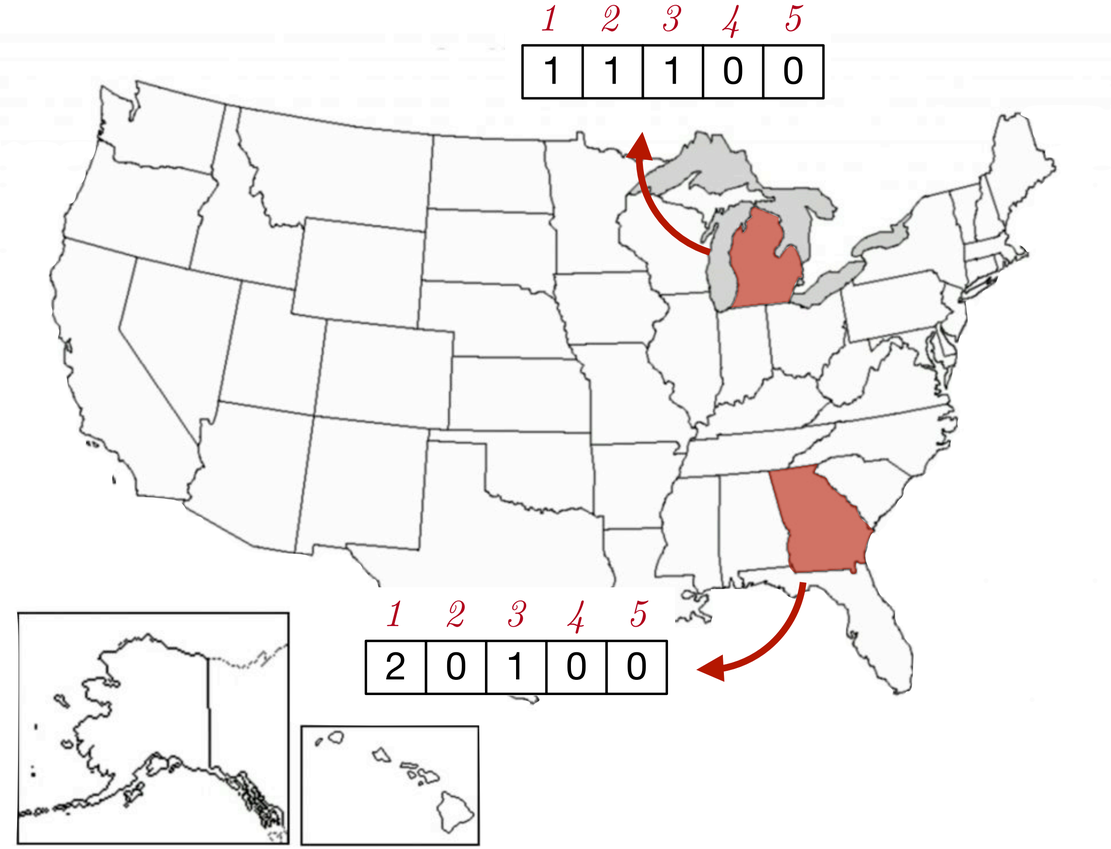
Dwork:06

The Census Group Size Release Problem

The Census Group Size Release Problem

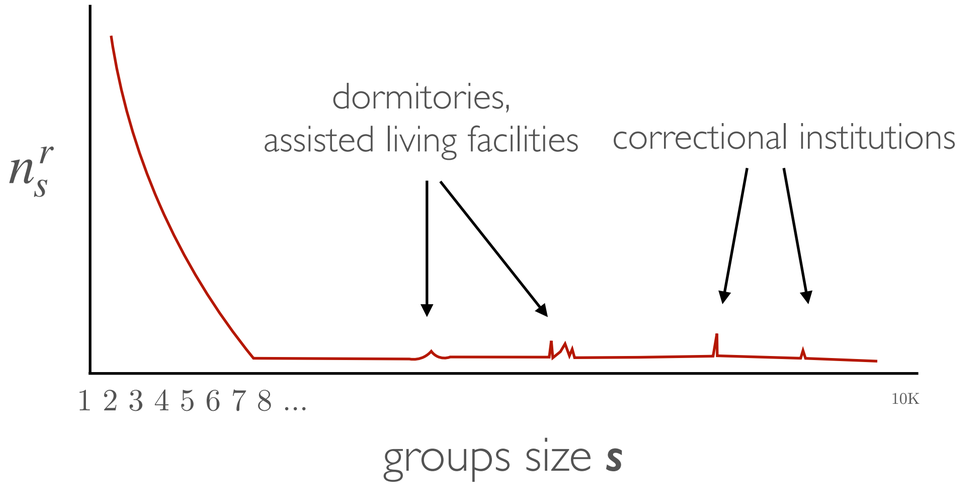
For every group size $s \in [N]$ (i.e., number of households of size 1, 2, 3, …) and region ***r*** (i.e., census block, counties, states) release the number n_s^r of groups of size ***s*** in region ***r*** while preserving privacy

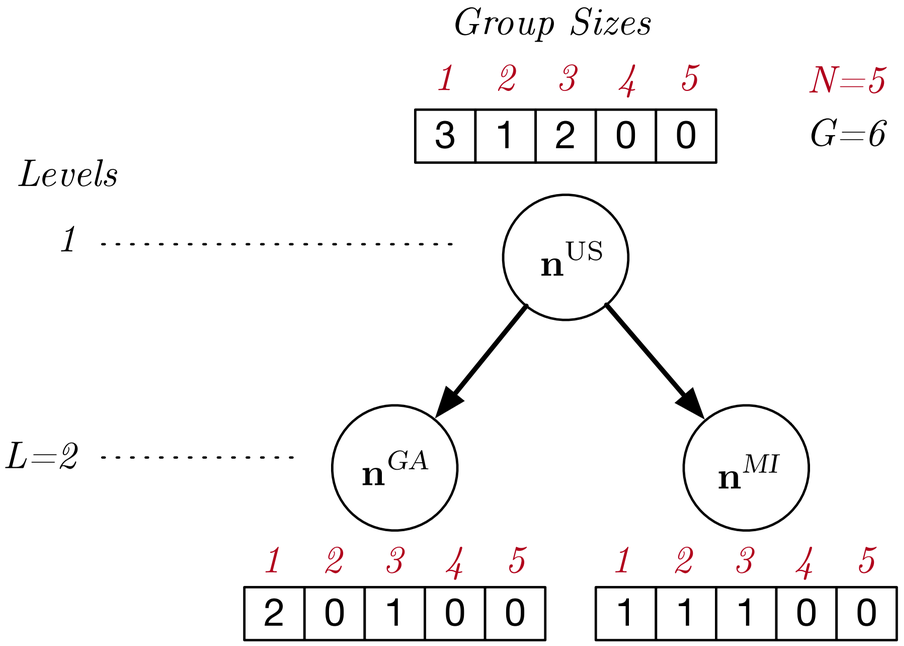




The Census Group Size Release Problem

For every group size $s \in [N]$ (i.e., number of households of size 1, 2, 3, …) and region ***r*** (i.e., census block, counties, states) release the number n_s^r of groups of size ***s*** in region ***r*** while preserving privacy





Requirements

- Requirements:

  1. **Privacy**

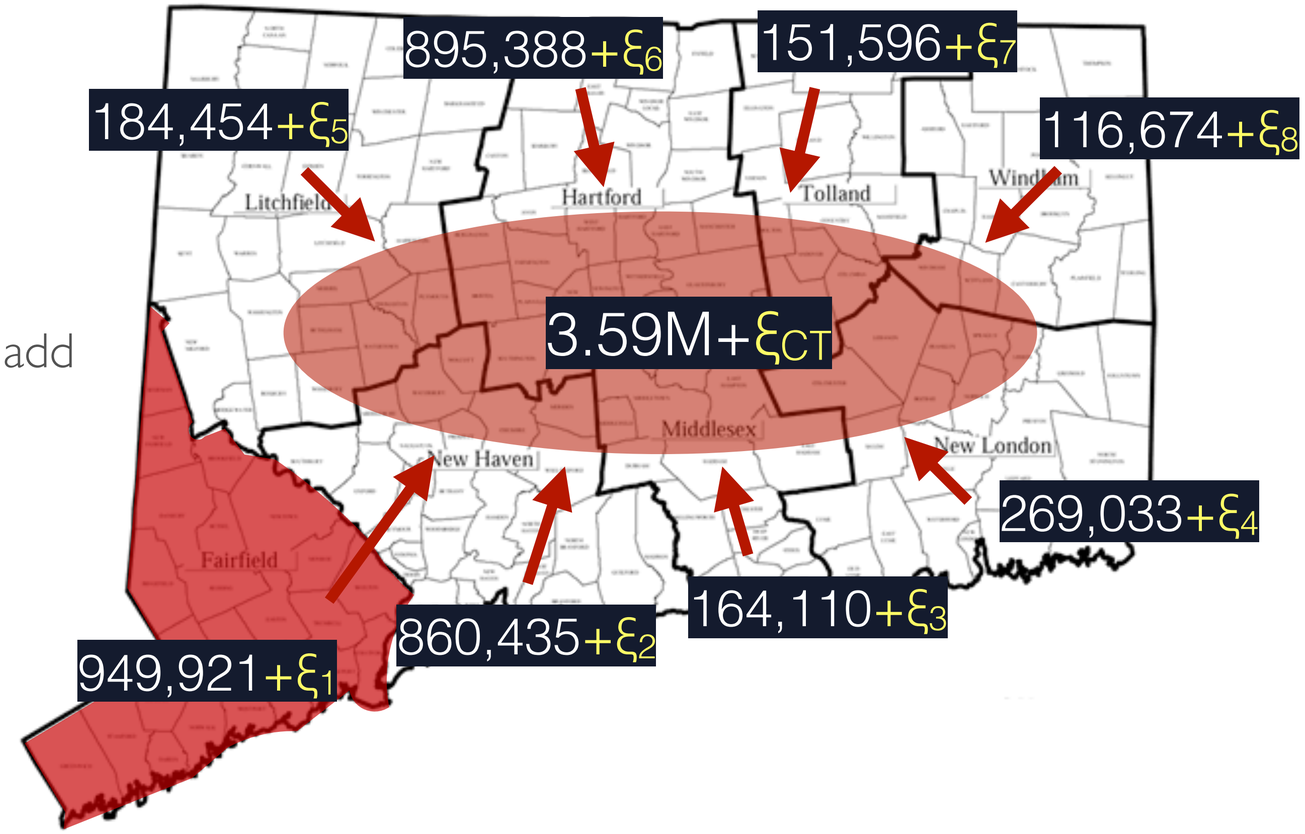
  2. Hierarchical **Consistency**

  3. **Validity**: The values $\tilde{n}_s^r$ are non-negative

  4. **Faithfulness**; The group size at each level need to add up to value G (which is a given public information)

- Noise is applied independently to each estimate.

- The noisy quantities at a "level" (e.g., state) are inconsistent with the sum of the noisy quantities at the "children levels" (e.g., counties of that state)



Georgia Tech

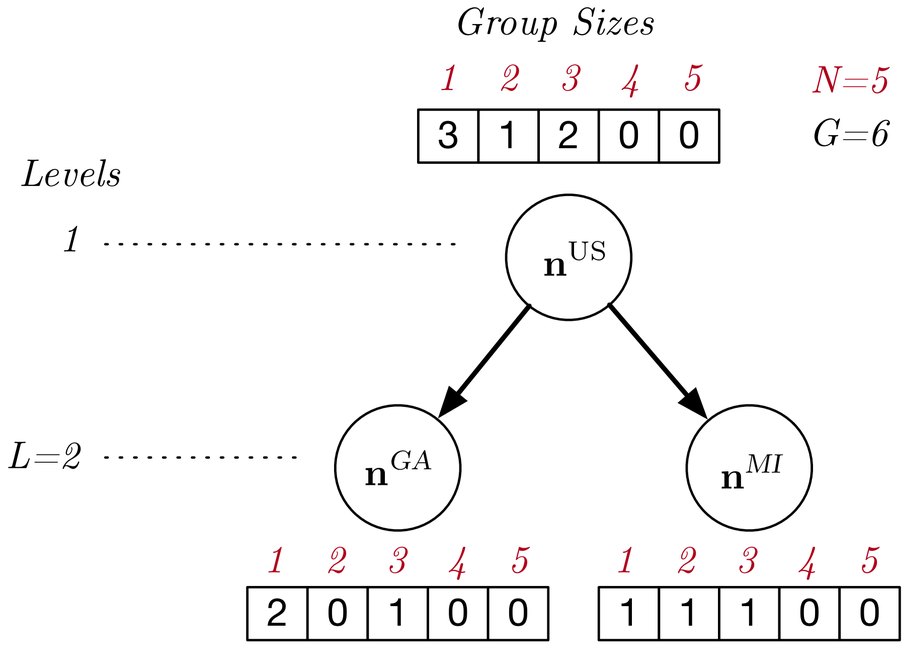
A Differentially Private Optimization Approach

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$

$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\begin{aligned}
&\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 && \text{(H1)}\\
&\text{s.t:} \quad \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} && \text{(H2)}\\
&\qquad \sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] && \text{(H3)}\\
&\qquad\qquad \hat{n}_s^r \in D_s^r \qquad \forall r \in \mathcal{R}, s \in [N] && \text{(H4)}
\end{aligned}$$



← Satisfies DP due to post-processing immunity

Georgia Tech

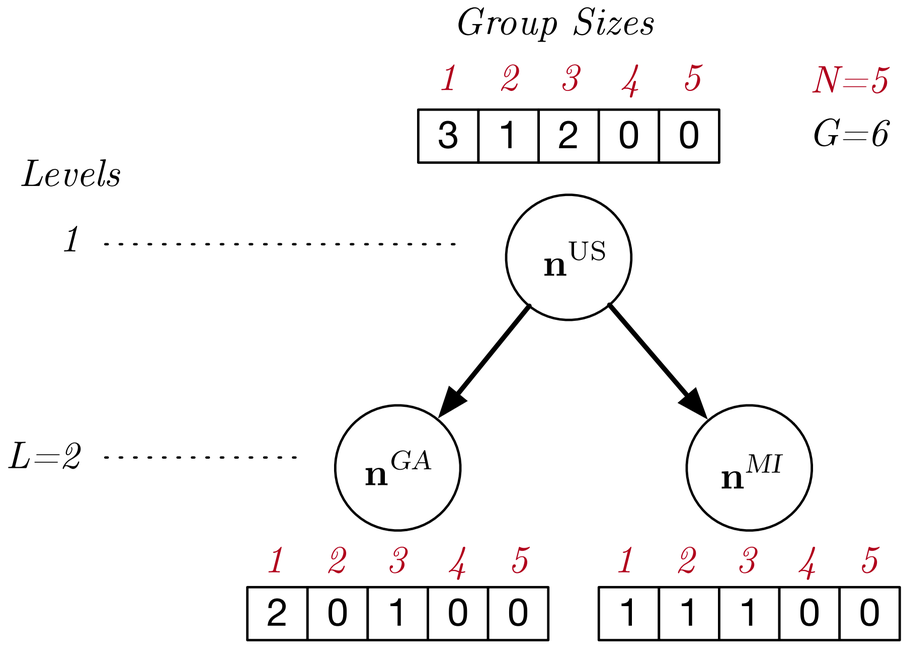
A Differentially Private Optimization Approach

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$

$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\begin{aligned}
\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad & \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 && \text{(H1)}\\
\text{s.t:} \quad & \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} && \text{(H2)}\\
& \sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] && \text{(H3)}\\
& \hat{n}_s^r \in D_s^r \qquad \forall r \in \mathcal{R}, s \in [N] && \text{(H4)}
\end{aligned}$$



← Satisfies DP due to post-processing immunity

Georgia Tech

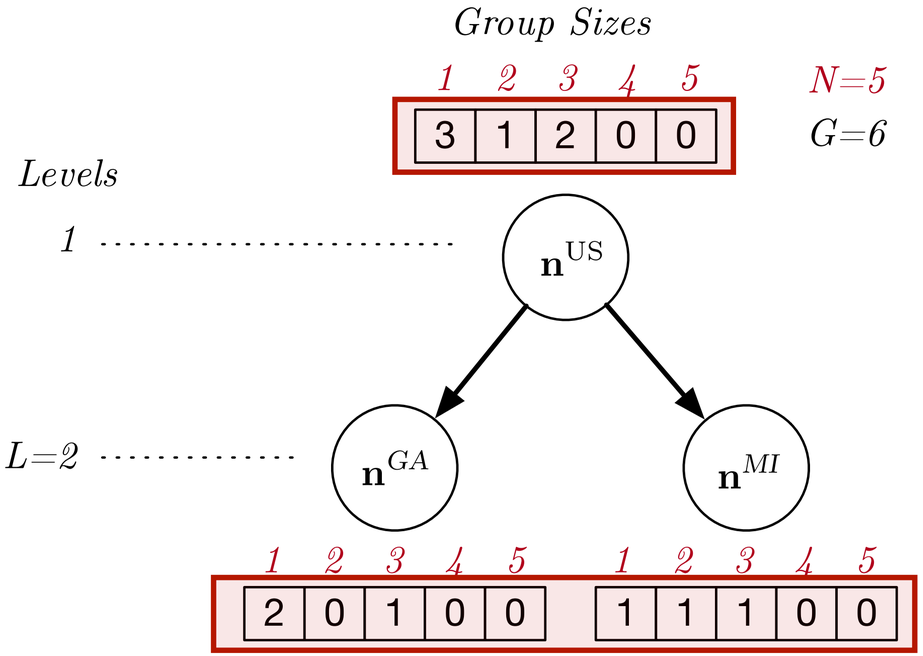
A Differentially Private Optimization Approach

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$

$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\begin{aligned}
&\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 && \text{(H1)}\\
&\text{s.t:} \quad \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} && \text{(H2)}\\
&\qquad \sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] && \text{(H3)}\\
&\qquad \hat{n}_s^r \in D_s^r \qquad \forall r \in \mathcal{R}, s \in [N] && \text{(H4)}
\end{aligned}$$



← Satisfies DP due to post-processing immunity

Georgia Tech

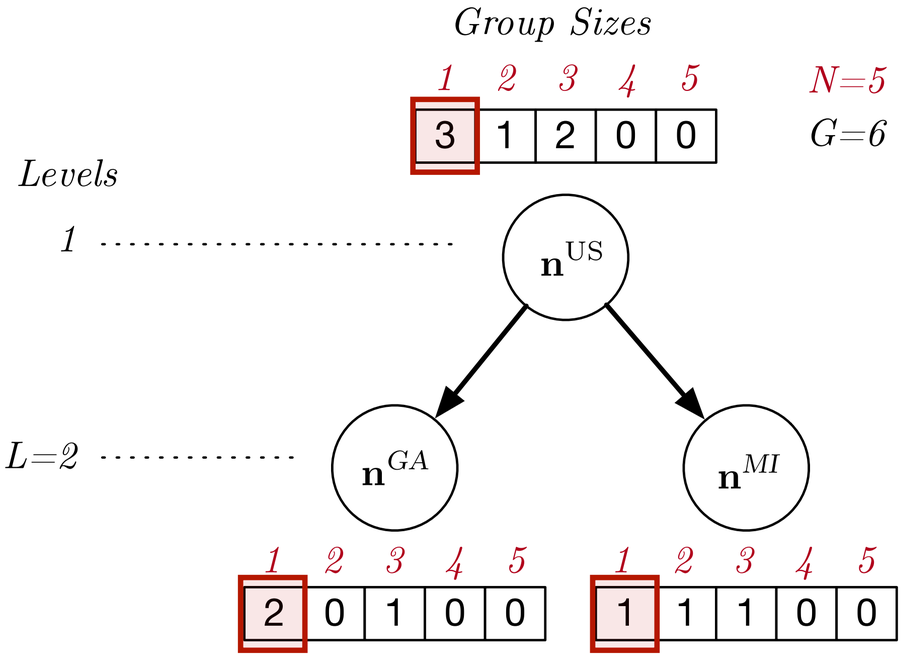
A Differentially Private Optimization Approach

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$

$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\begin{aligned}
&\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 && \text{(H1)}\\
&\text{s.t:} \quad \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} && \text{(H2)}\\
&\sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] && \text{(H3)}\\
&\hat{n}_s^r \in D_s^r \qquad \forall r \in \mathcal{R}, s \in [N] && \text{(H4)}
\end{aligned}$$



← Satisfies DP due to post-processing immunity

Georgia Tech

A Differentially Private Optimization Approach

2. Postprocess output $\tilde{\boldsymbol{n}}$ to enforce consistency

$$\underset{\{\hat{\boldsymbol{n}}^r\}_{r\in\mathcal{R}}}{\text{minimize}} \quad \sum_{r\in\mathcal{R}} \|\hat{\boldsymbol{n}}^r - \tilde{\boldsymbol{n}}^r\|_2^2 \qquad \text{(H1)}$$

$$\text{s.t:} \quad \sum_{s\in[N]} \hat{n}_s^r = G \quad \forall r \in \mathcal{R} \qquad \text{(H2)}$$

$$\sum_{c\in ch(r)} \hat{n}_s^c = \hat{n}_s^r \quad \forall r \in \mathcal{R}, s \in [N] \qquad \text{(H3)}$$

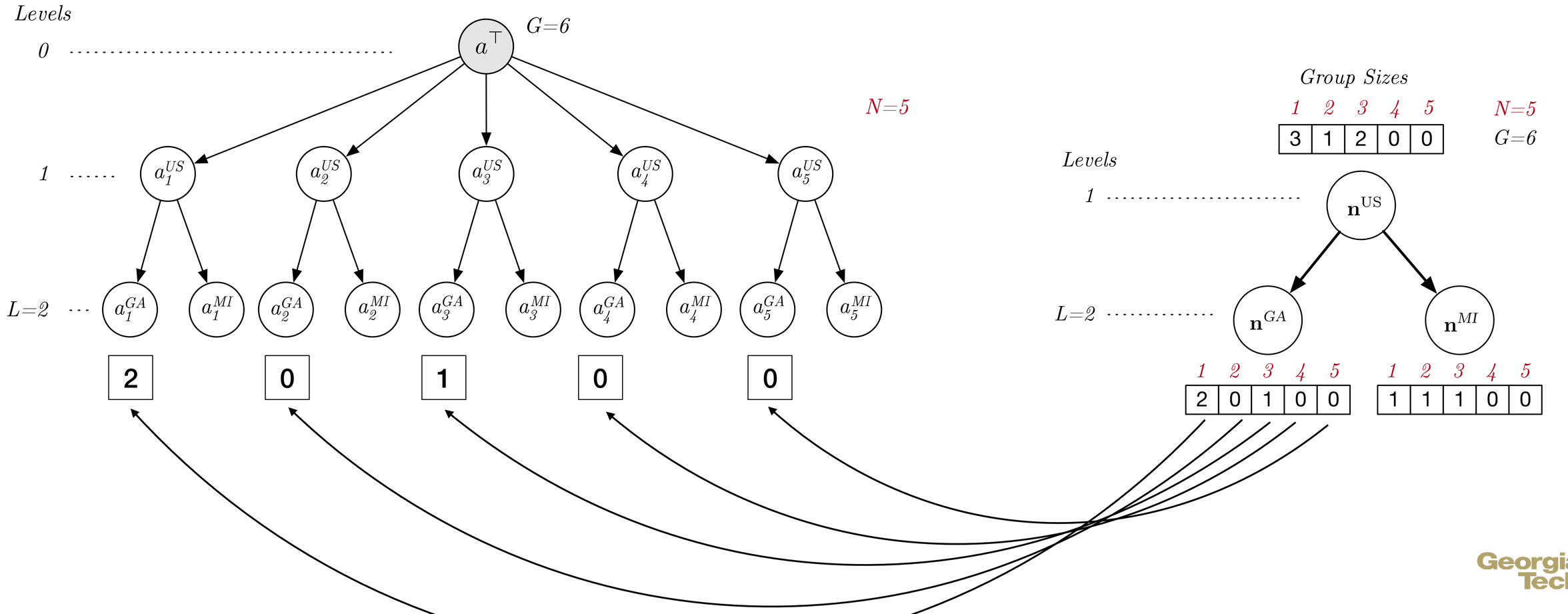
$$\hat{n}_s^r \in D_s^r \quad \forall r \in \mathcal{R}, s \in [N] \qquad \text{(H4)}$$

- Solving this QIP is intractable for the datasets of interest to the census bureau.
- Relax the integrality constraint.
- The resulting optimization problem becomes convex but presents two limitations:
  1. Its final solution may violate the **consistency** and **faithfulness** conditions
  2. The mechanism is still too slow for very large problems!

Paper Contribution

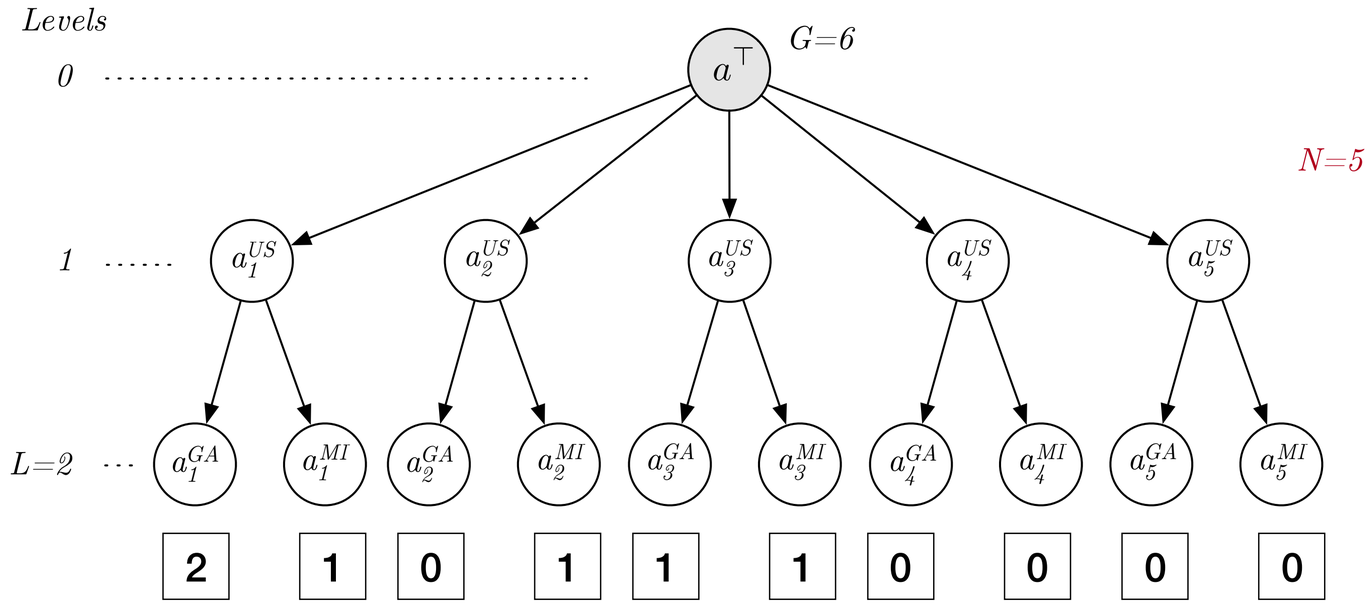
Exploiting the Problem Structure

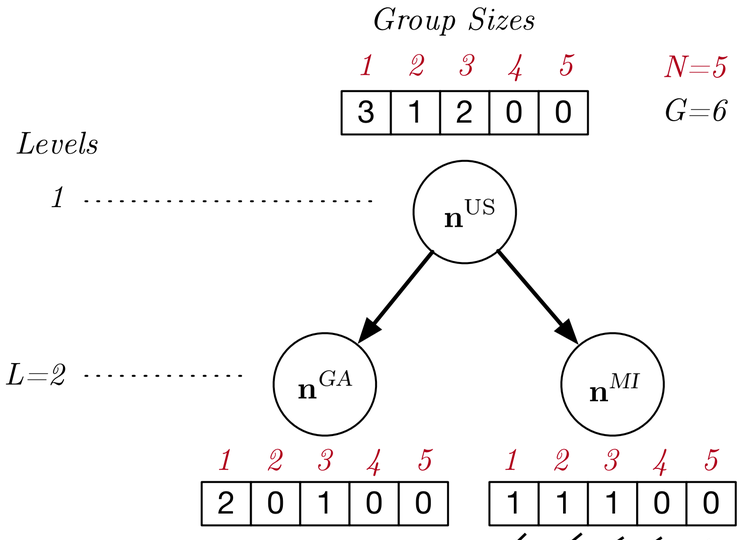
A Dynamic Programming Solution



Exploiting the Problem Structure

A Dynamic Programming Solution

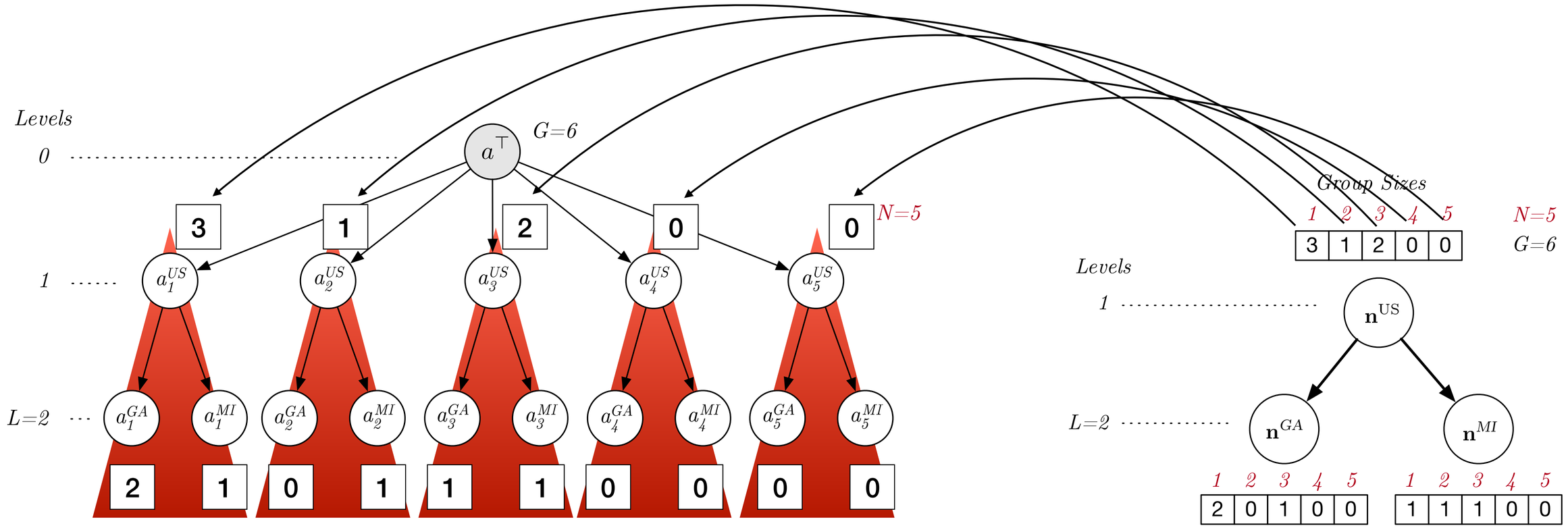




Georgia Tech

Exploiting the Problem Structure

A Dynamic Programming Solution



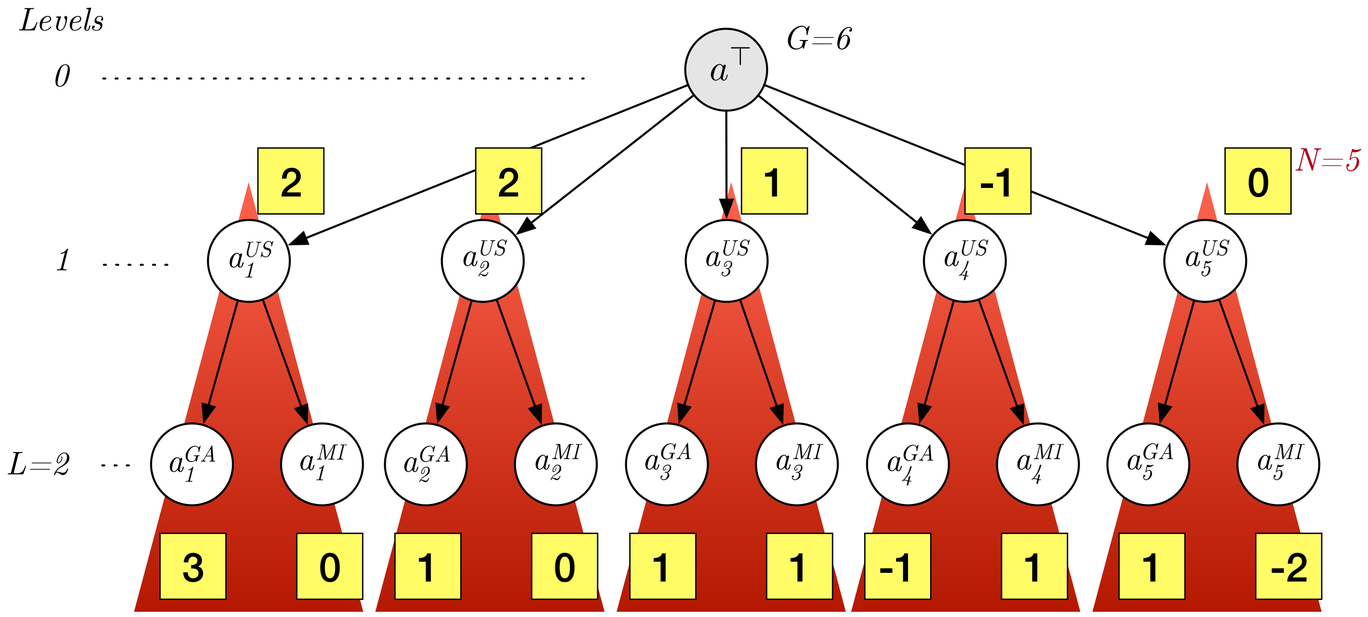
independent subtrees

Georgia Tech

Exploiting the Problem Structure

A Dynamic Programming Solution

1. Apply Geometrical Noise with parameter $\lambda = \frac{2L}{\epsilon}$



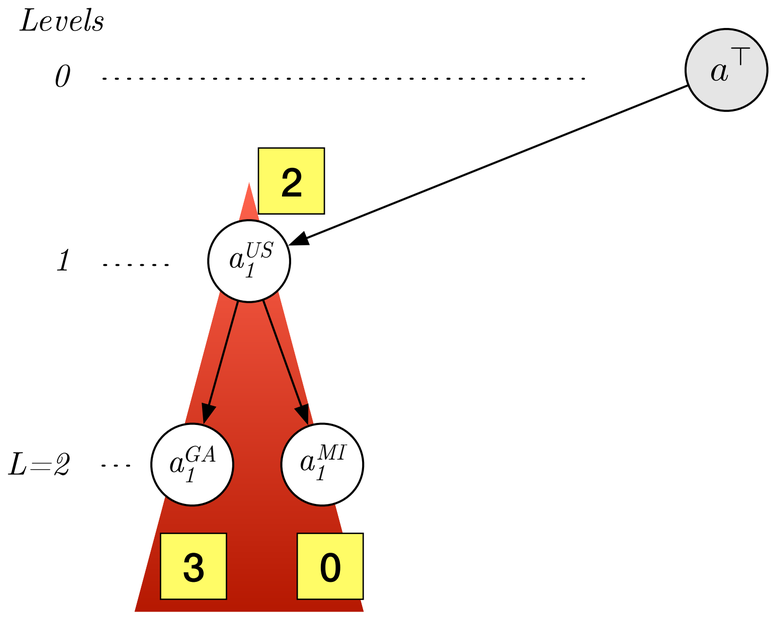
$$\tilde{\boldsymbol{n}}^r = \boldsymbol{n}^r + Geom\left(\frac{2L}{\epsilon}\right)^N$$

independent subtrees

Georgia Tech

Exploiting the Problem Structure

A Dynamic Programming Solution



2. Bottom-up phase

- Find new, group sizes $\hat{n}^r$ that satisfy the consistency properties.
- Each node of the tree, computes a table $\tau^r : D^r \rightarrow \mathbb{R}_+$ mapping values (group sizes) to costs.
- $\tau^r(v)$ is the optimal cost for $\hat{n}^r$ in the subtree rooted at region **r** when $\hat{n}^r = v$
- The optimal cost for $\tau^r(v)$ can be computed from the cost table τ^c of region **r** children $c \in ch(r)$

$$\tau^r(v) = \left(v - \tilde{n}^r\right)^2 + \qquad \text{(d1)}$$

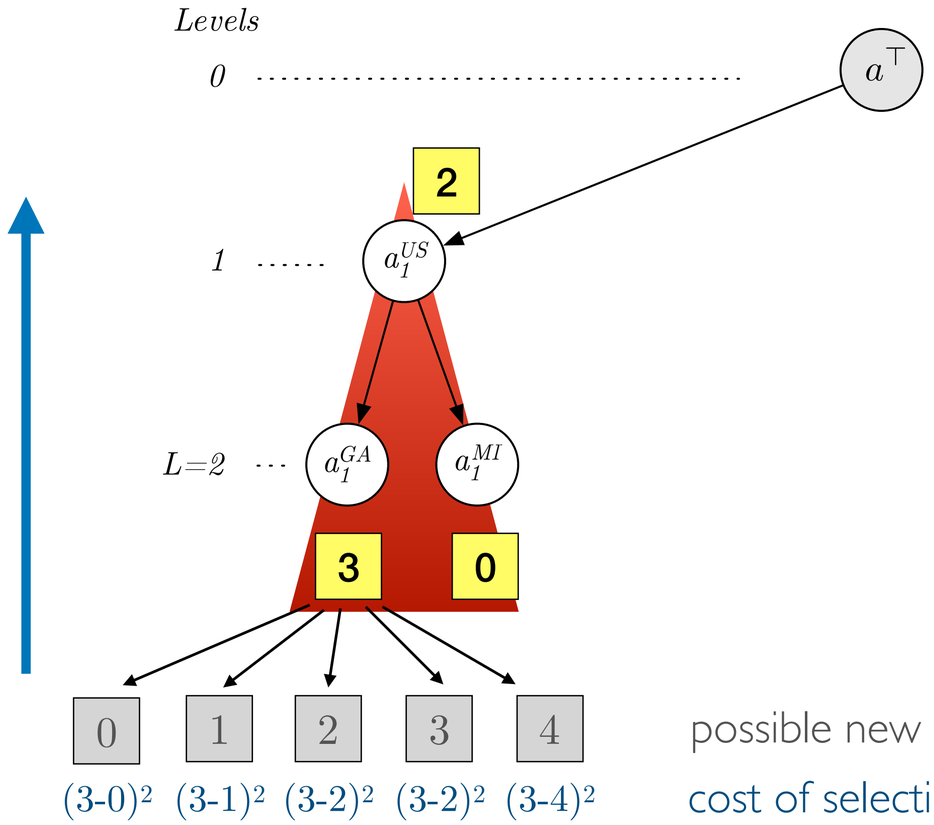
$$\phi^r(v) = \min_{\{x_c\}_{c \in ch(r)}} \sum_{c \in ch(r)} \tau^c(x_c) \qquad \text{(d2)}$$

$$\text{s.t.} \sum_{c \in ch(r)} x_c = v \qquad \text{(d3)}$$

$$x_c \in D^c \quad \forall c \in ch(r) \qquad \text{(d4)}$$

Exploiting the Problem Structure

A Dynamic Programming Solution



possible new values v for $\hat{n}_1^{GA}$

cost of selecting value v

2. Bottom-up phase

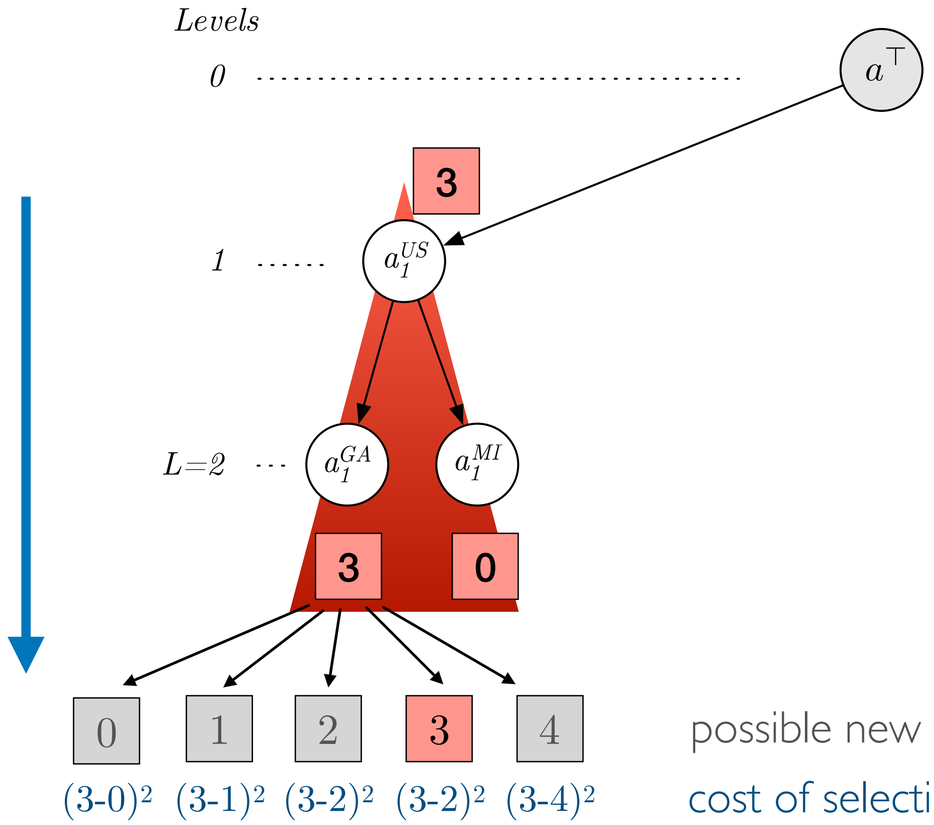
- Find new, group sizes $\hat{n}^r$ that satisfy the consistency properties.
- Each node of the tree, computes a table $\tau^r : D^r \rightarrow \mathbb{R}_+$ mapping values (group sizes) to costs.
- $\tau^r(v)$ is the optimal cost for $\hat{n}^r$ in the subtree rooted at region **r** when $\hat{n}^r = v$
- The optimal cost for $\tau^r(v)$ can be computed from the cost table τ^c of region **r** children $c \in ch(r)$

cost tables

v	τ_1^{GA}	τ_1^{MI}	τ_1^{US}
0	9	**0**	$4+\min(9+0)$
1	4	1	$1+\min(9+1;\ 4+0)$
2	1	4	$0+\min(0+4;\ 4+1;\ 1+0)$
3	**0**	9	$\mathbf{1}+\min(0+0;\ 1+1;\ 4+4;\ 9+9)$
4	1	16	$4+\min(1+0;\ 0+1;\ 1+4;\ \ldots)$

Exploiting the Problem Structure

A Dynamic Programming Solution



possible new values v for $\hat{n}_1^{GA}$

cost of selecting value v

3. Top-down phase

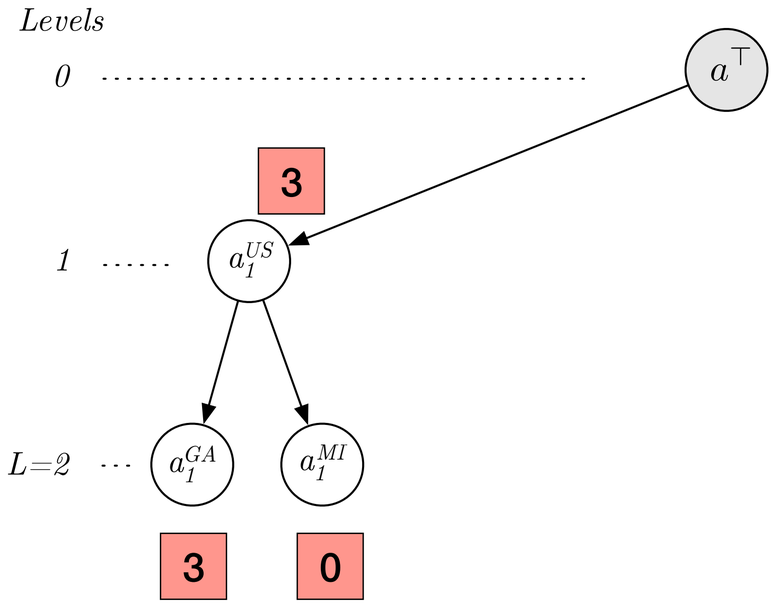
- Each node select the value v that minimizes its cost table
- Using the parent value, choice each children repeats the process

cost tables

v	τ_1^{GA}	τ_1^{MI}	τ_1^{US}
0	9	**0**	$4+\min(9+0)$
1	4	1	$1+\min(9+1;\ 4+0)$
2	1	4	$0+\min(0+4;\ 4+1;\ 1+0)$
3	**0**	9	$\mathbf{1}+\min(0+0;\ 1+1;\ 4+4;\ 9+9)$
4	1	16	$4+\min(1+0;\ 0+1;\ 1+4;\ \ldots)$

Exploiting the Problem Structure

A Dynamic Programming Solution



The issue

- The construction of the data hierarchy requires solving $O(|R|N\bar{D})$ optimization problems, with $\bar{D} = \max_{s,r} |D_s^r|$

R = regions, s = groups sizes (1, 2, 3 , …)

$$\boldsymbol{\tau}^r(v) = \left(v - \tilde{n}^r\right)^2 + \quad \text{(d1)}$$

$$\boldsymbol{\phi}^r(v) = \min_{\{x_c\}_{c \in ch(r)}} \sum_{c \in ch(r)} \boldsymbol{\tau}^c(x_c) \quad \text{(d2)}$$

$$\text{s.t.} \sum_{c \in ch(r)} x_c = v \quad \text{(d3)}$$

$$x_c \in D^c \quad \forall c \in ch(r) \quad \text{(d4)}$$

Exploiting the Cost Functions Structure

A Polynomial Time Mechanism

Main Result: The function ϕ_s^r used to compute the values $\tau_s^c(v)$ convex piecewise linear (CPWL)

$$v_c^k = \begin{cases} v_c^{k-1} + 1 \text{ if } \; c = \text{argmin}_c \, \tau^c(v_c^{k-1}+1) - \tau^c(v_c^{k-1}) \\ v_c^{k-1} \qquad\qquad \text{otherwise.} \end{cases}$$

$$\tau^r(v) = \left(v - \tilde{n}^r\right)^2 + \tag{d1}$$

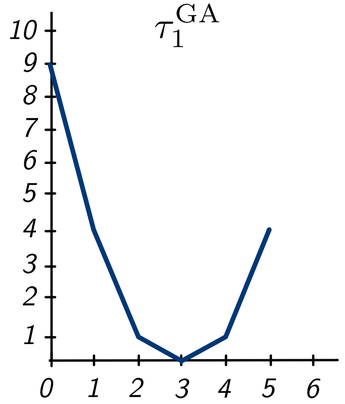
$$\phi^r(v) = \min_{\{x_c\}_{c\in ch(r)}} \sum_{c\in ch(r)} \tau^c(x_c) \tag{d2}$$

$$\text{s.t.} \sum_{c\in ch(r)} x_c = v \tag{d3}$$

$$x_c \in D^c \quad \forall c \in ch(r) \tag{d4}$$

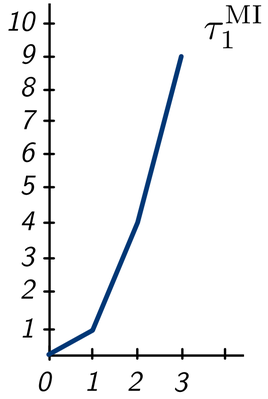
τ_1^{GA}

j	cost
0	9
1	4
2	1
3	0
4	1
5	4



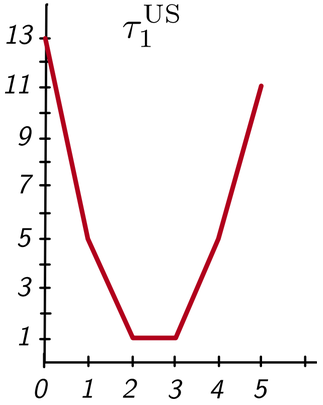
τ_1^{MI}

j	cost
0	0
1	1
2	4
3	9



τ_1^{US}

j	cost
0	4+(0+0)
1	1+(0+4)
2	0+(0+1)
3	1+(0+0)
4	4+(1+0)
5	9+(1+1)



Corollary: The cost table function of each node of the tree is CPWL

Exploiting the Cost Functions Structure

A Polynomial Time Mechanism

Main Result: The function ϕ_s^r used to compute the values $\tau_s^c(v)$ convex piecewise linear (CPWL)

$$v_c^k = \begin{cases} v_c^{k-1} + 1 \text{ if } \; c = \text{argmin}_c \, \tau^c(v_c^{k-1}+1) - \tau^c(v_c^{k-1}) \\ v_c^{k-1} \qquad \text{otherwise.} \end{cases}$$

$$\tau^r(v) = \left(v - \tilde{n}^r\right)^2 + \quad \text{(d1)}$$

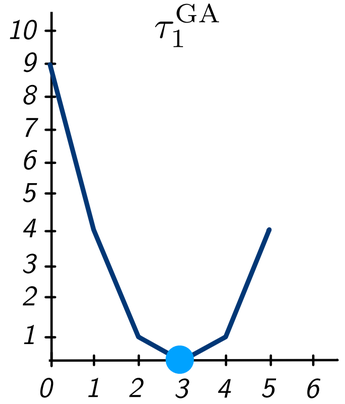
$$\phi^r(v) = \min_{\{x_c\}_{c\in ch(r)}} \sum_{c\in ch(r)} \tau^c(x_c) \quad \text{(d2)}$$

$$\text{s.t.} \sum_{c\in ch(r)} x_c = v \quad \text{(d3)}$$

$$x_c \in D^c \;\; \forall c \in ch(r) \quad \text{(d4)}$$

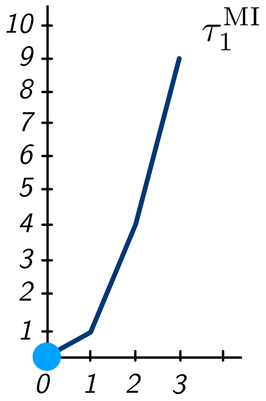
τ_1^{GA}

j	cost
0	9
1	4
2	1
3	0
4	1
5	4



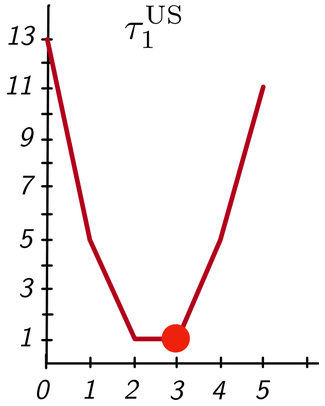
τ_1^{MI}

j	cost
0	0
1	1
2	4
3	9



τ_1^{US}

j	cost
0	4+(0+0)
1	1+(0+4)
2	0+(0+1)
3	1+(0+0)
4	4+(1+0)
5	9+(1+1)



Corollary: The cost table function of each node of the tree is CPWL

Exploiting the Cost Functions Structure

A Polynomial Time Mechanism

Main Result: The function ϕ_s^r used to compute the values $\tau_s^c(v)$ convex piecewise linear (CPWL)

$$v_c^k = \begin{cases} v_c^{k-1} + 1 \text{ if } c = \text{argmin}_c\, \tau^c(v_c^{k-1}+1) - \tau^c(v_c^{k-1}) \\ v_c^{k-1} \quad \text{otherwise.} \end{cases}$$

$$\tau^r(v) = \left(v - \tilde{n}^r\right)^2 + \quad \text{(d1)}$$

$$\phi^r(v) = \min_{\{x_c\}_{c\in ch(r)}} \sum_{c\in ch(r)} \tau^c(x_c) \quad \text{(d2)}$$

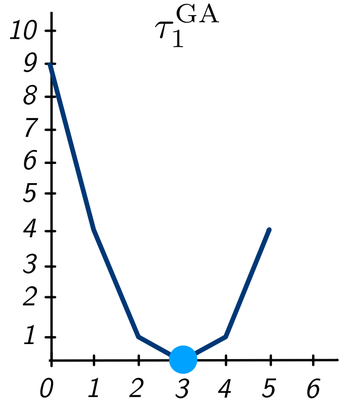
$$\text{s.t.} \sum_{c\in ch(r)} x_c = v \quad \text{(d3)}$$

$$x_c \in D^c \quad \forall c \in ch(r) \quad \text{(d4)}$$

τ_1^{GA}

j	cost
0	9
1	4
2	1
3	0
4	1
5	4

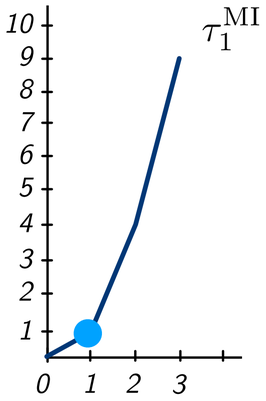
$\tau^{GA}(4) - \tau^{GA}(3) = 1$



τ_1^{MI}

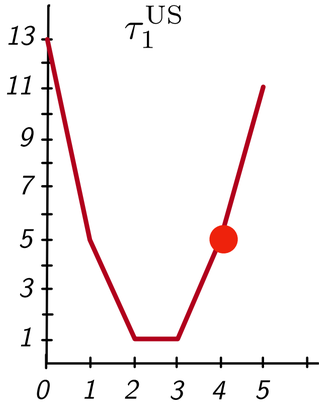
j	cost
0	0
1	1
2	4
3	9

$\tau^{MI}(1) - \tau^{MI}(0) = 1$



τ_1^{US}

j	cost
0	4+(0+0)
1	1+(0+4)
2	0+(0+1)
3	1+(0+0)
4	4+(1+0)
5	9+(1+1)



Corollary: The cost table function of each node of the tree is CPWL

Exploiting the Cost Functions Structure

A Polynomial Time Mechanism

Main Result: The function ϕ_s^r used to compute the values $\tau_s^c(v)$ convex piecewise linear (CPWL)

$$v_c^k = \begin{cases} v_c^{k-1} + 1 \text{ if } c = \text{argmin}_c\, \tau^c(v_c^{k-1}+1) - \tau^c(v_c^{k-1}) \\ v_c^{k-1} \quad \text{otherwise.} \end{cases}$$

$$\tau^r(v) = \left(v - \tilde{n}^r\right)^2 + \quad \text{(d1)}$$

$$\phi^r(v) = \min_{\{x_c\}_{c\in ch(r)}} \sum_{c\in ch(r)} \tau^c(x_c) \quad \text{(d2)}$$

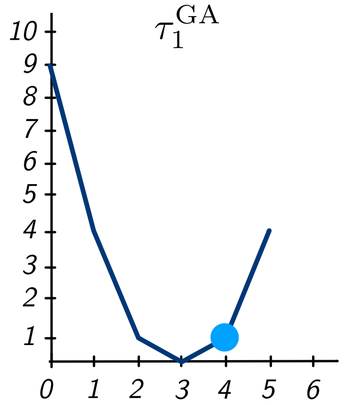
$$\text{s.t.} \sum_{c\in ch(r)} x_c = v \quad \text{(d3)}$$

$$x_c \in D^c \quad \forall c \in ch(r) \quad \text{(d4)}$$

τ_1^{GA}

j	cost
0	9
1	4
2	1
3	0
4	1
5	4

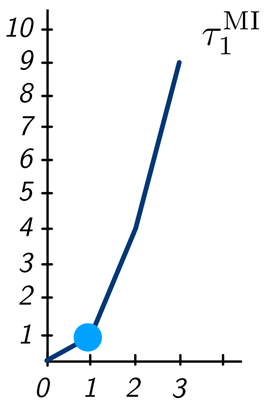
$\tau^{GA}(4) - \tau^{GA}(3) = 1$



τ_1^{MI}

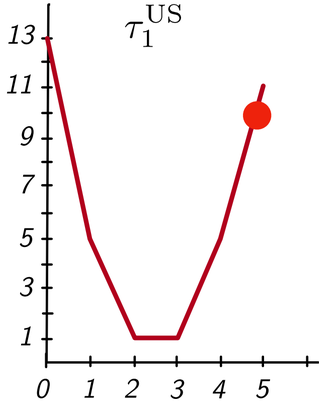
j	cost
0	0
1	1
2	4
3	9

$\tau^{MI}(2) - \tau^{MI}(1) = 3$



τ_1^{US}

j	cost
0	4+(0+0)
1	1+(0+4)
2	0+(0+1)
3	1+(0+0)
4	4+(1+0)
5	9+(1+1)



Corollary: The cost table function of each node of the tree is CPWL

Experimental Evaluation

Mechanisms & Dataset

Accuracy: Average L1 difference of the privacy-preserving data VS the original one

Runtime : in seconds

Mechanisms

- $\mathscr{M}_H$: The QIP mechanism
- $\mathscr{M}_{dp}^{OP}$: The dynamic-programming mechanism
- $\mathscr{M}_H^{dp}$: The polynomial-time dynamic programming counterpart

(extensions)

- $\mathscr{M}_c$: The QIP mechanism
- $\mathscr{M}_c^{dp}$: The polynomial-time dynamic programming counterpart

Datasets

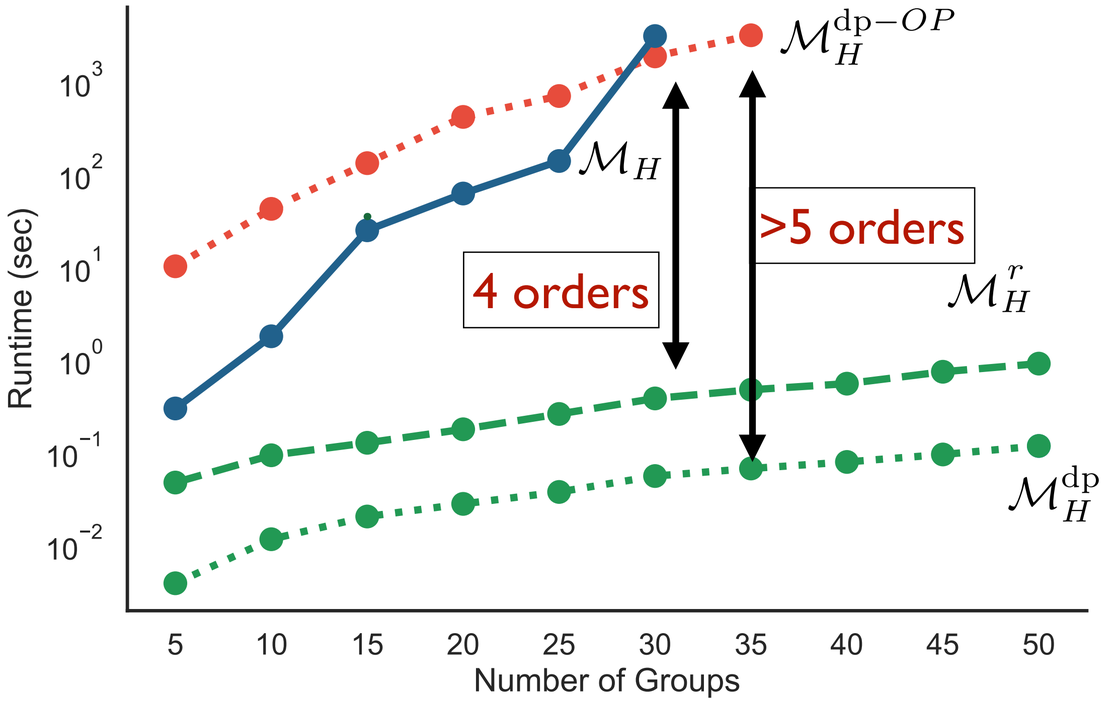
	US Census	NYC Taxi
Individuals	305,276,358	24,489,743
#Groups	117,630,445	13,282
Levels	3	3
#Leaves	7,592	3,973
N	1,000	13,282

Experimental Evaluation

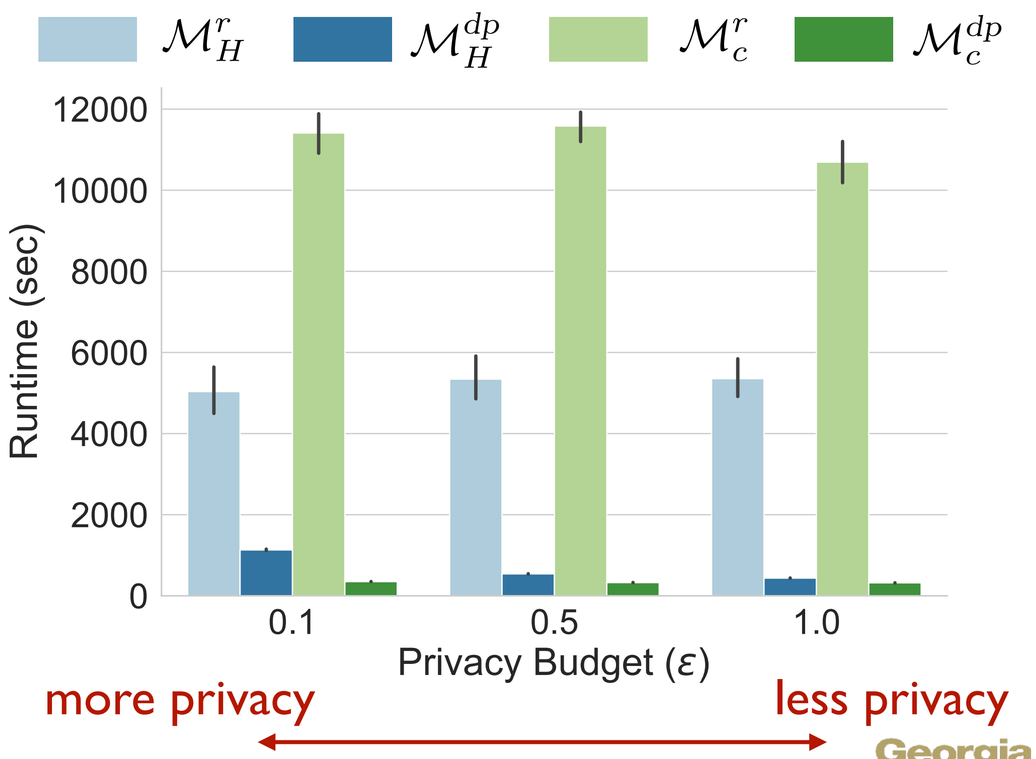
Runtime

Main Result: Exploiting convexity and structure of the problem provides up to **2 order magnitude improvements** w.r.t. the relaxed QIP method

Rule out the "naive" and QIP options



US Census Dataset

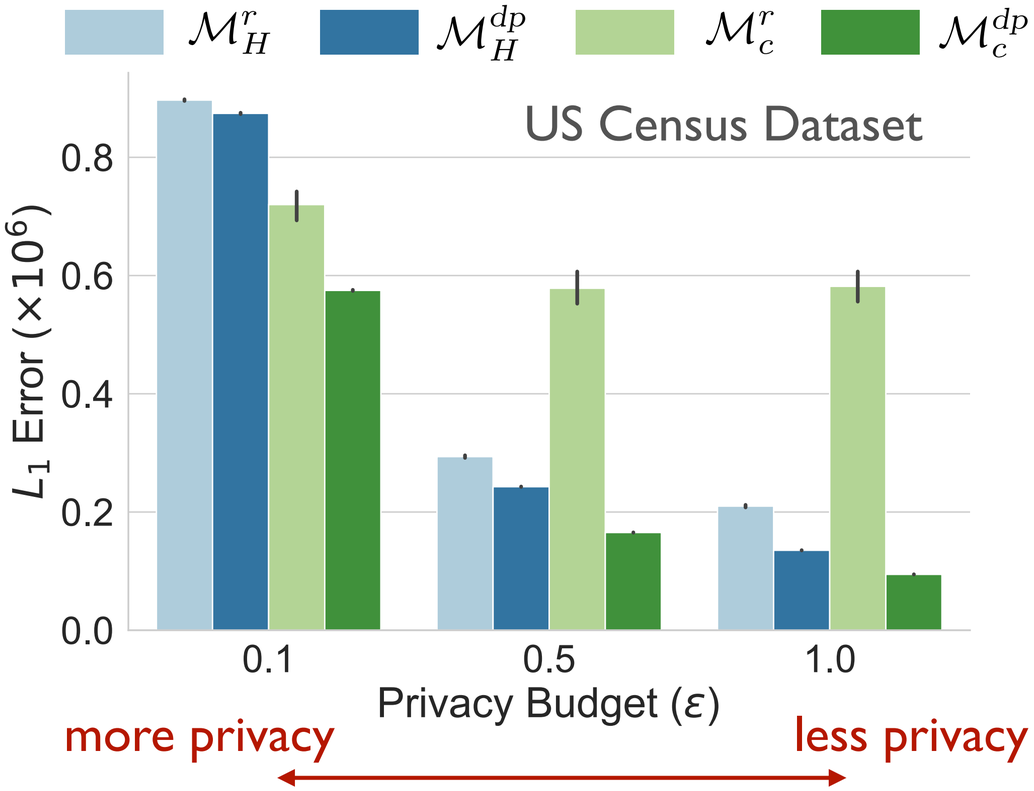


Georgia Tech

Experimental Evaluation

Accuracy

Main Result: The CPWL-dynamic programming versions produces **less errors** than the QP counterparts **violating no constraints**



		Taxi Data				Census Data			
		L_1 Errors ($\times 10^4$)			#CV	L_1 Errors ($\times 10^3$)			#CV
ϵ	**Alg**	Lev 1	Lev 2	Lev 3		Lev 1	Lev 2	Lev 3	
0.1	$\mathcal{M}_H^r$	25.4	158.7	904.4	18206	40.3	54.3	802.1	1966
	$\mathcal{M}_H^{dp}$	26.6	121.9	915.7	**0**	10.3	38.4	825.4	**0**
	$\mathcal{M}_c^r$	47.9	153.2	**551.6**	19460	23.1	64.5	632.2	1715
	$\mathcal{M}_c^{dp}$	**19.9**	**65.6**	644.3	**0**	**0.9**	**23.2**	**550.6**	**0**
0.5	$\mathcal{M}_H^r$	8.6	81.2	364.2	18591	39.4	37.9	216.3	1990
	$\mathcal{M}_H^{dp}$	5.5	31.0	408.9	**0**	2.4	9.4	230.8	**0**
	$\mathcal{M}_c^r$	46.7	153.5	450.7	19531	23.1	61.0	494.2	1718
	$\mathcal{M}_c^{dp}$	**4.0**	**16.4**	**352.9**	**0**	**0.2**	**5.8**	**159.1**	**0**
1.0	$\mathcal{M}_H^r$	7.7	77.2	**279.0**	18085	40.7	39.2	130.0	1989
	$\mathcal{M}_H^{dp}$	3.1	19.8	328.5	**0**	1.2	5.1	128.8	**0**
	$\mathcal{M}_c^r$	47.1	154.2	447.1	19706	24.1	63.0	494.5	1728
	$\mathcal{M}_c^{dp}$	**2.0**	**8.7**	307.8	**0**	**0.1**	**3.2**	**91.0**	**0**

Georgia Tech

CP4DP

How can CP help Differential Privacy?

- DP research - Largely dominated by theoretical results
- One of the longstanding difficulties of DP is releasing privacy-preserving data that satisfies constraints derived from the data (i.e., hierarchical datasets) or from a problem of interest (i.e., an optimization task)
- CP, instead, thrives with the presence of constraints!
- We believe that the CP community can contribute to

  1. Improve the applicability of DP for problems in which constraints arise
  2. Transfer the Programming Language expertise to generalize and simplify the use of DP algorithms in many domains of interest.

Conclusions

- The release of datasets containing sensitive information is central to a number of statistical analysis and machine learning tasks.
- Hierarchical datasets are of particular interest (e.g., those used by the US Census Bureau).
- Motivated by the DP challenges in guaranteeing the satisfaction of constraints arising from the data and by the dataset size.
- We proposed an efficient **constrained optimization-based** method to satisfy the problem constraints while preserving the desired privacy guarantees.
- Experiments on *very* large databases show the effectiveness and efficiency of our proposal.

Conclusions

- The release of datasets containing sensitive information is central to a number of statistical analysis and machine learning tasks.
- Hierarchical datasets are of particular interest (e.g., those used by the US Census Bureau).
- Motivated by the DP challenges in guaranteeing the satisfaction of constraints arising from the data and by the dataset size.
- We proposed an efficient **constrained optimization-based** method to satisfy the problem constraints while preserving the desired privacy guarantees.
- Experiments on *very* large databases show the effectiveness and efficiency of our proposal.

Thank you!

Ferdinando Fioretto
fioretto@gatech.edu

Pascal Van Hentenryck
pvh@isye.gatech.edu

References

[McKenna:18] McKenna, Laura. “Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing”, Working Papers 18-47, Center for Economic Studies, U.S. Census Bureau, Handle: RePEc:cen:wpaper:18-47, 2018

[Ramachandran:12] Ramachandran, Aditi, Lisa Singh, Edward Porter, and Frank Nagle. “Exploring Re-Identification Risks in Public Domains”, IEEE Conference on Privacy, Security and Trust. 2012

[Irit:03] Irit Dinur and Kobbi Nissim. Revealing Information while Preserving Privacy. PODS 2013