

PRIVACY-PRESERVING POLICY SYNTHESIS IN MARKOV DECISION PROCESSES

PARHAM GOHARI

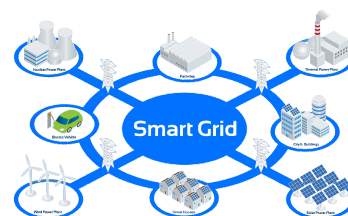
Joint work with: Bo Wu, Ufuk Topcu and Matthew Hale (University of Florida)

Privacy: A challenge in autonomous systems

- Decision-makers often collect sensitive data from the network members.
- Examples:
 - Autonomous driving
 - Smart power grids
 - Smart homes



Escript.com



subpng.com



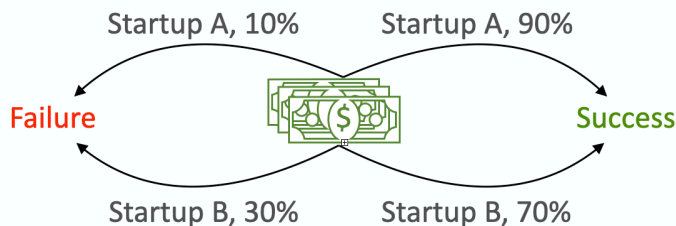
getvera.com

Environment dynamics as the sensitive data

- The environment dynamics reflect our model of the environment.



- Information regarding the environment dynamics may have \$\$\$ values!
- Example: Business firms must keep their market research data private from their competitors.



Privacy attacks on environment dynamics

- Various privacy attacks have been studied in reinforcement learning, for example:
 - On experience data for MC methods
 - On the underlying reward system
- A recent privacy attack infers the floor plans by observing the agent's actions with 95% precision [1].



[1] X. Pan *et al.*, "How you act tells a lot: privacy-leakage attack on deep reinforcement learning," [arXiv](#), 2019.

Challenge 1.

We need privacy in decision-making problems.

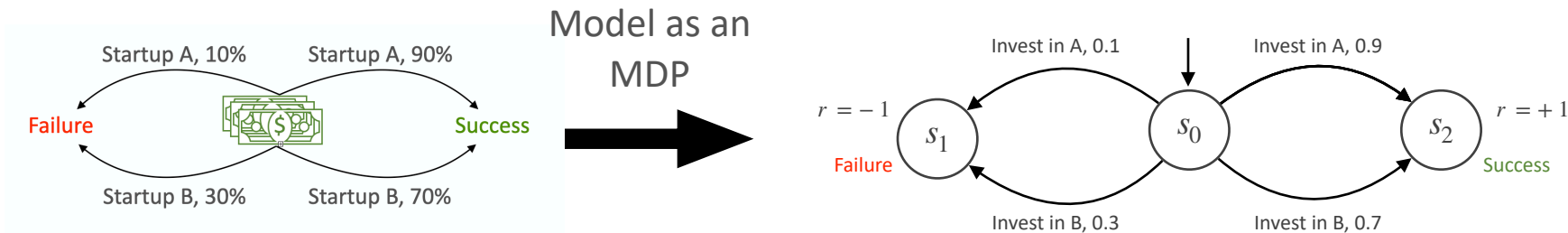
Challenge 2.

Sensitive data is the environment dynamics.

Challenge 3.

The actions must preserve the privacy of the environment dynamics.

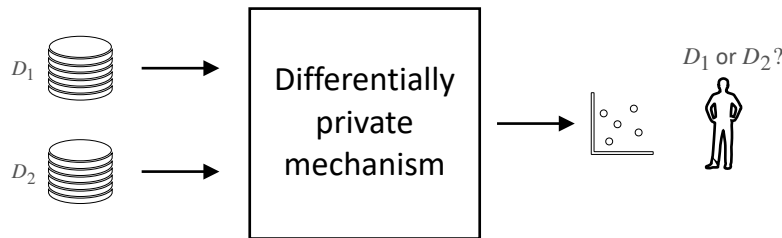
Markov Decision Processes (MDPs)



- Find a reward-maximizing policy based on the transition probabilities (policy synthesis).
- The policy *Invest in A* at s_0 reveals $\mathbb{P}[\text{success} \mid \text{invest in A}] \geq \mathbb{P}[\text{success} \mid \text{invest in B}]$

Differential privacy as the underlying privacy definition

- The intuition:



- Why differential privacy?
 - A well-defined quantitative definition
 - Immunity to **post-processing**
 - Robustness to **side information**

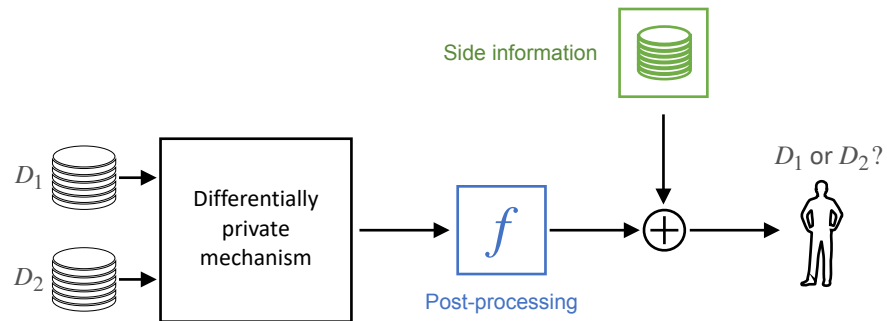
Differential privacy as the underlying privacy definition

- The intuition:



- Why differential privacy?

- A well-defined quantitative definition
- Immunity to **post-processing**
- Robustness to **side information**



Problem Statement

Find a policy synthesis algorithm that preserves the privacy of the transition probabilities, in the sense of differential privacy.

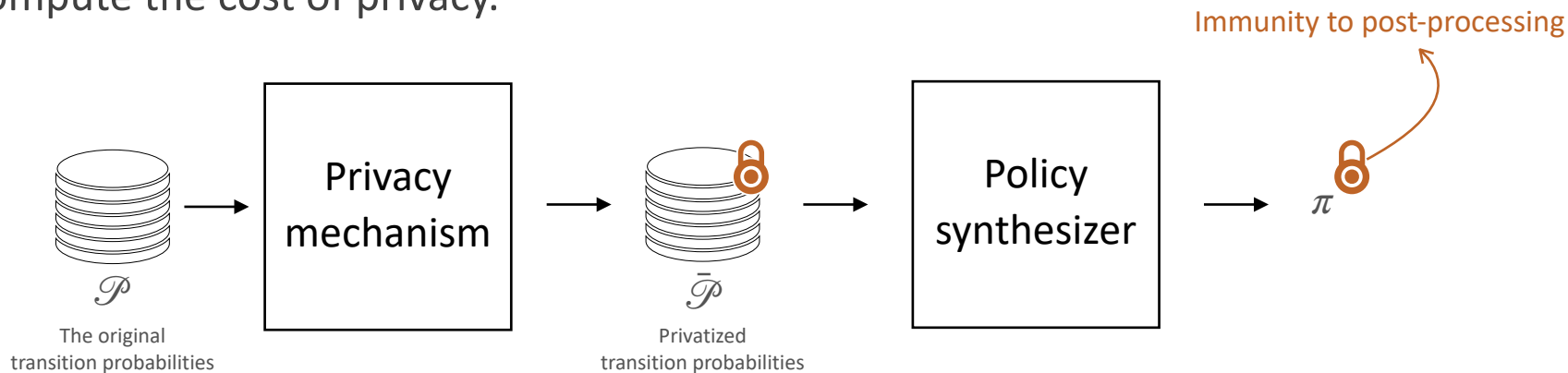
Question 1. How do we enforce differential privacy?

Question 2. How does privacy affect the optimality of the policy?

Question 3. Can we bound the suboptimality of the private policy?

What is our approach?

- Privatize the transition probabilities first.
- Synthesize a policy using the privatized transition probabilities.
- Compute the cost of privacy.



The Dirichlet mechanism

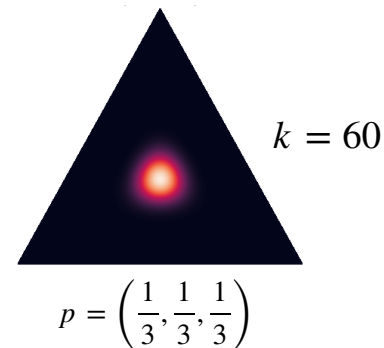
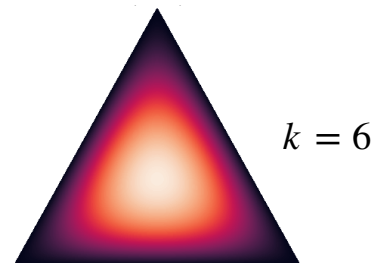
- A probabilistic mapping from $\Delta(n)$ to $\Delta(n)$ using the Dirichlet distribution.

- $\text{Dir}_k(p) = x$ with probability $\frac{1}{\text{B}(kp)} \prod_{i=1}^n x_i^{kp_i-1}$.

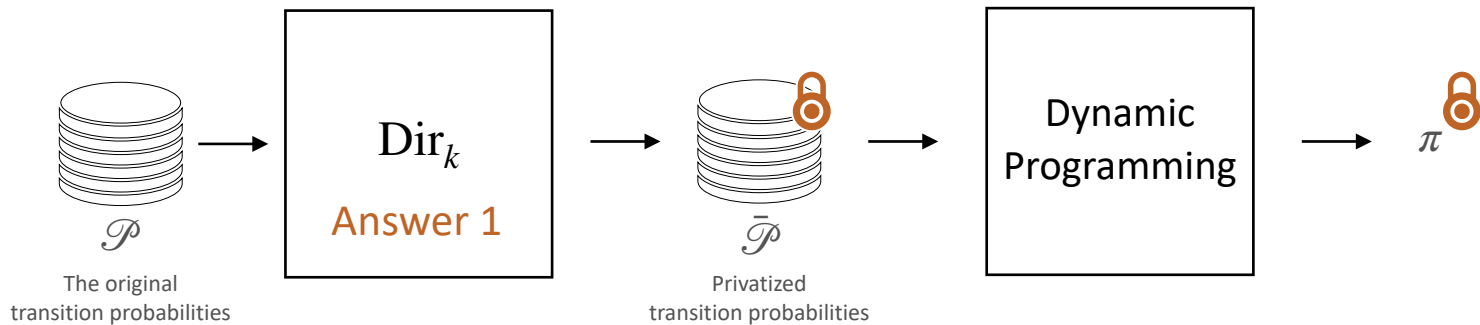
the multivariate beta function

- $\text{B}(kp) := \frac{\prod_{i=1}^n \Gamma(kp_i)}{\Gamma\left(k \sum_{i=1}^n p_i\right)}$ is the normalizing coefficient

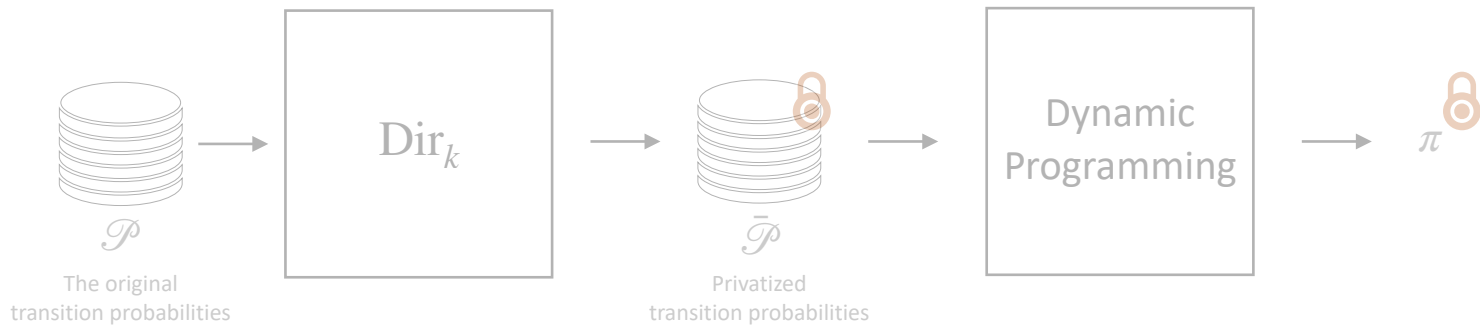
- The Dirichlet mechanism satisfies (ϵ, δ) -differential privacy [2].



Question 1. How do we enforce differential privacy?

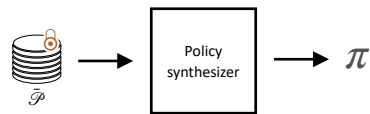


Question 1. How do we enforce differential privacy?



Question 2. How does privacy affect the optimality of the policy?

Dynamic programming



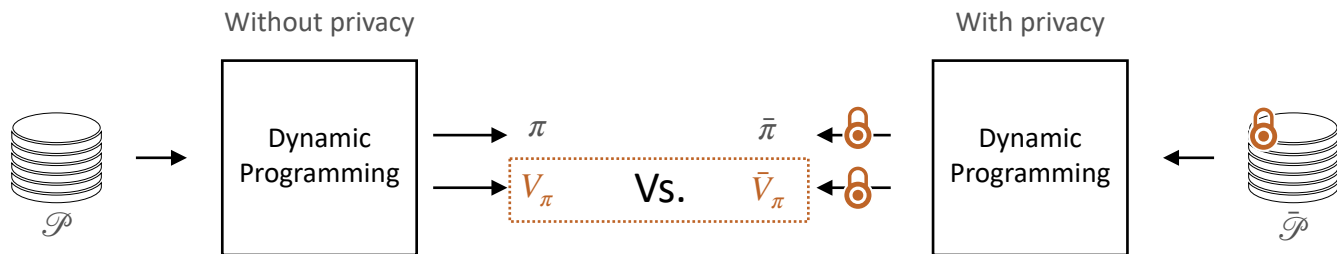
- Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, T)$ be the MDP representation of the environment.
- A policy $\pi : \mathcal{S} \mapsto \Delta(n)$ determines the decision rule at each state.
- The value of a policy π at state s is $V_t^\pi(s) = \mathbf{E} \left[\sum_{i=t}^T \gamma^{i-t} r_i \mid s_t = s \right]$.
- The optimal value and policy satisfy the Bellman condition of optimality:

$$V_t^*(s) = \max_{\pi} \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s, a, s') V_{t+1}^*(s') \right),$$

$$\pi_t^* \in \arg \max_{\pi} \sum_{s \in \mathcal{S}} \pi(a|s) \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s, a, s') V_{t+1}^*(s') \right).$$

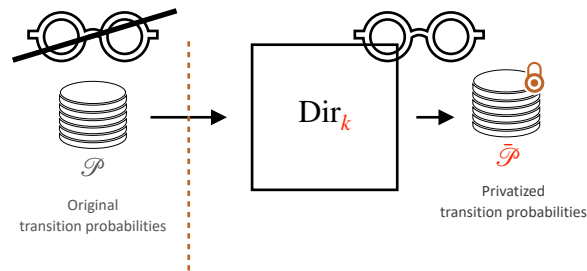
Cost of Privacy

- Captures the difference between the value function with and without privacy.



- Is **not** based on the sensitive data.

- Cost of privacy = $\mathbf{E} \left[\bar{V}_\pi - V_\pi \mid \bar{\mathcal{P}}, k \right]$



Main lemma: A concentration bound on Dir_k

- It is the first step in bounding the cost of privacy.

Lemma 1. (CONCENTRATION BOUND) [3]

For all $\beta > 0$ and $p \in \Delta(n)$, with probability at least $1 - \beta$,

$$\| \text{Dir}_k(p) - p \|_\infty \leq \underbrace{\sqrt{\frac{\log(1/\beta)}{2(k+1)}}}_{\alpha}.$$

- The concentration is explicitly affected by k .
- Large $k \rightarrow$ **higher** accuracy, however, **weaker** privacy protections.

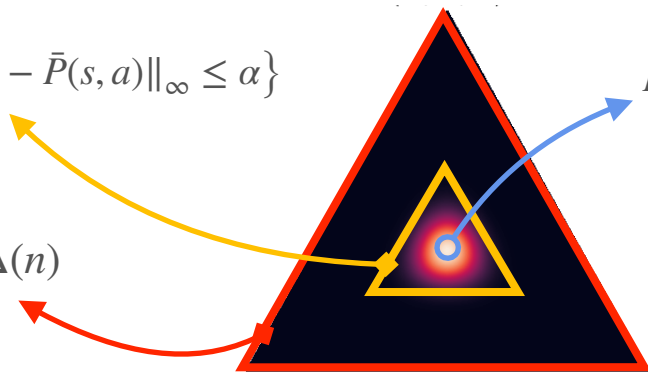
The set $\hat{\mathcal{P}}_{\alpha, \beta}$

- $\hat{\mathcal{P}}_{\alpha, \beta}(s, a)$ determines an estimation of $P(s, a)$ given $\bar{P}(s, a)$.

$$\mathcal{P}_1 := \{p \in \Delta(n) \mid \|p - \bar{P}(s, a)\|_{\infty} \leq \alpha\}$$

$$\mathcal{P}_2 := \Delta(n)$$

$$\bar{P}(s, a) = \text{Dir}_k(P(s, a))$$



- $\hat{\mathcal{P}}_{\alpha, \beta}(s, a) \rightarrow$ the set of all β -convex combinations of \mathcal{P}_1 and \mathcal{P}_2 ($\beta P_1 + (1 - \beta)P_2$).
- We show that: $\mathbf{E}[P(s, a) \mid \bar{P}(s, a), k] \in \mathcal{P}_{\alpha, \beta}(s, a)$ and $\bar{P}(s, a) \in \mathcal{P}_{\alpha, \beta}(s, a)$

Main result 1:

Theorem 1. (COST OF PRIVACY IN FINITE-HORIZON MDPS) [3]

Let $\bar{\pi}$ denote the policy with privacy protections. Define

$$\underline{v}_t^{\bar{\pi}}(s) := \sum_{a \in \mathcal{A}_s} \bar{\pi}(a | s) \left(r(s, a) + \gamma \min_{p \in \hat{\mathcal{P}}_{\alpha, \beta}(s, a)} \sum_{s' \in \mathcal{S}} p(s, a, s') \underline{v}_{t+1}^{\bar{\pi}}(s') \right),$$

$$\bar{v}_t^{\bar{\pi}}(s) := \sum_{a \in \mathcal{A}_s} \bar{\pi}(a | s) \left(r(s, a) + \gamma \max_{p \in \hat{\mathcal{P}}_{\alpha, \beta}(s, a)} \sum_{s' \in \mathcal{S}} p(s, a, s') \bar{v}_{t+1}^{\bar{\pi}}(s') \right).$$

Then,

$$\left| \text{cost of privacy} \right| \leq \bar{v}_t^{\bar{\pi}}(s) - \underline{v}_t^{\bar{\pi}}(s)$$

Finite horizon vs. infinite horizon

- The value of a policy π at state s is $V_{\infty}^{\pi}(s) = \mathbf{E} \left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i \mid s_t = s \right]$.
- The optimal value and policy satisfy the Bellman condition of optimality:

$$V^*(s) = \max_{\pi} \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s, a, s') V_{t+1}^*(s') \right),$$

$$\pi^* \in \arg \max_{\pi} \sum_{s \in \mathcal{A}} \pi(a|s) \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s, a, s') V_{t+1}^*(s') \right).$$

Main result 2:

Theorem 1. (COST OF PRIVACY IN INFINITE-HORIZON MDPS) [3]

Let $\underline{v}_\infty^{\bar{\pi}}$ and $\bar{v}_\infty^{\bar{\pi}}$ satisfy

$$\underline{v}_\infty^{\bar{\pi}}(s) = \sum_{a \in \mathcal{A}_s} \bar{\pi}(a | s) \left(r(s, a) + \gamma \min_{p \in \hat{\mathcal{P}}_{\alpha, \beta}(s, a)} \sum_{s' \in \mathcal{S}} p(s, a, s') \underline{v}_\infty^{\bar{\pi}}(s') \right),$$

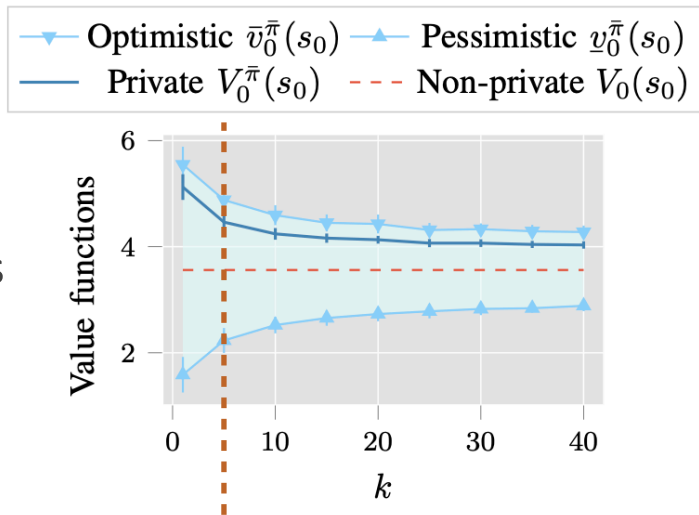
$$\bar{v}_\infty^{\bar{\pi}}(s) = \sum_{a \in \mathcal{A}_s} \bar{\pi}(a | s) \left(r(s, a) + \gamma \max_{p \in \hat{\mathcal{P}}_{\alpha, \beta}(s, a)} \sum_{s' \in \mathcal{S}} p(s, a, s') \bar{v}_\infty^{\bar{\pi}}(s') \right).$$

Then,

$$|\text{cost of privacy}| \leq \bar{v}_\infty^{\bar{\pi}}(s) - \underline{v}_\infty^{\bar{\pi}}(s)$$

Numerical Results

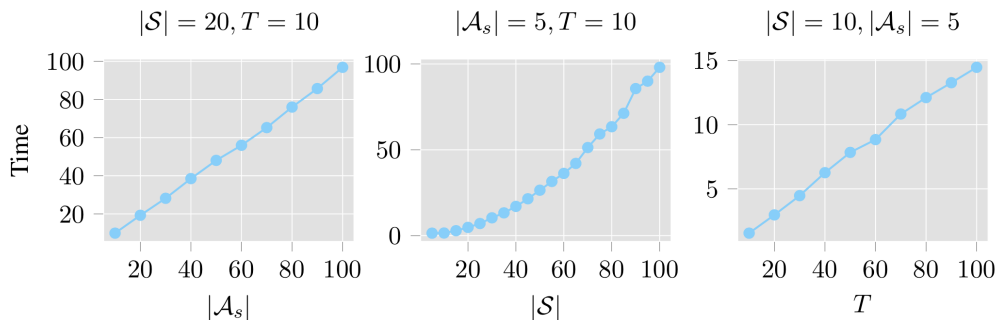
- We consider a 30-state 10-action MDP with random transition probabilities and rewards.
- Observe that an increase in k results in a lower cost of privacy.



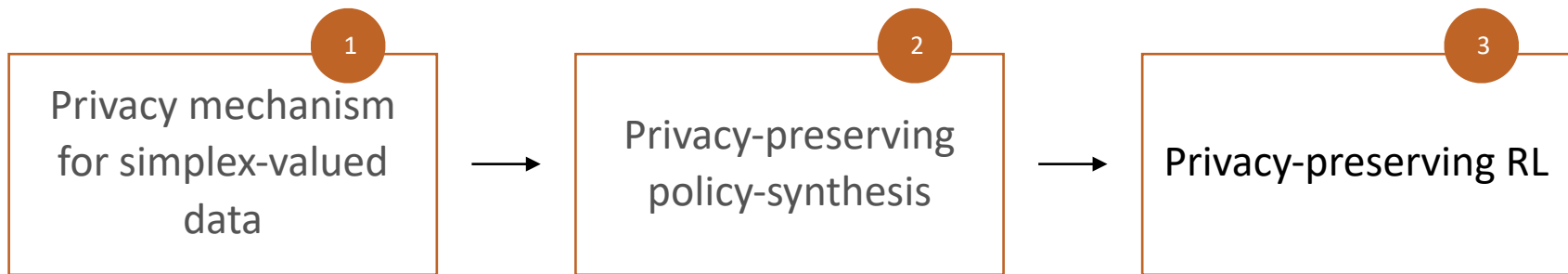
At $k = 5$:
 (2,0.02)-differentially privacy

Numerical Results: Computational complexity

	Computational complexity
Finite horizon	$\mathcal{O}(T \mathcal{S} ^{4.5} \mathcal{A})$
Infinite horizon	$\mathcal{O}(\mathcal{S} ^{4.5} \mathcal{A} \log(1/\eta))$



What's next?



Key takeaway

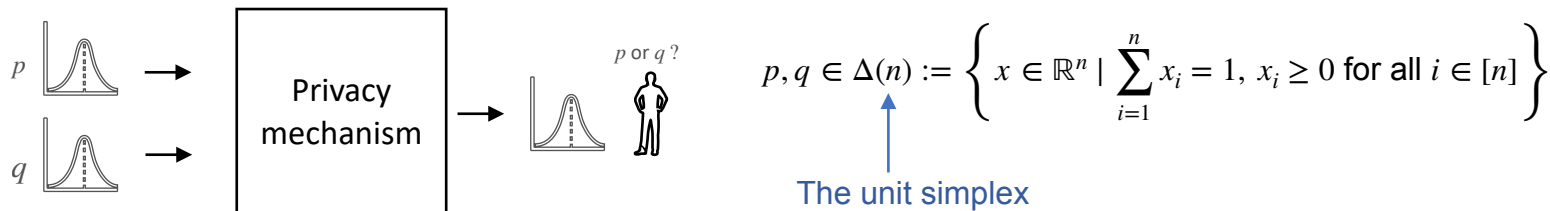
Using the Dirichlet mechanism, we found a differentially private policy-synthesis algorithm and we bounded the cost of privacy.

Thank you for your attention.

My email: pgohari@utexas.edu


Adjacency relationship

- The output of *similar* datasets must be approximately indistinguishable.
- Formally, similar datasets are defined by an adjacency relationship.



- Two vectors $p, q \in \Delta(n)$ are b -adjacent, denoted $p \stackrel{b}{\sim} q$, if there exist indices i, j such that

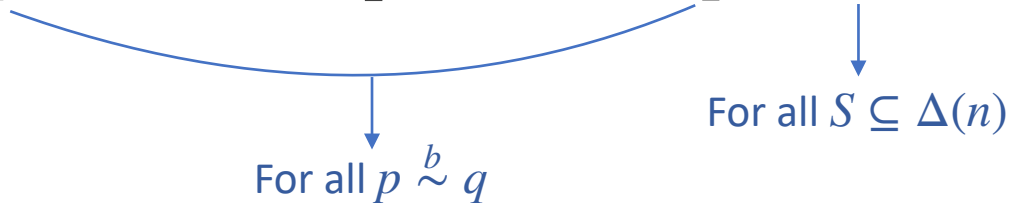
$$p_{-(i,j)} = q_{-(i,j)} \quad \text{and} \quad \|p - q\|_1 \leq b.$$


 A constant $\in (0, 1]$

Definition of differential privacy

- A mechanism M is (ϵ, δ) -differentially private if

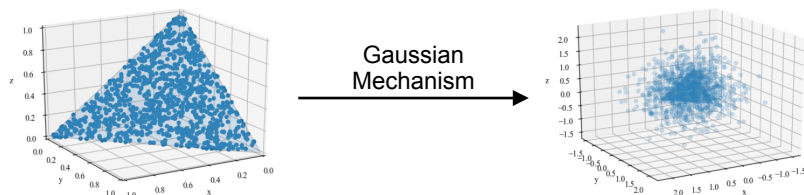
$$\mathbb{P}[M(p) \in S] \leq \exp(\epsilon) \cdot \mathbb{P}[M(q) \in S] + \delta.$$



- $\epsilon \rightarrow$ Level of privacy protections (typically $\epsilon \in [0, \log(3)]$).
- $\delta \rightarrow$ The probability of protection failure (typically $\delta \in [0, 0.1]$).

Why Dirichlet mechanism?

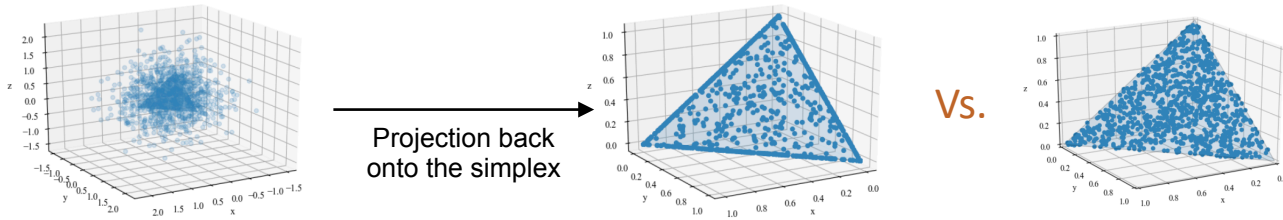
- Traditional methods add infinite-support noise to the entries of the dataset.
- Infinite-support noise breaks the special structure of the transition probabilities.



- Dynamic programming does not converge with transition probabilities outside the unit simplex.

Why projection is not a good idea?

- Projection back onto the simplex preserves differential privacy. However:



- Projection hurts the accuracy of the privacy mechanism.
- There is a need for a new privacy mechanism for simplex-valued data.