Approximate DBSCAN under Differential Privacy

Yuan Qiu and Ke Yi

yuan.qiu@cnrsatcreate.sg, yike@cse.ust.hk

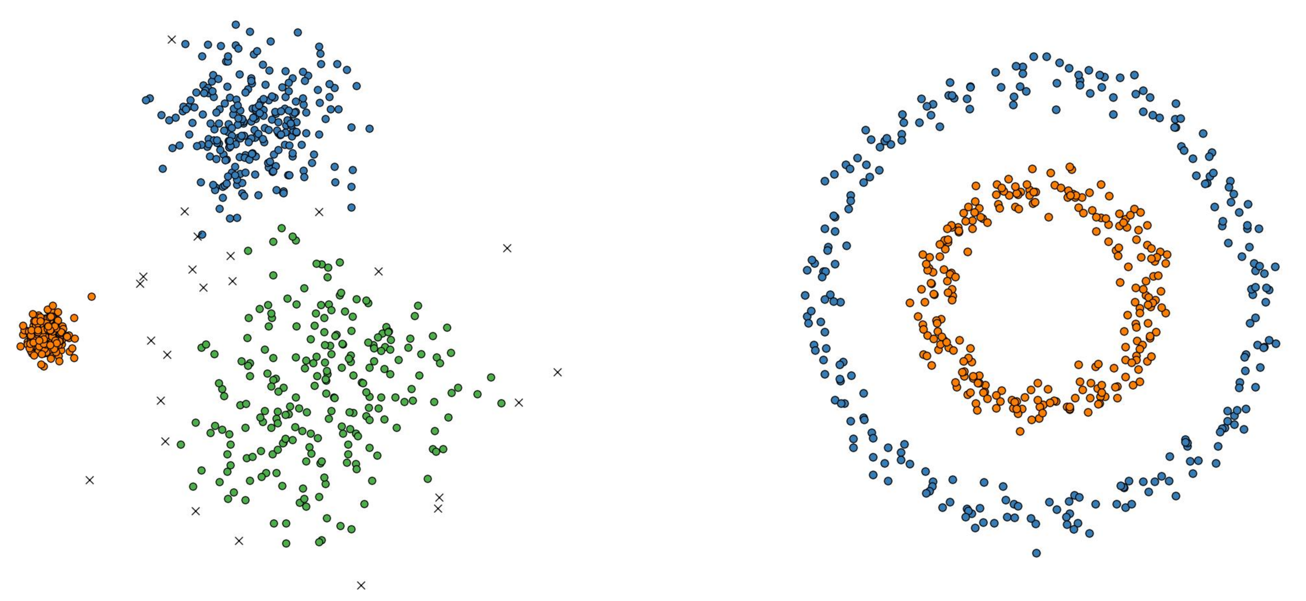
cnrs @CREATE Singapore

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Problem Definition

DBSCAN(α, MinPts):

- p is a **core point** if $B(p, \alpha)$ contains at leasat MinPts points
- p and q are **reachable** if dist(p, q) < α
- p and q are **connected** if they are directly or transitively reachable
- A **cluster** is a maximal set of mutually connected core points

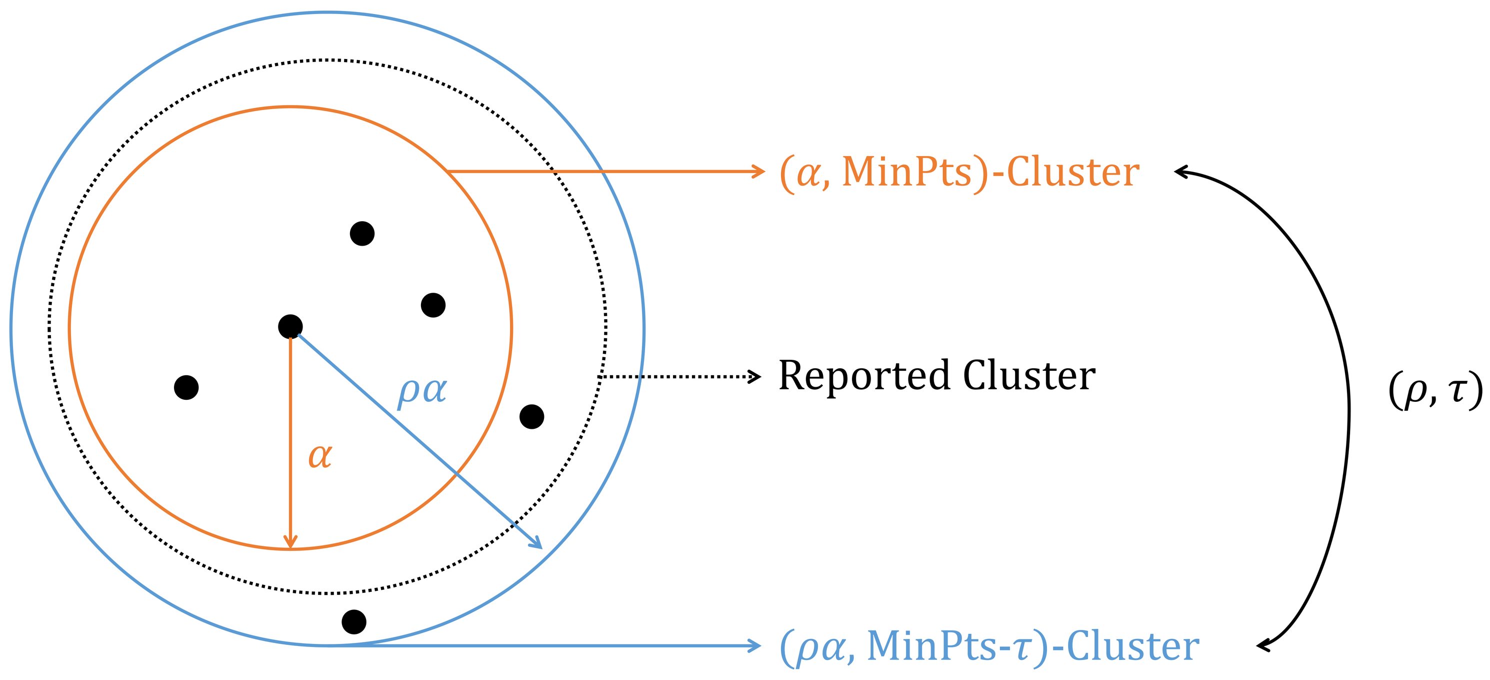


Differential Privacy(ε, δ):

- For any pair of neighboring datasets $P \sim P'$ and any subset of outputs $O \subseteq \mathcal{O}$, we should have

$$\Pr[\mathcal{M}(P) \in O] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(P) \in O] + \delta$$

Approximate DBSCAN [Gan and Tao '15]



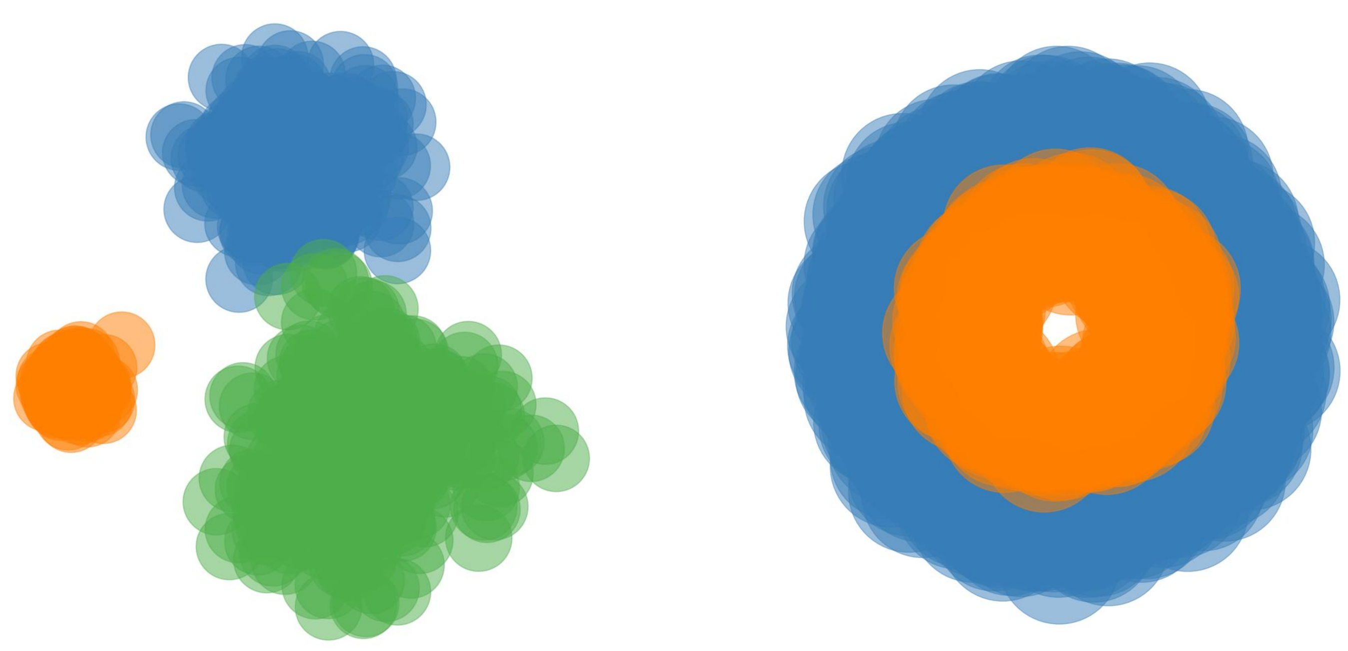
Approximate Cluster Spans

Negative Result 1:

- If an ε-DP mechanism (that outputs core points) is always (ρ, τ)-accurate with probability $1-\beta$, then $\beta \geq n/(n+e^{\varepsilon}) \approx 1$

New Approximation: Spans of Clusters

- Informally, the span is the union of α-neighborhood of core points

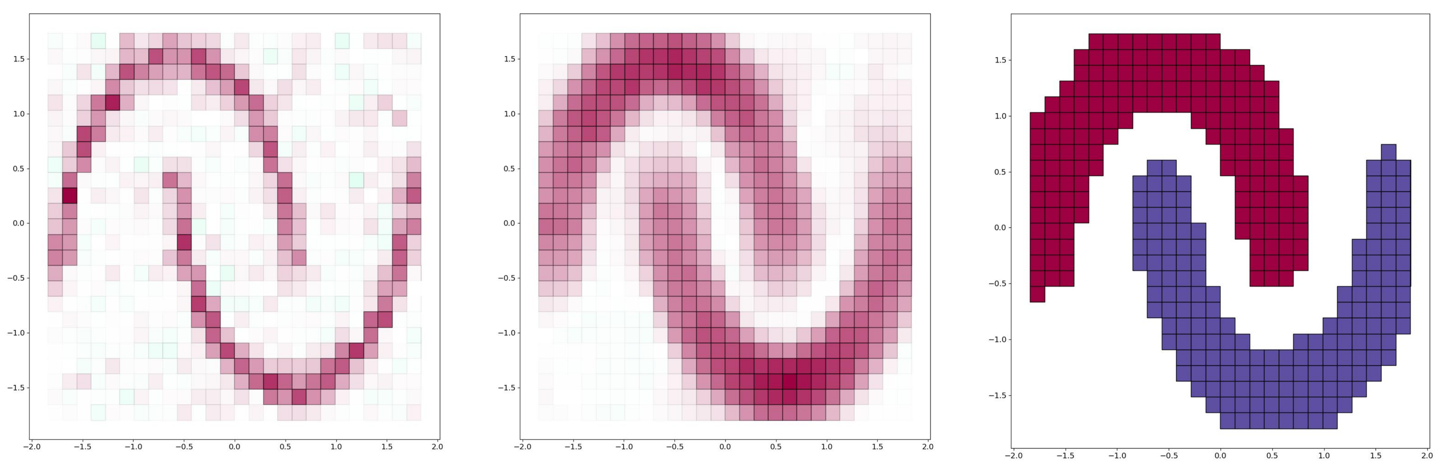


Negative Result 2 (Approximation lower bound):

- If an ε-DP mechanism (that outputs spans) is always (ρ, τ)-accurate with probability 0.9, then $\rho \geq 3$ and $\tau = \Omega(\frac{1}{\varepsilon}\log\frac{1}{\rho\alpha})$

DP Approximate DBSCAN

- Partition the space into cells of width $w \propto \alpha/\sqrt{d}$
- Release a DP histogram for the cell counts
- Post-process the histogram by computing neighbor sums and finding core cells
- Merge adjacent core cells and report approximate spans



Utility Guarantee (Approximation ratio upper bound):

- The DP-DBSCAN algorithm is (3+η, τ)-accurate for

$$\tau = O((1+\frac{8\sqrt{d}}{\eta})^{d} \cdot \frac{d}{\varepsilon}\log\frac{d}{\alpha\beta})$$

- For constant d and η, this matches the lower bound

Linear-Time Pure-DP Histsogram

- A naive histogram over universe X takes O(|X|) time
- We simulate a histogram that is equivalent to keeping only noisy frequencies above θ in a standard Laplace histogram:
  - For non-zero frequencies, add Laplace noise and keep if above θ
  - All the M zero-frequency entries share the same distribution:
    - Sample the number of non-zero entries $m \sim Bin(M, p)$
    - Sample m entries without replacement $J \subseteq X$
    - Sample m noises from the upper-tail of Laplace distribution

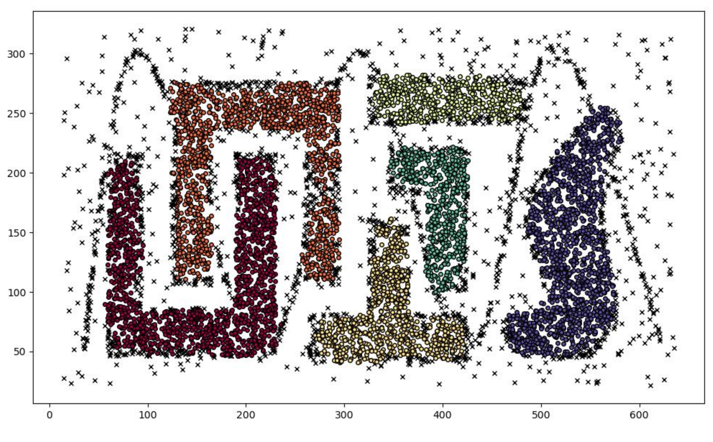
Complexity and Utility Guarantee:

- The histogram is ε-DP
- With high probability, it can be built in O(n) time and space
- Its simultaneous error is $O(\frac{1}{\varepsilon}\log|X|)$ for any entry

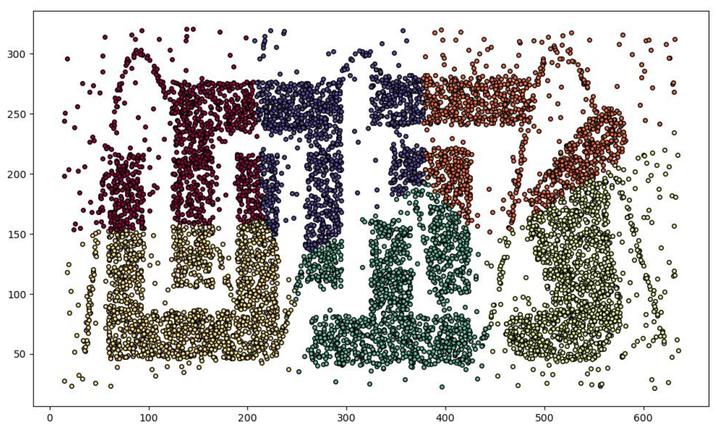
Comparison with Stability-based Histogram [Balcer and Vadhan '19]

- Both run in O(n) time
- Our histogram achieves pure-DP with error $O(\frac{1}{\varepsilon}\log|X|)$
- Existing work achieves approximate-DP with error $O(\frac{1}{\varepsilon}\log\frac{1}{\delta})$

Experiments



DP-DBSCAN



DP-KMeans

	Dataset	$\varepsilon = 1$ DP-DBSCAN	$\varepsilon = 1$ DP-Kmeans	$\varepsilon = \infty$ DBSCAN	$\varepsilon = \infty$ Kmeans
ARI	Circles	**0.94**	0.00	**0.98**	0.00
	Moons	**0.99**	0.51	**1.00**	0.47
	Blobs	**0.81**	0.79	0.55	**0.89**
	Cluto-t4	**0.64**	0.47	**0.95**	0.50
	Cluto-t5	**0.93**	0.69	**0.96**	0.78
	Cluto-t7	**0.52**	0.32	**0.76**	0.33
	HAR70+	**0.40**	0.19	**0.57**	0.23
AMI	Circles	**0.92**	0.00	**0.96**	0.00
	Moons	**0.99**	0.41	**1.00**	0.37
	Blobs	**0.83**	0.79	0.66	**0.87**
	Cluto-t4	**0.74**	0.59	**0.92**	0.61
	Cluto-t5	**0.92**	0.77	**0.95**	0.82
	Cluto-t7	**0.63**	0.54	**0.82**	0.56
	HAR70+	**0.45**	0.43	**0.54**	0.48