

Learning Partitions with Simple Queries

Minimizing Query Complexity, Adaptivity, and Size

EnCORE Collaboration Workshop
December 4, 2024

Hadley Black
UCSD



Arya Mazumdar
UCSD



Barna Saha
UCSD



Clustering via Crowdsourcing

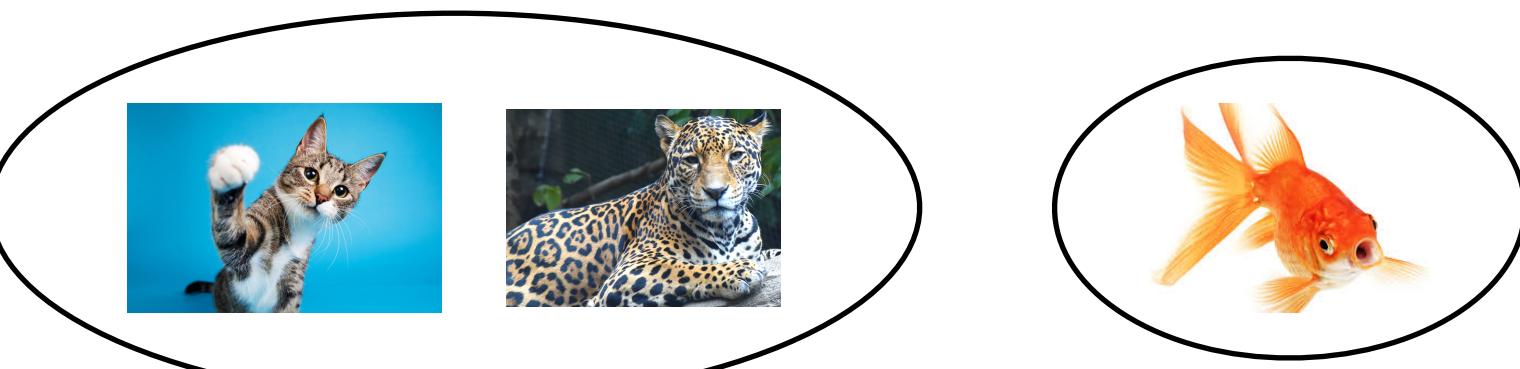
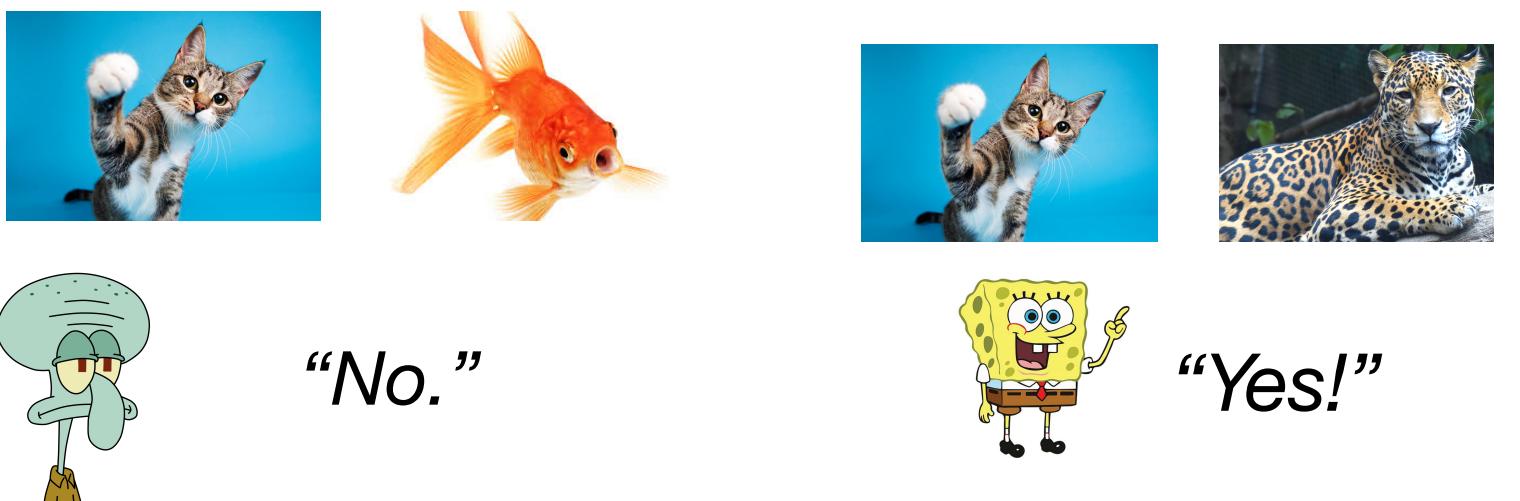
- Can we offload the work of computing a clustering by asking simple questions to external individuals?
- **Same-cluster queries:** Are these two points of the same type?

Wish list:

- (1) few queries
- (2) queries specified in few rounds

- Responses may be slow
- Want to parallelize queries

Query profile



Learned clustering

Learning Partitions with Pair Queries

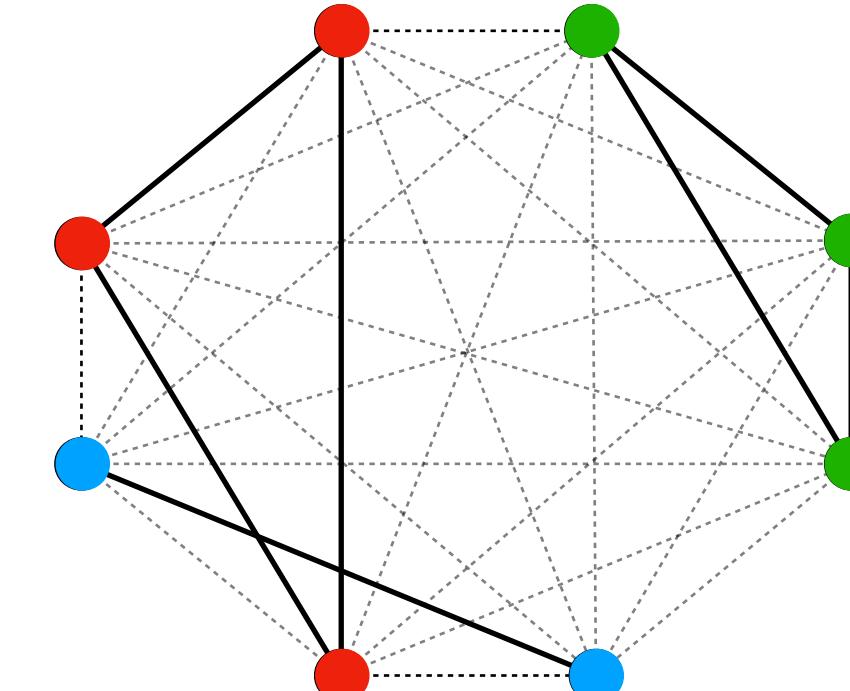
Mazumdar-Saha [NeurIPS 17], Mazumdar-Saha [AAAI 17], Mazumdar-Pal [NeurIPS 17], Mitzenmacher-Tsourakis [16], Saha-Subramanian [ESA 19], Pia-Ma-Tzamos [COLT 22], Bressan-Cesa-Bianchi-Lattanzi-Paudice [NeurIPS 20], Huleihal-Mazumdar-Médard-Pal [NeurIPS 19]

- Set U of n points
- Hidden k -partition $X_1 \sqcup \dots \sqcup X_k = U$
- How many pair queries to learn X_1, \dots, X_k exactly?

Non-adaptive

$$\Theta(n^2)$$

1 round



r rounds?

Theorem

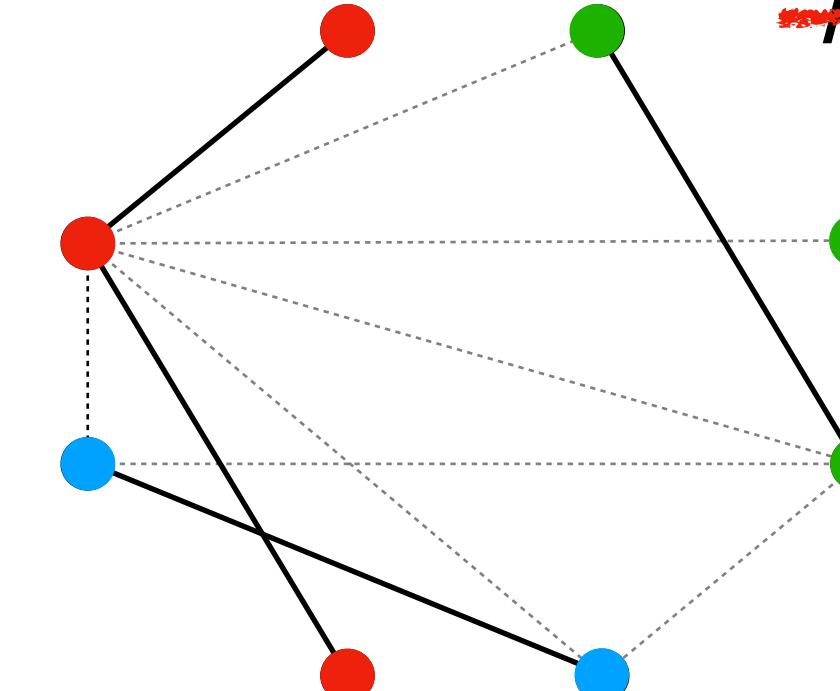
$$\Theta\left(n^{1+\frac{1}{2^r-1}} \cdot k^{1-\frac{1}{2^r-1}}\right)$$

*

* lower bound matches for $r = O(1)$

Adaptive

$$\Theta(nk)$$



$$O(\log \log n)$$

~~k rounds~~

Learning Partitions with **Subset** Queries

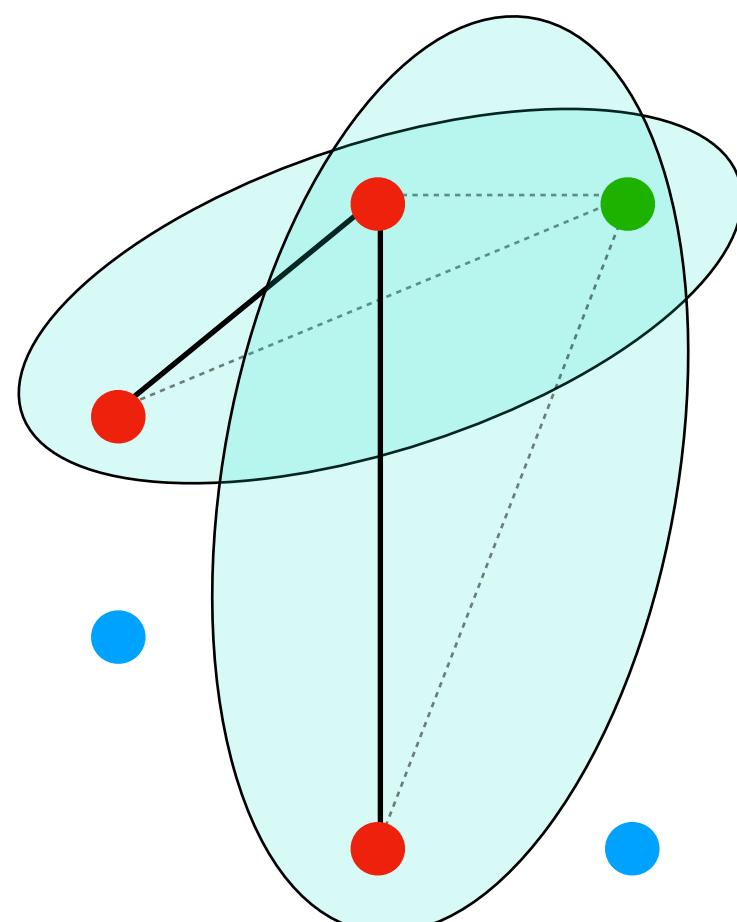
(Non-adaptively)

Chakrabarty-Liao [FSTTCS 24], Black-Lee-Mazumdar-Saha [NeurIPS 24]

- Set U of n points
- Hidden k -partition $X_1 \sqcup \dots \sqcup X_k = U$
- How many **subset queries of size at most s** to learn X_1, \dots, X_k exactly?

Strong

Returns full description of partition on S



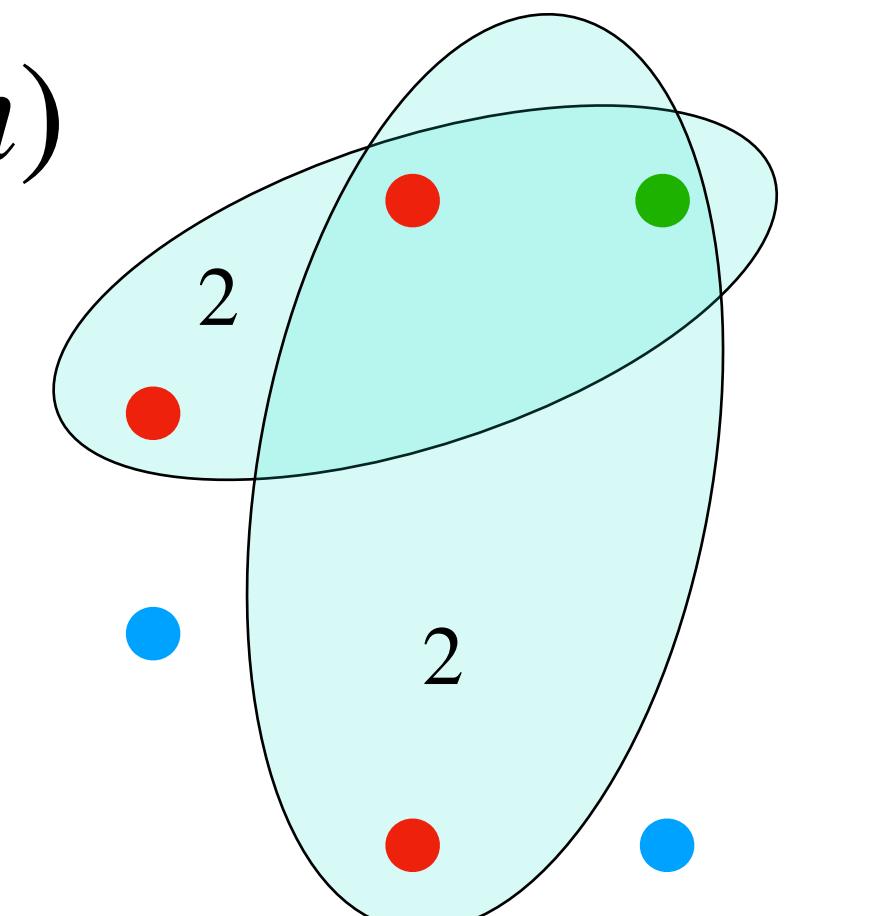
$$\Theta(n^2/s^2)$$

+ info theory

Weak

Returns # sets intersecting S

$$\Omega(\max(n^2/s^2), n)$$



Question

Is $O(n^2/s^2)$ weak queries possible for $s \leq \sqrt{n}$?

Theorem

$\widetilde{O}(n^2/s^2)$ non-adaptive weak queries for all $s \leq \sqrt{n}$