Differentially-private Sublinear-Time Clustering

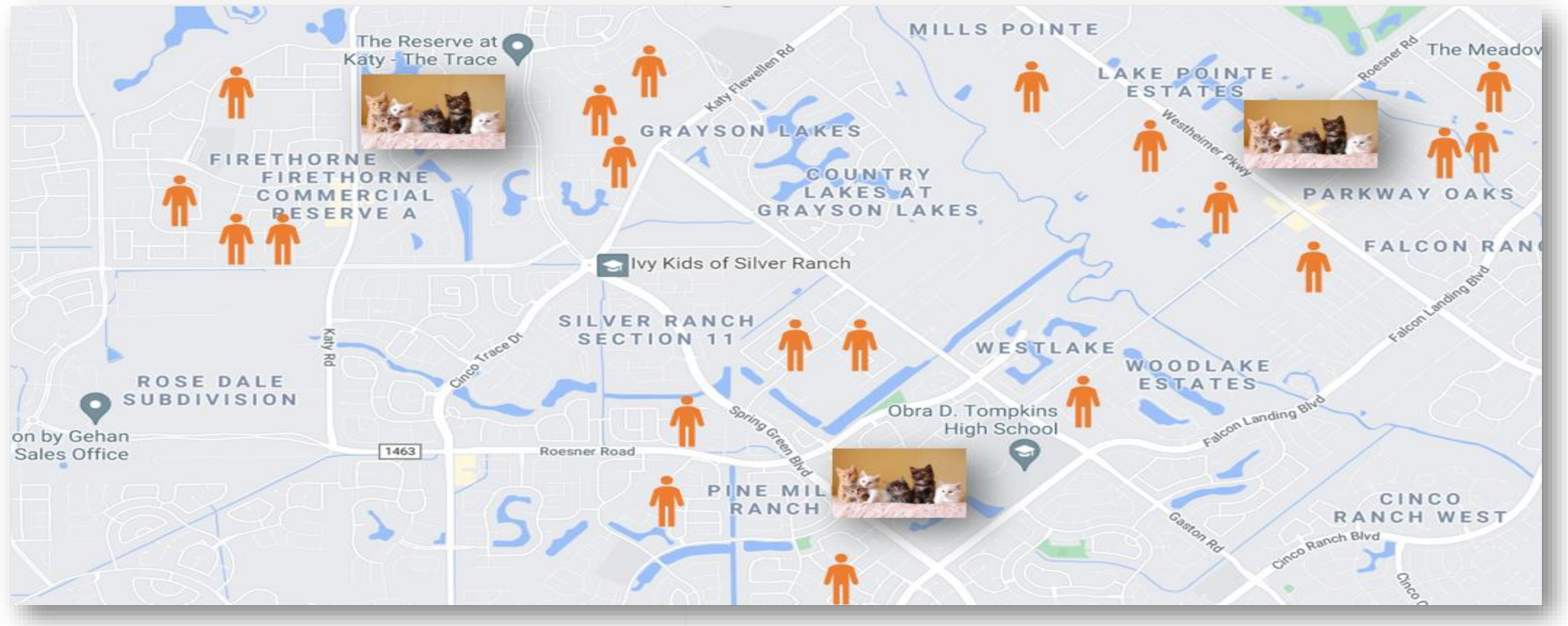
Jeremiah Blocki, Elena Grigorescu, Tamalika Mukherjee*

Department of Computer Science

Purdue University

DP Clustering Motivation

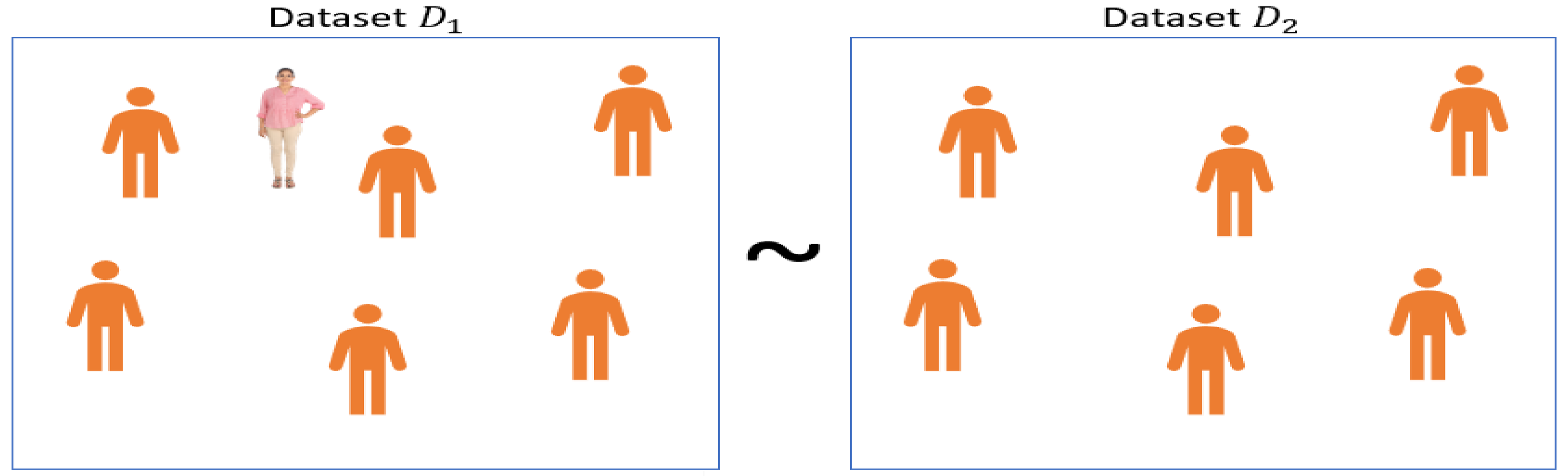
Cat Lovers Society wants to open some Cat Café centers close to its members.



K-median clustering. Input is set of member locations D, Output is cat cafes $c_1, c_2, \ldots, c_k$ such that $\sum_{x \in D} \min_i d(x, c_i)$

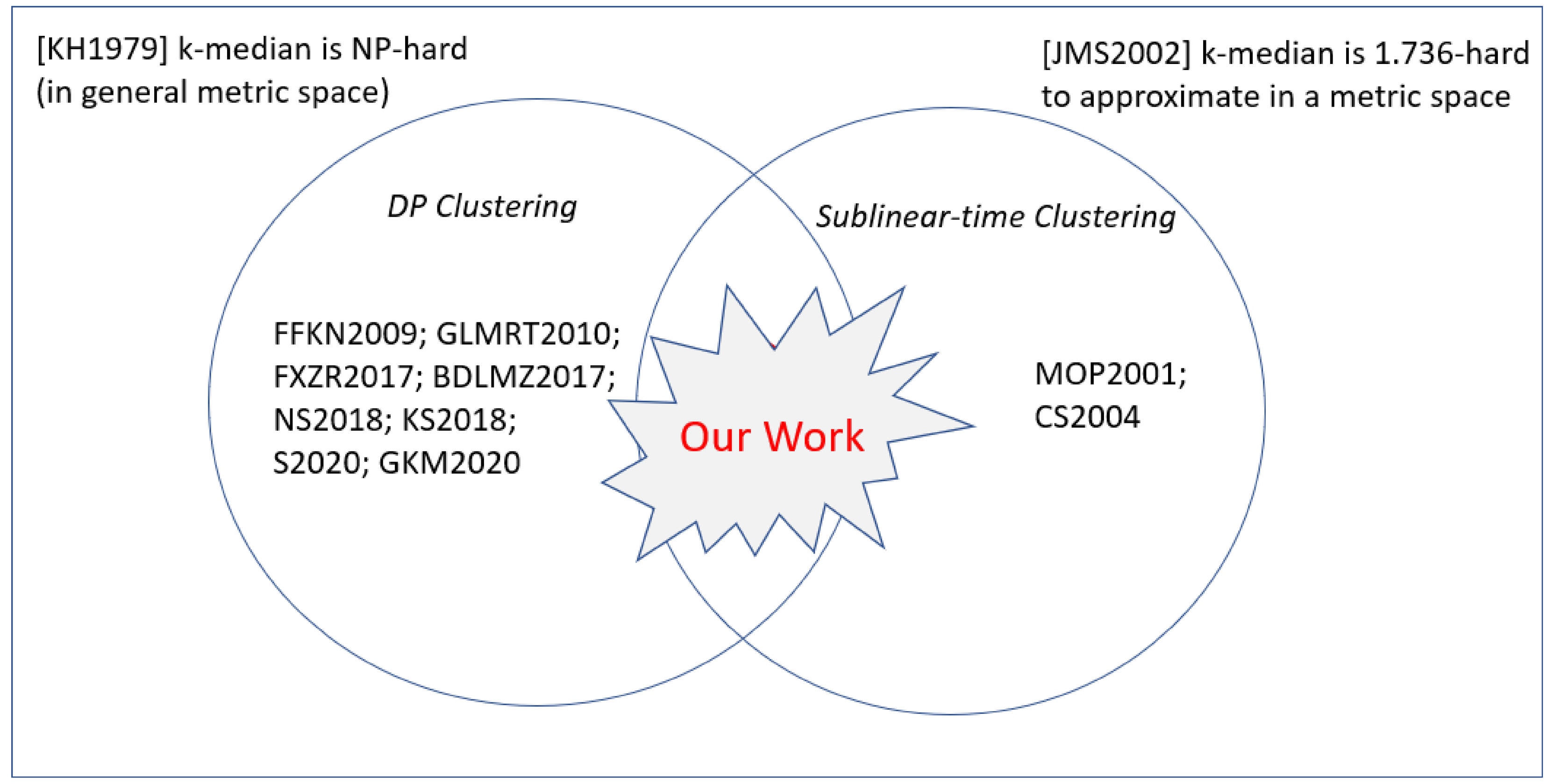
DP Motivation. Alice is a closet cat lover. Her partner Eve is a cat hater. Alice being a member of Cat Lovers Society is *sensitive information*.

DP Clustering

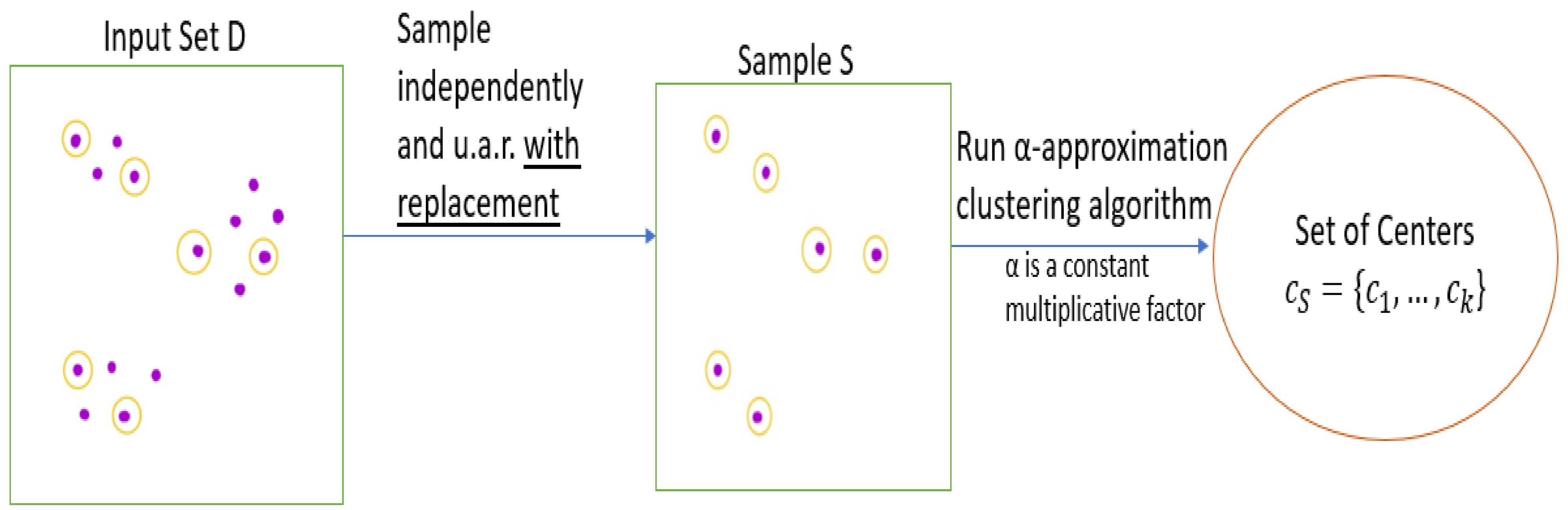


$\Pr[A(D_1) \in \ ] \leq e^{\varepsilon} \Pr[A(D_2) \in \ ] + \delta$

Related Work

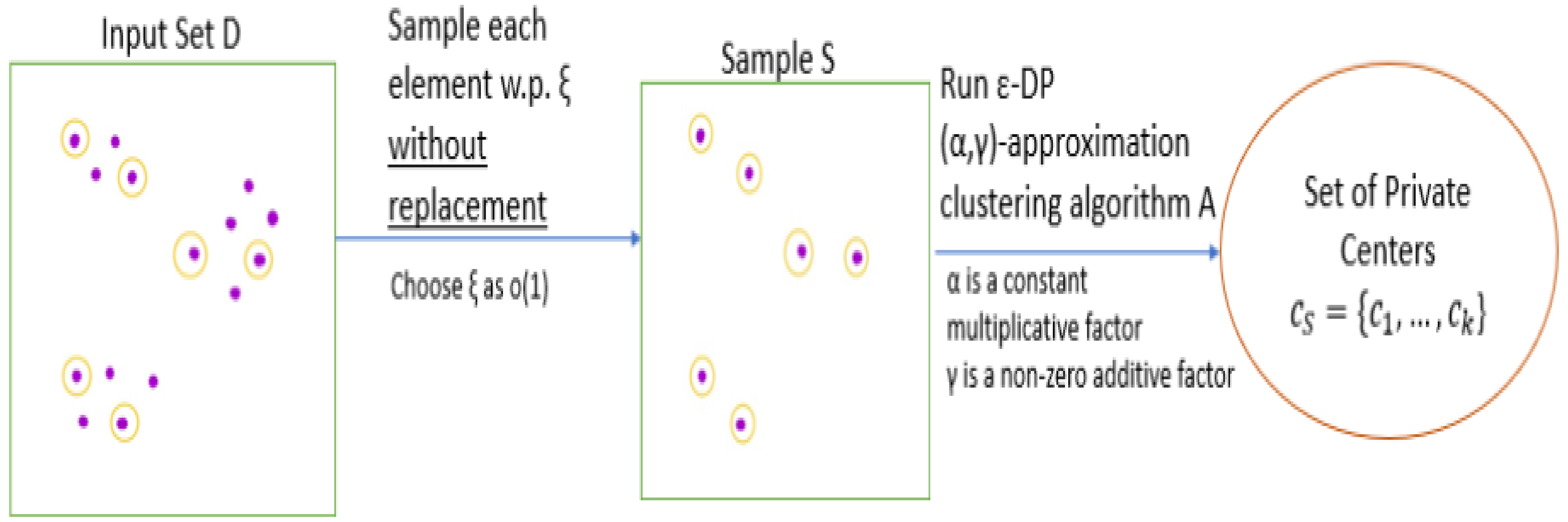


Framework for Sublinear-time Clustering



[MOP2001; CS2004] showed that for a small sample size.
Average cost of clustering on the sample S ≈ Average cost of clustering on the entire input set D

Framework for Sublinear-time DP Clustering



Challenges.
(1) Need to sample without replacement to preserve DP.
(2) DP Clustering algorithms are (α, γ)-approximate where $\gamma \neq 0$.

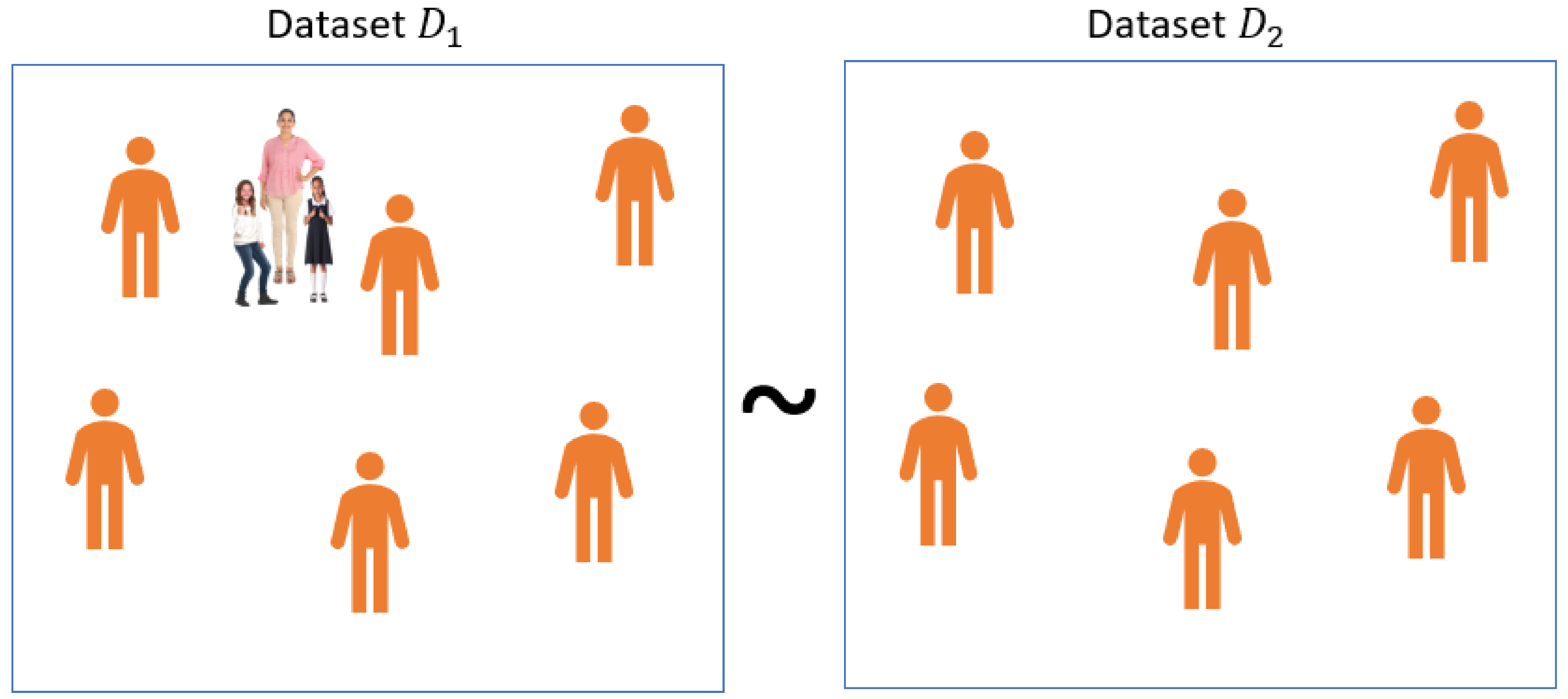
Sublinear-time DP Clustering Results

Assuming a DP (α, γ)-factor approx. k-median (or k-means) algorithm that runs in time $T(n)$ we can draw a sample S of size s = $poly(\alpha, k, \ln n)$ and obtain a k-median (or k-means) clustering $\hat{c}_S$ in time $T(s)$ such that with high probability

$$avg - cost(\hat{c}_S) \leq \alpha \cdot avg - cost(c_D) + \gamma + \epsilon$$

Where c_D is the optimum k-median (or k-means) clustering of input set D.

Group Privacy



(Naïve bound) An $(\varepsilon, 0)$-DP mechanism guarantees $(g\varepsilon, 0)$-group DP for group of size g elements.

Stronger Group Privacy for Sampling Algorithms

An algorithm that runs an $(\varepsilon, 0)$-DP mechanism on a subsample (each item sampled w.p. ξ) is $(T\varepsilon, \delta_{T,\xi,g})$-group DP for groups of size g.

where $T \in [0, g]$is a threshold, and
$\delta_{T,\xi,g} := \Pr[(\#samples\ from\ the\ group) > T]$.

$\delta_{T,\xi,g}$is often negligible even for $T \ll g$.
The guarantee of $(T\varepsilon, \delta_{T,\xi,g})$-group DP is then much stronger than the naive bound of $(g\varepsilon, 0)$-group DP .

References

1. Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang. Differentially private clusteringin high-dimensional Euclidean spaces. ICML, 2017
2. Artur Czumaj and Christian Sohler. Sublinear-time approximation algorithms for clustering via random sampling. ICALP, 2004.
3. Badih Ghazi, R. Kumar, and Pasin Manurangsi. Differentially private clustering: Tight approximation ratios. NeurIPS, 2020.
4. Anupam Gupta, Katrina Ligett, Frank McSherry, Aaron Roth, and Kunal Talwar. Differentially private combinatorial optimization. SODA 2010.
5. Kamal Jain, Mohammad Mahdian, and Amin Saberi. A new greedy approach for facility location problems. STOC, 2002.
6. Oded Kariv and S Louis Hakimi. An algorithmic approach to network location problems. ii: The p-medians. SIAM Journal on Applied Mathematics, 1979.
7. Haim Kaplan and Uri Stemmer. Differentially private k-means with constant multiplicative error. NeurIPS, 2018.
8. Nina Mishra, Dan Oblinger, and Leonard Pitt. Sublinear time approximate clustering. SODA, 2001.
9. Kobbi Nissim and Uri Stemmer. Clustering algorithms for the centralized and local models. ALT, 2018.
10. Dan Feldman, Amos Fiat, Haim Kaplan, and Kobbi Nissim. Private coresets. STOC, 2009.
11. Dan Feldman, Chongyuan Xiang, Ruihao Zhu, and Daniela Rus. Coresets for differentially private k-means clustering and applications to privacy in mobile sensor networks. IPSN, 2017
12. Uri Stemmer. Locally private k-means clustering. SODA, 2020