Contrastive Explainable Clustering with Differential Privacy

AAAI Track

Dũng Nguyen*
Department of Computer Science,
and Biocomplexity Institute,
University of Virginia
Charlottesville, VA, USA
dungn@virginia.edu

Ariel Vetzler*
Department of Computer Science,
Bar-Ilan University
Ramat Gan, Israel
arielvetzler1@gmail.com

Sarit Kraus
Department of Computer Science,
Bar-Ilan University
Ramat Gan, Israel
sarit@cs.biu.ac.il

Anil Vullikanti
Department of Computer Science,
and Biocomplexity Institute,
University of Virginia
Charlottesville, VA, USA
vsakumar@virginia.edu

ABSTRACT

This paper presents a novel approach to Explainable AI (XAI) that combines contrastive explanations with differential privacy for clustering algorithms. Focusing on k-median and k-means problems, we calculate contrastive explanations as the utility difference between original clustering and clustering with a centroid fixed to a specific data point. This method provides personalized insights into centroid placement. Our key contribution is demonstrating that these differentially private explanations achieve essentially the same utility bounds as non-private explanations. Experiments across various datasets show that our approach offers meaningful, privacy-preserving, and individually relevant explanations without significantly compromising clustering utility. This work advances privacy-aware machine learning by balancing data protection, explanation quality, and personalization in clustering tasks.

KEYWORDS

Explainable AI; Differential Privacy; Clustering

ACM Reference Format:
Dũng Nguyen, Ariel Vetzler, Sarit Kraus, and Anil Vullikanti. 2025. Contrastive Explainable Clustering with Differential Privacy: AAAI Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 22 pages.

1 INTRODUCTION

Different notions of clustering are fundamental primitives in several areas, including machine learning, data science, and operations research [34]. k-means and k-median clustering remain among the most important and widely used approaches, as demonstrated by recent advances in explainability, privacy, fairness, and contrastive learning [12, 25, 26, 42]. These problems often involve significant trade-offs between accessibility, resource allocation, and overall cost. For example, in emergency response planning, authorities must decide the optimal locations for ambulance stations to minimize response times across a city. Residents might question why an ambulance station isn't located closer to their neighborhood, especially if they feel imbalances in resource distribution. Similarly, in retail, customers might wonder why a new store is not placed near their area, despite being part of a high-demand demographic. Explainability in these contexts provides transparency into how and why certain decisions are made, addressing questions like: "Why was this location chosen instead of another?" This is particularly important in applications where the consequences of clustering decisions directly affect individuals or communities. [2, 31]

Such questions fall within the area of Explainable AI, which is a rapidly growing and vast area of research [4, 10, 24, 36, 37]. We focus on post-hoc explanations, especially contrastive explanations, e.g., [24, 33], which address "why P instead of Q?" questions. For example, in warehouse optimization, contrastive explanations clarify why a specific location was chosen as a distribution center, considering constraints like storage capacity or demand [42]. These methods are widely applied in multi-agent systems, reinforcement learning, and contrastive analysis [3, 4, 20, 37, 38]. In reinforcement learning, they explain actions by highlighting trade-offs, such as long-term rewards or risks [41].

Following the approach introduced in [33, 42], we explain clustering decisions by comparing the costs in two scenarios: $cost(S)$, the cost of an optimal clustering solution on the whole dataset $\mathbf{X} = \{x_1, \ldots, x_n\}$ and $cost(S^{(i)})$, the cost of a modified solution where we fix a center at a desired location requested by agent $x_i \in \mathbf{X}$. We explain the decision by showing $cost(S^{(i)}) - cost(S)$, i.e., how much the overall clustering cost increases when we force a center to be in a specific location. A higher clustering cost indicates worse performance. This comparison reveals the trade-offs in clustering: optimizing for one specific location often leads to a higher overall cost, meaning a worse solution for everyone else. By examining the difference between the optimal clustering cost and the cost of

*These authors contributed equally.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

the forced fixed clustering, we can understand why the algorithm chose certain locations for centers and not others. This approach helps people grasp the complex balancing act involved in clustering decisions, especially when trying to distribute resources fairly [42].

Data privacy is a crucial concern across various fields, and Differential Privacy (DP) is one of the most widely used and rigorous models for privacy [8]. We focus on the setting where the set of data points $\mathbf{X} = \{x_1, \ldots, x_n\}$ are private; for instance, in the ambulance center deployment problem, each x_i represents an individual requiring emergency services and seeking to keep their information private. There has been a lot of work on the design of differentially private solutions to clustering problems such as k-median and k-means in such a privacy model [1, 13, 15, 40].

While there has been significant progress in various domains of differential privacy, the intersection of explainability and differential privacy still needs to be explored. In clustering problems, building on the formalization of explanations for combinatorial problems we provide a private contrastive explanation to agent x_i by computing $\text{cost}(S_\epsilon^{(i)}) - \text{cost}(S_\epsilon)$. Here, S_ϵ represents a private solution using the privacy budget ϵ, while $S_\epsilon^{(i)}$ denotes a private solution, when we constrain one center to be at a location specified by agent x_i (this can be any point of interest, not necessarily the agent's own location). This difference quantifies how the clustering cost changes when accommodating agent i's position, offering a privacy-preserving explanation of the clustering decision. However, giving such a private contrastive explanation to each agent i naively using a private clustering algorithm would require a high privacy budget due to composition, which impact the accuracy, and lead to misleading or uninformative results. The central question of our research: *is it possible to offer each user an informative private contrastive explanation with a limited overall privacy budget?*

Our contributions.

1. We introduce the PrivEC problem, designed to formalize private contrastive explanations to all agents in clustering using k-median and k-means objectives.
2. We present an ϵ-DP mechanism, PrivateExplanations, which provides a contrastive explanation to each agent while ensuring the same utility bounds as private clustering in Euclidean spaces, offering personalized insights without compromising privacy or clustering quality. We use the private coreset technique of [13], which is an intermediate private data structure that preserves similar clustering costs as the original data.
3. We evaluate our methods on diverse datasets with varying distributions and feature dimensions. Our results demonstrate privacy-utility trade-offs comparable to private clustering, with low clustering errors even at reasonable privacy budgets, showcasing the effectiveness of our approach. Our research stands out by seamlessly integrating differential privacy into contrastive explanations, maintaining the quality of explanations even under privacy constraints. This work bridges the gap between privacy and explainability, marking a significant advancement in privacy-aware machine learning. A key technical contribution of our work is the derivation of rigorous bounds on the approximation factors for all contrastive explanations, ensuring their reliability and effectiveness. Due to space limitations, we only present major technical details in the main paper. We maintain a full, updated version of this paper with complete proofs and extended experimental results at [27].

2 RELATED WORK

Our work considers differential privacy for explainable AI in general (XAI) and Multi-agent explanations (XMASE) in particular, focusing on post-hoc contrastive explanations for clustering. We summarize some of the work directly related to our paper; additional discussion is presented in the Appendix, due to space limitations. Extensive experiments presented in [35] demonstrate non-negligible changes in explanations of black-box ML models through the introduction of privacy.

[29] considers feature-based explanations (e.g., SHAP) that can expose the top important features that a black-box model focuses on. To prevent such expose they introduced a new concept of achieving local differential privacy (LDP) in the explanations, and from that, they established a defense, called XRAND, against such attacks. They showed that their mechanism restricts the information that the adversary can learn about the top important features while maintaining the faithfulness of the explanations.

[14] study the security of contrastive explanations, and introduce the concept of the "explanation linkage attack", a potential vulnerability that arises when employing strategies to derive contrastive explanations. To address this concern, they put forth the notion of k-anonymous contrastive explanations. As the degree of privacy constraints increases, a discernible trade-off comes into play: the quality of explanations and, consequently, transparency are compromised.

Closer to our application is the work of [11], which investigates the privacy aspects of contrastive explanations in the context of team formation. They present a comprehensive framework that integrates team formation solutions with their corresponding explanations, while also addressing potential privacy concerns associated with these explanations. Additional evaluations are needed to determine the privacy of such heuristic-based methods.

There has been a lot of work on private clustering and facility location, starting with [15], which was followed by a lot of work on other clustering problems in different privacy models, e.g., [9, 16, 30, 39, 40]. [15] demonstrated that the additive error bound for points in a metric space involves an $O(\Delta k^2 \log n/\epsilon)$ term, where Δ is the space's diameter. Consequently, all subsequent work, including ours, assumes points are restricted to a unit ball.

We note that our problem has not been considered in any prior work in the XAI or differential privacy literature. The formulation we study here will likely be useful for other problems requiring private contrastive explanations.

3 PRELIMINARIES

Let $\mathbf{X} \subset \mathbb{R}^d$ denote a dataset consisting of d-dimensional points (referred as agents). We consider the notion of (k, p)-clustering, as defined by Definition 1.

Definition 1. *((k, p)-Clustering [13]). Given $k \in \mathbb{N}$, and a multi-set $\mathbf{X} = \{x_1, \ldots, x_n\}$ of points in the unit ball, a (k, p)-clustering is a set of k centers $\{c_1, \ldots, c_k\}$ minimizing $\text{cost}^p_{\mathbf{X}}(c_1, \ldots, c_k) = \sum_{i \in [n]} \min_{j \in [k]} \|x_i - c_j\|^p$.*

For $p = 1$ and $p = 2$, this corresponds to the k-median and k-means objectives, respectively. We drop the subscript $\mathbf{X}$ and superscript p, when it is clear from the context, and refer to the cost of a feasible clustering solution S by $\text{cost}(S)$.

Definition 2. *((w,t)-approximation). Given $k \in \mathbb{N}$, and a multiset $\mathbf{X} = \{x_1, \ldots, x_n\}$ of points in the unit ball, let $OPT_{\mathbf{X}}^{p,k} = \min_{c_1,\ldots,c_k \in \mathbb{R}^d} cost_{\mathbf{X}}^{p}(c_1, \ldots, c_k)$ denote the cost of an optimal (k,p)-clustering. We say $c_1, \ldots, c_k$ is a $(w,t)-$ approximation to a (k,p)-optimal clustering if $cost_{\mathbf{X}}^{p}(c_1, \ldots, c_k) \leq w \cdot OPT_{\mathbf{X}}^{p,k} + t$.*

Let OPT denote the cost of the optimal (k,p)-clustering, OPT_i denote the cost of the optimal (k,p)-clustering, with a center fixed at position z_i (the location chosen by agent i) and the remaining k-1 centers are chosen to optimize the objective. Let w', w'' denote the maximum approximation (w.r.t. OPT and OPT_i respectively) of non-private clustering algorithms. These factors will be specified in Sections 4.1 and 4.2.

A coreset (of some original set) is a set of points that, given any k centers, the cost of clustering of the original set is "roughly" the same as that of the coreset [13].

Definition 3. *For $\gamma, t > 0, p \geq 1, k, d \in \mathbb{N}$, a set X' is a (p,k,γ,t)-coreset of $X \subseteq \mathbb{R}^d$ if for every $C = \{c_1, \ldots, c_k\} \in R^d$, we have $(1-\gamma)cost_X^p(C) - t \leq cost_{X'}(C) \leq (1+\gamma)cost_X^p(C) + t$.*

Privacy model. We use the notion of differential privacy (DP), introduced in [8], which is a widely accepted formalization of privacy. A mechanism is DP if its output doesn't differ too much on "neighboring" datasets; this is formalized below.

Definition 4. *$\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-differentially private if for any neighboring datasets $X \sim X' \in \mathcal{X}$ and $S \subseteq \mathcal{Y}$,*

$$\Pr[\mathcal{M}(X) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(X') \in S] + \delta.$$

If $\delta = 0$, we say $\mathcal{M}$ is ϵ-differentially private.

We assume that the data points in $\mathbf{X}$ (i.e., users) are private, and say $\mathbf{X}, \mathbf{X}'$ are neighboring (denoted by $\mathbf{X} \sim \mathbf{X}'$) if they differ in one data point. When a value is disclosed to an individual agent i, it is imperative to treat the remaining clients in $\mathbf{X} - \{i\}$ as private entities.

Definition 5. *A mechanism $\mathcal{M}$ is ϵ-i-exclusion DP if, $\forall X, X' : i \in X, i \in X', X \setminus \{i\} \sim X' \setminus \{i\}$, and for all $S \subseteq Range(\mathcal{M})$:*

$$\Pr[\mathcal{M}(X) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(X') \in S].$$

We extend this to say that $\mathcal{M}$ is ϵ-Y-exclusion DP if the above holds $\forall X, X' : Y \subset X, Y \subset X', X \setminus Y \sim X' \setminus Y$.

We now define the PrivEC problem for providing private contrastive explanations, where each agent x_i seeks an explanation for a center fixed at a location of their choosing denoted by z_i.

Definition 6. ***Private and Explainable Clustering problem (PrivEC)*** *Given an instance $\mathbf{X} \subset \mathbb{R}^d$, clustering parameters k, p, and a contrastive set of points $Z \subset \mathbb{R}^d$, the goal is to output:*
***Private**: An ϵ-DP clustering solution S_ϵ (available to all)*
***Explainable**: For each agent $x_i \in X$, output $cost(S_\epsilon^{(i)})$ - $cost(S_\epsilon)$. $S_\epsilon^{(i)}$ is a private solution computed by the clustering algorithm with one centroid fixed at the position requested by agent i.*

We assume that $S_\epsilon^{(i)}$ is not revealed to any agent, but $cost(S_\epsilon^{(i)})$ - $cost(S_\epsilon)$ is released to agent i as contrastive explanation, which is ϵ-i-exclusion DP.

Lemma 1. *With probability at least $1-\beta$, $cost(S_\epsilon)$ (clustering cost) is a (w,t)-approximation of OPT, where[1]:*

$$w = w'(1+\alpha),$$
$$t = w' O_{p,\alpha}\left((k/\beta)^{O_{p,\alpha}(1)}.polylog(n/\beta)/\epsilon\right)$$

α is the approximation parameter for the utility of clustering and explanations. β is the failure probability of the utility guarantees of clustering and explanations.

Lemma 2. DP of fixed centroid yield additional cost. *Fix an i. If $OPT_i \geq w''(1+\alpha)OPT + t^{(i)}$, then with probability at least $1-2\beta$, $cost(S_\epsilon)$ and $cost(S_\epsilon^{(i)})$ computed by Algorithm PrivateExplanations satisfies that $cost(S_\epsilon^{(i)}) > cost(S_\epsilon)$.*

4 PRIVATEEXPLANATIONS MECHANISM

We design PrivateExplanation (Algorithm 1) for providing contrastive explanations for each agent. Specifically, it takes as inputs: (1) x_i which specifies the location of each agent i, and the contrastive location z_i for which they want an explanation, (2) original and target dimensions (d, d'), number of clusters (k), privacy budget ϵ, and ζ (explained later). The algorithm's key components are:

Algorithm 1 PrivateExplanation
Input: $(x_1, \ldots, x_n), (z_1, \ldots, z_n), d, d', k, \epsilon, \zeta$
Output: (ϵ, δ)-differentially private explanation for agents

$(x'_1, \ldots, x'_n) \leftarrow DimReduction((x_1, \ldots, x_n), d, d')$
$Z' \leftarrow DimReduction((z_1, \ldots, z_n), d, d')$
$Y \leftarrow \text{PrivateCoreset}^{\epsilon/2}(x'_1, \ldots, x'_n; \zeta)$
$(c'_1, \ldots, c'_k), cost(S'_\epsilon) \leftarrow NonPrivateApprox(Y, k)$
$cost(S_\epsilon) = RevertDimValue * cost(S'_\epsilon)$
$c \leftarrow \text{DimReverse}^{\epsilon/2}((c'_1, \ldots, c'_k), (x'_1, \ldots, x'_n))$
for $z'_i \in Z'$ **do**
  $cost(S'^{(i)}_\epsilon) \leftarrow NonPrivateApproxFC(Y, k, z'_i)$
  $cost(S^{(i)}_\epsilon) = RevertDimValue * cost(S'^{(i)}_\epsilon)$
end for
return $cost(S^{(i)}_\epsilon) - cost(S_\epsilon) | i \in range(1, \ldots, |X|)$

- **Dimension Reduction:** Using DimReduction from [13], we transform input data (in dimension d) to a lower-dimensional space d'. This reduction is crucial since our coreset algorithm is exponential in the dimension, but by reducing to logarithmic dimensions, it becomes polynomial-time.
- **Private Coreset:** We create a differentially private coreset Y using PrivateCoreset from [13], ensuring $\epsilon/2$ differential privacy; the coreset is defined in Definition 3.
- **Clustering:** The coreset is clustered using a non-private approximation algorithm (NonPrivateApprox). We can use a non-private clustering algorithm here since the coreset itself is already private, and by the Post-Processing property of differential privacy, the final result remains private.

[1]We use the notation $O_{p,\alpha}$ to explicitly ignore factors of p, α

- **Cost Scaling:** In line 5 of the algorithm, by multiplying by RevertDimValue = $(\log(n/\beta)/0.01)^{p/2}$ we scale the clustering cost ($cost(S'_\epsilon)$) (in the low-dimensional space) back to the original dimension ($\text{cost}(S_\epsilon)$) as shown in [21]. This reversal is necessary because while we computed costs in reduced dimensions for efficiency, we need the final cost in the original dimensions for accuracy.
- **Dimension Reverse:** Centroids are mapped back to the original space using DimReverse, maintaining ϵ-differential privacy.
- **Contrastive Explanations:** For each data point, we execute fixed-centroid clustering (NonPrivateApproxFC) on the coreset, constraining one centroid to a location chosen by the agent. This algorithm, detailed in Sections 4.1 and 4.2, is our key contribution as it modifies standard k-means and k-median algorithms to fix one centroid while maintaining their original utility bounds from literature, ensuring meaningful explanations. Without utility bounds, agents could challenge the validity of the explanation, arguing that the fixed centroid might degrade the clustering solution to an unacceptable extent. However, by guaranteeing the same utility bounds as the original algorithms, we ensure that the explanations are grounded in the vicinity of optimal clustering solutions, leaving no room for users to dispute the fairness or validity of the explanation. This alignment between explanation quality and clustering utility reinforces the trustworthiness of the algorithm and the insights it provides. After clustering, we apply RevertDimValue to transform the cost back to the original space ($\text{cost}(S_\epsilon^{(i)})$). By combining lines 5 and 9 of the algorithm, we derive the output: $\text{cost}(S_\epsilon^{(i)})$ - $\text{cost}(S_\epsilon)$ for each agent. This value captures the loss of optimality when fixating one centroid, quantifying how much the clustering quality degrades due to this constraint, serving as a contrastive explanation.

Theorem 1. DP of Explanation. *The solution $(c_1, \ldots, c_k)$ and $cost(S_\epsilon)$ computed by Algorithm PrivateExplanations are ϵ-DP. For all clients i and $S_\epsilon^{(i)}$ computed by Algorithm PrivateExplanations is ϵ-i-exclusion DP.*

Privacy analysis, as demonstrated in Theorem 1, we establishes the privacy guarantees of PrivateExplanations. Y coreset is $\epsilon/2$-differentially private as an output of $\epsilon/2$-DP algorithm. Consequently, $(c'_1, \ldots, c'_k)$ and $cost(S_\epsilon)$ maintain $\epsilon/2$-DP status, under the Post-Processing property.
Applying $\text{DimReverse}^{\epsilon/2}$ to find the centers in the original space, $c = \{c_1, \ldots, c_k\}$ is ϵ-DP by Composition theorem. For each i, $\text{cost}(S_\epsilon^{(i)})$ is produced by Post-Processing of Y with only z'_i, hence $cost(S_\epsilon^{(i)})$ satisfies ϵ-i-exclusion-DP.

Running Time Analysis.

Algorithm 1 has a total runtime of $O((k/\beta)^{O_{p,\alpha}(1)}\text{poly}(nd))$, which is polynomial in the input size. The key components contributing to this complexity include PrivateCoreset, DimReverse, and instances of (k, p)-clustering with and without fixed centers.
PrivateCoreset runs in $O((k/\beta)^{O_{p,\alpha}(1)}\text{poly}(n))$ time, as it sets $d' = O(p^4 \log(k/\beta))$ to satisfy the Dimension-Reduction Lemma (Appendix Section B) and uses Lemma 42 from [13]. DimReverse, which includes the FindCenter operation (detailed in the Appendix), has a time complexity of $O(\text{poly}(np))$ and is invoked k times. Additionally, we execute one standard (k, p)-clustering and $|X|$ instances of (k, p)-clustering with a fixed center. Together, these steps ensure the algorithm's overall polynomial runtime. All symbols used in this analysis are defined in Table 2 in the Appendix.

Theorem 2. *Assume there exist polynomial-time algorithms for (k, p)-clustering and (k, p)-clustering with a fixed center. The total running time of Algorithm 1 is $O((k/\beta)^{O_{p,\alpha}(1)} poly(nd))$.*

This computational complexity demonstrates that our algorithm is efficient for large datasets, balancing the additional overhead of fixed-centroid clustering with practical runtimes. Theorem 2 follows from the detailed steps, as PrivateCoreset and FindCenter contribute manageable computational overhead. Finally, the algorithm integrates a critical utility analysis to ensure robust performance. In the following sections, we present rigorous upper bounds and specific constraints for k-means and k-median, illustrating the practicality and effectiveness of our approach.

Utility Analysis. PrivateCoreset uses parameters ζ (which is a function of α) and privacy budget $\epsilon/2$, derived from [13] and detailed in our Appendix. This algorithm produces a coreset Y that ensures the clustering cost on Y closely approximates the cost on the projected dataset X'. Specifically, the approximation is within a $(1 + 0.1\alpha)$ factor, plus an additive $\widetilde{O}(\text{polylog}(n/\beta)/\epsilon)$ term. Then, by applying the Dimensional Reduction lemma (in the Appendix), which states that the cost of a specific clustering on X' (d'-dimensional space) is under some constant factor of the same clustering on X (d-dimensional space), we can bound the $cost(S_\epsilon{}^{(i)})$ by its optimal clustering OPT_i. We first state the approximation factor derived using [13], since this is used in our analysis.

Theorem 3. Cost of explanations due to privacy. *Fix an agent i. With probability at least $1 - \beta$, $cost(S_\epsilon^{(i)})$ computed by Algorithm PrivateExplanations is a (w, t)-approximation of OPT_i, with*

$$w = w''(1 + \alpha)$$
$$t = w'' O_{p,\alpha}\left((k/\beta)^{O_{p,\alpha}(1)}.polylog(n/\beta)/\epsilon\right)$$

As $S_\epsilon^{(i)}$ results from a randomized mechanism, its cost is higher than S_ϵ's most of the time with high probability, ensuring a positive private explanation.

Tight Approximation Ratios. The most challenging aspect of our analysis is determining the precise approximation factor w'' for k-means and k-median in the context of fixed-centroid clustering. In the following sections, we will present modifications to well-known k-median and k-means algorithms, adapting them for fixed-centroid clustering scenarios. We will then demonstrate that these modified algorithms achieve the same tight approximation factors. Formally, we show how the well-known utility bounds of k-means and k-median can be preserved while fixing one centroid to a requested location, ensuring the robustness of these algorithms under such constraints. Corollary 1 and Corollary 2 will conclude this section by presenting the specific, tight approximation ratios (w'') achieved after applying our NonPrivateApproxFC algorithm.

These corollaries will provide detailed confirmation of our algorithm's effectiveness in achieving these optimal approximation ratios within the constraints of differential privacy.

4.1 NonPrivateApproxFC for k-median

We have developed a non-private fixed centroid clustering algorithm, which we call NonPrivateApproxFC. This algorithm is an adaptation of [5]. In the following section, we will prove that our modified algorithm, which works with a fixed centroid (referred to as z), achieves an 8-approximation factor. To grasp how we adapted the algorithm to suit our needs, it's essential to understand the symbols used in [5]. In this section, we adopt the notation from [5] to avoid confusion with the symbols used in this paper, where d and d' denote the original and reduced dimensions, respectively. In their work, d_j represents the demand at each location $j \in N$, serving as a weight that reflects the importance of the location. N refers to the set of agents $1, \ldots, n$.

For the conventional k-median problem, each d_j is initially set to 1 for all $j \in N$. The term c_{ij} represents the cost of assigning any i to j, x_{ij} represents if location j is assigned to center i and y_i indicates if the location i is selected as a center.

We assume the fixed center is one of the input data points N. [5] demonstrates that the k-median problem can be formulated as an integer programming problem, and in order to adapt the algorithm we add a constraint in line 9 to treat z as a fixed centroid. This modification allows the algorithm to account for the fixed centroid requirement. We then relax the integer program (IP) into a linear program (LP) and show that it preserves the same utility bound as the original algorithm. By specifying that $y_z \geq 1$, we ensure that y_z is designated as a centroid in our linear programming formulation. Throughout the solution process, y_z remains fixed as a centroid.

$$\text{minimize} \sum_{i,j \in N} d_j c_{ij} x_{ij} \tag{1}$$

$$\text{s.t.} \sum_{i \in N} x_{ij} = 1 \text{ for each } j \in N; \sum_{j \in N} y_j = k \tag{2}$$

$$x_{ij} \leq y_i \text{ for each } i, j \in N \tag{3}$$

$$x_{ij}, y_i \geq 0 \text{ for each } i, j \in N \tag{4}$$

$$y_z, x_{zz} \geq 1 \text{ for a fixed } z \in N \tag{5}$$

Let $(\bar{x}, \bar{y})$ be a feasible solution of the LP relaxation and let $\bar{C}_j = \sum_{i \in N} c_{ij} \bar{x}_{ij}$ for each $j \in N$ as the total (fractional) cost of client j. Throughout the three steps, we demonstrate that solving this linear program with the added constraint does not introduce any additional approximation factor. The program is solved with the same efficiency and accuracy as it would be without the fixed centroid constraint.

The first step. We group nearby locations by their demands without increasing the cost of a feasible solution $(\bar{x}, \bar{y})$, such that locations with positive demands are relatively far from each other. By re-indexing, we get $\bar{C}_z \leq \bar{C}_1 \leq \bar{C}_2 \leq \ldots \bar{C}_n$.

We will show that it's always possible to position $\bar{C}_z$ as the first element of the list, i.e., $\bar{C}_z$ is equal to the minimum value of all $\bar{C}_j$. Recall that: $\bar{C}_z = \sum_{i \in N} c_{iz} \bar{x}_{iz} = \sum_{i \in N, i \neq z} c_{iz} \bar{x}_{iz} + c_{zz} \bar{x}_{zz} = 0$, since we know that $\sum_{i \in N} x_{iz} = 1, x_{zz} \geq 1$ and $c_{zz} = 0$.

The remaining work of the first step follows [5]. We first set the modified demands $d_j' \leftarrow d_j$. For $j \in N$, moving all demand of location j to a location $i < j$ s.t. $d_i' > 0$ and $c_{ij} \leq 4\bar{C}_j$, i.e., transferring all j's demand to a nearby location with existing positive demand. Demand shift occurs as follows: $d_i' \leftarrow d_i' + d_j', d_j' \leftarrow 0$. Since we initialize $d_z' = d_z = 1$, and we never move its demands away, it follows that $d_z' > 0$.

Let N' be the set of locations with positive demands $N' = \{j \in N, d_j' > 0\}$. A feasible solution to the original demands is also a feasible solution to the modified demands.

Lemma 3. *Locations $i, j \in N'$ satisfy: $c_{ij} > 4 \max(\bar{C}_i, \bar{C}_j)$.*

Proof. The lemma follows the demands moving step (in the first step of the algorithm): for every j to the right of i (which means $\bar{C}_j \geq \bar{C}_i$) and within the distance of $\bar{C}_j$ (that also covers all points within distance $\bar{C}_i$), we move all demands of j to i, hence j will not appear in N'. □

Lemma 4. *The cost of the fractional $(\bar{x}, \bar{y})$ for the input with modified demands is at most its cost for the original input.*

Proof. The cost of the LP $\bar{C}_{LP} = \sum_{j \in N} d_j \bar{C}_j$ and $\bar{C}'_{LP} = \sum_{j \in N} d_j' \bar{C}_j$. Since we move the demands from $\bar{C}_j$ to a location i with lower cost $\bar{C}_i \leq \bar{C}_j$ the contribution of such moved demands in $\bar{C}'$ is less than its contribution in $\bar{C}$, it follows that $\bar{C}'_{LP} \leq \bar{C}_{LP}$. □

The second step. We analyze the problem with modified demands d'. We will group fractional centers from the solution $(\bar{x}, \bar{y})$ to create a new solution (x', y') with cost at most $2\bar{C}_{LP}$ such that $y_i' = 0$ for each $i \notin N'$ and $y_i' \geq 1/2$ for each $i \in N'$. We also ensure that $y_z' \geq 1/2$ in this step, i.e., z will be a fractional center after this. A solution is called 1/2-restricted if $y_j \geq 1/2$ for any point $j \in N$ and $y_j = 0$ otherwise. This restriction balances the assignment of demand, ensuring that no single center dominates excessively. The concept of 1/2-restricted solutions is used to create more equitable distributions of demand and is key to transitioning to a $\{1/2, 1\}$-integral solution. The next lemma leverages this property:

Lemma 5. *For any 1/2-restricted solution (x', y') there exists a $\{1/2, 1\}$-integral solution with no greater cost.*

Proof. The cost of the $\frac{1}{2}$-restricted solution (by Lemma 7 of [5]) is:

$$C'_{LP} = \sum_{j \in N'} d_j' c_{s(j)j} - \sum_{j \in N'} d_j' c_{s(j)j} y_j', \tag{6}$$

Let $s(j)$ be j's closest neighbor location in N', the first term above is independent of y' and the minimum value of y_j' is 1/2. We now construct a $\{1/2, 1\}$-integral solution $(\hat{x}, \hat{y})$ with no greater cost. Sort the location $j \in N', j \neq z$ in the decreasing order of the weight $d_j' c_{s(j)j}$ and put z to the first of the sequence, set $\hat{y}_j = 1$ for the first $2k - n'$ locations and $\hat{y}_j = 1/2$ for the rest. By doing that, we minimize the cost by assigning heaviest weights $d_j' c_{s(j)j}$ to the maximum multiplier (i.e., 1) while assigning lightest weights $d_j' c_{s(j)j}$ to the minimum multiplier (i.e., 1/2) for each $j \in N', j \neq z$. Any feasible 1/2-restricted solution must have $y_z' = 1$ to satisfy the constraint of z so that the contribution of $\hat{y}_z$ is the same as its of y_z'. It follows that the cost of $(\hat{x}, \hat{y})$ is no more than the cost of (x', y'). □

The third step. This step is similar to the part of Step 3 of [5] that converts a $\{1/2, 1\}$-integral solution to an integral solution with the cost increases at most by 2. We note that there are two types of center $\hat{y}_j = 1/2$ and $\hat{y}_j = 1$, hence there are two different processes. All centers j with $\hat{y}_j = 1$ are kept while more than half of centers j with $\hat{y}_j = 1/2$ are removed. Since we show that $\hat{y}_z = 1$ in the previous step, z is always chosen by this step and hence guarantees the constraint of z.

Theorem 4. Approximation factor of fixed centroid k-median. *For the metric k-median problem, the algorithm above outputs an* 8*-approximation solution.*

Proof. It is obvious that the optimal of the LP relaxation is the lower bound of the optimal of the integer program. While constructing an integer solution for the LP relaxation with the modified demands, [5] states that there is a 1/2-restricted solution (x', y') which costs at most $2\bar{C}_{LP}$. And now the third step multiplies this cost by a factor of 2, making the cost of the solution (to the LP) at most $4\bar{C}_{LP}$. Transforming the integer solution of the modified demands to a solution of the original input adds an additive cost of $4\bar{C}_{LP}$ by Lemma 4 of [5] and the Theorem follows. □

Having demonstrated that our modification of [5] to execute fixed-centroid k-median instead of standard k-median yields an 8-factor approximation of the optimal solution, we can now proceed to prove that our private explanation closely approximates the optimal solution for the fixed-centroid scenario.

Corollary 1. *Running PrivateExplanations with NonPrivateApproxFC be the above K-median algorithm, with probability at least* $1 - \beta$, $S_\epsilon^{(i)}$ *is a* (w, t)*-approximation of* OPT_i*–the optimal K-median with a center fixed at position* z_i*, in which:*

$$w = 8(1 + \alpha)$$
$$t = 8O_{p,\alpha}\left((k/\beta)^{O_{p,\alpha}(1)}.polylog(n/\beta)/\epsilon\right).$$

4.2 NonPrivateApproxFC for k-means

In this section, we present our NonPrivateApproxFC algorithm for k-means with a fixed center. Based on [19], we achieve a 25-approximation. We will analyze this approximation factor in detail below. We adapt the work by [19] by adding a fixed center constraint to the single-swap heuristic algorithm. As in their result, we need to assume that we are given a discrete set of candidate centers C from which we choose k centers. The optimality is defined in the space of all feasible solutions in C, i.e., over all subsets of size k of C. We then present how to remove this assumption, with the cost of a small constant additive factor.

Definition 7. *Let* $O = (O_1, O_2, \ldots, O_k)$ *be the optimal clustering with* O_1 *be the cluster with the fixed center* z*. A set* $C \subset \mathbb{R}^d$ *is a* γ*-approximate candidate center set if there exists* $z \in \{c_1, c_2, \ldots, c_k\} \subseteq C$*, such that:* $cost(c_1, c_2, \ldots, c_k) \leq (1 + \gamma)cost(O)$.

Given $u, v \in \mathbb{R}^d$, let $\Delta(u, v)$ denote the squared Euclidean distance between u and v: $\Delta(u, v) = \text{dist}^2(u, v)$. For a set $S \subset \mathbb{R}^d$, the total squared distance between all points in S and a point v is given by $\Delta(S, v) = \sum_{u \in S} \Delta(u, v)$. Similarly, for a set $P \subset \mathbb{R}^d$, $\Delta_P(S)$ represents the total squared distance between each point $q \in P$ and its closest point $s_q \in S$. Here, q refers to an individual data point in set P, and s_q is its nearest neighbor in S. When the context is clear, we drop P for simplicity. This notation captures the essential relationships between points and their nearest centroids.

Let z be the fixed center that must be in the output. Let C be the set of candidate centers, that $z \in C$. We define **stability** in the context of k-means with a fixed center z as follows. We note that it differs from the definition of [19] such that we never swap out the fixed center z:

Definition 8. *A set* S *of* k *centers that contains the fixed center* z *is called* 1*-stable if:* $\Delta\big(S \setminus \{s\} \cup \{o\}\big) \geq \Delta(S)$*, for all* $s \in S \setminus \{z\}$, $o \in O \setminus \{z\}$.

Algorithm. We initialize $S^{(0)}$ as a set of k centers form C that $z \in S^{(0)}$. For each set $S^{(i)}$, we perform the swapping iteration:

- Select one center $s \in S^{(i)} \setminus z$
- Select one replaced center $s' \in C \setminus S^{(i)}$
- Let $S' = S^{(i)} \setminus s \cup s'$
- If S' reduces the distortion, $S^{(i+1)} = S'$. Else, $S^{(i+1)} = S^{(i)}$

We repeat the swapping iteration until $S = S^{(m)}$, i.e., after m iterations, is a 1-stable. Theorem 5 states the utility of an arbitrary 1-stable set, which is also the utility of our algorithm since it always outputs an 1-stable set. We note that if C is created with some errors γ to the actual optimal centroids, the utility bound of our algorithm is increased by the factor $\Theta(\gamma)$, i.e., ours is a $(25 + \Theta(\gamma))$-approximation to the actual optimal centroids.

Theorem 5. Approximation factor of fixed centroid k-mean. *If* S *is an* 1*-stable* k*-element set of centers,* $\Delta(S) \leq 25\Delta(O)$*. Furthermore, if* C *is a* $\frac{\gamma}{25}$*-approximate candidate center set,* S *is a* $(25 + \gamma)$*-approximate of the actual optimal centroids in the Euclidean space.*

Having demonstrated that our modification of [19] to execute fixed-centroid k-means instead of standard k-means yields a 25-factor approximation of the optimal solution, we can now proceed to prove that our private explanation closely approximates the optimal solution for the fixed-centroid scenario.

Corollary 2. *Running PrivateExplanations with NonPrivateApproxFC be the above k-means algorithm, with probability at least* $1 - \beta$, $S_\epsilon^{(i)}$ *is a* (w, t)*-approximation of* OPT_i*–the optimal k-means with a center fixed at position* x_i*, in which:*

$$w = (25 + \gamma)(1 + \alpha)$$
$$t = (25 + \gamma)O_{p,\alpha}\left((k/\beta)^{O_{p,\alpha}(1)}.polylog(n/\beta)/\epsilon\right).$$

With the utility bounds for k-means and k-median under the fixed-centroid constraint proven, it is clear that altering the original algorithms preserves the same utility bounds as their non-fixed counterparts. This ensures that accommodating fixed centroids does not compromise clustering quality. Notably, these bounds refer to clustering utility, not explanation bounds. By showing that fixed-k-means and fixed-k-median perform as effectively as standard versions, users can trust the quality of the explanations. Without these guarantees, users might question the validity of centroid placements. Our results ensure the explanations are based on clustering solutions that are as robust and reliable as the original algorithms.

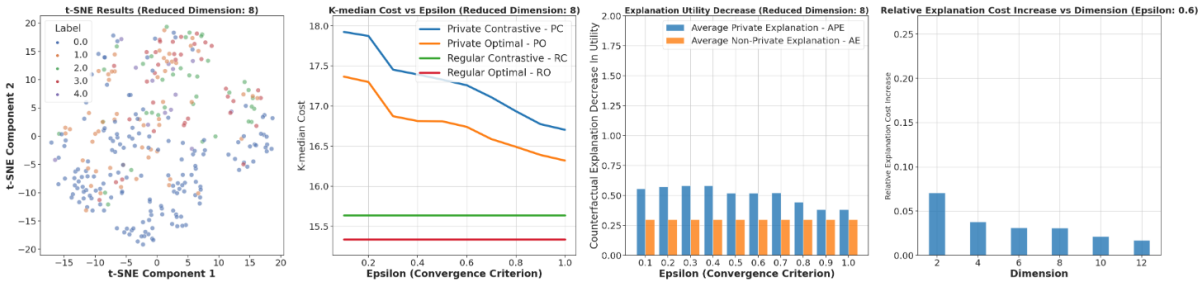

Figure 1: A visualization of our dataset (Heart Disease dataset from the UCI ML Repository), projected into an 8-dimensional space. (a) t-SNE of our data. (b) Comparison of k-median clustering with fixed and non-fixed centroids, both private and non-private. (c) Bar graph showing contrastive explanation differences for differential private and non-private k-median with a fixed centroid. (d) We fix the privacy budget of 0.6 while demonstrating the contrastive explanation across various dimensions.

5 EXPERIMENTS

Our study examines how the privacy budget ϵ affects the trade-off between privacy and accuracy, focusing on the quality of differentially private explanations. We use four key metrics: Private Optimal (PO, S_ϵ), Private Contrastive (PC, $S_\epsilon^{(i)}$), Regular Optimal (RO, OPT), and Regular Contrastive (RC, OPT_i), to compare clustering costs with and without fixed centroids in both private and non-private algorithms. To assess explanation quality, we define two derived metrics: Average Private Explanation (APE, PC - PO, cost($S_\epsilon^{(i)}$) - cost(S_ϵ)) and Average Explanation (AE, RC - RO, cost($S^{(i)}$) - cost(S)). APE measures utility loss in private clustering as an explanatory output, while AE provides a non-private baseline. These metrics help us evaluate the explanatory power of our approach. By analyzing these metrics across different ϵ values, we explore the balance between privacy and utility, highlighting the trade-offs in our differentially private clustering and explanation framework.

Datasets Our research utilizes a diverse set of datasets to demonstrate the versatility and effectiveness of our approach, as summarized in Table 1. We employed the Heart Disease dataset featuring 13 dimensions, and the Breast Cancer dataset with 30 features, including both numeric and categorical fields. Both datasets were taken from the UCI Machine Learning Repository Those higher-dimensional datasets were crucial in validating our theoretical framework. Additionally, we used two-dimensional activity-based population datasets from Charlottesville City and Albemarle County, Virginia, previously employed in mobile vaccine clinic deployment studies [23]. To complement these real-world datasets, we also generated a synthetic two-dimensional dataset. By testing our method on both high-dimensional and two-dimensional data, as well as on real and synthetic datasets, we showcase its robustness across different data complexities and origins.

Data Preprocessing: We normalized all datasets to fit within a unit ball to ensure consistency with prior work and standardize our analysis framework. While this normalization alters the absolute scale, it preserves the relative relationships between data points, which is crucial for clustering. The entire preprocessed dataset was used for analysis, as there is no ground truth labeling for a traditional train-test split in this unsupervised task.

Dimensionality Our study explored both 2D and higher-dimensional datasets. A crucial aspect of our methodology, DimReduction, employs Principal Component Analysis (PCA) for initial dimensionality reduction. This process normalizes the data and creates lower-dimensional representations. We performed extensive experiments, reducing high-dimensional datasets to various lower dimensions, including 2D, with additional low-dimension experiments. Remarkably, our results remained consistent across different reduced dimensionalities. Even when reducing data from 13 dimensions to 2, we observed similar trends and results as with other dimensional reductions, despite significant information loss. This consistency underscores our method's robustness across varying dimensions. By addressing high-dimensional data challenges through PCA reduction, we ensure our technique's applicability and efficiency across diverse dataset complexities, maintaining result integrity regardless of original data dimensionality.

Running Time Analysis: The computational complexity of our algorithm varies by clustering method. For k-means, we use the linear-time algorithm from [19], while the k-median approach relies on polynomial-time Linear Programming (LP). We have optimized performance with GPU parallelization, reducing execution times from minutes to seconds for both differentially private coresets and clustering tasks. Our method is **data-agnostic**, handling any data distribution efficiently, independent of sparsity. For reproducibility, we provide our code, experimental details, and pre-processed datasets in a public repository.

5.1 Experimental results

Figure 1 presents four key visualizations of our differentially private clustering and explanation framework across various dimensions. The t-SNE plot (leftmost) shows the 2D representation of our dataset, revealing potential clusters and patterns. The second plot illustrates K-median cost versus ϵ for our four metrics (PC,

Dataset	Dim	Size	Source
Heart Disease	13	303	UCI MLR
Breast Cancer Wisconsin	31	569	UCI MLR
Charlottesville	2	33K	[23]
Albemarle	2	74K	[23]
Synthetic dataset	2	1k	Generated

Table 1: Datasets used in our research

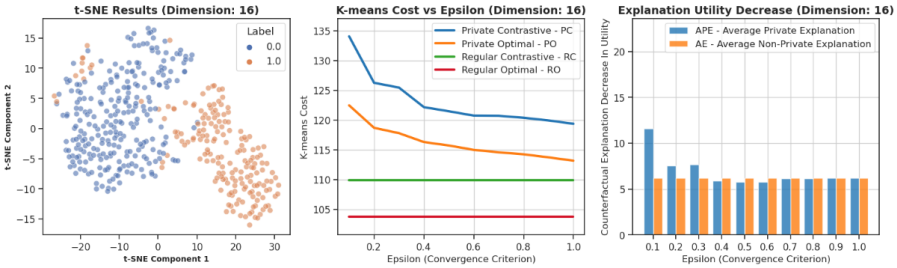

Figure 2: A visualization of another dataset (Breast Cancer dataset (30 features) from the UCI ML Repository), projected into a 16-dimensional space. (a) t-SNE of our data. (b) Comparison of k-means clustering with fixed and non-fixed centroids, both private and non-private. (c) Bar graph showing contrastive explanation differences for differential private and non-private k-means with a fixed centroid.

PO, RC, RO), demonstrating the privacy-accuracy trade-off and the consistency of our contrastive explanations.

In plots (b) and (c) of Figure 1, the x-axis represents the privacy budget ϵ, which we tested over the range $[0, 1]$ in intervals of 0.05. This granularity allows for a detailed analysis of the privacy-utility trade-off. Smaller values of ϵ enforce stronger privacy guarantees, as reflected in higher clustering costs for PC and PO. Conversely, as ϵ increases, these costs gradually decrease, highlighting the improved utility that comes with relaxed privacy constraints. Importantly, the observed trends demonstrate that the framework consistently balances privacy and utility, even at stricter privacy levels. Additionally, the stability of the non-private metrics (RC and RO) across the ϵ range provides a robust baseline for evaluating the performance of our private clustering and explanation methods.

As expected, both PC and PO costs decrease as ϵ increases, demonstrating the trade-off between privacy and accuracy. The non-private metrics (RC and RO) remain constant across ϵ values, serving as baselines for comparison. Notably, the gap between PC and PO remains relatively consistent, indicating that our contrastive explanations maintain their relative quality at different levels of privacy. The third plot illustrates the Explanation Utility for both private (Average Private Explanation) and non-private (Average Non-Private Explanation) scenarios across various ϵ values. This graph quantifies the difference in clustering cost between the optimal solution and the solution with a fixed centroid, representing our contrastive explanations. Notably, we observe that the Average Private Explanation remains relatively stable across different ϵ values. This stability is crucial as it indicates that the quality of our contrastive explanations in the private setting is consistent, regardless of the privacy budget. The consistent performance across different ϵ values underscores the robustness of our method, providing reliable explanations even under strict privacy constraints. The rightmost plot demonstrates the difference between PC and PO across dimensions for a fixed ϵ, illustrating our method's scalability with dimensionality.

Figure 2 follows a similar format but uses another high-dimensional dataset. This dataset was reduced from 30 dimensions to 16 to test the robustness of our approach on different datasets and higher dimensions. Unlike Figure 1, this figure presents results obtained using the k-means algorithm. Furthermore, we extended our experiments to include other reduced dimensions for both k-means and k-median, with the detailed results provided in the Appendix. These additional experiments further validate the adaptability and robustness of our framework across different clustering methods and dimensionality settings.

6 CONCLUSIONS

Our work explores the design of private explanations for clustering, particularly focusing on the k-median and k-means objectives for Euclidean datasets. We formalize this as the PRIVEC problem, where each agent receives a contrastive explanation corresponding to the loss in utility they experience when a cluster centroid is placed at a strategic position chosen by the agent. Our algorithm provides explanations to each user while maintaining the same approximation factor as private clustering, within a predefined privacy budget. The related work in this domain has shown the development of algorithms for contrastive explanations, but our contribution stands out by integrating differential privacy guarantees.

Our experiments demonstrate the resilience of our approach. Despite the added layer of providing differentially private explanations on top of differentially private clustering, the quality of our explanations remains uncompromised. The extended experiments on all our datasets further validate our approach's efficacy. The balance between privacy and utility, the robustness of contrastive explanations, and the negligible impact of ϵ on explainability were consistent across datasets. These findings underscore the potential of our method for diverse real-world applications.

Our approach is not restricted to k-means and k-median but can be applied to other clustering algorithms as well. The methodology leverages fundamental principles common to many clustering techniques, such as centroids and utility functions. As long as a clustering algorithm defines centroids and evaluates clustering quality using these metrics, our approach can be adapted to provide privacy-preserving contrastive explanations. This adaptability makes it suitable for extending to other paradigms, such as density-based or hierarchical clustering, extending its applicability to various datasets and contexts.

ACKNOWLEDGMENTS

This research is partially supported by the Israel Ministry of Innovation, Science & Technology grant 1001818511, NSF grants CCF-1918656, CNS-2317193, IIS-2331315, and CDC MIND U01CK000589.

REFERENCES

[1] Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang. 2017. Differentially Private Clustering in High-Dimensional Euclidean Spaces. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 322–331. http://proceedings.mlr.press/v70/balcan17a.html

[2] Szymon Bobek, Michal Kuk, Maciej Szelążek, and Grzegorz J Nalepa. 2022. Enhancing cluster analysis with explainable AI and multidimensional cluster prototypes. *IEEE Access* 10 (2022), 101556–101574.

[3] K Boggess, S Kraus, and L Feng. 2022. Toward Policy Explanations for Multi-Agent Reinforcement Learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[4] Kayla Boggess, Sarit Kraus, and Lu Feng. 2023. Explainable Multi-Agent Reinforcement Learning for Temporal Queries. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*.

[5] Moses Charikar, Sudipto Guha, Éva Tardos, and David B Shmoys. 1999. A constant-factor approximation algorithm for the k-median problem. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*. 1–10.

[6] Hongjie Chen, Vincent Cohen-Addad, Tommaso d'Orsi, Alessandro Epasto, Jacob Imola, David Steurer, and Stefan Tiegel. 2023. Private estimation algorithms for stochastic block models and mixture models. *Advances in Neural Information Processing Systems* 36 (2023), 68134–68183.

[7] Sanjoy Dasgupta and Anupam Gupta. 2003. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms* 22, 1 (2003), 60–65.

[8] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3–4 (aug 2014), 211–407. https://doi.org/10.1561/0400000042

[9] Dan Feldman, Chongyuan Xiang, Ruihao Zhu, and Daniela Rus. 2017. Coresets for differentially private k-means clustering and applications to privacy in mobile sensor networks. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 3–15.

[10] Mira Finkelstein, Lucy Liu, Yoav Kolumbus, David C Parkes, Jeffrey S Rosenschein, Sarah Keren, et al. 2022. Explainable Reinforcement Learning via Model Transforms. *Advances in Neural Information Processing Systems* 35 (2022), 34039–34051.

[11] Athina Georgara, Juan Antonio Rodríguez-Aguilar, and Carles Sierra. 2022. Privacy-Aware Explanations for Team Formation. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 543–552.

[12] Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. 2021. Socially fair k-means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 438–448.

[13] Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. 2020. Differentially Private Clustering: Tight Approximation Ratios. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 340, 15 pages.

[14] Sofie Goethals, Kenneth Sörensen, and David Martens. 2022. The privacy issue of counterfactual explanations: explanation linkage attacks. *arXiv preprint arXiv:2210.12051* (2022).

[15] Anupam Gupta, Katrina Ligett, Frank McSherry, Aaron Roth, and Kunal Talwar. 2010. Differentially Private Combinatorial Optimization. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*. SIAM, 1106–1125. https://doi.org/10.1137/1.9781611973075.90

[16] Zhiyi Huang and Jinyan Liu. 2018. Optimal differentially private algorithms for k-means clustering. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. 395–408.

[17] Tianxi Ji, Changqing Luo, Yifan Guo, Jinlong Ji, Weixian Liao, and Pan Li. 2019. Differentially private community detection in attributed social networks. In *Asian Conference on Machine Learning*. PMLR, 16–31.

[18] William B Johnson and Joram Lindenstrauss. 1984. Extensions of Lipschitz mappings into a Hilbert space. In *Conference on Modern Analysis and Probability*, Vol. 26. American Mathematical Society, 189–206.

[19] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. 2002. A local search approximation algorithm for k-means clustering. In *Proceedings of the eighteenth annual symposium on Computational geometry*. 10–18.

[20] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2493–2500.

[21] Konstantin Makarychev, Yury Makarychev, and Ilya Razenshteyn. 2019. Performance of Johnson-Lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 1027–1038.

[22] Jiří Matoušek. 2000. On approximate geometric k-clustering. *Discrete & Computational Geometry* 24, 1 (2000), 61–84.

[23] Zakaria Mehrab, Mandy L Wilson, Serina Chang, Galen Harrison, Bryan Lewis, Alex Telionis, Justin Crow, Dennis Kim, Scott Spillmann, Kate Peters, et al. 2022. Data-Driven Real-Time Strategic Placement of Mobile Vaccine Distribution Sites. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12573–12579.

[24] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[25] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. 2020. Explainable k-means and k-medians clustering. In *International conference on machine learning*. PMLR, 7055–7065.

[26] James Newling and François Fleuret. 2016. Fast k-means with accurate bounds. In *International Conference on Machine Learning*. PMLR, 936–944.

[27] Dung Nguyen, Ariel Vetzler, Sarit Kraus, and Anil Vullikanti. 2024. Contrastive explainable clustering with differential privacy. arXiv:2406.04610 [cs.LG] https://arxiv.org/abs/2406.04610

[28] Dung Nguyen and Anil Kumar Vullikanti. 2024. Differentially private exact recovery for stochastic block models. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 37798–37839. https://proceedings.mlr.press/v235/nguyen24j.html

[29] Truc Nguyen, Phung Lai, Hai Phan, and My T Thai. 2023. XRand: Differentially Private Defense against Explanation-Guided Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11873–11881.

[30] Kobbi Nissim and Uri Stemmer. 2018. Clustering algorithms for the centralized and local models. In *Algorithmic Learning Theory*. PMLR, 619–653.

[31] Sariel Ofek and Amit Somech. 2024. Explaining Black-Box Clustering Pipelines With Cluster-Explorer. *arXiv preprint arXiv:2412.20446* (2024).

[32] Neel Patel, Reza Shokri, and Yair Zick. 2022. Model explanations with differential privacy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1895–1904.

[33] Alberto Pozanco, Francesca Mosca, Parisa Zehtabi, Daniele Magazzeni, and Sarit Kraus. 2022. Explaining preference-driven schedules: the expres framework. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 32. 710–718.

[34] Chandan K Reddy. 2018. *Data clustering: algorithms and applications*. Chapman and Hall/CRC.

[35] Saifullah Saifullah, Dominique Mercier, Adriano Lucieri, Andreas Dengel, and Sheraz Ahmed. 2022. Privacy meets explainability: A comprehensive impact benchmark. *arXiv preprint arXiv:2211.04110* (2022).

[36] Sören Schleibaum, Lu Feng, Sarit Kraus, and Jörg P Müller. 2024. ADESSE: Advice Explanations in Complex Repeated Decision-Making Environments. *arXiv preprint arXiv:2405.20705* (2024).

[37] Sarath Sreedharan, Utkarsh Soni, Mudit Verma, Siddharth Srivastava, and Subbarao Kambhampati. 2020. Bridging the gap: Providing post-hoc symbolic explanations for sequential decision-making problems with inscrutable representations. *arXiv preprint arXiv:2002.01080* (2020).

[38] Sarath Sreedharan, Siddharth Srivastava, and Subbarao Kambhampati. 2021. Using state abstractions to compute personalized contrastive explanations for AI agent behavior. *Artificial Intelligence* 301 (2021), 103570.

[39] Uri Stemmer. 2020. Locally Private k-Means Clustering. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms, SODA 2020, Salt Lake City, UT, USA, January 5-8, 2020*, Shuchi Chawla (Ed.). SIAM, 548–559. https://doi.org/10.1137/1.9781611975994.33

[40] Uri Stemmer and Haim Kaplan. 2018. Differentially private k-means with constant multiplicative error. *Advances in Neural Information Processing Systems* 31 (2018).

[41] Jasper van der Waa, Jurriaan van Diggelen, Karel van den Bosch, and Mark Neerincx. 2018. Contrastive explanations for reinforcement learning in terms of expected consequences. *arXiv preprint arXiv:1807.08706* (2018).

[42] Parisa Zehtabi, Alberto Pozanco, Ayala Bolch, Daniel Borrajo, and Sarit Kraus. 2024. Contrastive Explanations of Centralized Multi-agent Optimization Solutions. In *Proceedings of the International Conference on Automated Planning and Scheduling*, Vol. 34. 671–679.

Notations	Definitions	Note
$X \sim X'$	Two datasets X and X' differ by at most 1 element	
α	Approximation parameter for the utility of clustering and explanations	
β	Failure probability of the utility guarantees of clustering and explanations	
ζ	Parameter to control the utility of the private coreset: $\zeta = 0.01\left(\frac{\alpha}{10\lambda_{p,\alpha/2}}\right)^{1/p}$	
$\lambda_{p,\alpha/2}$	Definition 9	
$O_{p,\alpha}$	Big O notation that explicit ignore factors of p and α	
w, t	Approximation factors of the utility of our private explanations	Def. 2
w''	Approximation factors of the non-private clustering algorithm when one centroid is fixed	
OPT	Cost of the optimal clustering	
OPT_i	Cost the the optimal clustering when one centroid is fixed at requested location of agent i	
d, d'	The # of dimensions of the original and projected spaces	
S_ϵ	Private clustering solution with privacy budget ϵ	
$S_\epsilon^{(i)}$	Private solution computed by the clustering algorithm while Fixing one centroid to the requested position of agent i	
NonPrivateApprox	Any (not necessarily private) clustering algorithm with approximation factor $w' \leq w''$	
NonPrivateApproxFC	Any (not necessarily private) clustering algorithm with one centroid fixed as request, with approximation factor w''	
X	Input dataset in the original space	
X'	Projected dataset in the d'-dimensional space	in Theorem 3
Y	The private coreset	
$\mathcal{S}$	Projection from $\mathbb{R}^d$ to $\mathcal{S}$	

Table 2: Summary of repeatedly used notations and their definitions

A RELATED WORK: ADDITIONAL DETAILS

Our work considers differential privacy for explainable AI in general (XAI) and Multi-agent explanations (XMASE) in particular, focusing on post-hoc contrastive explanations for clustering.

Extensive experiments presented in [35] demonstrate non-negligible changes in explanations of black-box ML models through the introduction of privacy. The findings in [32] corroborate these observations regarding explanations for black-box feature-based models. These explanations involve creating local approximations of the model's behavior around specific points of interest, potentially utilizing sensitive data. In order to safeguard the privacy of the data used during the local approximation process of an eXplainable Artificial Intelligence (XAI) module, the researchers have devised an innovative adaptive differentially private algorithm. This algorithm is designed to determine the minimum privacy budget required to generate accurate explanations effectively. The study undertakes a comprehensive evaluation, employing both empirical and analytical methods, to assess how the introduction of randomness inherent in differential privacy algorithms impacts the faithfulness of the model explanations.

[29] considers feature-based explanations (e.g., SHAP) that can expose the top important features that a black-box model focuses on. To prevent such expose they introduced a new concept of achieving local differential privacy (LDP) in the explanations, and from that, they established a defense, called XRAND, against such attacks. They showed that their mechanism restricts the information that the adversary can learn about the top important features while maintaining the faithfulness of the explanations.

The analysis presented in restatable[14] considers security concerning contrastive explanations. The authors introduced the concept of the "explanation linkage attack", a potential vulnerability that arises when employing instance-based strategies to derive contrastive explanations. To address this concern, they put forth the notion of k-anonymous contrastive explanations. Furthermore, the study highlights the intricate balance between transparency, fairness, and privacy when incorporating k-anonymous explanations. As the degree of privacy constraints is heightened, a discernible trade-off comes into play: the quality of explanations and, consequently, transparency are compromised.

Amongst the three types of eXplainable AI mentioned earlier, the maintenance of privacy during explanation generation incurs a certain cost. This cost remains even if an expense was previously borne during the creation of the original model. However, in our proposed methodology for generating contrastive explanations in clustering scenarios, once the cost of upholding differential privacy in the initial solution is paid, no additional expenses are requisite to ensure differential privacy during the explanation generation phase.

Closer to our application is the study that investigates the privacy aspects concerning contrastive explanations in the context of team formation [11]. In this study, the authors present a comprehensive framework that integrates team formation solutions with their corresponding explanations, while also addressing potential privacy concerns associated with these explanations. To accomplish this, the authors introduce a privacy breach detector (PBD) that is designed to evaluate whether the provision of an explanation might lead to privacy breaches. The PBD consists of two main components: (a) A belief updater (BU), calculates the posterior beliefs that a user is likely to form after receiving

the explanation. (b) A privacy checker (PC), examines whether the user's expected posterior beliefs surpass a specified belief threshold, indicating a potential privacy breach. However, the research is still in its preliminary stages and needs a detailed evaluation of the privacy breach detector.

Our contribution includes the development of comprehensive algorithms for generating contrastive explanations with differential privacy guarantees. We have successfully demonstrated the effectiveness of these algorithms by providing rigorous proof for their privacy guarantees and conducting extensive experiments that showcased their accuracy and utility. In particular, we have shown the validity of our private explanations for clustering based on the k-median and k-means objectives for Euclidean datasets. Moreover, our algorithms have been proven to have the same accuracy bounds as the best private clustering methods, even though they provide explanations for all users, within a bounded privacy budget. Notably, our experiments in the dedicated experiments section reveal that the epsilon budget has minimal impact on the explainability of our results, further highlighting the robustness of our approach.

There has been a lot of work on private clustering and facility location, starting with [15], which was followed by a lot of work on other clustering problems in different privacy models, e.g., [9, 16, 30, 39, 40]. [15] demonstrated that the additive error bound for points in a metric space involves an $O(\Delta k^2 \log n/\epsilon)$ term, where Δ is the space's diameter. Consequently, all subsequent work, including ours, assumes points are restricted to a unit ball. In addition, there has been extensive work on a closely related problem, in the context of private clustering on graphs or networked data, often mentioned as community detection [6, 17, 28].

Algorithm 2 DimReduction
Input: $(x_1, x_2, \ldots, x_n), d, d', \beta$
Output: $(x'_1, \ldots, x'_n)$ low-dimensional space dataset.

$\Lambda = \sqrt{\frac{0.01d}{\log(n/\beta)d'}}$
for $i \in \{1, .., n\}$ **do**
  $\tilde{x}_i \leftarrow \Pi_{\mathcal{S}}(x_i)$
  if $\|\tilde{x}_i\| \leq 1/\Lambda$ **then**
    $x'_i = \Lambda \tilde{x}_i$
  else
    $x'_i = 0$
  end if
end for
return $(x'_1, \ldots, x'_n)$

Algorithm 3 DimReverse
Input: $(c'_1, \ldots, c'_k), (x'_1, \ldots, x'_n)$
Output: $(c_1, \ldots, c_k)$ Private Centroids in high dimension

$\mathbf{X}_1, \ldots, \mathbf{X}_k \leftarrow$ the partition induced by $(c'_1, \ldots, c'_k)$ on $(x'_1, \ldots, x'_n)$
for $j \in \{1, \ldots, k\}$ **do**
  $c_j \leftarrow \textsc{FindCenter}^{\epsilon/2}(\mathbf{X}_j)$
end for
return $(c_1, \ldots, c_k)$

B ADDITIONAL PROOFS FOR PRIVATEEXPLANATIONS

Theorem 1. DP of Explanation. *The solution $(c_1, \ldots, c_k)$ and cost(S_ϵ) computed by Algorithm PrivateExplanations are ϵ-DP. For all clients i and $S_\epsilon^{(i)}$ computed by Algorithm PrivateExplanations is ϵ-i-exclusion DP.*

Proof. It follows that cost(S_ϵ) is the direct results of Y, which is $\epsilon/2$-differentially private coreset. By the post-processing property, cost(S_ϵ) is $\epsilon/2$-DP (which implies ϵ-DP).

The output c of the DimReverse algorithm is $\epsilon/2$-differentially private with respect to the input $(X_1, X_2, \ldots, X_k)$, where X_i represents the data points in cluster i. The overall process achieves ϵ-differential privacy through composition, as $(X_1, X_2, \ldots, X_k)$ is partially derived from Y, which itself is $\epsilon/2$-differentially private.

For each explanations $S_\epsilon^{(i)}$, let $X, X' : X \setminus \{x_i\} \sim X' \setminus \{x_i\}$, i.e., X and X' are any two neighbor datasets that differ at exact one data point that is not agent i. Let $S_\epsilon^{(i)}(Y^X)$ be the value of $S_\epsilon^{(i)}$ with input dataset X (and Y^X as the private coreset of X respectively). Note that we specify the dataset X (and X') as a parameter of $S_\epsilon^{(i)}$ to highlight the original dataset where the explanation comes from (either X or X'). Fix

any set S, let $T = \{Y : S_\epsilon^{(i)}(Y) \in S\}$,i.e., the set of coresets Y that make $S_\epsilon^{(i)}(Y) \in S$. Since $X \sim X'$, we have:

$$\Pr[S_\epsilon^{(i)}(Y^X) \in S] = \Pr[Y^X \in T] \tag{7}$$

$$\leq e^{\epsilon/2} \Pr[Y^{X'} \in T] \tag{8}$$

$$= e^{\epsilon/2} \Pr[S_\epsilon^{(i)}(Y^{X'}) \in S], \tag{9}$$

which implies that $\text{cost}(S_\epsilon^{(i)}) - \text{cost}(S_\epsilon)$ is ϵ-x_i-exclusion DP, since $\text{cost}(S_\epsilon)$ is $\epsilon/2$-DP (which implies $\epsilon/2$-x_i-exclusion DP). □

Lemma 2. DP of fixed centroid yield additional cost. *Fix an i. If $OPT_i \geq w''(1+\alpha)OPT + t^{(i)}$, then with probability at least $1-2\beta$, cost(S_ϵ) and cost($S_\epsilon^{(i)}$) computed by Algorithm PrivateExplanations satisfies that $cost(S_\epsilon^{(i)}) > cost(S_\epsilon)$.*

Proof. By the result of Lemma 6, with probability $1-2\beta$ we have:

$$cost(S_\epsilon^{(i)}) \geq OPT_i \tag{10}$$

$$\geq w''(1+\alpha)OPT + t^{(i)} \tag{11}$$

$$\geq w''(1+\alpha)\frac{cost(S_\epsilon) - \Omega_{p,\alpha,w''}\left(\frac{(k/\beta)^{O_{p,\alpha}(1)}}{\epsilon}.\text{polylog}(n/\beta)\right)}{w''(1+\alpha)} + t^{(i)} \tag{12}$$

$$= cost(S_\epsilon) + t^{(i)} - \Omega_{p,\alpha,w''}\left(\frac{(k/\beta)^{O_{p,\alpha}(1)}}{\epsilon}.\text{polylog}(n/\beta)\right). \tag{13}$$

Set $t^{(i)} = \Omega_{p,\alpha,w''}\left(\frac{(k/\beta)^{O_{p,\alpha}(1)}}{\epsilon}.\text{polylog}(n/\beta)\right)$ and the Lemma follows. □

Definition 9. *For $p \geq 1, \alpha > 0$, $\lambda_{p,\alpha/2} \stackrel{def}{=} \frac{1+\alpha/2}{((1+\alpha/2)^{1/p}-1)^p}$.*

Lemma 6. *(Johnson-Lindenstrauss (JL) Lemma [7, 18]) Let v be any d-dimensional vector. Let $\mathcal{S}$ denote a random d'-dimensional subspace of $\mathbb{R}^d$ and let Π_S denote the projection from $\mathbb{R}^d$ to $\mathcal{S}$. Then, for any $\tau \in (0,1)$ we have*

$$\Pr\left[\|v\|_2 \approx_{1+\tau} \sqrt{d/d'}\|\Pi_{\mathcal{S}}(v)\|_2\right] \geq 1 - 2\exp\left(-\frac{d'\tau^2}{100}\right) \tag{14}$$

Lemma 7. *(Dimensionality Reduction for (k,p)-Cluster [21]) For every $\beta > 0, \tilde{\alpha} < 1, p \geq 1, k \in \mathbb{N}$, there exists $d' = O_{\tilde{\alpha}}(p^4 \log(k/\beta))$. Let $\mathcal{S}$ be a random d-dimensional subspace of $\mathbb{R}^d$ and $\Pi_{\mathcal{S}}$ denote the projection from $\mathbb{R}^d$ to $\mathcal{S}$. With probability $1-\beta$, the following holds for every partition $\mathcal{X} = (X_1, \ldots, X_k)$ of X:*

$$cost^p(\mathcal{X}) \approx_{1+\tilde{\alpha}} (d/d')^{p/2} cost^p(\Pi_{\mathcal{S}}(\mathcal{X})), \tag{15}$$

where $\Pi_{\mathcal{S}}(\mathcal{X})$ denotes the partition $(\Pi_{\mathcal{S}}(X_1), \ldots, \Pi_{\mathcal{S}}(X_k))$.

Theorem 3. Cost of explanations due to privacy. *Fix an agent i. With probability at least $1-\beta$, cost($S_\epsilon^{(i)}$) computed by Algorithm PrivateExplanations is a (w,t)-approximation of OPT_i, with*

$$w = w''(1+\alpha)$$

$$t = w'' O_{p,\alpha}\left((k/\beta)^{O_{p,\alpha}(1)}.polylog(n/\beta)/\epsilon\right)$$

Proof. Let $\tilde{\mathbf{X}} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$, i.e., the projected data after applying the transformation $\Pi_{\mathcal{S}}$. Let $X = (x'_1, x'_2, \ldots, x'_n)$, i.e., $\tilde{\mathbf{X}}$ after being clipped by $1/\Lambda$. By setting $\alpha' = 0.1\alpha$, and applying Lemma 7, we have:

$$OPT_i^{\tilde{d}} \leq \left(\frac{d'}{d}\right)^{p/2} (1+0.1\alpha) OPT_i^d. \tag{16}$$

By standard concentration, it can be proved that $\|x'_i\| \leq 1/\Lambda$ with probability $0.1\beta/n$ as follows:
Using Lemma 6, we have:

$$\Pr\left[\|x\| > \frac{1}{1+\tau}\sqrt{d/d'}\|x'\|\right] \geq 1 - 2\exp(-d'\tau^2/100). \tag{17}$$

Since x is in the unit ball, $\|x\| < 1$, which leads to:

$$\Pr\left[\|x'\| < (1+\tau)\sqrt{d'/d}\right] \geq 1 - 2\exp(-d'\tau^2/100). \tag{18}$$

Setting $\tau = \sqrt{\frac{log(n/\beta)}{0.01}} - 1$, $\Lambda = \frac{1}{1+\zeta}\sqrt{d/d'} = \sqrt{\frac{0.01}{\log(n/\beta)} \cdot \frac{d}{d'}}$, we have:

$$\begin{aligned}
\Pr[\|x'\| < 1/\Lambda] &\geq 1 - 2\exp(-\frac{d'\tau^2}{100}) && (19)\\
&> 1 - 2\exp(-d'\log(n/\beta)) && (20)\\
&> 1 - 2\beta/n, && (21)
\end{aligned}$$

By union bound on all i, then with probability at least $1 - 2\beta$, $x'_i = \Lambda\tilde{x}_i$ for all i. Since Y is the output of PrivateCoreset with input X' and ζ, then by Theorem 38 of [13], Y is a $(0.1\alpha, t)$-coreset of X' (with probability at least $1 - \beta$), with $\alpha = \frac{(100\zeta)^p}{10\lambda_{p,\alpha/2}}$ and t' as:

$$t' = O_{p,\alpha}\left(\frac{2^{O_{p,\alpha}(d')}k^2\log^2 n}{\epsilon}\log\left(\frac{n}{\beta}\right) + 1\right). \tag{22}$$

We note that alternatively, given a target approximation parameter α, we can set $\zeta = 0.01\left(\frac{\alpha}{10\lambda_{p,\alpha/2}}\right)^{1/p}$.
Let $(y_1, y_2, \ldots, y_k)$ be the solution of NonPrivateApproxFC in PrivateExplanations for a fixed i, $(y_1^*, y_2^*, \ldots, y_k^*)$ be the optimal solution of the clustering with fixed center at x'_i on X, OPT_Y be the optimal cost of the clustering with fixed center at x'_i on Y. By the w'-approximation property of NonPrivateApproximationFC, we have:

$$\begin{aligned}
cost_Y(y_1, y_2, \ldots, y_k) &\leq w' OPT_Y && (23)\\
&\leq w' cost_Y(y_1^*, y_2^*, \ldots, y_k^*) && (24)\\
&\leq w'(1+0.1\alpha) cost_{\mathbf{X}'_{1..n}}(y_1^*, y_2^*, \ldots, y_k^*) + w't' && (25)\\
&= w'(1+0.1\alpha) OPT_i^{d'} + w't'. && (26)
\end{aligned}$$

Composing with Lemma 7, we have:

$$\begin{aligned}
cost_Y(y_1, y_2, \ldots, y_k) &\leq w'(1+0.1\alpha) OPT_i^{d'} + w't' && (27)\\
&\leq \Lambda^p w'(1+0.1\alpha) OPT_i^{\tilde{d}} + w't' && (28)\\
&\leq \Lambda^p w'(1+0.1\alpha)(1+0.1\alpha)\left(\frac{d'}{d}\right)^{p/2} OPT_i^d + w't' && (29)\\
&\leq w'(1+\alpha) OPT_i^d \left(\frac{0.01}{\log(n/\beta)}\right)^{p/2} + w't', \text{ since } \Lambda^2 d'/d = \Theta(1/\log(n/\beta)) && (30)
\end{aligned}$$

Finally, since

$$cost(S_\epsilon^{(i)}) = cost_Y(y_1, y_2, \ldots, y_k)\left(\frac{\log(n/\beta)}{0.01}\right)^{p/2} \tag{31}$$

$$\leq w'(1+\alpha)OPT_i^d + \Theta(w't'(\log(n/\beta)^{p/2}) \tag{32}$$

$$\overset{(a)}{\leq} w'(1+\alpha)OPT_i^d + w'\Theta\left(\frac{2^{O_{p,\alpha}(d')}k^2\log^2 n}{\epsilon}\log\left(\frac{n}{\beta}\right)(\log(n/\beta))^{p/2}\right) \tag{33}$$

$$\overset{(b)}{\leq} w'(1+\alpha)OPT_i^d + w'\Theta\left(\frac{(k/\beta)^{O_{p,\alpha}(1)}k^2\log^2 n}{\epsilon}\log\left(\frac{n}{\beta}\right)(\log(n/\beta))^{p/2}\right) \tag{34}$$

$$\overset{(c)}{\leq} w'(1+\alpha)OPT_i^d + w'\Theta\left(\frac{(k/\beta)^{O_{p,\alpha}(1)}(k/\beta)^2\log^2(n/\beta)}{\epsilon}\log\left(\frac{n}{\beta}\right)(\log(n/\beta))^{p/2}\right) \tag{35}$$

$$= w'(1+\alpha)OPT_i^d + w'\Theta\left(\frac{(k/\beta)^{O_{p,\alpha}(1)}}{\epsilon}.\text{polylog}(n/\beta)\right), \tag{36}$$

$$\tag{37}$$

where in (a) we substitute the value of t', and (b) is because $d' = O_\alpha(p^4\log(n/\beta))$, and (c) is because $\beta < 1$, and the Lemma follows. □

C ADDITIONAL PROOFS FOR k-MEDIAN ALGORITHM

Definition 10. *The solution of the k-median problem (with demands and a center fixed at a location z can be formulated as finding the optimal solution of the following Integer program (IP):*

$$minimize \sum_{i,j\in N} d_j c_{ij} x_{ij} \tag{38}$$

$$subject\ to \sum_{i\in N} x_{ij} = 1\ for\ each\ j \in N \tag{39}$$

$$x_{ij} \leq y_i\ for\ each\ i, j \in N \tag{40}$$

$$\sum_{j\in N} y_i = k \tag{41}$$

$$x_{ij} \in \{0,1\}\ for\ each\ i, j \in N \tag{42}$$

$$y_i \in \{0,1\}\ for\ each\ i \in N \tag{43}$$

$$y_z = 1\ for\ a\ fixed\ z \in N \tag{44}$$

$$x_{zz} = 1\ for\ a\ fixed\ z \in N. \tag{45}$$

Lemma 8. *Locations $i, j \in N'$ satisfy: $c_{ij} > 4\max(\bar{C}_i, \bar{C}_j)$.*

Proof. The lemma follows the demands moving step (in the first step of the algorithm): for every j to the right of i (which means $\bar{C}_j \geq \bar{C}_i$) and within the distance of $\bar{C}_j$ (that also covers all points within distance $\bar{C}_i$), we move all demands of j to i, hence j will not appear in N'. □

D ADDITIONAL PROOFS FOR k-MEANS ALGORITHM

Lemma 9. *(Lemma 2.1 of [19]) Given a finite subset S of points in $\mathbb{R}^d$, let c be the centroid of S. Then for any $c' \in \mathbb{R}^d$, $\Delta(S, c') = \Delta(S, c) + |S|\Delta(c, c')$.*

Lemma 10. *Let S be 1-stable set and O be the optimal set of k centers, we have $\Delta(O) - 3\Delta(S) + 2R \geq 0$, where $R = \sum_{q\in P}\Delta(q, s_{o_q})$.*

Proof. Since S is 1-stable, we have for each swap pair:

$$\sum_{q\in N_O(o)}\left(\Delta(q,o) - \Delta(q,s_q)\right) \tag{46}$$

$$+ \sum_{q\in N_S(s)\setminus N_O(o)}\left(\Delta(q,s_{o_q}) - \Delta(q,s)\right) \geq 0. \tag{47}$$

We will sum up the inequality above for all swap pairs. For the left term, the sum is overall $o \in O$:

$$\sum_{o\in O}\sum_{q\in N_O(o)}\big(\Delta(q,o)-\Delta(q,s_q)\big) \tag{48}$$

$$=\sum_{q\in P}\big(\Delta(q,o)-\Delta(q,s_q)\big), \tag{49}$$

Since each $o \in O$ will appear exactly once, and $\cup_{o\in O} q \in N_O(o)$ will cover all points in P.
For the right term, the sum is over all s that is being swapped out. We note that each s can be swapped out at most twice, hence:

$$\sum_{s \text{ being swapped out}}\sum_{q\in N_S(s)}\big(\Delta(q,s_{o_q})-\Delta(q,s)\big) \tag{50}$$

$$\le 2\sum_{q\in P}\big(\Delta(q,s_{o_q})-\Delta(q,s)\big) \tag{51}$$

When we combine the two terms, we have:

$$\sum_{q\in P}\big(\Delta(q,o)-\Delta(q,s_q)\big)+2\sum_{q\in P}\big(\Delta(q,s_{o_q})-\Delta(q,s)\big)\ge 0 \tag{52}$$

$$\sum_{q\in P}\Delta(q,o_q)-3\sum_{q\in P}\Delta(s,s_q)+2\sum_{q\in P}\Delta(q,s_{o_q})\ge 0 \tag{53}$$

$$\Delta(O)-3\Delta(S)+2R\ge 0, \tag{54}$$

and the Lemma follows. □

Lemma 11. *(Proof in Lemma 2.2 & 2.3 of [19]) Let* $\alpha^2=\frac{\Delta(S)}{\Delta(O)}$*, we have* $\sum_{q\in P} dist(q,o_q)dist(q,s_q)\le\frac{\Delta(S)}{\alpha}$

Lemma 12. *With R and* α *defined as above:* $R\le 2\Delta(O)+(1+2/\alpha)\Delta(S)$.

Proof. By Lemma 10, we have:

$$R=\sum_{q\in P}\Delta(q,s_{o_q}) \tag{55}$$

$$=\sum_{o\in O}\sum_{q\in N_O(o)}\Delta(q,s_o) \tag{56}$$

$$=\sum_{o\in O\setminus\sigma}\sum_{q\in N_O(o)}\Delta(q,s_o)+\sum_{q\in N_O(\sigma)}\Delta(q,\sigma) \tag{57}$$

$$=\sum_{o\in O\setminus\sigma}\Delta(N_O(o),s_o)+\Delta(N_O(\sigma),\sigma) \tag{58}$$

$$\overset{(a)}{=}\sum_{o\in O\setminus\sigma}\big(\Delta(N_O(o),o)+|N_O(o)|\Delta(o,s_o)\big)+\Delta(N_O(\sigma),\sigma) \tag{59}$$

$$=\sum_{o\in O\setminus\sigma}\sum_{q\in N_O(o)}\big(\Delta(q,o)+\Delta(o,s_o)\big)+\sum_{q\in N_O(\sigma)}\Delta(q,\sigma)+\Delta(\sigma,o_\sigma) \tag{60}$$

$$=\sum_{o\in O}\sum_{q\in N_O(o)}\big(\Delta(q,o)+\Delta(o,s_o)\big) \tag{61}$$

$$\overset{(b)}{=}\sum_{o\in O}\sum_{q\in N_O(o)}\big(\Delta(q,o)+\Delta(o,s_q)\big) \tag{62}$$

$$\le\sum_{q\in P}\big(\Delta(q,o_q)+\Delta(o_q,s_q)\big) \tag{63}$$

$$\overset{(c)}{\le}\Delta(O)+\sum_{q\in P}(dist(o_q,q)+dist(q,s_q))^2 \tag{64}$$

$$=2\Delta(O)+\Delta(S)+2\sum_{q\in P}dist(q,o_q)dist(q,s_q) \tag{65}$$

$$\overset{(d)}{\le}2\Delta(O)+\Delta(S)+(2/\alpha)\Delta(S), \tag{66}$$

where (a) is because Lemma 9 applies for all $o \in O \setminus \sigma$, (b) is because $\Delta(o, s_o) \leq \Delta(o, s_q)$, (c) is because the triangle inequality applies for $\Delta(o_q, s_q)$ and (d) is because of Lemma 11 and the Lemma follows. □

D.1 Swap pairs mapping

In this section, we describe the swap pairs mapping scheme for the k-means with a fixed center algorithm. We adapt the scheme of [22] to accommodate the fixed center. We discuss the modifications in Section 4.2. Here we discuss the complete mapping scheme.

At the last iteration of the algorithm, we always have a candidate set of centers S that is 1-stable, i.e., no single feasible swap can decrease its cost. We then analyze some hypothetical swapping schemes, in which we try to swap a center $s \in S$ with an optimal center $o \in O$. We utilize the fact that such single swaps do not decrease the cost to create some relationships between $\Delta(S)$ and $\Delta(O)$–the optimal cost. Particularly, these relationships are stated in Lemma 10 and Lemma 12.

Let σ be the fixed center. We note that $\sigma \in S$ and $\sigma \in O$. Let s_o be the closest center in S for an optimal center $o \in O$, which means o is captured by s_o. It follows that $s_\sigma = \sigma$. A center $s \in S$ may capture no optimal center (we call it lonely). We partition both S and O into $S_1, \ldots, S_r$ and $O_1, \ldots, O_r$ that $|S_i| = |O_i|$ for all i.

We construct each pair of partitions S_i, O_i as follows: let s be a non-lonely center, $O_i = \{o \in O : s_o = s\}$, i.e., O_i is the set of all optimal centers that are captured by s. Now, we compose s with $|O_i| - 1$ lonely centers (which are not partitioned into any group from S) to form S_i. It is clear that $|S_i| = |O_i| \geq 1$.

We then generate swap pairs for each pair of partitions S_i, O_i by the following cases:

- $|S_i| = |O_i| = 1$: let $S_i = \{s\}, O_i = \{o\}$, generate a swap pair $\{s, o\}$.
- $|S_i| = |O_i| = m > 1$: let $S_i = \{s, s_{1..m-1}\}$ in which $s_{1..m-1}$ are $m-1$ lonely centers, let $O_i = \{o_{1..m}\}$, generate $m-1$ swap pairs $\{s_j, o_j\}$ for $j = 1..m-1$. Also, we generate a swap pair of $\{s_1, o_m\}$. Please note that s does not belong to any swap pair, each o_j belongs to exactly one swap pair, and each s_j belongs to at most two swap pairs.

We then guarantee the following 3 properties of our swap pairs:

(1) each $o \in O$ is swapped in exactly once
(2) each $s \in S$ is swapped out at most twice
(3) for each swap pair $\{s, o\}$, s either captures only o, or s is lonely (captures nothing).

D.2 γ-approximate candidate center set for fixed-center k-means.

We describe how to generate a γ-approximate candidate center set for k-means with fixed center σ for a dataset $X \subset \mathbb{R}^d$. From the result of [22], we create a set C' which is a γ-approximation centroid set of X. We will prove that $C = C' \cup \{\sigma\}$ forms a γ-approximate candidate center set for k-means with fixed center σ.

Definition 11. *Let $S \subset \mathbb{R}^d$ be a finite set with its centroid $c(S)$. A γ-tolerance ball of S is the ball centered at $c(S)$ and has radius of $\frac{\gamma}{3}\rho(S)$.*

Definition 12. *Let $X \subset \mathbb{R}^d$ be a finite set. A finite set $C' \in \mathbb{R}^d$ is a γ-approximation centroid set of X if C' intersects the γ-tolerance ball of each nonempty $S \subseteq X$.*

Lemma 13. *(Theorem 4.4 of [22]) We can compute C'–a γ-approximation centroid set of X that has size of $O(n\gamma^{-d}\log(1/\gamma))$ in time $O(n \log n + n\gamma^{-d}\log(1/\gamma))$.*

Theorem 6. *Let $C = C' \cup \{\sigma\}$, in which C' is a γ-approximation centroid set computed as Lemma 13, then C is a γ-approximate candidate center set for k-means with fixed center σ.*

Proof. Let $O = (O_1, O_2, \ldots, O_k)$ be the optimal clustering in which O_1 is the cluster whose center is σ (we denote it as O_σ). For any $S \subset \mathbb{R}^d$, we define $cost_S(c) = \sum_{x \in S} \|x - c\|^2$ and $cost(S) = cost_S(c(S))$ in which $c(S)$ is the centroid of S. By Definition 7, we will prove that there exists a set $c_1, c_2, \ldots, c_k \subset C$ and $c_1 = \sigma$ such that $cost(c_1, c_2, \ldots, c_k) \leq (1+\gamma)cost(O)$. We adapt the analysis of [22] for the special center σ–which is not a centroid as other centers in k-means.

First, we analyze the optimal cost. For any cluster except O_σ, its center is also the centroid $c(O_i)$ of the cluster, while O_σ must have center σ:

$$cost(O) = \sum_{x \in O_\sigma} \|x - \sigma\|^2 + \sum_{i=2..k} \sum_{x \in O_i} \|x - c(O_i)\|^2 \tag{67}$$

$$= cost_{O\sigma}(\sigma) + \sum_{i=2..k} cost(O_i) \tag{68}$$

Now, we construct $\{c_1, \ldots, c_k\}$ as follows: setting $c_1 = \sigma$, for $i = 2..k$, $c_i \in C'$ is the candidate center that intersects the γ-tolerance ball of cluster O_i. For O_σ, $cost_{O_\sigma}(\sigma) = cost(O_\sigma)$. For other clusters, $cost_{O_i}(c_i) \leq (1+\gamma)cost(O_i)$ as below:

$$cost_{O_i}(c_i) = \sum_{x \in O_i} \|x - c_i\|^2 \tag{69}$$

$$\leq \sum_{x \in O_i} (\|x - c(O_i)\| + \|c(O_i) - c_i\|)^2 \tag{70}$$

$$= cost(O_i) + 2\|c_i - c(O_i)\| \sum_{x \in O_i} \|x - c(O_i)\| + |O_i|\|c_i - c(O_i)\|^2 \tag{71}$$

$$\leq cost(O_i) + 2\gamma/3\rho(O_i)\sqrt{|O_i|}\sqrt{cost(O_i)} + |O_i|(\gamma/3\rho(O_i))^2 \tag{72}$$

$$\leq cost(O_i) + (2/3)\gamma cost(O_i) + (\gamma^2/9)cost(O_i) \tag{73}$$

$$\leq (1+\gamma/3)^2 cost(O_i) \tag{74}$$

$$\leq (1+\gamma)cost(O_i). \tag{75}$$

Let $(S_1, S_2, \ldots, S_k)$ be the Voronoi partition with centers $(c_1, c_2, \ldots, c_k)$, i.e., S_i are points in the Voronoi region of c_i in the Voronoi diagram created by $c_1, \ldots, c_k$, we have:

$$cost(c_1, c_2, \ldots, c_k) = cost_{S_1}(\sigma) + \sum_{i=2..k} cost(S_i) \tag{76}$$

$$\overset{(a)}{\leq} cost_{S_1}(\sigma) + \sum_{i=2..k} cost_{S_i}(c_i) \tag{77}$$

$$\overset{(b)}{\leq} cost_{O_\sigma}(\sigma) + \sum_{i=2..k} cost_{O_i}(c_i) \tag{78}$$

$$\overset{(c)}{\leq} cost_{O_\sigma}(\sigma) + (1+\gamma) \sum_{i=2..k} cost(O_i) \tag{79}$$

$$\leq (1+\gamma)cost(O), \tag{80}$$

where (a) is because $cost(S_i)$ implies its minimal cost for any center, (b) is because S_is are picked by Voronoi partition which minimies the cost over k partitions of seletecd k centers, and (c) is because $cost_{O_i}(c_i) \leq (1+\gamma)cost(O_i)$ as we proved above, and the Theorem follows. □

D.3 Additional details on experiments

D.3.1 Datasets and Experimental Setup

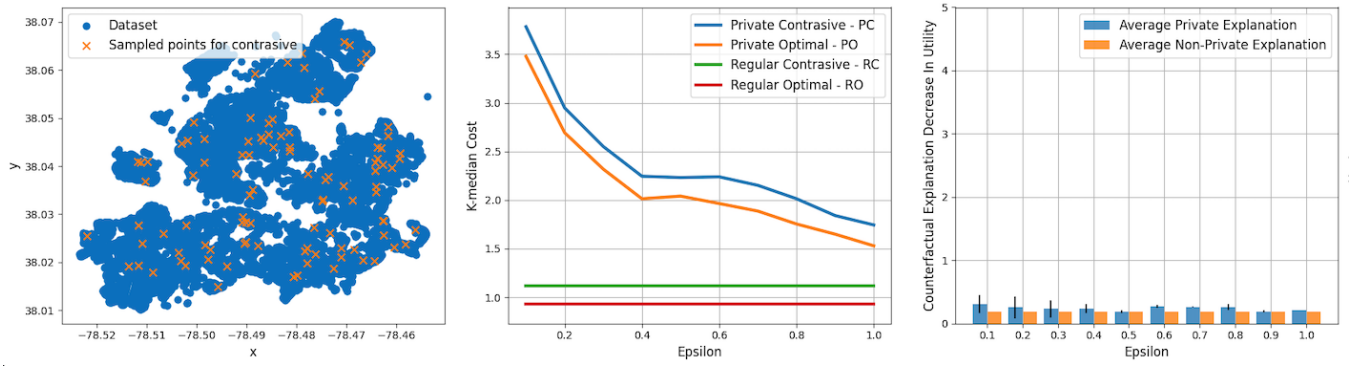

Figure 3: A detailed visualization of our dataset (Charlottesville County, Virginia) and analysis includes (a) A scatter plot of the full dataset with 100 randomly selected points for contrastive analysis, chosen to provide a more comfortable and manageable subset for explanation purposes. (b) Comparison of k-median clustering with fixed and non-fixed centroids, both private and non-private. (c) Bar graph showing contrastive explanation differences for differential private and non-private k-median with a fixed centroid.

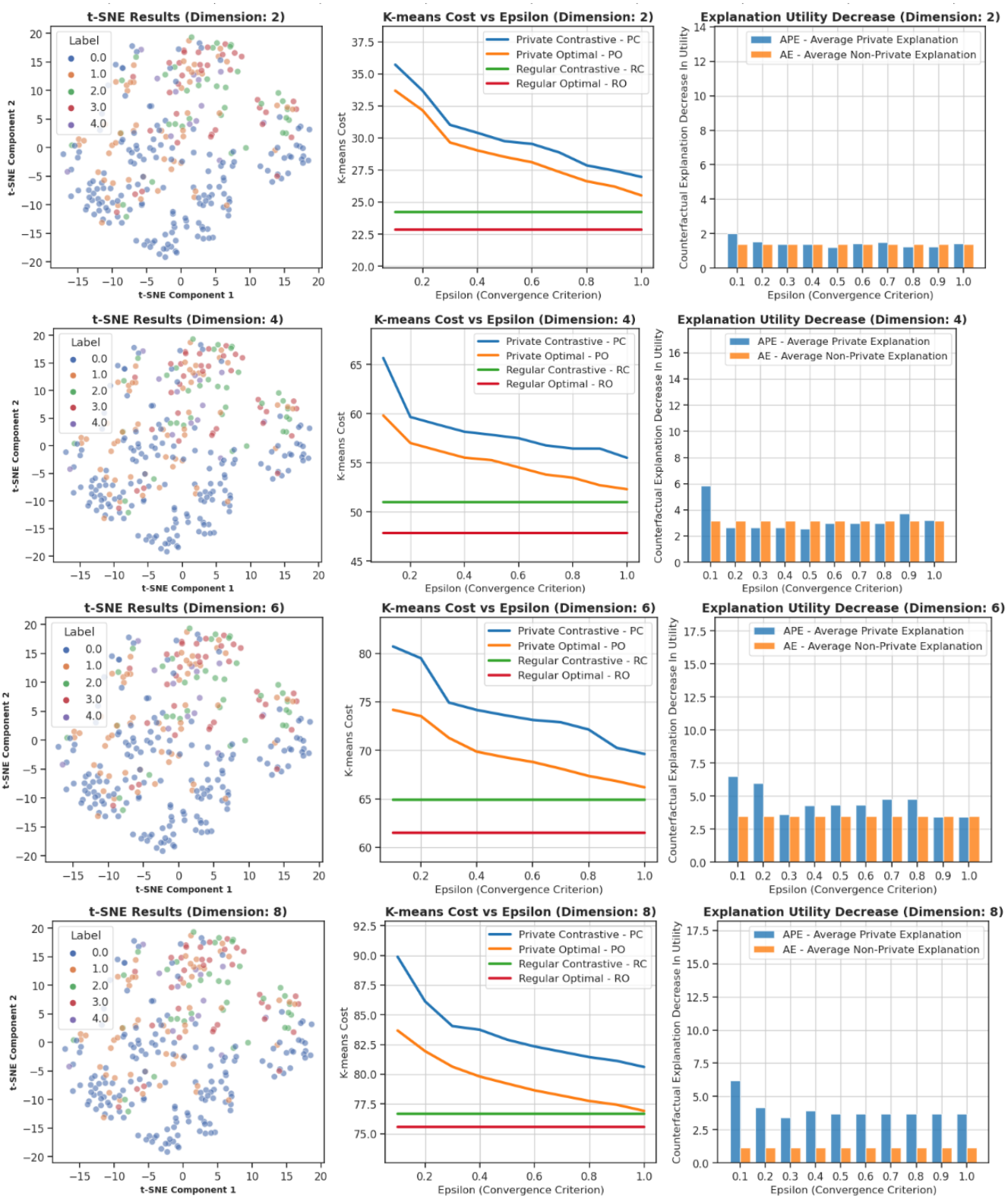

Figure 4: The figure presents visualizations of the dataset (Heart Disease, UCI MLR) reduced to four dimensions (2, 4, 6, and 8), along with the corresponding analysis (a) A t-SNE scatter plot illustrating the high-dimensional data. (b) Comparison of k-means clustering with fixed and non-fixed centroids, both private and non-private. (c) Bar graph showing contrastive explanation differences for differential private and non-private k-means with a fixed centroid.

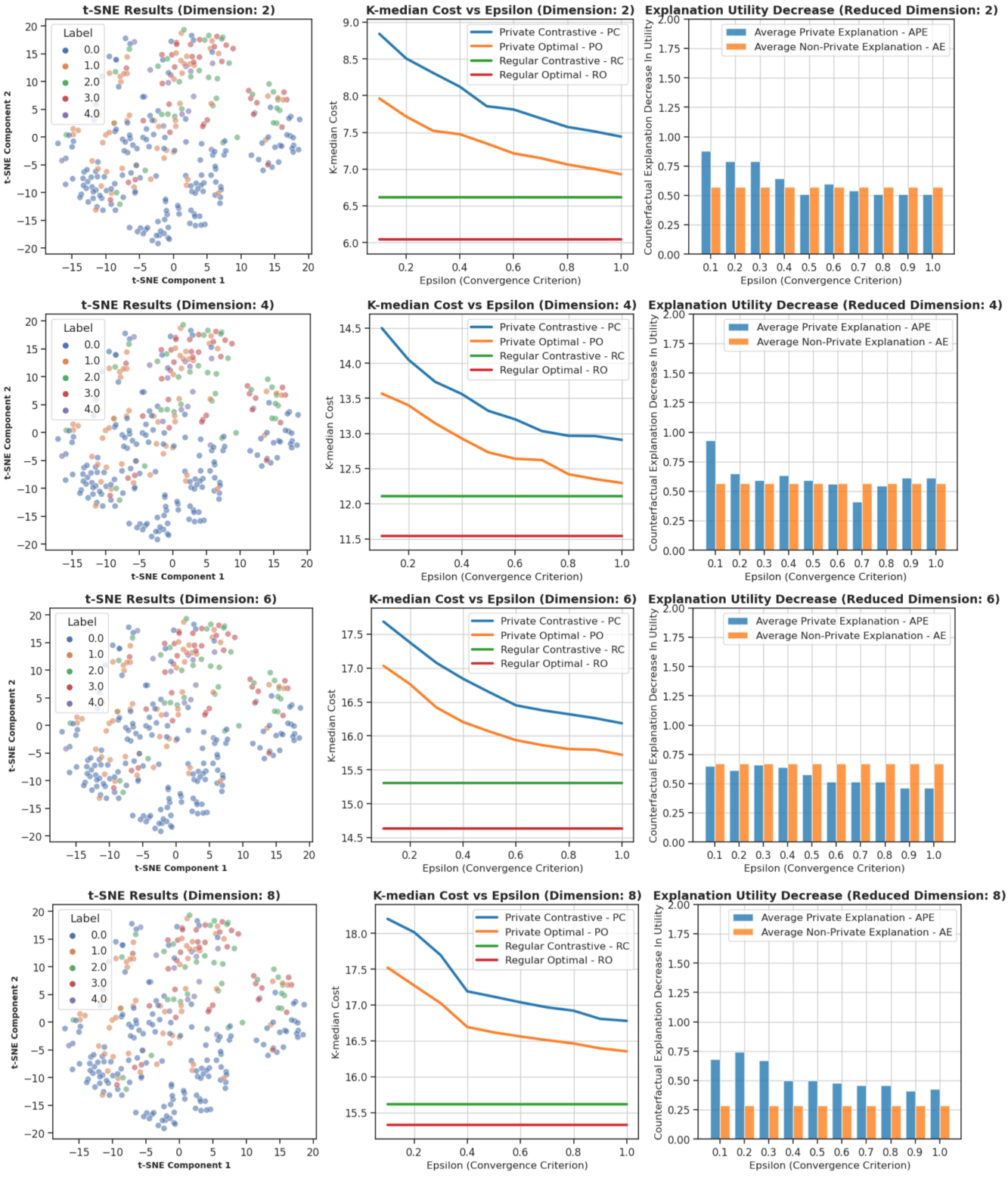

Figure 5: The figure presents further visualizations of the dataset (Heart Disease, UCI MLR) introduced in the main paper, reduced to four dimensions (2, 4, 6, and 8), along with the corresponding analysis (a) A t-SNE scatter plot illustrating the high-dimensional data. (b) Comparison of k-median clustering with fixed and non-fixed centroids, both private and non-private. (c) Bar graph showing contrastive explanation differences for differential private and non-private k-median with a fixed centroid.

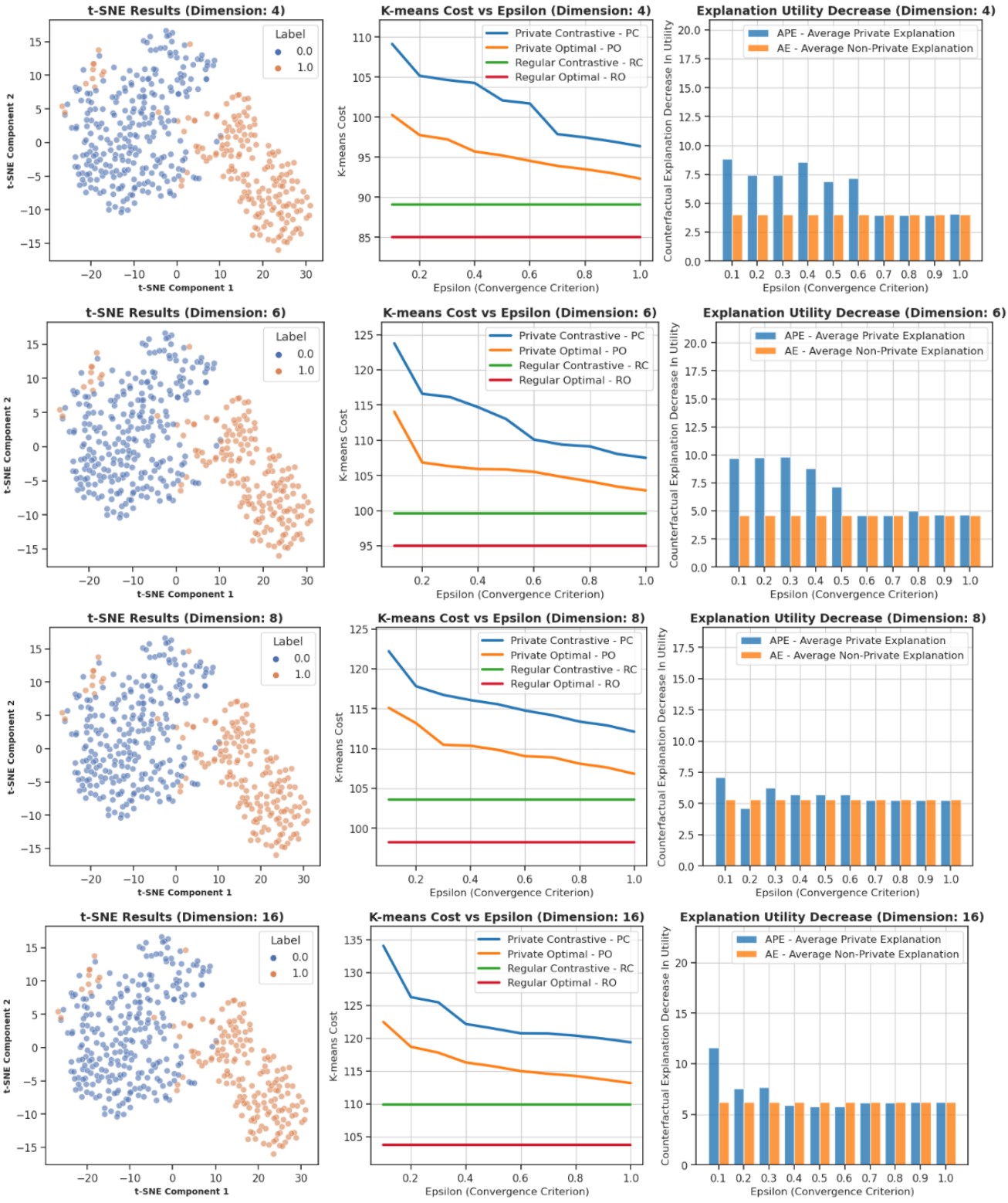

Figure 6: The figure presents visualizations of the dataset (Breast Cancer - 30 features, UCI MLR) reduced to four dimensions (4, 6, 8, and 16), along with the corresponding analysis (a) A t-SNE scatter plot illustrating the high-dimensional data. (b) Comparison of k-means clustering with fixed and non-fixed centroids, both private and non-private. (c) Bar graph showing contrastive explanation differences for differential private and non-private k-means with a fixed centroid.

In the main body of the paper, we focused on the Heart Disease dataset from the UCI Machine Learning Repository, which was reduced from 13 features to 8 features for visualization purposes. However, we have also conducted experiments on the same dataset with different dimensionality reductions to analyze the impact on our results. Furthermore, to demonstrate the robustness of our work, we have applied our analysis to additional datasets, including synthetic datasets and real-world datasets from Charlottesville, Albemarle and Breast Cancer Wisconsin dataset including 30 features. This appendix provides a detailed discussion of these experiments and their outcomes.

Real-World Datasets

To demonstrate the robustness and real-world applicability of our approach, we conducted experiments on real-world datasets: the Charlottesville City dataset, the Albemarle County dataset, Heart Disease and Breast Cancer Wisconsin.

Charlottesville City Dataset

The Charlottesville City dataset is part of the synthetic U.S. population, as described in [Chen et al., 2021] and [Barrett et al., 2009]. This dataset consists of approximately 33,000 individuals and around 5,600 activity locations visited by these individuals. The locations represent places where individuals perform various activities, providing insights into human mobility patterns and social interactions within the city.

Albemarle County Dataset

The Albemarle County dataset is another real-world dataset used in our experiments. This dataset is significantly larger than the Charlottesville City dataset, comprising about 74,000 individuals. The increased sample size allows us to evaluate the scalability and performance of our approach when applied to larger, more complex datasets.

The Albemarle County dataset contains information about individuals' activities and the locations they visit, similar to the Charlottesville City dataset. This dataset provides a comprehensive representation of human mobility patterns and social interactions within the county.

By using these real-world datasets, we aim to validate the effectiveness and practicality of our methodology in real-life scenarios. The diverse characteristics of these datasets, such as the number of individuals and activity locations, enable us to assess the robustness and generalizability of our approach.

In our experiments, we applied our methodology to all datasets and compared the results to those obtained from the synthetic datasets. The consistency of results across these real-world datasets further reinforces the reliability and potential of our approach for real-world applications.

Synthetic 2D Dataset: We carefully created a synthetic dataset that mimics the properties of real-world datasets, striking a balance between realism and controlled variability. This dataset consists of 1000 uniformly distributed data points in a 2D space, with a range similar to the real datasets we analyzed.

The primary motivation behind this synthetic dataset is to provide a sandbox environment free from the unpredictable noise and anomalies of real-world data. This controlled setting is pivotal in understanding the core effects of differential privacy mechanisms, isolating them from external confounding factors. The dataset is a foundational tool in our experiments, allowing us to draw comparisons and validate our methodologies before applying them to more complex, real-world scenarios.

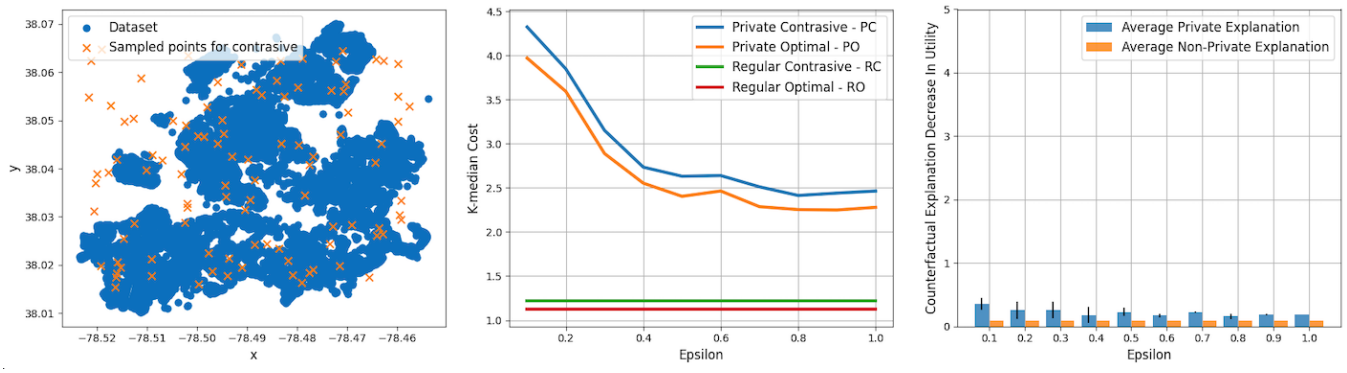

Figure 7: A detailed visualization of our synthetic dataset, measured by the same metrics as the other real-world datasets.

D.3.2 Experimental Results

Impact of ϵ on Private Optimal (PO) and Private Contrastive (PC):. We observed consistent trends and patterns in all our data sets, including Charlottesville city, Albemarle county, and notably the Heart Disease dataset and the Breast Cancer dataset from the UCI Machine Learning Repository. As the value of ϵ increased to prioritize accuracy, we observed a gradual reduction in privacy protection. However, in line with

our hypothesis, the impact of the epsilon budget on the explainability of our outcomes remained minimal. This consistency held true across all dimensions of the Heart Disease dataset, reinforcing the robustness of our findings across diverse data sources and attributes.

D.3.3 Performance Evaluation:

For each ϵ value, we conducted 100 different runs for each dataset. The average results were consistent with our findings across all datasets. It's essential to note that these multiple invocations were solely for performance evaluation. In real-world applications, invoking private algorithms multiple times could degrade the privacy guarantee.

Consistency in Contrastive Explanations across Datasets: Despite the distinct scales between our different datasets, we observed consistent patterns in the contrastive explanations. Specifically, as illustrated in all Figures - (b), contrastive explanations remained largely unaffected by variations in the ϵ budget. This consistency further reinforces our hypothesis that the epsilon budget has a negligible influence on the explainability of our outcomes, even when applied to datasets of different scales.

D.3.4 Conclusion:

The extended experiments on all our datasets further validate our approach's efficacy. The balance between privacy and utility, the robustness of contrastive explanations, and the negligible impact of ϵ on explainability were consistent across datasets. These findings underscore the potential of our method for diverse real-world applications.