# Topological Federated Clustering via Gravitational Potential Fields Under Local Differential Privacy

**Yunbo Long**[*1], **Jiaquan Zhang**[*2,4], **Xi Chen**[2,3†], **Alexandra Brintrup**[1,5]

[1]Department of Engineering, University of Cambridge
[2]Artificial Intelligence Innovation and Incubation Institute, Fudan University
[3]Shanghai Academy of AI for Science
[4]Shanghai Innovation Institute
[5]The Alan Turing Institute

## Abstract

Clustering non-independent and identically distributed (non-IID) data under local differential privacy (LDP) in federated settings presents a critical challenge: preserving privacy while maintaining accuracy without iterative communication. Existing one-shot methods rely on unstable pairwise centroid distances or neighborhood rankings, degrading severely under strong LDP noise and data heterogeneity. We present Gravitational Federated Clustering (GFC), a novel approach to privacy-preserving federated clustering that overcomes the limitations of distance-based methods under varying LDP. Addressing the critical challenge of clustering non-IID data with diverse privacy guarantees, GFC transforms privatized client centroids into a global gravitational potential field where true cluster centers emerge as topologically persistent singularities. Our framework introduces two key innovations: (1) a client-side compactness-aware perturbation mechanism that encodes local cluster geometry as "mass" values, and (2) a server-side topological aggregation phase that extracts stable centroids through persistent homology analysis of the potential field's superlevel sets. Theoretically, we establish a closed-form bound between the privacy budget $\epsilon$ and centroid estimation error, proving the potential field's Lipschitz smoothing properties exponentially suppress noise in high-density regions. Empirically, GFC outperforms state-of-the-art methods on ten benchmarks, especially under strong LDP constraints ($\epsilon < 1$), while maintaining comparable performance at lower privacy budgets. By reformulating federated clustering as a topological persistence problem in a synthetic physics-inspired space, GFC achieves unprecedented privacy-accuracy trade-offs without iterative communication, providing a new perspective for privacy-preserving distributed learning.

**Code** — https://github.com/Yunbo-max/Topological-GFC

## Introduction

Federated learning (FL) has emerged as a crucial privacy-preserving paradigm for training machine learning models on decentralized data (McMahan et al. 2017). Although initially developed for supervised tasks, FL is increasingly being adapted to unsupervised learning scenarios, particularly clustering, to analyze distributed unlabeled data. Standard federated clustering methods, such as federated k-means (Dennis, Li, and Smith 2021a), exchange aggregated data representations (e.g., cluster centroids) rather than raw client data (Garst and Reinders 2024). While this prevents direct exposure of individual data points, it fails to provide formal privacy guarantees. For example, a hospital with few pediatric cancer patients may publish a centroid so close to individual records that adversaries can reconstruct original diagnoses, e.g. in extreme conditions only one patient is recorded, hence the centroid is this individual's data. Differential privacy (DP) provides rigorous privacy guarantees by injecting calibrated noise into computations. While centralized DP (Demelius, Kern, and Trügler 2025) protects only the final aggregated output, local differential privacy (LDP) (Xia et al. 2020) enforces privacy at the client level by requiring data randomization before any communication occurs. This is critical for federated learning scenarios where the server cannot be trusted. However, LDP poses significant challenges for clustering algorithms: the client-side noise required to achieve strong privacy ($\varepsilon < 1.0$) often disrupts the geometric relationships between points that clustering relies upon, leading to degraded utility.

Recently, FedDP-KMeans demonstrated the effectiveness of federated clustering under differential privacy by proposing an improved cluster initialization method combined with DP-Lloyds (Scott, Lampert, and Saulpic 2025). However, its multi-round communication exacerbates privacy costs due to composition, as each Lloyd's iteration requires fresh client-server interaction. This results in increased latency and bandwidth overhead. Such constraints explain the preference for one-shot methods in production FL systems, where communication efficiency is critical—especially given heterogeneous client connectivity. For instance, K-FED (Dennis, Li, and Smith 2021a) clusters local centroids to derive global ones, while MUFC (Pan et al. 2023) enhances performance on imbalanced, non-IID data. The key challenge for one-shot federated clustering lies in accurately estimating global centroids under varying privacy budgets while limiting communication to a single

[*]These authors contributed equally.
[†]Xi Chen is the corresponding author

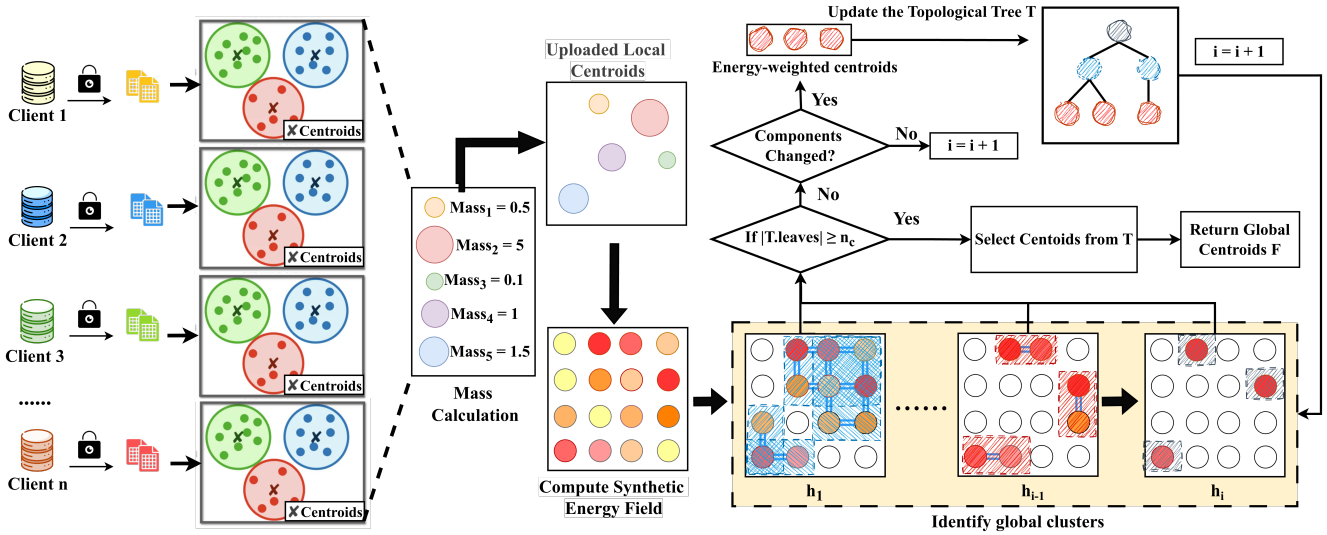



Figure 1: Gravitational Federated Clustering Pipeline.

round—eliminating the possibility of iterative refinement. Existing theoretical guarantees rely heavily on the assumption that geometric properties(connectedness and distance) remain intact after noise injection(Wang, Pang, and Pedrycz 2024). However, this assumption breaks down in practical settings with strict privacy budgets (e.g., $\epsilon < 1$), where the noise magnitude becomes large enough to severely distort the underlying data structure, as evidenced by the catastrophic failure of baseline methods in Figure 3.

Therefore, we propose Gravitational Federated Clustering (GFC), a one-shot federated clustering method designed to handle varying Local Differential Privacy (LDP) constraints. Unlike existing approaches that rely on noise-corrupted distance metrics or raw centroids, GFC reformulates clustering as a dynamic topological feature extraction process for global centroid search. By constructing a gravitational potential field from privatized client data (augmented with synthetic data), GFC robustly adapts to privacy budgets ranging from $\epsilon = 1000$ (weak privacy) to $\epsilon = 0.01$ (extremely strong privacy). To summarize, our main contributions are as follows:

- We propose the first one-shot federated clustering method GFC, that effectively handles varying LDP by modeling clustering as a topological persistence problem within a gravitational potential field. This avoids reliance on noise-sensitive distance metrics.
- We provide a theoretical analysis of topological feature extraction via homotopy equivalence and introduce an efficient tree-splitting algorithm to recover cluster structures from the potential field. The final centroids are then derived from the leaves of the constructed tree.
- Extensive experiments on ten real-world federated benchmarks show that GFC outperforms state-of-the-art one-shot methods under varying differential privacy (DP) constraints, particularly for small privacy budgets ($\varepsilon < 1$). We further present a *privacy-accuracy boundary analysis* to quantify the relationship between centroid accuracy and the privacy budget.

## Related Work

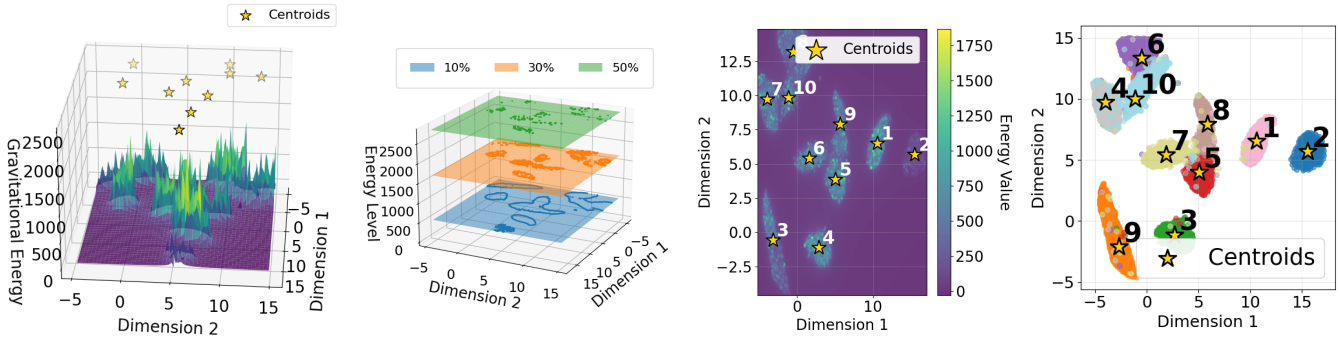
Our work addresses the central challenge of achieving accurate clustering under Local Differential Privacy (LDP) in federated settings.

### Federated Clustering

Federated clustering enables decentralized data analysis by preserving privacy, avoiding raw data centralization. Early iterative methods like federated k-means required many communication rounds to refine centroids (Garst and Reinders 2024). To enhance efficiency, one-shot methods became standard, where clients perform local clustering and send results (e.g., centroids or density cores) to the server for a single aggregation. Key examples include K-FED (Dennis, Li, and Smith 2021a), MUFC (Pan et al. 2023), and NNFC (Chen et al. 2024), which use advanced aggregation to handle non-IID and imbalanced data. To address data heterogeneity, Clustered FL (CFL) groups clients by distribution to train specialized models (Sattler, Müller, and Samek 2020; Ghosh et al. 2021; Long et al. 2025), enhancing performance and fairness (Zhang et al. 2024; Gupta et al. 2024). However, these methods are vulnerable under strong privacy constraints.

### Differential Privacy in Federated Clustering

In differential privacy, bounded sensitivity is enforced by clipping each data point to satisfy $|\mathbf{x}_i|_1 \leq \Delta$, where $\Delta$ controls the maximum influence of any single sample and $\epsilon$ determines the privacy-utility trade-off. Clients then release privatized data as $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\eta}_i$, with $\boldsymbol{\eta}_i \sim \text{Laplace}(0, \Delta/\epsilon)$, and we refer to $\epsilon$-differential privacy as DP with noise $\eta \sim$




(a) 3D Plot (b) Contours (c) 2D Gravitational Field (d) Clustering

Figure 2: Examples of Topological Analysis for GFC on MNIST Data Visualized via UMAP Projection

Laplace$(0, 1/\epsilon)$. To ensure formal privacy guarantees in federated clustering with an untrusted server, we adopt client-level local differential privacy (LDP), which treats the entire client dataset $\mathcal{D}_m$ as the privacy unit, following (Scott, Lampert, and Saulpic 2025). Unlike data-point-level DP, which protects individual samples, client-level LDP safeguards the whole dataset without scaling noise per data point, simplifying sensitivity control. However, this stronger privacy model introduces a conflict with clustering accuracy: under small $\epsilon$, the injected noise dominates pairwise distances, scaling with $\mathcal{O}(d/\epsilon^2)$ and destroying the geometric structure necessary for meaningful clustering. Consequently, existing private federated clustering methods maintain accuracy only under weak privacy budgets ($\epsilon > 1$), offering insufficient protection for sensitive applications. Our work, Gravitational Federated Clustering (GFC), is designed specifically to overcome this accuracy collapse by creating a robust representation of data topology that is resilient to the high levels of noise required by strong LDP. here amke this concise

## Gravitational Federated Clustering

We introduce Gravitational Federated Clustering (GFC), shown in Figure 1, a novel one-shot federated clustering framework designed for robustness against the noise introduced by differential privacy. Traditional federated clustering methods that rely on precise distance calculations or nearest-neighbour graphs are often destabilized by the perturbations required to protect data privacy. GFC overcomes this limitation by reframing the clustering problem through the lens of physics and topology. It models the decentralized, noisy client data as a system of point masses and identifies the true cluster centers by finding the most topologically significant features of the resulting gravitational potential field.

The algorithm operates in two main phases: (1) client-side processing, where local centroids are extracted under differential privacy, and (2) server-side global aggregation, where a continuous energy landscape is constructed and analyzed using tools from Topological Analysis to reveal the global cluster structure. See the details of the GFC implementation in Algorithm 1.

### Client-Side Processing

Each client in the federated system generates privacy-preserving local centroids without revealing its raw data. Let client $m$ possess a local dataset $\mathcal{D}_m = \{\mathbf{x}_i\}_{i=1}^{N_m}$, where $N_m$ is the size of client $m$'s dataset and each data point $\mathbf{x}_i \in \mathbb{R}^d$.

To protect privacy, each client applies $\epsilon$-LDP at the record level, as shown in Algorithm 1 (Line 3). Calibrated Laplace noise is added to each data point $\mathbf{x}_i \in \mathcal{D}_m$ to produce a noisy dataset $\tilde{\mathcal{D}}_m$. The noise is scaled by $\Delta$, the $L_1$-sensitivity of the data.

After privatizing the data, the client computes the "Local Centroids" by running k-means on the noisy dataset $\tilde{\mathcal{D}}_m$. The $i$-th client's $j$-th cluster is denoted $\mathcal{C}_{ij}$ with its centroid $\mathbf{c}_{ij}$. To characterize the cluster's density, we compute its mass:

$$w_{ij} = \exp\left(-\frac{\sum_{\mathbf{x}\in\mathcal{C}_{ij}} \|\mathbf{x} - \mathbf{c}_{ij}\|^2}{2\sigma_i^2}\right) \tag{1}$$

where $\sigma_i^2$ is the variance of $\|\mathbf{x}_\mu - \mathbf{x}_\nu\|^2$ for distinct $\mathbf{x}_\mu, \mathbf{x}_\nu \in \mathcal{D}_i$.

In a single communication round, each client transmits its set of local centroids and their corresponding masses, $(\mathbf{c}_{ij}, w_{ij})$, to the central server and take the union, denoted $\mathcal{S}$. After re-indexing, we obtain $\mathcal{S} = \{\mathbf{c}_\alpha, w_\alpha\}_\alpha$.

### Server-Side Global Aggregation

Upon receiving the weighted local centroids from all $m$ clients, the server performs a topological analysis to identify the final $N$ global cluster centers.

**Gravitational Potential Field Construction** The server first defines the bounding box of the received centroids. It then generates a synthetic dataset $G$ by uniformly sampling $\alpha k$ points from within these bounds. The potential energy field is then evaluated at each of these synthetic points $\mathbf{g}_j \in G$:

$$E(\mathbf{g}_j) = \sum_{(\mathbf{c}_\alpha, w_\alpha)\in\mathcal{S}} \frac{w_\alpha}{\|\mathbf{c}_\alpha - \mathbf{g}_j\|^p + \delta}, \tag{2}$$

where $\delta$ is a small constant to avoid division by zero and $p$ is a hyperparameter (typically $p = 2$ for a Coulomb-like

potential) that controls the field's decay. Note: For high-dimensional data, this field is typically constructed in a low-dimensional embedding (e.g., via UMAP) to mitigate computational cost.

**Gravitational Potential Field Construction** The server constructs a potential energy field over the data space $\mathbb{R}^d$. The potential energy $E(y)$ at any point $y \in \mathbb{R}^d$ is defined by summing the contributions of each point mass:

$$E(y) = \sum_{(c_\alpha, w_\alpha) \in \mathcal{S}} \frac{w_\alpha}{\|c_\alpha - y\|^2 + \delta}, \tag{3}$$

where $\delta$ is a small constant to avoid division by zero

In this formulation, dense regions of client centroids (the point masses) create high-potential peaks in the energy landscape, with more massive (more compact) centroids contributing more significantly. These peaks are hypothesized to correspond to the true global cluster centers.

**Topological Feature Extraction** To robustly identify cluster centers in the presence of differential privacy noise, we employ topological persistence analysis on the gravitational potential field. The method constructs a filtration through super-level sets $\mathcal{F}_h = \{\mathbf{g} \in G : E(\mathbf{g}) \geq h\}$, where $G$ represents the synthetic grid points and $h$ varies across energy thresholds. The function $\pi_0(\mathcal{F}_h)$, known as the *zeroth homotopy group*, counts the number of *path-connected components* in each super-level set. These components directly correspond to the distinct cluster peaks at energy level $h$.

The algorithm tracks these 0D-topological features (the components) as $h$ decreases in a process known as persistent homology. This allows us to distinguish stable, persistent peaks (true clusters) from minor, noisy peaks that merge quickly. The threshold sequence $H$ is generated by $\Phi(E(G), \alpha \cdot |\mathcal{S}|)$, scaling with data density through $\alpha \cdot |\mathcal{S}|$ where $|\mathcal{S}|$ is the total number of uploaded centroids. For each threshold $h_i \in H$, connected components are computed using the function $\text{CC}(F_{h_i})$, which constructs an adjacency matrix based on neighborhood radius $r$ and identifies connected components through graph traversal. For each persistent component $L$ identified in the merge tree, we compute an energy-weighted centroid $\mu_L = \frac{\sum_{\mathbf{g} \in L} E(\mathbf{g})\mathbf{g}}{\sum_{\mathbf{g} \in L} E(\mathbf{g})}$, naturally emphasizing regions of high potential energy that align with true cluster centers. The leaves of the resulting merge tree represent the most persistent peaks, which we identify as the final global cluster centers. We employ a merge tree for efficient $O(m \log m)$ persistence computation, where $m$ is the number of synthetic points, avoiding costly recomputation at each threshold. If fewer than $n_c$ persistent peaks are found, a fallback mechanism hierarchically selects high-energy components followed by maximum-energy synthetic points to guarantee $n_c$ clusters. (See details in Appendix).

Algorithm 1: Gravitational Federated Clustering (GFC)

**Input**: Distributed datasets $D = \{D_1, \ldots, D_C\}$ across $C$ clients, Privacy budget $\epsilon$, Local clusters per client $k$, Target clusters $n_c$, Softening factor $\delta$, Synthetic multiplier $\alpha$, Local centroids' mass variance $\sigma$, Neighborhood radius $r$.
**Output**: Global centroids $F$

**Client Phase:**
**for** each client $c \in \{1, \ldots, C\}$ **do**
  $\tilde{D}_c \leftarrow D_c + \text{Lap}(0, \Delta/\epsilon)$ {LDP protection}
  $\{C_{c1}, \ldots, C_{ck}\} \leftarrow \text{KMeans}(\tilde{D}_c, k)$
  **for** each centroid $C_{ci}$ **do**
    $w_{ci} \leftarrow \exp(-\frac{\sum_{x \in C_{ci}} \|x - C_{ci}\|^2}{2\sigma^2})$ {Mass calculation}
  **end for**
  Upload $\{(C_{ci}, w_{ci})\}_{i=1}^k$ to server
**end for**
**Server Phase:**
$\mathcal{S} \leftarrow \bigcup_c \{(C_{ci}, w_{ci})\}_{i=1}^k$ {Aggregate centroids}
Generate synthetic data $G$ with $\alpha \cdot |\mathcal{S}|$ data points within data bounds $\mathcal{B} = [\min(\mathcal{S}), \max(\mathcal{S})]$
**for** each $g_j \in G$ **do**
  $E(g_j) \leftarrow \sum_{(C_i, w_i) \in \mathcal{S}} \frac{w_i}{\|g_j - C_i\|^2 + \delta}$ {Potential field}
**end for**
$H \leftarrow \Phi(E(G), \alpha \cdot |\mathcal{S}|)$ {Threshold sequence}
$\mathcal{T}$.init() {Initialize empty topological tree}
**for** each $h_i \in H$ **do**
  $F_{h_i} \leftarrow \{g_j | E(g_j) \leq h_i\}$ {Sub-level set at threshold $h_i$}
  $\mathcal{C}_{\text{new}} \leftarrow \text{CC}(F_{h_i})$ {Connected components}
  **if** $\mathcal{C}_{\text{new}} \neq \mathcal{T}$.leaves() **then**
    $\mathcal{T}$.add_nodes($\mathcal{C}_{\text{new}}$) {Update tree structure}
    **for** each new leaf $L \in \mathcal{T}$.new_leaves() **do**
      $\mu_L \leftarrow \frac{\sum_{g_j \in L} E(g_j) g_j}{\sum_{g_j \in L} E(g_j)}$ {Energy-weighted centroid}
      $\mathcal{T}$.store($L, \mu_L$) {Store cluster candidate}
    **end for**
  **end if**
  **if** $|\mathcal{T}.\text{leaves}()| \geq n_c$ **then**
    BREAK {Stop when sufficient clusters found}
  **end if**
**end for**
**if** $|\mathcal{T}.\text{leaves}()| < n_c$ **then**
  $F \leftarrow \mathcal{T}$.isolated_paths() {Get isolated components}
**end if**
**if** $|F| < n_c$ **then**
  $F \leftarrow F \cup \mathcal{T}$.top_energy_leaves($n_c - |F|$) {Add high-energy leaves}
**end if**
**if** $|F| < n_c$ **then**
  Add top $n_c - |F|$ candidates by $E(g_j)$ {Final fallback}
**end if**
**return** $F$

## Experimental Settings

**Non-IID Datasets Settings.** We construct non-IID federated scenarios by partitioning datasets into *num_clients* subsets with heterogeneous label distributions. Our partitioning strategy applies $k$-means clustering followed by a system-

| $\varepsilon$ | Metric | Method | Small to Medium-sized Datasets | | | | | | | Large Datasets | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Seeds | Thyroid | Breast | Heart | Gesture | Abalone | Waveform | Celltype | MNIST | Postures |
| 0.1 | ARI | K-Fed | 22.65±20.37 | **43.39±18.66** | 1.89±3.69 | NA | 0.24±0.81 | 11.16±2.61 | 30.40±6.69 | 17.32±2.24 | 39.56±4.19 | 2.54±0.30 |
| | | MUFC | 27.20±16.59 | 42.85±17.24 | 1.75±4.13 | 2.04±7.53 | 10.07±8.59 | 8.01±5.27 | **33.80±7.80** | 17.27±2.27 | 34.13±3.02 | 2.86±0.15 |
| | | NNFC | 1.36±5.67 | 25.63±15.89 | 0.67±2.48 | 1.18±4.10 | 3.92±8.33 | NA | 5.14±7.25 | 8.98±7.66 | 8.23±8.30 | 0.95±1.06 |
| | | GFC(Ours) | **41.07±5.93** | 43.38±9.93 | **2.19±3.92** | **20.04±10.54** | **13.49±11.06** | **12.95±4.09** | 23.8±6.31 | **17.91±3.44** | **56.53±5.95** | **2.98±0.42** |
| | | Improv. | 50.99% ↑ | 0.02% ↓ | 15.87% ↑ | 882.35% ↑ | 33.96% ↑ | 16.04% ↑ | 29.58% ↓ | 3.41% ↑ | 42.89% ↑ | 4.20% ↑ |
| | NMI | K-Fed | 28.84±22.16 | 39.40±12.00 | 1.61±2.42 | 0.02±0.05 | 0.78±2.06 | 11.88±1.82 | 35.69±5.46 | 16.09±0.93 | 63.86±3.76 | 3.14±0.24 |
| | | MUFC | 36.97±15.17 | 38.64±10.89 | 1.12±2.65 | 2.54±6.20 | 14.06±6.28 | 9.32±4.34 | **40.21±4.41** | 15.91±1.10 | 58.26±2.56 | 3.12±0.34 |
| | | NNFC | 2.48±7.17 | 25.16±13.94 | 0.58±1.09 | 1.19±3.40 | 4.46±7.79 | NA | 6.28±7.53 | 9.02±7.26 | 20.39±18.80 | 1.38±1.06 |
| | | GFC(Ours) | **46.09±8.40** | **40.40±10.78** | **1.67±3.09** | **14.79±9.51** | **17.47±6.14** | **13.38±2.78** | 25.58±6.13 | **16.11±1.30** | **72.25±3.91** | **3.33±0.48** |
| | | Improv. | 24.67% ↑ | 2.54% ↑ | 3.73% ↑ | 482.28% ↑ | 24.25% ↑ | 12.63% ↑ | 36.38% ↓ | 1.24% ↑ | 13.14% ↑ | 6.05% ↑ |
| 0.05 | ARI | K-Fed | 8.74±16.04 | 25.28±12.44 | 0.99±2.24 | 0.04±0.16 | NA | 8.04±4.70 | 26.64±7.62 | 16.05±3.33 | 26.68±5.32 | 2.30±0.60 |
| | | MUFC | 18.05±20.36 | 26.62±16.22 | 0.27±0.96 | 0.66±2.64 | 3.37±4.73 | 7.20±3.63 | **28.94±5.60** | 14.64±3.15 | 26.78±5.80 | 2.85±0.22 |
| | | NNFC | 0.69±2.59 | 23.10±11.95 | 0.10±0.45 | NA | 1.63±4.19 | NA | NA | 0.87±3.65 | 2.36±4.91 | NA |
| | | GFC(Ours) | **36.96±8.13** | **29.24±12.64** | **1.79±4.22** | **12.35±11.49** | **7.66±9.53** | **10.88±5.51** | 21.9±7.24 | **16.91±3.57** | **56.23±5.99** | **2.99±0.46** |
| | | Improv. | 104.76% ↑ | 9.84% ↑ | 80.81% ↑ | 1771.21% ↑ | 127.30% ↑ | 35.32% ↑ | 24.33% ↓ | 5.36% ↑ | 109.97% ↑ | 4.91% ↑ |
| | NMI | K-Fed | 12.48±19.17 | 28.91±8.80 | 0.75±1.41 | 0.03±0.13 | 0.34±1.11 | 9.88±3.61 | 29.73±6.06 | 15.10±1.63 | 53.20±3.32 | 2.84±0.48 |
| | | MUFC | 22.45±23.47 | 28.92±10.03 | 0.58±1.63 | 0.82±2.39 | 5.80±5.84 | 9.67±3.39 | **32.24±4.74** | 14.45±1.65 | 53.51±4.37 | 3.09±0.32 |
| | | NNFC | 2.18±6.39 | 23.17±9.53 | 0.14±0.60 | NA | 2.02±4.27 | NA | 0.02±0.03 | 1.16±3.40 | 6.23±12.56 | NA |
| | | GFC(Ours) | **42.22±8.60** | **31.75±9.40** | **0.82±2.12** | **11.21±10.15** | **11.78±7.28** | **11.45±4.05** | 19.86±5.98 | **15.78±2.03** | **72.15±4.18** | **3.22±0.54** |
| | | Improv. | 88.06% ↑ | 9.79% ↑ | 9.33% ↑ | 1511.12% ↑ | 103.10% ↑ | 15.89% ↑ | 38.40% ↓ | 4.50% ↑ | 34.83% ↑ | 4.20% ↑ |
| 0.01 | ARI | K-Fed | 2.86±10.87 | 18.55±13.95 | NA | NA | 0.12±0.53 | 2.92±5.05 | 2.25±5.58 | 13.61±3.60 | 19.21±8.81 | 1.97±0.82 |
| | | MUFC | 0.89±3.48 | 14.34±12.23 | 0.03±0.13 | NA | 1.01±3.49 | 1.69±3.81 | 3.12±6.22 | 14.38±3.75 | 19.34±8.57 | **2.85±0.54** |
| | | NNFC | NA | 4.97±9.06 | NA | NA | NA | NA | NA | 6.3±0.7 | NA | NA |
| | | GFC(Ours) | **5.10±14.16** | **18.92±13.84** | **0.27±0.95** | **0.22±0.95** | **8.82±10.28** | **7.82±6.48** | **9.64±9.23** | **14.54±4.23** | **52.79±4.42** | **2.79±0.50** |
| | | Improv. | 78.32% ↑ | 1.99% ↑ | 800.00% ↑ | $\infty$ ↑ | 773.27% ↑ | 167.81% ↑ | 208.97% ↑ | 1.11% ↑ | 162.62% ↑ | 2.10% ↓ |
| | NMI | K-Fed | 4.03±12.95 | 18.51±13.98 | NA | 0.04±0.17 | 0.21±0.92 | 4.51±6.05 | 3.50±5.84 | 14.57±1.60 | 42.13±12.28 | 2.58±0.69 |
| | | MUFC | 2.23±0.07 | 13.84±10.63 | 0.01±0.05 | NA | 1.80±4.57 | 2.79±4.09 | 4.59±6.37 | 14.53±1.74 | 42.34±12.08 | 3.17±0.41 |
| | | NNFC | NA | 4.93±8.84 | NA | NA | 0.31±1.36 | NA | NA | NA | NA | NA |
| | | GFC(Ours) | **5.80±16.32** | **19.14±13.38** | **0.26±0.95** | **0.63±0.76** | **9.69±6.33** | **9.57±5.52** | **11.68±8.17** | **15.27±3.43** | **69.63±2.90** | **3.20±0.46** |
| | | Improv. | 43.92% ↑ | 3.40% ↑ | 2500.00% ↑ | 1475.00% ↑ | 438.33% ↑ | 112.20% ↑ | 154.46% ↑ | 4.80% ↑ | 64.5% ↑ | 0.95% ↑ |

Table 1: Performance comparison of GFC, NN-FC, K-Fed, and MUFC using Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Results, shown as mean $\pm$ standard deviation over 10 seeds, are evaluated under different privacy budgets $\varepsilon$, where a smaller $\varepsilon$ (e.g., 0.01) indicates stronger privacy. Higher values denote better performance; best results are highlighted in **blue**.

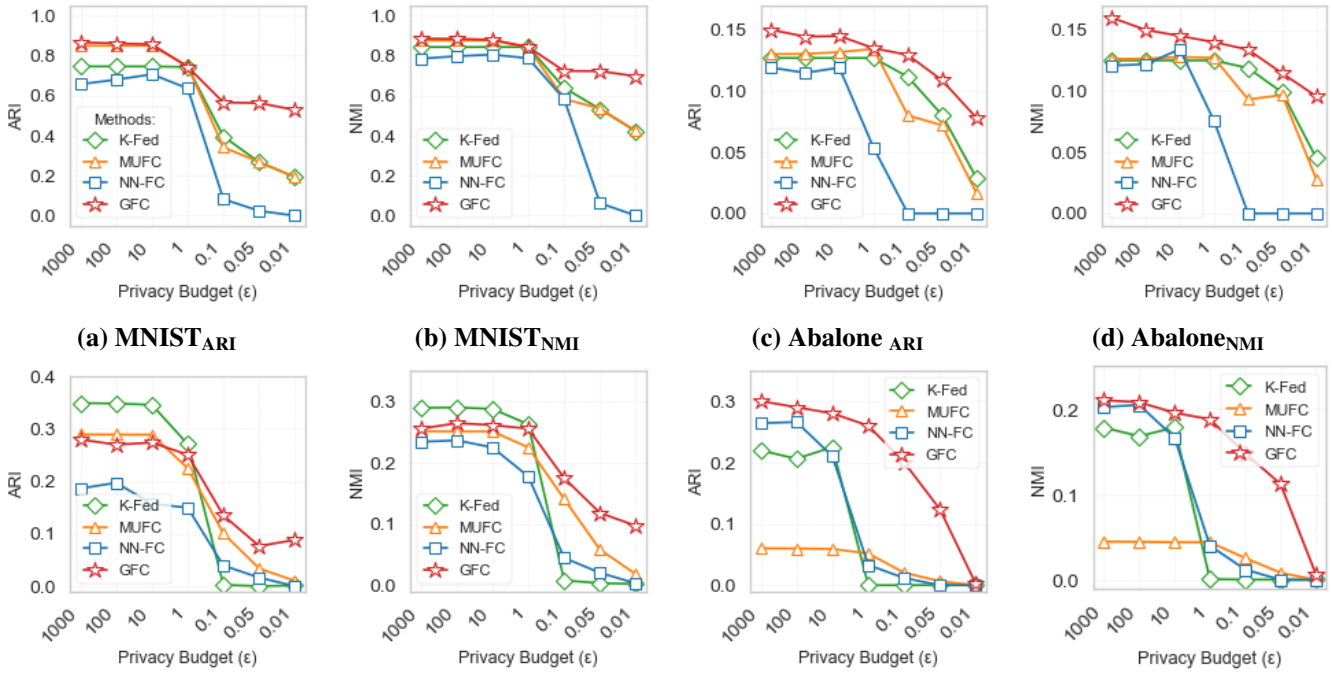
| Dataset | Samples | Dimensions | Clusters |
|---|---|---|---|
| Postures | 74,975 | 15 | 5 |
| MNIST | 70,000 | 784 | 10 |
| Celltype | 12,009 | 10 | 4 |
| Waveform | 5,000 | 21 | 3 |
| Abalone | 4,177 | 8 | 3 |
| Gesture | 1,747 | 18 | 5 |
| Heart | 303 | 13 | 2 |
| Breast | 277 | 9 | 2 |
| Thyroid | 215 | 5 | 3 |
| Seeds | 210 | 7 | 3 |

Table 2: Diversity of real-world datasets used for benchmarking, ordered by the number of samples.

atic allocation process that ensures diverse label distributions across clients while maintaining balanced cluster representation. The detailed partitioning algorithm is provided in the Appendix.

**Baselines.** We compare against three state-of-the-art one-shot federated clustering methods: k-FED (Dennis, Li, and Smith 2021b) for non-IID data handling, MUFC (Pan et al. 2022) for privacy through unlearning mechanisms, and NN-FC (Wang, Pang, and Pedrycz 2024) as the current SOTA using differential privacy and nearest-neighbor relationships. These cover key dimensions of accuracy, privacy, and heterogeneity adaptation.

**Experimental Setup and Evaluation.** We evaluate our GFC method against baselines using two robust clustering metrics: Adjusted Rand Index (ARI) (Steinley 2004) for cluster similarity and Normalized Mutual Information (NMI) (McDaid, Greene, and Hurley 2011) for label agreement, accounting for label permutations and cluster size imbalances. All methods are tested under identical conditions across benchmark datasets, with 20 random seeds to assess consistency (reported as mean $\pm$ std). To analyze privacy-accuracy trade-offs, we fix $\Delta$ while varying $\epsilon \in [1000, 100, 10, 1, 0.1, 0.05, 0.01]$, measuring performance degradation as privacy strengthens. Beyond assessing clustering performance (ARI/NMI) under varying privacy




**(a) MNIST$_{\text{ARI}}$** **(b) MNIST$_{\text{NMI}}$** **(c) Abalone $_{\text{ARI}}$** **(d) Abalone$_{\text{NMI}}$**

**(e) Gestures$_{\text{ARI}}$** **(f) Gestures $_{\text{NMI}}$** **(g) Heart $_{\text{ARI}}$** **(h) Heart$_{\text{NMI}}$**

Figure 3: Performance comparison of GFC, NN-FC, K-Fed, and MUFC across varying privacy budgets $\varepsilon$ (smaller $\varepsilon$ denotes stronger privacy). Mean ARI scores are shown in (a, c, e, g), while corresponding mean NMI scores are in (b, d, f, h). GFC results are highlighted in red.

budgets, we conduct multi-dimensional analysis: (1) sensitivity to initial centroid count $k_0$ across dataset scales, (2) computational efficiency across diverse datasets, (3) ablation studies on gravitational parameters ($\delta \in [0.0001, 0.1, 100]$, synthetic data multiplier $\alpha \in [1, 2, 5, 10]$) to quantify their impact on topological analysis, and (4) scalability tests with client numbers $C$ ranging from 10 to 100.

**Hyperparameters Finetuning.** To ensure a fair comparison, all baseline methods utilize their originally published hyperparameters. For our method (GFC), we establish a systematic heuristic for hyperparameter selection grounded in the dataset's intrinsic properties and the allocated privacy budget $\epsilon$, which avoids extensive manual tuning. The number of local centroids $k$ scales with the dataset size $n$ as $k(n) = 15 + n/500$. The gravitational smoothing parameter is set to $\delta(\epsilon) = 500 \cdot e^{-5\epsilon}$, and the synthetic data multiplier follows $\alpha(\epsilon) = 2 + 20/(\epsilon + 1)$, ensuring both adapt to the privacy-utility trade-off. Finally, the neighborhood radius $r$ is dynamically determined as the 1$^{\text{st}}$ percentile of non-zero distances between the received centroids, capturing the data's characteristic local scale. This principled approach, informed by persistent homology stability, allows for robust and reproducible configuration.

## Experimental Results

**Performance under Strong LDP Constraints.** The results in Table 1 demonstrate that GFC consistently outperforms all state-of-the-art (SOTA) methods in accuracy, as measured by ARI and NMI, across datasets of varying sizes under privacy budgets ($\varepsilon$) ranging from 0.1 to 0.01. For small datasets (e.g., Breast, Heart, and Gesture), GFC achieves significantly better performance under low privacy budgets, whereas other methods either fail (producing NA results) or yield near-zero accuracy. In medium and large datasets (e.g., Abalone and MNIST), GFC delivers substantial improvements: up to 167.81% (Abalone) and 162.62% (MNIST) in ARI under $\varepsilon = 0.01$ compared to baseline methods. Furthermore, while existing SOTA methods exhibit high variance (reflecting instability under strong DP constraints), GFC maintains lower variance across most datasets, even as noise injection increases with stricter privacy budgets. On the high-dimensional Waveform dataset, GFC is competitive at moderate privacy and becomes the top performer at $\epsilon = 0.01$, where other methods fail. These results confirm GFC's accuracy and robustness in the federated clustering.

**Performance under Varying Local DP.** To evaluate the generalization of GFC under different privacy constraints, we test its clustering performance across a wide range of privacy budgets ($\varepsilon$), from 1000 (nearly non-private) to 0.01 (highly private), using both the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) metrics. As illustrated in Figure 3, GFC achieves comparable or even superior performance to baselines under high privacy budgets ($\varepsilon \geq 1$). As noise levels increase with stricter privacy constraints ($\varepsilon < 1$), GFC maintains relatively high accuracy with only a gradual decline, while competing methods suf-

fer sudden performance drops. This advantage is particularly evident on large-scale datasets (e.g., MNIST) and small-to-medium datasets (e.g., Abalone and Heart), where GFC consistently outperforms state-of-the-art approaches across all privacy levels. These results highlight GFC's robustness under varying privacy constraints, making it a reliable choice for federated clustering. See further results in Appendix.

## Ablation Study

**Sensitivity to Local Centroids Number** $n$ We observed that under the privacy budget 0.1, GFC maintains stable performance across a wide range of local centroid counts (n), showing consistent accuracy on diverse datasets. Detailed results are listed in the appendix.

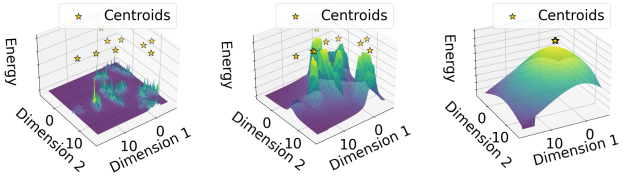
**FC Efficiency** We compare the efficiency-accuracy trade-off between *k*-Fed, MUFC, NN-FC, and our GFC framework (detailed in the appendix). While GFC exhibits comparable runtime performance to state-of-the-art (SOTA) methods—closely matching their speed on both small and large datasets—it achieves significantly higher clustering accuracy. Specifically, GFC maintains near-identical computational efficiency to k-fed, which is usually the fastest method (within 5% time difference) but improves the average ARI by 167.81% in the abalone dataset. This demonstrates that our energy-based formulation introduces comparable communication time cost while substantially enhancing federated clustering quality. Detailed runtime-accuracy curves are provided in the Appendix.

**Number of Synthetic Data Points** The number of synthetic data points ($n$) plays a crucial role in determining the energy field's approximation quality, with its impact varying significantly across privacy budgets. Our evaluation of $n \in \{1\times, 2\times, 10\times, 100\times\}$ under $\epsilon \in \{1000, 0.01\}$ (measured via ARI/NMI metrics in Table 3) reveals distinct patterns: For $\epsilon = 1000$, small $n$ ($\leq 1\times$) causes undersampling and unstable cluster boundaries, while moderate $n$ ($\approx 2\times$ to $5\times$) achieves optimal balance between computational cost and field fidelity, and large $n$ ($10\times$) shows diminishing returns with increased communication overhead. However, under strict privacy ($\epsilon = 0.01$), larger $n$ ($10\times$) becomes essential as high-energy centroids aggregate more closely, requiring denser sampling to maintain clustering accuracy - demonstrating how the optimal $n$ depends on both the dataset characteristics and privacy requirements. More Results are shown in the Appendix.

**Gravitational Field Construction** We analyze the effect of the smoothing parameter $\delta$ (tested at $1 \times 10^{-4}$, $1 \times 10^{-1}$, $1 \times 10^{2}$) on cluster formation and noise robustness in our potential field model. And $\delta$ controls the field's smoothness - smaller values create sharper energy field wells around synthetic points $\mathbf{s}_i$, while larger values produce smoother energy landscapes. Figure 4 demonstrates this on MNIST clustering under low privacy regimes ($\epsilon = 1000$), showing $\delta$'s impact on cluster separation and noise tolerance. Delta can be used to control the smoothness of the energy field under varying private budgets. Extended results appear in the appendix.

| $\epsilon$ | Metric | Number of Synthetic Data ($n$) | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 5 | 10 |
| 1000 | ARI | 0.82 | 0.88 | 0.86 | 0.37 |
| | NMI | 0.88 | 0.89 | 0.88 | 0.56 |
| 0.01 | ARI | 0.10 | 0.11 | 0.15 | 0.53 |
| | NMI | 0.34 | 0.35 | 0.34 | 0.70 |

Table 3: Impact of Synthetic Data ($n$) and Privacy Budget ($\epsilon$) on Clustering Quality. Strong DP ($\epsilon = 0.01$) requires a higher $n$ to compensate for noisier gradients.




(a) delta=$1 \times 10^{-4}$ (b) delta=$1 \times 10^{-1}$ (c) delta=$1 \times 10^{2}$

Figure 4: Impact of $\delta$ for GFC on MNIST Data

**Scalability to Large Client Populations** Our framework demonstrates strong scalability, maintaining stable and high clustering performance when extended to 100 or 1000 clients (see Appendix for full results). This scalability highlights our framework's practical applicability for real-world distributed systems with numerous participants.

## Discussion and Limitations

Despite its strong performance, our framework has several limitations that point to future research directions. First, while our proposed heuristic methods automate much of the hyperparameter tuning, the model's performance can still be sensitive to these settings, particularly with high-dimensional datasets exhibiting highly non-IID client data distributions. Second, regarding scalability, although the algorithm is designed for a large number of clients, its communication overhead scales linearly with the client count. This becomes a significant bottleneck at an ultra-large scale (e.g., ten thousands of clients), as the server must aggregate information from every participant in each round.

## Conclusions

We presented Gravitational Federated Clustering (GFC), a novel one-shot approach that reformulates private federated clustering as a topological persistence problem in a synthetic potential field. By encoding local cluster geometries as gravitational masses and extracting centroids through persistent homology analysis, GFC achieves: (1) (1) better robustness across wide-range privacy budgets ($\epsilon = 0.01$ to 1000) while superior accuracy on ARI and NMI under strong LDP ($\epsilon < 1$); (2) provable noise suppression via the potential field's Lipschitz smoothing properties; and (3) elimination of iterative communication by proposing an one-shot method.

## Acknowledgments

The computations in this research were performed using the CFFF platform of Fudan University.

## References

Chen, L.; Zhao, J.; Fan, J.-W.; Chen, H.; Zhao, Z.; Wang, G.; and Wang, C. 2024. One-Shot Secure Federated K-Means Clustering Based on Density Cores. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Demelius, L.; Kern, R.; and Trügler, A. 2025. Recent advances of differential privacy in centralized deep learning: A systematic survey. *ACM Computing Surveys*, 57(6): 1–28.

Dennis, D. K.; Li, T.; and Smith, V. 2021a. Heterogeneity for the Win: One-Shot Federated Clustering. In *International Conference on Machine Learning*, 2611–2620. PMLR.

Dennis, D. K.; Li, T.; and Smith, V. 2021b. Heterogeneity for the win: One-shot federated clustering. In *International conference on machine learning*, 2611–2620. PMLR.

Garst, S.; and Reinders, M. 2024. Federated K-Means Clustering. *arXiv preprint arXiv:2310.01195*.

Ghosh, A.; Chung, J.; Yin, D.; and Ramchandran, K. 2021. An Efficient Framework for Clustered Federated Learning. *arXiv preprint arXiv:2006.04088*.

Gupta, S.; Tarushi; Wangzes, T.; and Jain, S. 2024. Fair Federated Data Clustering through Personalization: Bridging the Gap between Diverse Data Distributions. *arXiv preprint arXiv:2407.04302*.

Long, Y.; Xu, L.; Zheng, G.; and Brintrup, A. 2025. PA-CFL: Privacy-Adaptive Clustered Federated Learning for Transformer-Based Sales Forecasting on Heterogeneous Retail Data. arXiv:2503.12220.

McDaid, A. F.; Greene, D.; and Hurley, N. 2011. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*.

McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Pan, C.; Sima, J.; Prakash, S.; Rana, V.; and Milenkovic, O. 2022. Machine unlearning of federated clusters. *arXiv preprint arXiv:2210.16424*.

Pan, C.; Sima, J.; Prakash, S.; Rana, V.; and Milenkovic, O. 2023. Machine Unlearning of Federated Clusters. In *International Conference on Learning Representations*.

Sattler, F.; Müller, K.-R.; and Samek, W. 2020. Clustered Federated Learning: Model-Agnostic Distributed Multi-Task Optimization under Privacy Constraints. *IEEE Transactions on Neural Networks and Learning Systems*.

Scott, J.; Lampert, C. H.; and Saulpic, D. 2025. Differentially Private Federated $k$-Means Clustering with Server-Side Data. *arXiv preprint arXiv:2506.05408*.

Steinley, D. 2004. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3): 386.

Wang, Y.; Pang, W.; and Pedrycz, W. 2024. One-Shot Federated Clustering Based on Stable Distance Relationships. *IEEE Transactions on Industrial Informatics*.

Xia, C.; Hua, J.; Tong, W.; and Zhong, S. 2020. Distributed K-Means clustering guaranteeing local differential privacy. *Computers & Security*, 90: 101699.

Zhang, Y.; Wang, X.; Shen, S.; Wang, Y.; and Cheng, L. 2024. Enhancing Group Fairness in Federated Learning through Personalization. *arXiv preprint arXiv:2407.19331*.