CASE WESTERN RESERVE UNIVERSITY
Case School of Engineering
Computer and Data Sciences

Privacy-Preserving Optimal Parameter Selection for Collaborative Clustering

Authors: Maryam Ghasemian, Erman Ayday
Presenter on be half of the authors: Alexander Nemecek

International Workshop on Data Privacy Management (DPM 2024) , September 19, 2024

INTRODUCTION

❑ Overview of Clustering:

- *Definition:* Clustering is an unsupervised machine learning technique used to group similar data points based on certain features.
- *Importance:* Fundamental for pattern recognition, data analysis, and segmentation.

❑ Collaborative Clustering:

- *Definition:* Multiple data owners collaborate by sharing data to improve clustering outcomes.
- *Challenge:* Ensuring privacy while achieving effective clustering.

❑ Purpose of the Study:

- To explore how to select optimal clustering parameters in a privacy-preserving manner.

PROBLEM STATEMENT

❑ Objective:

- Develop a method for selecting optimal clustering parameters in a privacy-preserving manner.

❑ Key Challenge:

- **Existing Approaches:**
  - Many existing works rely on pre-defined clustering algorithms and a fixed number of clusters.
  - These methods often apply encryption techniques to protect data privacy.
- **Limitations:**
  - Pre-selecting the number of clusters and the algorithm may not be suitable for all datasets, leading to suboptimal clustering results.

❑ Our Contribution:

- We focus on identifying the optimal clustering algorithm and the corresponding hyperparameters within a privacy-preserving framework.
- Our approach addresses the gap by allowing flexibility in the choice of clustering parameters, tailored to the specific data being analyzed, while still ensuring robust privacy protection.

CLUSTERING ALGOLITHMS

- **K-Means (Partitioning-based)**:
  - Divides data into K non-overlapping clusters.
  - Minimizes the sum of distances between data points and their respective cluster centroids.
- **Hierarchical Clustering (HC)**:
  - Builds a hierarchy of clusters.
  - Can be agglomerative (bottom-up) or divisive (top-down).
  - Useful for data with a hierarchical structure.
- **Gaussian Mixture Models (GMM, Distribution-based)**:
  - Assumes data is generated from a mixture of Gaussian distributions.
  - Flexible with complex cluster structures.
- **DBSCAN (Density-based)**:
  - Identifies clusters based on the density of data points.
  - Effective in finding arbitrarily shaped clusters.
  - Marks points in low-density regions as outliers.

DIFFERENTIAL PRIVACY

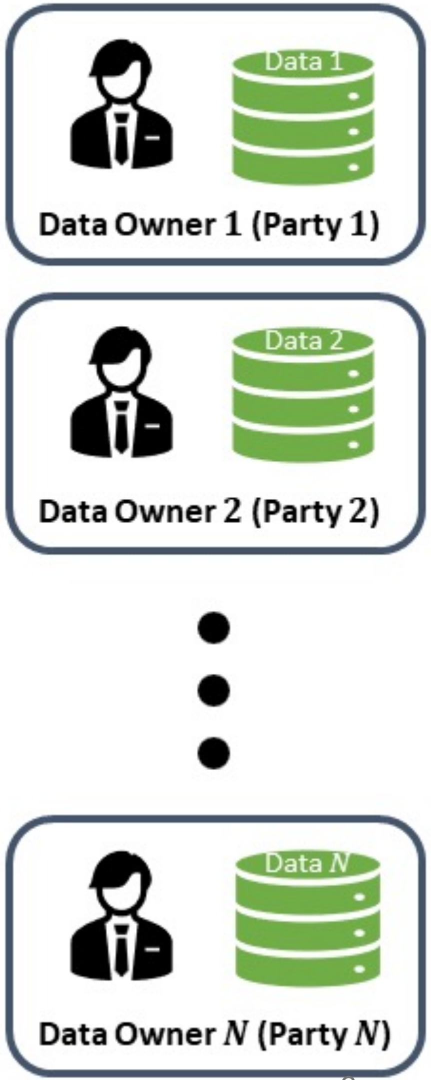
- **Overview**:
  - Differential Privacy (DP) is a framework to ensure that the output of a computation does not compromise the privacy of individuals in the dataset.

- **Local Differential Privacy (LDP)**:
  - A stronger form of DP where each data owner perturbs their data before sharing it.
  - Ensures that even if the perturbed data is intercepted, it cannot easily reveal the original data.

- **Randomized Response Mechanism:**
  - *Explanation*: Introduces noise into data in a controlled manner, providing plausible deniability.
  - *Purpose*: To protect individual privacy while allowing aggregate data analysis.

SYSTEM MODEL

- **Roles**:
  - **Data Owners (Researchers)**: Collaborate in clustering while maintaining data privacy.
  - **Semi-Trusted Server**: Acts as a third-party intermediary to assist in identifying the optimal clustering algorithm and hyper-parameters.
- **Focus**:
  - **Preliminary Stages**: The approach is applied before the actual clustering to determine optimal conditions.
- **Objective**:
  - Identify the best clustering algorithm and input parameters for collaborative clustering among multiple data owners.
- **Data Sharing**:
  - Data owners share selectively perturbed, *differentially private* data with the server.
  - The server analyzes the noisy data and recommends the most suitable clustering algorithm and corresponding hyper-parameters.



THREAT MODEL

- **Server**:
  - **Semi-Honest Behavior**: The server might attempt to infer sensitive information but follows the protocol.
  - **Risks**: Potential privacy violations, including:
    - ***Membership Inference****:* Inferring whether a specific record is part of a dataset.
    - ***Deanonymization****:* Linking anonymized data to real identities.
    - ***Attribute Inference****:* Deducing sensitive attributes from data.
  - **Mitigation**: Only a small, differentially-private portion of the data is shared, significantly reducing the risk of these attacks.
- **Data Owners**:
  - **Honest-but-Curious**: Data owners are cooperative but may be interested in learning about each other's data.
- **Assumption**: This is a cooperative environment, focusing on algorithm and parameter selection, while other literature handles more adversarial scenarios.

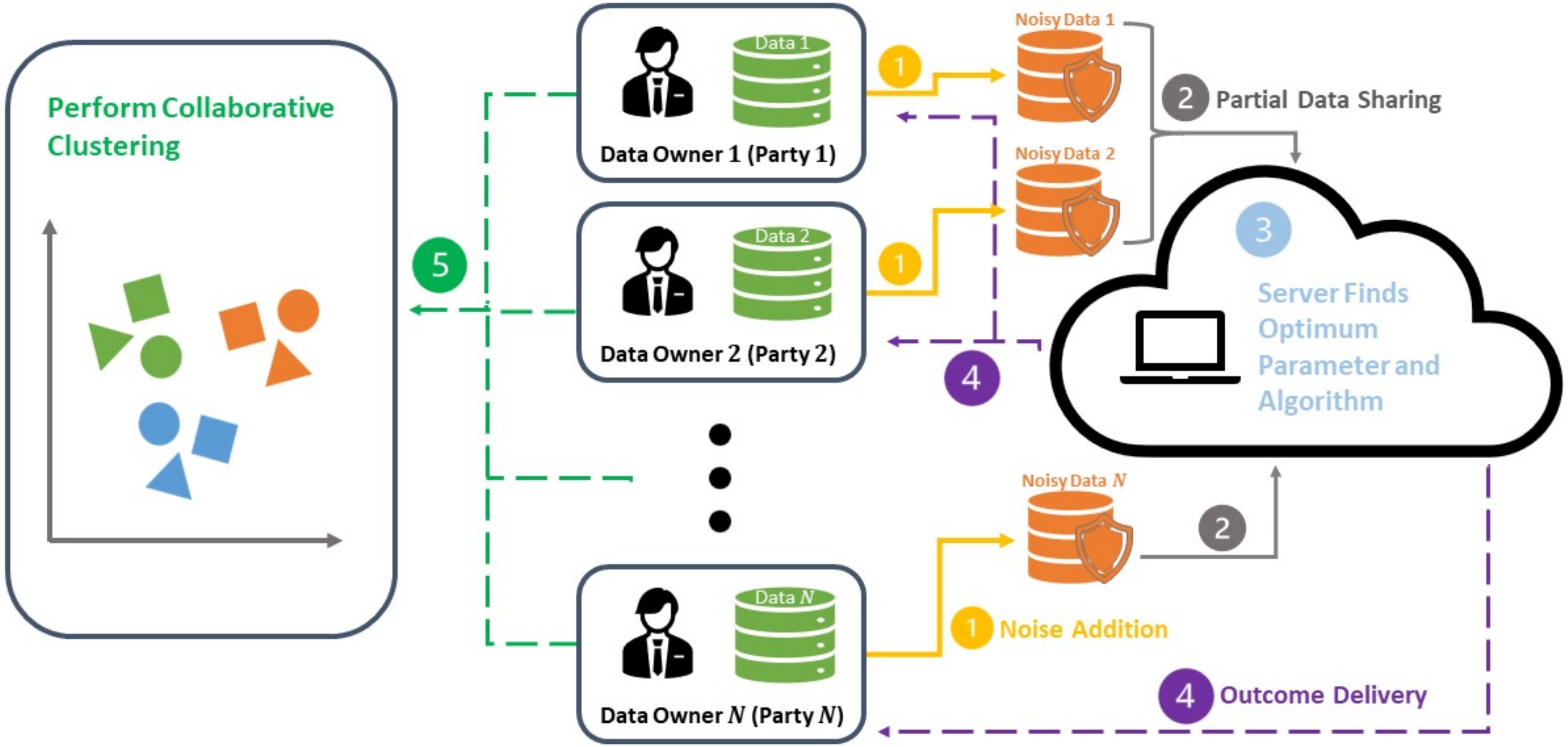
PROPOSED SOLUTION

1 Data owners add noise to their data using Randomized Response.

2 They share a portion of this differentially private data with the server.

3 The server analyzes the data and recommends the best algorithm and parameters.

4 Data owners receive these recommendations and 5 proceed with clustering.



EVALUATION SETUP

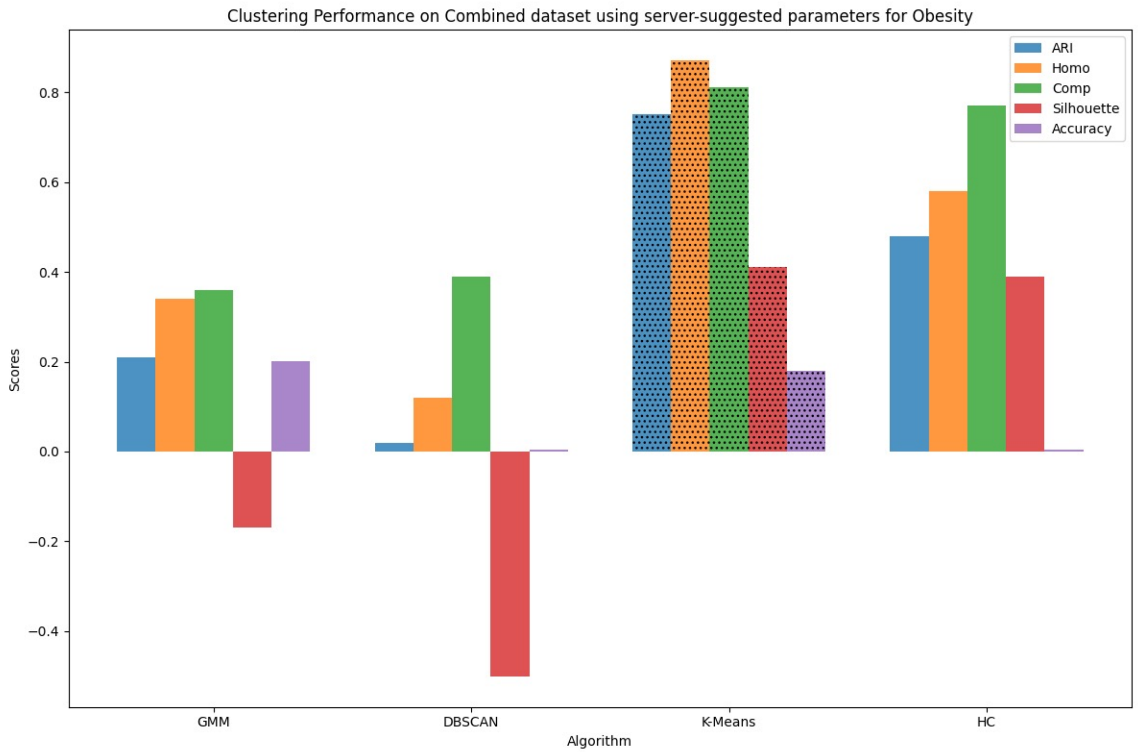
- **Datasets Used**:
  - *Obesity Dataset*: 2,111 records with 17 features, focusing on diet and physical condition.
  - *Extended Iris Dataset*: 1,200 rows with 20 features, providing detailed biological information about the iris flower.

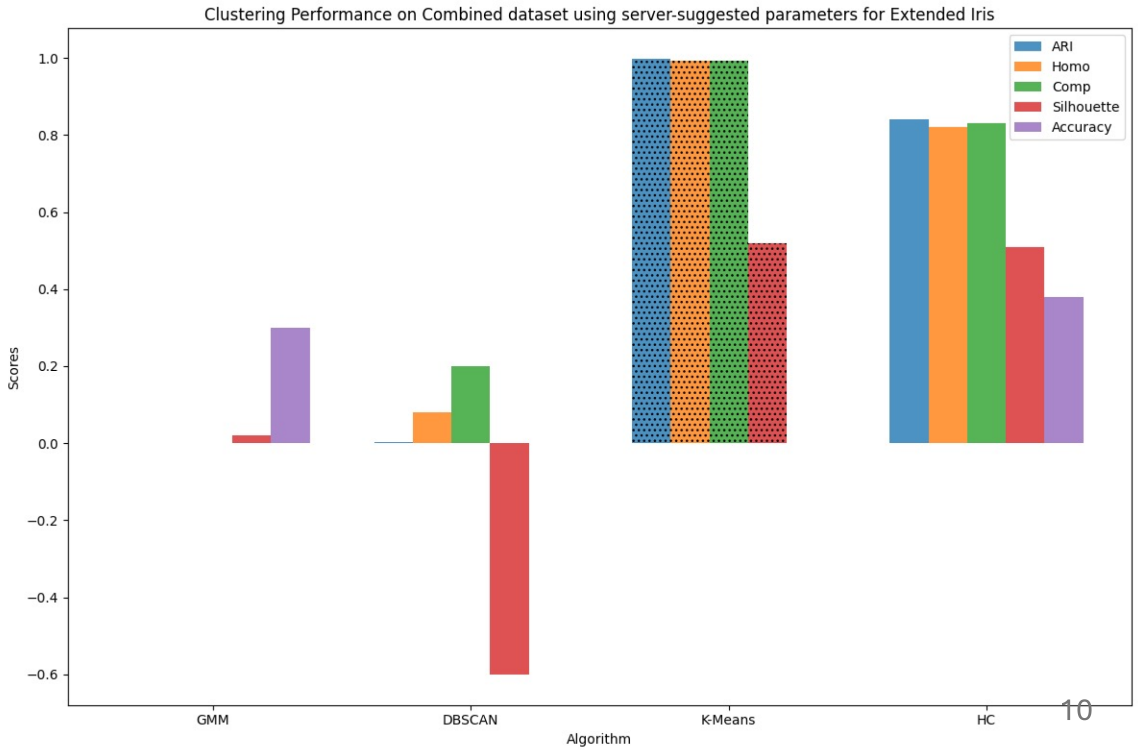
- **Evaluation Metrics**:
  - ***Adjusted Rand Index (ARI)*****:** Measures the similarity between the predicted and true clusters.
  - ***Silhouette Score*****:** Assesses how similar data points are within their clusters.
  - ***Calinski-Harabasz Index (CH)*****:** Evaluates the ratio of between-cluster dispersion to within-cluster dispersion.
  - ***Classification Accuracy:*** Although unusual for clustering, used here to assess how well clusters match known labels

SERVER RECOMENDATION

Table 4: Server Suggestions for Clustering Input Parameters: Recommendations for various clustering algorithms based on 10% shared noisy data ($\epsilon = 0.1$).

Dataset	Algorithm	Data shared to Server	ϵ	K or Eps	Silhouette	CH
Dataset #1	GMM	10%	0.1	k = 8	0.34	301.30
	DBSCAN	10%	0.1	k = 10, Eps = 1	-	-
	K-Means	**10%**	**0.1**	**k = 8**	**0.36**	**318.13**
	HC	10%	0.1	k = 8	0.31	237.61
Dataset #2	GMM	10%	0.1	k = 3	0.23	46.88
	DBSCAN	10%	0.1	k = 6, Eps = 7	-	-
	K-Means	**10%**	**0.1**	**k = 3**	**0.36**	**61.92**
	HC	10%	0.1	k = 3	0.37	51.57





RESULTS-CLUSTERING QUALITY

❑ Impact of Privacy Parameter (ϵ):

- *Explanation*: ϵ controls the level of noise added; lower ϵ means more noise and higher privacy.

- *Results*:
  - Server's recommendations remained stable across different ϵ values.
  - Clustering quality, as measured by ARI and Silhouette Score, was largely unaffected by noise.

Table 5: Differential Impact of Privacy Levels on Clustering Algorithms in the dataset #1. This table explores the performance variations (measured through ARI, Silhouette, and Accuracy) of four distinct clustering algorithms (K-Means, HC, GMM, DBSCAN) at different privacy budget levels (ϵ = 0.1, 1, 5) with a consistent data sharing percentage (10%).

Algorithm	Shared	ϵ	K	ARI	Silhouette	Accuracy
K-Means	10%	0.1	k = 8	0.75	0.41	0.18
K-Means	10%	1	k = 8	0.75	0.41	0.18
K-Means	10%	5	k = 7	1	0.44	0.15
HC	10%	0.1	k = 8	0.481	0.39	0.005
HC	10%	1	k = 7	0.482	0.41	0.17
HC	10%	5	k = 8	0.482	0.41	0.005
GMM	10%	0.1	k = 6	0.185	-0.0143	0.201
GMM	10%	1	k = 8	0.2069	-0.072	0.05
GMM	10%	5	k = 6	0.2008	-0.007	0.14
DBSCAN	10%	0.1	k = 10	0.017	-0.504	0.005
DBSCAN	10%	1	k = 10	0.017	-0.504	0.005
DBSCAN	10%	5	k = 10	0.017	-0.504	0.005

RESULTS-DATA SHARING

Effect of Data Sharing Volume:

- *Experiment*: Compared the clustering outcomes when 10%, 30%, and 50% of the data were shared.
- *Findings*:
  - The server's recommendations were consistent regardless of the amount of data shared.
  - Clustering results (ARI, Silhouette, CH Index) were robust to changes in the data sharing volume.

Table 7: Impact of Data Sharing Proportions on Clustering Algorithms' Performance in the dataset #1. This table evaluates how different proportions of data shared with the server (10%, 30%, 50%) influence the clustering outcomes (ARI, Silhouette, and Accuracy) for various algorithms (K-Means, HC, GMM, DBSCAN) at a fixed privacy parameter ($\epsilon = 0.1$).

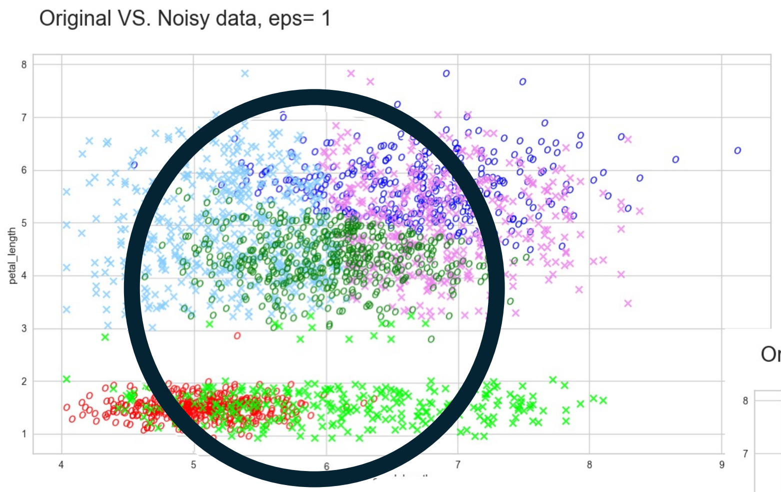
Algorithm	Shared	ϵ	K	ARI	Silhouette	Accuracy
K-Means	10%	0.1	k = 8	0.75	0.41	0.18
K-Means	30%	0.1	k = 8	0.75	0.41	0.18
K-Means	50%	0.1	k = 8	0.75	0.41	0.18
HC	10%	0.1	k = 8	0.481	0.39	0.005
HC	30%	0.1	k = 8	0.481	0.39	0.005
HC	50%	0.1	k = 8	0.0.481	0.39	0.005
GMM	10%	0.1	k = 6	0.185	-0.143	0.201
GMM	30%	0.1	k = 8	0.175	-0.111	0.18
GMM	50%	0.1	k = 5	0.169	-0.001	0.23
DBSCAN	10%	0.1	k = 10	0.017	-0.504	0.005
DBSCAN	30%	0.1	k = 10	0.017	-0.504	0.005
DBSCAN	50%	0.1	k = 10	0.017	-0.504	0.005

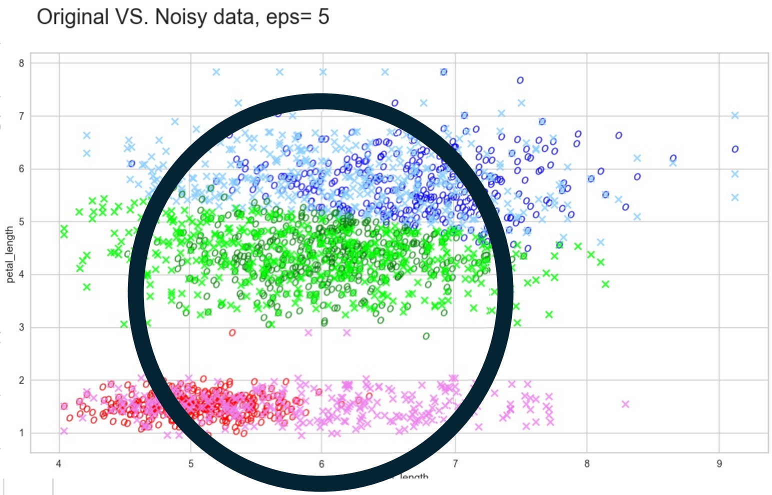
Conclusion: Effective clustering can be achieved even with minimal data sharing, enhancing privacy.

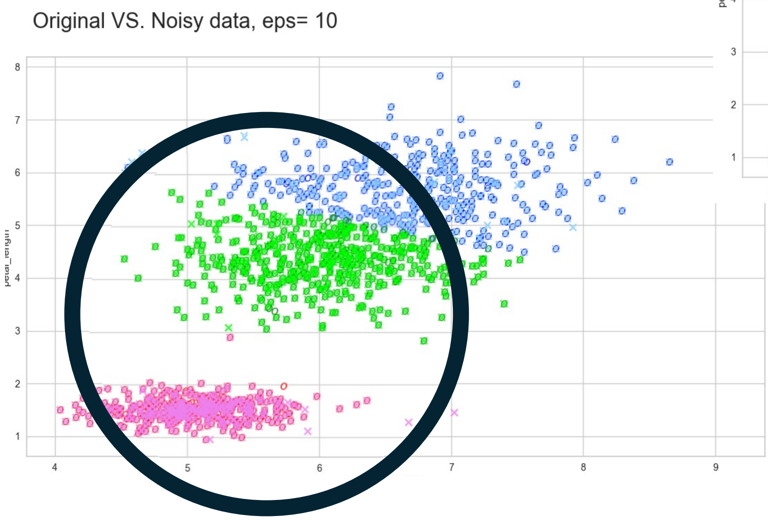
CASE WESTERN RESERVE UNIVERSITY
Case School of Engineering
Computer and Data Sciences

EFFECT of RANDOMIZED RESPONSE

- **Purpose of Randomized Response**:
  - *Add Noise to Data*: Introduces noise to individual data points to enhance privacy.
  - *Preserve Data Structure*: Despite noise, the underlying structure and gaps between clusters are preserved.
- **Key Observations**:
  - *Maintaining Cluster Gaps*: The RR mechanism effectively maintains the separation (gaps) between clusters.
  - *Impact of ϵ*: Different levels of the privacy parameter ϵ affect the amount of noise, but the distinctiveness between clusters remains.

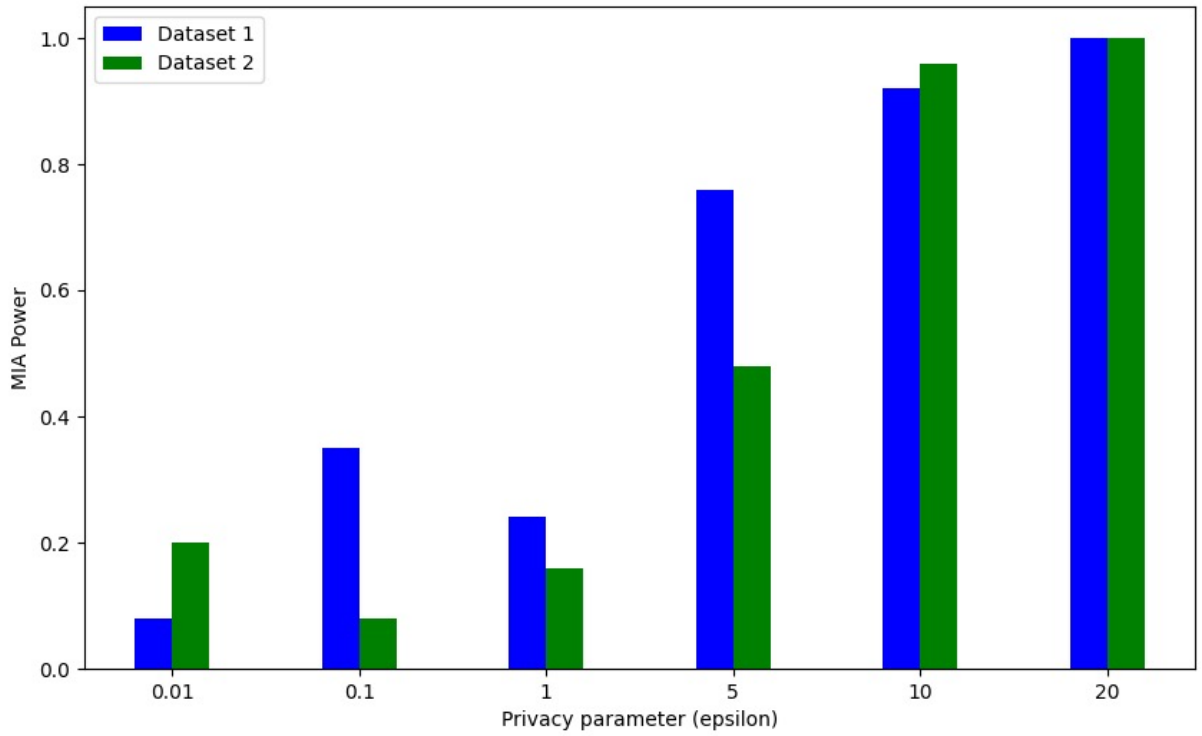






PRIVACY ANALYSIS

- **Membership Inference Attack**:
  - *Risk Analysis*: As ϵ increases (less noise), the risk of membership inference attacks also increases.



CONCLUSION

❑ **Summary of Contributions**:

- Developed a privacy-preserving framework for optimal parameter selection in collaborative clustering.
- Demonstrated the effectiveness of the proposed method through robust evaluation.

❑ **Future Work**:

- Explore other clustering algorithms and privacy mechanisms.
- Investigate further into mitigating risks associated with membership inference attacks.
- Expand the framework to more complex and diverse datasets.

Contact Maryam Ghasemian with any Questions:

Email: maryam.ghasemian@case.edu