Published: 22 June 2025

SIGMOD/PODS '25: International Conference on Management of Data
June 22 - 27, 2025
Berlin, Germany

Conference Sponsors:
SIGMOD

SHORT-PAPER

Demonstration of DPClustX: Differentially Private Explanations for Clusters

AMIR GILAD, Hebrew University of Jerusalem, Jerusalem, Israel

T. MILO, Tel Aviv University, Tel Aviv-Yafo, Tel Aviv District, Israel

KATHY RAZMADZE, Tel Aviv University, Tel Aviv-Yafo, Tel Aviv District, Israel

RON ZADICARIO, Tel Aviv University, Tel Aviv-Yafo, Tel Aviv District, Israel

Open Access Support provided by:

Tel Aviv University

Hebrew University of Jerusalem

SIGMOD/PODS '25: Companion of the 2025 International Conference on Management of Data (June 2025)
https://doi.org/10.1145/3722212.3725102
ISBN: 9798400715648

Demonstration of DPClustX: Differentially Private Explanations for Clusters

Amir Gilad
Hebrew University
Jerusalem, Israel
amirg@cs.huji.ac.il

Tova Milo
Tel Aviv University
Tel Aviv, Israel
milo@post.tau.ac.il

Kathy Razmadze
Tel Aviv University
Tel Aviv, Israel
kathyr@mail.tau.ac.il

Ron Zadicario
Tel Aviv University
Tel Aviv, Israel
ronzadicario@mail.tau.ac.il

Abstract

We present DPClustX, a framework designed to generate Differential Privacy (DP)-compliant histogram-based explanations for black-box clustering results. Interpreting clustering results is challenging even without privacy constraints, and the challenge is amplified under DP, as analysts receive noisy query responses. By addressing these challenges, DPClustX enables the selection of high-quality explaining attributes and the generation of informative histograms that balance privacy and utility. Compatible with any DP-clustering algorithm, our framework provides explanations that uncover significant patterns even under strict privacy constraints. This demonstration highlights DPClustX's capabilities using real datasets, illustrating its practical utility in sensitive data analysis and its potential for enhancing the interpretability of clustering methods.

CCS Concepts

• Security and privacy → Privacy-preserving protocols; **• Information systems → Data analytics**.

Keywords

Differential Privacy, Interpretability, Explanations, Clustering

ACM Reference Format:
Amir Gilad, Tova Milo, Kathy Razmadze, and Ron Zadicario. 2025. Demonstration of DPClustX: Differentially Private Explanations for Clusters. In *Companion of the 2025 International Conference on Management of Data (SIGMOD-Companion '25), June 22–27, 2025, Berlin, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3722212.3725102

1 Introduction

Sensitive data collection is ubiquitous in modern society, spanning domains such as fitness tracking, healthcare, financial services, and insurance. This wealth of data necessitates specialized handling to ensure its safe and secure usage without exposing individuals to potential harm. Differential Privacy (DP) [7] has emerged as the gold standard for addressing these challenges, enabling data analysis while bounding privacy leakage. DP has been widely adopted by companies and governmental organizations due to its ability to ensure privacy protection while maintaining analytical utility.

Clustering, a fundamental data analysis task, is commonly used to uncover patterns in complex datasets. However, interpreting clustering results poses significant challenges, and often necessitates an extensive analytical process. Interpreting clustering results under DP is even more challenging, as analysts are provided with noisy responses to queries, and longer, manual exploration sessions require additional noise to meet privacy constraints.

The integration of DP into clustering explanations introduces several challenges. First, existing methods rely on quality measures that require significant distortion to satisfy DP, making it difficult to generate reliable explanations. Second, traditional approaches often generate a pool of possible explanations before selecting the best ones, a process that incurs excessive privacy costs. Finally, balancing utility, interpretability, and privacy remains a critical but underexplored challenge.

To bridge the gap between privacy and interpretability, we present DPClustX, a novel system for generating histogram-based explanations of clustering results under DP. A unique feature of our system is its integration with a Large Language Model (LLM) plugin, which provides human-readable, Natural Language explanations accompanying the histograms. DPClustX is implemented as a Python package with an intuitive interface, allowing analysts to easily generate DP explanations for clustering results. We illustrate the use of DPClustX through the following example.

Example 1.1. Consider a healthcare analyst working with the Diabetes dataset [1], aiming to identify groups of patients with similar medical records, and assess their diabetes risk. While a clustering algorithm may privately provide cluster centers, it does not offer additional insights about the clusters. The analyst identifies Cluster 3 as having the highest diabetes rate. The textual summary of the histogram explanation for Cluster 3 generated by DPClustX is:

> **LLM:** The 'Income' column values differ significantly. Values below 25K are 7.3 times more frequent in Cluster 3 (84.8%) than in the remaining data (11.6%).

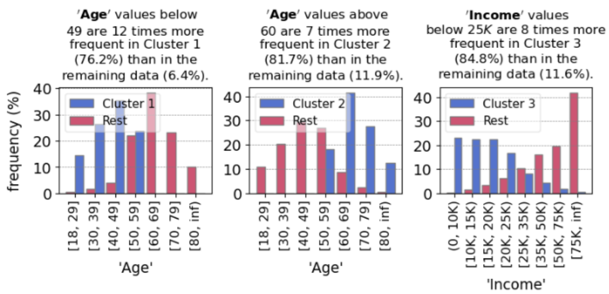
with an example of the full explanation shown in Figure 1a. The explanation indicates that Cluster 3 consists primarily of individuals with lower income, revealing a significant pattern captured by the clustering algorithm. Comparing the 'Income' distribution of Cluster 3 with the remaining data, we see that values outside Cluster 3 are concentrated in the higher ranges, peaking at $[75K, \infty)$. In contrast, Cluster 3 values differ greatly, being concentrated in the lower range, peaking at $(0, 10K)$. This suggests that the clustering algorithm groups individuals with lower incomes in Cluster 3.

DPClustX addresses the limitations of prior work with several key innovations. First, it introduces DP-tailored quality functions that minimize distortion and allow for reliable, high-utility explanations. Second, it reduces privacy costs by directly evaluating attribute quality and selecting high-quality attributes through a

SIGMOD-Companion '25, June 22–27, 2025, Berlin, Germany

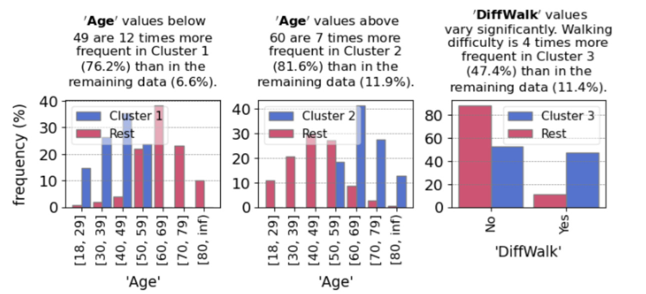
ACM ISBN 979-8-4007-1564-8/2025/06
https://doi.org/10.1145/3722212.3725102

```python
from DPClustX import ClusteringExplainer
explainer = ClusteringExplainer(clustered_df, mode='admin')
explainer.explain()
```



(a) Output in admin mode, showing the 'true' explanation.

```python
from DPClustX import ClusteringExplainer
explainer = ClusteringExplainer(clustered_df, mode='user')
explainer.explain(eps_topcomb=0.001)
```



(b) Output in user mode with default privacy budgets for $\varepsilon_{CandList}$, ε_{Hist}, and with a lower budget of 0.001 for $\varepsilon_{TopComb}$.

Figure 1: Examining the impact of privacy on the generated explanation with the Diabetes dataset.

novel DP mechanism, focusing only on the most relevant ones. This approach ensures that the system requires less noise to generate private histograms. Third, DPClustX leverages the One-shot Top-k mechanism [6] to efficiently construct candidate attribute sets, enabling real-time, interactive analysis. By integrating an LLM plugin, DPClustX enhances interpretability by converting histogram-based explanations into natural language summaries. This feature makes the system particularly accessible for non-expert users while maintaining the strict privacy guarantees of DP.

Demonstration Overview. This demonstration highlights the capabilities of DPClustX using real-world datasets. Participants can interactively:

- Explore clustering explanations under varying DP budgets.
- Compare private and non-private explanations to understand the trade-offs between privacy and utility.
- Experiment with different clustering methods, including DP-k-means, k-modes, and Gaussian Mixture Models.

Related work. Existing methods for explaining clustering algorithms (e.g., [10]) focus primarily on non-private, white-box approaches. While these methods offer interpretable summaries by highlighting how clusters differ, they are not directly adaptable to the DP setting. Histogram-based explanations for black-box clustering methods (e.g., [4]) provide a more general approach but fail to address the privacy costs associated with generating explanations for all attributes and clusters. DPClustX advances the state of the art by addressing these limitations and demonstrating the feasibility of privacy-preserving clustering explanations in practical scenarios.

2 Technical Background

We now give the necessary background for our algorithm. Given a single table schema $R(A_1, \ldots, A_d)$, R is a relation name and $\mathcal{A} = \{A_1, \ldots, A_d\}$ denotes the set of attributes. Each attribute A_i has discrete, finite, and data-independent domain $\mathrm{dom}(A_i)$. The full domain of R is $\mathrm{dom}(R) = \mathrm{dom}(A_1) \times \cdots \times \mathrm{dom}(A_d)$. An instance (dataset) D of a relation R is a finite bag (multiset) whose elements are tuples in $\mathrm{dom}(R)$. For a dataset D and an attribute A, we let $h_A(D) \in (\mathbb{N} \cup \{0\})^{\mathrm{dom}(A)}$ denote the histogram of the dataset D over attribute A. That is, for any $a \in \mathrm{dom}(A)$, $h_A(D)[a]$ is the number of occurrences of value a in the column $\pi_A(D)$. For visualizations, we use normalized histograms, where each count is replaced by its proportion.

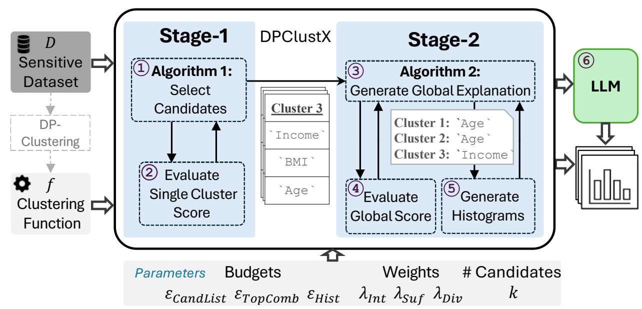


Figure 2: DPClustX **explanation generation pipeline.**

Histogram-based explanations (HBE). A clustering of D is a partition into disjoint subsets $\{D_c \subseteq D \mid c \in C\}$ each assigned with a cluster label $c \in C$. A *single-cluster explanation candidate* for D_c consists of two histograms on a specified attribute: the histogram of the cluster values, and the histogram of the values outside the cluster. Formally, it is a tuple $(c, A, h_A(D \setminus D_c), h_A(D_c))$ where $A \in \mathcal{A}$. A *global explanation candidate* is a set of single cluster explanation candidates, with a candidate for each cluster.

Differential Privacy. We consider generating clustering explanations under *differential privacy* (DP) [7] to protect sensitive information. DP enables querying a private database without compromising privacy by ensuring that the distribution of outputs does not significantly change when the algorithm is applied to any two distinct but *neighboring datasets*. Formally, a randomized clustering explanation mechanism $\mathcal{M}$ is said to satisfy ε-DP if for any neighboring datasets $D{\sim}D'$ (D' can be obtained from D by removing or adding one tuple), any clustering function $f : \mathrm{dom}(R) \to C$, and any set of possible outputs $S \subseteq Range(\mathcal{M})$, $\mathbf{Pr}[\mathcal{M}(D, f) \in S] \leq e^{\varepsilon} \cdot \mathbf{Pr}[\mathcal{M}(D', f) \in S]$, where the probability is over the randomness of $\mathcal{M}$.

2.1 Quality Functions for Private Explanations

DPClustX focuses on three prominent quality aspects of histogram-based explanations considered in prior work: Sufficiency, interestingness , and diversity. We prove that these measures are highly sensitive, making them unsuitable for a DP algorithm. These results motivate the design of low-sensitivity variants that capture the essence of each measure while minimizing distortion for DP compliance. For all the provided metrics below, the formal definitions and sensitivity analysis of our adapted measures are in [2].

Sufficiency. To ensure that a given explanation is relevant only to its associated cluster, we adopt the notion of sufficiency from prior work in the non-private setting (e.g., [4]). Informally, the sufficiency of an explanation at $t \in D$ is defined as the fraction of tuples assigned the same cluster as t, out of all tuples for which the explanation of t "holds". A global sufficiency score is obtained by averaging the values at all tuples. To quantify the extent to which an explanation "holds for" a tuple $t \in D_c$, prior work uses the following probability value derived from the HBE: the probability that a uniformly random tuple sampled from D conditioned on having the same value in A_c as t, belongs to D_c, where A_c is the attribute used to explain the cluster D_c. Unfortunately, this function is too sensitive to be of use. We propose a low-sensitivity variant and prove that it preserves the original attribute ranking.

Interestingness. The interestingness of an HBE is quantified as the distributional shift of the given attribute values between the cluster and the full-dataset, with the global interestingness being the average across all clusters (e.g [4, 5]). There are many ways to quantify distance between probability distributions. However, prior work focuses on distance metrics which are too sensitive to be of use in a DP algorithm. Intuitively, when a cluster is small, the removal of a single tuple from it can significantly change the cluster's internal distribution, which may lead to a large change in the distribution distance. To account for these issues, we introduce a low-sensitivity variant that rescales scores relative to cluster sizes, mitigating the impact of small clusters while preserving attribute ranking for each cluster.

Diversity. This measure quantifies the overall distinctiveness among explanations. Diversity is initially defined for a pair of single-cluster explanations and then generalized to global explanations by averaging all pairwise diversities. It attains its maximum value of 1 when the explanations use different attributes, and equals their distribution distance when they share the same attribute [4]. Since diversity relies on sensitive distance metrics, existing definitions are unsuitable for the DP setting, and cannot be applied directly. We introduce a low-sensitivity variant, scaling pairwise distances by cluster sizes, thus allowing identification of diverse explanations despite the noise injected for DP.

Overall score. The *global* score combines all individual measures for a comprehensive assessment of explanation quality. It is defined as a weighted average of the three components mentioned earlier, with adjustable weights for specific needs or equal weights by default. Additionally, DPClustX uses a *single-cluster* score function, combining both interestingness and sufficiency, to privately filter high-quality candidates for each cluster.

3 Algorithm and System Architecture

DPClustX is designed to generate privacy-preserving explanations for clustering results while maintaining high utility. The algorithm comprises two key stages, illustrated in Figure 2, that ensure efficient, high-quality, and interpretable outputs under differential privacy (DP).

Candidate set generation. To address the vast search space for clustering explanations, DPClustX constructs a high-quality candidate set of explanations for each cluster. This step, illustrated as Stage 1 in Figure 2, begins by adapting a single-cluster score function to rank explanations by their quality. The score function is tailored to the DP setting to preserve utility even with the added noise. To privately select the Top-k explanations for each cluster, DPClustX leverages the One-shot Top-k mechanism [6]. This mechanism ensures computational efficiency and minimizes privacy budget consumption, making the system suitable for interactive use. The resulting candidate pool forms the foundation for the global explanation generated in the next stage.

Private global explanation. Generating private explanations for all candidates and subsequently selecting the top-scoring ones can be inefficient and wasteful in terms of privacy budget. To address this, DPClustX uses a novel DP approach that ranks attributes by their explanatory quality and employs a DP selection mechanism to identify globally high-quality attributes.

As shown in Stage 2 of Figure 2, only the selected attributes are used to generate noisy histograms. This significantly reduces the amount of noise required for DP compliance, ensuring that the final explanations remain interpretable and informative while meeting privacy guarantees.

These stages are underpinned by rigorous DP and utility guarantees, ensuring that the explanations generated align closely with their non-private counterparts while preserving privacy.

Implementation. DPClustX is implemented in Python 3.9, leveraging the Pandas and NumPy libraries. The Geometric mechanism for DP histogram generation is implemented using DiffPrivLib [8]. Additionally, the system integrates an LLM to automatically generate textual descriptions for each histogram-based explanation, enhancing interpretability for users.

The system supports two operational modes:

- Admin Mode: Allows access to the "ground truth" explanation generated without privacy guarantees (no privacy budget applied). In Stage 1, DPClustX selects the true Top-k attributes for each cluster. In Stage 2, the true quality-maximizing attributes are selected for the global explanation, outputting undistorted histograms on those attributes. Admins can choose to execute certain steps privately and others non-privately, as each component of the framework has an individual privacy parameter.
- End User Mode: Enforces privacy constraints, generating noisy explanations to ensure DP compliance. A default budget is applied if none is specified.

DPClustX is an open-source library, packaged for easy installation via pip, enabling researchers and practitioners to seamlessly integrate it into their workflows. The source code, along with detailed documentation and examples, is publicly available [2] to encourage community contributions and adoption.

4 Demonstration

We present DPClustX, a framework designed to generate privacy-preserving histogram-based explanations for clustering outcomes. The demonstration showcases DPClustX's capabilities through interactive, real-time experiments using the US Census [9] and Diabetes [1] datasets. The demonstration utilizes a Jupyter notebook to present an application of the system within a real-world data exploration setting. Each phase emphasizes a specific functionality and includes a real-world use case that illustrates the system's practical value[3].

Phase 1: System overview and interactive exploration. Participants will be introduced to DPClustX's architecture and interface via a pre-prepared Jupyter notebook. The system will be instantiated in admin mode, providing access to ground truth explanations. Using the Diabetes dataset, pre-clustered into groups, participants will interact with the system to examine how individual attributes contribute to clustering results. We highlight the transformation of raw clustering results into interpretable histograms that explain the role of selected attributes in defining each cluster (Figure 1a). Participants will adjust system parameters, like diversity and sufficiency weights, to explore how different configurations affect the explanation. They may also modify the attribute set considered for each cluster to observe different patterns.

Healthcare Domain. A healthcare analyst aims to understand the factors driving a clustering of the diabetes patient dataset. Using DPClustX, they discover that specific attributes like "Age" and "Income" play key roles in defining clusters. By adjusting the attributes considered, the analyst explores different explanations to ensure that the clusters provide interpretable insights for policymaking, such as identifying high-risk groups for targeted interventions. Participants will replicate this process and observe how different configurations affect the generated explanations. This hands-on demonstration emphasizes DPClustX's flexibility and its ability to capture significant patterns.

Phase 2: Impact of differential privacy on explanations. In this phase, participants will evaluate how DP affects explanation quality. Switching between admin mode (non-private) and user mode (privacy-preserving), attendees will explore the impact of privacy budgets on the generated explanations. Leveraging the Diabetes dataset for experimentation, users can compare DP explanations with non-private ones. For example, explanations generated with default privacy budgets are nearly identical to non-private ones. Further reducing the privacy budget shows that DP explanations continue to closely align with the non-private counterparts (Figure 1). This highlights the system's usability, enabling users to privately explore clustering results without DP noise distorting the explanations. Furthermore, participants will examine explanations generated using a direct DP adaptation of an existing non-private approach [4], thereby recognizing the impracticality of sensitive HBE quality functions under DP.

Healthcare Domain cont. A government health agency is tasked with publishing anonymized insights from a diabetes dataset while preserving patient privacy. The agency uses DPClustX to generate DP explanations that reveal trends without leaking sensitive data. For instance, a cluster predominantly composed of older patients with mobility issues may be identified using DP explanations. Participants will adjust privacy parameters (e.g., $\varepsilon\_TopComb$) to observe their influence on the utility of explanations. This phase demonstrates how DPClustX balances privacy guarantees with the need for actionable insights.

Phase 3: Comparing clustering methods. In this phase, participants will explore the versatility of DPClustX by applying it to the US Census and experimenting with multiple clustering methods, including DP-k-means, non-private k-means, and Gaussian Mixture Models. The system generates HBEs for each clustering result, enabling participants to visually and quantitatively compare the clustering behaviors. For instance, k-modes may highlight categorical features such as "Marital Status," while DP-k-means emphasizes numeric attributes like "Age." These comparisons allow participants to assess the strengths and weaknesses of each method in grouping data and explain the role of key attributes in the clustering results.

Social Science domain. A sociologist seeks to analyze patterns in income distribution across demographic groups using the US Census dataset. For example, DPClustX may identify groups with low income and high education, revealing areas for targeted policy interventions. By providing DP explanations, DPClustX ensures that sensitive demographic details remain protected while still offering actionable insights. Participants will replicate this analysis and evaluate how clustering methods and privacy settings affect the generated explanations, illustrating DPClustX's adaptability to different clustering techniques and datasets.

Acknowledgments

This research was supported by the Israel Science Foundation (ISF) under grant 2707/22 of the Breakthrough Research Grant (BRG) Program. The work of Amir Gilad was funded by the Israel Science Foundation (ISF) under grant 1702/24, the Scharf-Ullman Endowment, and the Alon Scholarship.

References

[1] 2022. Diabetes Health Indicators Dataset. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.
[2] 2024. DPClustX Git Repository. https://anonymous.4open.science/r/DPClustX/
[3] 2025. Video of demonstration scenario. https://youtu.be/LeKkVdlyiXo?si=8QV5h7tl3H1ZIwfJ.
[4] Roni Copul, Nave Frost, Tova Milo, and Kathy Razmadze. 2024. TabEE: Tabular Embeddings Explanations. *Proceedings of the ACM on Management of Data* 2, 1 (2024), 1–26.
[5] Daniel Deutch, Amir Gilad, Tova Milo, Amit Mualem, and Amit Somech. 2022. FEDEX: An Explainability Framework for Data Exploration Steps. *Proceedings of the VLDB Endowment* 15, 13 (2022), 3854–3868.
[6] David Durfee and Ryan M Rogers. 2019. Practical differentially private top-k selection with pay-what-you-get composition. *Advances in Neural Information Processing Systems* 32 (2019).
[7] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.
[8] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. 2019. Diffprivlib: the IBM differential privacy library. *arXiv preprint arXiv:1907.02444* (2019).
[9] Chris Meek, Bo Thiesson, and David Heckerman. 2001. US Census Data (1990). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5VP42.
[10] Sariel Tutay and Amit Somech. 2023. Cluster-Explorer: An interactive Framework for Explaining Black-Box Clustering Results. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 5106–5110. https://doi.org/10.1145/3583780.3614734