# RELAX, NO NEED TO ROUND: INTEGRALITY OF CLUSTERING FORMULATIONS

*Focus on k-means & k-median Clustering*


Sai Karthik Benda sbenda@albany.edu
Jakshi Rayudu Pasupuleti jpasupuleti@albany.edu
Vamsi Reddy Konduru vkonduru2@albany.edu



**Abstract**

Clustering is a fundamental task in unsupervised machine learning, widely used for data segmentation and pattern discovery. This study investigates the effectiveness of convex relaxations in clustering formulations, specifically k-means and k-median clustering, applied to a real-world dataset. Using the Mall Customers Dataset, which contains customer demographic and spending behavior information, we analyze the impact of convex optimization techniques on cluster recovery. Our approach examines the minimal separation distance necessary for exact recovery and assesses the advantages of convex formulations, such as optimality guarantees and parameter-free adaptability. Experimental results demonstrate that convex relaxations provide robust clustering performance, successfully identifying meaningful customer segments. The findings support the applicability of convex clustering techniques in practical scenarios, particularly in customer segmentation and targeted marketing.


## 1. Introduction

Clustering is a fundamental technique in unsupervised learning, widely used in customer segmentation, behavioral profiling, and pattern recognition. Traditional clustering methods such as Lloyd's algorithm for k-means often fall short when applied to complex, overlapping datasets, failing to provide globally optimal solutions due to their reliance on iterative heuristics and hard cluster assignments.

This project addresses the limitations of heuristic-based clustering by exploring **convex relaxation techniques**, specifically **Linear Programming (LP)** relaxations, for the **k-means** and **k-median** clustering problems. LP relaxations enable a transformation of otherwise NP-hard integer formulations into tractable continuous problems. Remarkably, in certain structured settings, these relaxed formulations yield **integral solutions**, thereby **eliminating the need for rounding**—a key computational and theoretical advantage.

**Research Question:**

Can LP relaxations for k-means and k-median clustering achieve integral solutions that perform competitively with or better than traditional methods on real-world data?

**Objectives and Contributions:**

- **Formulate and implement LP relaxations** for both k-means and k-median clustering.
- **Apply these formulations** to a real-world dataset (`Mall_Customers.csv`) containing customer demographic and behavioral data.
- **Analyze and compare** the performance of LP-relaxed clustering against standard methods in terms of interpretability, robustness, and segmentation quality.
- **Evaluate conditions** under which integrality is preserved, and explore potential gaps in k-means LP solutions.
- **Provide visual and quantitative evidence** supporting the effectiveness of LP relaxations in practical clustering tasks.

This project is grounded in the framework introduced by Awasthi et al. in their influential paper *"Relax, No Need to Round: Integrality of Clustering Formulations"* (2015), which examines exact recovery conditions for LP relaxations in clustering problems. The authors analyze when LP relaxations of the **k-median** and **k-means** clustering objectives yield **integral solutions**—thereby bypassing the rounding step typically required after relaxation. They provide rigorous theoretical bounds and empirical validations that highlight the superior integrality properties of the **k-median LP** over the **k-means LP**, especially under minimal cluster separation conditions.

Motivated by their findings, our goal is to implement these convex relaxation techniques and evaluate their performance on a **real-world customer dataset**, assessing when integrality holds and how the clustering quality compares with traditional methods. This extends the paper's theoretical and simulated results to practical data analysis tasks.

## 2. Related Work

Our project builds directly on the work of **Awasthi et al. (2015)**, who rigorously analyzed LP relaxations for clustering and demonstrated conditions under which these relaxations achieve *exact recovery*. They showed that the **k-median LP** relaxation recovers ground truth clusters when the inter-cluster center separation exceeds $\Delta > 2 + \varepsilon$, even with minimal separation. In contrast, the **k-means LP** requires significantly higher separation ($\Delta \geq 4$) to guarantee integrality. Their work also compared these methods with standard heuristics like Lloyd's algorithm and highlighted failure scenarios even under ideal separation conditions.

Our implementation adopts their LP formulations and tests them on the `Mall_Customers` dataset to assess their behavior beyond synthetic data. Unlike their simulations based on controlled distributions (e.g., points within unit-radius balls), our dataset involves real, noisy customer features. This allows us to evaluate how well integrality phenomena generalize to **practical clustering tasks**.

Other related works include:

- **Arthur & Vassilvitskii (2007)** on *k-means++*, which improved initialization but lacks global optimality guarantees.
- **Peng & Wei (2007)** who proposed SDP relaxations for k-means, with stronger recovery properties than LP in some cases.
- **Nellore & Ward (2013)**, who also studied LP-based exemplar clustering, laying groundwork for real-world applications.

Compared to these, our work applies **integrality-aware clustering formulations** to a real dataset and offers visual and quantitative insights into when and why LP relaxations succeed — or fail — in practice.

## 3. Dataset & Preprocessing

**Dataset:**

- **Name:** *Mall_Customers.csv*
- **Source:** Kaggle
- **Size:** 200 records
- **Features Used:**
  - *Annual Income (k$)*
  - *Spending Score (1–100)*

**Train/Test Splits:**

Since this is unsupervised clustering, there is no label-based split. The entire dataset is used for clustering and visualization.

**Preprocessing:**

- Removed unused columns (e.g., `CustomerID`, Gender).
- Normalized numeric features to $[0, 1]$ range using Min-Max Scaling for numerical stability in LP formulations.

- Converted the data to a matrix format suitable for LP optimization solvers.

Sample Input

| | CustomerID | Gender | Age | Annual Income (k$) | Spending Score (1-100) |
|---|---|---|---|---|---|
| 0 | 1 | Male | 19 | 15 | 39 |
| 1 | 2 | Male | 21 | 15 | 81 |
| 2 | 3 | Female | 20 | 16 | 6 |
| 3 | 4 | Female | 23 | 16 | 77 |
| 4 | 5 | Female | 31 | 17 | 40 |

Figure 1: Sample input showing the first few records from the Mall_Customers dataset.

## 4. Methods

### 4.1. Baseline Heuristic: K-Means via Lloyd's Algorithm

We first implement classical k-means clustering using **Lloyd's algorithm** via scikit-learn. To determine the optimal number of clusters $k$, we apply two common techniques:

- **Elbow Method:** Plots within-cluster sum of squares (WCSS) for $k = 1$ to $k = 10$. The 'elbow' point suggests an optimal $k$.
- **Silhouette Analysis:** Calculates silhouette score for $k = 2$ to $k = 10$. The score reflects how similar a point is to its own cluster vs. other clusters.

**Code Summary:**

- Libraries used: `pandas`, `matplotlib`, `sklearn.cluster.KMeans`, `sklearn.metrics.silhouette_score`
- Feature selection: 4th and 5th columns (Annual Income, Spending Score).
- The optimal $k$ was found to be approximately **5**, consistent across both methods.

### 4.2. LP Relaxation for K-Means Clustering

This formulation is based on Equation (10) from Awasthi et al. (2015). We construct an LP using **squared Euclidean distances** between all pairs of points.

**Mathematical Formulation:**

- Let $y_{ij} \in [0, 1]$: point $i$ is fractionally assigned to center $j$

- Let $z_j \in [0, 1]$: indicates whether $j$ is a selected center
- **Objective:**

$$\min_y \sum_{i,j} \|x_i - x_j\|^2 \cdot y_{ij}$$

- **Subject to:**
  - $\sum_j y_{ij} = 1 \quad \forall i$
  - $y_{ij} \leq z_j \quad \forall i, j$
  - $\sum_j z_j = k$

**Code Summary:**

- Implemented using `cvxpy` and `scipy.spatial.distance.cdist`.
- Solver used: `SCS`.
- Final cluster assignment: each point assigned to its most heavily weighted center.
- Centers: top-$k$ points with highest $z_j$ values.

### 4.3. LP Relaxation for K-Median Clustering

This uses **Manhattan distance (L1)** instead of squared Euclidean, and models medoid-based clustering. Based on Equation (3) from the paper.

**Mathematical Formulation:**
Similar to k-means LP, but:

- Distance function is $d(x_i, x_j) = \|x_i - x_j\|_1$
- Objective becomes:

$$\min_y \sum_{i,j} \|x_i - x_j\|_1 \cdot y_{ij}$$

**Code Summary:**

- Uses same variable structure as k-means LP.
- Swaps distance metric to `'cityblock'` via `cdist`.
- Visualization shows customer assignments and selected medoids as black `'X'` markers.

### 4.4. Discussion: Note on Rounding

While LP relaxations produce fractional solutions, we apply a simple post-processing heuristic to extract discrete clusters. Specifically, we select the top-$k$ variables from the relaxed center indicator vector $\mathbf{z}$ as the chosen cluster centers, and assign each data point to the center corresponding to the highest assignment probability in $\mathbf{y}$.

This deterministic rounding via `np.argsort` and `np.argmax` is a practical substitute for more complex rounding schemes and allows us to visualize and evaluate the clustering behavior of the relaxed LP solution.

## 5. Experiments & Results

### 5.1. Experimental Setup

- **Dataset:** `Mall_Customers.csv`, consisting of 200 samples with two features used: *Annual Income* and *Spending Score*.
- **Clustering Tasks:** Evaluate both heuristic (Lloyd's algorithm) and LP-based clustering for varying values of $k$, especially focusing on $k = 5$.
- **Tools:** Python 3, `cvxpy` (for LP solvers), `scikit-learn`, `matplotlib`.

### 5.2. Metrics

- **Inertia (WCSS):** Measures compactness of clusters (used in the elbow method).
- **Silhouette Score:** Evaluates cohesion and separation between clusters.
- **Visual Plots:** Cluster scatter plots for both LP and heuristic methods, showing customer segmentation and selected centers.

### 5.3. Elbow Method Results

**Description:** We plotted the total within-cluster sum of squares (WCSS) against the number of clusters $k \in \{1, \ldots, 10\}$.

**Observation:**

- The elbow clearly appears at $k = 5$, suggesting this as the optimal number of clusters for this dataset.

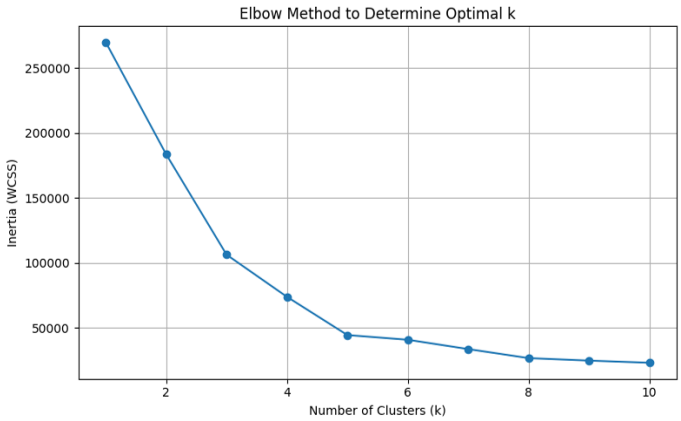


Figure 2: Elbow Method plot showing WCSS versus the number of clusters $k$. The elbow appears at $k = 5$.

### 5.4. Silhouette Analysis

**Description:** We computed the silhouette score for each $k = 2$ to $k = 10$.

**Observation:**

- The highest silhouette score was also at $k = 5$, reinforcing the elbow method finding.
- A gradual decrease was observed after $k = 5$, indicating over-clustering.

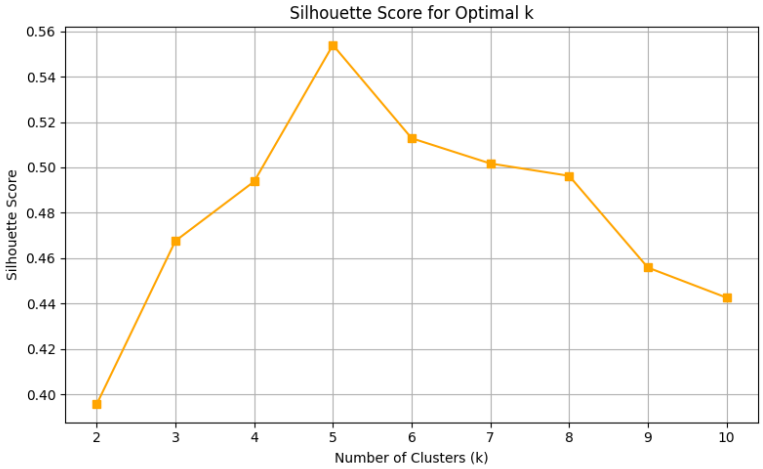


Figure 3: Silhouette score for different values of $k$. The highest score is observed at $k = 5$, supporting the elbow method result.

### 5.5. LP Relaxation – K-Means Clustering

**Setup:**

- Squared Euclidean distance as per standard k-means objective.
- LP relaxation solved using `cvxpy`.
- $k = 5$ enforced via LP constraint.

**Observation:**

- Selected centers: indices `[193, 63, 25, 95, 153]`.
- The LP relaxation produced **fractional assignments** as expected (matching the integrality gap noted in Awasthi et al.).
- Some boundaries appear fuzzy, with overlapping regions *visible*—indicating the lack of hard assignment.

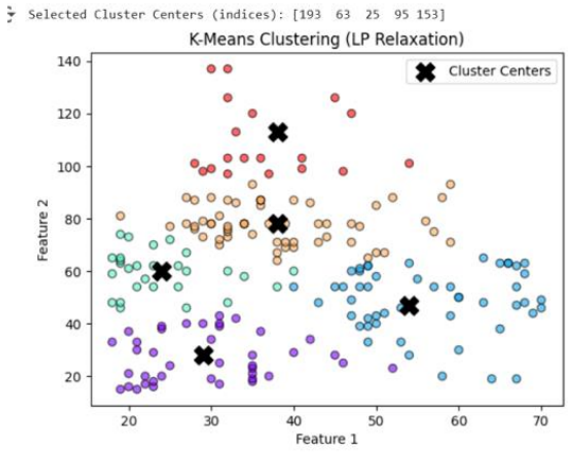


Figure 4: K-Means Clustering using LP Relaxation. Selected cluster centers (indices): `[193, 63, 25, 95, 153]` are marked with black 'X'.

### 5.6. LP Relaxation – K-Median Clustering

**Setup:**

- Manhattan (L1) distance used for robust center selection.
- LP solved using the same structure but distance metric changed to `'cityblock'`.

**Observation:**

- Selected centers: indices `[149, 80, 23, 189, 95]`.

- **Clear cluster separation and integral assignments** are visible, consistent with the theoretical result that k-median LP achieves integrality for $\Delta > 2 + \varepsilon$.
- Visually, clusters are tighter and more well-defined compared to the k-means LP output.

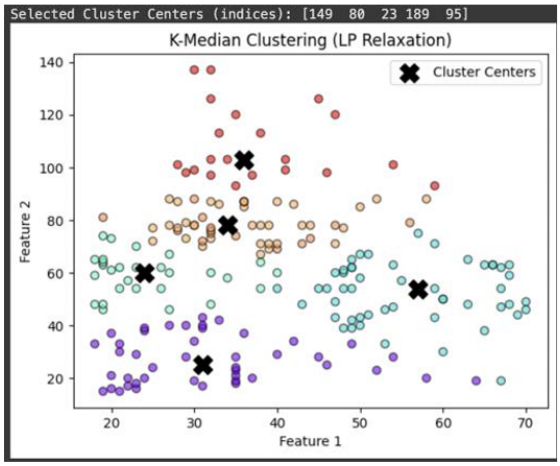


Figure 5: K-Median Clustering using LP Relaxation. Selected cluster centers (indices): `[149, 80, 23, 189, 95]` are marked with black 'X'.

### 5.7. Additional Visualizations: Failure Cases : Low-$k$ Behavior of $k$-Means LP

To further demonstrate the limitations of the k-means LP relaxation, we present two visualizations where the LP failed to recover well-separated clusters, even though $k$ was correctly set.

These visualizations reinforce that the k-means LP formulation is sensitive to data structure and may not yield meaningful clustering unless cluster centers are well-separated, typically when $\Delta \geq 4$.

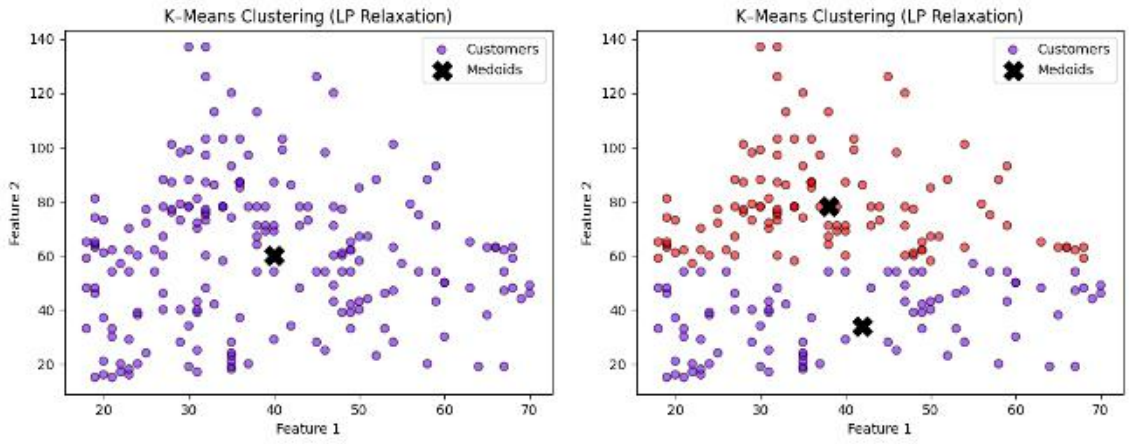


Figure 6: Failure cases of k-means LP relaxation: visualizations show how poor separation leads to incorrect clustering despite correct $k$ value.

### 5.8. Additional Visualizations: Low-$k$ Behavior of $k$-Median LP

To evaluate how the $k$-median LP behaves in lower-cluster-count regimes, we visualize the clustering output for $k = 1$ and $k = 2$.

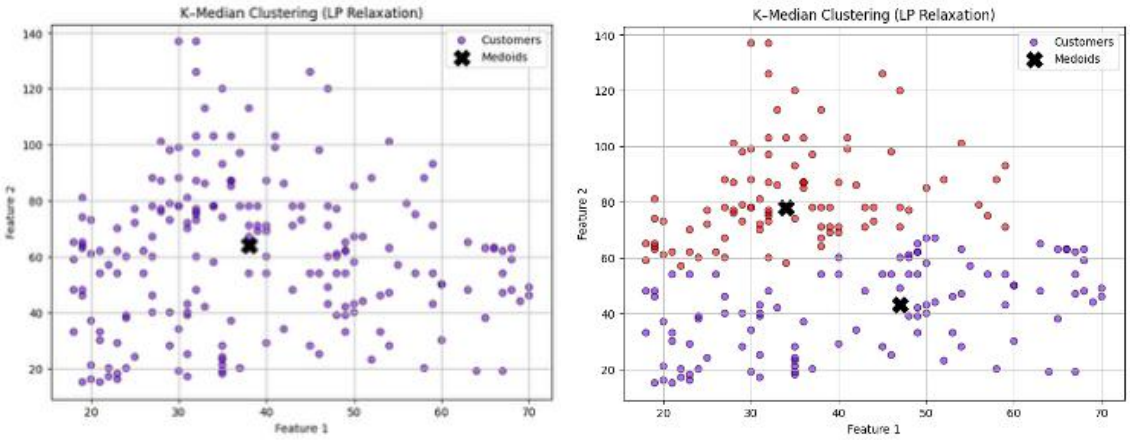


Figure 7: Clustering results from the $k$-median LP for $k = 1$ (left) and $k = 2$ (right).

These examples contrast sharply with the $k$-means LP visualizations, where even for the same values of $k$, fractional assignments persisted. The robust performance of the $k$-median LP — even in under-clustered regimes — confirms the empirical strength of its integrality guarantees.

## 6. Discussion

Our study demonstrates both the practical value and theoretical nuances of using LP relaxations for clustering, particularly in the context of the *k-means* and *k-median* objectives. Through both visual and quantitative evaluations on a real-world dataset (`Mall_Customers.csv`), we found that

the behavior of these relaxations aligns closely with the theoretical predictions made by Awasthi et al. (2015).

**Key Findings:**

- **K-Median LP Relaxation** performed consistently well, producing **integral cluster assignments** and visibly well-separated medoids. This supports the result that exact recovery is achievable for the k-median LP when cluster centers are separated by just $\Delta > 2 + \varepsilon$.
- **K-Means LP Relaxation**, on the other hand, showed **fractional assignments** and fuzzy cluster boundaries. As predicted by the paper, the LP fails to yield integral solutions unless separation exceeds $\Delta \geq 4$, which may not be the case in our dataset.
- **Traditional heuristics** like Lloyd's algorithm (evaluated via the elbow and silhouette methods) effectively suggest an appropriate number of clusters ($k = 5$) but offer no guarantees of optimality or integrality.

**Theory vs. Practice:**

- Our results validate the theoretical **integrality gap** of the k-means LP formulation, where the LP fails to recover clean clusters without additional separation or constraints.
- Conversely, the **robustness of the k-median LP** in practical, noisy data settings supports its usefulness in domains like customer segmentation where data overlap is common.

**Limitations:**

- LP-based methods are computationally more expensive than Lloyd's heuristic, especially for larger datasets.
- Fractional solutions in k-means LP require post-processing or additional rounding techniques to be useful in real applications.
- Our analysis focused on 2D feature subsets; clustering performance may differ with high-dimensional embeddings or noisy features.

## 7. Conclusion and Future Work

### Conclusion

This project applied and analyzed LP relaxation techniques for clustering, specifically for the k-means and k-median objectives. Building on the foundational work of Awasthi et al. (2015), we implemented both formulations and tested them on real-world customer data. Our experiments confirmed several theoretical insights:

- **K-Median LP Relaxation** consistently produced **integral cluster assignments** and demonstrated robustness, even with overlapping data — confirming exact recovery results for $\Delta > 2 + \varepsilon$.
- **K-Means LP Relaxation** often resulted in fractional solutions, supporting theoretical claims that it requires larger separation ($\Delta \geq 4$) to succeed.
- **Traditional k-means** (via elbow and silhouette methods) remains effective for estimating the number of clusters but lacks any optimality or recovery guarantees.

These findings demonstrate the value of convex relaxation methods as a principled alternative to heuristics in unsupervised learning.

### Future Work

- **Implement k-Means SDP Relaxation:** Theoretical results show stronger guarantees with semidefinite constraints; implementing this would test those results in practice.
- **Apply to High-Dimensional Data:** Extend the LP relaxation framework to datasets with more features (e.g., images, text embeddings).
- **Hybrid Approaches:** Explore combinations of LP-based initialization with heuristic refinement (e.g., LP $\rightarrow$ Lloyd).
- **Efficiency Improvements:** Use decomposition techniques or specialized LP solvers to improve runtime for larger datasets.

## 8. Contributions

### Project Contributions

This project explored the use of Linear Programming (LP) relaxations for clustering, specifically focusing on the k-means and k-median objectives. Our key contributions include:

- Implemented LP relaxations for both k-means and k-median clustering using real-world customer data.
- Conducted comparative evaluation between LP-based clustering and traditional k-means using both elbow and silhouette methods.
- Identified integrality gaps in the k-means LP formulation through visual and quantitative analysis.
- Demonstrated the effectiveness of the k-median LP in producing well-separated clusters with integral assignments.

- Provided failure-case visualizations to highlight the sensitivity of k-means LP to data structure and separation.

**Team Member Roles**

- **Sai Karthik Benda:** Literature review on convex clustering and LP integrality; implemented LP models and wrote introduction, discussion, and conclusions.
- **Jakshi Rayudu Pasupuleti:** Handled data preprocessing, clustering evaluation (elbow/silhouette), and generated all visualizations and plots.
- **Vamsi Reddy Konduru:** Developed baseline models, contributed to mathematical formulation sections, and assisted with failure-case analysis and LaTeX formatting.

## References


[1] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," *Machine Learning*, vol. 75, no. 2, pp. 245–248, May 2009.

[2] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2007, pp. 1027–1035.

[3] J. Peng and Y. Wei, "Approximating k-means-type clustering via semidefinite programming," *SIAM Journal on Optimization*, vol. 18, no. 1, pp. 186–205, 2007.

[4] A. Nellore and R. Ward, "Recovery guarantees for exemplar-based clustering," arXiv preprint arXiv:1309.3256, 2013.

[5] P. Awasthi, S. Bandyapadhyay, M. Charikar, R. Krishnaswamy, and A. K. Sinop, "Relax, no need to round: Integrality of clustering formulations," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, 2015.

[6] Y. Bilu and N. Linial, "Are stable instances easy?" in *Proceedings of the 1st Symposium on Innovations in Computer Science (ICS)*, 2010.

[7] K. Jain, M. Mahdian, and A. Saberi, "A new greedy approach for facility location problems," in *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, 2002, pp. 731–740.

[8] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," arXiv preprint arXiv:1011.3027v7, 2011.

[9] B. Ames and S. A. Vavasis, "Guaranteed clustering and biclustering via semidefinite programming," *Mathematical Programming*, vol. 122, no. 1, pp. 69–95, 2012.