Fairness in Clustering: A Study in Replication

Victor Huang, Sophie Boileau, Avery Hall, Jeremiah Mensah, Armira Nance, Muno Siyakurima, Brie Sloves

Advised by Professor Layla Oesper, Department of Computer Science, Carleton College

1. Context

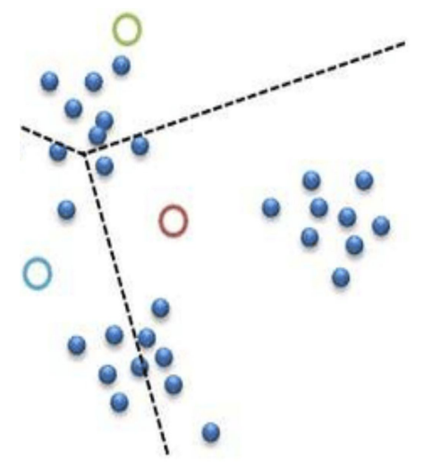
Clustering is a fundamental concept in machine learning and data analysis. It partitions n data points into k clusters, with the primary aim of uncovering inherent patterns or structures within a dataset. For our group's project, we want to explore how different clustering methods perform given a standardized definition of fairness.

2. Fairness

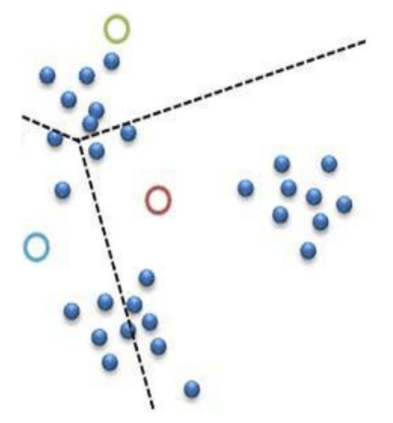
In this project, we adopt the fairness definition from the paper 'Fair Clustering Through Fairlets'. Fair clusters are those that maintain the same attribute ratio as the original dataset. For instance if gender is the attribute, a fair cluster would retain the same male to non-male ratio found in the original dataset.

3. *k* Means Clustering

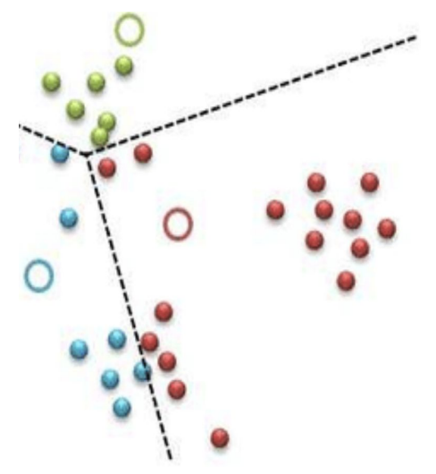
1. Begin by choosing *k* random data points to be centroids



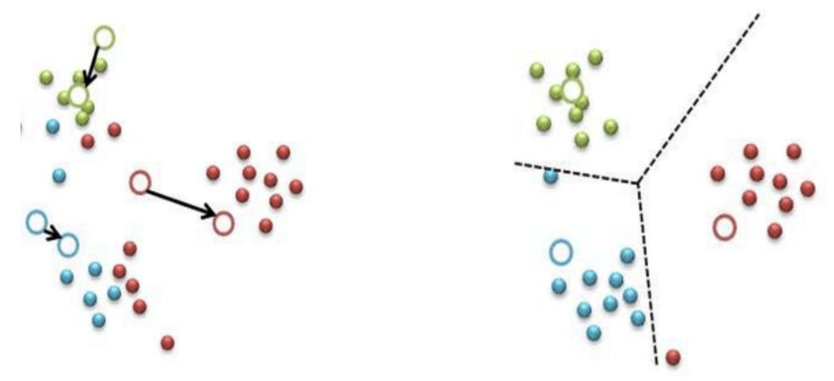
2. For each of the *n* data points, calculate the Euclidean distance between that point and each of the *k* centroids (in this illustration k = 3)



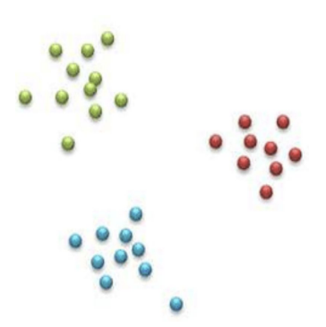
3. Assign the data point to the cluster with the nearest centroid



4. Then, for each cluster, find the mean of all data points in that cluster and update the location of the centroid.



5. Repeat the data point assignment and centroid update step until convergence is reached (i.e. the clusters are unchanging).



4. K Means++

Q: What is *k*-means++?

A: *k*-means++ represents an enhanced version of the basic *k*-means clustering algorithm

Q: How to run *k*-means++

A: (Steps shown as follows)
1. Start by selecting a random data point from
the dataset as the first centroid.

2. Compute the distance between each data point and the nearest existing centroid.

3. Square the distances found from step 2

4. Select the next centroid based on the probabilities calculated in step 3. Data points with higher squared distances (farther from existing centroids) have a higher probability of being chosen as the next centroid

5. Repeat steps 2 to 4 iteratively until 'K' centroids have been chosen.

Q: Why use k-means++

A: As opposed to the regular k-means algorithm that randomly selected *k* data points as initial centroids, *k*-means++ initializes centroids systematically and intentionally. This resolves instability issues, lowers computational costs, and results in even distributions and higher quality clustering

5. Fairlets

Q: What are Fairlets?

A: Fairlets are minimal sets that preserve the balance of protected attributes from the overall dataset. Any dataset can be decomposed into fairlets, then clustered according to traditional algorithms.

Q: What differentiates Fairlets?

A: Fairlets includes protected classes (race, gender, etc.) where each protected class must have approximately equal representation in every cluster.

Q: How are protected classes "protected"

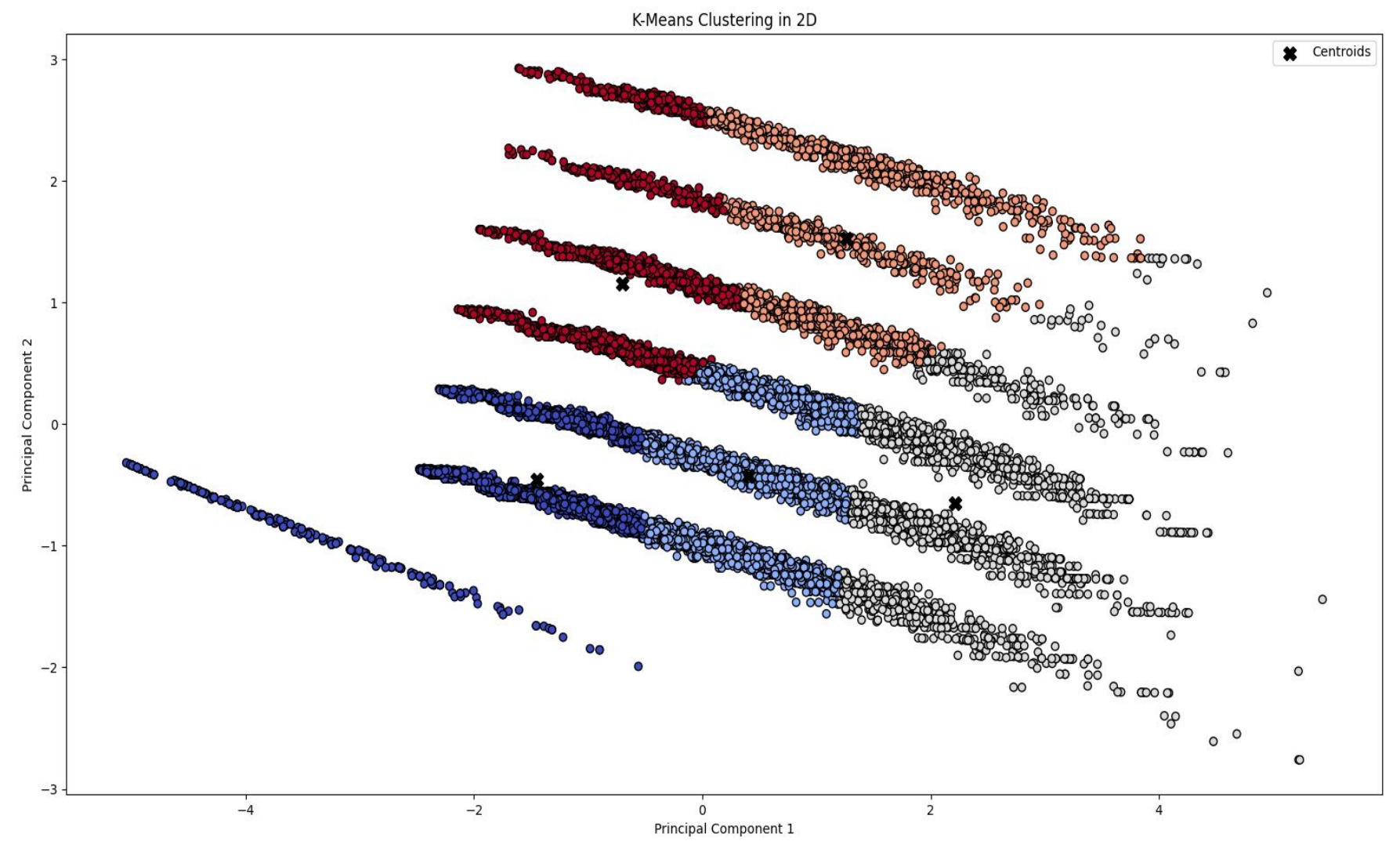
A: Let X be a set of points in a metric space, which we cluster into disjoint subsets C = {C1, ... , Ck}. The balance of cluster Y is:

$$\text{balance}(Y) = \min\left(\frac{\#\text{RED}(Y)}{\#\text{BLUE}(Y)}, \frac{\#\text{BLUE}(Y)}{\#\text{RED}(Y)}\right) \in [0,1]$$

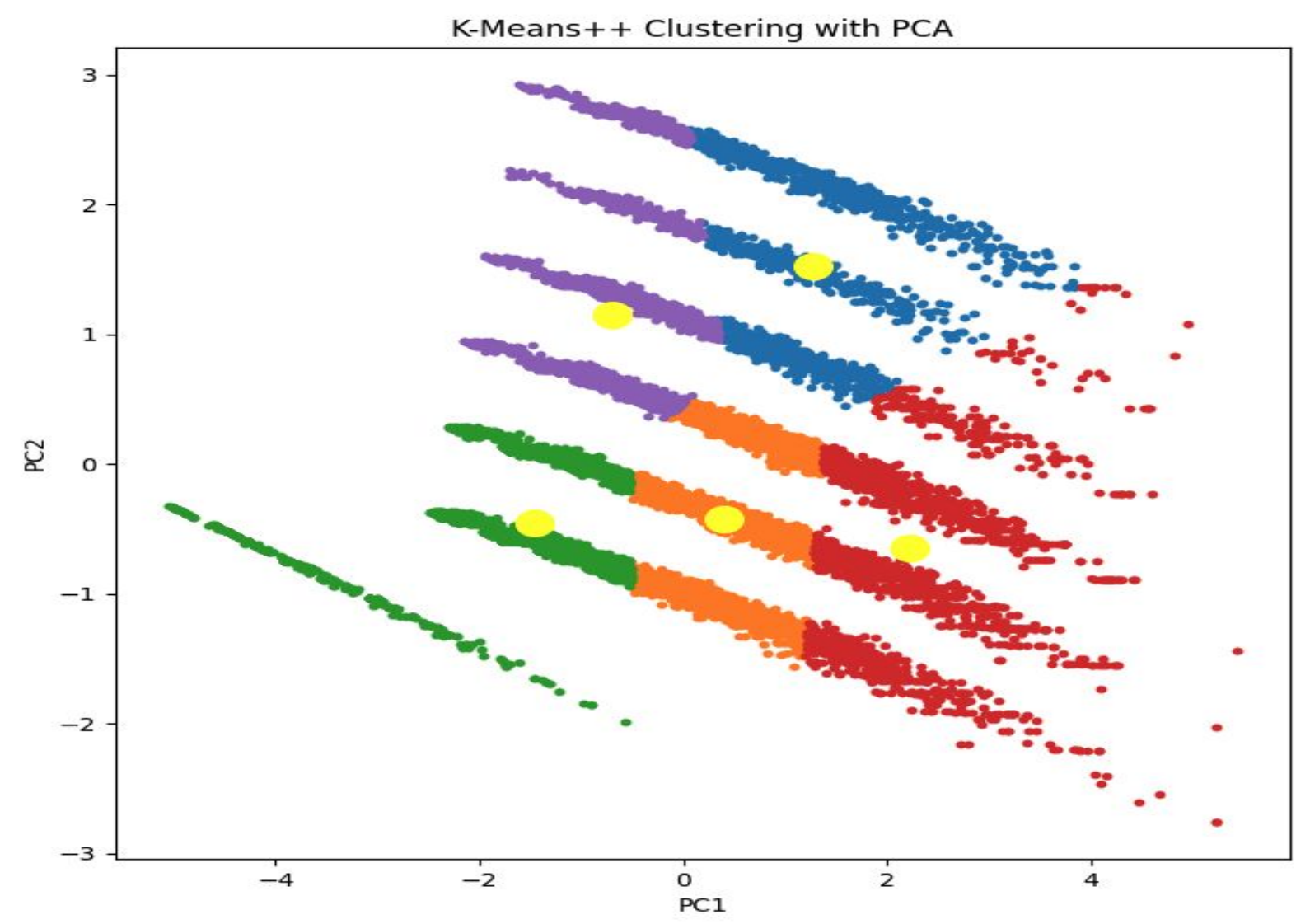
and the balance of a clustering C is

$$\text{balance}(\mathcal{C}) = \min_{C \in \mathcal{C}} \text{balance}(C)$$

6. Clustering Results



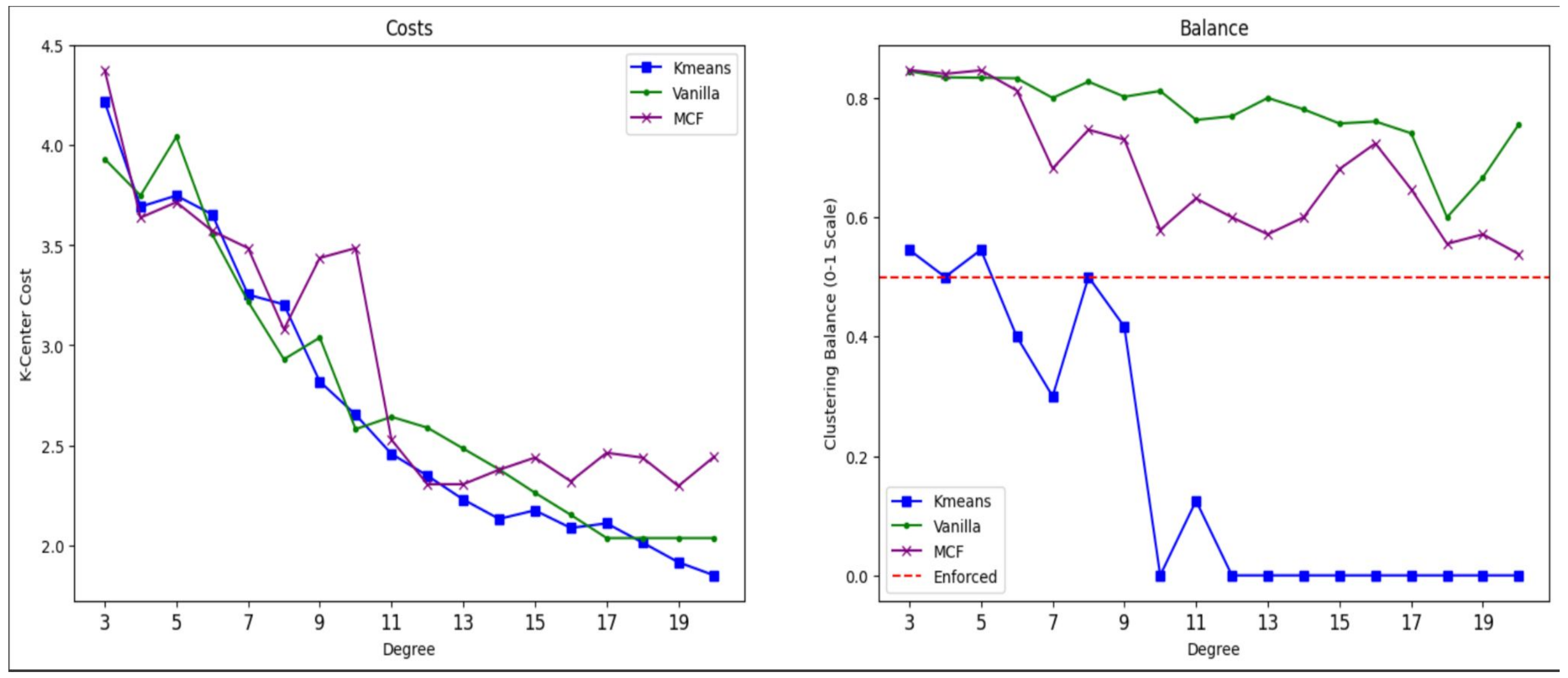
Basic K-means clustering Pictured here is a clustering of our four attributes into 5 clusters. Red, Orange, Light Blue, Dark Blue, and Grey. The clusters center around each of the black centroids. (Source : Muno Siyakurima)



K-means++ clustering Pictured here is a clustering of our four attributes into 5 clusters. Red, Green, Blue, Orange, and Purple. The clusters center around each of the black centroids. (Source: Jeremiah Mensah)

Goals:
- Analyze four key attributes that have high levels of intersectionality: Parent Education, Socioeconomic Status, Household Members per Income, and Hours of Extracurricular Activity
- Compare clustering methods: K-Means, K-Means++, Fairlets, and MCF Fairlets
- Examine how Race and Gender affect the balance value and k-center cost.

7. Comparing Method Results



Comparing regression of k-centers costs over different degrees of K with vanilla fairlets, minimum cost flow (MCF) fairlets, and basic k-means. On the right graph 'enforce' is a measure of fairness.

8. Conclusions

K-means vs K-means++
In our cluster analysis, we applied both K-Means++ and Basic K-Means to group data based on the attributes outlined in the Clustering Results section. Both K-means and K-means++ resulted in similar initial centroid placements, but the final clustering outcomes differed. Notably, K-means++ yielded centroids that were more widely distributed across the data space, aligning with the advantageous properties associated with K-means++. However when running each algorithm multiple times (not pictured here) they both produced inconsistent clusterings which could be due to the random initialization step present in both algorithms.

Fairness Comparison
We further investigated the impact of fairness in clustering using K-Means, Vanilla Fairlets, K-Means++, and MCF Fairlets. On the left side of the line graphs, MCF Fairlets, K-Means, and Vanilla Fairlets exhibited similar k-center costs until degree 8 - 10 which indicates MCF and Vanilla Fairlets don't have much impact on k-center costs. On the right side of the graphs, we introduced Gender as a balancing attribute. Notably, MCF Fairlets and Vanilla Fairlets consistently outperformed K-Means and K-Means++ in terms of enforcing fairness. These results lead us to conclude that, based on our dataset and our defined fairness criteria, MCF and Vanilla Fairlets are more effective in clustering data fairly with respect to Gender.

9. Future Work

Implement Socially Fair k Clustering: A more "human-centric" algorithm that is at odds with algorithms that prioritize proportionality. Rather than minimizing the average clustering cost over an entire dataset, we minimize the average clustering cost across different demographic groups in the dataset.

Digging Deeper into Intersectionality: So far, we have assumed that each data point can only belong to one demographic group or protected attribute. There are plenty of human-centered applications where this is not the case:
- Gender is not binary
- Many people identify with more than one racial group
- Unique minority experiences (disabled, impoverished, etc.) can not be lumped together

Citations

Fair Clustering through Fairlets - ACM Digital Library, dl.acm.org/doi/pdf/10.5555/3295222.3295256. Accessed 3 Nov. 2023.

Martins, Meacutenica V., et al. "Early Prediction of Student's Performance in Higher Education: A Case Study." *SpringerLink*, Springer International Publishing, 1 Jan. 1970, link.springer.com/chapter/10.1007/978-3-030-72657-7_16.

Ferrare, Joseph. "Intergenerational Education Mobility Trends by Race and Gender in the United States." *openICPSR*, Inter-university Consortium for Political and Social Research (ICPSR), 27 Aug. 2019, www.openicpsr.org/openicpsr/project/111586/version/V1/view?path=%2Fopenicpsr%2F111586%2Ffcr%3Aversions%2FV1&type=project.