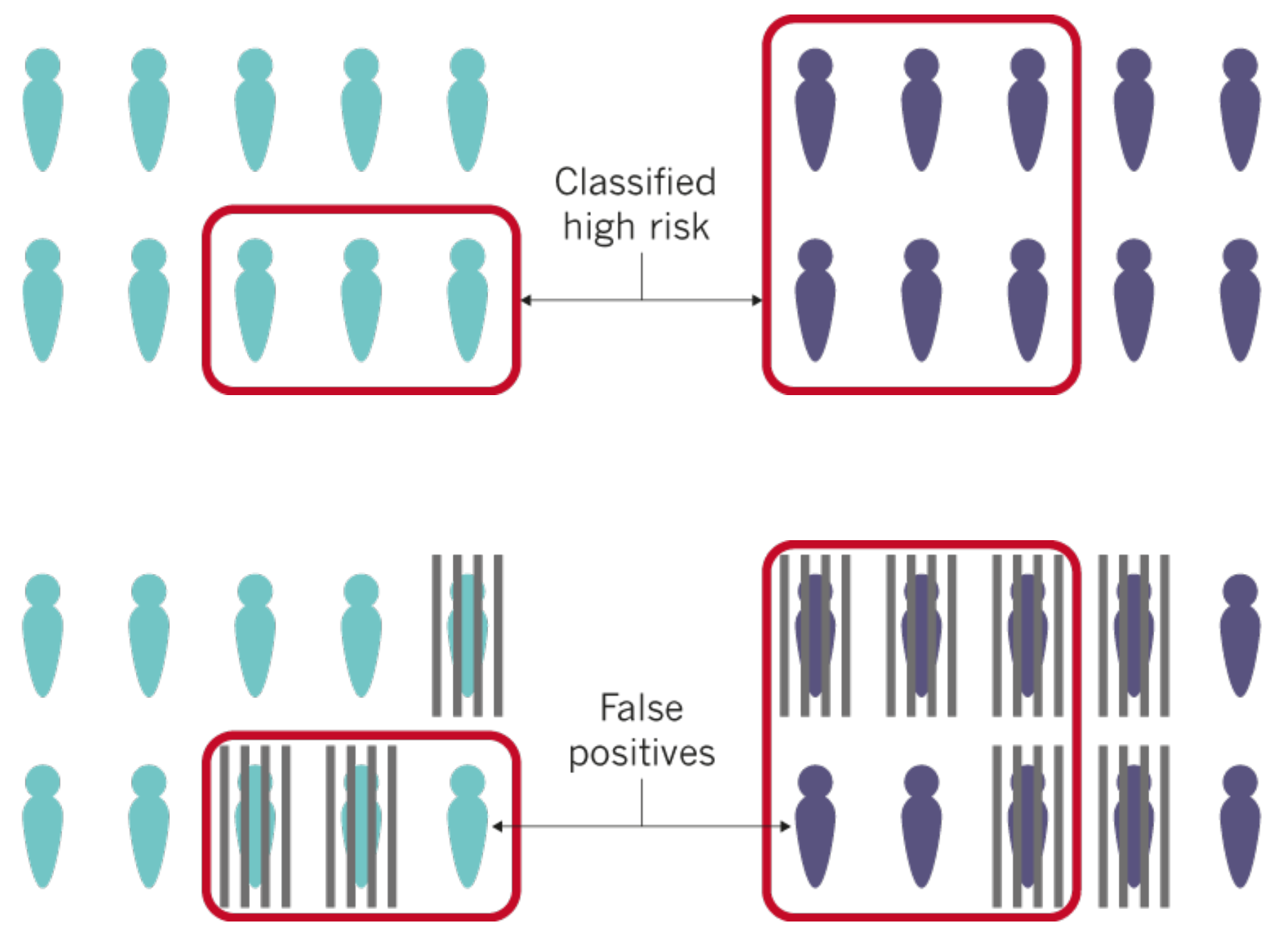
Fair Clustering

Student: Meenal Jhajharia | Mentor: Prof. Shobha Bagai

An augmented version of the K-clustering algorithm with a fairness parameter, such that each point has a "colour", and the representation of each colour in any cluster is constrained. It prevents any one colour from dominating a cluster. In the real world, colour can be any feature we want to protect from under or over representation, such as gender, race, party-affiliation, etc.



Source - Bias Detectives: The researchers striving to make algorithms fair [Nature]

0 Motivation

Understanding and reducing Algorithmic Bias, for example COMPAS software used in US Courts for parole decisions is racially biased towards Black People

1 Objective

Colour-blind clustering, which does not take a protected attribute into its decision making, may still result in very unfair clustering, which gives rise to the need for explicit fair clustering algorithms where the representation fraction is bound. This defines a fairness notion which puts a bound 1/alpha on the fraction of points of any colour in a given cluster.

- Develop a suitable algorithm for K-center clustering with a computational upper bound on cost of clustering, given a linear constraint to ensure fairness.
- Implement the algorithm on real world datasets and present empirical evaluations with existing algorithms.

3 Methodology

Implement a greedy clustering algorithm to get baseline bound λ, then solve a Linear Program Relaxation of the problem with an experimental solution λ/2 to obtain a fractional solution and a set of facilities F. *(with the fairness constraint intact)*

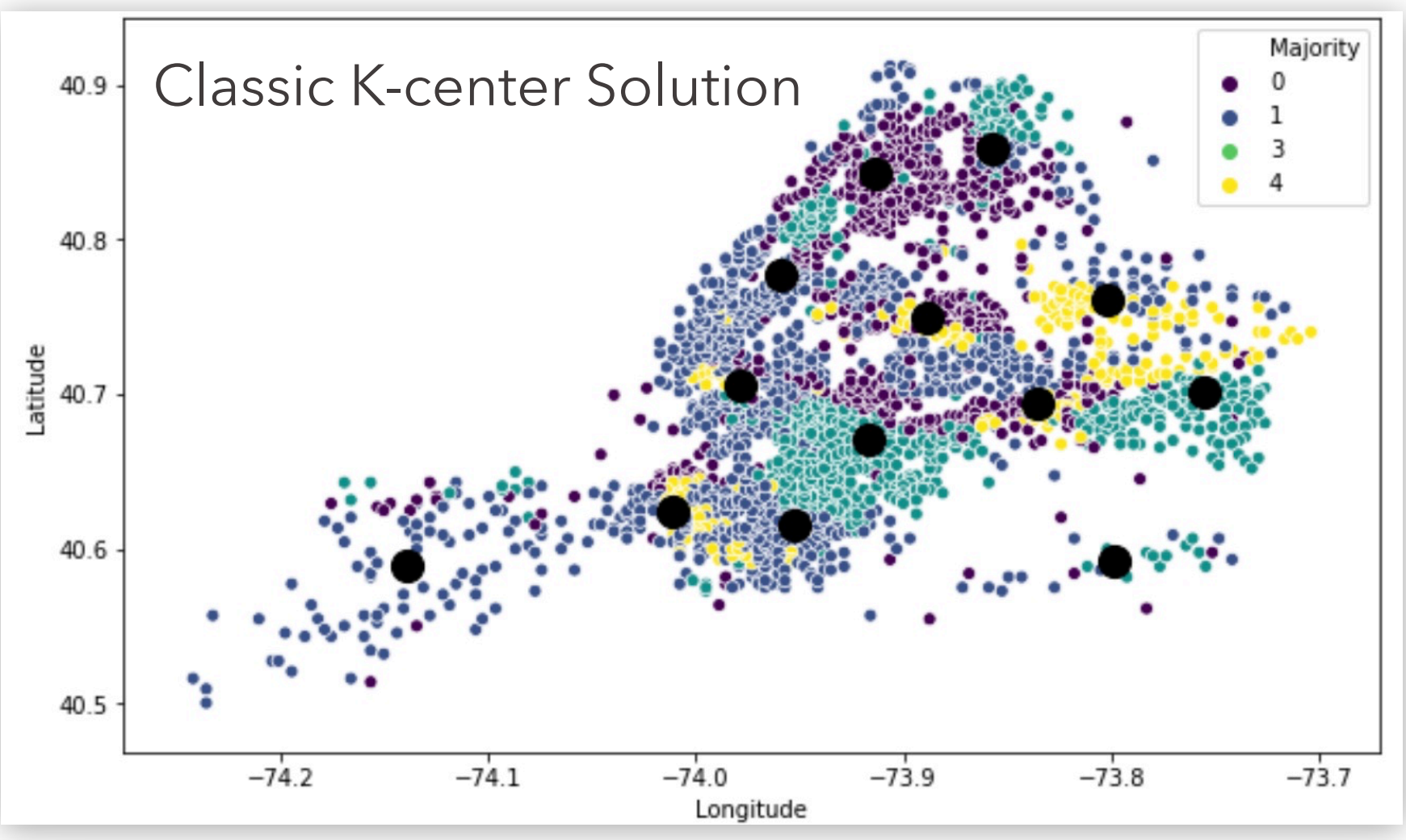
Obtain F′ from F by removing facilities that do not have a distance of at least 2λ`` from all other facilities in the subset. This set must be maximised, such that there must not be any facility in F not chosen to be in F′, but is at least 2λ″ from all other facilities in F′.

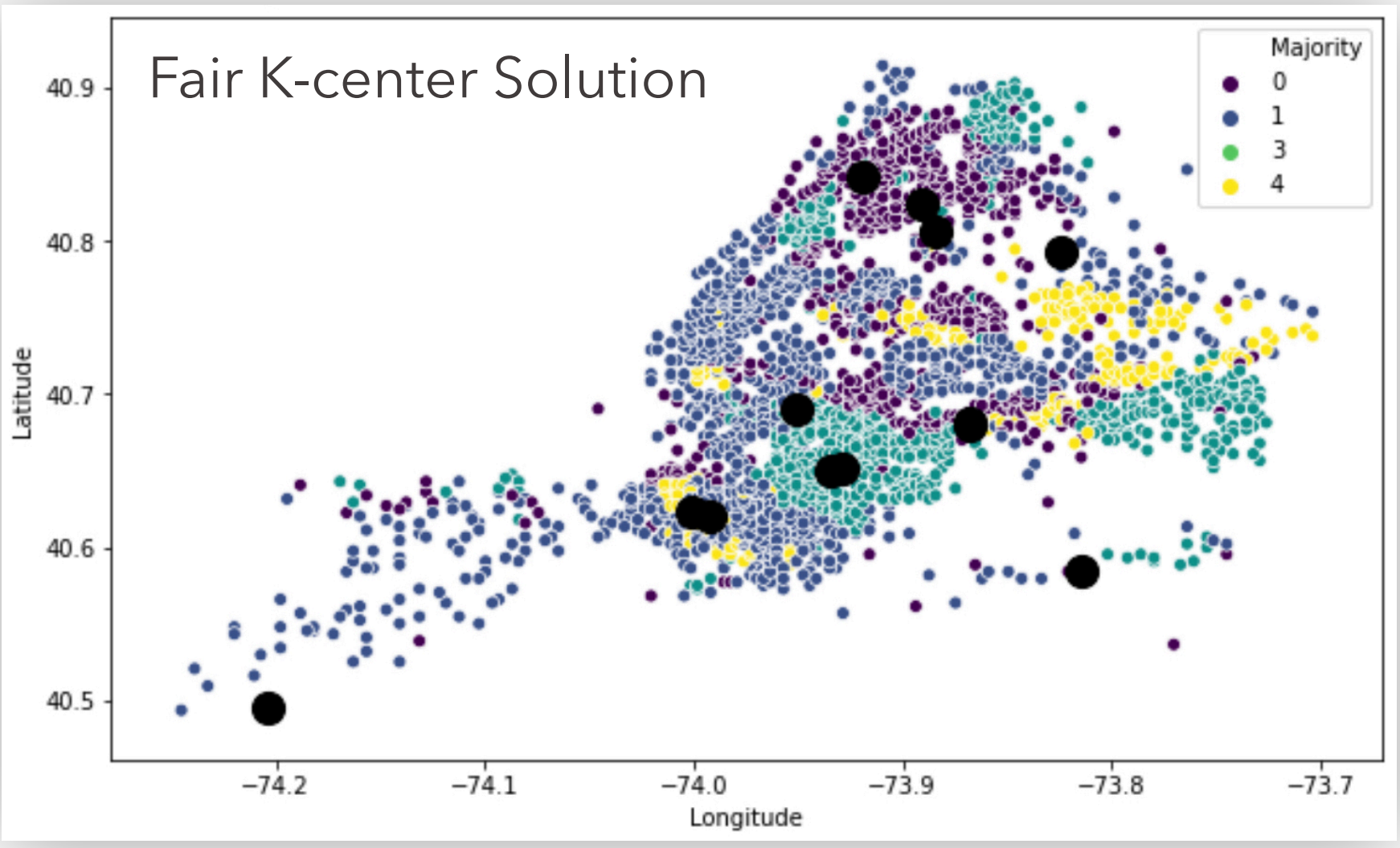
Choose the smallest possible λ`` such that the number of facilities is less than or equal to k. Since all clients are at most λ′ from their initially assigned facilities, and the new facility is at most 2λ″ away, the rerouted assignments must have cost within 2λ″ +λ′. *(Improved Upper Bound)*

Re-assign clients to facilities in F and Construct a maximum flow network with integer edge capacities to obtain an integer assignment for every client

4 Results

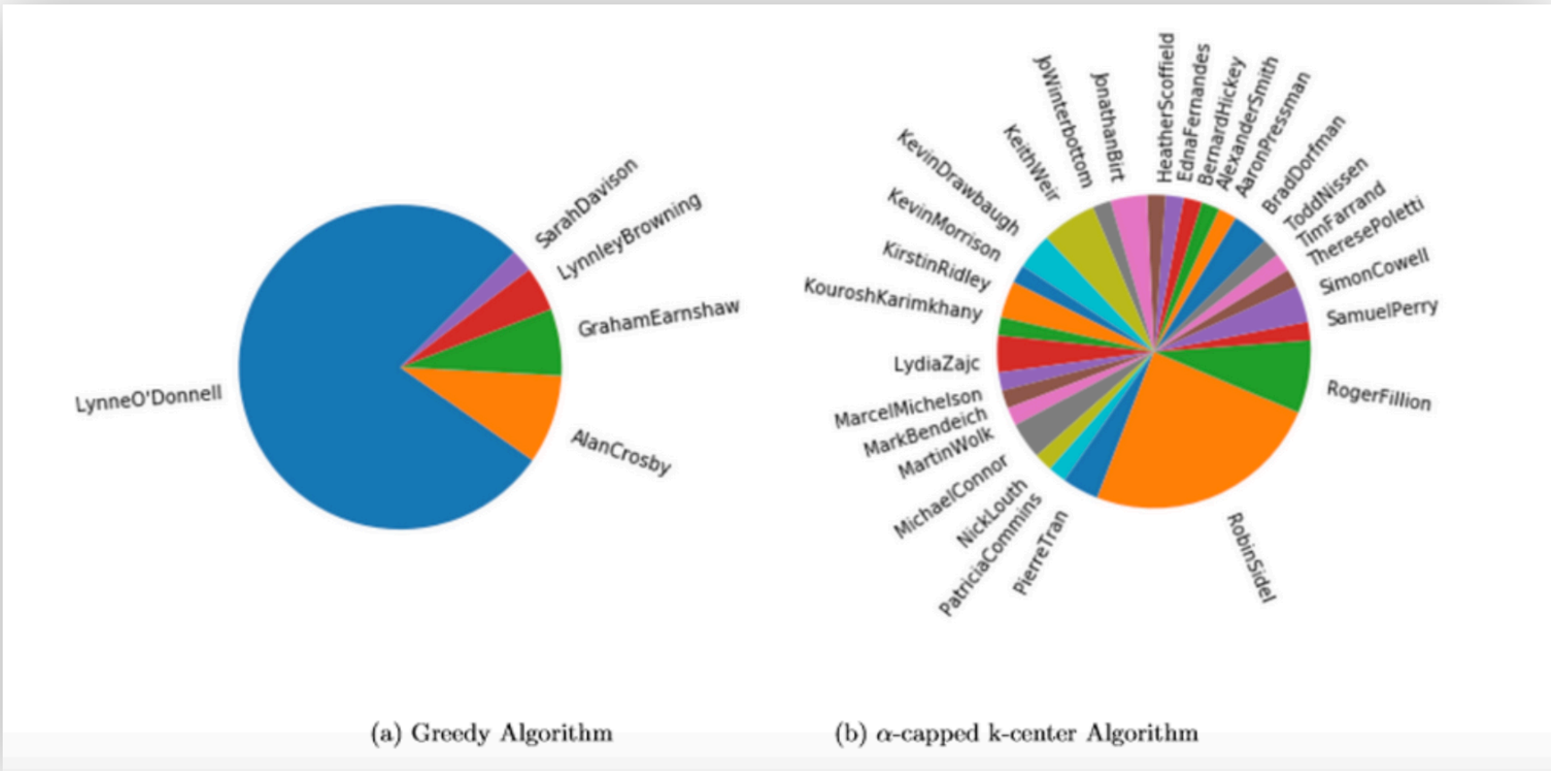
NYC Census Data consists of points as city blocks(each block is a point) with protected attribute(colour of the points) - race, we wish to cluster New York City such that each cluster contains no more than a third of any city block of a certain racial majority. Each number/colour pair correspond to a racial category:
0 -> Hispanic, 1 -> White,
3 -> Black, 4 -> Asian





α	3λ	Original Cost	$\lambda+2\lambda''$	Our Cost
0.5	0.27	0.205	0.127	0.120
0.33	0.27	0.222	0.134	0.122
0.25	0.27	0.202	0.12	0.1173

The Reuters dataset was created by transforming 2500 English language texts by 50 authors into 10-dimensional vectors using Gensim's Doc2Vec package. we initially set k = 25 and alpha = 0.5, to first examine the significant case where no cluster has one significantly dominating author. We found that lambda = 15 is the minimum value that produces a feasible solution (not exact).[1]



(a) Greedy Algorithm (b) α-capped k-center Algorithm

α	3λ	Original Cost	$\lambda+2\lambda''$	Our Cost
0.5	45	36.21	30.4	27.54
0.25	45	34.8	30.5	26.46
0.1	45	35.44	31	28.89

Lastly, we also evaluate our algorithm on Synthetic Data generated with Scikit-Learn and obtain similar results

5 Conclusion and Future Work

Clustering is used in a myriad of settings, from detecting fake news to creating targeted ads, this paper adds to the body of work exploring this domain. This study revisited the application of the k-center clustering problem and improved upon this model by generating a new, tighter upper bound specific to the solution returned by the LP relaxation. We tested our model on both synthetic datasets and real-world datasets, yielding extensive results that were compared based on objective function value and upper bound tightness. In future it would be worth exploring computationally efficient ways to solve the large LP problem so we can obtain an exact minimum value of lambda. With some modifications this algorithm can be extended to other versions of center based clustering problems like K-medoid and K-median

[1] The LP runs in time quadratic to the number of data points, which is too high to fine tune the value.