COMP 3801 - Final Report

Stable Approximations in Clustering

"Approximate Clustering without the Approximation"
by Balcan, Blum, and Gupta

Aashna Verma

Why ?

Protein by functions | Images by Subject | Articles by topic | Social Networks

01 **Clustering is NP hard in general**

02 **Approximation algorithms:** k-median, k-means, min-sum

03 **Cost ≠ Structure**

What if we assume the data is well-behaved?

Preliminaries

$$\mathcal{M} = (X, d)$$

Metric space M, where X is all the possible points and d is the distance function satisfying triangle inequality

$$\Phi(\mathcal{C}) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)$$

Objective function for k-median clustering. k-median wants clusters that minimize travel distance.

What is Approximation-Stability ?

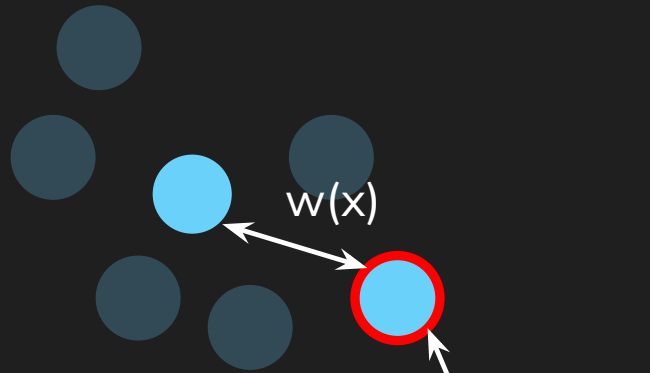
Definition: (c, ε)-approximation-stability

The assumption that in a stable dataset, any clustering close in cost is also close in structure.

- **c = cost closeness**
- **ε = structural closeness**
- **Stable datasets have “geometric separation”**

Approximately stable

Un-stable


w(x)
w2(x)

Using (c, ε)-approximation-stability

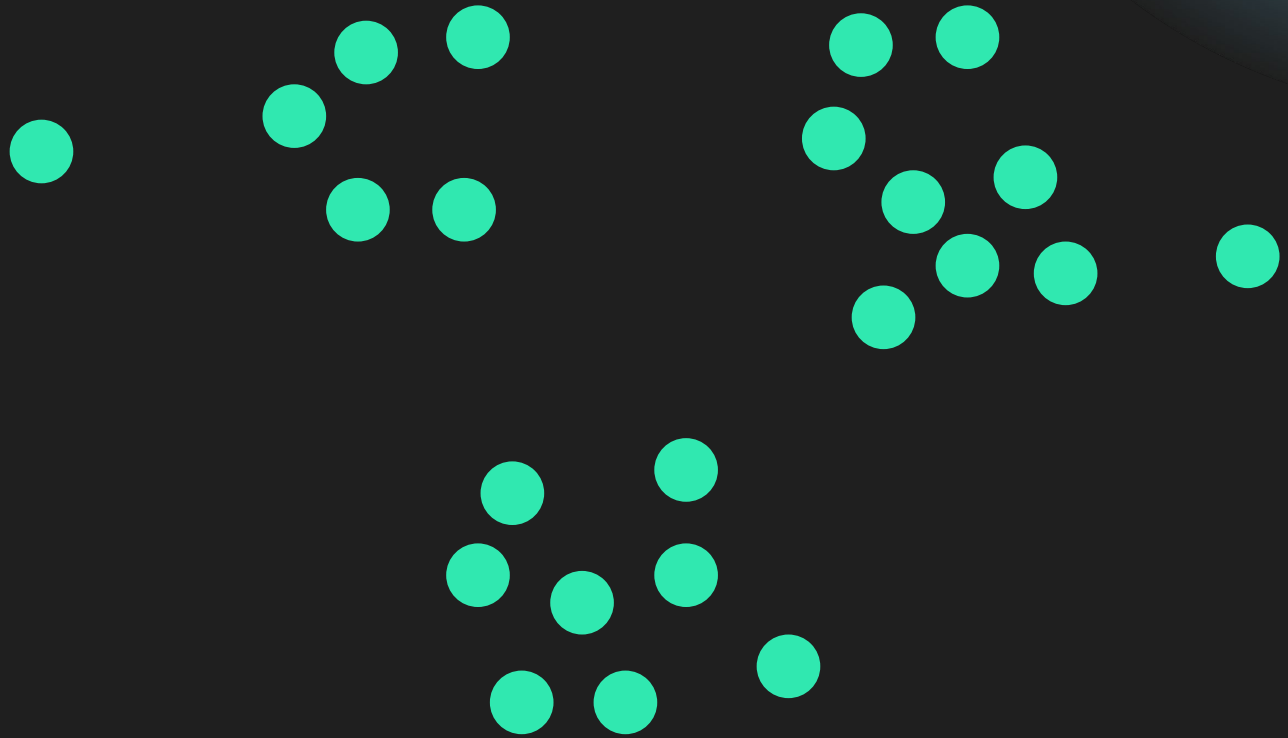
Where α ≤ 1, ε > 0

$$(1+\alpha, \epsilon) - approximation - stability$$

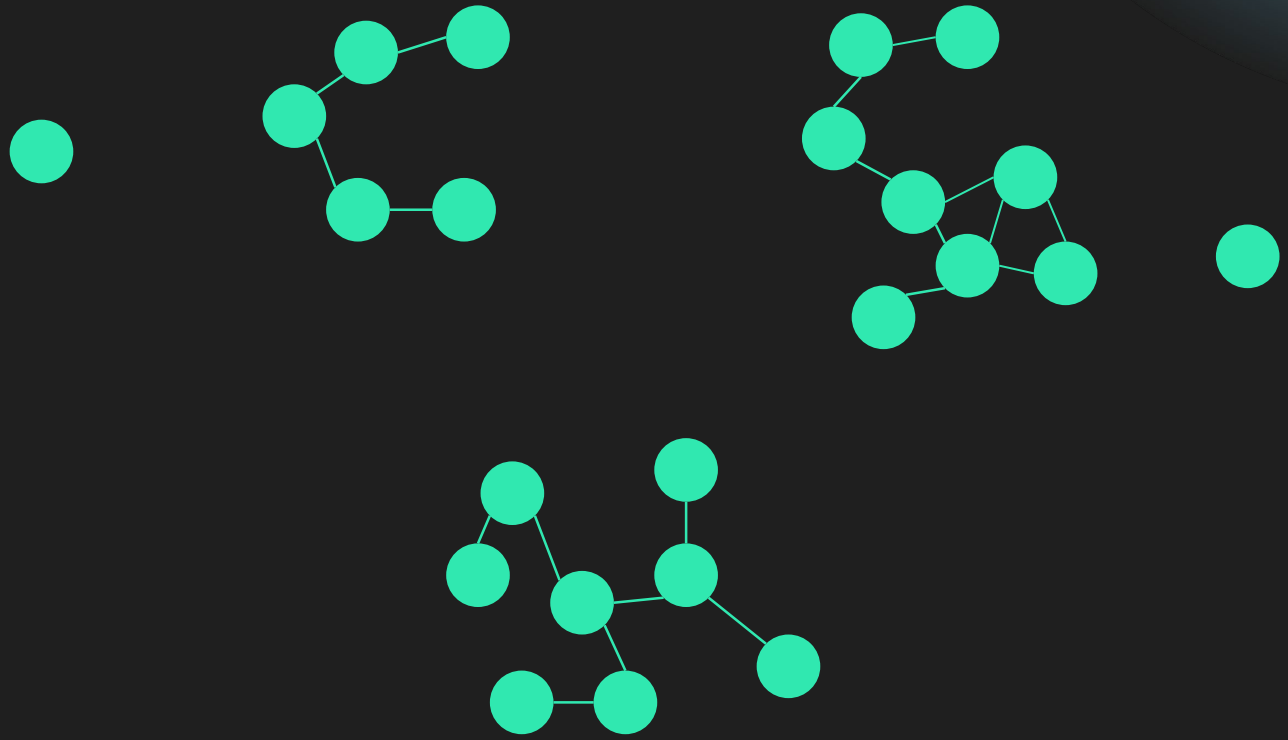
Most points have a large separation between these quantities

Algorithm overview

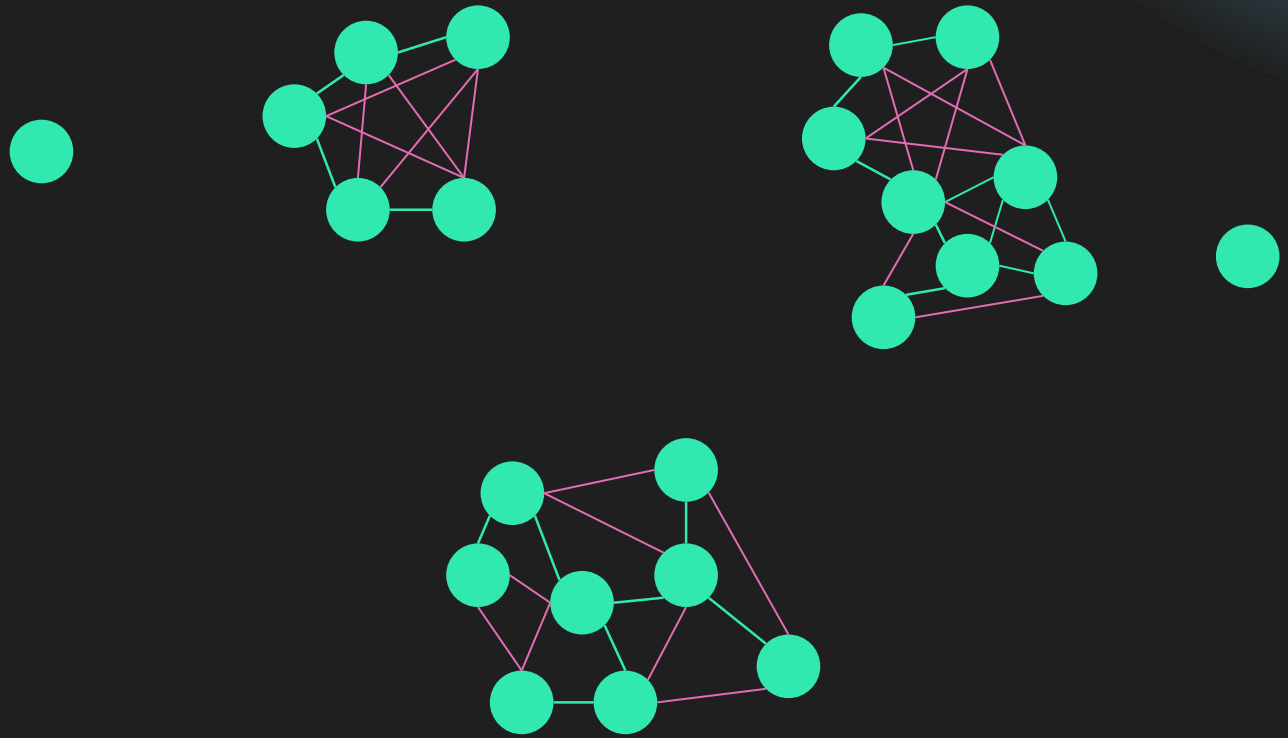
1. Build G: the threshold graph
2. Build H: connect x,y if they share ≥ b common neighbors
3. Take k largest components
4. Reassign points by median distance



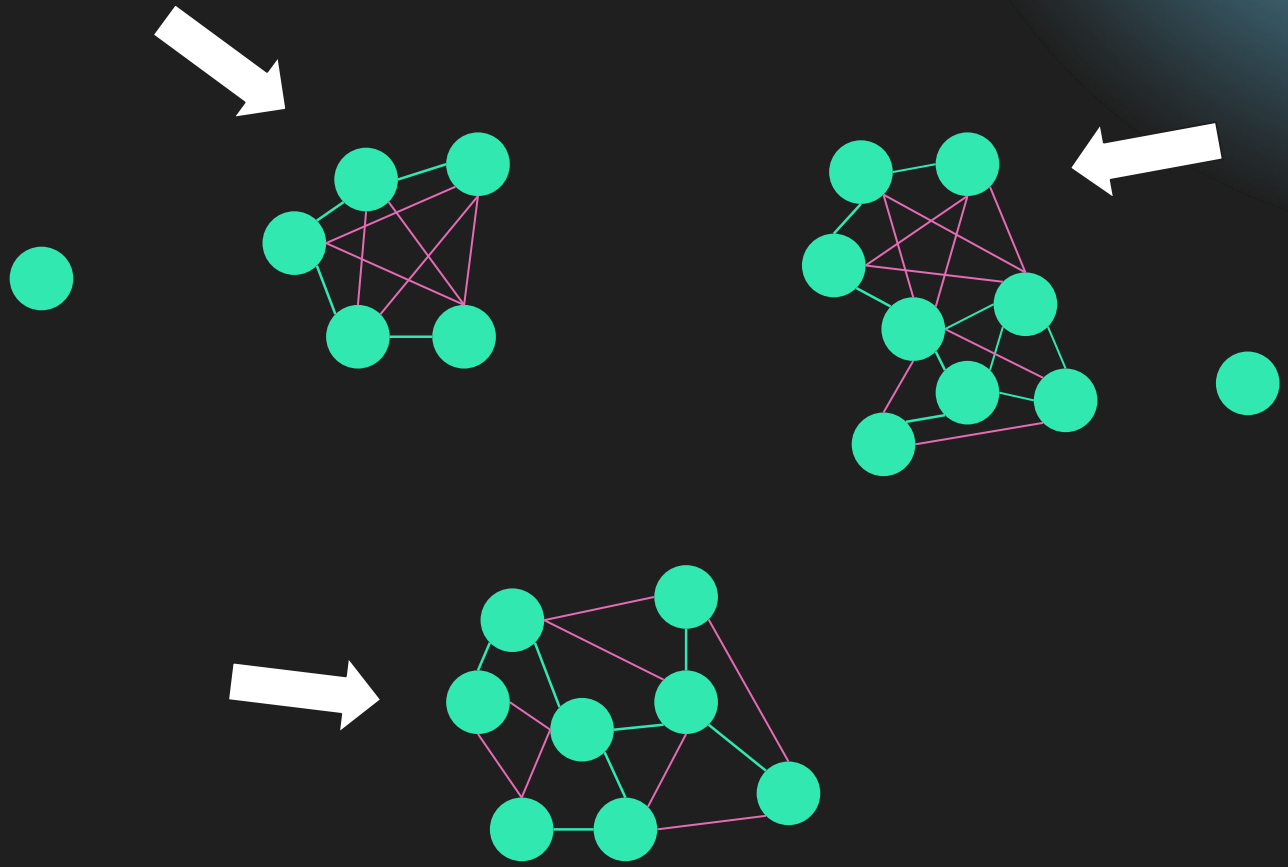
1.



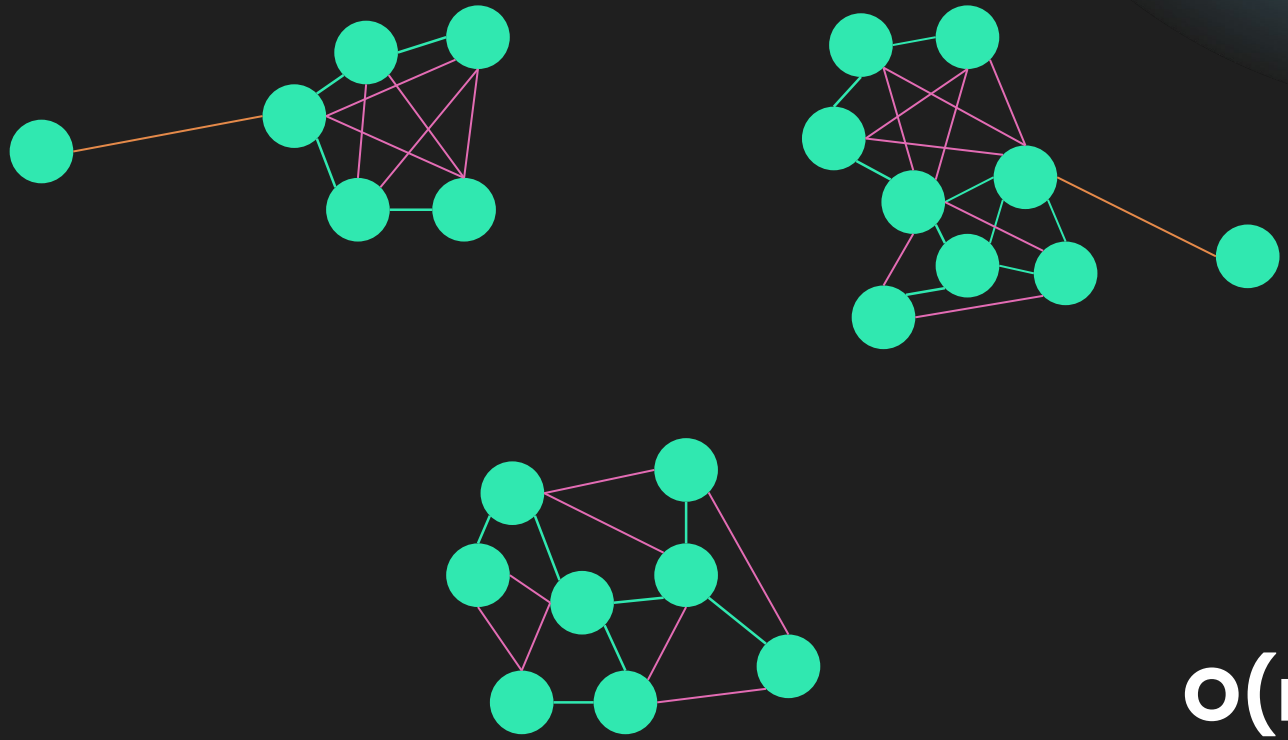
2.



3.



4.



$O(n^2)$

TLDR

1. Stability provides structure that algorithms can exploit
2. Hard clustering objectives become easy under mild assumptions
3. Threshold graphs capture cluster geometry better than optimization
4. Theoretical clustering can be solved without solving k-median

References

Balcan, M.-F., Blum, A., & Gupta, A. (2009). Approximate Clustering without the Approximation. Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA).