Approximate Correlation Clustering using Same-Cluster Queries

Ragesh Jaiswal

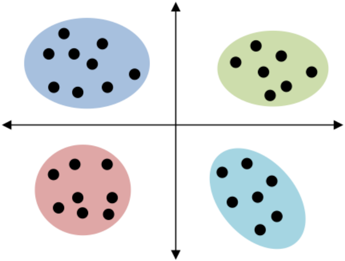
CSE, IIT Delhi

LATIN Talk, April 19, 2018

[Joint work with Nir Ailon (Technion) and Anup Bhattacharya (IITD)]

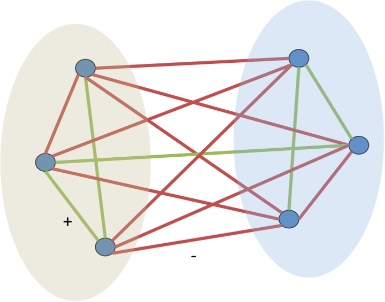
Clustering

- Clustering is the task of partitioning a given set of objects into *clusters* such that similar objects are in the same group (cluster) and dissimilar objects are in different groups.



Correlation Clustering

- Correlation clustering: Objects are represented as vertices in a complete graph with $\pm$ labeled edges. Edges labeled $+$ denote similarity and those labeled $-$ denote dissimilarity. The goal is to find a clustering of vertices that maximises agreements (MaxAgree) or minimise disagreements (MinDisAgree).



Correlation Clustering

MaxAgree

Given a complete graph with $\pm$ labeled edges, find a clustering of the vertices such that objective function Φ is maximized, where
$\Phi =$ sum of $+$ edges within clusters and $-$ edges across clusters.

MinDisAgree

Given a complete graph with $\pm$ labeled edges, find a clustering of the vertices such that objective function Ψ is minimised, where
$\Psi =$ sum of $-$ edges within clusters and $+$ edges across clusters.

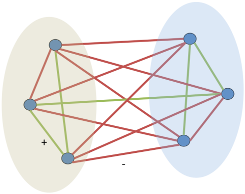


Figure: $\Phi = 12$ and $\Psi = 3$.

Correlation Clustering

MaxAgree

Given a complete graph with $\pm$ labeled edges, find a clustering of the vertices such that objective function Φ is maximized, where $\Phi =$ sum of $+$ edges within clusters and $-$ edges across clusters.

- NP-hard [BBC04]
- There is a PTAS for the problem [BBC04]

MinDisAgree

Given a complete graph with $\pm$ labeled edges, find a clustering of the vertices such that objective function Ψ is minimised, where $\Psi =$ sum of $-$ edges within clusters and $+$ edges across clusters.

- APX-hard [CGW05]
- Constant factor approximation algorithms [BBC04, CGW05]

Correlation Clustering

MaxAgree[k]

Given a complete graph with $\pm$ labeled edges and k, find a clustering of the vertices such that objective function Φ is maximized, where $\Phi=$ sum of $+$ edges within clusters and $-$ edges across clusters.

MinDisAgree[k]

Given a complete graph with $\pm$ labeled edges and k, find a clustering of the vertices such that objective function Ψ is minimised, where $\Psi=$ sum of $-$ edges within clusters and $+$ edges across clusters.

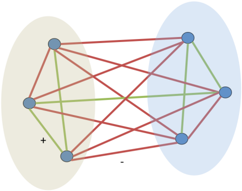


Figure: $\Phi = 12$ and $\Psi = 3$ for $k = 2$.

Correlation Clustering

MaxAgree[k]

Given a complete graph with $\pm$ labeled edges and k, find a clustering of the vertices such that objective function Φ is maximized, where $\Phi =$ sum of $+$ edges within clusters and $-$ edges across clusters.

- NP-hard for $k \geq 2$ [SST04].
- PTAS for any k (since there is a PTAS for MaxAgree).

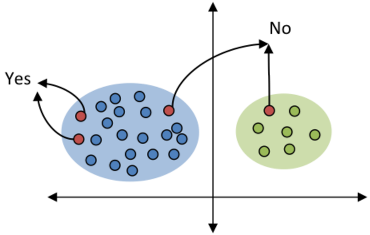
MinDisAgree[k]

Given a complete graph with $\pm$ labeled edges and k, find a clustering of the vertices such that objective function Ψ is minimised, where $\Psi =$ sum of $-$ edges within clusters and $+$ edges across clusters.

- NP-hard for $k \geq 2$ [SST04].
- PTAS for constant k with running time $n^{O(9^k/\varepsilon^2)} \log n$ [GG06].

- "*Beyond worst-case*"
    - Separating mixture of Gaussians.
    - Clustering under separation in the context of k-means clustering.
    - Clustering in semi-supervised setting where the clustering algorithm is allowed to make "*queries*" during its execution.

- "*Beyond worst-case*"
  - Mixture of Gaussians.
  - Clustering under separation.
  - Clustering in semi-supervised setting where the clustering algorithm is allowed to make "*queries*" during its execution.
    - Semi-Supervised Active Clustering (SSAC) [AKBD16]: In the context of the k-**means problem**, the clustering algorithm is given the dataset $X \subset \mathbb{R}^d$ and integer k (as in the classical setting) and it can make same-cluster queries.



- SSAC framework: Same-cluster queries for correlation clustering.

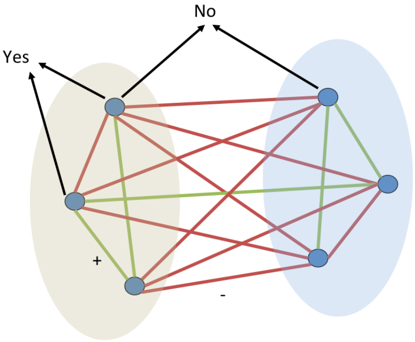


Figure: SSAC framework: same-cluster queries

- SSAC framework: Same-cluster queries for correlation clustering.

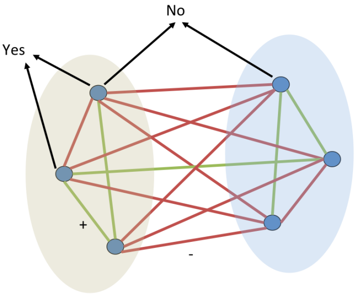


Figure: SSAC framework: same-cluster queries

- A limited number of such queries (or some weaker version) may be feasible in certain settings.
- So, understanding the power and limitations of this idea may open interesting future directions.

- Clearly, we can output optimal clustering using $O(n^2)$ same-cluster queries. Can we cluster using fewer queries?
- The following result is already known for the SSAC setting in the context of k-means problem.

Theorem (Informally stated theorem from [AKBD16])

There is a randomised algorithm that runs in time $O(kn \log n)$ and makes $O(k^2 \log k + k \log n)$ same-cluster queries and returns the optimal k-means clustering for any dataset $X \subseteq \mathbb{R}^d$ that satisfies some separation guarantee.

- The following result is already known for the SSAC setting in the context of k-means problem.

Theorem (Informally stated theorem from [AKBD16])

There is a randomised algorithm that runs in time $O(kn \log n)$ and makes $O(k^2 \log k + k \log n)$ same-cluster queries and returns the optimal k-means clustering for any dataset $X \subseteq \mathbb{R}^d$ that satisfies some separation guarantee.

- Ailon *et al.* [ABJK18] extend the above results to approximation setting while removing the separation condition with:
  - Running time: $O(nd \cdot poly(k/\varepsilon))$
  - # same-cluster queries: $poly(k/\varepsilon)$ (independent of n)
- Question: Can we obtain similar results for correlation clustering?

MinDisAgree[k] within SSAC

MinDisAgree[k]

Given a complete graph with $\pm$ labeled edges and k, find a clustering of the vertices such that objective function Ψ is minimised, where
$\Psi =$ sum of $-$ edges within clusters and $+$ edges across clusters.

- $(1+\varepsilon)$-approximate algorithm with running time $n^{O\left(\frac{9^k}{\varepsilon^2}\right)} \log n$ [GG06].

Theorem (Main result – upper bound)

There is a randomised query algorithm that runs in time $O(poly(\frac{k}{\varepsilon}) \cdot n \log n)$ and makes $O(poly(\frac{k}{\varepsilon}) \cdot \log n)$ same-cluster queries and outputs a $(1+\varepsilon)$-approximate solution for MinDisAgree[k].

MinDisAgree[k] within SSAC

- $(1+\varepsilon)$-approximate algorithm with running time $n^{O\left(\frac{9^k}{\varepsilon^2}\right)} \log n$ [GG06].

Theorem (Main result – upper bound)

There is a randomised query algorithm that runs in time $O(poly(\frac{k}{\varepsilon}) \cdot n \log n)$ and makes $O(poly(\frac{k}{\varepsilon}) \cdot \log n)$ same-cluster queries and outputs a $(1+\varepsilon)$-approximate solution for MinDisAgree[k].

Theorem (Main result - running time lower bound)

If the Exponential Time Hypothesis (ETH) holds, then there is a constant $\delta > 0$ such that any $(1+\delta)$-approximation algorithm for MinDisAgree[k] *runs in time $2^{\Omega(\frac{k}{poly \log k})}$-time.*

MinDisAgree[k] within SSAC

- $(1+\varepsilon)$-approximate algorithm with running time $n^{O\left(\frac{9^k}{\varepsilon^2}\right)} \log n$ [GG06].

Theorem (Main result – upper bound)

There is a randomised query algorithm that runs in time $O(poly(\frac{k}{\varepsilon}) \cdot n \log n)$ and makes $O(poly(\frac{k}{\varepsilon}) \cdot \log n)$ same-cluster queries and outputs a $(1+\varepsilon)$-approximate solution for MinDisAgree[k].

Theorem (Main result - running time lower bound)

If the Exponential Time Hypothesis (ETH) holds, then there is a constant $\delta > 0$ such that any $(1+\delta)$-approximation algorithm for MinDisAgree[k] *runs in time $2^{\Omega(\frac{k}{poly \log k})}$-time.*

Theorem (Main result - query lower bound)

If the Exponential Time Hypothesis (ETH) holds, then there is a constant $\delta > 0$ such that any $(1+\delta)$-approximation algorithm for MinDisAgree[k] *within the SSAC framework that runs in polynomial time makes $\Omega(\frac{k}{poly \log k})$ same-cluster queries.*

MinDisAgree[k] within SSAC

Theorem (Main result - running time lower bound)

If the Exponential Time Hypothesis (ETH) holds, then there is a constant $\delta > 0$ such that any $(1+\delta)$-approximation algorithm for MinDisAgree[k] *runs in time $2^{\Omega(\frac{k}{poly \log k})}$-time.*

Chain of reductions for lower bounds

- ETH $\xrightarrow{Dinur\ \ PCP}$ E3-SAT
- E3-SAT $\rightarrow$ NAE6-SAT
- NAE6-SAT $\rightarrow$ NAE3-SAT
- NAE3-SAT $\rightarrow$ Monotone NAE3-SAT
- Monotone NAE3-SAT $\rightarrow$ 2-colorability of 3-uniform bounded degree hypergraph.
- 2-colorability of 3-uniform bounded degree hypergraph $\xrightarrow{[CGW05]}$ MinDisAgree[k]

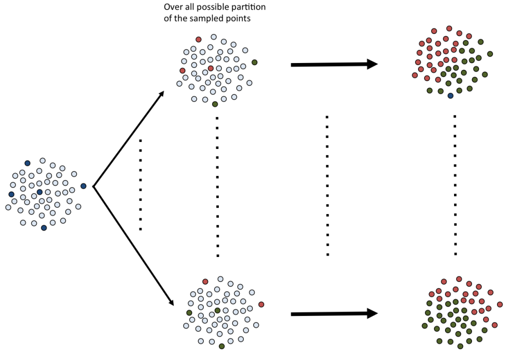
MinDisAgree[k] within SSAC

Theorem (Main result – upper bound)

There is a randomised query algorithm that runs in time $O(poly(\frac{k}{\varepsilon}) \cdot n \log n)$ and makes $O(poly(\frac{k}{\varepsilon}) \cdot \log n)$ same-cluster queries and outputs a $(1+\varepsilon)$-approximate solution for MinDisAgree[k].

Main ideas

- Through a simple observation about PTAS of Giotis and Guruswami[GG06].



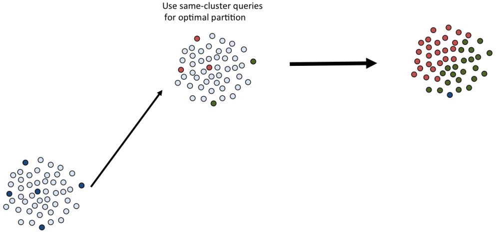
MinDisAgree[k] within SSAC

Theorem (Main result – upper bound)

There is a randomised *query* *algorithm that runs in time* $O(poly(\frac{k}{\varepsilon}) \cdot n \log n)$ *and makes* $O(poly(\frac{k}{\varepsilon}) \cdot \log n)$ *same-cluster queries and outputs a* $(1+\varepsilon)$*-approximate solution for* MinDisAgree[k].

Main ideas

- Through a simple observation about PTAS of Giotis and Guruswami[GG06].



Future Directions

- Future directions:
  - Gap in query upper and lower bounds.
  - Faulty-query setting.

References I

Nir Ailon, Anup Bhattacharya, Ragesh Jaiswal, and Amit Kumar, *Approximate clustering with same-cluster quesries*, Proceedings of the ninth Innovations in Theoretical Computer Science (ITCS'18), 2018.

Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David, *Clustering with same-cluster queries*, Advances in neural information processing systems, 2016, pp. 3216–3224.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla, *Correlation clustering*, Machine Learning **56** (2004), no. 1-3, 89–113.

Moses Charikar, Venkatesan Guruswami, and Anthony Wirth, *Clustering with qualitative information*, Journal of Computer and System Sciences **71** (2005), no. 3, 360–383.

Ioannis Giotis and Venkatesan Guruswami, *Correlation clustering with a fixed number of clusters*, Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm, Society for Industrial and Applied Mathematics, 2006, pp. 1167–1176.

Ron Shamir, Roded Sharan, and Dekel Tsur, *Cluster graph modification problems*, Discrete Applied Mathematics **144** (2004), no. 1, 173 – 182, Discrete Mathematics and Data Mining.

Thank you