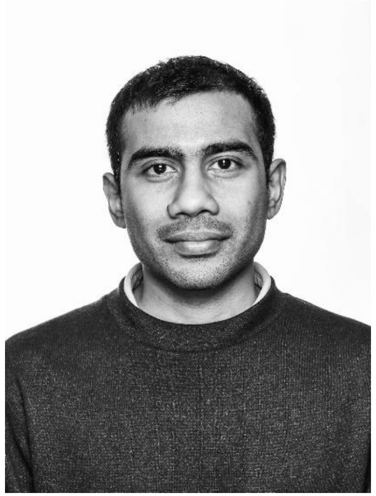
Towards Better-than-2 Approximation for Constrained Correlation Clustering

Andreas Kalavas
Max Planck Institute for Informatics, Germany
Archimedes/Athena RC, Greece

Evangelos Kipouridis
Max Planck Institute for Informatics, Germany

Nithin Varma
University of Cologne, Germany



Correlation Clustering

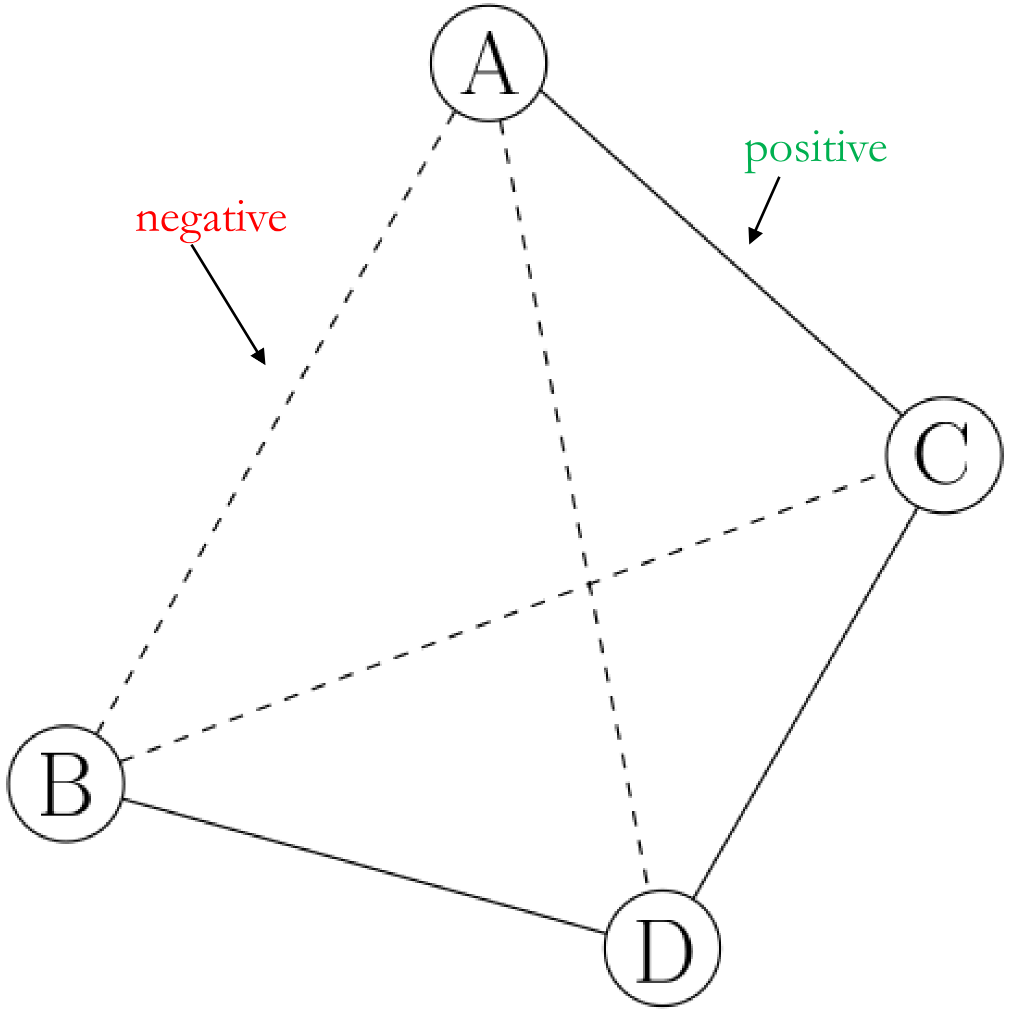
Input:

- Complete graph with preferences on each edge (positive or negative)

Correlation Clustering

Input:

- Complete graph with preferences on each edge (positive or negative)



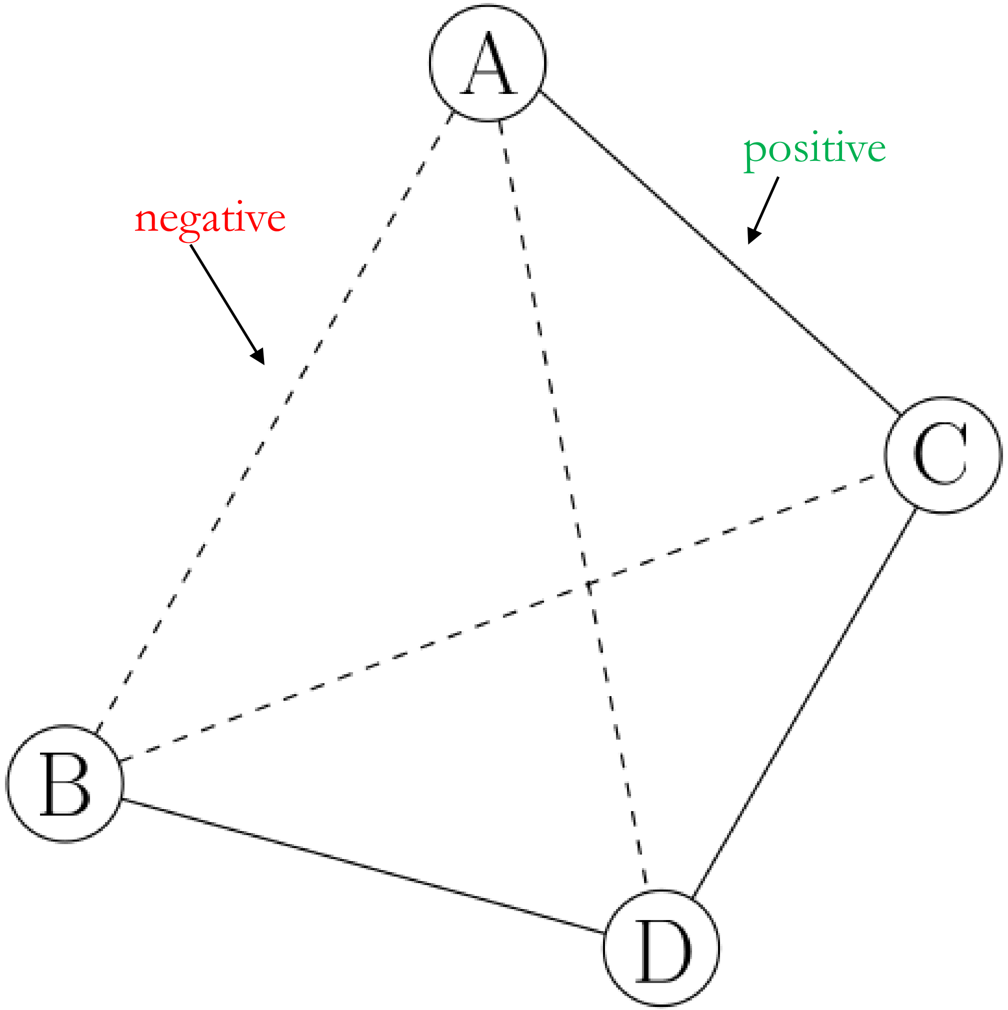
Correlation Clustering

Input:

- Complete graph with preferences on each edge (positive or negative)

Output:

- A **partition** of nodes violating the least number of pairwise preferences



Correlation Clustering

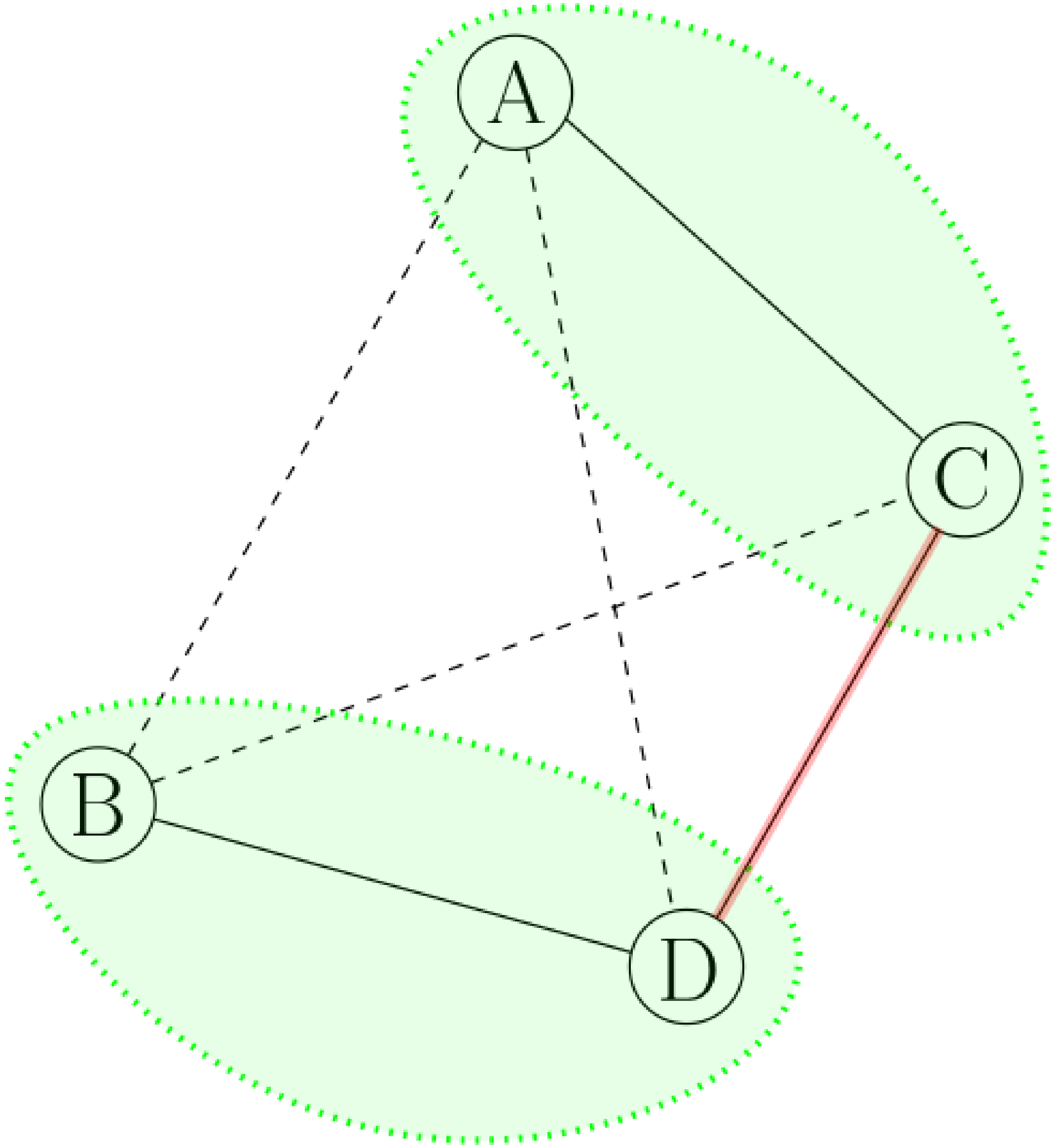
Input:

- Complete graph with preferences on each edge (positive or negative)

Output:

- A **partition** of nodes violating the least number of pairwise preferences

Example: Clustering violates only the preference of CD ⇒ Cost = 1



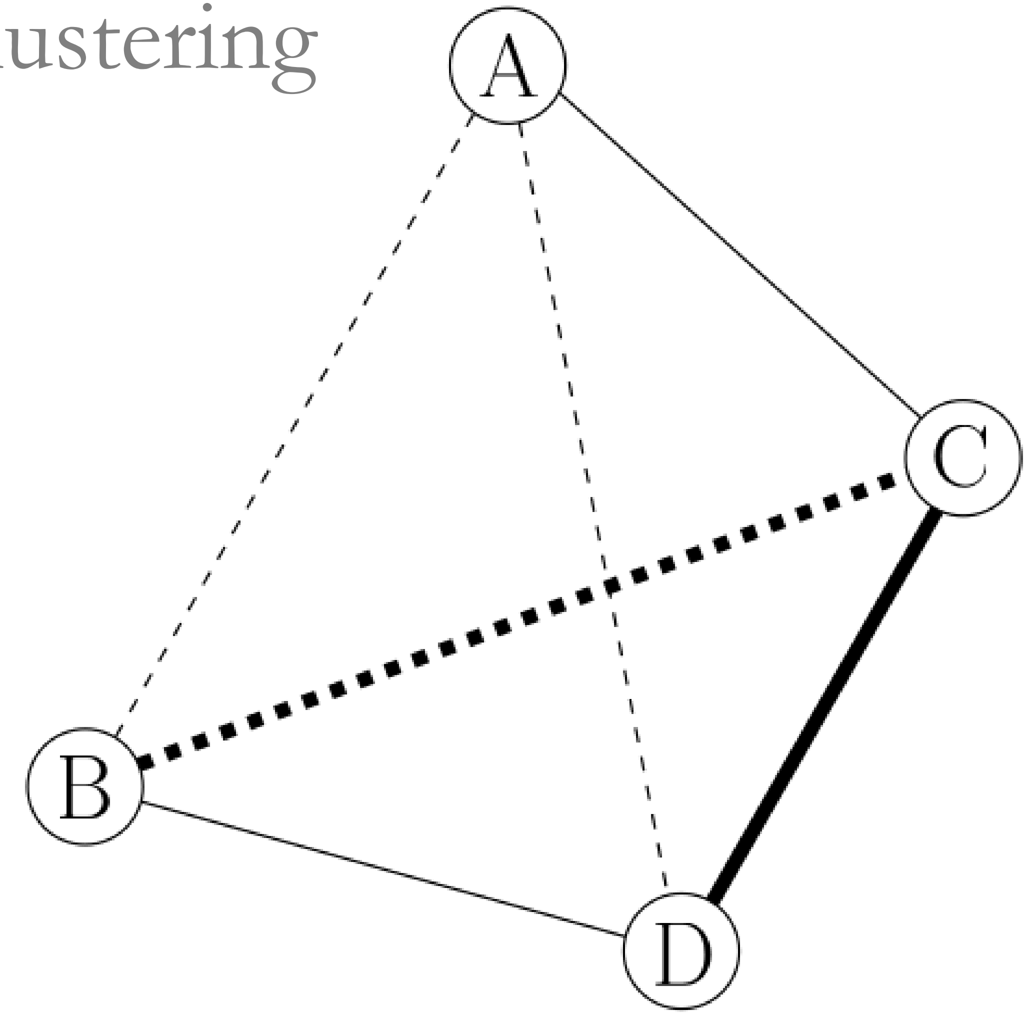
Constrained Correlation Clustering

Input:

- Complete graph with preferences on each edge (positive or negative)
- Some preferences are **hard constraints**

Output:

- A **partition** of nodes violating the least number of pairwise preferences and **satisfies** all **hard constraints**



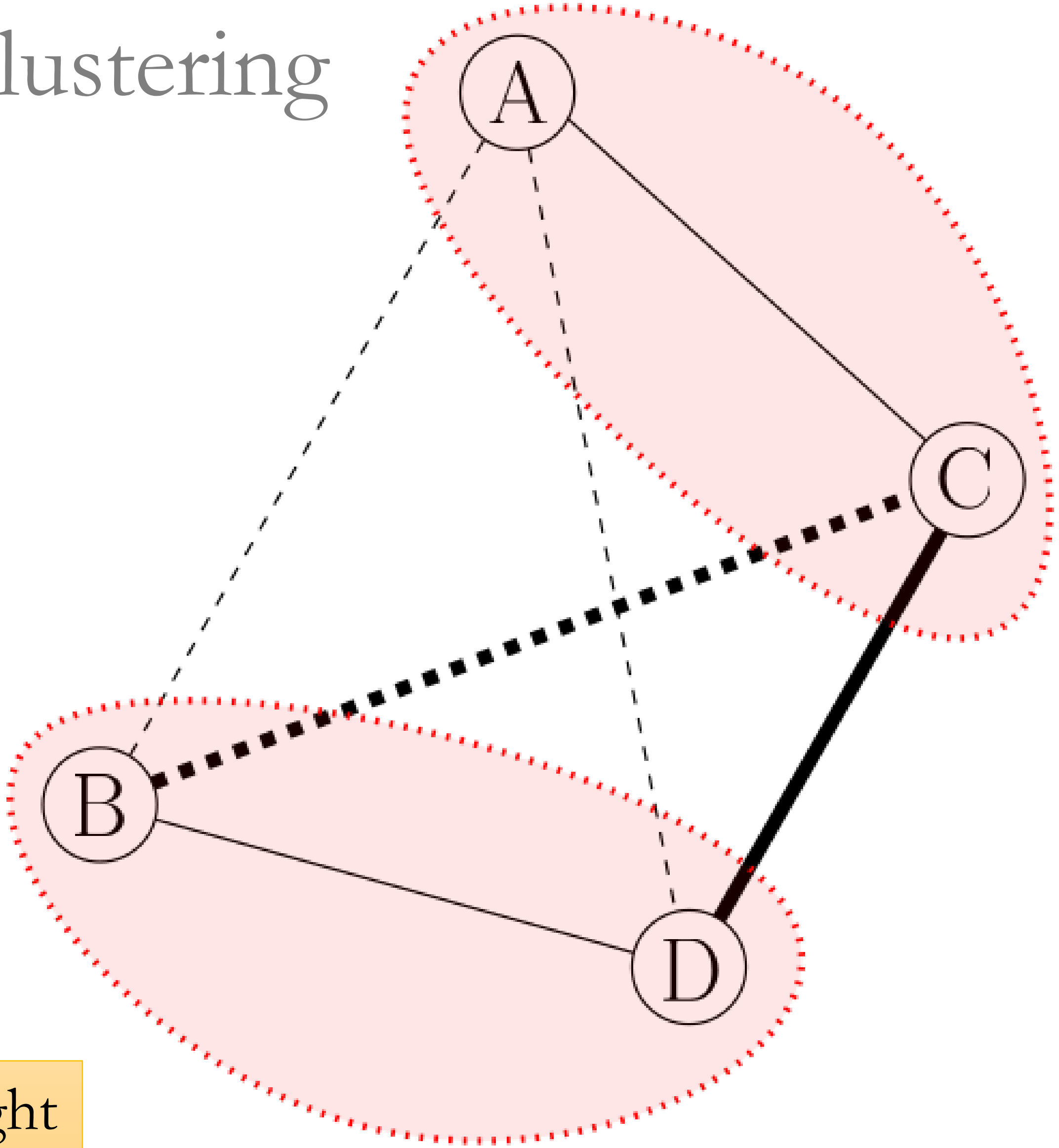
Constrained Correlation Clustering

Input:

- Complete graph with preferences on each edge (positive or negative)
- Some preferences are **hard constraints**

Output:

- A **partition** of nodes violating the least number of pairwise preferences and **satisfies** all **hard constraints**



Best clustering for **unconstrained** instance might **not** be **feasible**

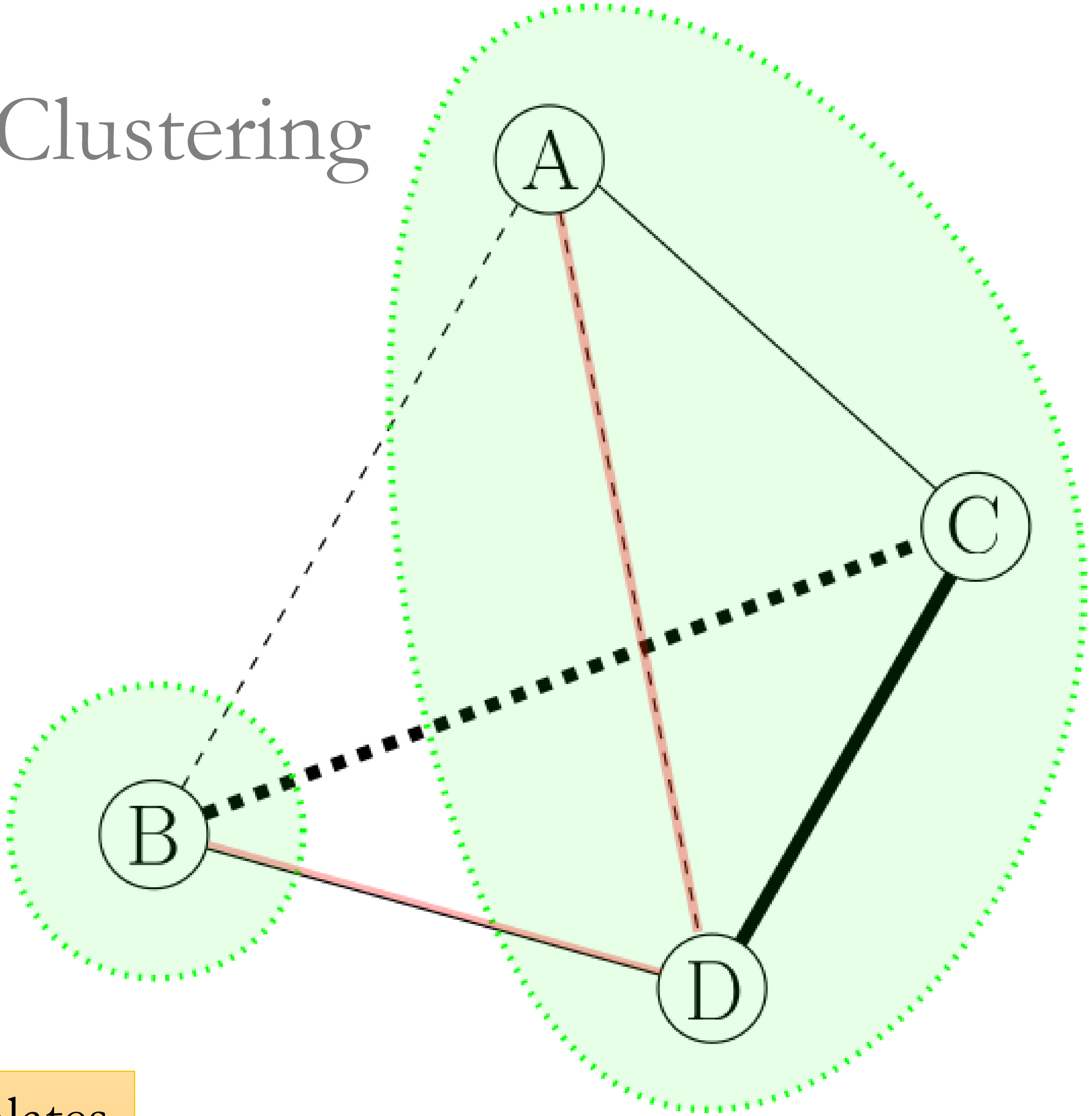
Constrained Correlation Clustering

Input:

- Complete graph with preferences on each edge (positive or negative)
- Some preferences are **hard constraints**

Output:

- A **partition** of nodes violating the least number of pairwise preferences and **satisfies** all **hard constraints**



Example: Feasible constrained clustering violates preferences of AD, BD ⇒ Cost = 2

Motivation

- No need to specify number of clusters in advance
  (in contrast e.g. with k-means)

Motivation

- No need to specify number of clusters in advance
  (in contrast e.g. with k-means)
- Incorporates expert knowledge

Motivation

- No need to specify number of clusters in advance (in contrast e.g. with k-means)
- Incorporates expert knowledge
- Semi-supervised learning

Motivation

- No need to specify number of clusters in advance
  (in contrast e.g. with k-means)
- Incorporates expert knowledge
- Semi-supervised learning

Example:

- Crosslingual Link Detection
  [Gael & Zhu, IJCAI '07]

Motivation

- No need to specify number of clusters in advance
  (in contrast e.g. with k-means)
- Incorporates expert knowledge
- Semi-supervised learning

Example:

- Crosslingual Link Detection
  [Gael & Zhu, IJCAI '07]
- Pool of news articles – in **multiple languages**

Motivation

- No need to specify number of clusters in advance (in contrast e.g. with k-means)
- Incorporates expert knowledge
- Semi-supervised learning

Example:

- Crosslingual Link Detection [Gael & Zhu, IJCAI '07]
- Pool of news articles – in **multiple languages**
- Identify articles that report on the same news event

Motivation

- No need to specify number of clusters in advance
  (in contrast e.g. with k-means)
- Incorporates expert knowledge
- Semi-supervised learning

Example:

- Crosslingual Link Detection
  [Gael & Zhu, IJCAI '07]
- Pool of news articles – in **multiple languages**
- Identify articles that report on the same news event
- **Expert** speaks one of these languages – very **reliable**!

Motivation

- No need to specify number of clusters in advance
  (in contrast e.g. with k-means)
- Incorporates expert knowledge
- Semi-supervised learning

Example:

- Crosslingual Link Detection
  [Gael & Zhu, IJCAI '07]
- Pool of news articles – in **multiple languages**
- Identify articles that report on the same news event
- **Expert** speaks one of these languages – very **reliable**!

History of the Problem

Correlation Clustering

NP-hard

- Introduced by Bansal, Blum and Chawla [Bansal et al., Machine Learning '04]

History of the Problem

Correlation Clustering

NP-hard

- Introduced by Bansal, Blum and Chawla [Bansal et al., Machine Learning '04]

⋮

- ~ 2-approximation (Solving the standard Linear Programming (LP) formulation) [Chawla et al., STOC '15]

⋮

- ~ 1.8-approximation (Local Search) [Cohen-Addad et al., STOC '24]
- ~ 1.4-approximation (Exponentially large LP) [Cao et al., STOC '24]
- Results for Streaming, MPC, Online, …

History of the Problem

Correlation Clustering

NP-hard

- Introduced by Bansal, Blum and Chawla [Bansal et al., Machine Learning '04]

⋮

- ~ 2-approximation (Solving the standard Linear Programming (LP) formulation) [Chawla et al., STOC '15]

⋮

- ~ 1.8-approximation (Local Search) [Cohen-Addad et al., STOC '24]
- ~ 1.4-approximation (Exponentially large LP) [Cao et al., STOC '24]
- Results for Streaming, MPC, Online, …

Constrained Correlation Clustering

- 3-approximation (LP) [van Zuylen & Williamson, SODA '07]
- 16-approximation (Faster combinatorial algorithm) [Fischer et al., STACS '25]

This Work

Novel way of **combining** large **LP** and **Local Search** techniques

This Work

Novel way of **combining** large **LP** and **Local Search** techniques

Conceptually simple better-than-2 approximation for Correlation Clustering

This Work

Novel way of **combining** large **LP** and **Local Search** techniques

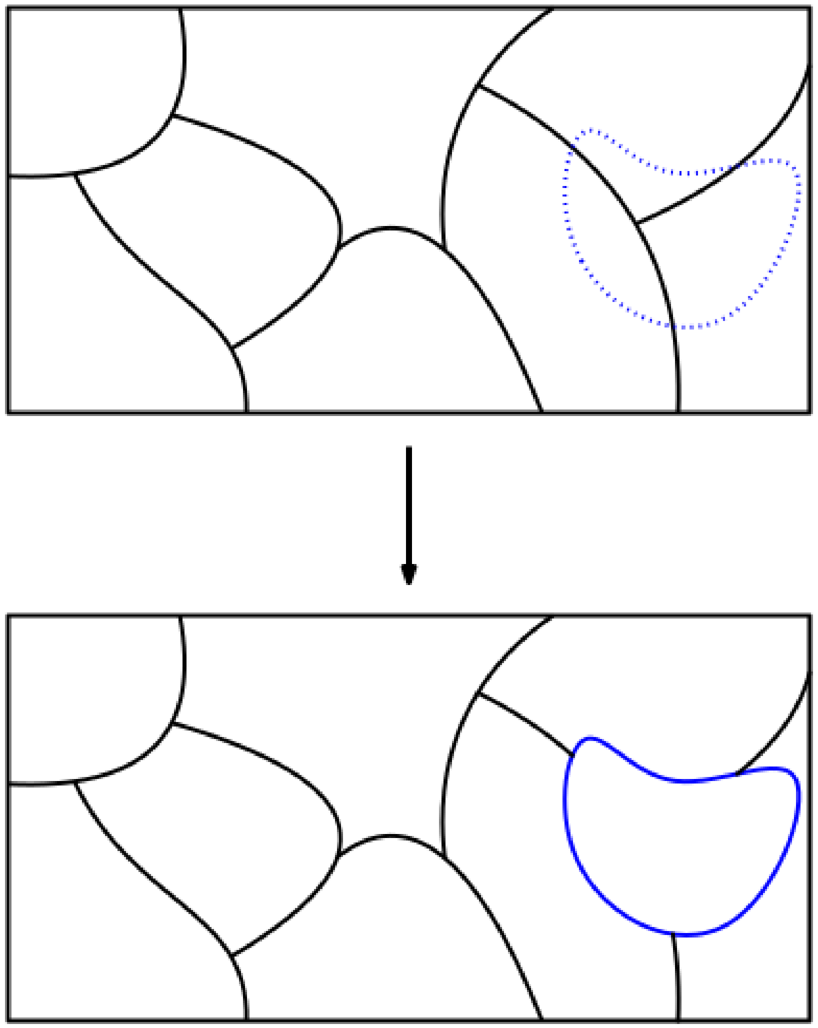
Conceptually simple better-than-2 approximation for Correlation Clustering

Better-than-2 approximation for Constrained Correlation Clustering, <u>conditional on solution to a large Constrained LP</u>

Main Idea

Local Search move: **swap** a **cluster** in the clustering

Local Search

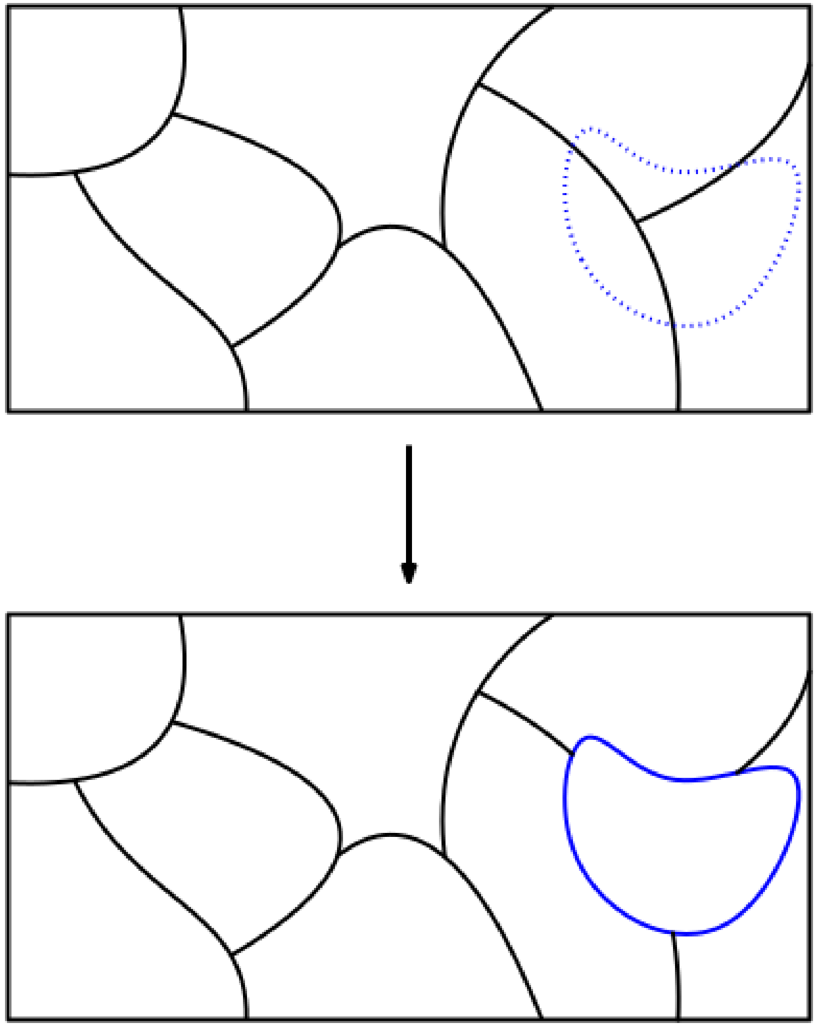


Main Idea

Local Search move: **swap** a **cluster** in the clustering

Problem: Exponentially many possible clusters

Local Search



Main Idea

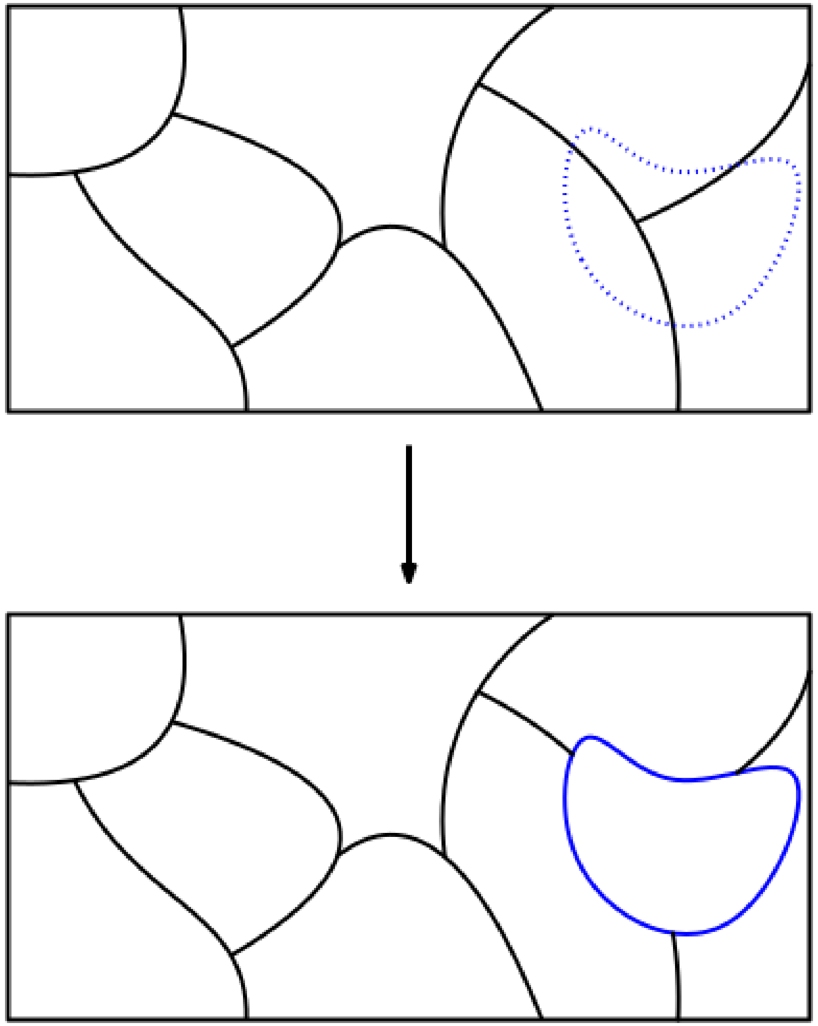
Problem: Exponentially many possible clusters

Local Search move: **swap** a **cluster** in the clustering

Solution of the "large" LP:

- ***poly***-sized family of clusters

Local Search



Main Idea

Problem: Exponentially many possible clusters

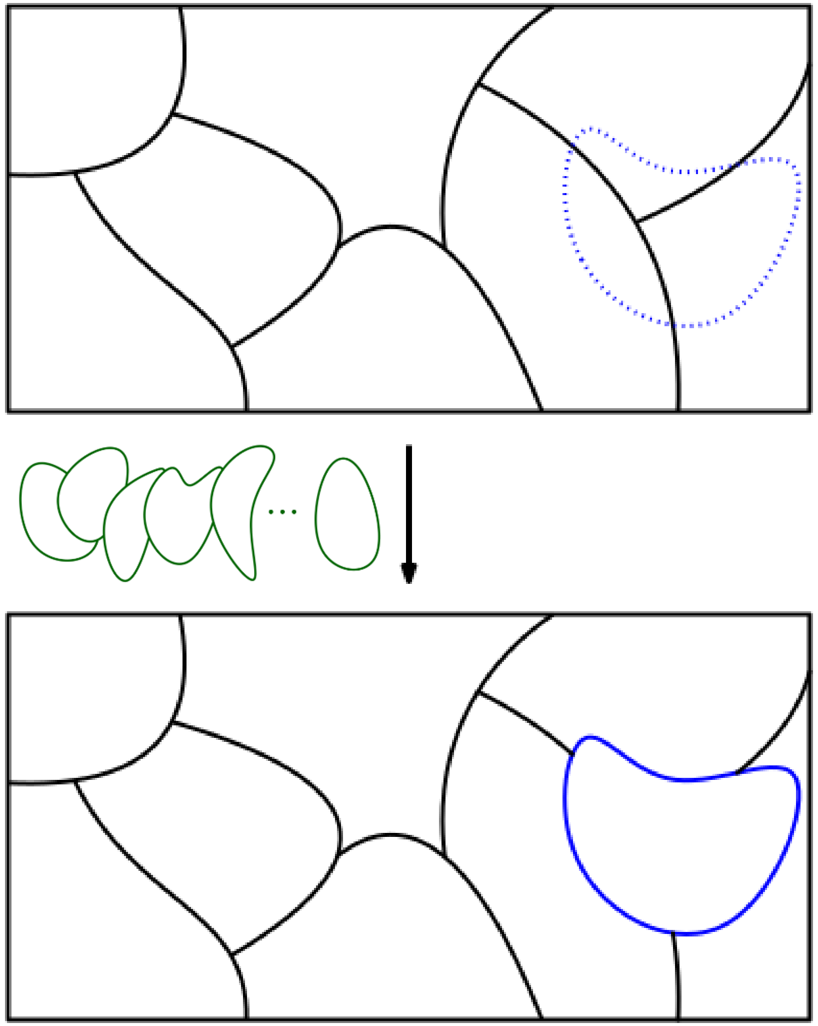
Local Search move: **swap** a **cluster** in the clustering

Solution of the "large" LP:

- ***poly***-sized family of clusters

Use it to **guide** Local Search

Local Search



Main Idea

Problem: Exponentially many possible clusters

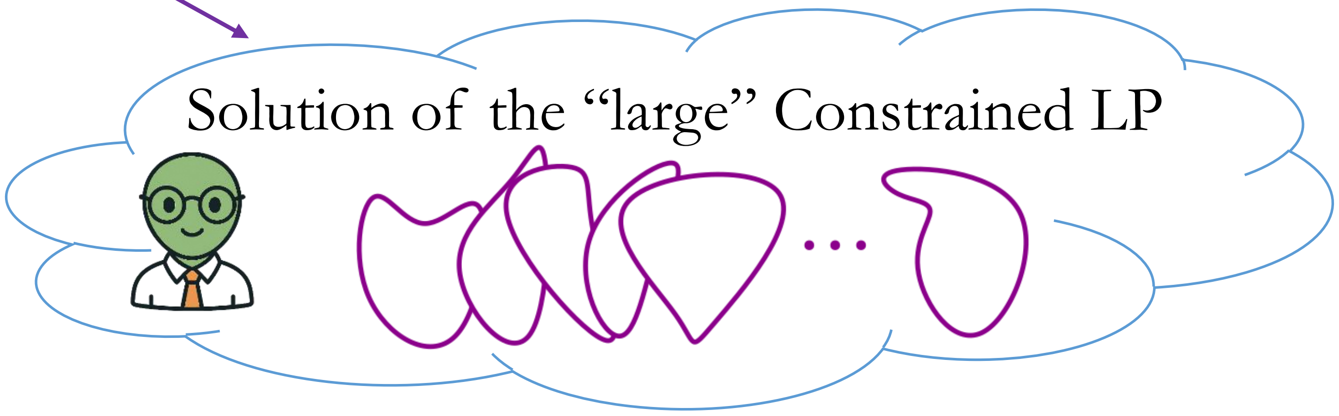
Local Search move: **swap** a **cluster** in the clustering

Solution of the "large" LP:

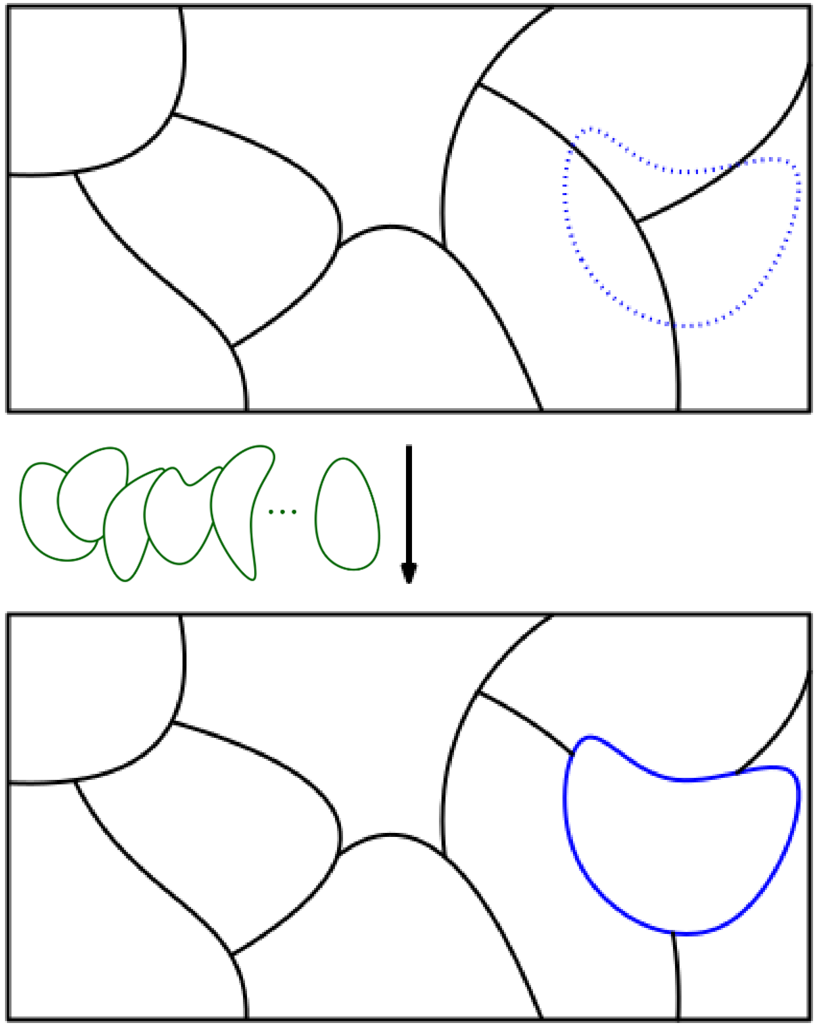
- ***poly***-sized family of clusters

Use it to **guide** Local Search

*Assume we have this



Local Search



Main Idea

Problem: Exponentially many possible clusters

Local Search move: **swap** a **cluster** in the clustering

Solution of the "large" LP:
- ***poly***-sized family of clusters

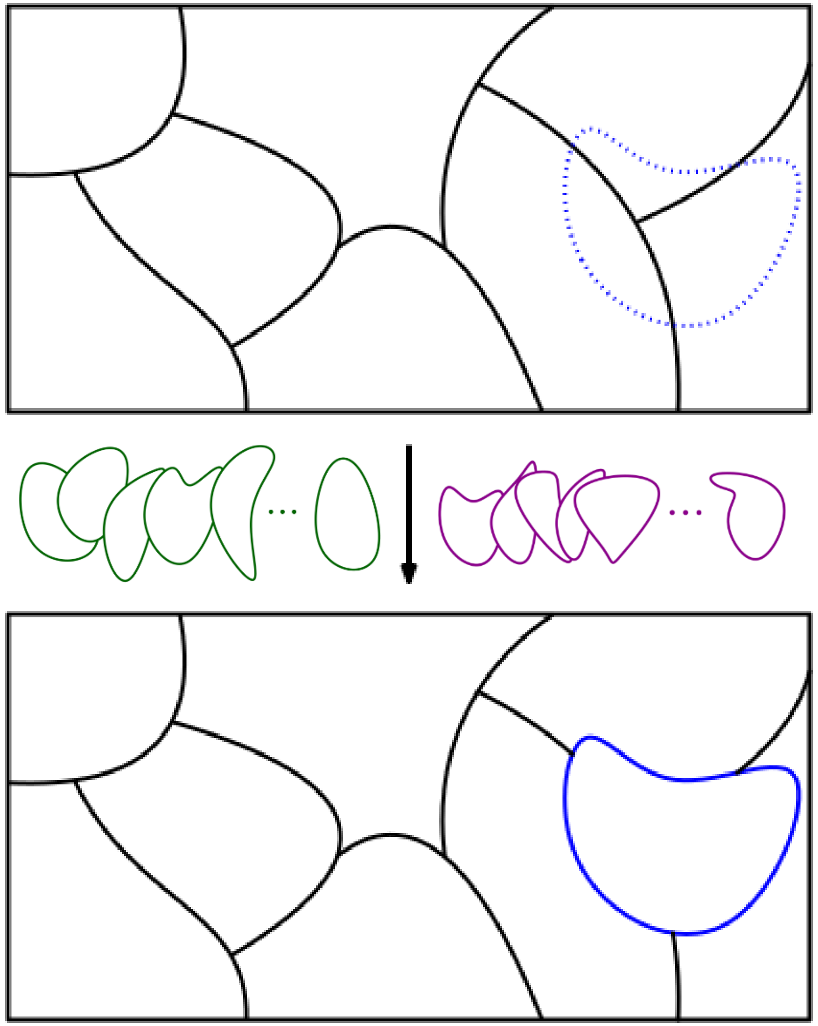
Use it to **guide** Local Search

*Assume we have this

Solution of the "large" Constrained LP

Guide Local Search, get solution for Constrained instance

Local Search



Main Idea

Problem: Exponentially many possible clusters

Local Search move: **swap** a **cluster** in the clustering

Solution of the "large" LP:

➢ ***poly***-sized family of clusters

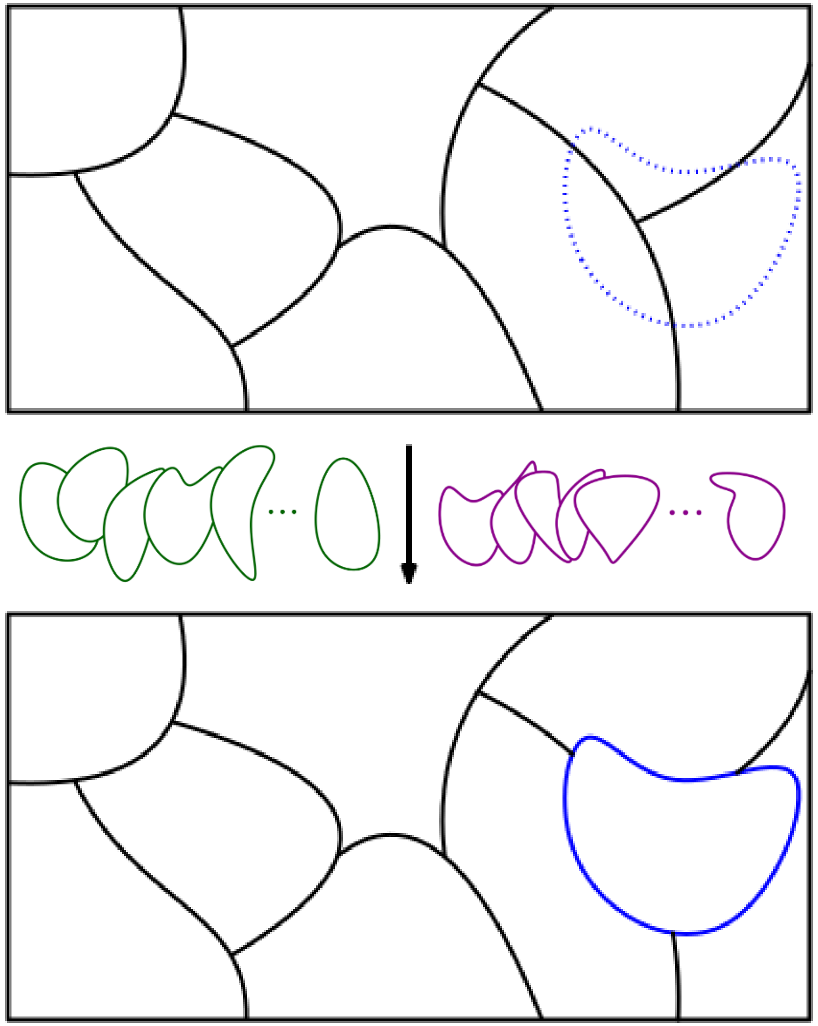
Use it to **guide** Local Search

*Assume we have this

Solution of the "large" Constrained LP

Guide Local Search, get solution for Constrained instance

Local Search



Result: Better-than-2 approximation for both variants!

Open Directions

- Efficient solution to the "large" Constrained LP
- Broader applications of combining LP and Local Search techniques
- Inapproximability of Constrained Correlation Clustering

Open Directions

- Efficient solution to the “large” Constrained LP
- Broader applications of combining LP and Local Search techniques
- Inapproximability of Constrained Correlation Clustering

Thank You!