Semi-supervised Spectral Clustering with automatic propagation of pairwise constraints

Nicolas Voiron [1] , Alexandre Benoit [1] , Andreï Filip [2] , Patrick Lambert [1] et Bogdan Ionescu [2]

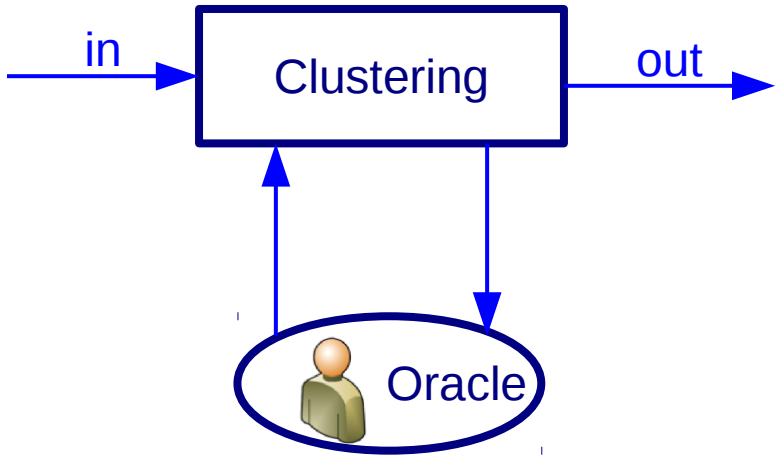
1 : LISTIC, University Savoie Mont Blanc, 74940, Annecy le Vieux, France
{nicolas.voiron, alexandre.benoit, patrick.lambert}@univ-smb.fr

2 : LAPI, University Politehnica of Bucharest, 061071, Bucharest, Romania
{afilip, bionescu}@alpha.imag.pub.ro

Table of contents

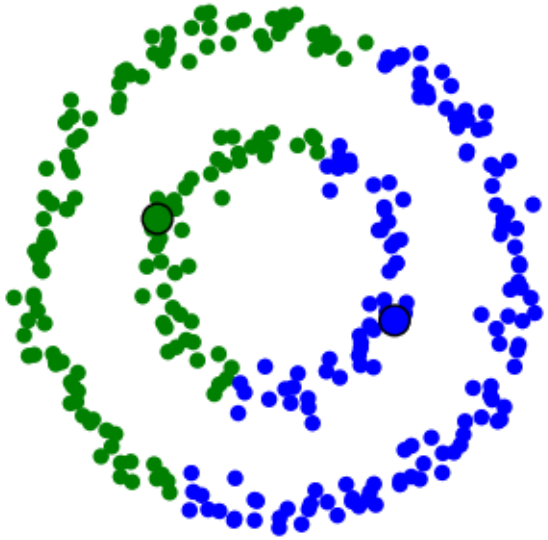
Semi-supervised clustering

- How to guide an automatic clustering ?
  - An iterative process
  - Knowledge added at each iteration with calls to an Oracle

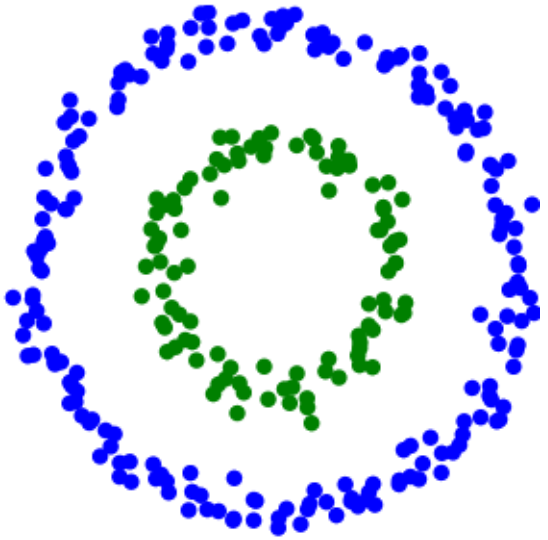


Semi-supervised Spectral Clustering

- Which clustering method to choose ?
  - We focus on Spectral Clustering that works by connectivity identification and is able to identify non-convex clusters



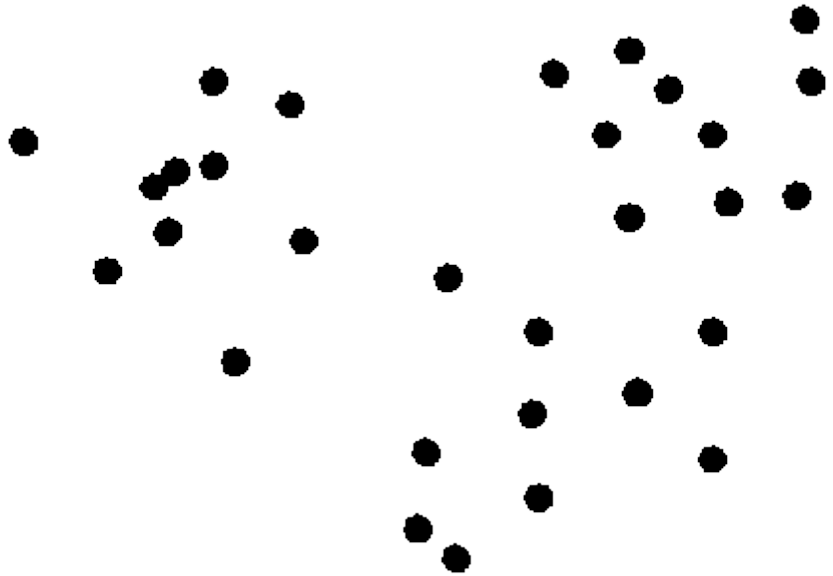
K-means clustering



Spectral Clustering

Semi-supervised Spectral Clustering with pairwise constraints

- How can an Oracle contribute ?
  - Absolute class annotation is harder than comparison class
  - Pairwise constraints Must Link and Cannot Link easily indicate if two objects belongs or not to the same class



Context summary and contributions

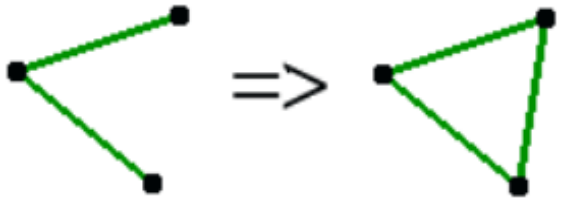
- Specifications:
  - An iterative process
  - Knowledge added at each iteration with calls to an Oracle
- Technical choice:
  - Spectral Clustering
  - Supervision by adding pairwise constraints
    Must Link and Cannot Link

- Contributions:
  - Accelerate the clustering process by propagating links (constraints)

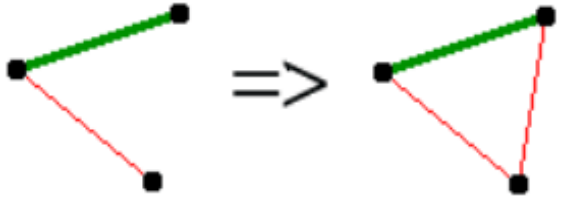
Table of contents

Contraint propagation state of the art

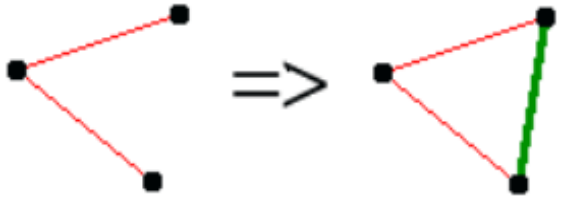
- Rule 1: ML+ML=>ML



- Rule 2: ML+CL=>CL

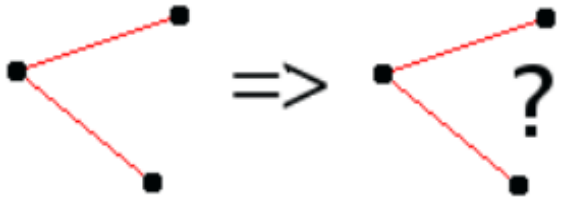


- Rule 3:
  - 2 classes : CL+CL=>ML

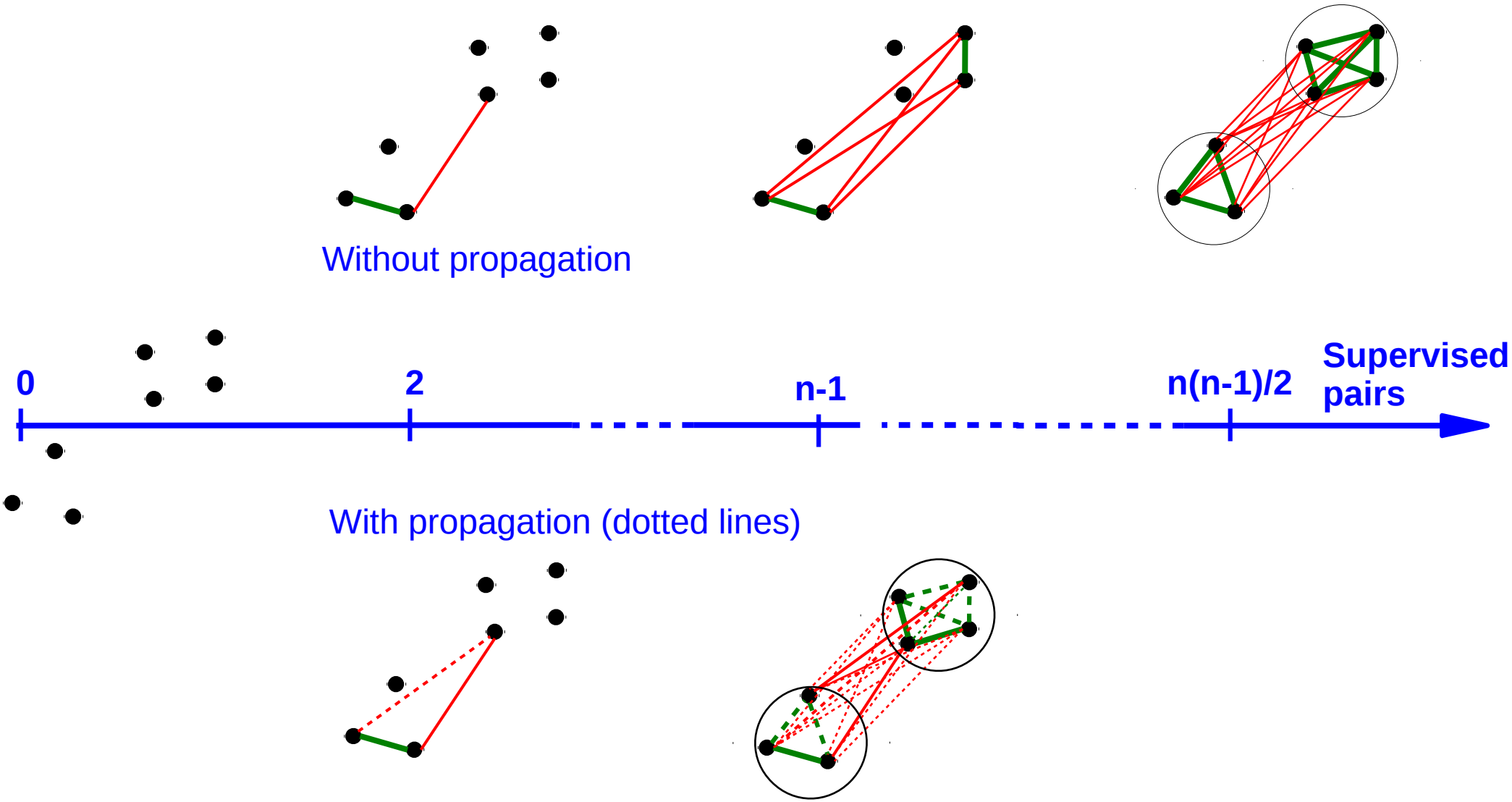


Unresolved issue

- n classes : CL+CL=>?

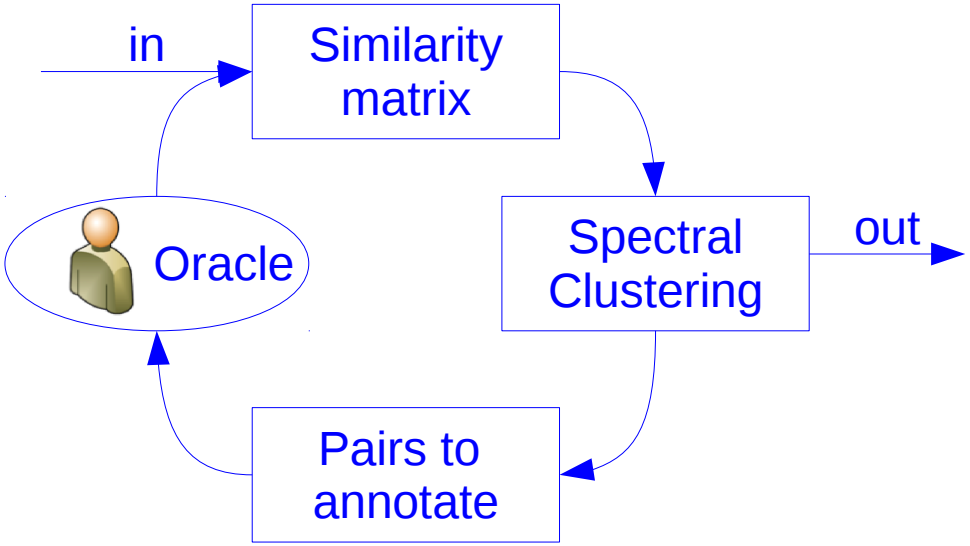


Taking advantage of the propagation



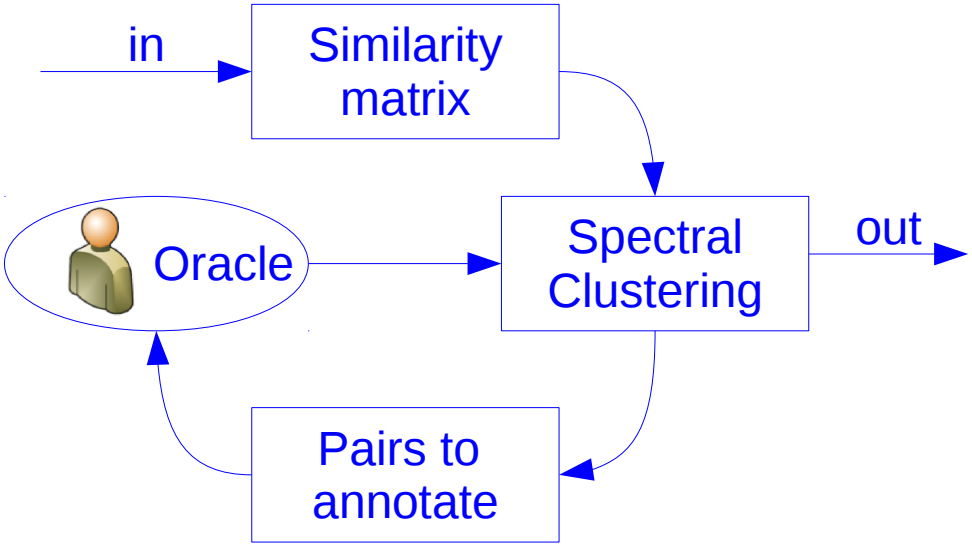
Oracle knowledge introduction

- **Active Clustering**
  => *Constraints not always respected*
  => *Fast*



C. Xiong, D. Johnson, J. J. Corso
« Active Clustering with Model-Based Uncertainty Reduction »
CoRR 2014

- **COSC**
  => Better *(up to perfect) constraint respect*
  => *Slow*



S. S. Rangapuram and M. Hein
« Constrained 1-spectral clustering »
in Proceedings of the 15th International Conference on AISTATS 2012

Table of contents

1st benefit of constraint propagation

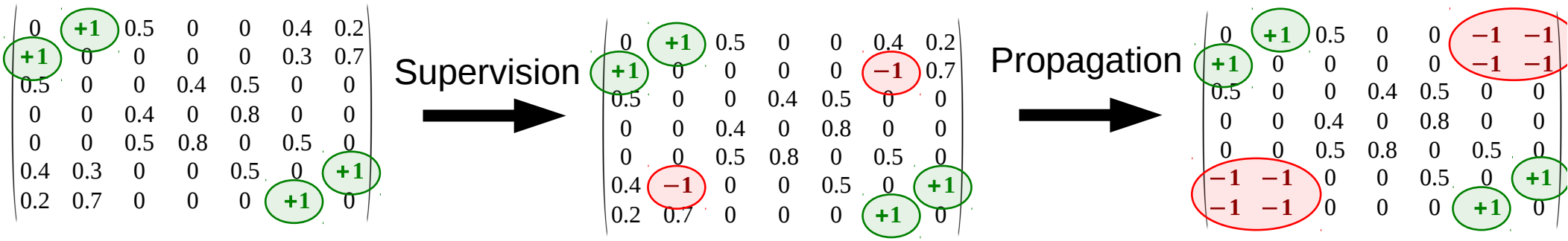
- Only call the Oracle when necessary to reduce annotation and clustering computationnal costs

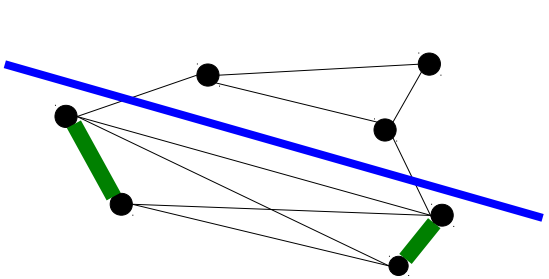
Without propagation:
useless Oracle calls

With propagation:
optimized Oracle calls

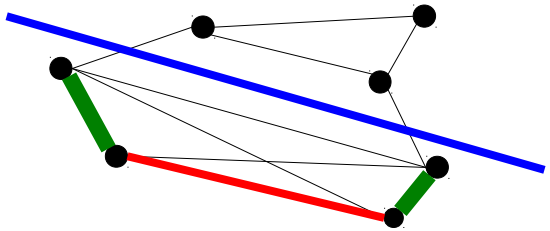
2nd benefit of constraint propagation

- Some methods like « Active Clustering » [Xiong & al] add constraints in the adjacency matrix. Constraints are not necessarily respected by the Spectral Clustering step. Propagation amplifies the constraint respect for those methods

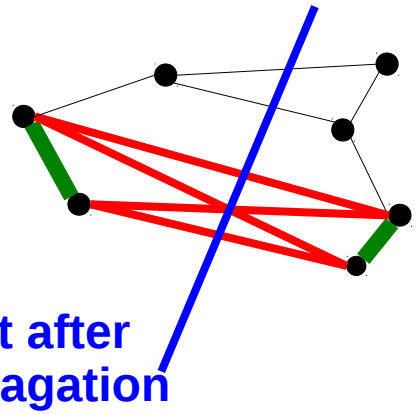




Initial cut



Cut without propagation

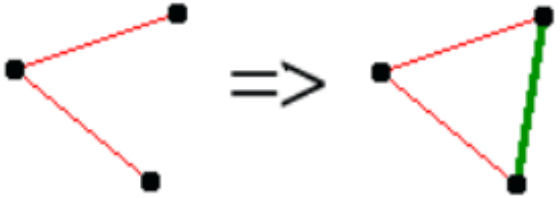


Cut after propagation

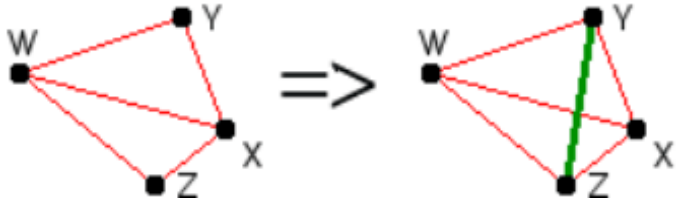
Our contribution: 3rd rule generalization

- **CL+CL => ?**

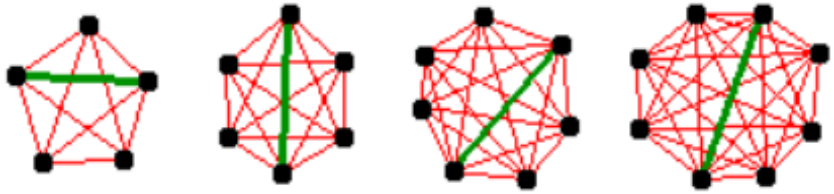
2 classes



3 classes



4-7 classes



- Large number of configurations to inspect (time consuming)
+ But those configurations are frequent (more CL than ML)
+ Cascade effect can be expected

The entire process of our method

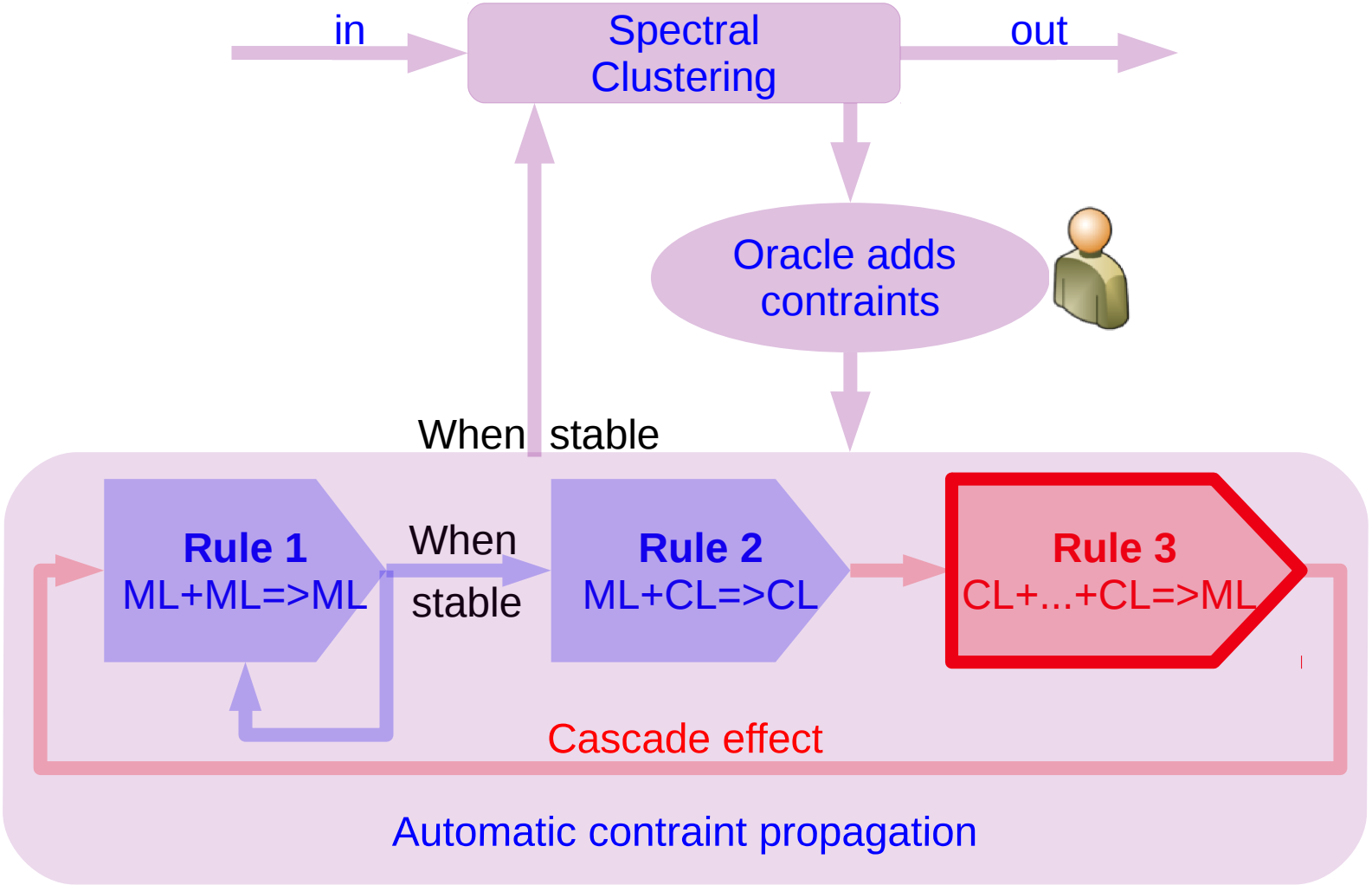
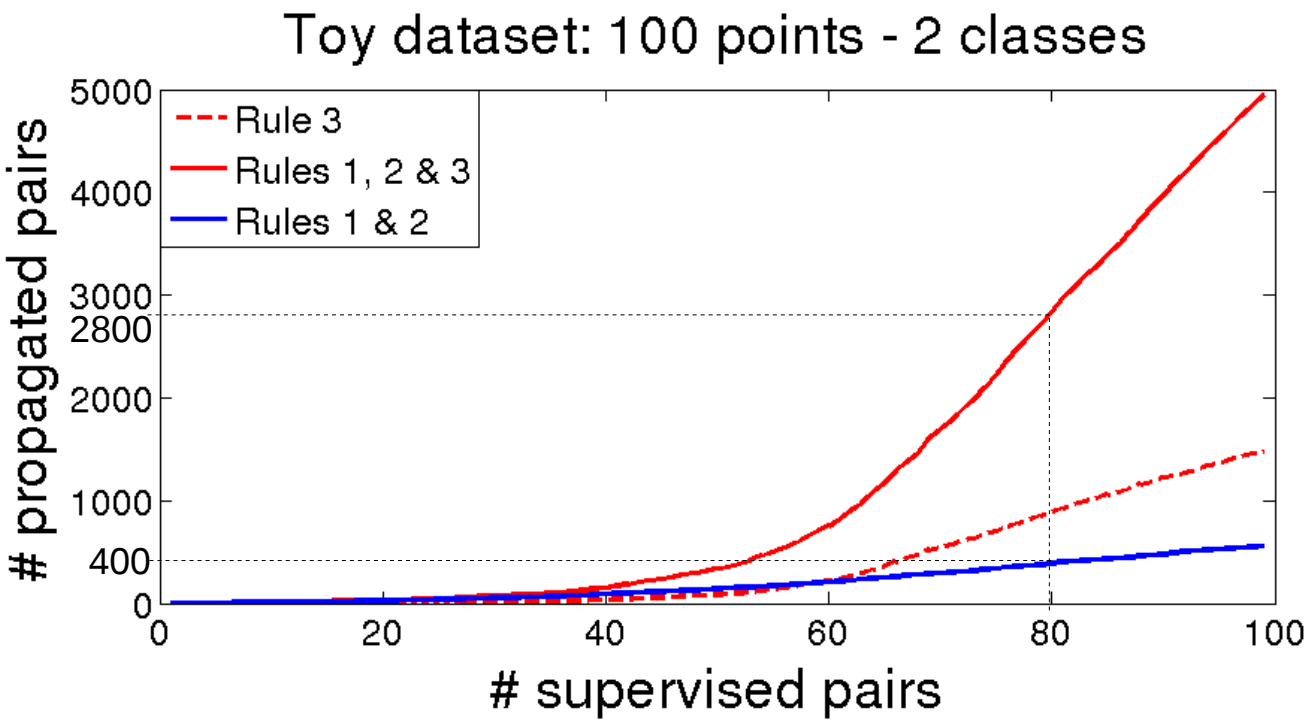
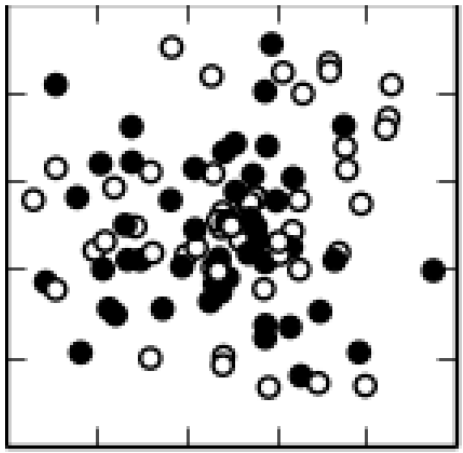


Table of contents

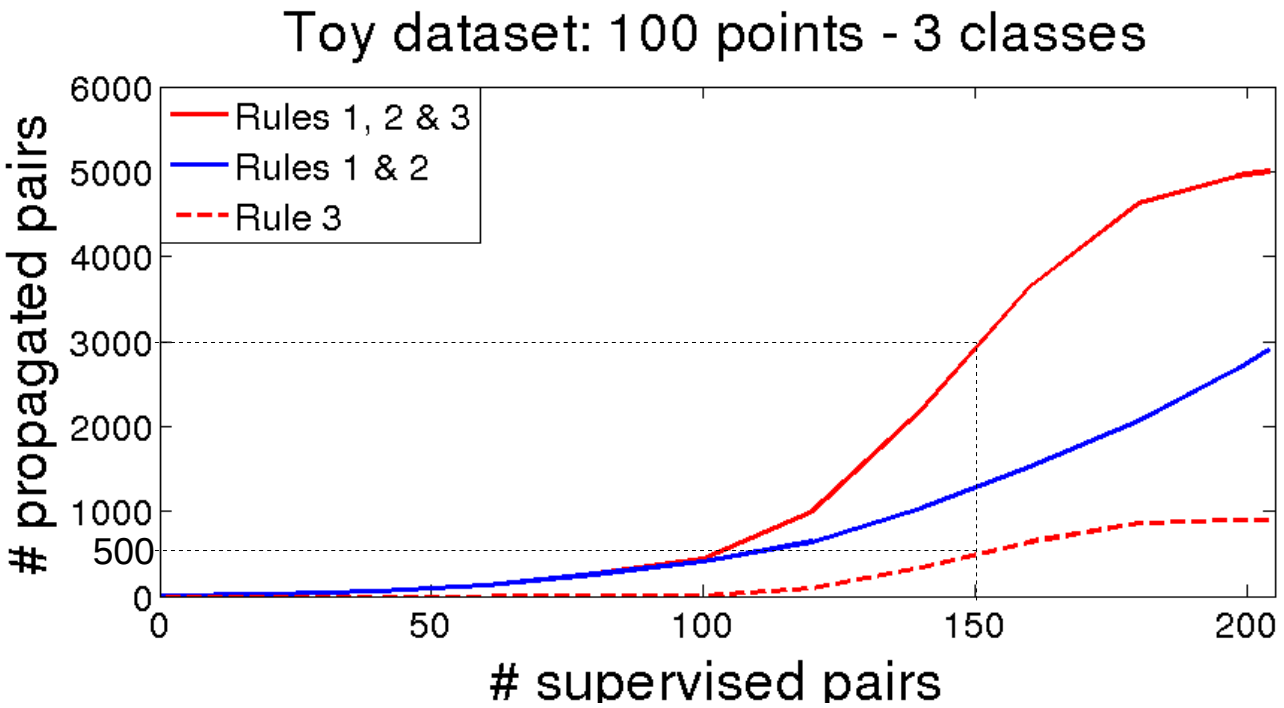
Propagation impact

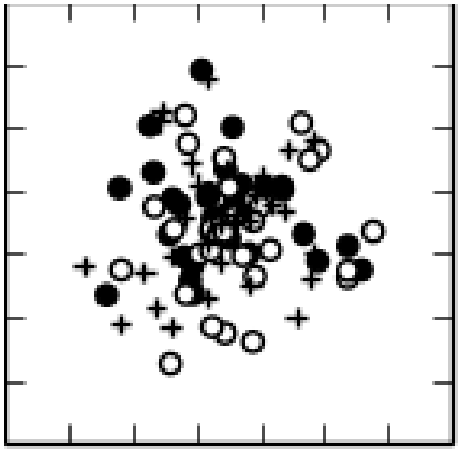




- Propagation requires a sufficient set of connected objects
- Significant impact of the third rule

Propagation impact

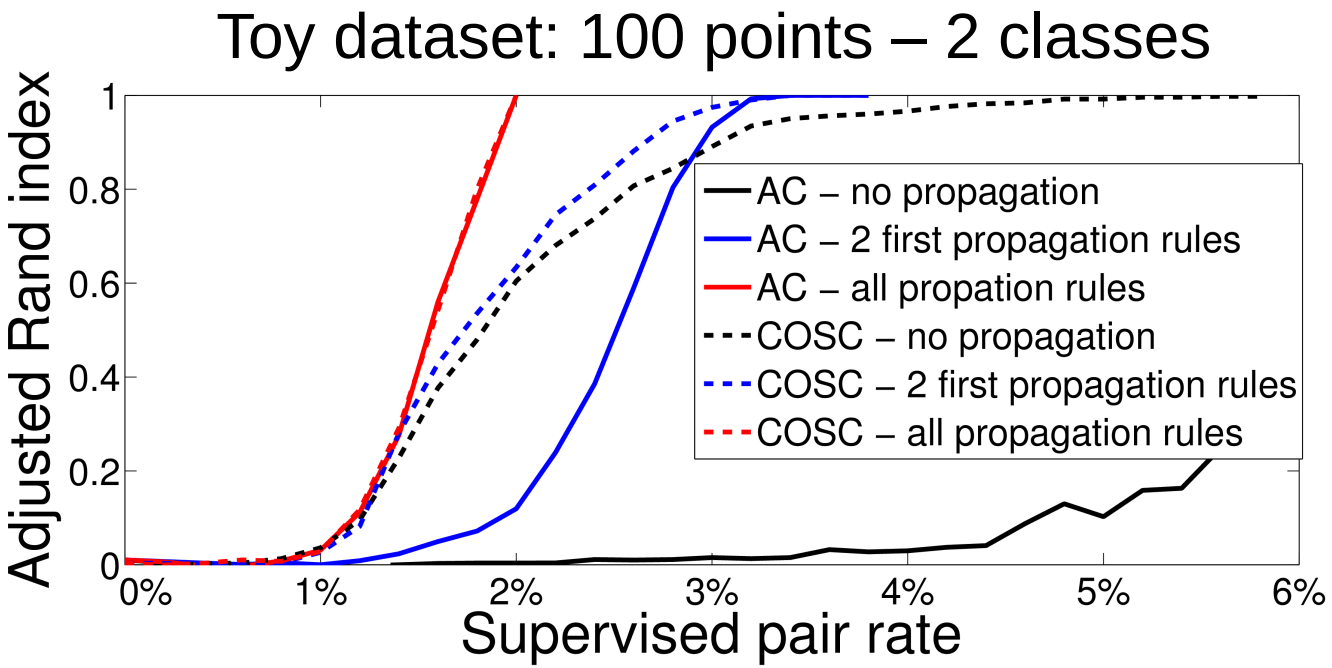


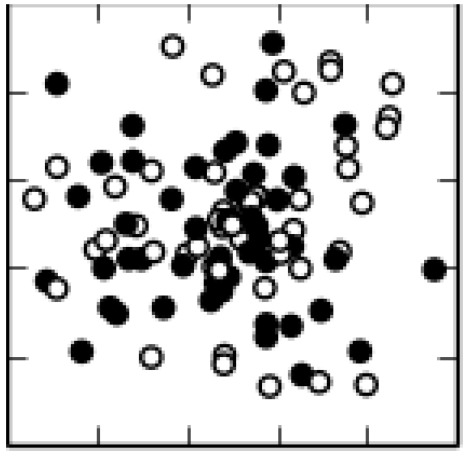


- Similar results with a 3 classes problem

Bi-partitionning quality of built data

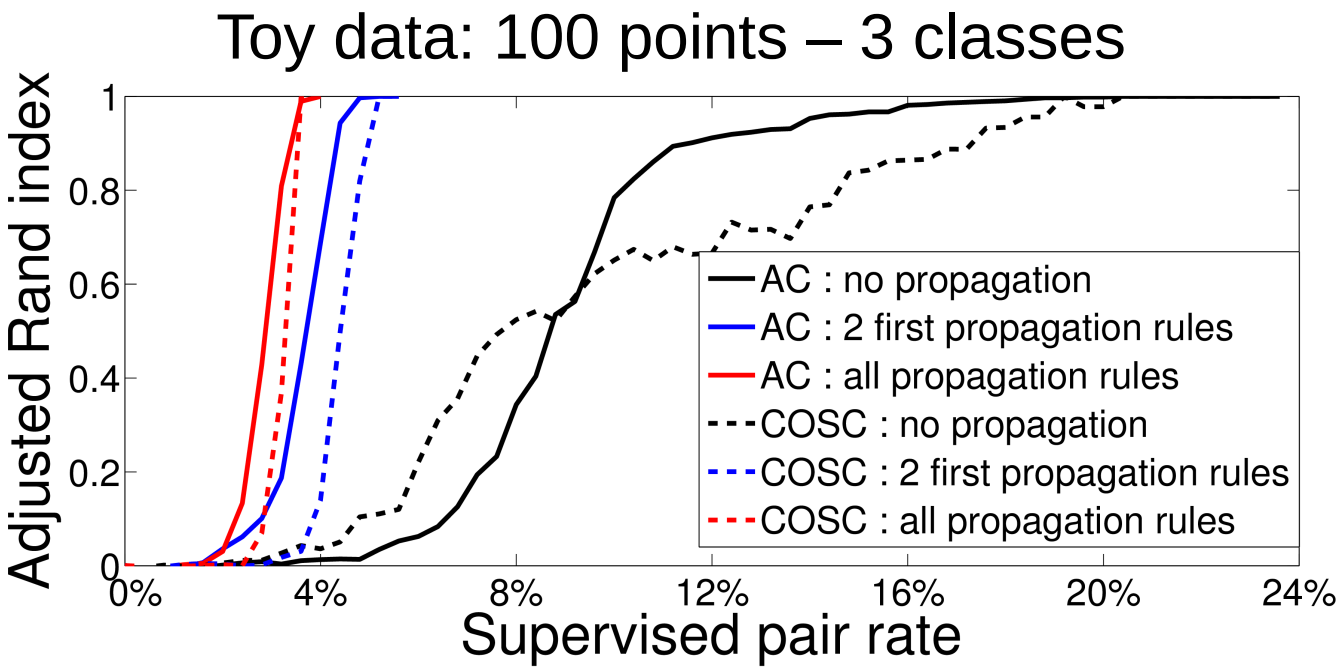
- A better quality score and faster
  - Enforce both AC and COSC methods
  - A simpler clustering method (AC) can reach a more complex method (COSC) partition quality

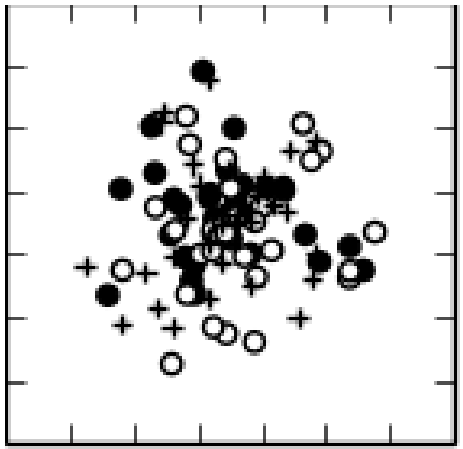




Tri-partitionning quality of built data

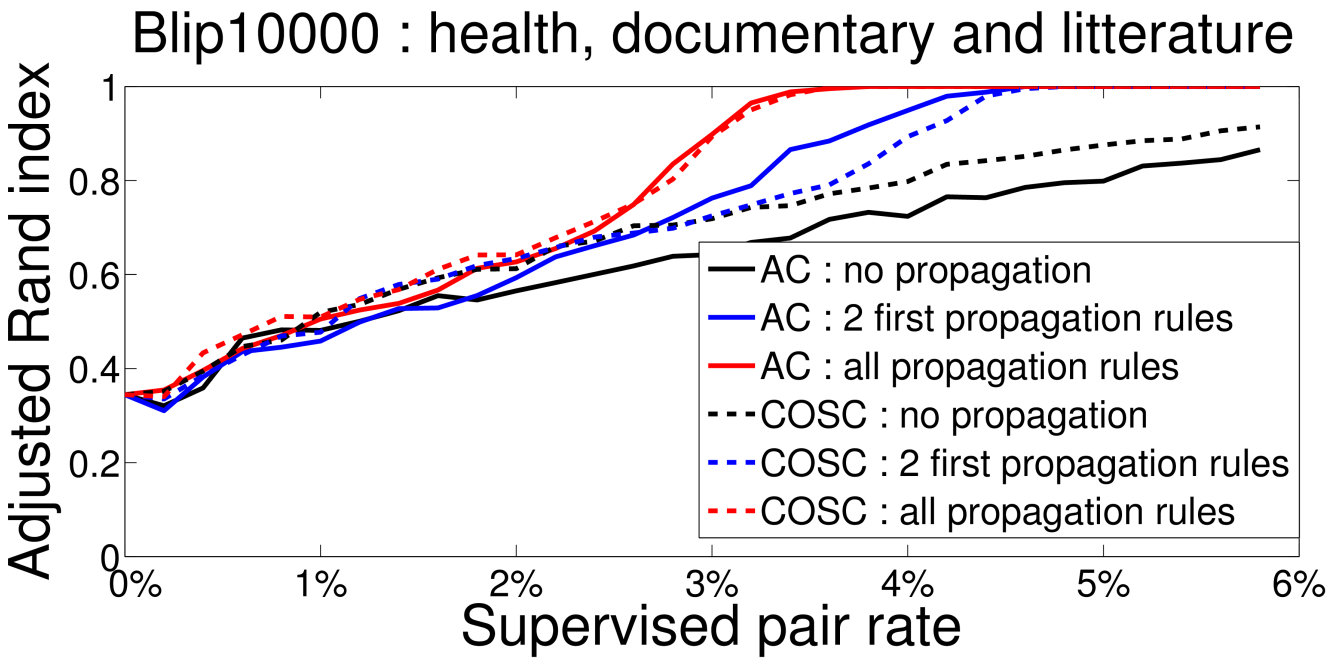
- A better quality score and faster
  - Enforce both AC and COSC methods
  - AC is better than COSC in a 3 class problem





Tri-partitionning of MediaEval data

- Sample with 100 videos taken into 3 classes of MediaEval challenge (Blip10000 - Audio Features)
- A better quality score and faster
  - Still enforce both AC and COSC methods
  - The total propagation raise the AC method up to the COSC method score



Multi-partitionning of MediaEval data

- Sample with 5127 videos taken into the 26 classes of MediaEval challenge (Blip10000 - Audio Features)
- A better quality score and faster
  - Still enforce both AC and COSC methods
  - Third rule requires optimized algorithms

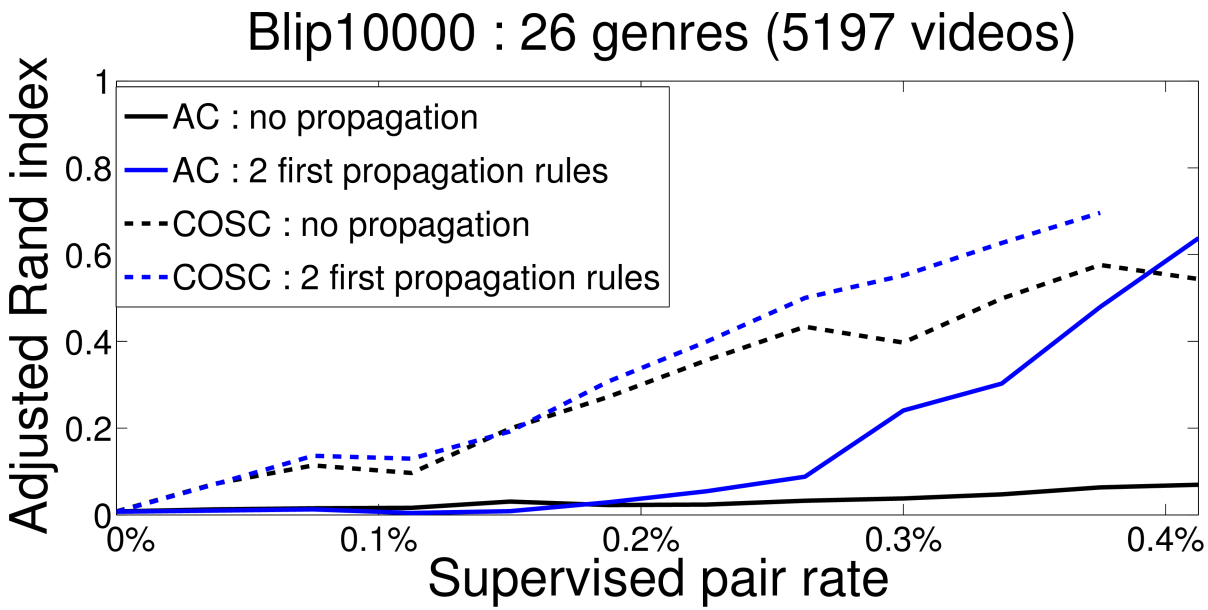


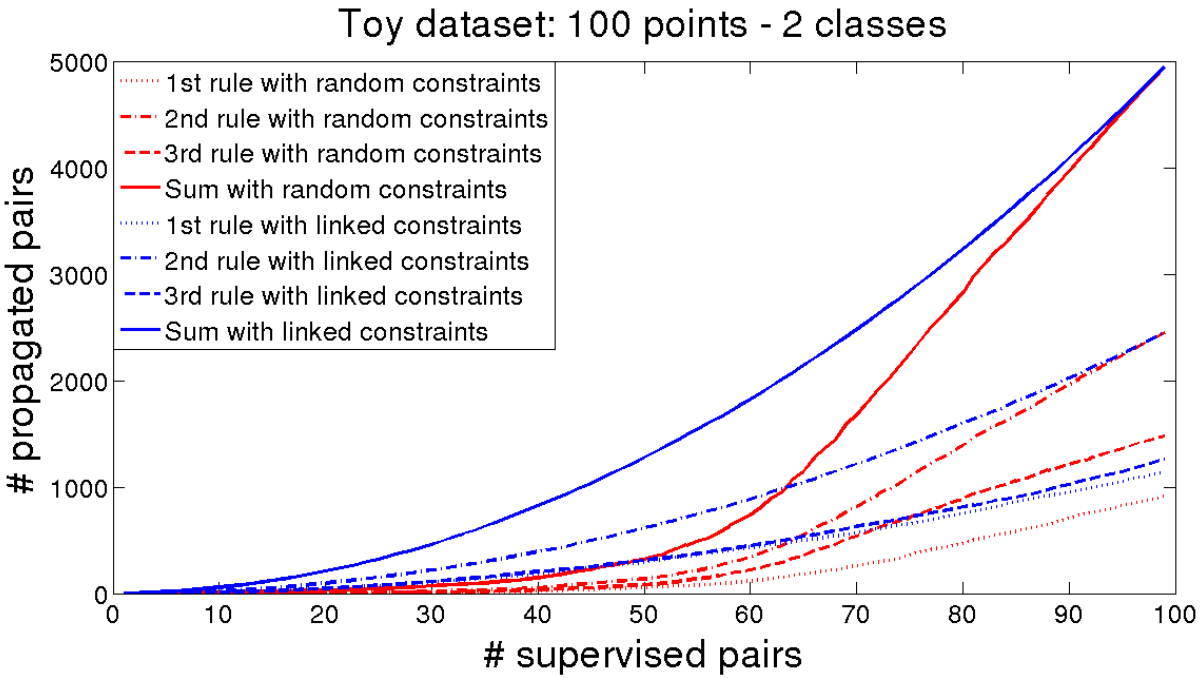
Table of contents

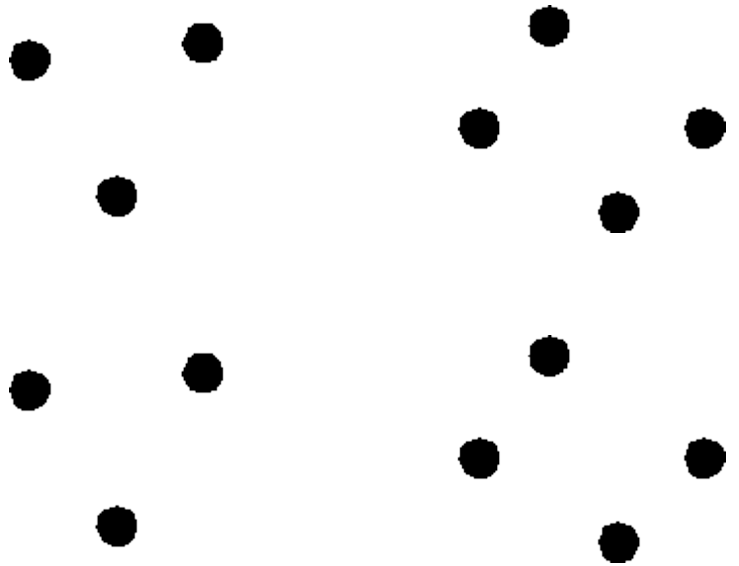
Conclusion and perspectives

- Conclusion
  - \+ Strong clustering enhancement by the use of propagation
  - \+ Benefit of the 3rd rule generalization
  - \+ Enforce simple clustering methods
  - \- 3rd rule is costly
- Perspectives
  - ✗ Improving propagation algorithms for computationnal cost and scaling step
  - ✗ Add constraint selection strategies amplifying the propagation benefits
  - ✗ Build up a real-time annotation and propagation framework
  - ✗ Experiment propagation with other clustering methods

A first strategy perspective

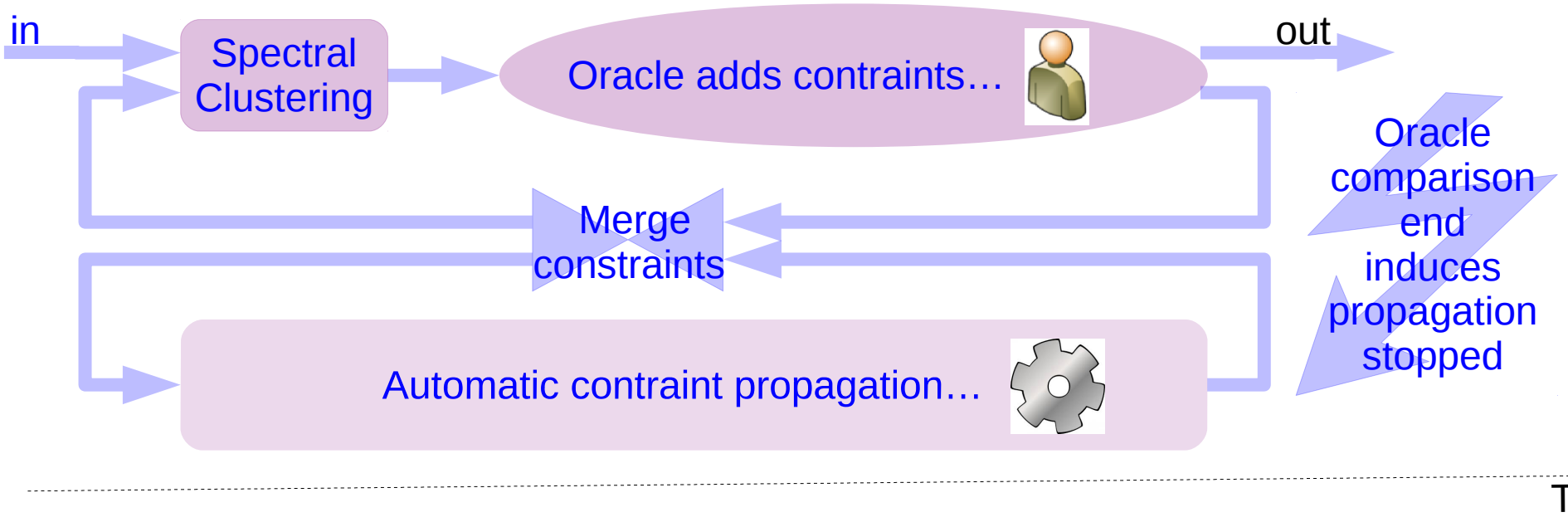
- We call “random link selection” a first strategy amplifying the propagation benefit
- We restrict random pair selection to the subset of pairs having only one object connected to an already supervised pair
- On green curves we can see that the random linked selection strategy boosts propagation for all the rules from the beginning



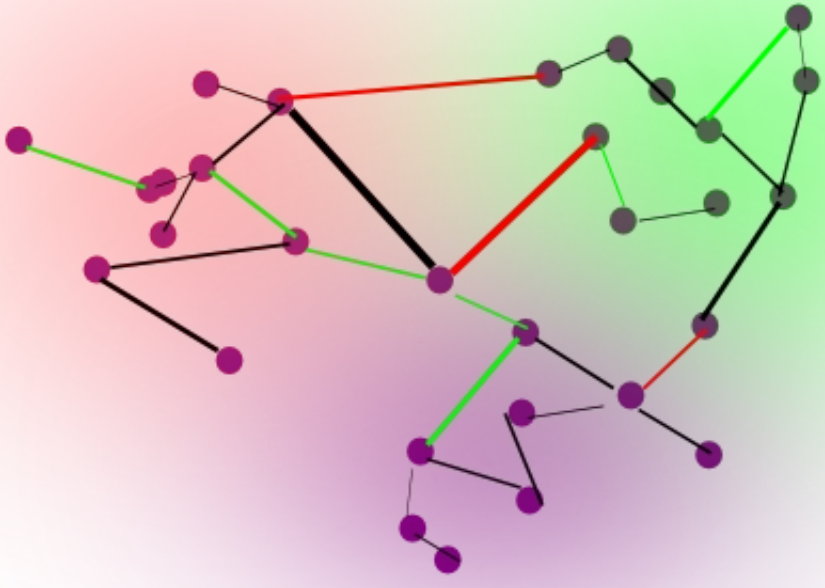


The applicative perspective

- With a lot of objects and classes, the propagation is expensive
- We don't want the Oracle waiting too much
- We can consider a process in which the Oracle comparison time is used to process the automatic constraint propagation
- After the clustering step, the Oracle supervises the constraints. During this time, the automatic propagation is proceeded.



Thanks for your attention



Questions