Clustering Spectral semi-supervisé avec propagation automatique des contraintes par paires

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

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

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

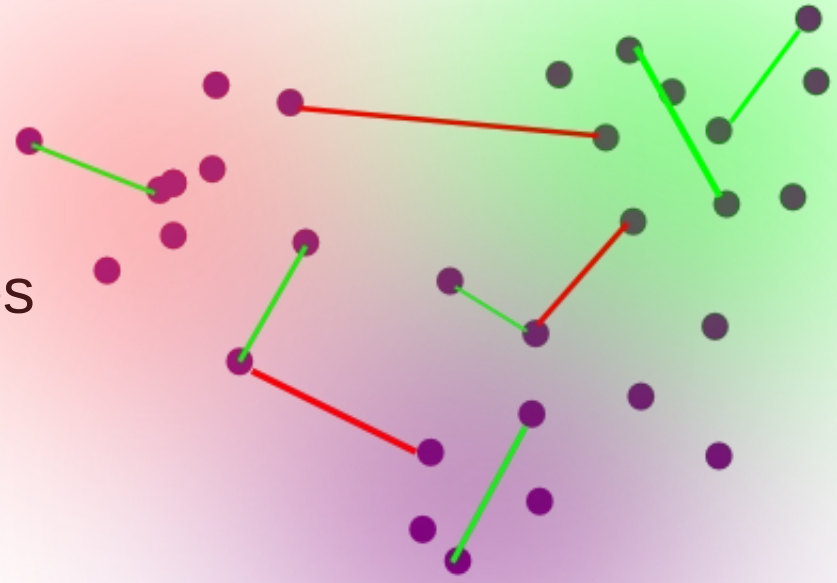
Plan

1. Le contexte
2. L'existant
3. Notre contribution
   - La généralisation des contraintes
   - Le processus complet de notre méthode
4. Les résultats
   - Impact de la propagation
   - Qualité du clustering obtenu avec des données synthétiques
   - Qualité du clustering obtenu avec des données réelles
5. Conclusion

Classification semi-supervisée

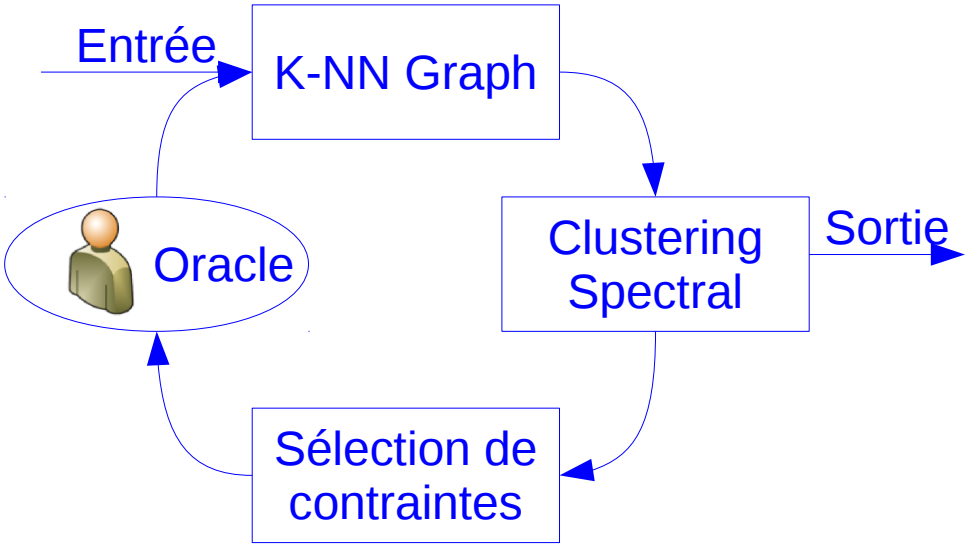
- Caractéristiques
  - Processus itératif
  - Ajout de connaissance à chaque itération : appel à un Oracle
- Choix technique :
  - Clustering Spectral
  - Supervision par ajout de contraintes

    Must Link et Cannot Link

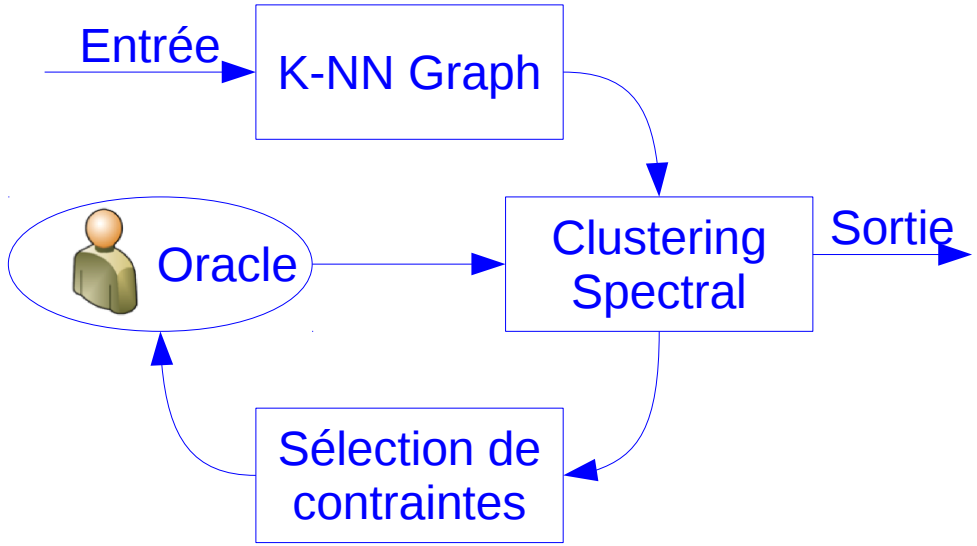


Intégration de la connaissance

- dans la matrice d'adjacence : **Active Clustering** [Corso&al] => *l'oracle guide, contraintes pas forcément respectées*



- dans un problème d'optimisation **COSC** [Hein&al] => *un meilleur (voir même parfait) respect des contraintes*

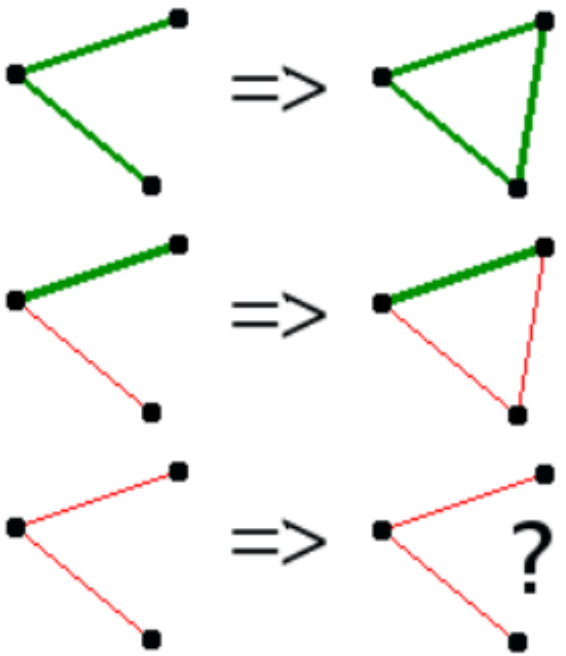


- **[Corso&al]** C. Xiong, D. Johnson, J. J. Corso « Active Clustering with Model-Based Uncertainty Reduction », CoRR 2014

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

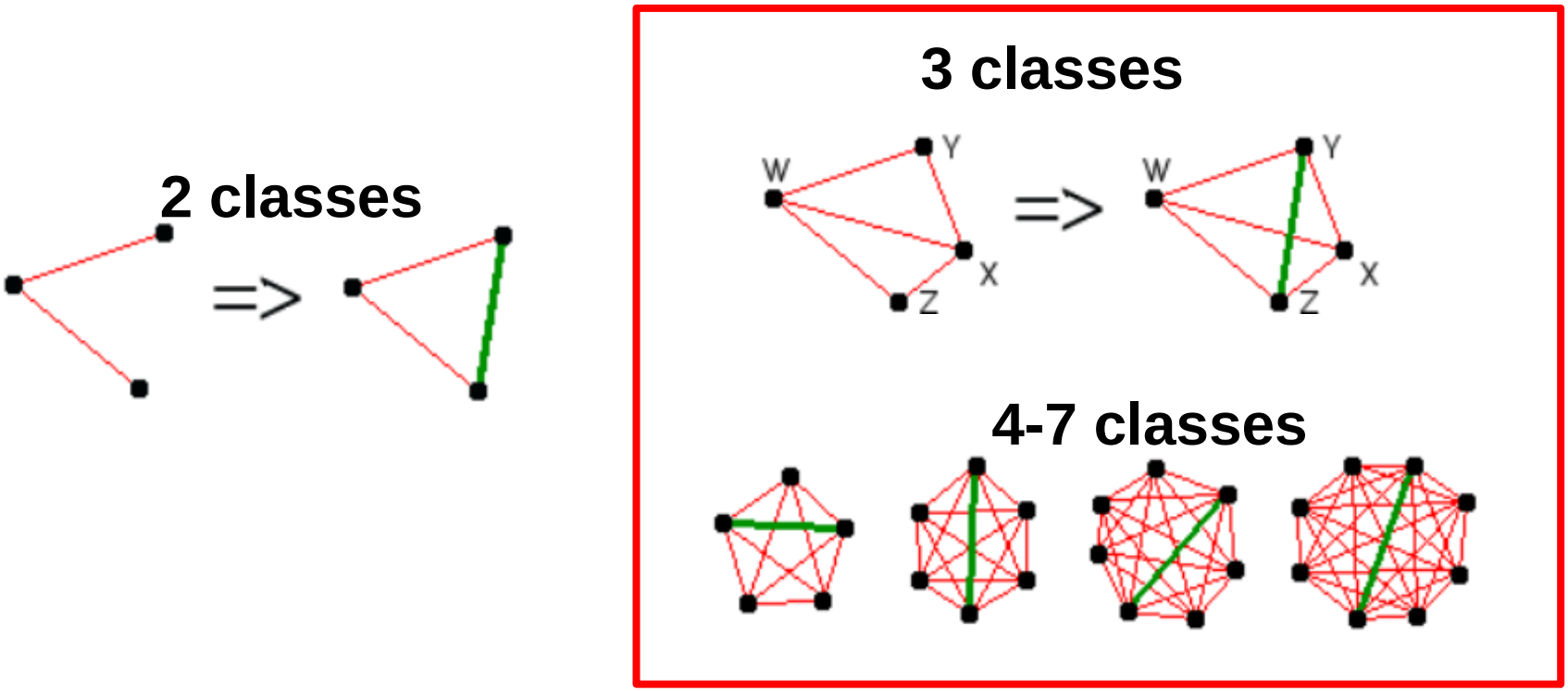
Propagation des contraintes

- Propager les contraintes :
  - Réduire la sollicitation des experts
  - Guider le Clustering Spectral pour obtenir une meilleure qualité de partition
- État de l'art :
  - Règle 1 : ML+ML=>ML
  - Règle 2 : ML+CL=>CL
  - Règle 3 : CL+CL=>?



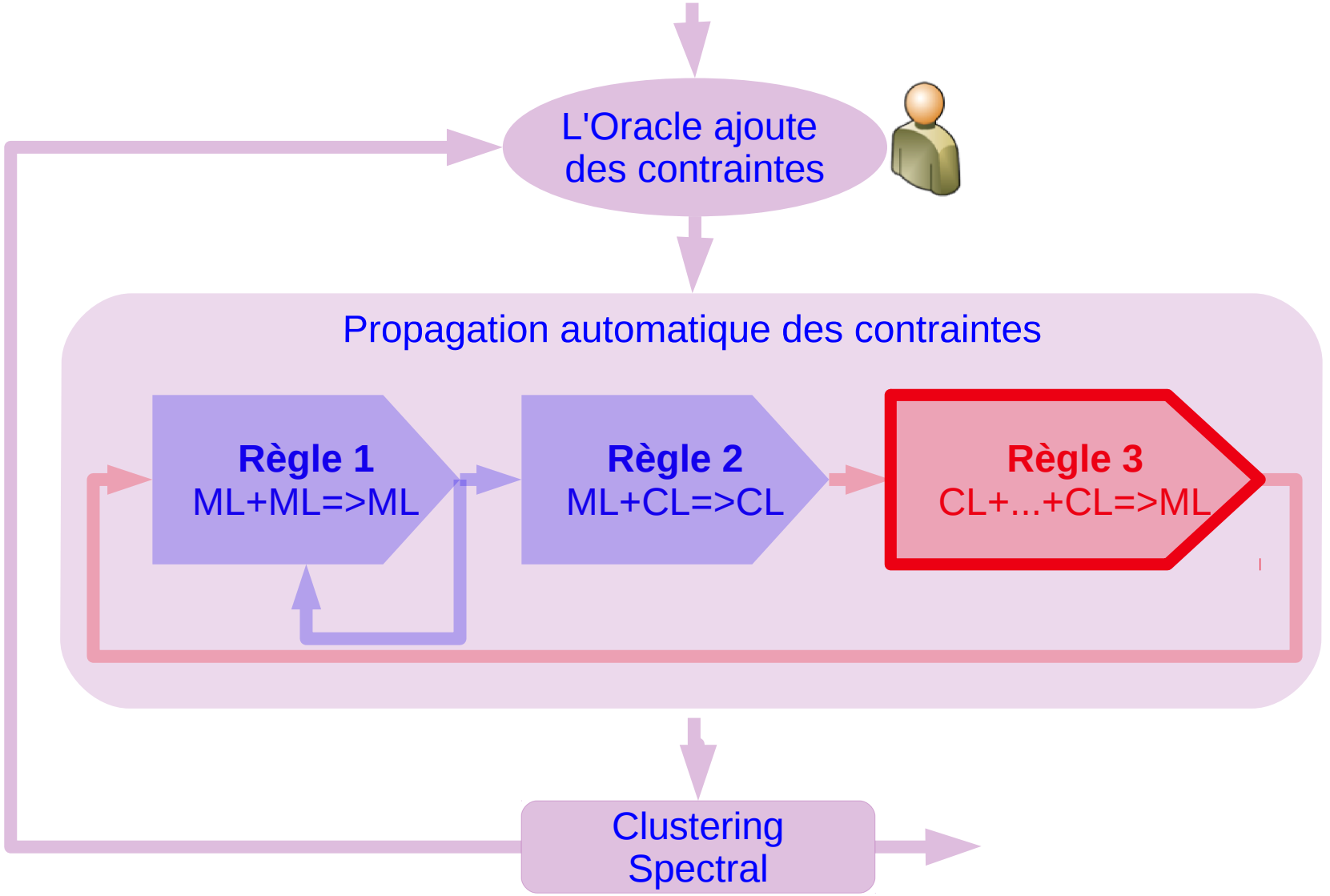
Propagation des contraintes

- **Généralisation de la 3ème règle CL+CL**



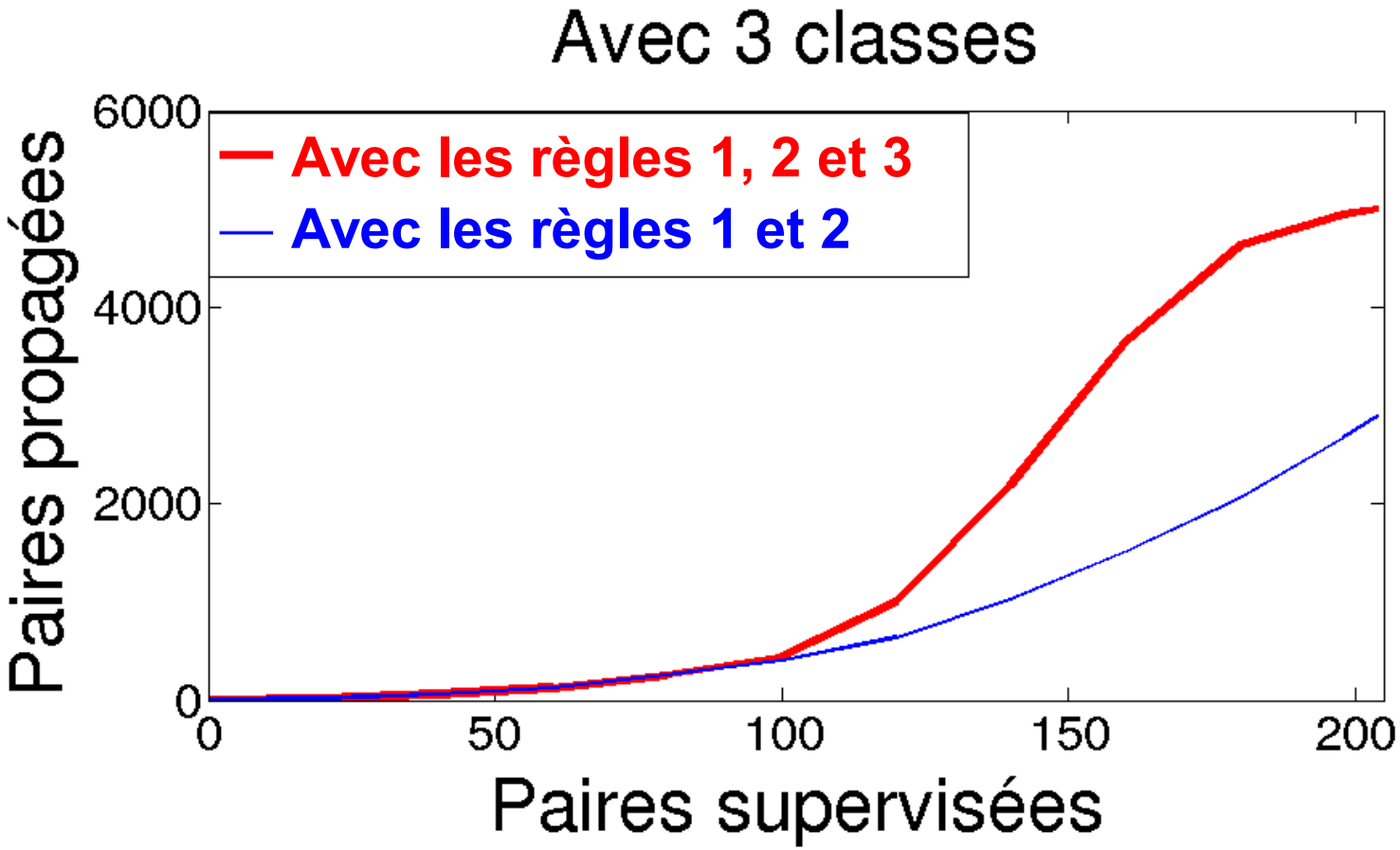
- Des cas finalement nombreux … mais pouvant devenir coûteux à rechercher lorsque le nombre de classes devient grand

Processus complet de notre méthode



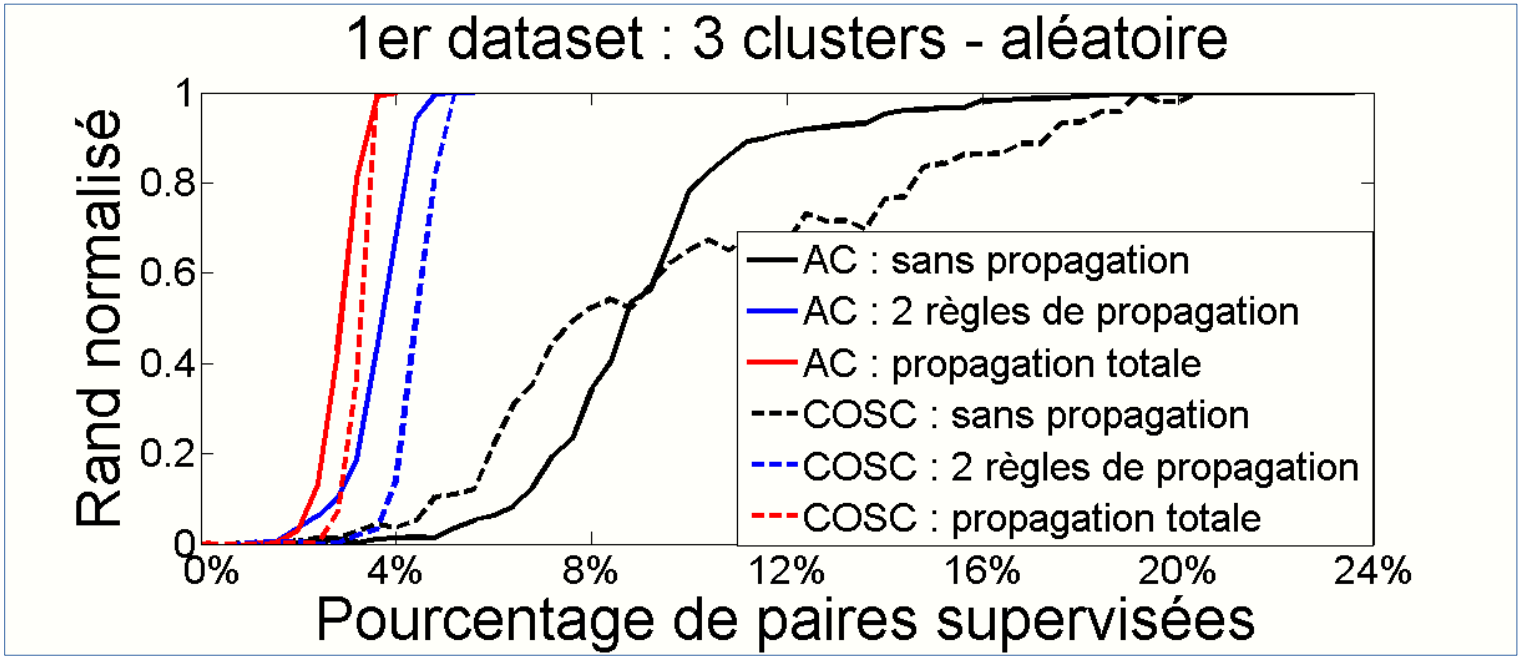
Impact de la propagation

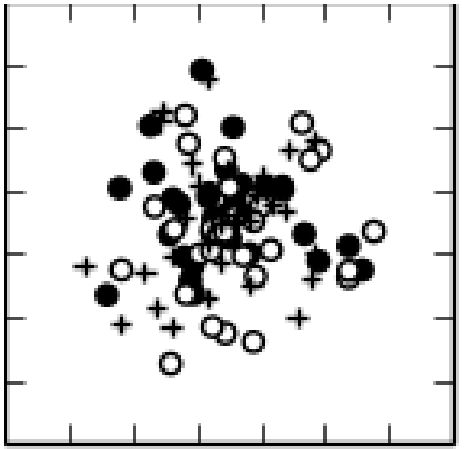
- 100 points équirépartis en 3 classes



Tri-partitionnement de données synthétiques

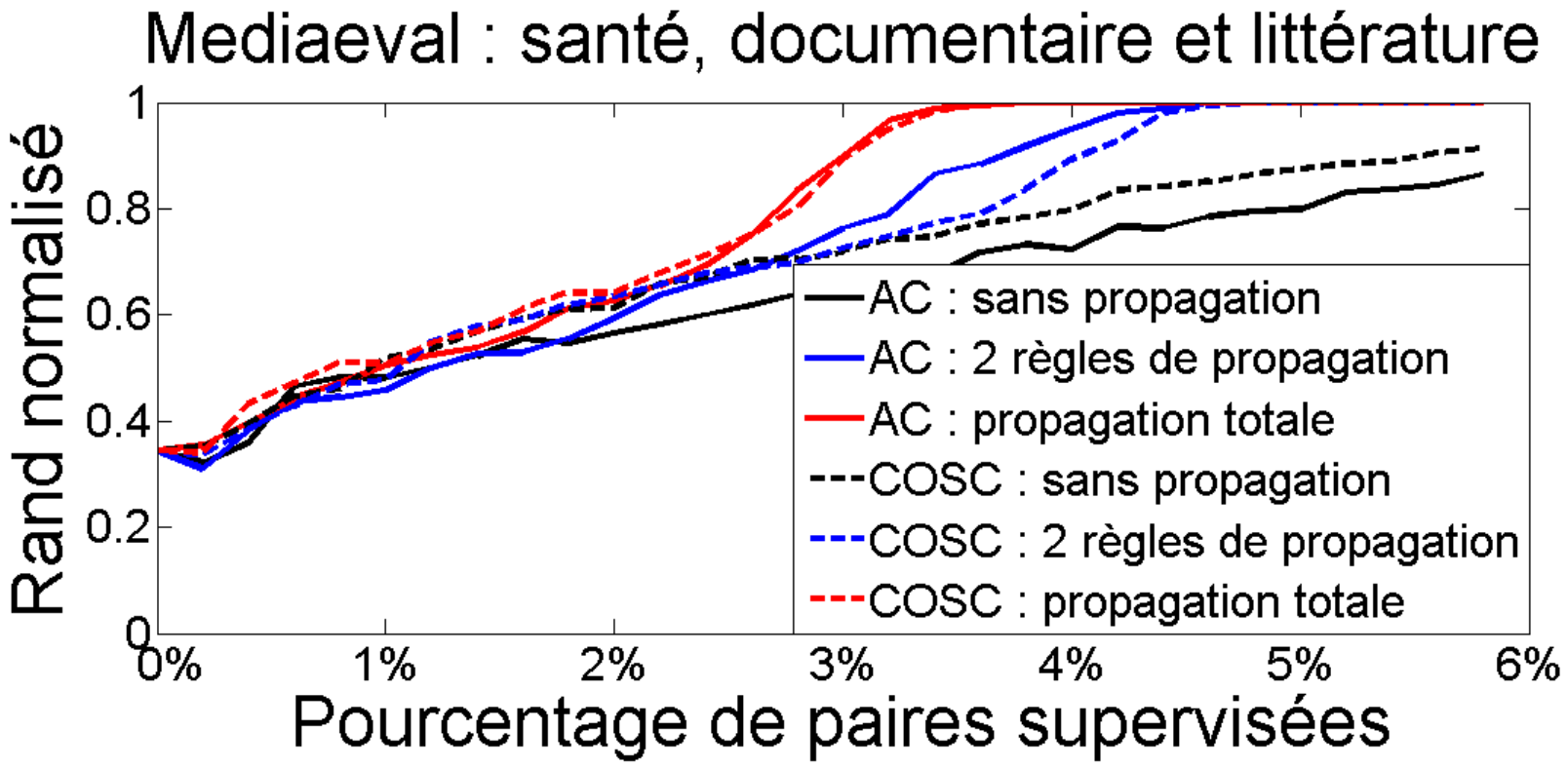
- Une meilleure qualité de partition, plus vite
  - Valable pour les deux méthodes AC et COSC
  - Exemple avec 100 points répartis en 3 classes aléatoires





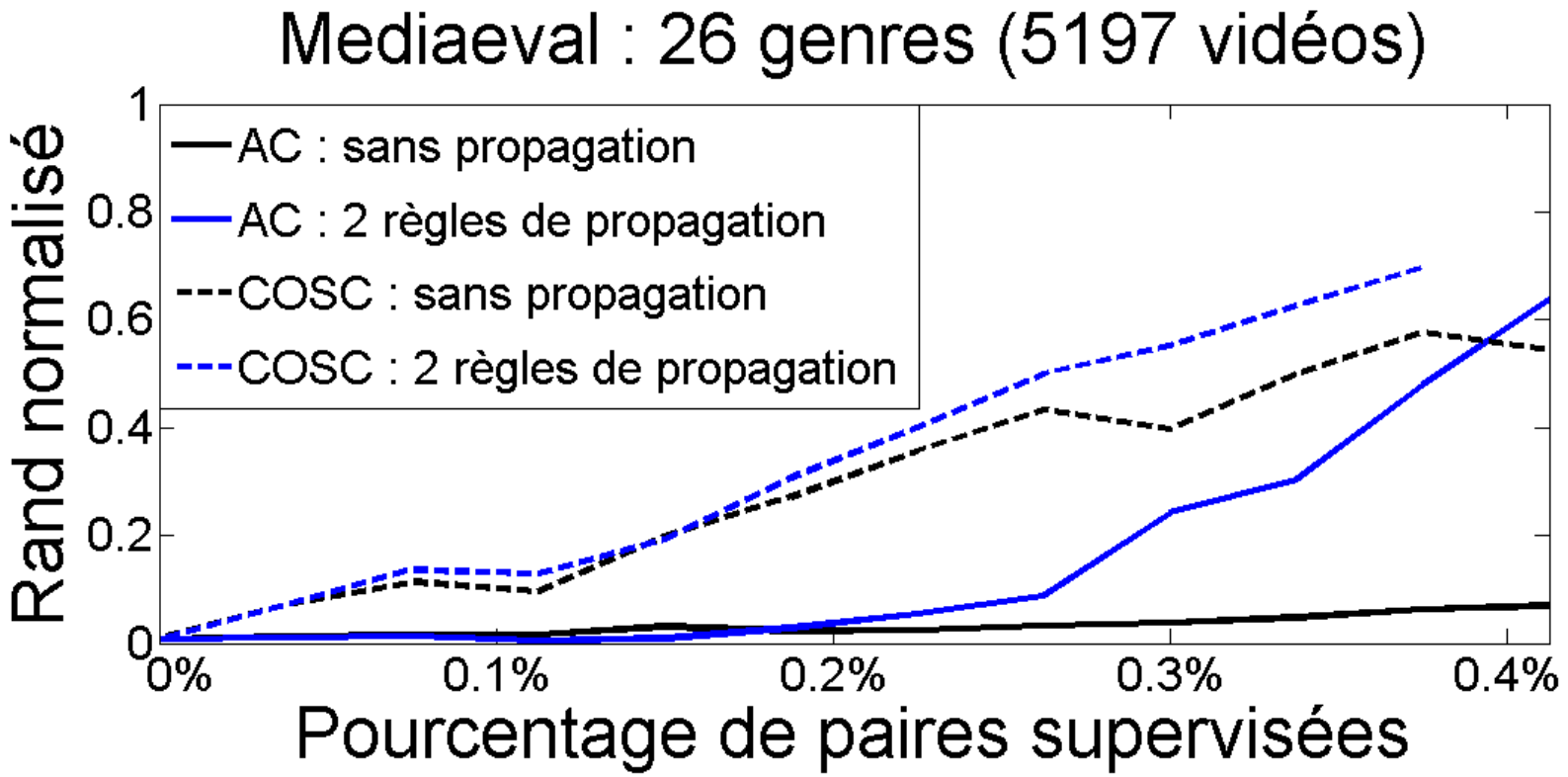
Tri-partitionnement des données réelles

- Une meilleure qualité de partition, plus vite
  - Valable pour deux méthodes AC et COSC
  - Exemple avec 100 vidéos réparties en 3 classes



Multi-partitionnement des données réelles

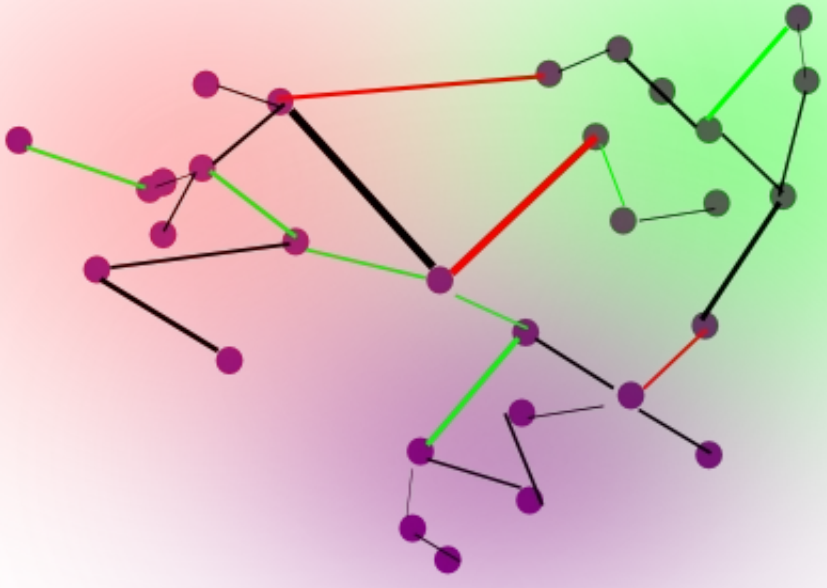
- Une meilleure qualité de partition, plus vite
  - Valable pour deux méthodes AC et COSC
  - Exemple avec 5127 vidéos réparties en 26 classes



Conclusion et perspectives

- Conclusion
  - Bénéfice de la propagation
  - Intérêt de la 3ème règle et de sa généralisation
  - Grâce à la propagation des contraintes, l'AC est aussi efficace que COSC
  - Propagation complète selon la 3ème règle coûteuse dans le cas d'un multi partitionnement en n classes avec n grand
- Perspectives
  - Amélioration des algorithmes de propagation en terme de coût de calcul et passage à l'échelle
  - Mettre en place des stratégies de sélection des contraintes à soumettre à l'Oracle amplifiant les gains obtenus par la propagation
  - Évaluer l'impact de la propagation généralisée sur d'autres méthodes de clustering

Merci de votre attention



Questions