Méthodes de classification organisée pour la recherche de communautés dans les réseaux sociaux

Nathalie Villa-Vialaneix & Fabrice Rossi
http://www.nathalievilla.org

Institut de Mathématiques de Toulouse, Toulouse School of Economics
& IUT STID (Carcassonne) - Université de Perpignan

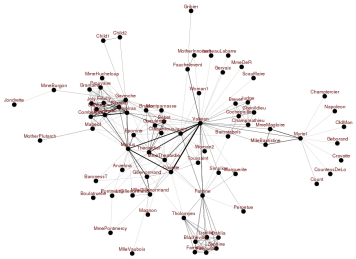
Journée Satellite STID aux JdS09

Bordeaux, 29 mai 2009

Sommaire

Université de Toulouse

Graphes et réseaux

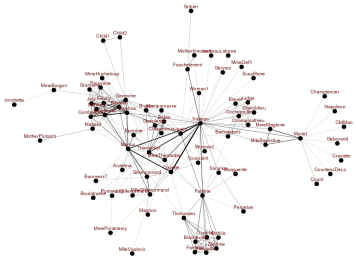


Cet exposé se propose de présenter des méthodes d'analyse de structures se présentant sous la forme de **données relationnelles** entre individus. Ici, on supposera les données représentées par un **graphe pondéré non orienté** : on notera donc G **un graphe**

- de **sommets** $V = \{x_1, \ldots, x_n\}$ (et de taille n)

Université de Toulouse

Graphes et réseaux

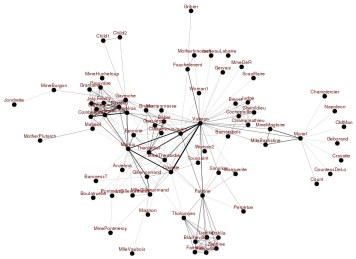


Cet exposé se propose de présenter des méthodes d'analyse de structures se présentant sous la forme de **données relationnelles** entre individus. Ici, on supposera les données représentées par un **graphe pondéré non orienté** : on notera donc G **un graphe**

- de **sommets** $V = \{x_1, \ldots, x_n\}$ (et de taille n)
- dont l'ensemble des **arêtes** est noté E. E est donc un sous-ensemble de $V \times V$

Université de Toulouse

Graphes et réseaux

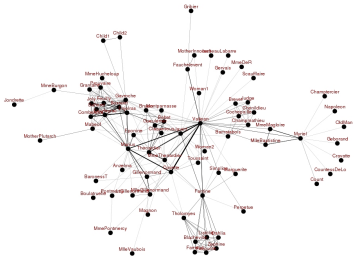


Cet exposé se propose de présenter des méthodes d'analyse de structures se présentant sous la forme de **données relationnelles** entre individus. Ici, on supposera les données représentées par un **graphe pondéré non orienté** : on notera donc G **un graphe**

- de **sommets** $V = \{x_1, \ldots, x_n\}$ (et de taille n)
- dont l'ensemble des **arêtes** est noté E. E est donc un sous-ensemble de $V \times V$
- dont les arêtes sont pondérées par la **matrice de poids** W telle que

$$\forall\, i, j = 1, \ldots, n,\ w_{ii} = 0, \qquad w_{ij} = w_{ji} \geq 0, \qquad w_{ij} > 0 \Leftrightarrow (x_i, x_j) \in E$$

Université de Toulouse



Cet exposé se propose de présenter des méthodes d'analyse de structures se présentant sous la forme de **données relationnelles** entre individus. Ici, on supposera les données représentées par un **graphe pondéré non orienté** : on notera donc G **un graphe**

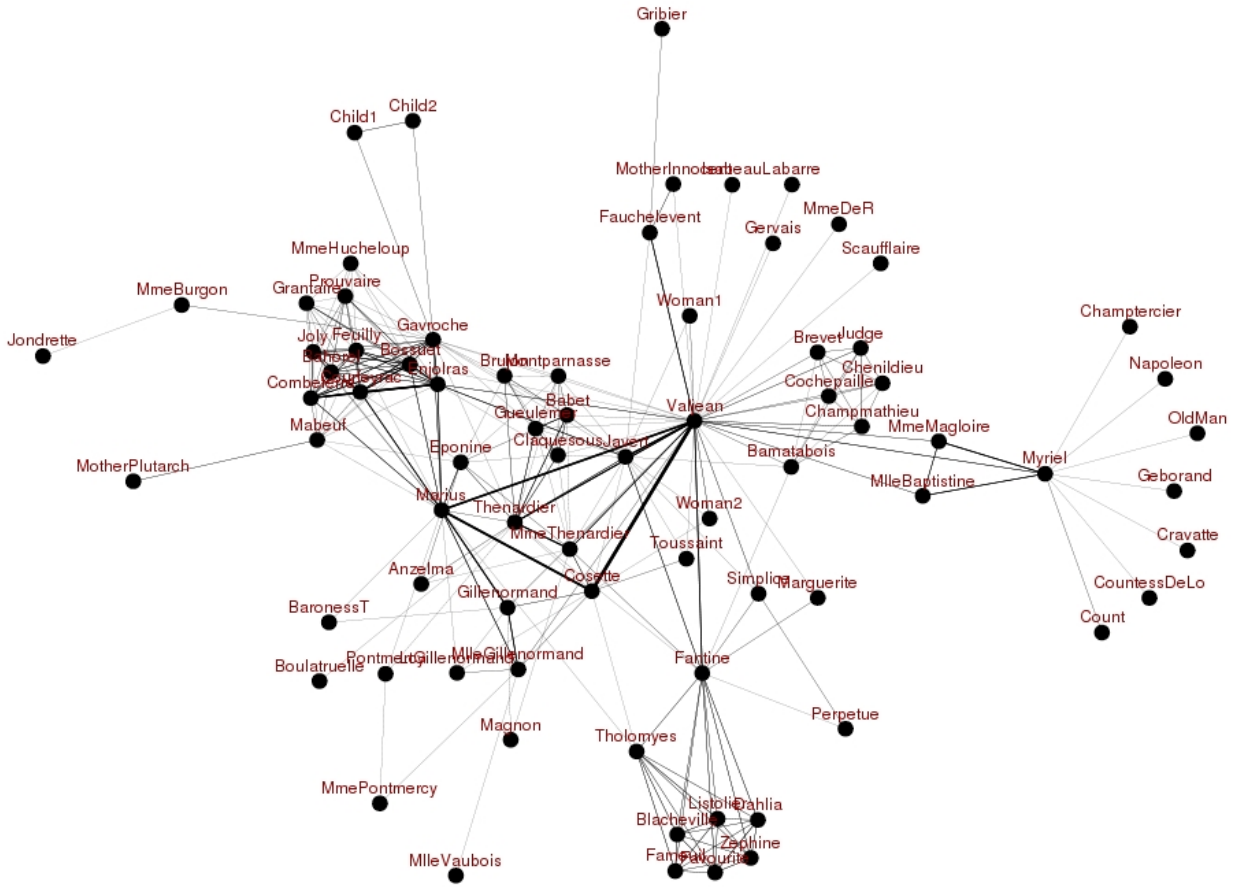
- de **sommets** $V = \{x_1, \ldots, x_n\}$ (et de taille n)
- dont l'ensemble des **arêtes** est noté E. E est donc un sous-ensemble de $V \times V$
- dont les arêtes sont pondérées par la **matrice de poids** W telle que

$$\forall\, i, j = 1, \ldots, n,\ w_{ii} = 0, \qquad w_{ij} = w_{ji} \geq 0, \qquad w_{ij} > 0 \Leftrightarrow (x_i, x_j) \in E$$

Remarque : Dans un **graphe non pondéré**, on convient que $w_{ij} \in \{0; 1\}$.

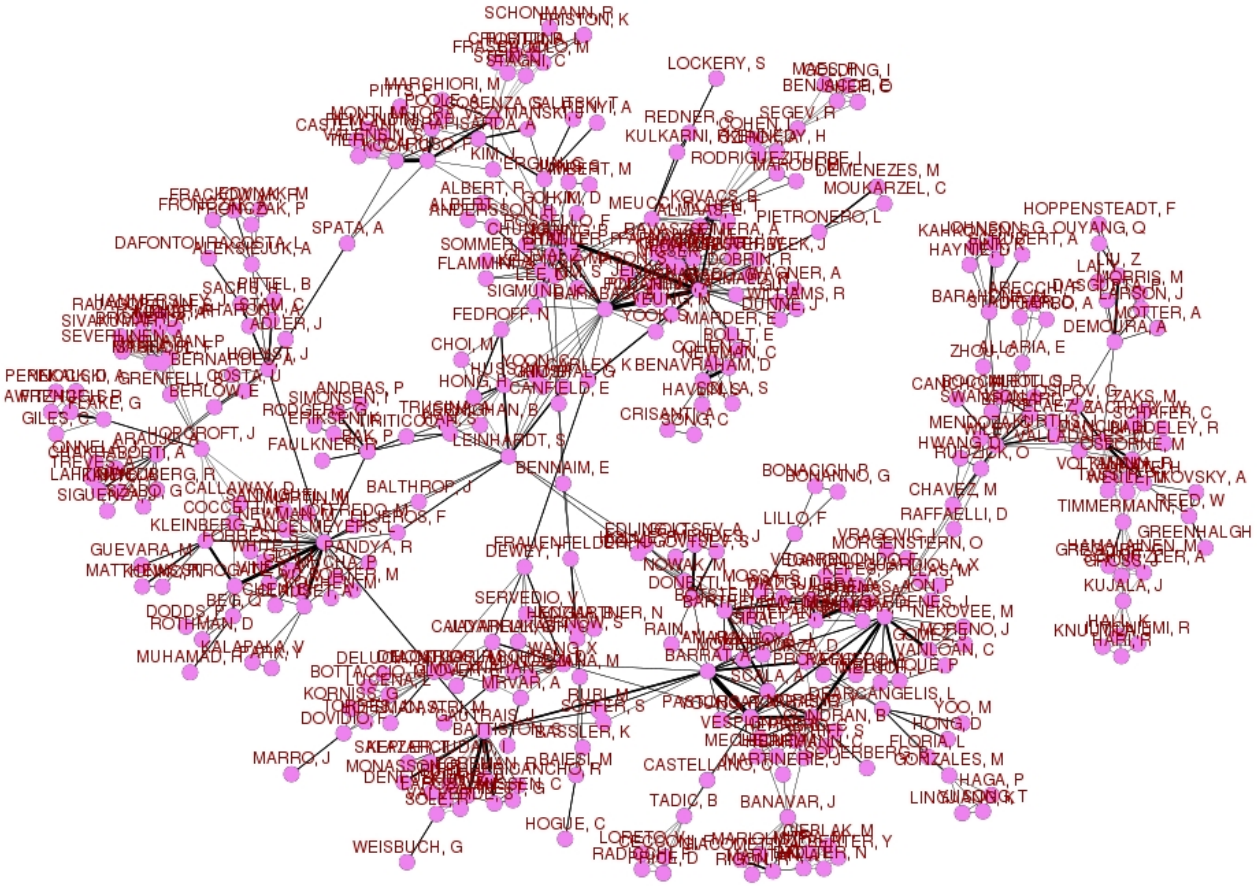
Quelques exemples de graphes I

“Les misérables” (V. Hugo) : 77 sommets, 254 arêtes.



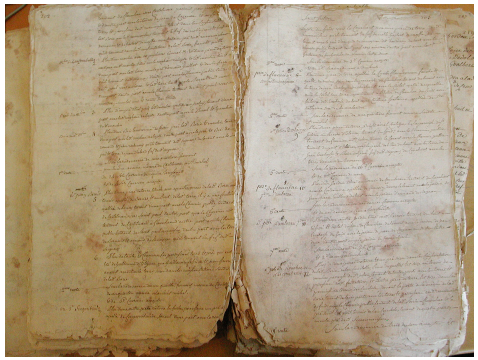
Quelques exemples de graphes II

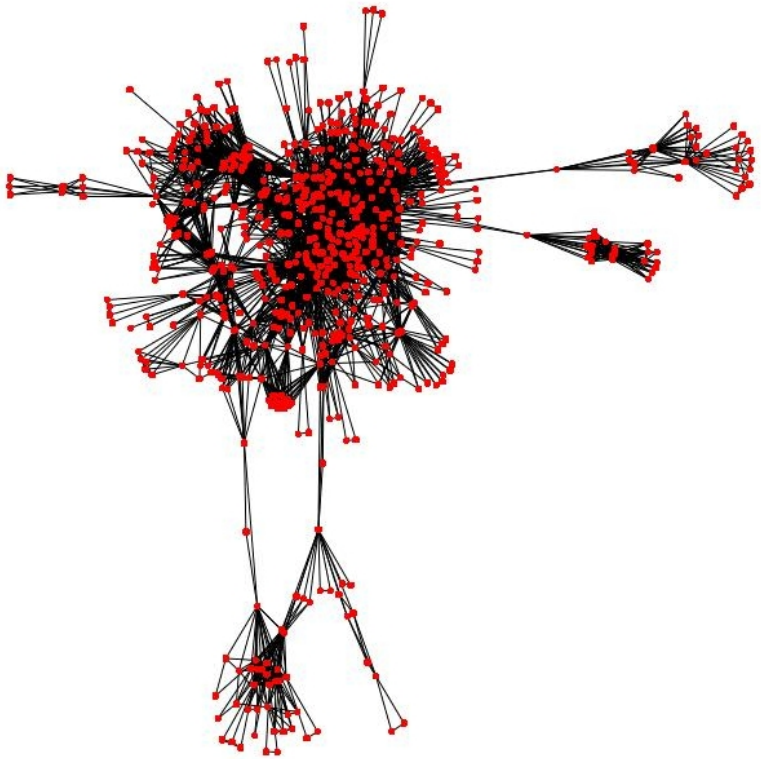
Réseau de collaborations scientifiques extrait de **[Newman, 2006]** :
379 sommets, 914 arêtes.



Quelques exemples de graphes III

Réseau issu d'un grand corpus médiéval extrait de **[Boulet et al., 2008]** : 615 sommets, 4 193 arêtes.

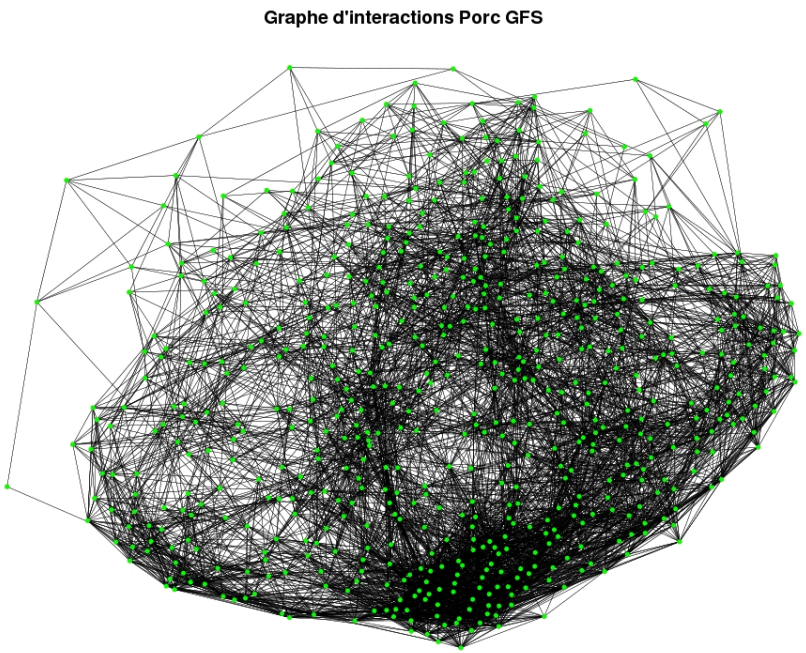


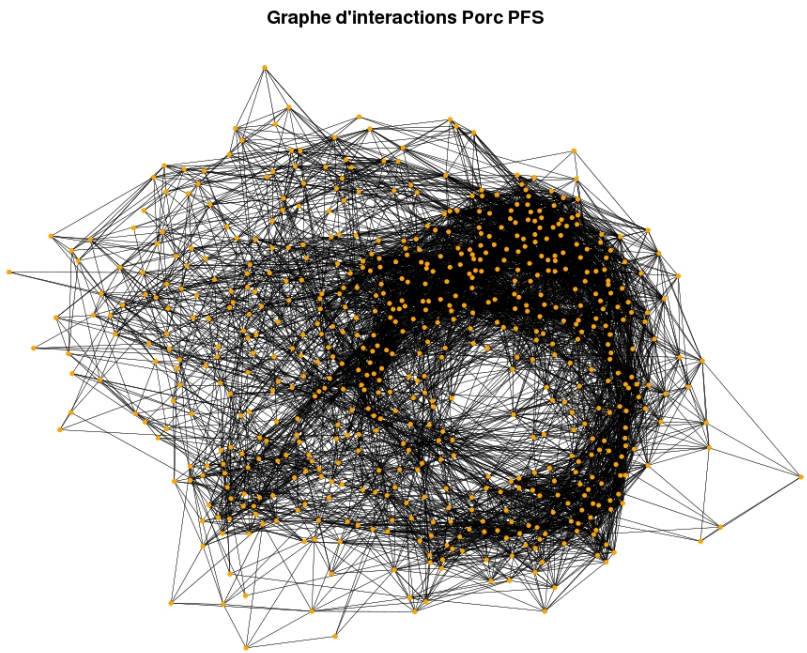


Université de Toulouse

Quelques exemples de graphes IV

Réseaux d'interactions génétiques (collaboration en cours avec Magali San Cristobal et Gwenola Tosser-Klopp de l'INRA de Castanet) : 622 sommets, 13 710 et 16 086 arêtes.





Université de Toulouse

Effet petit monde

Ces graphes ont en commun d'être caractérisés par :

- une **faible densité** : $\mathcal{D} = \frac{\sharp\{w_{ij}>0\}}{n(n-1)} < 10$ %;
- une **forte connectivité locale** :
  $C = \frac{1}{n}\sum_{i=1}^{n} \mathcal{D}(\text{Graphe des voisins de } x_i)$.

Effet petit monde

Ces graphes ont en commun d'être caractérisés par :

- une **faible densité** : $\mathcal{D} = \frac{\sharp\{w_{ij}>0\}}{n(n-1)} < 10\ \%$;
- une **forte connectivité locale** :
  $\mathcal{C} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{D}(\text{Graphe des voisins de } x_i)$.

Graphe	$\mathcal{D}$	$\mathcal{C}$
Misérables	8,7 %	49,9 %
NetScience	1,3 %	22,1 %
Médiéval	2,2 %	77,0 %
Porc GFS	3,5 %	47,0 %
Porc PFS	4,2 %	47,8 %

Université de Toulouse

Effet petit monde

Ces graphes ont en commun d'être caractérisés par :

- une **faible densité** : $\mathcal{D} = \frac{\sharp\{w_{ij}>0\}}{n(n-1)} < 10\ \%$;
- une **forte connectivité locale** :
  $\mathcal{C} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{D}(\text{Graphe des voisins de } x_i)$.

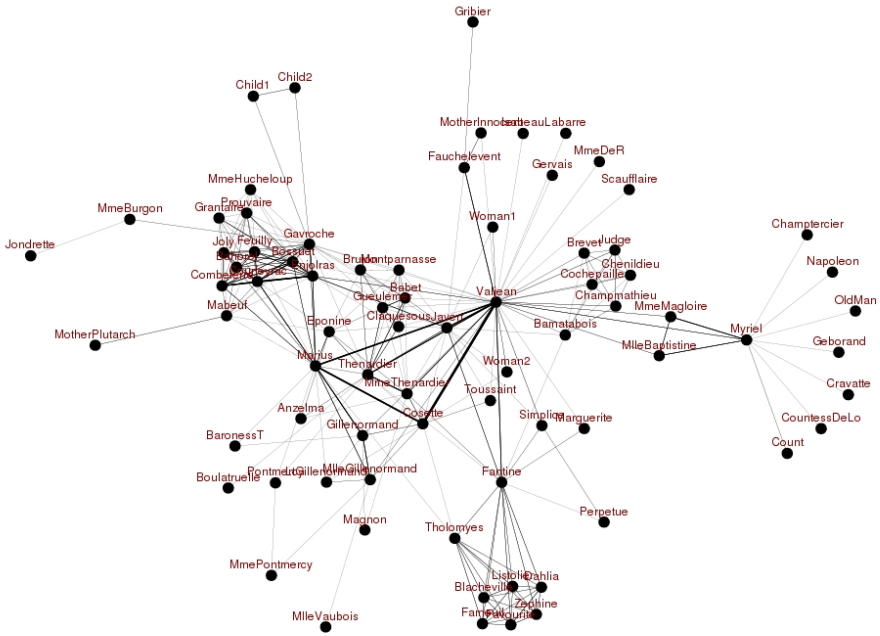
Graphe	$\mathcal{D}$	$\mathcal{C}$
Misérables	8,7 %	49,9 %
NetScience	1,3 %	22,1 %
Médiéval	2,2 %	77,0 %
Porc GFS	3,5 %	47,0 %
Porc PFS	4,2 %	47,8 %

$\Rightarrow$ Méthodes de **classification de sommets** pour retrouver des groupes homogènes faiblement connectés entre eux. Ces méthodes ont pour objectifs :

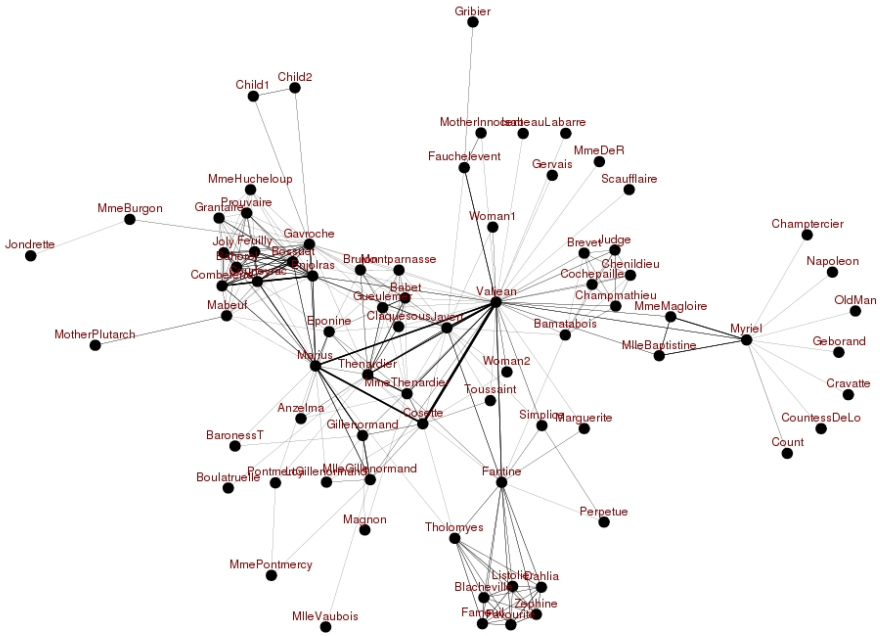
- de **simplifier la structure** de grands graphes ;
- de permettre une **représentation facilement interprétable** de ces graphes.

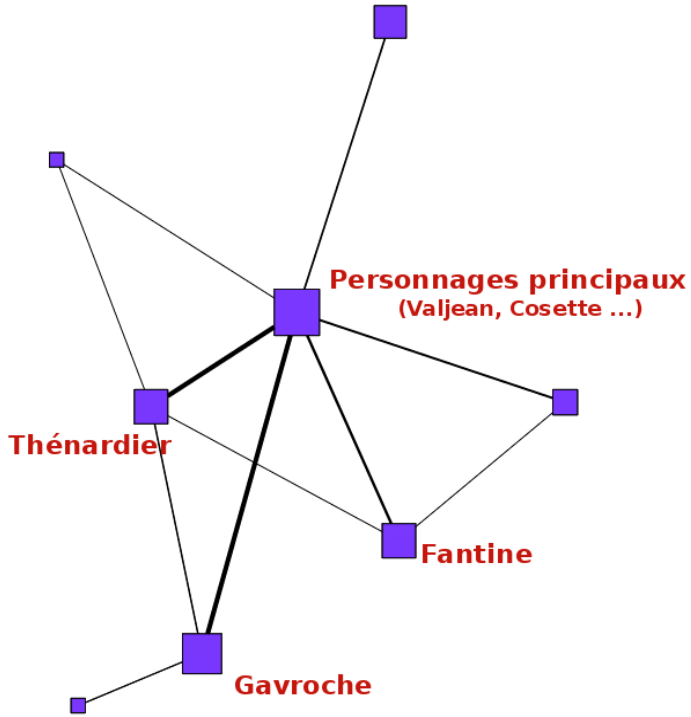
Université de Toulouse

Approche classification puis représentation



Approche classification puis représentation





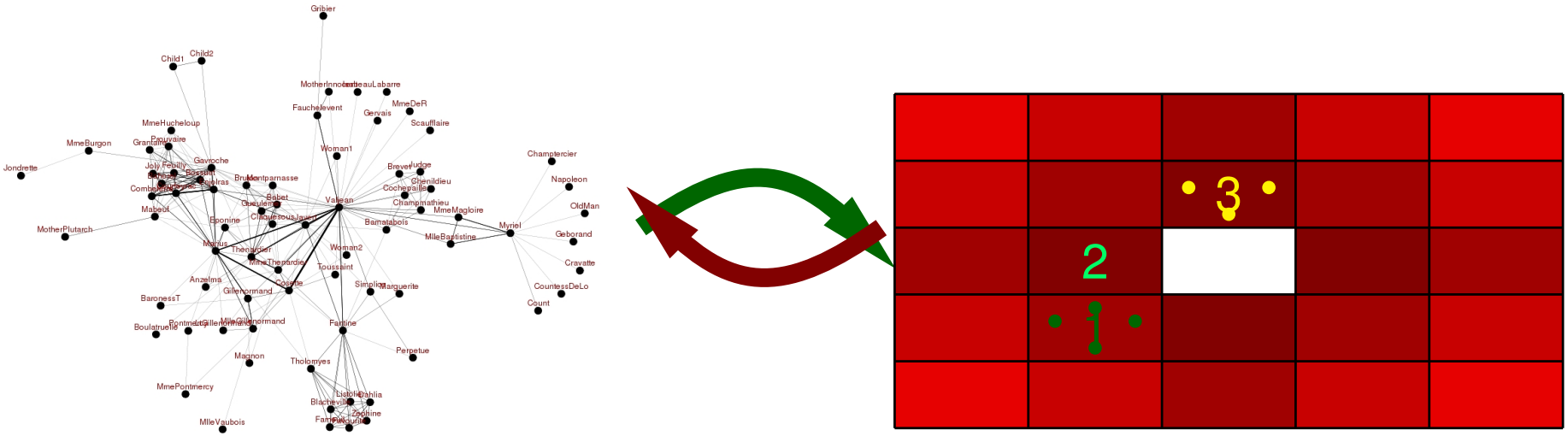
Approche “2 en 1” : organisation sur une grille

Au lieu de pratiquer une classification des sommets, on peut utiliser des algorithmes de type Carte de Kohonen pour **classer les sommets sur une grille organisée selon une topologie a priori** :

Université de Toulouse

Approche “2 en 1” : organisation sur une grille

Au lieu de pratiquer une classification des sommets, on peut utiliser des algorithmes de type Carte de Kohonen pour **classer les sommets sur une grille organisée selon une topologie a priori** :

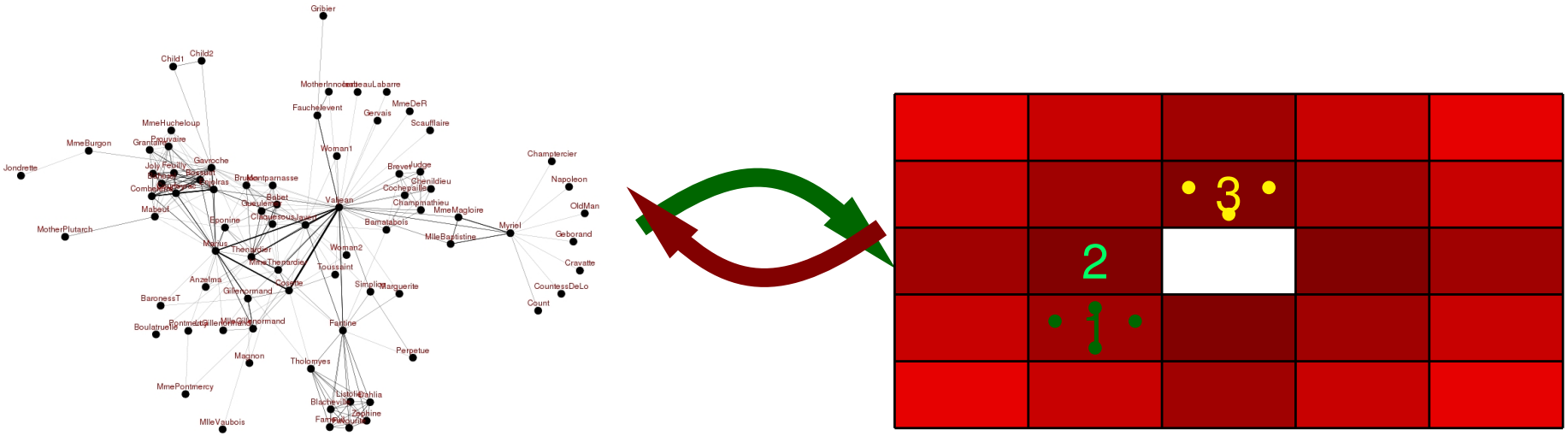


Les sommets doivent être classés de manière à ce que les sommets de la classe 1 soient plus proches dans le graphe de ceux de la classe 2 que de ceux de la classe 3.

Université de Toulouse

Approche “2 en 1” : organisation sur une grille

Au lieu de pratiquer une classification des sommets, on peut utiliser des algorithmes de type Carte de Kohonen pour **classer les sommets sur une grille organisée selon une topologie a priori** :



Les sommets doivent être classés de manière à ce que les sommets de la classe 1 soient plus proches dans le graphe de ceux de la classe 2 que de ceux de la classe 3.

⇒ **Classification** et **représentation** par placement sur une grille.

Approche 1 : Batch kernel SOM [Boulet et al., 2008]

Principe : Utiliser un **noyau** entre sommets du graphe pour jouer le rôle de produit scalaire. Une fois cette métrique définie, appliquer l'algorithme "Self-Organizing map" habituel de $\mathbb{R}^n$.

Université de Toulouse

Approche 1 : Batch kernel SOM [Boulet et al., 2008]

Principe : Utiliser un **noyau** entre sommets du graphe pour jouer le rôle de produit scalaire. Une fois cette métrique définie, appliquer l'algorithme "Self-Organizing map" habituel de $\mathbb{R}^n$.

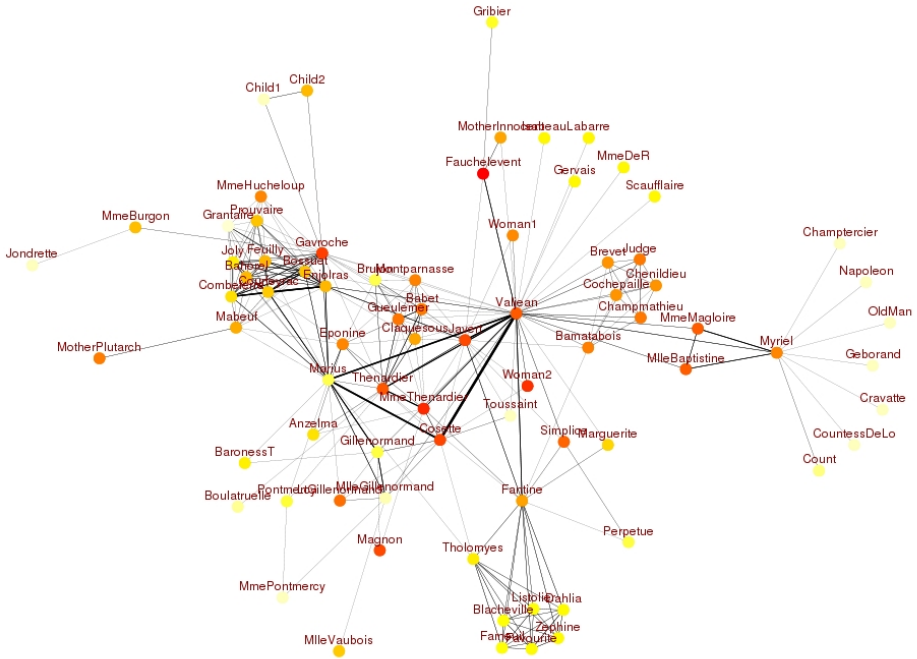
Quel noyau ? Une **régularisation du Laplacien** du graphe, $r(L)$ (avec $r(l) = e^{-\beta l}$, $r(l) = l^+$, ...) car la métrique induite par le Laplacien est en relation avec le problème de **minimisation du nombre d'arêtes entre les classes** ("graph cut optimization").

Université de Toulouse

Approche 1 : Batch kernel SOM [Boulet et al., 2008]

Principe : Utiliser un **noyau** entre sommets du graphe pour jouer le rôle de produit scalaire. Une fois cette métrique définie, appliquer l'algorithme "Self-Organizing map" habituel de $\mathbb{R}^n$.

Quel noyau ? Une **régularisation du Laplacien** du graphe, $r(L)$ (avec $r(l) = e^{-\beta l}$, $r(l) = l^+$, ...) car la métrique induite par le Laplacien est en relation avec le problème de **minimisation du nombre d'arêtes entre les classes** ("graph cut optimization").



Approche 2 : Modularité [Rossi and Villa, 2008]

Optimiser un critère qui coupe “plus facilement” les arêtes reliées aux sommets à fort degré :

$$Q = \frac{1}{2m} \sum_{i,j=1}^{n} \left(w_{ij} - \frac{d_i d_j}{2m} \right) \mathbb{I}_{f(x_i)=f(x_j)}$$

où $d_i = \sum_{j \neq i} w_{ij}$ et $m = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}$
(cf **[Newman and Girvan, 2004]** pour plus de détails)

Approche 2 : Modularité [Rossi and Villa, 2008]

Optimiser un critère qui coupe “plus facilement” les arêtes reliées aux sommets à fort degré :

$$Q = \frac{1}{2m} \sum_{i,j=1}^{n} \left(w_{ij} - \frac{d_i d_j}{2m} \right) \mathbb{I}_{f(x_i)=f(x_j)}$$

où $d_i = \sum_{j \neq i} w_{ij}$ et $m = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}$
(cf **[Newman and Girvan, 2004]** pour plus de détails)
ou sa version organisée :

$$Q_S = \frac{1}{2m} \sum_{i \neq j} \left(w_{ij} - \frac{d_i d_j}{2m} \right) h(d(f(x_i), f(x_j))).$$

Université de Toulouse

Approche 2 : Modularité [Rossi and Villa, 2008]

Optimiser un critère qui coupe “plus facilement” les arêtes reliées aux sommets à fort degré :

$$Q = \frac{1}{2m} \sum_{i,j=1}^{n} \left(w_{ij} - \frac{d_i d_j}{2m} \right) \mathbb{I}_{f(x_i)=f(x_j)}$$

où $d_i = \sum_{j \neq i} w_{ij}$ et $m = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}$
(cf **[Newman and Girvan, 2004]** pour plus de détails)
ou sa version organisée :

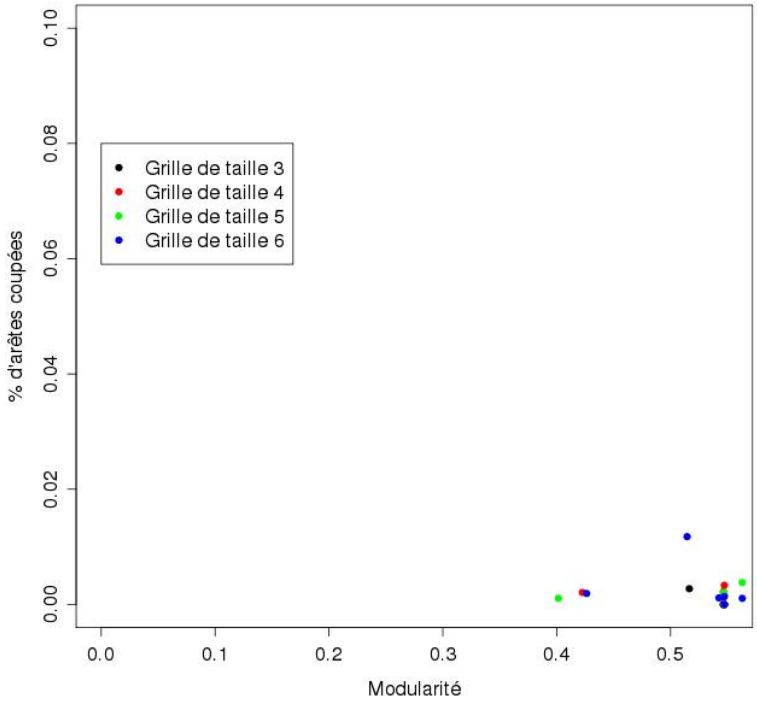
$$Q_S = \frac{1}{2m} \sum_{i \neq j} \left(w_{ij} - \frac{d_i d_j}{2m} \right) h(d(f(x_i), f(x_j))).$$

Optimisation par **recuit déterministe**.

Université de Toulouse

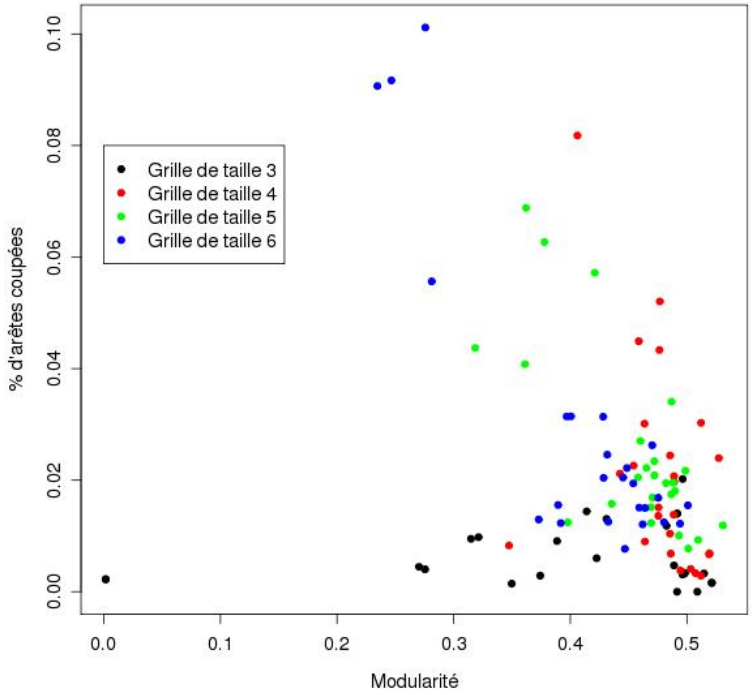
Exemple d'application au graphe "Les misérables"

Optimisation de la modularité



Kernel SOM

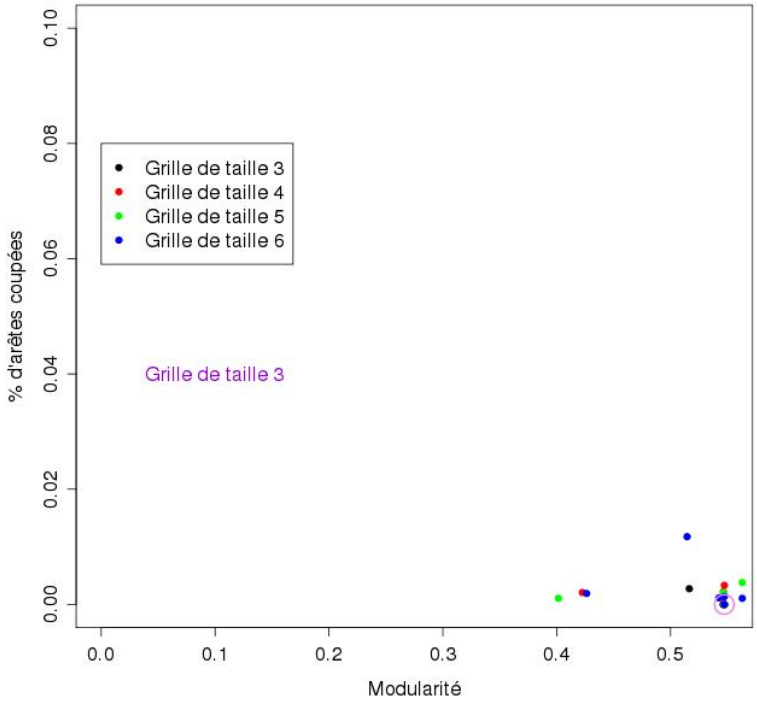
(Inverse généralisée du Laplacien)



Université de Toulouse

Exemple d'application au graphe "Les misérables"

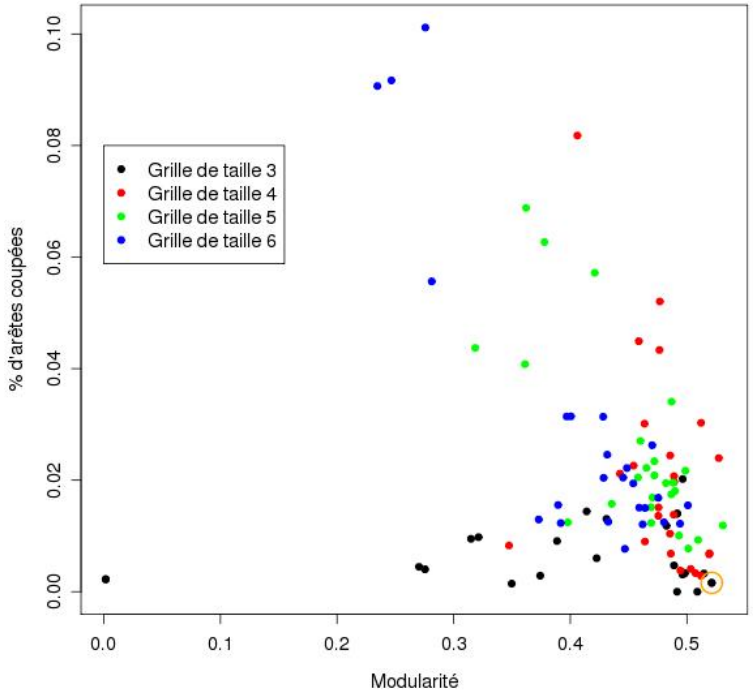
Optimisation de la modularité



0,547 / 0

Kernel SOM

(Inverse généralisée du Laplacien)

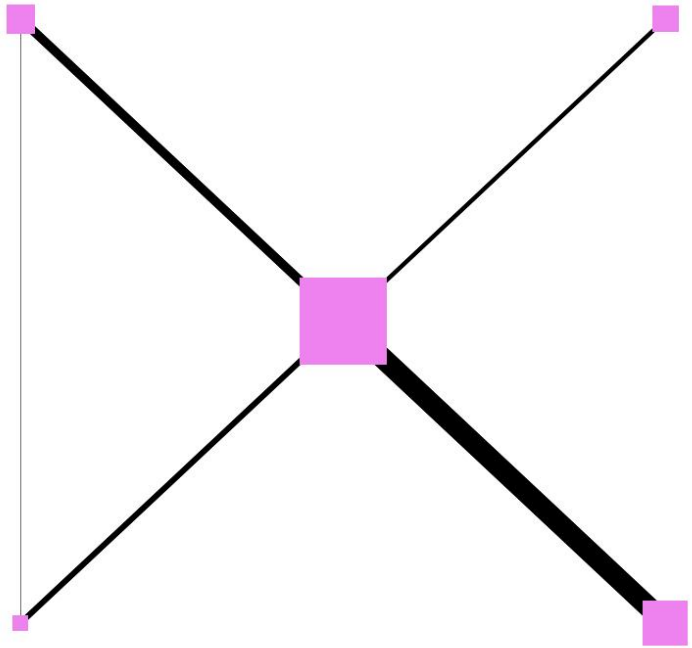


0,521 / 0,00159

Université de Toulouse

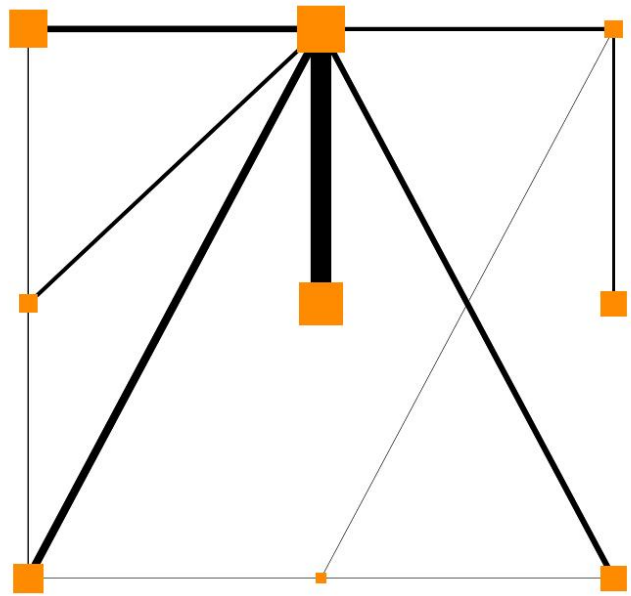
Exemple d'application au graphe "Les misérables"

Optimisation de la modularité



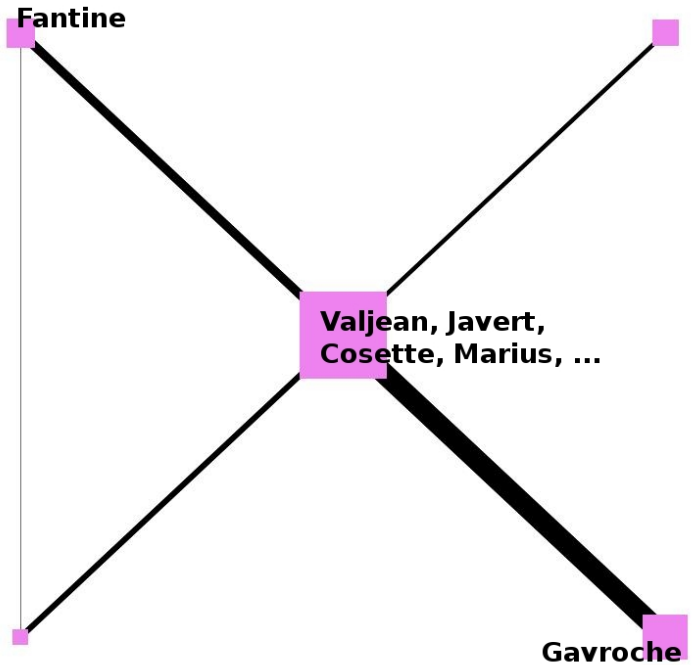
Kernel SOM

(Inverse généralisée du Laplacien)



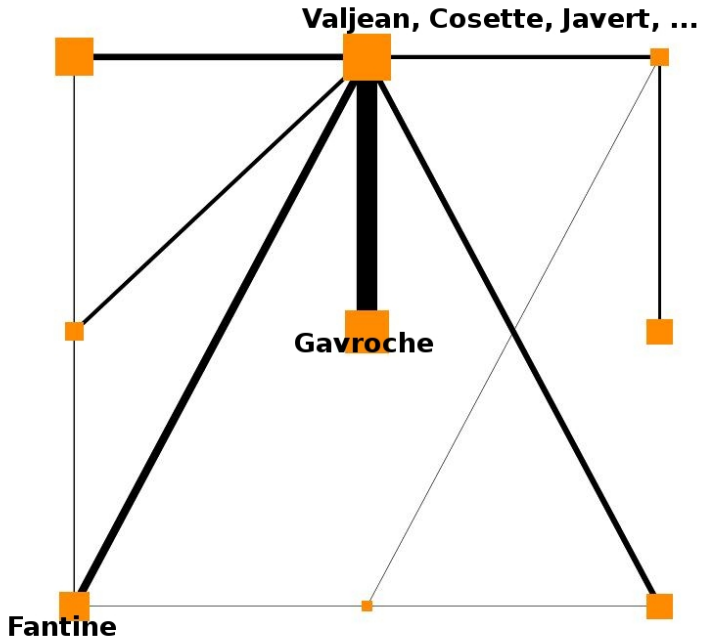
Exemple d'application au graphe "Les misérables"

Optimisation de la modularité



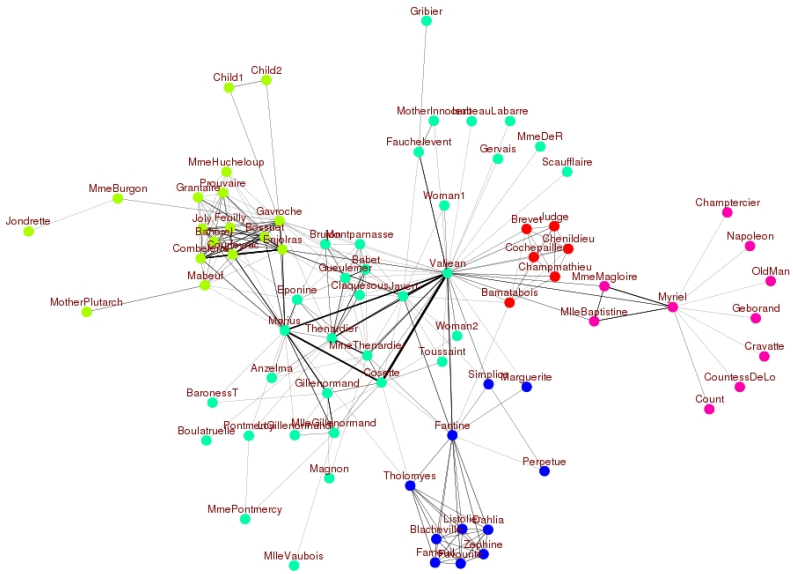
Kernel SOM

(Inverse généralisée du Laplacien)



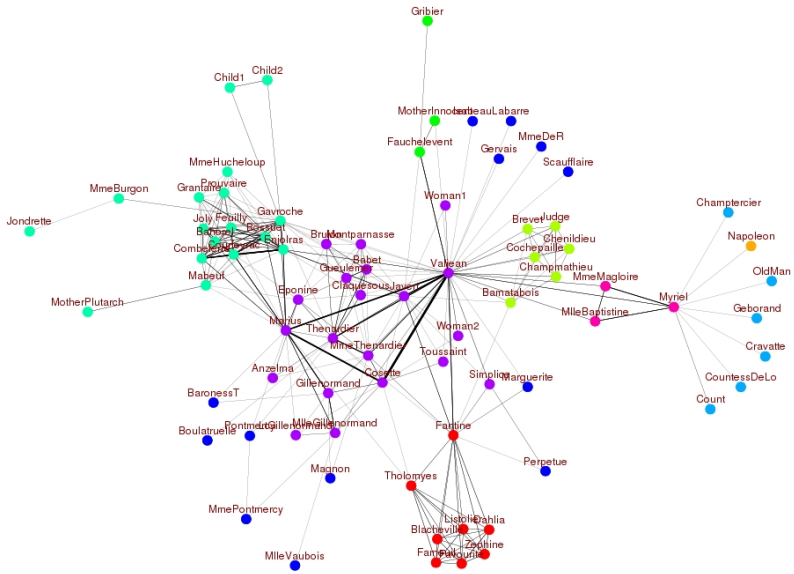
Exemple d'application au graphe "Les misérables"

Optimisation de la modularité



Kernel SOM

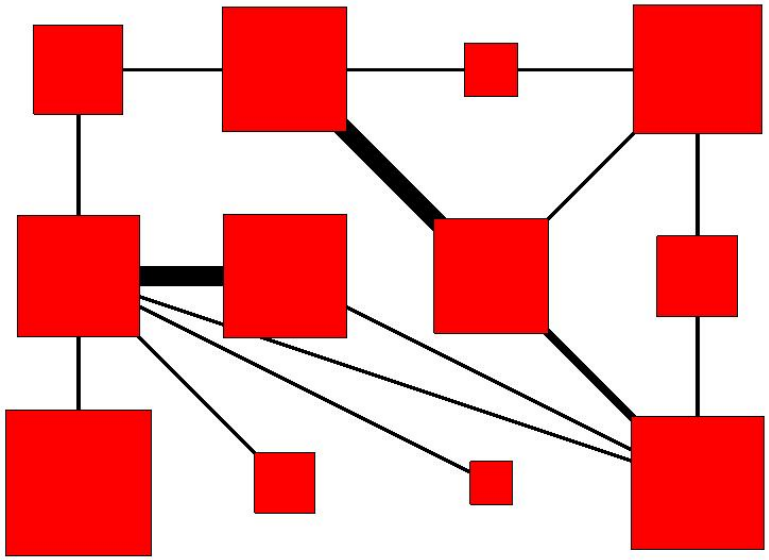
(Inverse généralisée du Laplacien)



Université de Toulouse

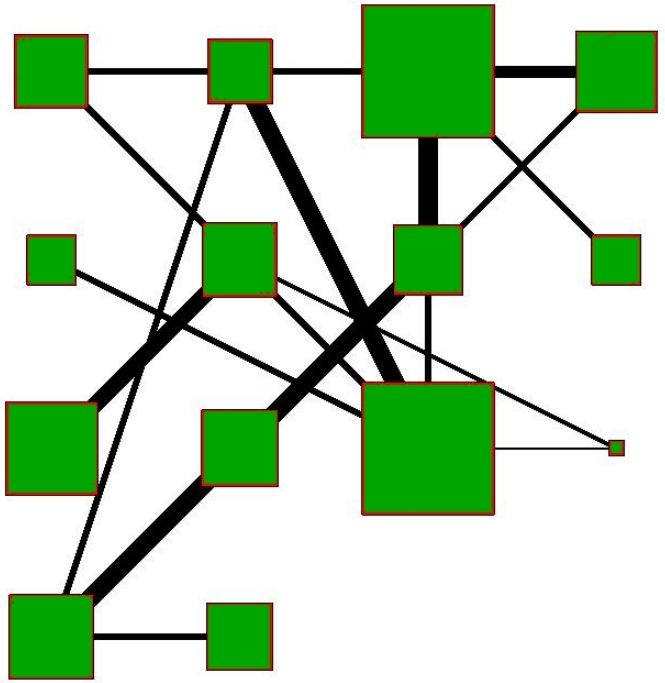
Exemple d'application au réseau de collaborations scientifiques

Recuit déterministe



Modularité = 0,836
% d'arêtes coupées sur la carte : 0

Kernel SOM (Inv. généralisée)

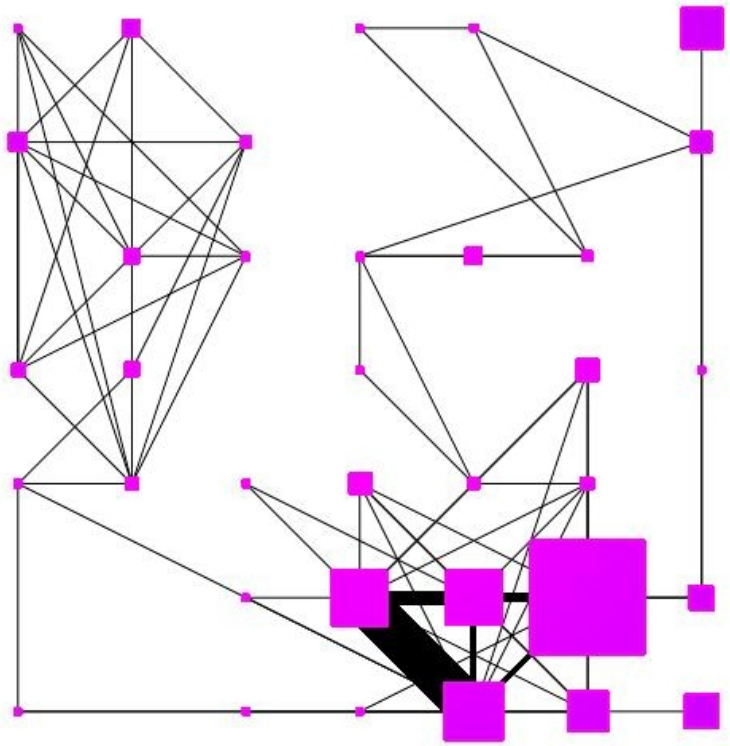


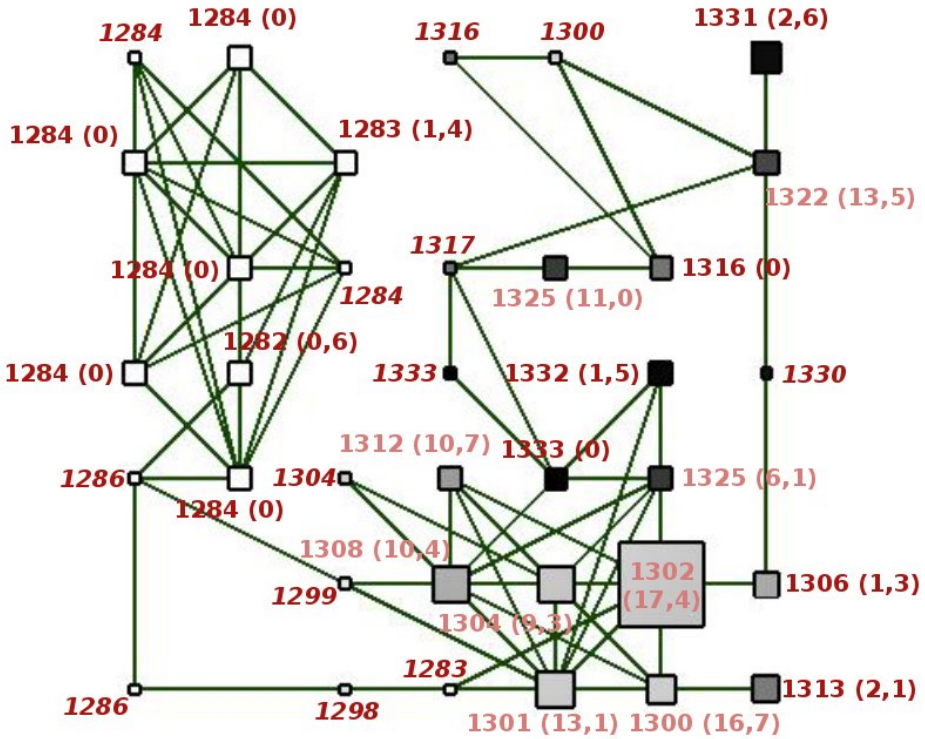
Modularité = 0,816
% d'arêtes coupées : 0,04

Université de Toulouse

Exemple d'application au réseau social médiéval

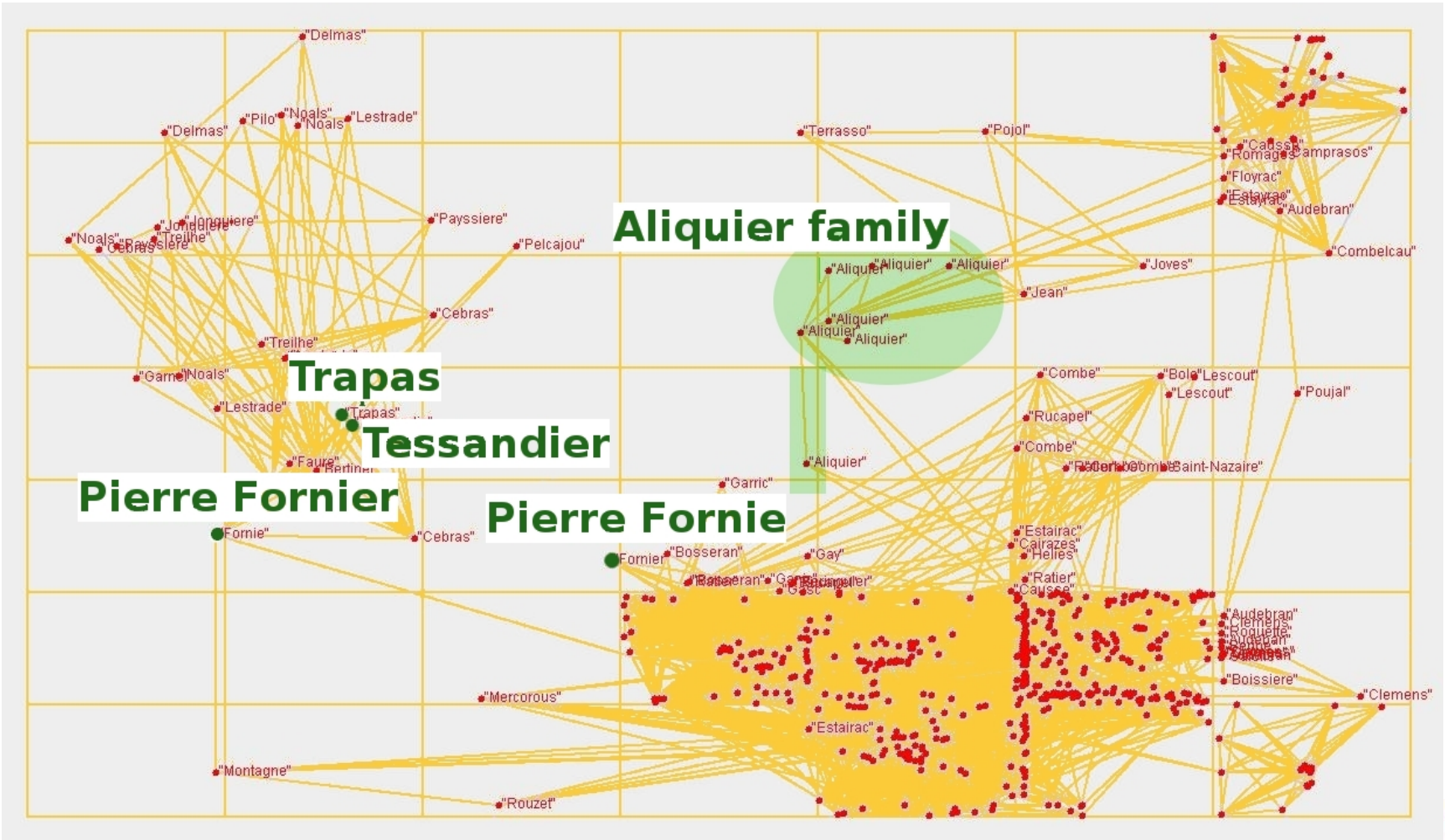
Kernel SOM avec noyau de la chaleur





Exemple d'application au réseau social médiéval

Kernel SOM avec noyau de la chaleur



Conclusion et perspectives

Conclusion : Méthodes d'organisation permettent d'obtenir une vision simplifiée d'un graphe complexe.

Conclusion et perspectives

Conclusion : Méthodes d'organisation permettent d'obtenir une vision simplifiée d'un graphe complexe.

Beaucoup de questions à aborder...

- Comparaison des approches
- Quels critères de choix et de qualité pour ces organisations ?
- ...

Université de Toulouse

Boulet, R., Jouve, B., Rossi, F., and Villa, N. (2008).

Batch kernel SOM and related laplacian methods for social network analysis.

Neurocomputing, 71(7-9):1257–1273.

Newman, M. (2006).

Finding community structure in networks using the eigenvectors of matrices.

Physical Review, E, 74(036104).

Newman, M. and Girvan, M. (2004).

Finding and evaluating community structure in networks.

Physical Review, E, 69:026113.

Rossi, F. and Villa, N. (2008).

Topologically ordered graph clustering via deterministic annealing.

In *Proceedings of ESANN 2009*, pages 529–534, Bruges, Belgium.

Université de Toulouse