Un modèle par blocs pour la classification de graphes temporels

Fabrice Rossi, Romain Guigourès et Marc Boullé

SAMM (Université Paris 1) et Orange Labs (Lannion)

Séminaire de statistiques du Cnam

Contexte et objectifs

Grands graphes temporels

- réseaux d'interactions sociales : emails, appels téléphoniques, collaborations scientifiques
- réseaux informationnels : citations entre documents (brevets, articles, textes de loi), liens entre site web
- réseaux d'usage : flux d'objets d'un site à un autre

Analyse exploratoire

- résumer un grand graphe
- chercher des schémas d'interaction (communautés, *hub* et *authorities*, structure quasi-bipartie, etc.)

Contexte et objectifs

Grands graphes temporels

- réseaux d'interactions sociales : emails, appels téléphoniques, collaborations scientifiques
- réseaux informationnels : citations entre documents (brevets, articles, textes de loi), liens entre site web
- réseaux d'usage : flux d'objets d'un site à un autre

Analyse exploratoire

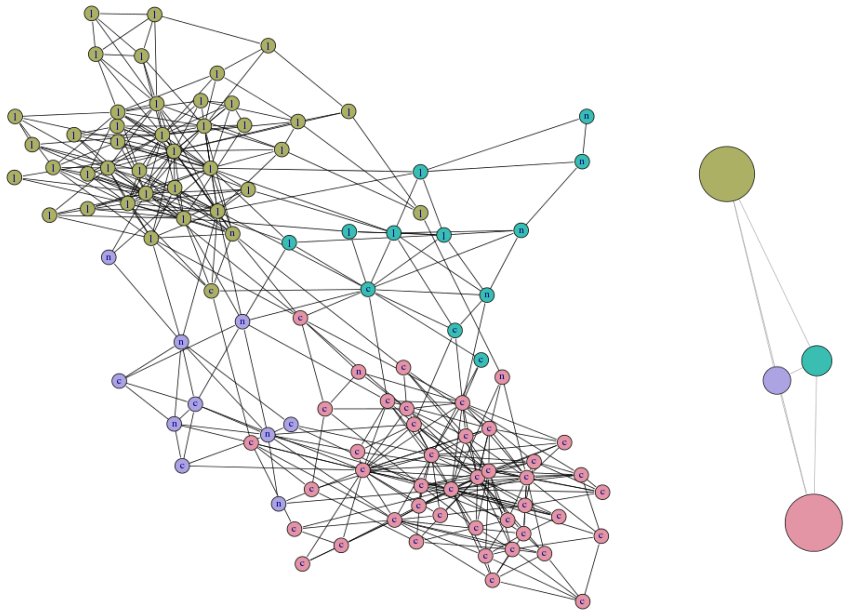
- résumer un grand graphe
- chercher des schémas d'interaction (communautés, *hub* et *authorities*, structure quasi-bipartie, etc.)
- solution par classification des sommets

Intérêts exploratoires

Graphe image

Graphe induit par une partition :

- un sommet pour chaque classe
- liens entre sommets : agrégation des liens entre les sommets sous-jacents



Identification de schémas

- *hub* : classes dont les sommets sont connectés à toutes les autres classes
- *bipartie* : classes avec des connexions externes mais pas internes

Point de vue probabiliste

Graphe « aléatoire »

- variable aléatoire W à valeurs dans l'ensemble des matrices d'adjacences ($\mathbb{R}^{+N\times N}$ ou $\{0,1\}^{N\times N}$, par exemple)
- le graphe est une réalisation de W
- graphe sans structure (Erdős–Rényi) : arêtes indépendantes, $P(W_{ij}) = \lambda$.

Principe de modélisation

- choisir une forme paramétrique simple pour $P(W)$
- adapter les paramètres pour rendre le graphe observé vraisemblable au sens de cette forme
- difficulté : structure $\Rightarrow$ dépendance entre W_{ij} et W_{ik}

Maximisation de modularité

Girvan & Newman, 2004

Principe

- graphe non orienté pondéré (poids W_{ij}, degrés $k_i = \sum_j W_{ij}$, poids total m)
- maximisation sur $(C_k)_{1 \leq k \leq C}$ de

$$Q(C) = \frac{1}{2m} \sum_{k=1}^{C} \sum_{i \in C_k, j \in C_k} \left(W_{ij} - \frac{k_i k_j}{2m} \right)$$

Modèle explicatif de $P(W)$

$$P(W_{ij} = 1 | i \in C_i, j \in C_j) =$$

	C_1	C_2	$\cdots$	C_C
C_1	$\gg \frac{k_i k_j}{2m}$	$\ll \frac{k_i k_j}{2m}$	$\cdots$	$\ll \frac{k_i k_j}{2m}$
C_2	$\ll$	$\gg$	$\cdots$	$\ll$
	$\cdots\cdots\cdots\cdots\cdots$			
C_k	$\ll$	$\ll$	$\cdots$	$\gg$

Stochastic Block-Model

Nowicki & Snijders, 2001

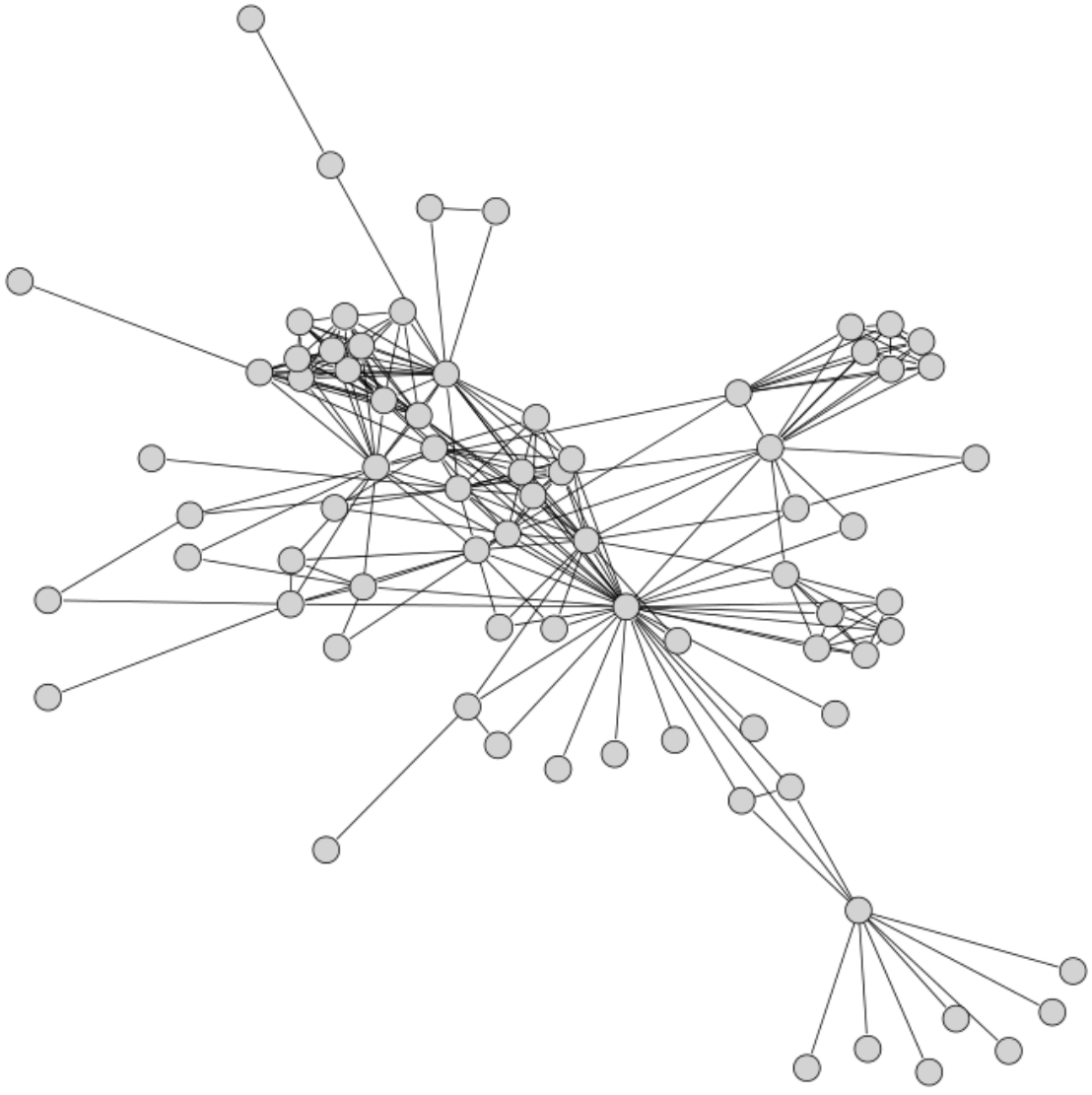
Principe

- graphe binaire $W_{ij} \in \{0, 1\}$
- modèle génératif : sachant les classes $(C_i)_i$ des sommets, les W_{ij} sont indépendants et de distribution une Bernoulli de paramètre γ_{C_i, C_j}

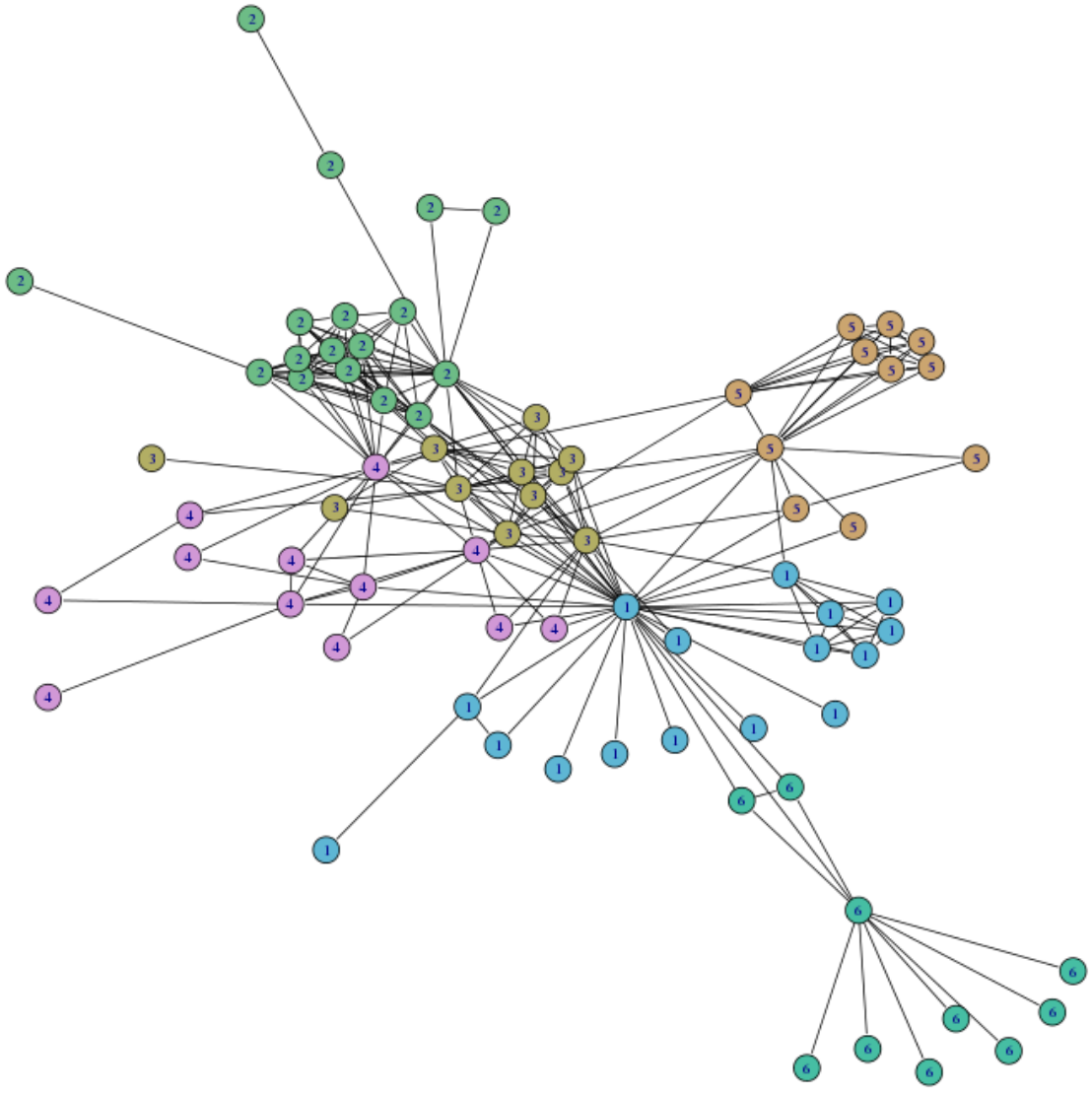
Modèle génératif

$$P(W_{ij} = 1 | i \in C_i, j \in C_j) = \begin{array}{c|cccc} & C_1 & C_2 & \cdots & C_C \\ \hline C_1 & \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1C} \\ C_2 & \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2C} \\ & \cdots & \cdots & \cdots & \cdots \\ C_k & \gamma_{k1} & \gamma_{k2} & \cdots & \gamma_{CC} \end{array}$$

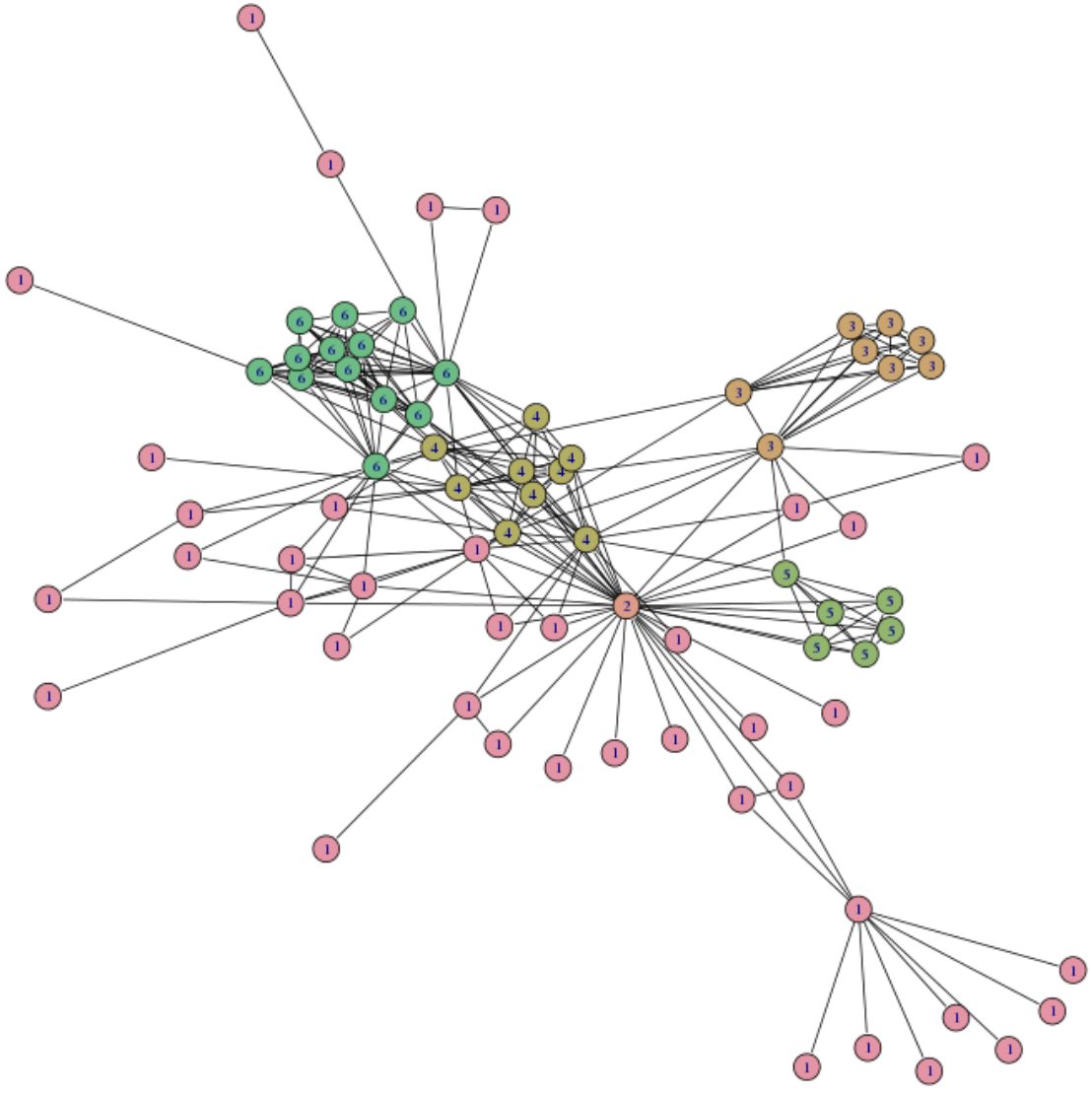
Exemple de traitements



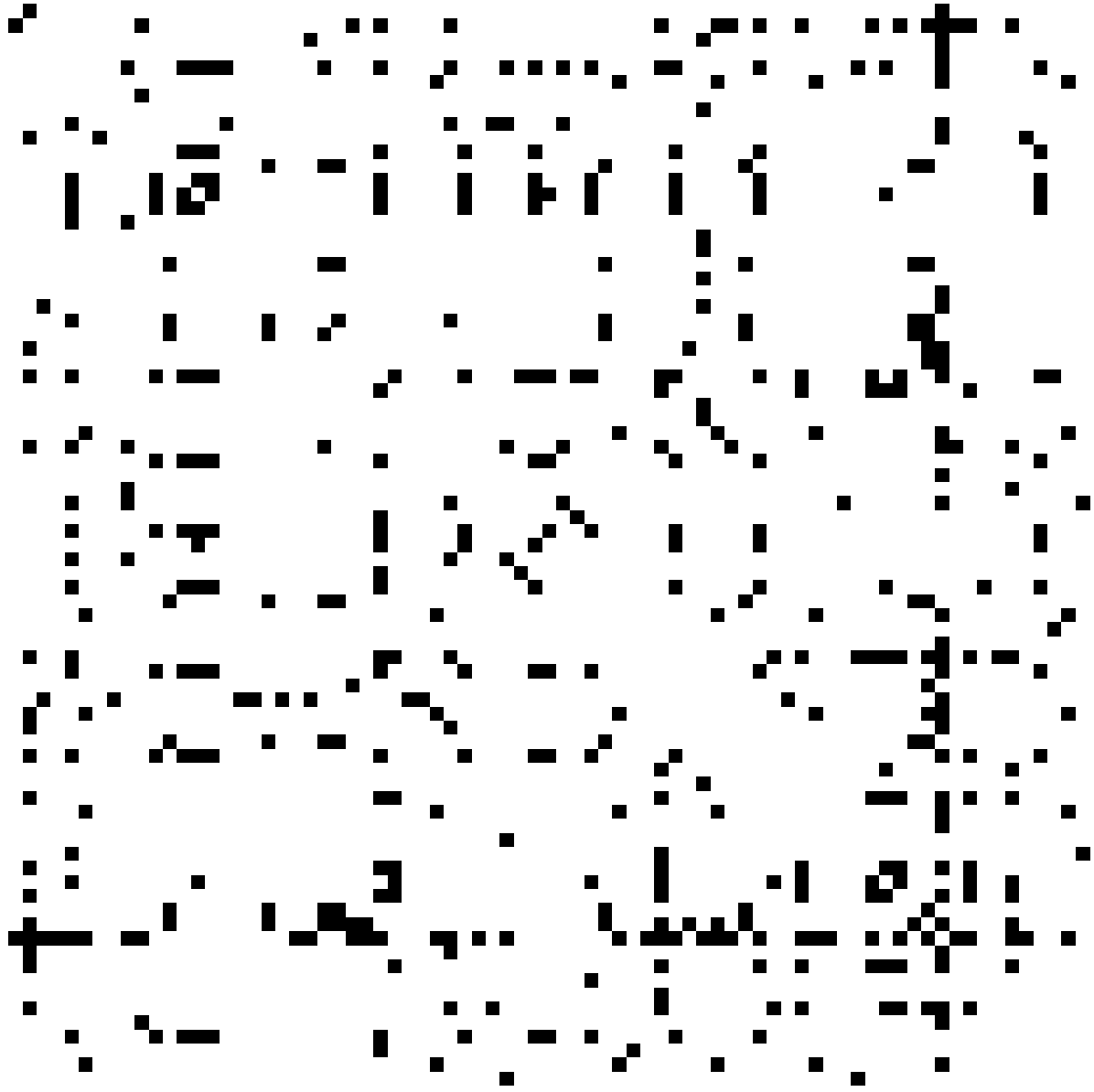
Maximum de modularité



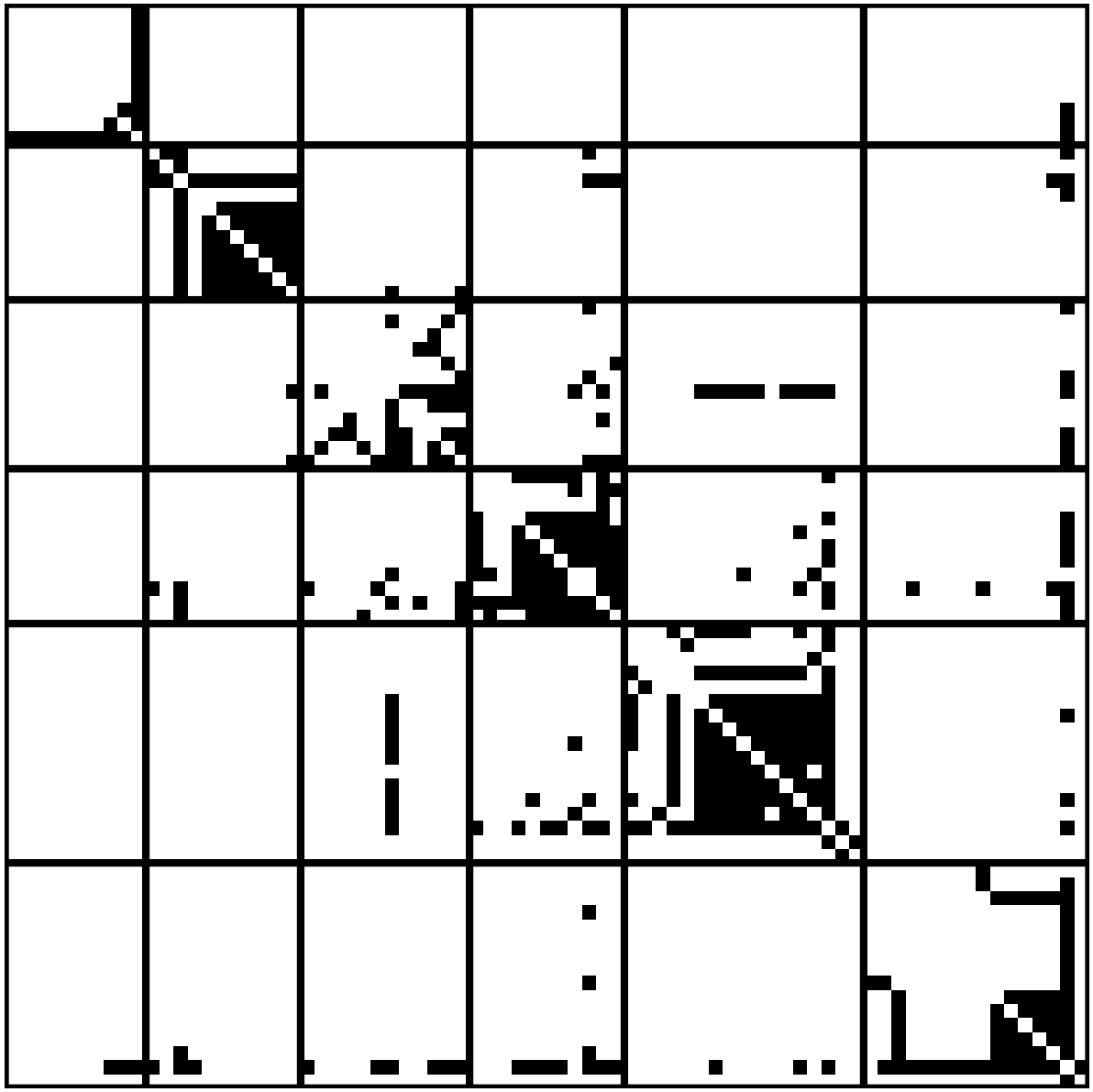
Stochastic block-model



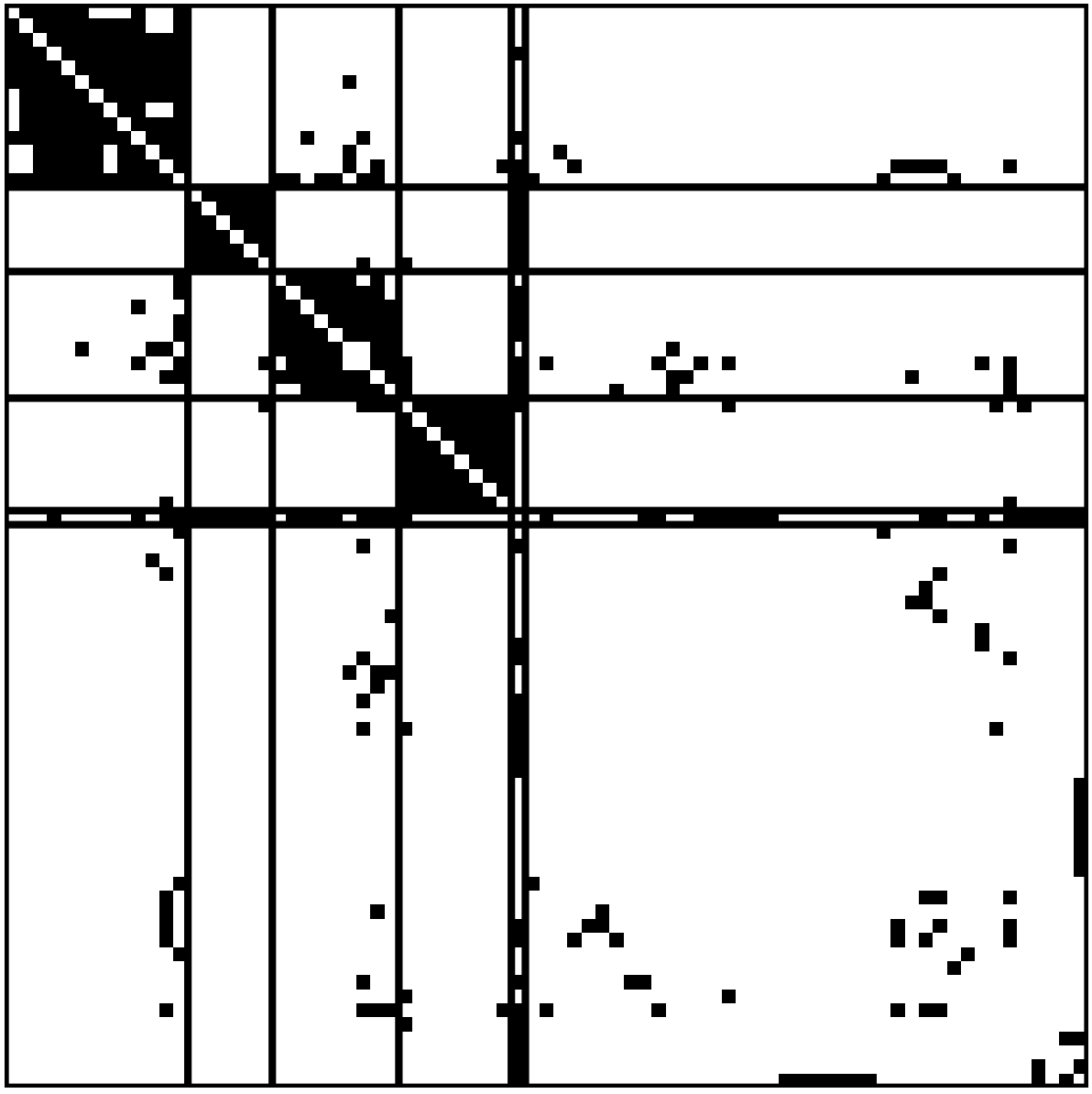
Matrice d'adjacence



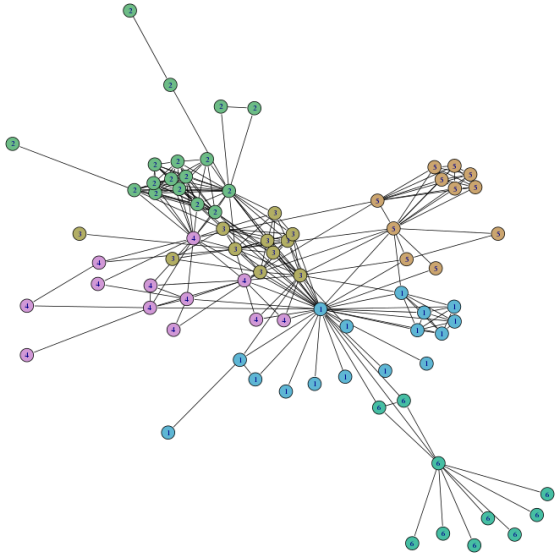
Maximum de modularité

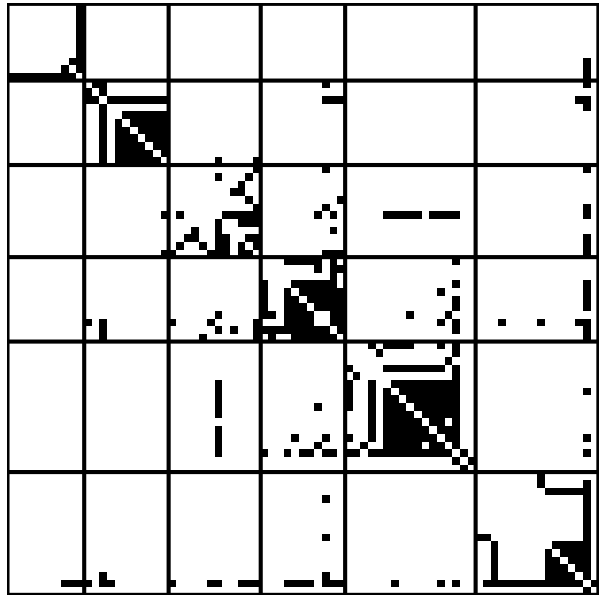


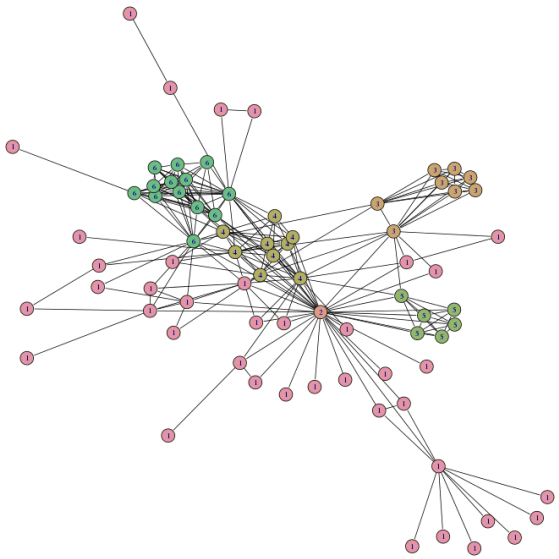
Stochastic block-model

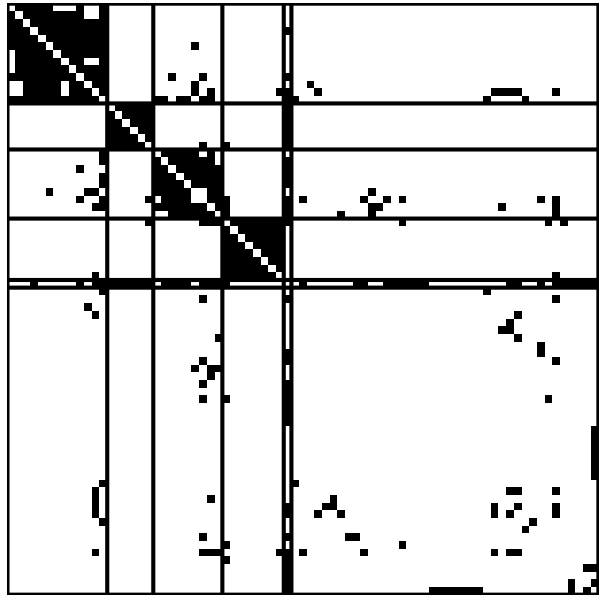


Comparaison









Comparaison

Classification de graphes

- Modularité :
  - \+ algorithmes rapides
  - \+ mélange de degrés possible dans une classe
  - \- s'intéresse seulement à la diagonale
  - \- homogénéité moyenne, même sur la diagonale
- SBM :
  - \+ modèle très riche
  - +/- classes homogènes au sens des degrés
  - \- algorithmes lents

Notre objectif

- Graphes temporels : arêtes avec une date
- Grands graphes
- Structures en bloc arbitraire (pas seulement la diagonale)

Graphes temporels

Modèle des données

- un ensemble $\mathcal{E}$ d'émetteurs, un ensemble $\mathcal{R}$ de récepteurs et un intervalle d'instants d'observation $\mathcal{T} = [a, b]$
- un ensemble de m triplets $G = (e, r, t)_{1 \leq i \leq m}$ éléments de $\mathcal{E} \times \mathcal{R} \times \mathcal{T}$: t est l'instant auquel e a interagi avec r (dans ce sens)
- multi-graphe orienté avec estampilles temporelles

Modèle recherché

- Une partition de $\mathcal{E}$, P_E, une partition de $\mathcal{R}$, P_R et une partition de $\mathcal{T}$ en intervalles contigus P_T, telles que la répartition des arêtes à l'intérieur de chaque blocs $C_e \times C_r \times C_t$ soit « uniforme » (sans structure)
- on « empile » des SBM

Interprétation

Évolution temporelle

- bi-classification : classes différentes pour les émetteurs et les récepteurs
- les classes des sommets sont fixes dans le temps
- chaque classe temporelle correspond à un modèle en blocs stable

	R_1	R_2
E_1	0.9	0.75
E_2	0.1	0.8
E_3	0.05	0

$[t_1, t_2]$

	R_1	R_2
E_1	0.8	0.5
E_2	0.1	0.6
E_3	0.1	0.3

$]t_2, t_3]$

	R_1	R_2
E_1	0.7	0.25
E_2	0.1	0.3
E_3	0.5	0.5

$]t_3, t_4]$

Modularité ?

- incompatible avec ce type de modèles
- classes changeantes, modèle diagonal fixé

Estimation du modèle

Difficultés

- choix d'un critère qualité et optimisation de celui-ci
- choix des paramètres : ici trois nombres de classes

Choix de modèle Bayésien

- D : données, M : modèle génératif
- $P(M|D) = \frac{P(D|M)P(M)}{P(D)}$
- maximum à posteriori

$$M^* = \arg\max_M P(D|M)P(M)$$

- on remplace le choix des paramètres par le choix d'une distribution (un à priori) sur les paramètres

Modèle génératif

Approche MODL

Principes

- distributions uniformes dans les structures choisies
- modélisation hiérarchique
- indépendance seulement au sein des niveaux de la hiérarchie

À priori sur les paramètres du modèle

- $|\mathcal{E}|$, $|\mathcal{R}|$ et m sont donnés
- $K_E \sim \mathcal{U}(\{1, \ldots, |\mathcal{E}|\})$, $K_R \sim \mathcal{U}(\{1, \ldots, |\mathcal{R}|\})$, $K_T \sim \mathcal{U}(\{1, \ldots, m\})$
- $P_E \sim \mathcal{U}(\mathcal{P}_{K_E}(\{1, \ldots, |\mathcal{E}|\}))$ et $P_R \sim \mathcal{U}(\mathcal{P}_{K_R}(\{1, \ldots, |\mathcal{R}|\}))$
- choix des effectifs de chaque bloc uniforme dans l'ensemble des répartitions de m arêtes dans les $K_E \times K_R \times K_T$ blocs
- choix uniforme indépendant de la répartition des arêtes sur les sommets (basée sur la répartition préalable des arêtes) au sein de chaque classe (sur $\mathcal{E}$ et sur $\mathcal{R}$)

Exemple

1. 10 émetteurs, 10 récepteurs, 50 messages

Exemple

1. 10 émetteurs, 10 récepteurs, 50 messages
2. 3 classes, 2 classes, 3 classes

	R_1	R_2
E_1		
E_2		
E_3		

I_1

	R_1	R_2
E_1		
E_2		
E_3		

I_2

	R_1	R_2
E_1		
E_2		
E_3		

I_3

Exemple

1. 10 émetteurs, 10 récepteurs, 50 messages
2. 3 classes, 2 classes, 3 classes

	R_1	R_2
E_1		
E_2		
E_3		

I_1

	R_1	R_2
E_1		
E_2		
E_3		

I_2

	R_1	R_2
E_1		
E_2		
E_3		

I_3

3. $E_1 = \{1, 2, 3\}$, $E_2 = \{4, 5\}$, $E_3 = \{6, 7, 8, 9, 10\}$, $R_1 = \{1, 2, 3, 4\}$ et $R_2 = \{5, 6, 7, 8, 9, 10\}$

Exemple

1. 10 émetteurs, 10 récepteurs, 50 messages
2. 3 classes, 2 classes, 3 classes

	R_1	R_2
E_1		
E_2		
E_3		

I_1

	R_1	R_2
E_1		
E_2		
E_3		

I_2

	R_1	R_2
E_1		
E_2		
E_3		

I_3

3. $E_1 = \{1, 2, 3\}$, $E_2 = \{4, 5\}$, $E_3 = \{6, 7, 8, 9, 10\}$, $R_1 = \{1, 2, 3, 4\}$ et $R_2 = \{5, 6, 7, 8, 9, 10\}$
4. Répartition des 50 messages dans les 18 blocs

	R_1	R_2
E_1	5	1
E_2	2	0
E_3	4	0

I_1

	R_1	R_2
E_1	2	2
E_2	2	5
E_3	5	5

I_2

	R_1	R_2
E_1	0	0
E_2	1	0
E_3	1	15

I_3

Exemple

4. Répartition des 50 messages dans les 18 blocs

	R_1	R_2
E_1	5	1
E_2	2	0
E_3	4	0

I_1

	R_1	R_2
E_1	2	2
E_2	2	5
E_3	5	5

I_2

	R_1	R_2
E_1	0	0
E_2	1	0
E_3	1	15

I_3

Exemple

4. Répartition des 50 messages dans les 18 blocs

	R_1	R_2
E_1	5	1
E_2	2	0
E_3	4	0

I_1

	R_1	R_2
E_1	2	2
E_2	2	5
E_3	5	5

I_2

	R_1	R_2
E_1	0	0
E_2	1	0
E_3	1	15

I_3

5. Répartition du degré sortant 10 sur $E_1 = \{1, 2, 3\}$, etc.

E_1	E_2	E_3	R_1	R_2
10	10	30	22	28

sommet de $\mathcal{E}$	1	2	3	4	5	6	7	8	9	10
degré	3	6	1	2	8	4	12	5	8	1

sommet de $\mathcal{R}$	1	2	3	4	5	6	7	8	9	10
degré	8	6	1	7	4	13	5	1	2	3

Vraisemblance

Approche MODL

Principes

- tirage global (pas arête par arête)
- identité forte : triplets numérotés (en dehors des estampilles)
- toujours uniforme et hiérarchique

Distribution des données

- choix uniforme de la répartition des m triplets dans les $K_E \times K_R \times K_T$ blocs (en respectant les effectifs)
- choix uniformes indépendants des répartitions des triplets de chaque classe sur les sommets (en respectant les contraintes de degré)
- choix uniformes indépendants de la répartition des estampilles temporelles dans chaque intervalle de temps

Exemple

6. découpage de $\{1, 2, \ldots, 50\}$ en 13 blocs de tailles 5, 1, etc. selon

	R_1	R_2
E_1	5	1
E_2	2	0
E_3	4	0

I_1

	R_1	R_2
E_1	2	2
E_2	2	5
E_3	5	5

I_2

	R_1	R_2
E_1	0	0
E_2	1	0
E_3	1	15

I_3

   par exemple $E_1 \times R_1 \times I_1 = \{1, 2, 27, 32, 50\}$, $E_1 \times R_2 \times I_1 = \{5\}$, $E_1 \times R_1 \times I_2 = \{8, 49\}$, $E_1 \times R_2 \times I_2 = \{25, 37\}$, etc.

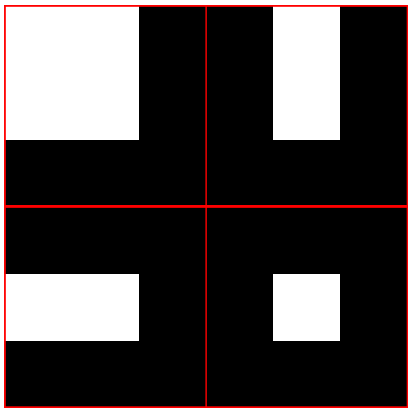
7. agrégation puis répartition sur les émetteurs (resp. récepteurs) : pour E_1, on a $\{1, 2, 5, 8, 25, 27, 32, 37, 49, 50\}$ et par exemple :

sommet de E_1	1	2	3
degré	3	6	1
identifiants	$\{5, 32, 37\}$	$\{2, 8, 25, 27, 49, 50\}$	$\{1\}$

8. ordre arbitraire des estampilles dans les I_l

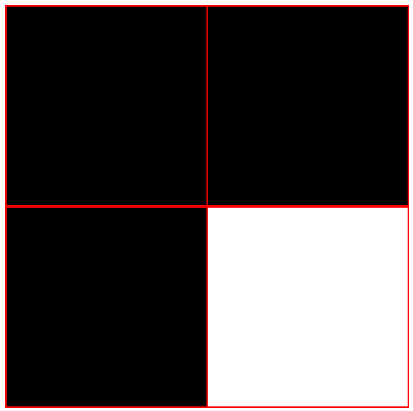
Distributions favorisées

Blocs vides



	R_1	R_2
E_1	5	7
E_2	7	8

88 597 190 167 200



	R_1	R_2
E_1	9	9
E_2	9	0

227 873 431 500

- mêmes paramètres : 2×2 classes, répartitions aléatoires des 27 arêtes
- plus de configurations à gauche, donc vraisemblance plus faible

Disributions favorisées

Tranches fines

- par ex. si $K_E = 1$, nb de configurations : $\frac{m!}{\prod_e d^{out}(e)!}$
- en général : $\frac{\prod_{l=1}^{K_E} \left(\sum_{e \in C_l^E} d^{out}(e) \right)!}{\prod_e d^{out}(e)!}$
- donc
  - le nombre de configurations diminue avec K_E
  - on a intérêt à mélanger les degrés
- même phénomène sur le temps

Régularisation

- mais un grand K_E rend chaque partition improbable
- et un grand $K_E \times K_R \times K_T$ rend chaque répartition improbable

Critère final

$$
\begin{aligned}
-\log P(D|M)P(M) = \log|\mathcal{E}| + \log|\mathcal{R}| + \log m + \underbrace{\log \mathcal{B}(|\mathcal{E}|, K_E) + \log \mathcal{B}(|\mathcal{R}|, K_R)}_{\text{partitions}} \\
+ \underbrace{\log \binom{m + K_E K_R K_T - 1}{K_E K_R K_T - 1}}_{\text{nb arêtes}} + \sum_{l=1}^{K_E} \underbrace{\log \binom{d^{out}(C_l^E) + |C_l^E| - 1}{|C_l^E| - 1}}_{\text{degrés dans } C_l^E} \\
+ \sum_{r=1}^{K_R} \underbrace{\log \binom{d^{in}(C_r^R) + |C_r^R| - 1}{|C_r^R| - 1}}_{\text{degrés dans } C_r^R} + \underbrace{\log(m!) - \sum_{l,r,s} \log(m_{l,r,s}!)}_{\text{arêtes}} \\
+ \underbrace{\sum_{l=1}^{K_E} \log(d^{out}(C_l^E)!) - \sum_{e} \log(d^{out}(e)!)}_{\text{arêtes dans les } C_E} \\
+ \underbrace{\sum_{r=1}^{K_R} \log(d^{in}(C_r^R)!) - \sum_{r} \log(d^{in}(r)!)}_{\text{arêtes dans les } C_R} + \underbrace{\sum_{s} \log(|I_s|!)}_{\text{temps}}
\end{aligned}
$$

Critère final

$$
\begin{aligned}
-\log P(D|M)P(M) = {} & \log|\mathcal{E}| + \log|\mathcal{R}| + \log m + \underbrace{\log \mathcal{B}(|\mathcal{E}|, K_E) + \log \mathcal{B}(|\mathcal{R}|, K_R)}_{\text{partitions}} \\
& + \underbrace{\log \binom{m + K_E K_R K_T - 1}{K_E K_R K_T - 1}}_{\text{nb arêtes}} + \sum_{l=1}^{K_E} \underbrace{\log \binom{d^{out}(C_l^E) + |C_l^E| - 1}{|C_l^E| - 1}}_{\text{degrés dans } C_l^E} \\
& + \sum_{r=1}^{K_R} \underbrace{\log \binom{d^{in}(C_r^R) + |C_r^R| - 1}{|C_r^R| - 1}}_{\text{degrés dans } C_r^R} + \underbrace{\log(m!) - \sum_{l,r,s} \log(m_{l,r,s}!)}_{\text{arêtes}} \\
& + \underbrace{\sum_{l=1}^{K_E} \log(d^{out}(C_l^E)!) - \sum_{e} \log(d^{out}(e)!)}_{\text{arêtes dans les } C_E} \\
& + \underbrace{\sum_{r=1}^{K_R} \log(d^{in}(C_r^R)!) - \sum_{r} \log(d^{in}(r)!)}_{\text{arêtes dans les } C_R} + \underbrace{\sum_{s} \log(|I_s|!)}_{\text{temps}}
\end{aligned}
$$

Critère final

$$
\begin{aligned}
-\log P(D|M)P(M) = {} & \log|\mathcal{E}| + \log|\mathcal{R}| + \log m + \underbrace{\log \mathcal{B}(|\mathcal{E}|, K_E) + \log \mathcal{B}(|\mathcal{R}|, K_R)}_{\text{partitions}} \\
& + \underbrace{\log \binom{m + K_E K_R K_T - 1}{K_E K_R K_T - 1}}_{\text{nb arêtes}} + \sum_{l=1}^{K_E} \underbrace{\log \binom{d^{out}(C_l^E) + |C_l^E| - 1}{|C_l^E| - 1}}_{\text{degrés dans } C_l^E} \\
& + \sum_{r=1}^{K_R} \underbrace{\log \binom{d^{in}(C_r^R) + |C_r^R| - 1}{|C_r^R| - 1}}_{\text{degrés dans } C_r^R} + \underbrace{\log(m!) - \sum_{l,r,s} \log(m_{l,r,s}!)}_{\text{arêtes}} \\
& + \underbrace{\sum_{l=1}^{K_E} \log(d^{out}(C_l^E)!) - \sum_{e} \log(d^{out}(e)!)}_{\text{arêtes dans les } C_E} \\
& + \underbrace{\sum_{r=1}^{K_R} \log(d^{in}(C_r^R)!) - \sum_{r} \log(d^{in}(r)!)}_{\text{arêtes dans les } C_R} + \underbrace{\sum_{s} \log(|I_s|!)}_{\text{temps}}
\end{aligned}
$$

Optimisation

Problème combinatoire

- espace gigantesque pour les modèles
- critère discret

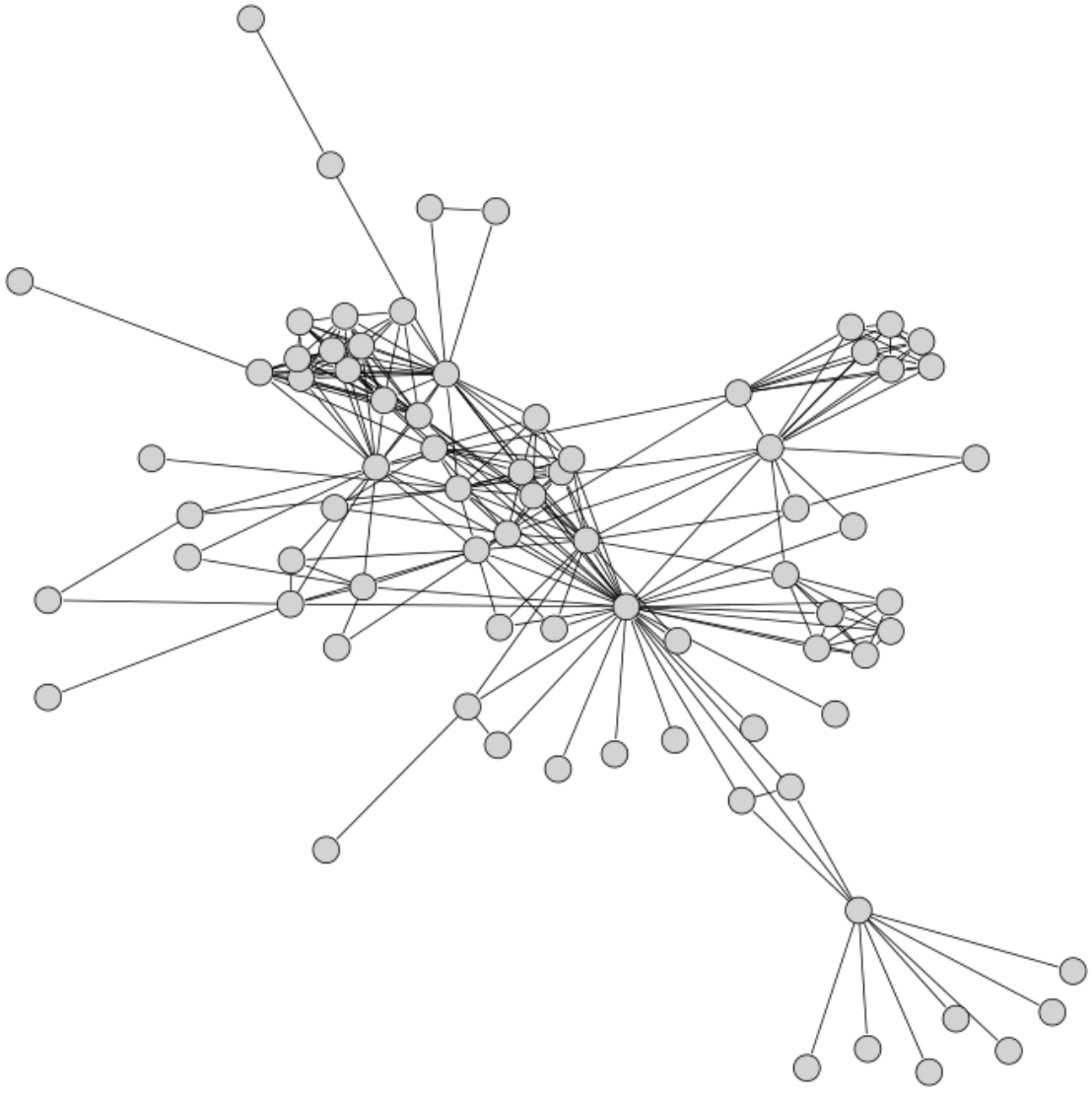
Heuristique simple

- fusion gloutonne de blocs :
  - démarrage totalement raffiné
  - meilleure fusion à chaque étape
- avec des structures de données adaptées : coût en $O(m)$ pour l'évaluation d'un modèle et en $O(m\sqrt{m}\log m)$ pour une fusion complète

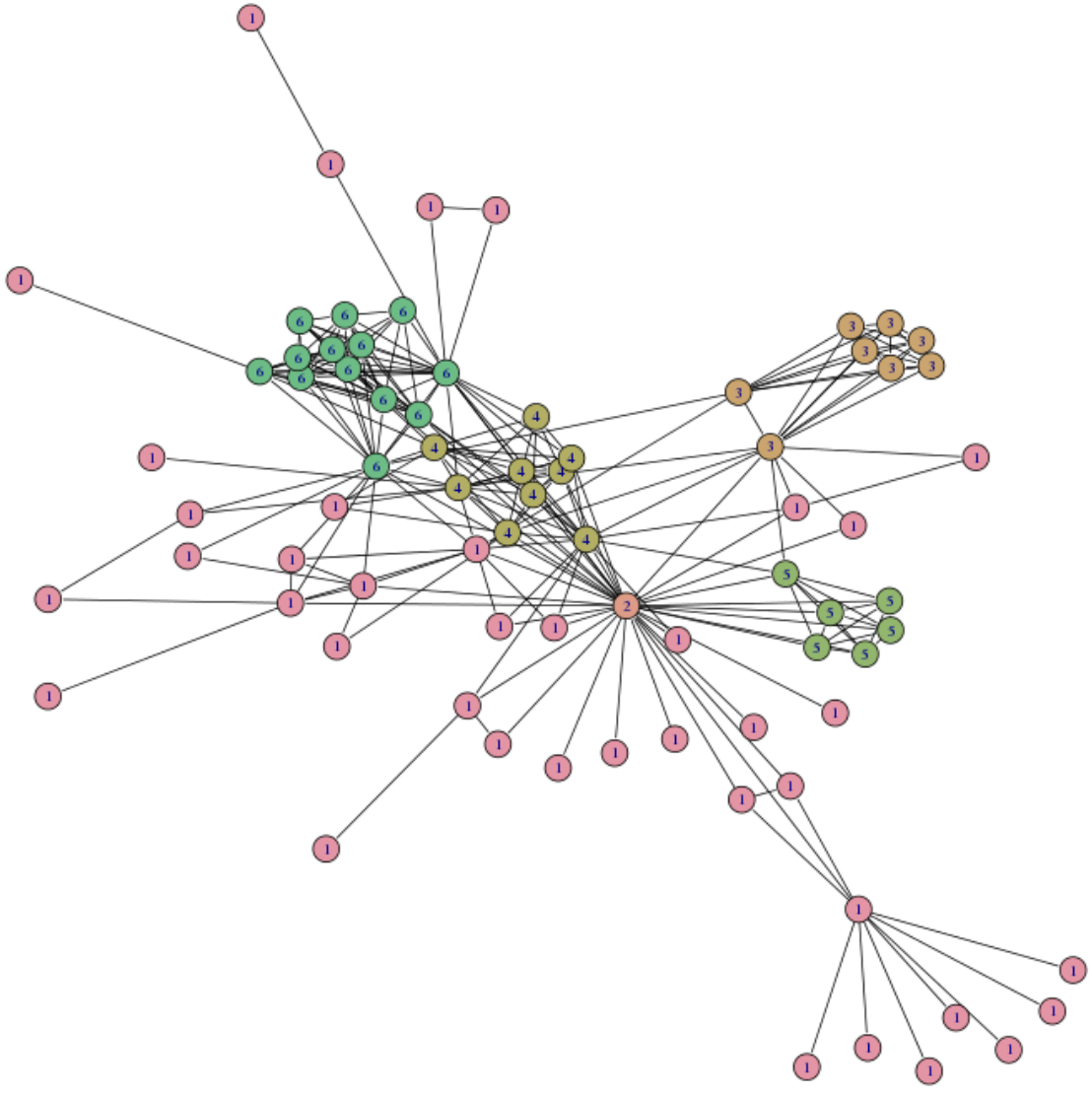
Améliorations

- ré-optimisation locale des partitions (variable par variable, par échanges)
- exploration gloutonne à partir de partitions semi-aléatoires

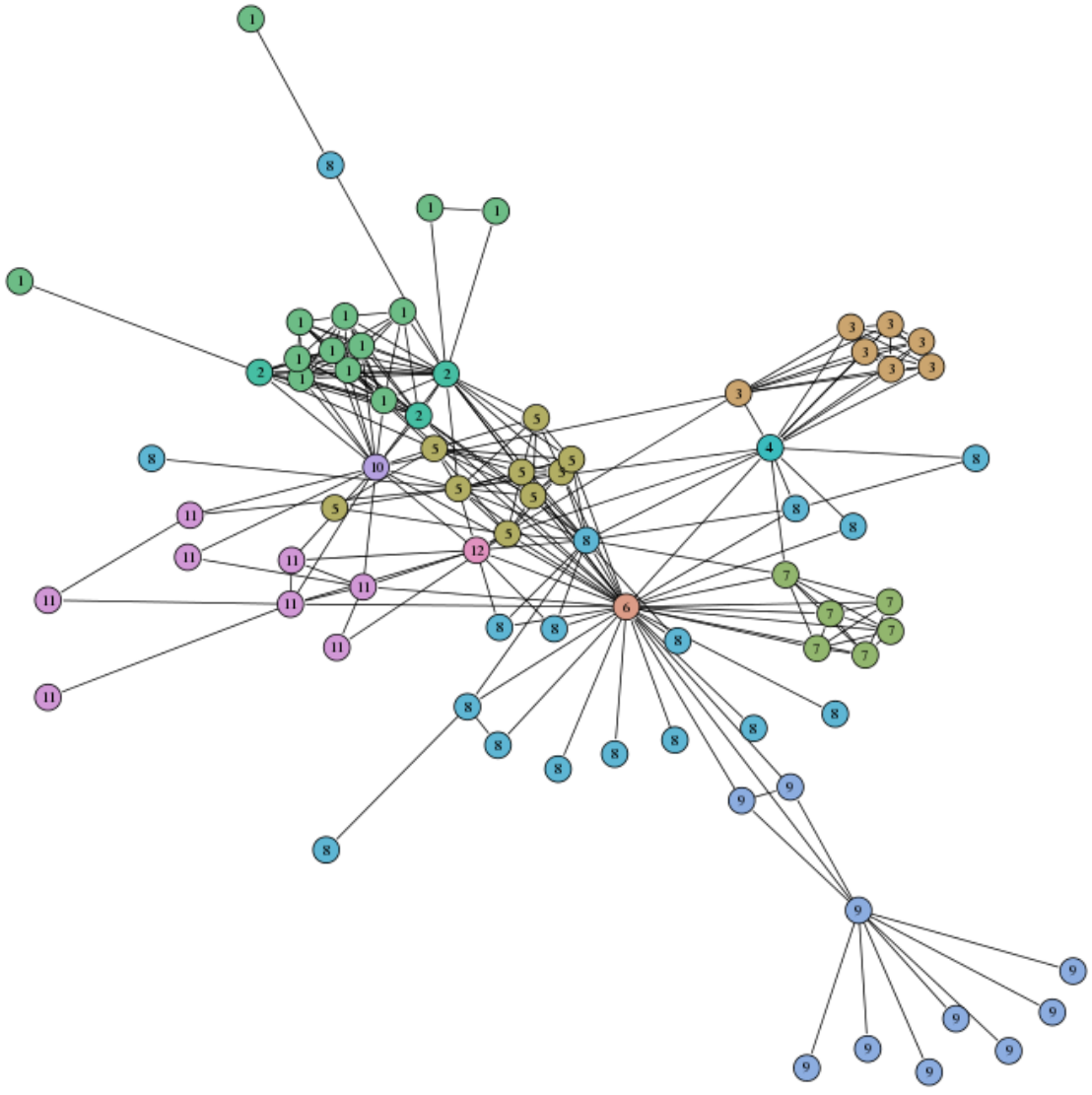
Exemple (non temporel)



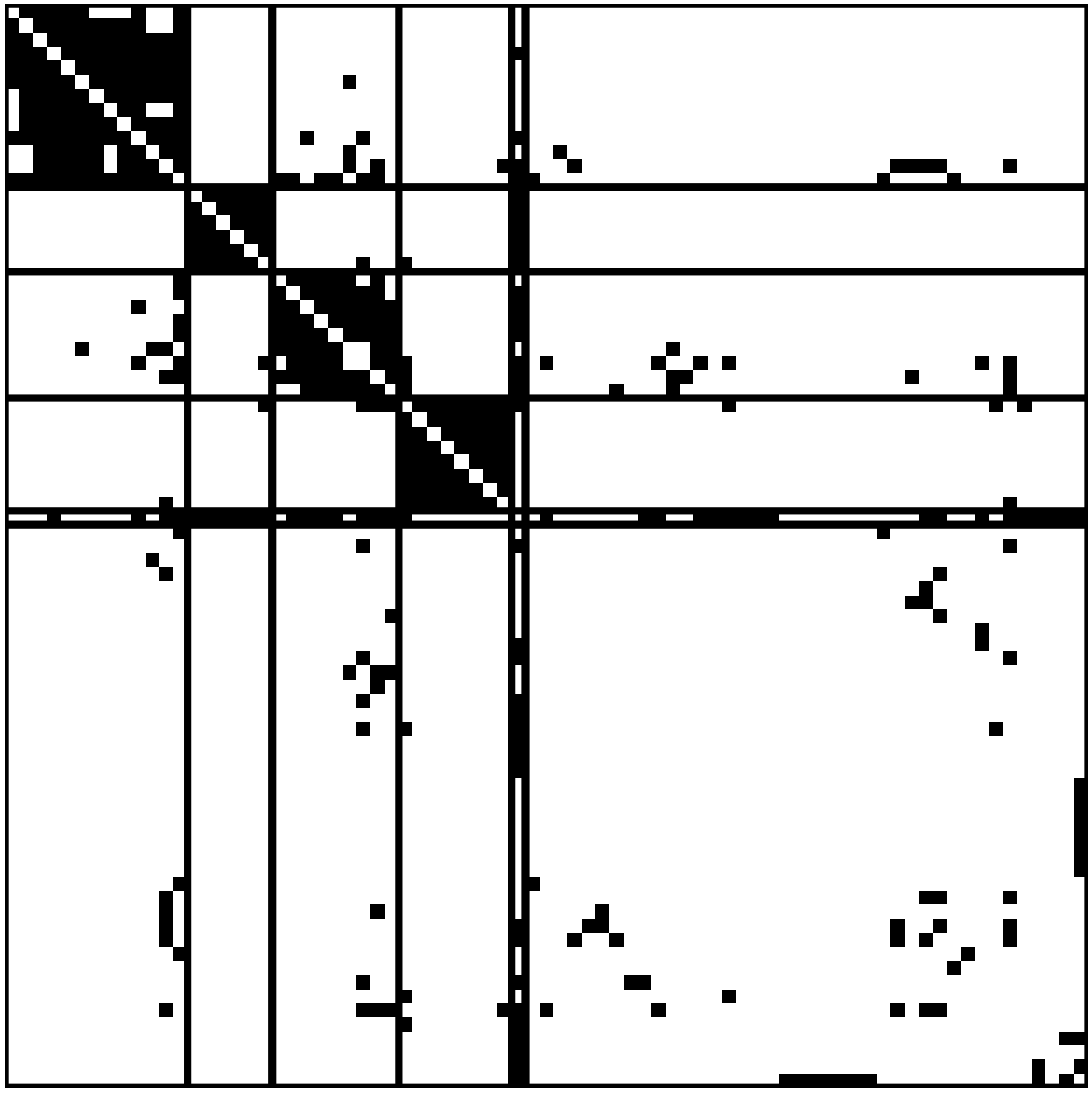
Stochastic block-model



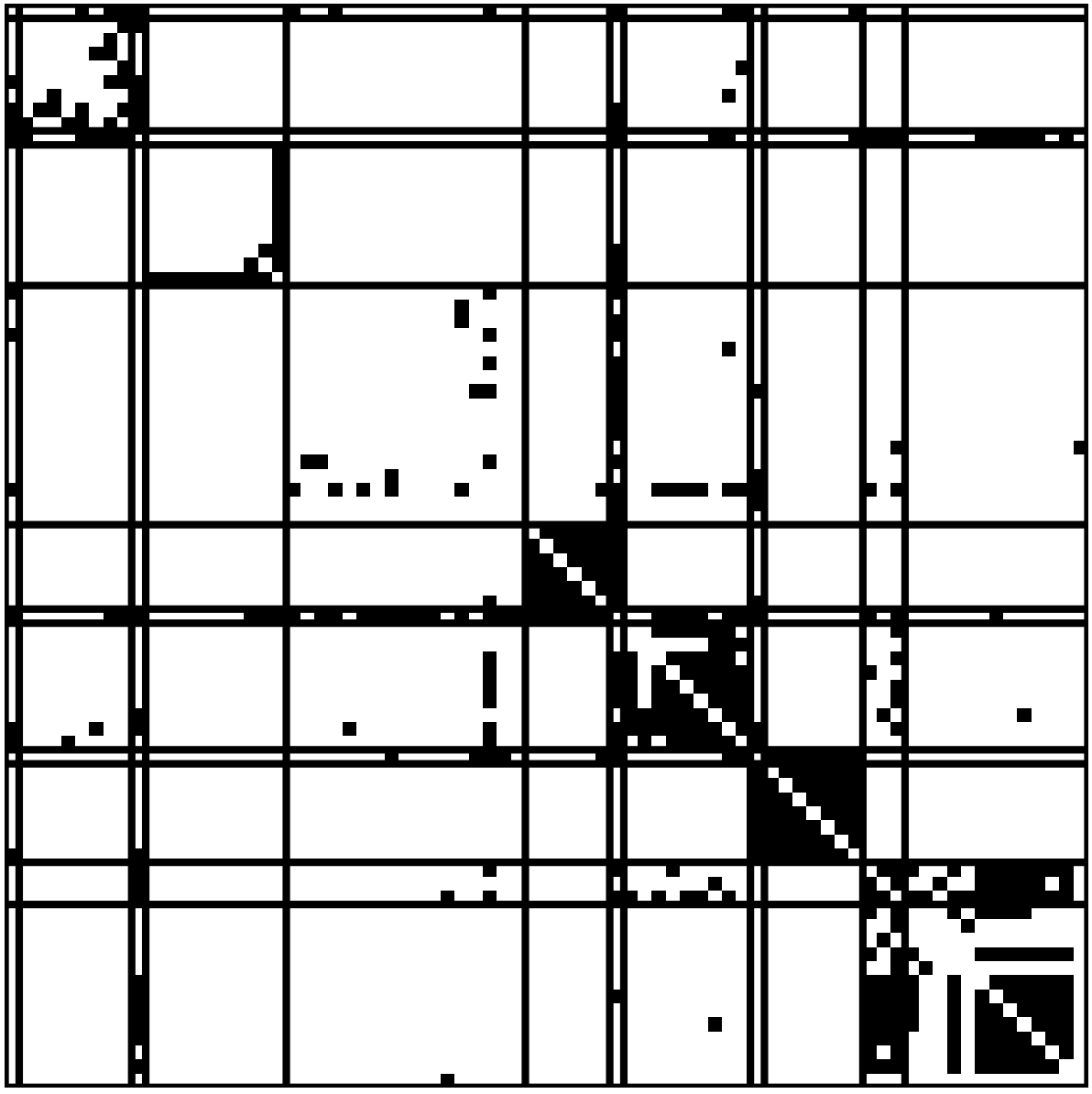
MODL



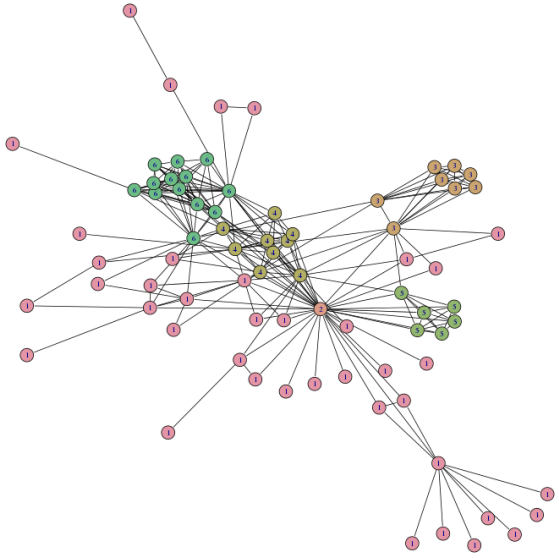
Stochastic block-model

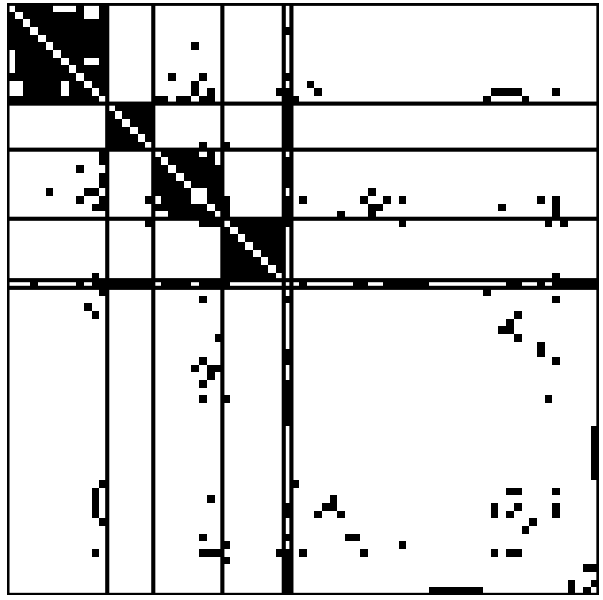


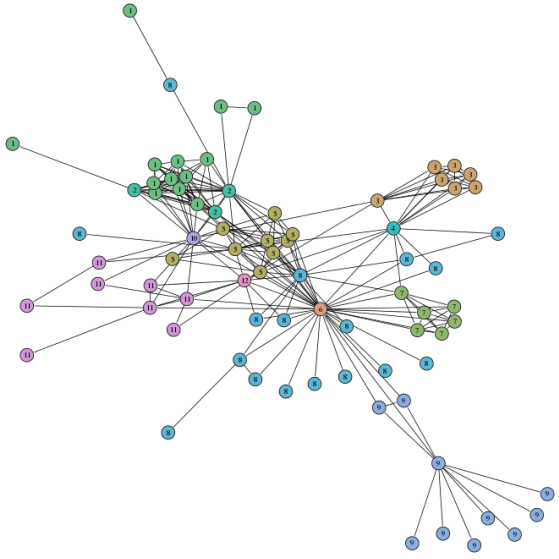
MODL

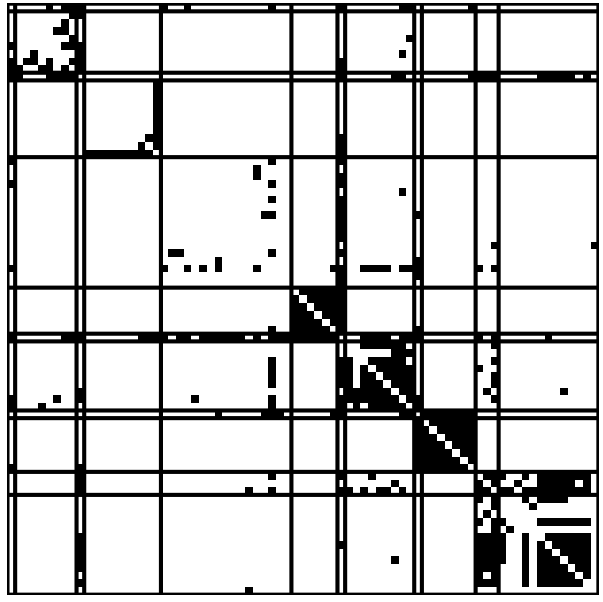


Comparaison





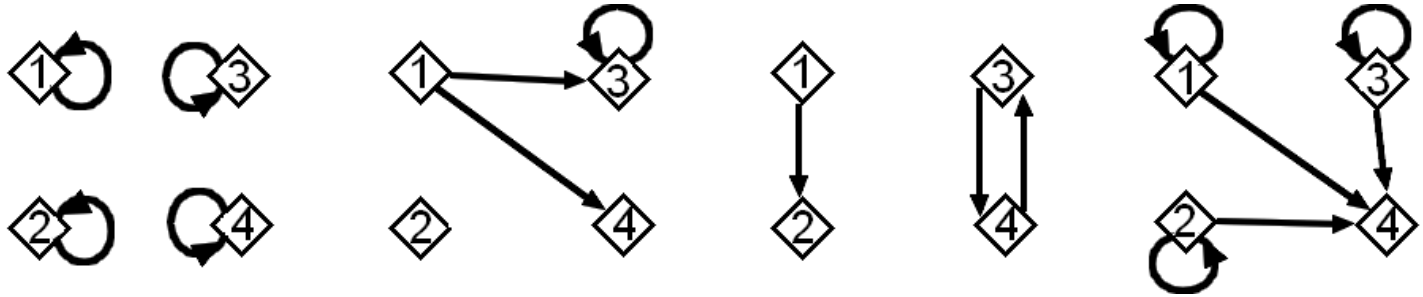




Expériences

Données synthétiques

- structure en blocs



$[0, 20[$ $[20, 30[$ $[30, 60[$ $[60, 100]$

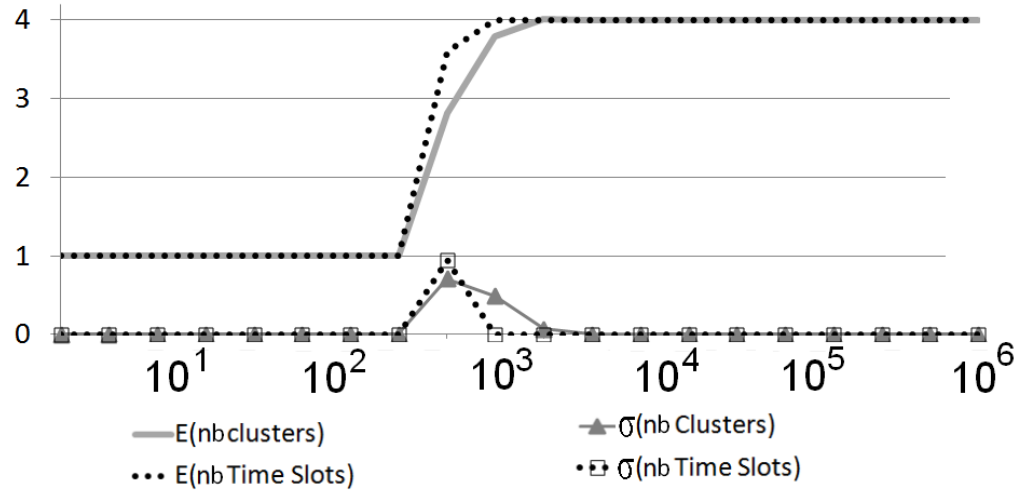
- taille des classes

classe	1	2	3	4
effectif	5	5	10	20

- connexion selon le schéma puis 30 % de re-connexions aléatoires
- essais avec un nombre croissant d'interactions (m)

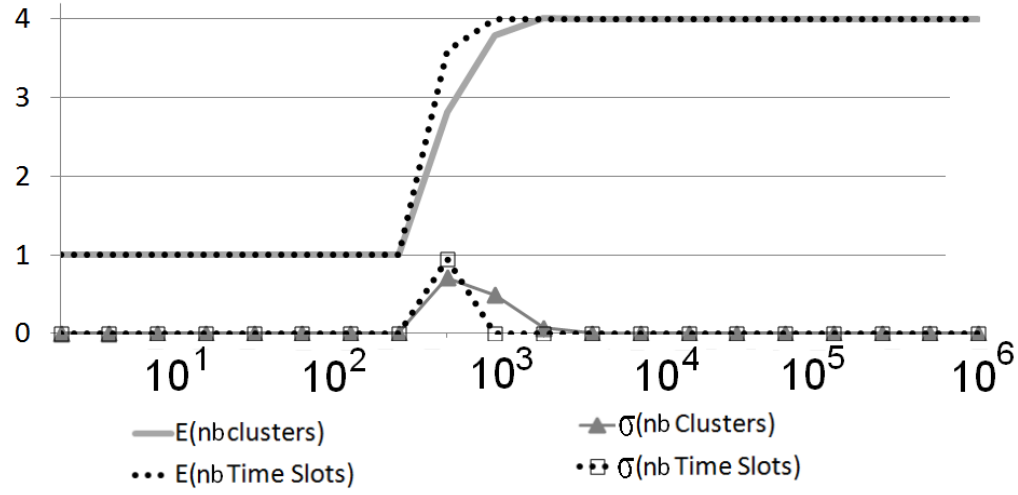
Résultats

1. Selon le modèle

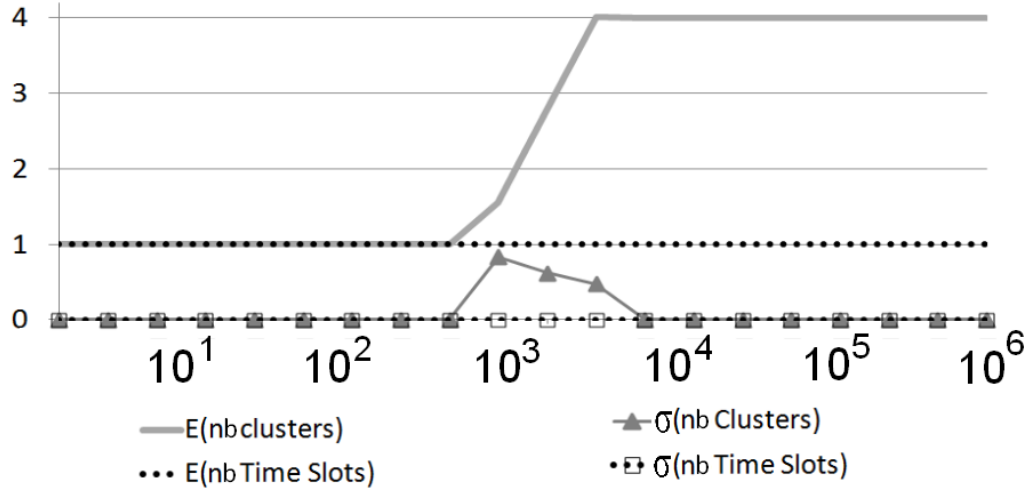


Résultats

1. Selon le modèle



2. Avec suppression de la structure temporelle



Données réelles

Vélos à Londres

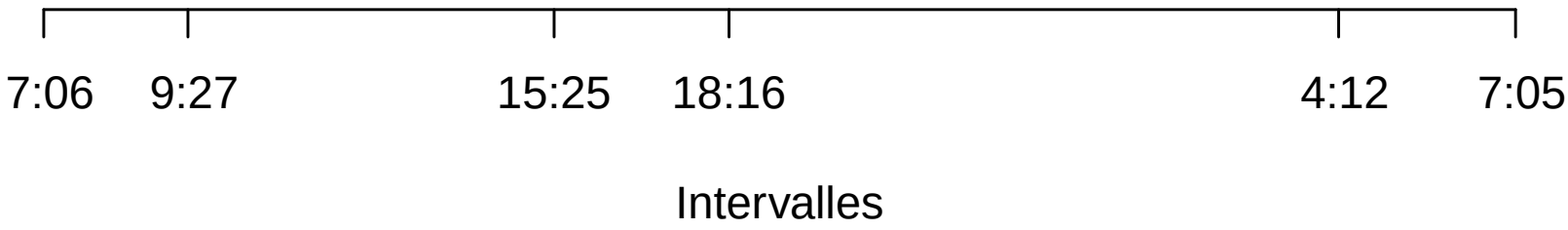
- système similaire au Vélib
- 488 stations
- enregistrement du 31 mai 2011 au 4 février 2012 : 4,8 millions de trajets

Analyse

- aspect stationnaire : prise en compte de l'heure seulement (précision à la minute)
- analyse de l'heure de départ
- temps de calcul de l'ordre de 50 minutes avec 4,5 Go de mémoire :
  - 296 classes sources, 281 classes cibles
  - 5 intervalles de temps

Analyse

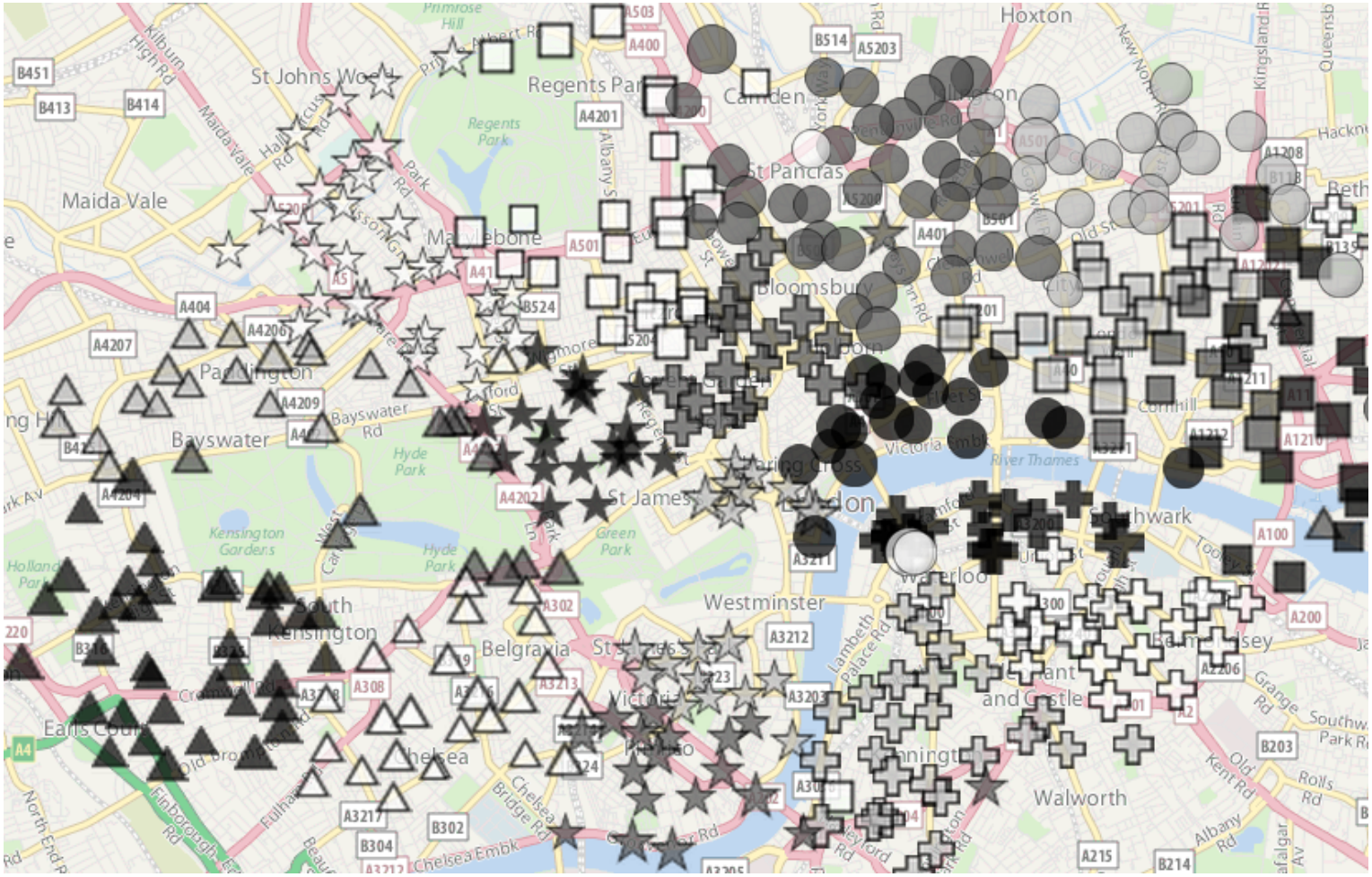
Intervalles de temps



Trop de clusters

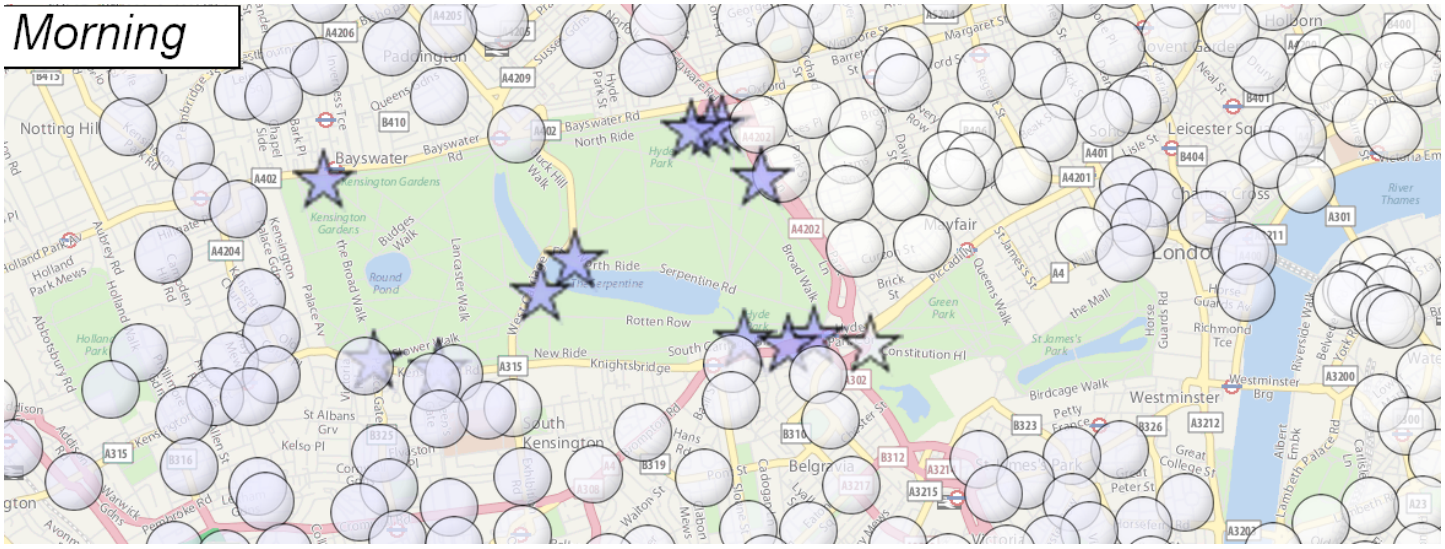
- Phénomène classique avec MODL : estimation de densité plutôt que classification
- Volume de données énorme $\Rightarrow$ justifie des classes très fines
- Simplification gloutonne :
  - même principe que l'algorithme de départ
  - choisit automatiquement la partition à simplifier

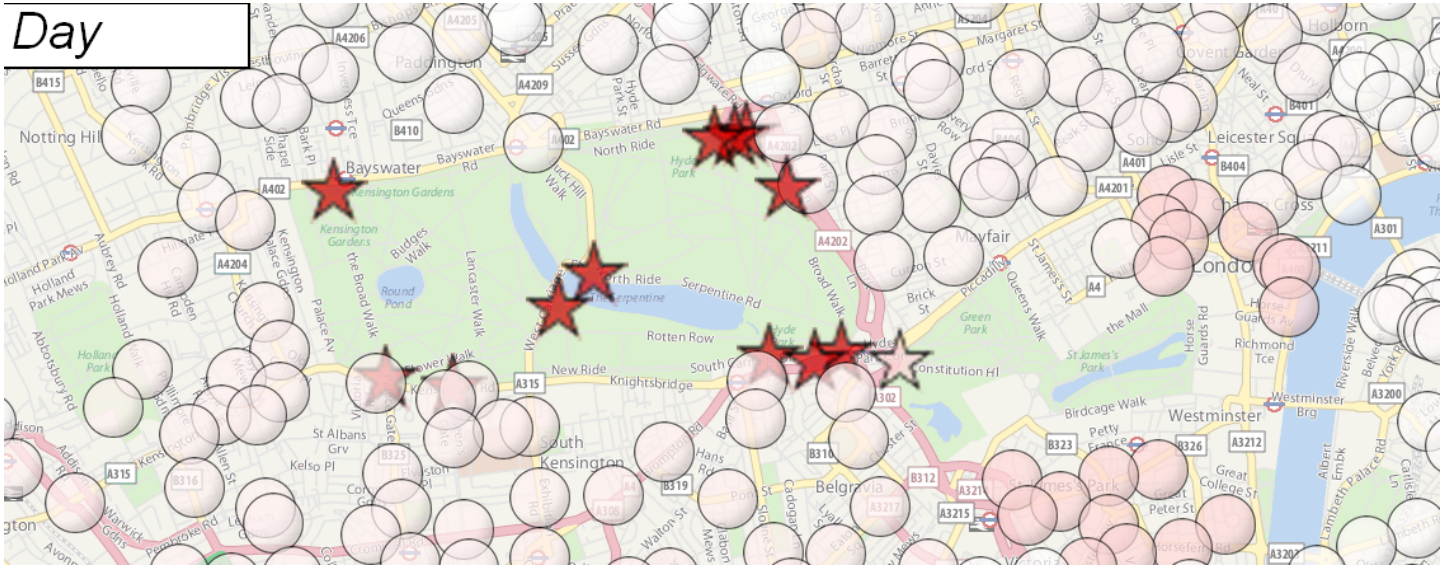
Analyse simplifiée



Réduction à 20 classes de stations (mais toujours 5 intervalles de temps)

Comparaisons





Conclusion

Résumé

- modèle en blocs basé sur MODL :
  - détecte des structures complexes
  - adapté aux gros volumes de données
- découpage temporel automatique

Perspectives

- comparaison avec d'autres techniques sur des petits graphes
- arcs pondérés de façon continue ?
- modèles très fins sur de grosses données : simplification gloutonne mais peut-on faire mieux ?
- exploitation des résultats : découverte de schémas « intéressants » ?