Détection de communautés dans des réseaux d'information utilisant liens et attributs

DAVID COMBE

sous la direction de Ch. LARGERON, E. EGYED-ZSIGMOND *, M. GÉRY

Laboratoire Hubert Curien, Université de Saint-Étienne
*LIRIS, Université de Lyon

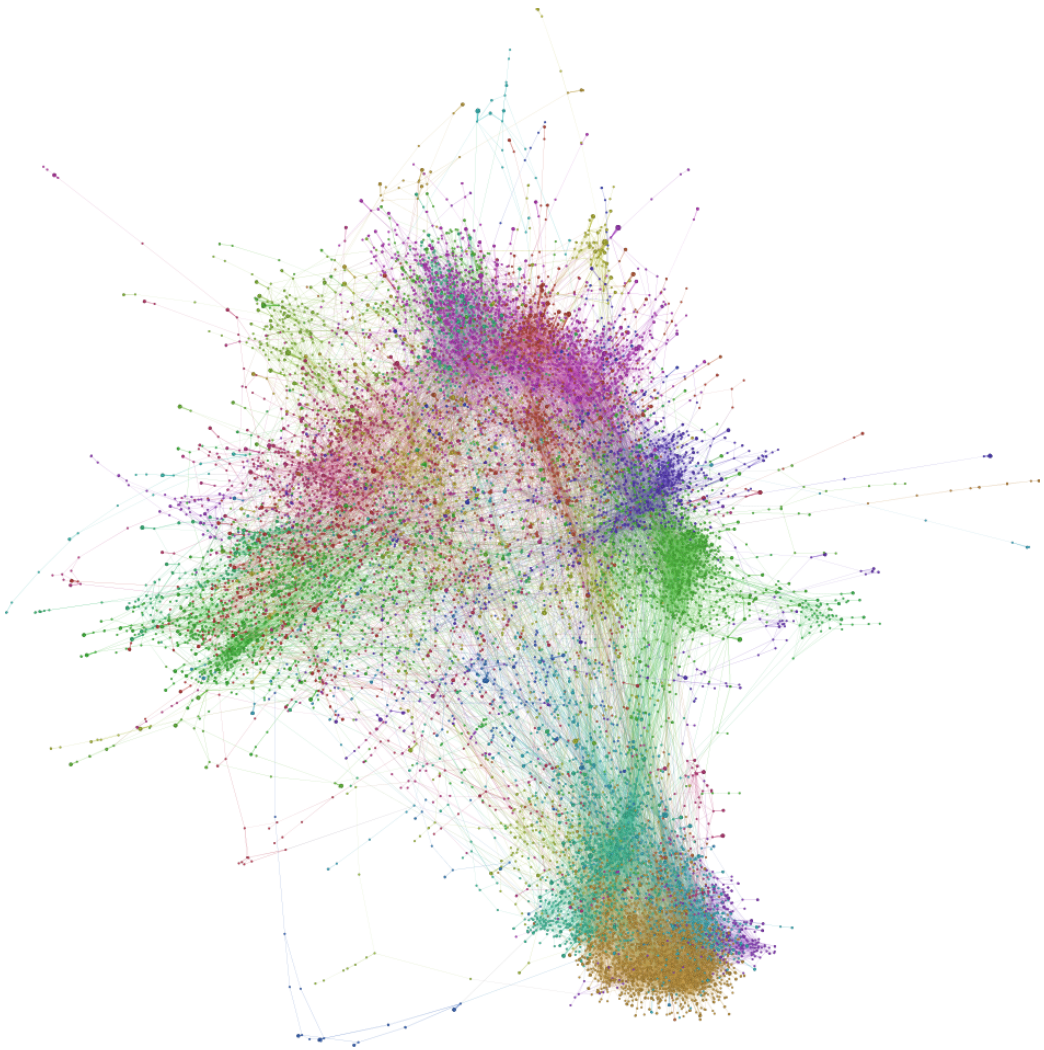
15 octobre 2013

david.combe@univ-st-etienne.fr

Réseaux sociaux

- Définition [Wasserman et al., 1994]
  "Un ou des ensembles finis d'entités ainsi que la ou les relations définies entre elles."
- Emergence des réseaux issus du Web 2.0
  - Myspace, Facebook, Twitter, LinkedIn, Instagram, etc
- Apparition de nouvelles applications [Gartner, 2008]
  - Identification d'individus influents
  - Détection de tendances émergentes
  - Recherche de communautés
- Regain d'intérêt pour l'analyse des réseaux sociaux

Exemple de réseau social



Un réseau bibliographique (PubMed)

Réseau d'information

Définition [Han *et al.*, 2009]
Un réseau d'information hétérogène est un réseau composé d'entités et de liens où :

- les sommet/liens peuvent être de différents types,
- chaque sommet/lien peut avoir un poids,
- chaque sommet/lien peut être caractérisé par des informations attachées (étiquettes, attributs numériques, information textuelle...).

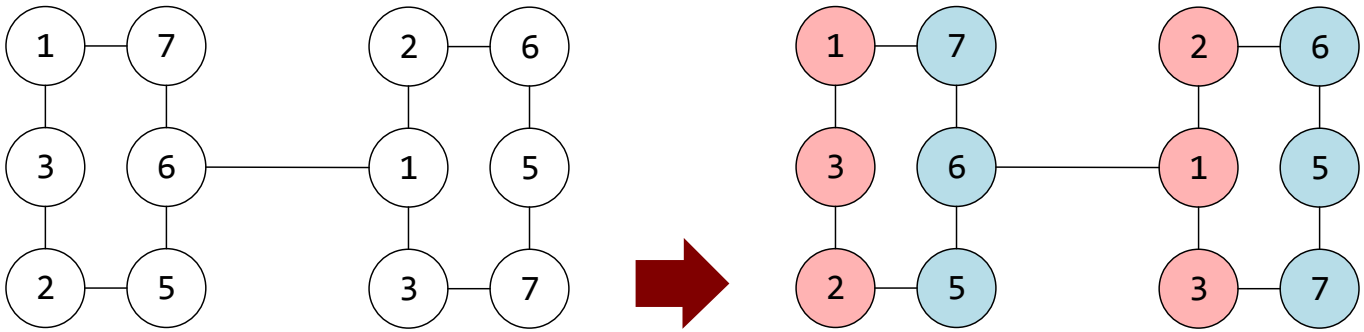
Problématique de la thèse

Détection de communautés dans un réseau d'information

Classification automatique

Construction de la partition d'éléments décrits par des vecteurs

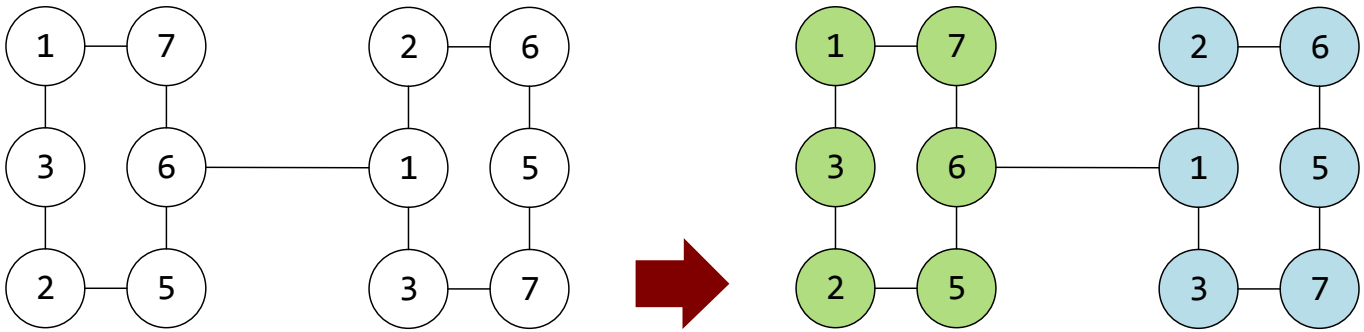
- Méthodes hiérarchiques [Ward, 1963]
- Nuées dynamiques [Forgy, 1967]
- **K-means** [MacQueen, 1967]
- etc.



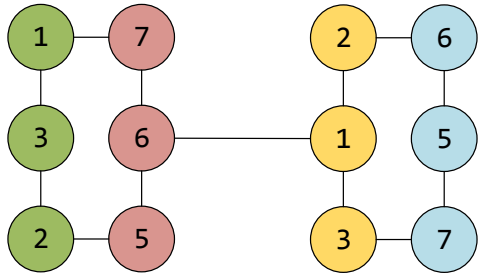
Détection de communautés

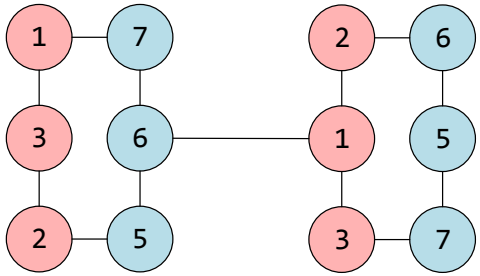
Construction de la partition à partir de données relationnelles

- Coupure minimum [Flake et al., 2003]
- Algorithme de Newman utilisant l'intermédiarité [Newman, 2004]
- **Méthode de Louvain** [V.D. Blondel *et al.*, 2008]
- etc.

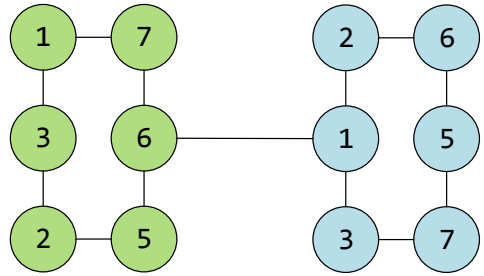


Construction de la partition à partir de données vectorielles et relationnelles





Classification selon les attributs



Classification selon les relations

Plan

Plan

Détection de communautés dans un réseau social

Étant donné un réseau social représenté par un graphe $G = (V, E)$ où

- V : l'ensemble fini des sommets de G
- $E \subset V \times V$: l'ensemble des arêtes de G
- A : matrice d'adjacence de G

il s'agit de définir une partition $\mathscr{P} = \{C_1, \ldots, C_r\}$ de V en r classes :

- $\bigcup_{k \in \{1,\ldots,r\}} C_k = V$
- $C_k \cap C_l = \emptyset, \forall\ 1 \leq k < l \leq r$
- $C_k \neq \emptyset, \forall k \in \{1, \ldots, r\}$

telle que :

- les sommets à l'intérieur d'une même classe soient fortement connectés
- les sommets de classes différentes soient peu connectés

Détection de communautés dans un réseau social

Étant donné un réseau social représenté par un graphe $G = (V, E)$ où

- V : l'ensemble fini des sommets de G
- $E \subset V \times V$: l'ensemble des arêtes de G
- A : matrice d'adjacence de G

il s'agit de définir une partition $\mathscr{P} = \{C_1, \ldots, C_r\}$ de V en r classes :

- $\bigcup_{k \in \{1,\ldots,r\}} C_k = V$
- $C_k \cap C_l = \emptyset, \forall\ 1 \leq k < l \leq r$
- $C_k \neq \emptyset, \forall k \in \{1, \ldots, r\}$

telle que :

- les sommets à l'intérieur d'une même classe soient fortement connectés
- les sommets de classes différentes soient peu connectés

Détection de communautés dans un réseau d'information

Graphe avec attributs [Zhou *et al.*, 2009]

- Étant donné un graphe $G = (V, E)$ dont **tout sommet est associé à un vecteur d'attributs**

Il s'agit de définir une partition $\mathscr{P} = \{C_1, \ldots, C_r\}$ de V en r classes telle que :

- les sommets à l'intérieur d'une même classe soient fortement connectés **et soient proches en termes d'attributs**
- les sommets de classes différentes soient peu connectés **et soient différents en termes d'attributs**

Détection de communautés dans un réseau d'information

Graphe avec attributs [Zhou *et al.*, 2009]

- Étant donné un graphe $G = (V, E)$ dont **tout sommet est associé à un vecteur d'attributs**

Il s'agit de définir une partition $\mathscr{P} = \{C_1, \ldots, C_r\}$ de V en r classes telle que :

- les sommets à l'intérieur d'une même classe soient fortement connectés **et soient proches en termes d'attributs**
- les sommets de classes différentes soient peu connectés **et soient différents en termes d'attributs**

Plan

Approches méthodologiques

- Exploitation des attributs puis des relations : enrichissement du graphe
  - [Steinhaeuser et al., 2008] Valuation des arêtes à l'aide d'une distance définie sur les attributs
  - [Zhou et al., 2009] Ajout de sommets et d'arêtes basés sur les attributs
- Exploitation des relations puis des attributs
  - [Li et al., 2008] Regroupement des communautés en fonction des attributs

Approches méthodologiques

- Exploitation des attributs puis des relations : enrichissement du graphe
  - [Steinhaeuser et al., 2008] Valuation des arêtes à l'aide d'une distance définie sur les attributs
  - [Zhou et al., 2009] Ajout de sommets et d'arêtes basés sur les attributs
- Exploitation des relations puis des attributs
  - [Li et al., 2008] Regroupement des communautés en fonction des attributs

Approches méthodologiques (2)

- Exploitation conjointe des relations et des attributs
  - [Ester et al. 2006, Moser *et al.*, 2007] NetScan, JointClust : K-means avec des contraintes de connexion des classes
  - [Handcock *et al.*, 2007] Modélisation à partir d'inférence statistique
  - Extensions de la méthode de Louvain
    - [Cruz-Gomez *et al.*, 2011] Utilisation de la notion d'entropie
    - [Dang, 2012] Combinaison de similarités locales
    - [Combe *et al.*, 2013] ToTeM, 2Mod-Louvain

Approches méthodologiques (2)

- Exploitation conjointe des relations et des attributs
  - [Ester et al. 2006, Moser *et al.*, 2007] NetScan, JointClust : K-means avec des contraintes de connexion des classes
  - [Handcock *et al.*, 2007] Modélisation à partir d'inférence statistique
  - Extensions de la méthode de Louvain
    - [Cruz-Gomez *et al.*, 2011] Utilisation de la notion d'entropie
    - [Dang, 2012] Combinaison de similarités locales
    - [Combe *et al.*, 2013] ToTeM, 2Mod-Louvain

Plan

ToTeM

- Une extension de la méthode de Louvain,
- prenant en compte des attributs vectoriels,
- optimisant un nouveau critère global lors de la phase itérative,
- proposant une redéfinition de la phase de fusion des communautés.

Modularité

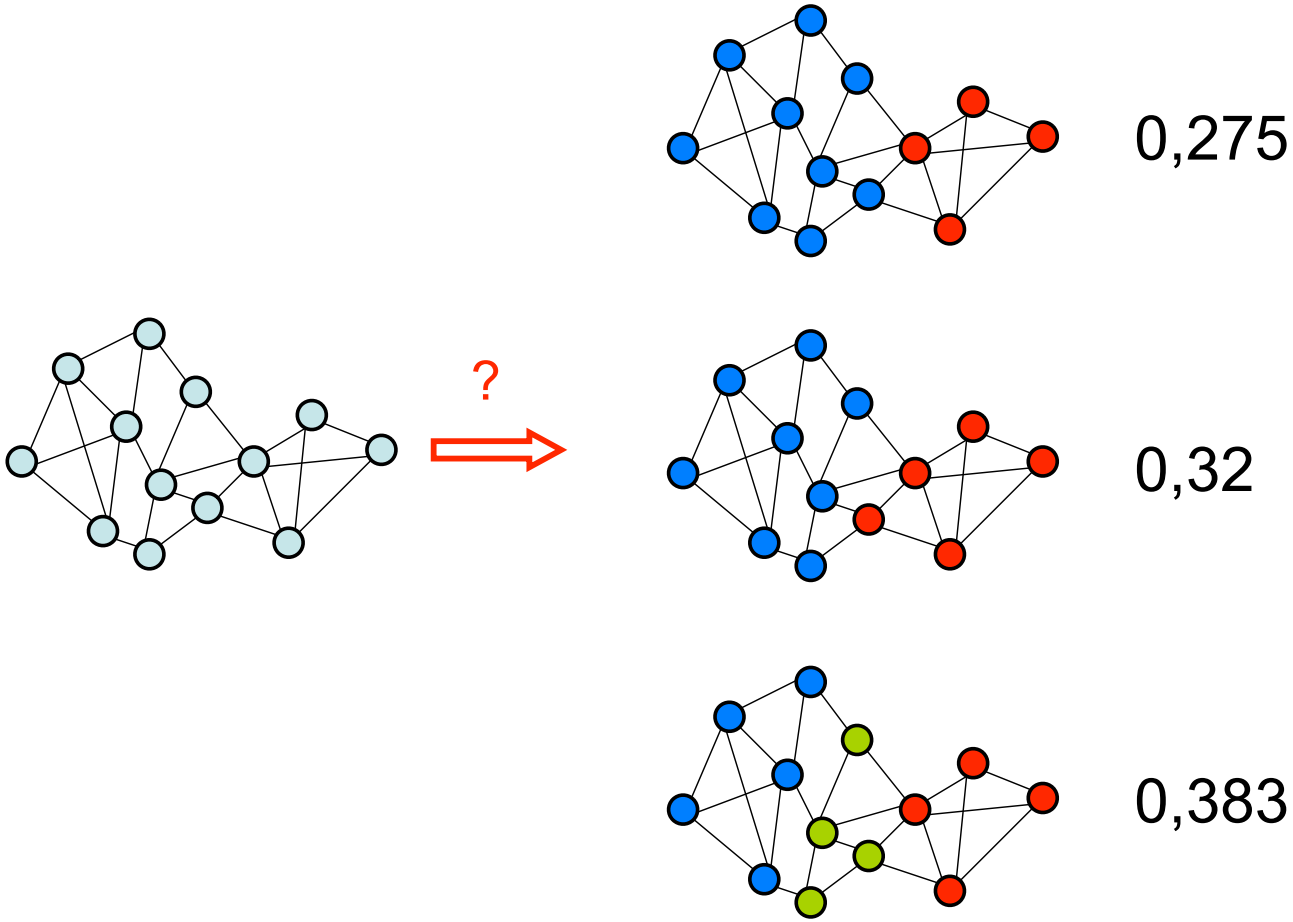
Mesure de qualité d'une partition $\mathscr{P}$ par rapport aux liens, variant entre -1 et 1 [Newman et Girvan, 2004].

$$\frac{\text{poids des arêtes dans les communautés}}{\text{poids total des arêtes}} - \frac{\text{poids des arêtes attendues dans les communautés dans le graphe aléatoire}}{\text{poids total des arêtes}}$$

$$Q_{NG}(\mathscr{P}) = \sum_{(i,i')\in V\times V} \left[\left(\frac{A_{ii'}}{2M} - \frac{k_i}{2M} \cdot \frac{k_{i'}}{2M} \right) \cdot \delta(c_i, c_{i'}) \right]$$

où M est la somme des poids des liens, k_i est le degré du sommet i et δ est la fonction de Kronecker.

Exemple



Choix de la partition optimisant la modularité (Blondel 2012)

Qualité d'une partition $\mathscr{P}$

Inertie interclasses

Qualité de $\mathscr{P}$ par rapport aux attributs

$$I_{inter}(\mathscr{P}) = \sum_{l=1,r} m_l \|g_l - g\|^2$$

où g est le centre de gravité de V, g_l est le centre de gravité de la classe l et m_l le poids de la classe C_l.

Qualité globale de $\mathscr{P}$ par rapport aux attributs et aux liens :

$$CG(\mathscr{P}) = \frac{I_{inter}(\mathscr{P})}{|\mathscr{P}| \cdot I(V)} \cdot Q_{NG}(\mathscr{P})$$

où $I(V)$ est l'inertie de V

Qualité d'une partition $\mathscr{P}$

Inertie interclasses

Qualité de $\mathscr{P}$ par rapport aux attributs

$$I_{inter}(\mathscr{P}) = \sum_{l=1,r} m_l \|g_l - g\|^2$$

où g est le centre de gravité de V, g_l est le centre de gravité de la classe l et m_l le poids de la classe C_l.

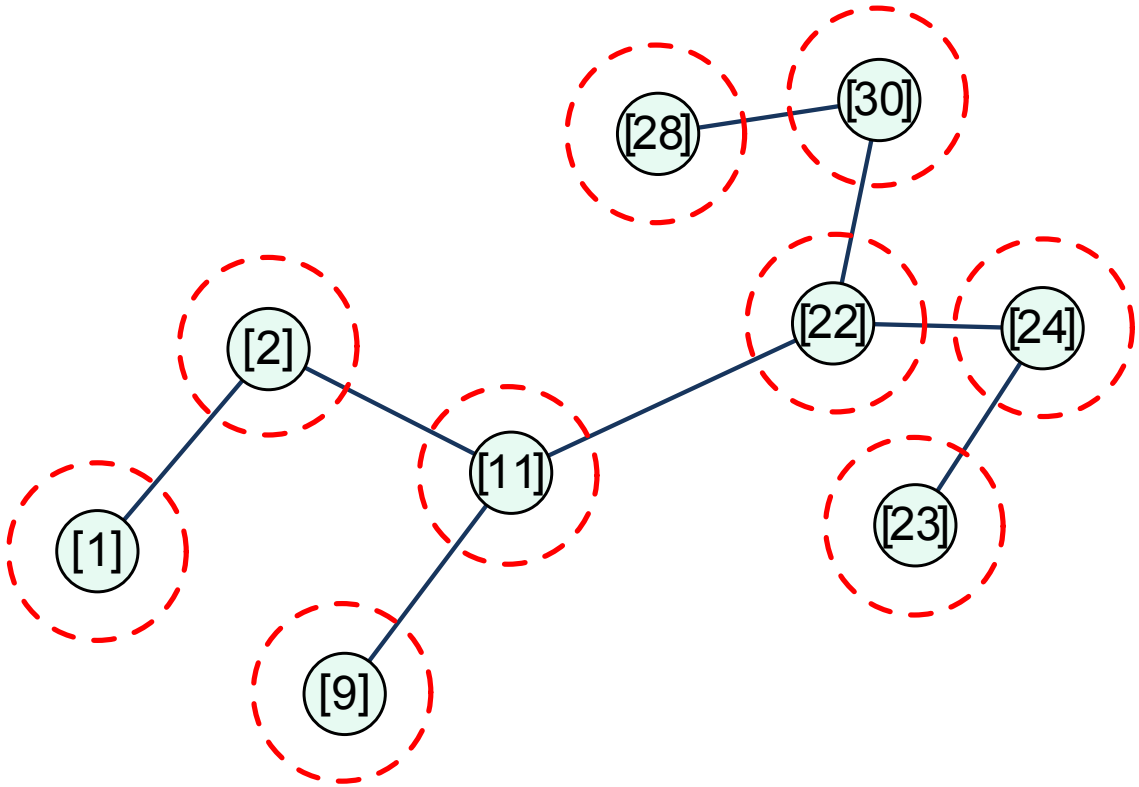
Qualité globale de $\mathscr{P}$ par rapport aux attributs et aux liens :

$$CG(\mathscr{P}) = \frac{I_{inter}(\mathscr{P})}{|\mathscr{P}| \cdot I(V)} \cdot Q_{NG}(\mathscr{P})$$

où $I(V)$ est l'inertie de V

Algorithme ToTeM

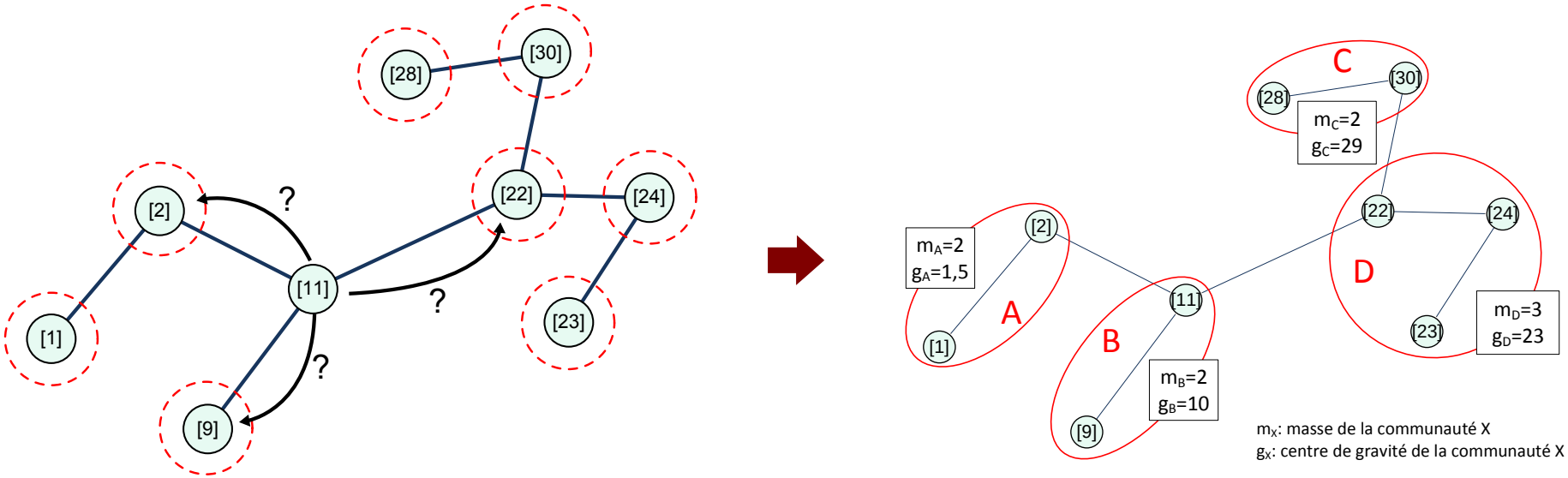
- Initialisation : chaque sommet constitue une communauté



Algorithme ToTeM

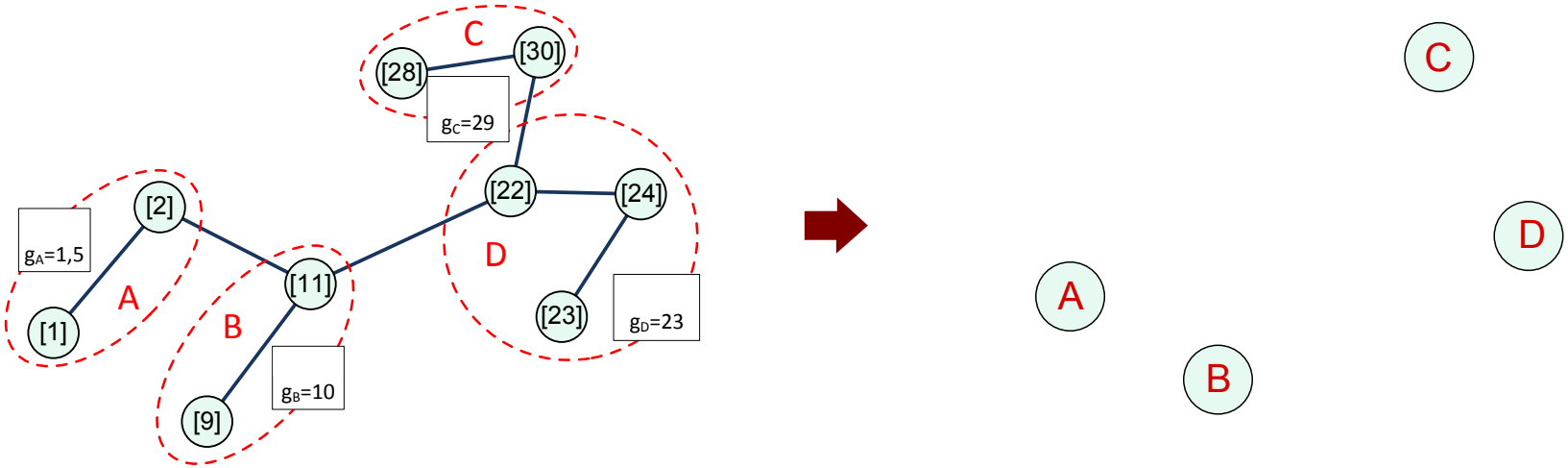
- **Phase itérative :**
  Répéter
  - Pour tout sommet i, insérer i dans la communauté voisine qui maximise le critère global

  jusqu'à ce qu'un maximum local soit atteint



Algorithme ToTeM

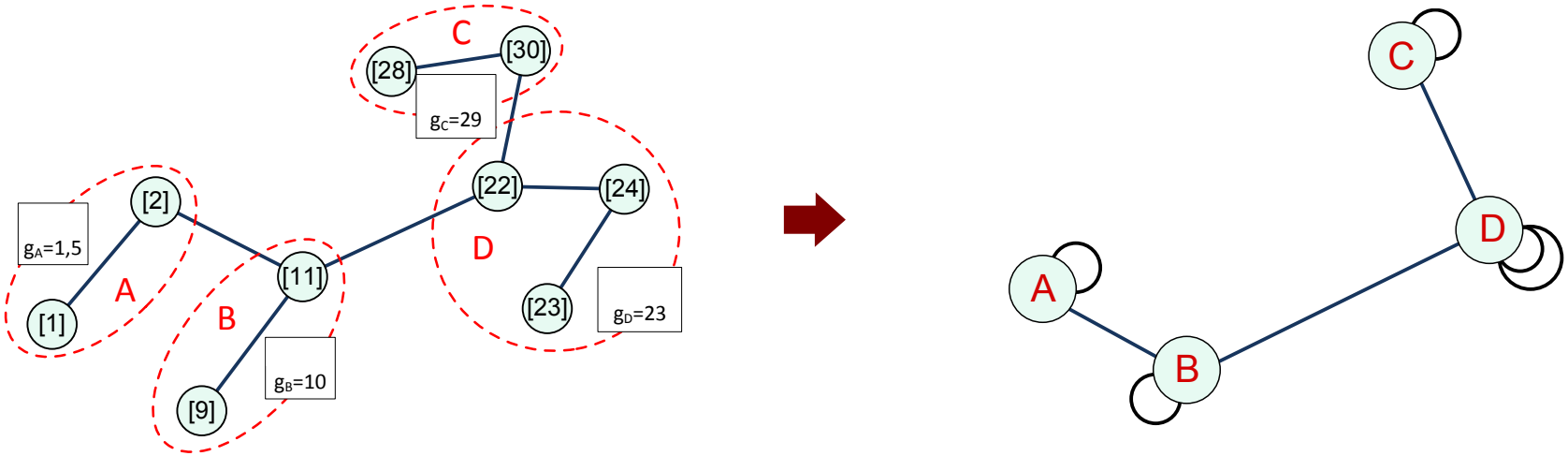
- **Phase de fusion**
  Construction d'un nouveau graphe $G' = (V', E')$ à partir de la partition $\mathscr{P}'$
  - Chaque sommet v de G' correspond à une classe C de $\mathscr{P}'$
  - La valuation de l'arête entre deux sommets v_x et v_y de G' est la somme des valuations entre les sommets des classes correspondantes
  - Le vecteur d'attributs associé à v est le centre de gravité de C
  - Le poids du sommet est celui de la classe



Phase itérative et phase de fusion alternées jusqu'à convergence du critère

Algorithme ToTeM

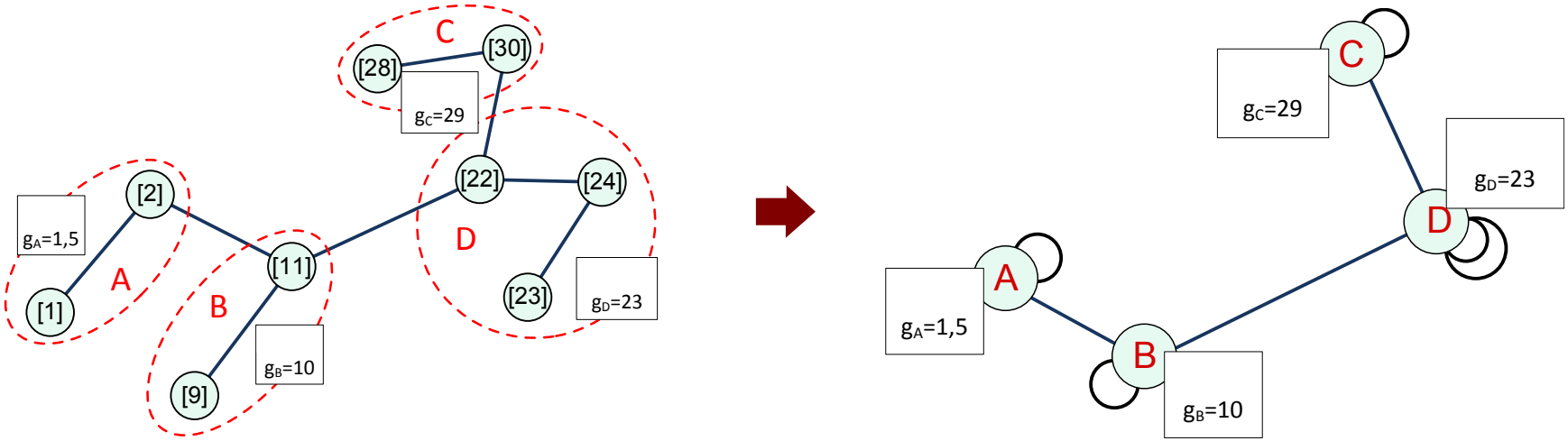
- **Phase de fusion**
  Construction d'un nouveau graphe $G' = (V', E')$ à partir de la partition $\mathscr{P}'$
  - Chaque sommet v de G' correspond à une classe C de $\mathscr{P}'$
  - La valuation de l'arête entre deux sommets v_x et v_y de G' est la somme des valuations entre les sommets des classes correspondantes
  - Le vecteur d'attributs associé à v est le centre de gravité de C
  - Le poids du sommet est celui de la classe



Phase itérative et phase de fusion alternées jusqu'à convergence du critère

Algorithme ToTeM

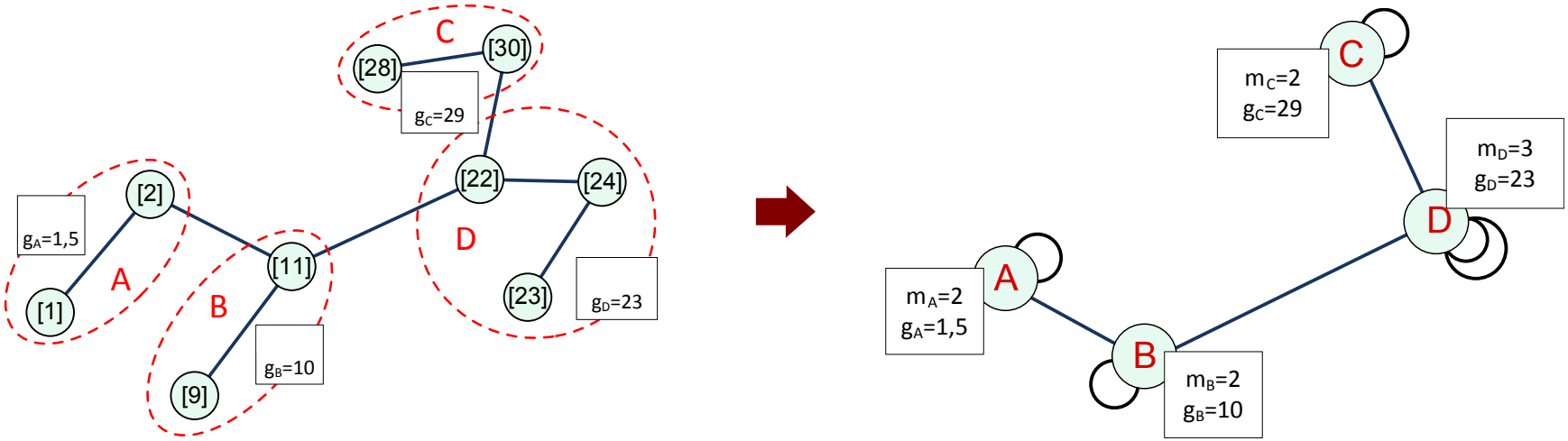
- **Phase de fusion**
  Construction d'un nouveau graphe $G' = (V', E')$ à partir de la partition $\mathscr{P}'$
  - Chaque sommet v de G' correspond à une classe C de $\mathscr{P}'$
  - La valuation de l'arête entre deux sommets v_x et v_y de G' est la somme des valuations entre les sommets des classes correspondantes
  - Le vecteur d'attributs associé à v est le centre de gravité de C
  - Le poids du sommet est celui de la classe



Phase itérative et phase de fusion alternées jusqu'à convergence du critère

Algorithme ToTeM

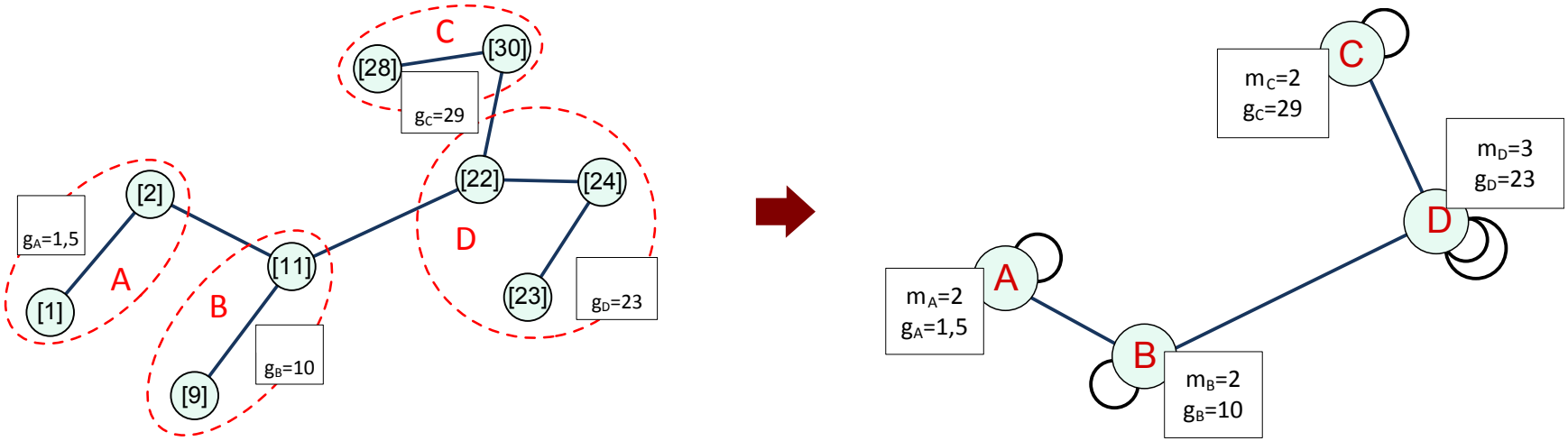
- **Phase de fusion**
  Construction d'un nouveau graphe $G' = (V', E')$ à partir de la partition $\mathscr{P}'$
  - Chaque sommet v de G' correspond à une classe C de $\mathscr{P}'$
  - La valuation de l'arête entre deux sommets v_x et v_y de G' est la somme des valuations entre les sommets des classes correspondantes
  - Le vecteur d'attributs associé à v est le centre de gravité de C
  - Le poids du sommet est celui de la classe



Phase itérative et phase de fusion alternées jusqu'à convergence du critère

Algorithme ToTeM

- **Phase de fusion**
  Construction d'un nouveau graphe $G' = (V', E')$ à partir de la partition $\mathscr{P}'$
  - Chaque sommet v de G' correspond à une classe C de $\mathscr{P}'$
  - La valuation de l'arête entre deux sommets v_x et v_y de G' est la somme des valuations entre les sommets des classes correspondantes
  - Le vecteur d'attributs associé à v est le centre de gravité de C
  - Le poids du sommet est celui de la classe



Phase itérative et phase de fusion alternées jusqu'à convergence du critère

Inertie interclasses

Avantage : calcul optimisé du gain d'inertie à partir d'une information locale.

Inertie interclasses suite au changement de classe d'un sommet
Pour un sommet u passant d'une classe A à une classe B :

$$\begin{aligned}\Delta I_{inter} =& I_{inter}(\mathscr{P}') - I_{inter}(\mathscr{P}) \\ =& (m_A - m_u) \cdot \left\| g_{A\setminus\{u\}} - g \right\|^2 + (m_B + m_u) \cdot \left\| g_{B\cup\{u\}} - g \right\|^2 \\ & - m_A \cdot \|g_A - g\|^2 - m_B \cdot \|g_B - g\|^2\end{aligned}$$

où g_A est le centre de gravité de la classe A,
m_A est la masse associée à la classe A.

Inconvénient : l'inertie est adaptée à la comparaison de partitions ayant le même nombre de classes.

Comparaison de partitions avec nombres de classes différents

- Indice de Calinski
  - utilisé pour déterminer le nombre optimum de classes

$$CH(\mathscr{P}) = \frac{I_{inter}(\mathscr{P})/(|\mathscr{P}| - 1)}{I_{intra}(\mathscr{P})/(|V| - |\mathscr{P}|)} \quad (1)$$

- Probabilité critique du test de Fisher-Snedecor.
  Mesure de l'écart par rapport à une distribution aléatoire des éléments au sein de la partition.

$$PC = P(F(|\mathscr{P}| - 1, |V| - |\mathscr{P}|) > F_{\mathscr{P}}) \quad (2)$$

Inconvénient : manque de précision pour évaluer l'évolution induite par un changement local.

Comparaison de partitions avec nombres de classes différents

- Indice de Calinski
  - utilisé pour déterminer le nombre optimum de classes

$$CH(\mathscr{P}) = \frac{I_{inter}(\mathscr{P})/(|\mathscr{P}| - 1)}{I_{intra}(\mathscr{P})/(|V| - |\mathscr{P}|)} \quad (1)$$

- Probabilité critique du test de Fisher-Snedecor.
  Mesure de l'écart par rapport à une distribution aléatoire des éléments au sein de la partition.

$$PC = P(F(|\mathscr{P}| - 1, |V| - |\mathscr{P}|) > F_{\mathscr{P}}) \quad (2)$$

Inconvénient : manque de précision pour évaluer l'évolution induite par un changement local.

Comparaison de partitions avec nombres de classes différents

- Indice de Calinski
  - utilisé pour déterminer le nombre optimum de classes

$$CH(\mathscr{P}) = \frac{I_{inter}(\mathscr{P})/(|\mathscr{P}| - 1)}{I_{intra}(\mathscr{P})/(|V| - |\mathscr{P}|)} \quad (1)$$

- Probabilité critique du test de Fisher-Snedecor.
  Mesure de l'écart par rapport à une distribution aléatoire des éléments au sein de la partition.

$$PC = P(F(|\mathscr{P}| - 1, |V| - |\mathscr{P}|) > F_{\mathscr{P}}) \quad (2)$$

Inconvénient : manque de précision pour évaluer l'évolution induite par un changement local.

Plan

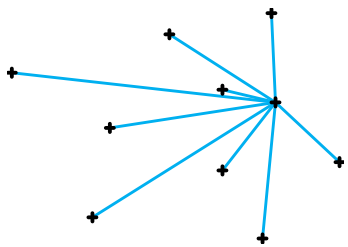
Modularité basée sur l'inertie

Définition

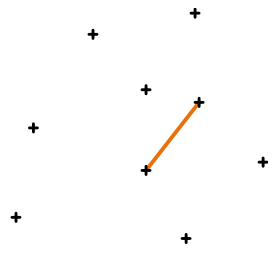
$$Q_{inertie}(\mathscr{P}) = \sum_{(i,i')\in V\times V} \left[\left(\frac{I(V,v)\cdot I(V,v')}{(2N\cdot I(V))^2} - \frac{\|v-v'\|^2}{2N\cdot I(V)} \right) \cdot \delta\left(c_i, c_{i'}\right) \right]$$

où $I(V)$ est l'inertie totale,
et $I(V,i)$ est l'inertie de V par rapport à $i \in V$.

Inertie par rapport à un sommet



Norme entre deux sommets



Modularité basée sur l'inertie : propriétés

- Varie entre -1 et 1, comme la modularité,
- Insensible à une transformation linéaire appliquée à l'ensemble des vecteurs,
- Insensible au nombre de classes de la partition,
- Calculable à partir de l'information locale.

Le critère utilisé dans 2Mod-Louvain est
$Q_{NG} + Q_{inertie}$.

Modularité basée sur l'inertie : propriétés

- Varie entre -1 et 1, comme la modularité,
- Insensible à une transformation linéaire appliquée à l'ensemble des vecteurs,
- Insensible au nombre de classes de la partition,
- Calculable à partir de l'information locale.

Le critère utilisé dans 2Mod-Louvain est
$Q_{NG} + Q_{inertie}$.

Plan

Réseaux réels : 4 sessions (1)

Construction d'un réseau à partir de 2 conférences, SAC 2009 et IJCAI 2009, où :

- $G = (V, E)$, $|V| = 99$, $|E| = 2623$
- chaque auteur est un sommet, décrit par le résumé de ses articles représentés par *tf-idf*
- il existe un lien entre 2 auteurs s'il existe au moins un journal ou une conférence dans lequel ils ont tous les deux publiés, entre 2007 et 2009, même sans être coauteurs (DBLP).

Réseaux réels : 4 sessions (2)

Effectifs des sessions :

Session et conférence de rattachement	Effectif
A Bioinformatique (SAC)	24
B Robotique (SAC)	16
C Robotique (IJCAI)	38
D Contraintes (IJCAI)	21
Effectif du jeu de données	99

3 partitions différentes :

- 2 conférences : SAC et IJCAI 2009 (partition P_S)
- 3 thématiques : Bioinformatique, Robotique et Contraintes (partition P_T)
- 4 sessions (partition P_{TS})

Réseaux réels : 4 sessions (2)

Effectifs des sessions :

Session et conférence de rattachement	Effectif
A Bioinformatique (SAC)	24
B Robotique (SAC)	16
C Robotique (IJCAI)	38
D Contraintes (IJCAI)	21
Effectif du jeu de données	99

3 partitions différentes :

- 2 conférences : SAC et IJCAI 2009 (partition P_S)
- 3 thématiques : Bioinformatique, Robotique et Contraintes (partition P_T)
- 4 sessions (partition P_{TS})

Réseaux réels : 4 sessions (2)

Effectifs des sessions :

Session et conférence de rattachement	Effectif
A Bioinformatique (SAC)	24
B Robotique (SAC)	16
C Robotique (IJCAI)	38
D Contraintes (IJCAI)	21
Effectif du jeu de données	99

3 partitions différentes :

- 2 conférences : SAC et IJCAI 2009 (partition P_S)
- 3 thématiques : Bioinformatique, Robotique et Contraintes (partition P_T)
- 4 sessions (partition P_{TS})

Réseaux réels : 4 sessions (2)

Effectifs des sessions :

Session et conférence de rattachement	Effectif
A Bioinformatique (SAC)	24
B Robotique (SAC)	16
C Robotique (IJCAI)	38
D Contraintes (IJCAI)	21
Effectif du jeu de données	99

3 partitions différentes :

- 2 conférences : SAC et IJCAI 2009 (partition P_S)
- 3 thématiques : Bioinformatique, Robotique et Contraintes (partition P_T)
- 4 sessions (partition P_{TS})

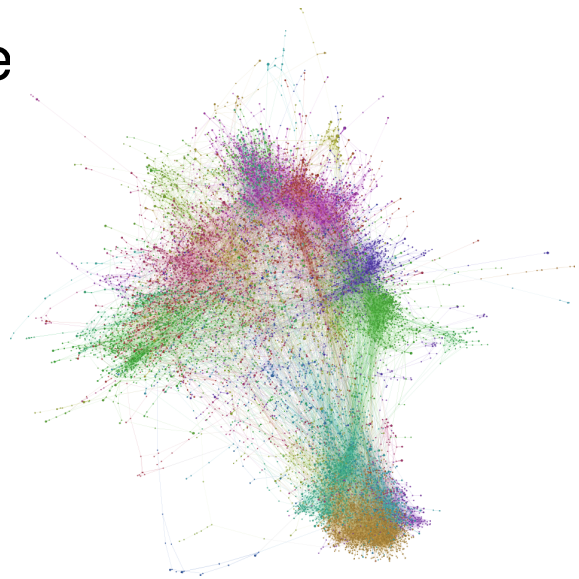
Réseaux réels : 4 sessions (3)

Modèle	Précision vis-à-vis de : P_T	P_S	P_{TS}
K-means	0,87	-	0,69
Méthode de Louvain	-	1,00	0,63
ToTeM	-	-	0,63
2Mod-Louvain	-	-	0,63

Pubmed-Diabètes

Jeu de données **Pubmed-Diabètes** [Sen, 2008]
Définition du réseau :

- 19 717 publications scientifiques traitant du diabète (V),
- reliées par la relation de citation,
- représentés par un vecteur de pondérations *tf-idf* avec 500 termes.
- Publications réparties en trois catégorie
  - diabète de type 1
  - diabète de type 2
  - induit médicalement / expérimental



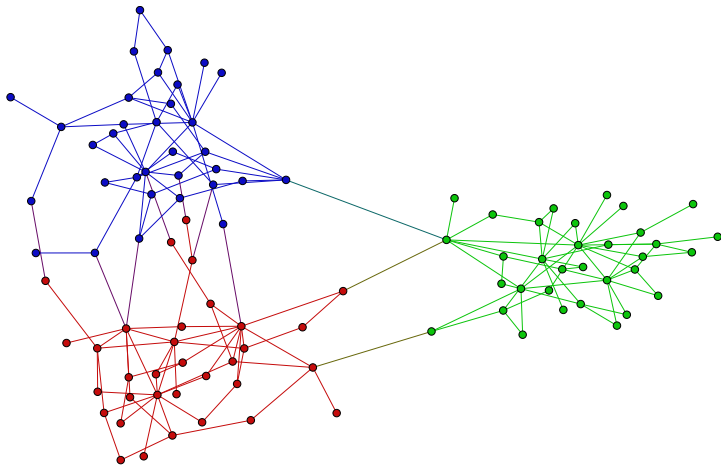
Réseaux réels : Pubmed-Diabètes

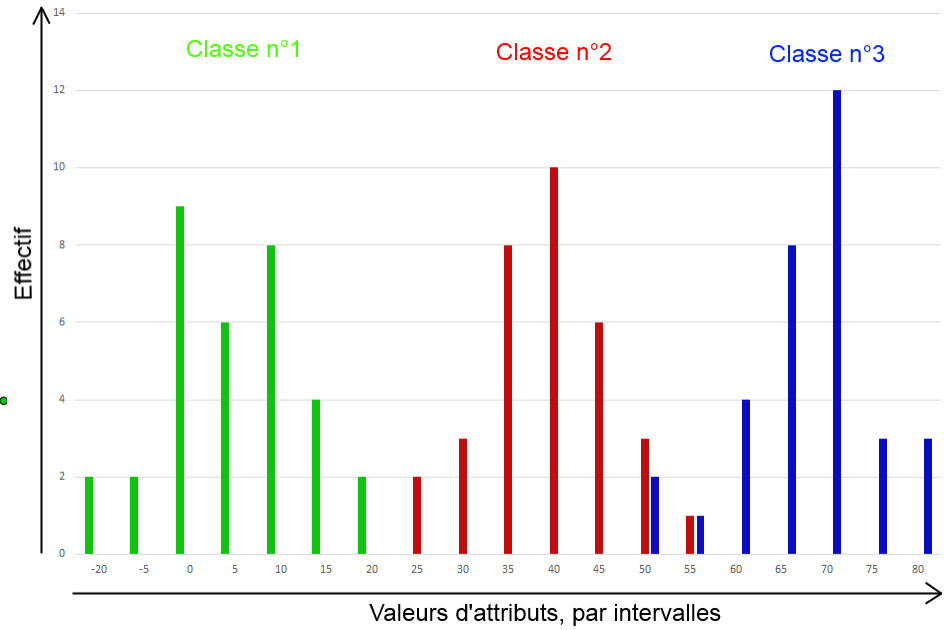
	K-means (k=3)	Louvain	ToTeM	2Mod-Louvain
NMI	**0,27**	0,23	0,20	0,24
V-Mesure	0,18	0,20	0,20	**0,21**
Homogénéité	0,10	0,13	**0,21**	0,14
Complétude	**0,69**	0,39	0,20	0,43

Données

Génération à l'aide d'un modèle de graphe à attributs [Dang et al. 2012]

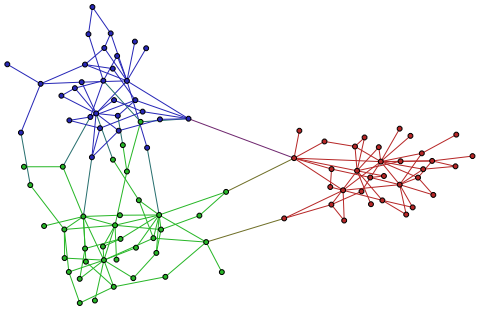
$G = (V, E)$ avec $|V| = 99$ et $|E| = 168$



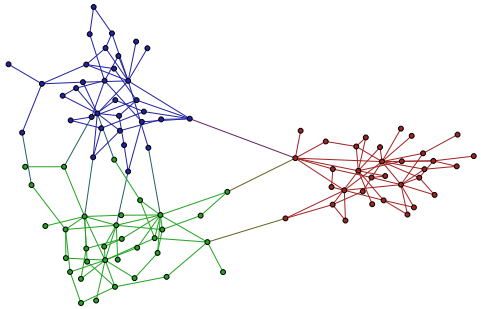


$|C1| = |C2| = |C3| = 33$ $N_{C1}(10, 7)$ $N_{C2}(40, 7)$ $N_{C3}(70, 7)$

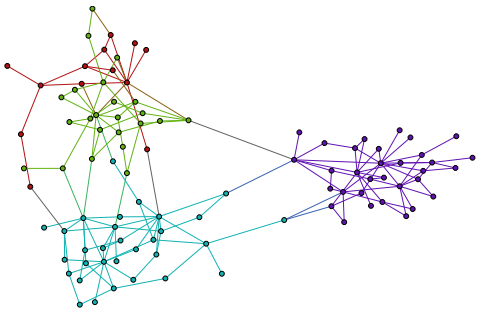
Résultats



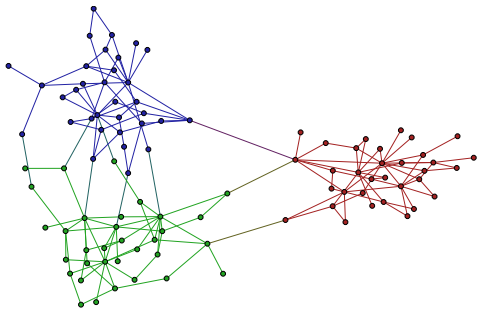
K-means
avec $k = 3$



ToTeM



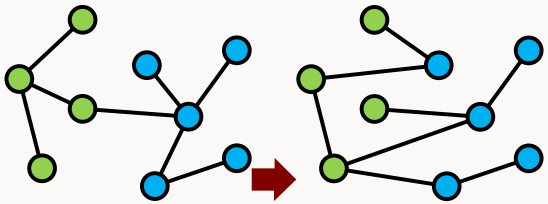
Méthode de Louvain



2Mod-Louvain

	K-means	Louvain	ToTeM	2Mod-Louvain
Nombre de classes	(3)	4	3	3
Taux de biens classés	0,96	0,83	0,95	**0,98**
Info. Mut. Norm. (NMI)	0,90	0,78	0,86	**0,93**

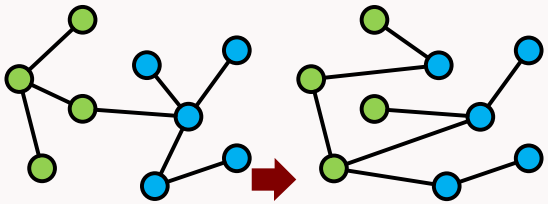
Mesure des conséquences de différentes évolutions du réseau



Dégradation de l'information relationnelle

$degr_{rel} \in (0; 0,25; 0,50)$

Mesure des conséquences de différentes évolutions du réseau



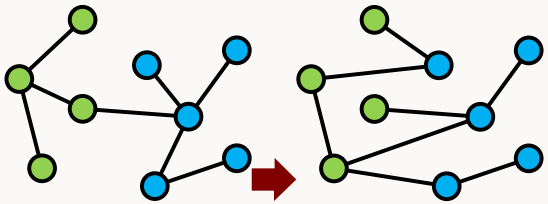
Dégradation de l'information relationnelle

$degr_{rel} \in (0; 0,25; 0,50)$

Dégradation des attributs

$\sigma \in (7; 10; 12)$

Mesure des conséquences de différentes évolutions du réseau

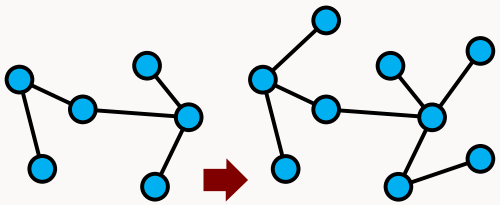


Dégradation de l'information relationnelle

$degr_{rel} \in (0; 0,25; 0,50)$

Dégradation des attributs

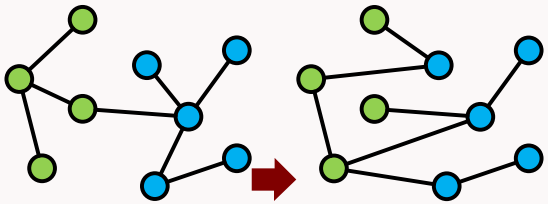
$\sigma \in (7; 10; 12)$



Augmentation de la taille du réseau

$|V| \in (99; 999; 5001)$

Mesure des conséquences de différentes évolutions du réseau

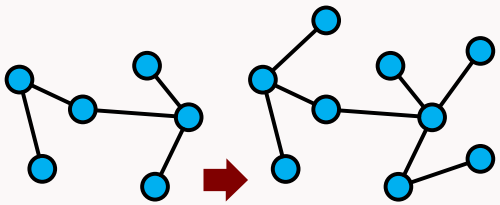


Dégradation de l'information relationnelle

$degr_{rel} \in (0; 0,25; 0,50)$

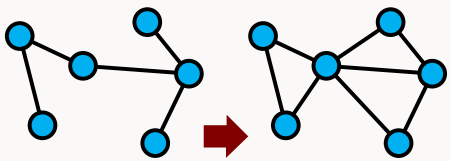
Dégradation des attributs

$\sigma \in (7; 10; 12)$



Augmentation de la taille du réseau

$|V| \in (99; 999; 5001)$



Augmentation du nombre d'arêtes

$|E| \in (168; 315; 508)$

Résultats face à diverses dégradations du réseau (NMI)

NMI	Louvain	K-means	ToTeM	2Mod-Louvain
Graphe de référence				
R	0,78	0,88	0,86	**0,93**
Dégradation de l'information relationnelle ($degr_{rel} = 0$ pour R)				
$degr_{rel} = 0,25$	0,22		0,49	**0,60**
$degr_{rel} = 0,5$	0,12		**0,38**	0,35
Dégradation des attributs ($\sigma = 7$ pour R)				
$\sigma = 10$		0,72	0,82	**0,89**
$\sigma = 12$		0,64	0,57	**0,93**
Augmentation de la taille du réseau ($\|V\| = 99$ pour R)				
$\|V\| = 999$	0,60	**0,88**	0,85	0,80
$\|V\| = 5001$	0,59	**0,89**	0,37	0,77
Augmentation du nombre d'arêtes ($\|E\| = 168$ pour R)				
$\|E\| = 315$	**0,85**		0,80	0,81
$\|E\| = 508$	0,88		**0,92**	**0,92**

Résultats face à diverses dégradations (Taux de bien classés)

TBC	Louvain		K-means	ToTeM		2Mod-Louvain	
	TBC	#classes	TBC	TBC	#classes	TBC	#classes
Graphe de référence							
R	84	4	96	97	3	**98**	3
Dégradation de l'information relationnelle							
$degr_{rel} = 0,25$	0,33	8	NA	0,18	30	**0,78**	5
$degr_{rel} = 0,5$	0,23	9	NA	0,14	36	**0,63**	6
Dégradation des attributs							
$\sigma = 10$	NA		0,90	0,95	3	**0,96**	3
$\sigma = 12$	NA		0,87	0,20	26	**0,98**	3
Augmentation de la taille du réseau							
$\|V\| = 999$	0,50	11	**0,97**	**0,97**	3	0,84	4
$\|V\| = 5001$	0,40	12	**0,98**	0,005	1 518	0,85	4
Augmentation du nombre d'arêtes							
$\|E\| = 315$	**0,96**	3	NA	0,95	3	0,94	3
$\|E\| = 508$	0,97	3	NA	**0,98**	3	**0,98**	3

Plan

Conclusion

Contributions :

- Détection de communautés dans un graphe à attributs à valeurs réelles
- ToTeM : basée sur l'optimisation de l'inertie interclasses et de la modularité.
- 2Mod-Louvain : basée sur une mesure de modularité adaptée aux données vectorielles
- Bons résultats sur les expérimentations

Perspectives

- Pondération automatisée des divers types d'information dans le réseau, dans un cadre supervisé
- Meilleur passage à l'échelle
  - Étude d'heuristiques plus adaptées à la modularité basée sur l'inertie (matrice dense)
- Adaptation aux graphes orientés
- Développement des applications de la modularité basée sur l'inertie

Publications

COMBE, D., LARGERON, C., EGYED-ZSIGMOND, E., & GÉRY, M. (2013). ToTeM : une méthode de détection de communautés adaptée aux réseaux dŠinformation. In *Extraction et gestion des connaissances (EGC 2013)* (pp. 305-310).

COMBE, D., LARGERON, C., EGYED-ZSIGMOND, E., & GÉRY, M. (2012a). Détection de communautés dans des réseaux scientifiques à partir de données relationnelles et textuelles. In *4ième conférence sur les modèles et l'analyse des réseaux : Approches mathématiques et informatiques (MARAMI)*.

COMBE, D., LARGERON, C., EGYED-ZSIGMOND, E., & GÉRY, M. (2012b). Combining relations and text in scientific network clustering. In *First International Workshop on Semantic Social Network Analyis and Design at IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1280-1285).

COMBE, D., LARGERON, C., EGYED-ZSIGMOND, E., & GÉRY, M. (2012c). Getting clusters from structure data and attribute data. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 710-712).

COMBE, D., LARGERON, C., EGYED-ZSIGMOND, E., & GÉRY, M. (2010). A comparative study of social network analysis tools. In *Web Intelligence Virtual Enterprise 2010*.

Merci pour votre attention.

Références (1/3)

- [K. Steinhaeuser et al., 2008] Steinhaeuser, K., & Chawla, N. V. (2008). Community detection in a large real-world social network. Social Computing, Behavioral Modeling, and Prediction, (pp. 168-175).
- [Y.H. Zhou et al., 2009] Zhou, Y., Cheng, H., & Yu, J. X. (2009). Graph clustering based on structural/attribute similarities. Proceedings of the VLDB Endowment, 2(1), (pp. 718-729).
- [Li et al., 2008] Li, H., Nie, Z., Lee, W.-C. W., Giles, C. L., & Wen, J.-R. (2008). Scalable Community Discovery on Textual Data with Relations. Proceedings of the 17th ACM conference on Information and knowledge management (pp. 1203-1212).
- [M. Ester et al.,2006] Ester, M., Ge, R., Gao, B. J., Hu, Z., & Ben-Moshe, B. (2006). Joint Cluster Analysis of Attribute Data and Relationship Data : the Connected k-Center Problem. SIAM International Conference on Data Mining (pp. 25-46). ACM Press.

Références (2/3)

- [F. Moser et al., 2007] Moser, F., Ge, R., & Ester, M. (2007). Joint Cluster Analysis of Attribute and Relationship Data Without A-Priori Specification of the Number of Clusters. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (p. 510).
- [V.D. Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics : Theory and Experiment.
- [Newman et al., 2004] Newman, M., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical review E, 69(2), (pp. 1-16).
- [Combe et al., 2013] Combe, D., Largeron, C., Egyed-Zsigmond, E., & Géry, M. (2013). ToTeM : une méthode de détection de communautés adaptée aux réseaux d'information. Extraction et gestion des connaissances (EGC 2013) (pp. 305-310).
- [Wasserman et al., 1994] Wasserman, S., & Faust, K. (1994). Social network analysis : Methods and applications. Cambridge University Press.
- [Han et al., 2009] Sun Y., Ha J., Zhaoy P., Yi Z., Chengz H., Wu T. RankClus : Integrating Clustering with Ranking for Heterogeneous Information Network Analysis. In Proceedings of the 12th International Conference on Extending Database Technology : Advances in Database Technology (pp. 565-576). ACM.

Références (3/3)

- [MacQueen, 1967] MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (pp. 281Ű297). University of California Press.
- [Ward, 1963] Ward, J. H. Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301), (pp. 236Ű244).
- [Newman, 2004] Newman, M. Detecting community structure in networks. The European Physical Journal B-Condensed Matter and Complex Systems, 38(2), (pp. 321Ű330).
- [Flake et al., 2003] Flake, G. W., Tarjan, R. E., Tsioutsiouliklis, K. Graph clustering and minimum cut trees. Internet Mathematics, 1(4), (pp. 385Ű408).
- [Sen et al., 2008] Sen, P., Namata, G., Bilgic, M., Getoor, L. Collective classification in network data. AI magazine.

Modularité de Newman et Girvan

$$Q_{NG}(\mathscr{P}) = \sum_{(i,i')\in V\times V} \left[\left(\frac{A_{ii'}}{2M} - \frac{k_i}{2M} \cdot \frac{k_{i'}}{2M} \right) \cdot \delta(c_i, c_{i'}) \right]$$

où M est la somme des poids des liens, k_i est le degré du sommet i et δ est la fonction de Kronecker.

Modularité basée sur l'inertie

$$Q_{inertie}(\mathscr{P}) = \sum_{(i,i')\in V\cdot V} \left[\left(\frac{I(V,i)\cdot I(V,i')}{(2N\cdot I(V))^2} - \frac{\|i-i'\|^2}{2N\cdot I(V)} \right) \cdot \delta\left(c_i, c_{i'}\right) \right]$$

où $I(V)$ est l'inertie totale,
et $I(V,i)$ est l'inertie par rapport au point $i \in V$.

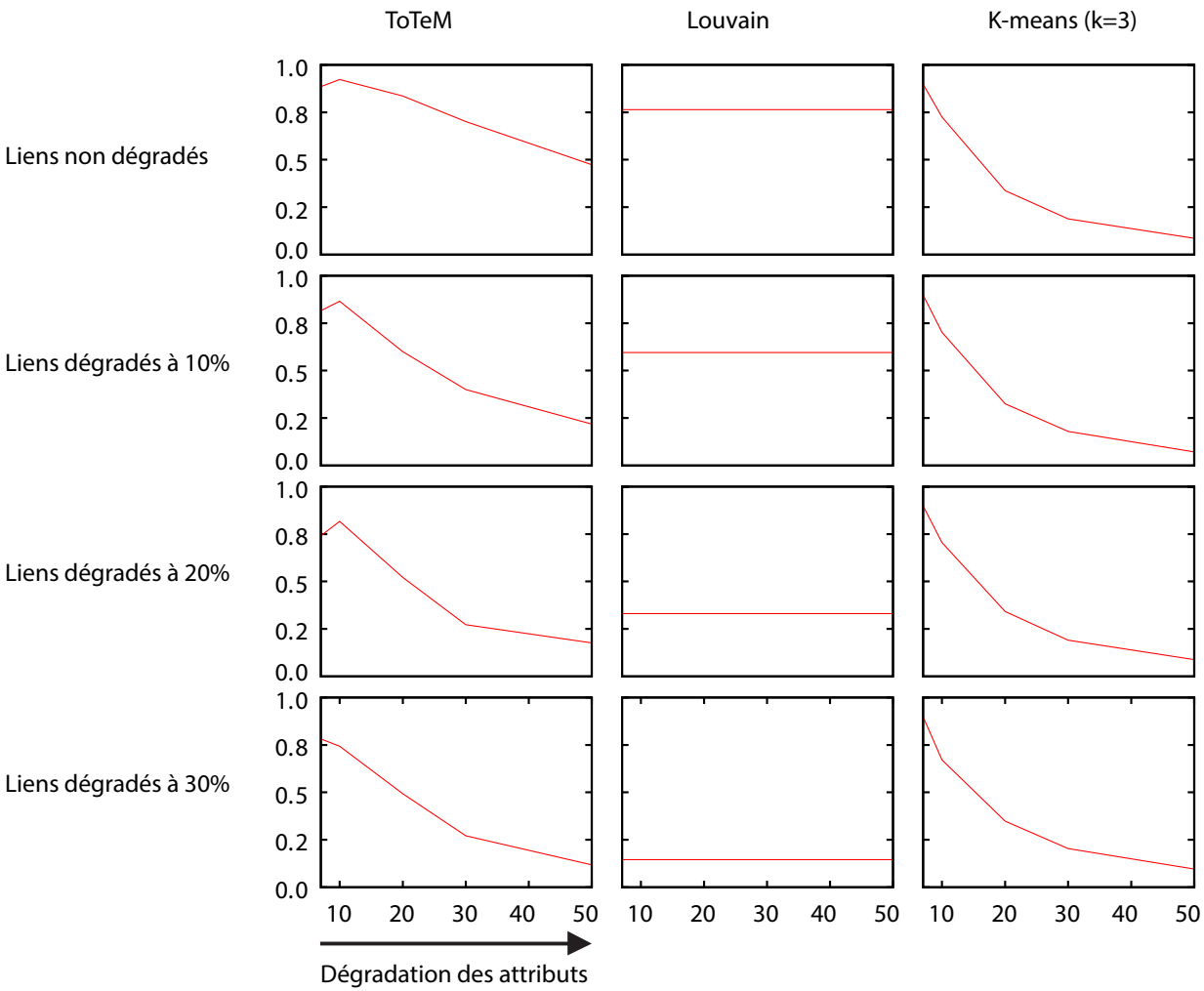
Résultats face à diverses dégradations du réseau

- une dégradation des liens ou des attributs va pénaliser le résultat de la combinaison
- une augmentation du nombre de sommets produit une augmentation importante du nombre de classes produites

Dégradation sur les liens et les attributs simultanément

1. Génération d'un réseau de référence.
2. Dégradation **simultanée** des liens et des attributs.
3. Calcul des résultats sur la base du nombre de sommets bien classés et de l'information mutuelle normalisée (NMI).

Dégradation sur les liens et les attributs simultanément (NMI)



Problèmes de ToTeM

- L'inertie interclasses a une approche différente de la modularité dans le cadre de leur optimisation, qu'a priori aucune normalisation ne peut corriger.
  Exemple selon le critère de modularité, il y a des partitions moins bonnes que la partition discrète.
- Augmentation forte du nombre de classes produites lorsque le réseau est plus grand

On propose de répondre à ce problème par l'introduction d'un critère sur les données vectorielles qui se comporte de manière similaire à la modularité.

Problèmes de ToTeM

- L'inertie interclasses a une approche différente de la modularité dans le cadre de leur optimisation, qu'a priori aucune normalisation ne peut corriger.
  Exemple selon le critère de modularité, il y a des partitions moins bonnes que la partition discrète.
- Augmentation forte du nombre de classes produites lorsque le réseau est plus grand

On propose de répondre à ce problème par l'introduction d'un critère sur les données vectorielles qui se comporte de manière similaire à la modularité.

Complexité des propositions

Parallélisation

Pas de parallélisation de la méthode de Louvain connue à ce jour (déjà rapide donc pas considéré comme nécessaire).
Hadoop sur des graphes : Giraph ?

NMI	ToTeM	2Mod-Louvain	Louvain	K-means
Graphe de référence				
R	0,861	**0,930**	0,784	0,906
Dégradation de l'information relationnelle				
R.1.1	0,489	**0,603**	0,220	
R.1.2	**0,377**	0,353	0,118	
Dégradation des attributs				
R.2.1	**0,819**	0,885		0,747
R.2.2	0,567	**0,930**		0,589
Augmentation de la taille du réseau				
R.3.1	0,854	0,800	0,597	**0,879**
R.3.2	0,376	0,774	0,576	**0,890**
Augmentation du nombre d'arêtes				
R.4.1	0,807	0,816	**0,848**	
R.4.2	**0,917**	**0,917**	0,876	

NMI	Louvain	K-means	ToTeM	2Mod-Louvain
Graphe de référence				
R 0,784	0,883	0,861	**0,930**	
Dégradation de l'information relationnelle				
R.1.1	0,220		0,489	**0,603**
R.1.2	0,118		**0,377**	0,353
Dégradation des attributs				
R.2.1		0,721	0,819	**0,885**
R.2.2		0,637	0,567	**0,930**
Augmentation de la taille du réseau				
R.3.1	0,597	**0,880**	0,854	0,800
R.3.2	0,586	**0,892**	0,376	0,774
Augmentation du nombre d'arêtes				
R.4.1	**0,848**		0,807	0,816
R.4.2	0,876		**0,917**	**0,917**

Résultats face à diverses dégradations (Taux de bien classés)

TBC	Louvain		K-means	ToTeM		2Mod-Louvain	
	TBC (%)	#classes	TBC (%)	TBC (%)	#classes	TBC (%)	#classes
Graphe de référence							
R	84	4	96	97	3	**98**	3
Dégradation de l'information relationnelle ($degr_{rel} = 0$ pour R)							
$degr_{rel} = 0,25$	33	8	NA	18	30	**78**	5
$degr_{rel} = 0,5$	23	9	NA	14	36	**63**	6
Étalement des distributions ($\sigma = 7$ pour R)							
$\sigma = 10$	NA		90	95	3	**96**	3
$\sigma = 12$	NA		87	20	26	**98**	3
Augmentation de la taille du réseau ($\|V\| = 99$ pour R)							
$\|V\| = 999$	50	11	**97**	**97**	3	84	4
$\|V\| = 5001$	40	12	**98**	0,5	1 518	85	4
Augmentation du nombre d'arêtes ($\|E\| = 168$ pour R)							
$\|E\| = 315$	**96**	3	NA	95	3	94	3
$\|E\| = 508$	97	3	NA	**98**	3	**98**	3

Résultats face à diverses dégradations (Taux de bien classés)

TBC	Louvain		K-means	ToTeM		2Mod-Louvain	
	TBC (%)	#classes	TBC (%)	TBC (%)	#classes	TBC (%)	#classes
Graphe de référence							
R	84	4	96	97	3	**98**	3
Dégradation de l'information relationnelle							
$degr_{rel} = 0,25$	33	8	NA	18	30	**78**	5
$degr_{rel} = 0,5$	23	9	NA	14	36	**63**	6
Étalement des distributions							
$\sigma = 10$	NA		90	95	3	**96**	3
$\sigma = 12$	NA		87	20	26	**98**	3
Augmentation de la taille du réseau							
$\|V\| = 999$	50	11	**97**	**97**	3	84	4
$\|V\| = 5001$	40	12	**98**	0,5	1 518	85	4
Augmentation du nombre d'arêtes							
$\|E\| = 315$	**96**	3	NA	95	3	94	3
$\|E\| = 508$	97	3	NA	**98**	3	**98**	3