Community detection in attributed graphs

CHRISTINE LARGERON

Laboratoire Hubert Curien, Université de Saint-Étienne

Journée Graphes et Systèmes sociaux (JGSS) - Avignon

18 Mars 2016

Plan

Plan

Réseau social

- Definition [Wasserman1994]
  ”Social network: finite set or sets of actors and the relation or relations defined on them”
  - Actors : entités, individus, organisations, etc.
  - Relations: amicales, professionnelles, etc.
- Exemples de réseaux
  - Sciences sociales : collaborations, économiques, échanges, etc.
  - Biologie : réseaux de neurones, gènes, protéines,
  - Linguistique : synonymie, co-occurrences, etc

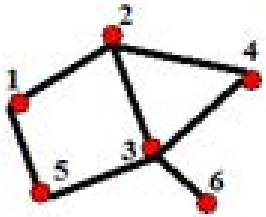
Réseau social

Représentation du réseau social par un graphe $G = (V, E)$

- Chaque entité est un sommet du graphe
  V : l'ensemble fini des sommets de G
- Il existe un lien (arc ou arête) entre deux sommets s'il y a une relation entre les entités correspondantes
  $E \subset V \times V$: l'ensemble des arêtes de G
- A : matrice d'adjacence de G

G = (E,V) with E= {1,2,3,4,5,6}

and V = {(1,2), (1,5), (2,3), (2,4), (3,4), (3,5),(3,6)}



$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Information network

- Definition [Sun2009]
  "A network where each node represents an entity (e.g. actor in a social network) and each link (e.g. tie) a relationship between entities"
  - Each node may have attributes, labels and weights.
  - Link may carry rich semantic information.
- Réseau homogène vs. hétérogène
  - Réseau homogène
    - Un seul type d'entités et un seul type de relation
    - Modèle simple de réseau : WWW
  - Réseau multi-types, hétérogènes
    - Plusieurs types d'entités et / ou de relations
    - Réseau médical : patients, médecins, maladies, traitements
- Tâches
  - Détection de communautés

Détection de communautés

Qu'est ce que c'est une communauté ?

@D.J. Wilson (UNC Chapel Ill)

Détection de communautés

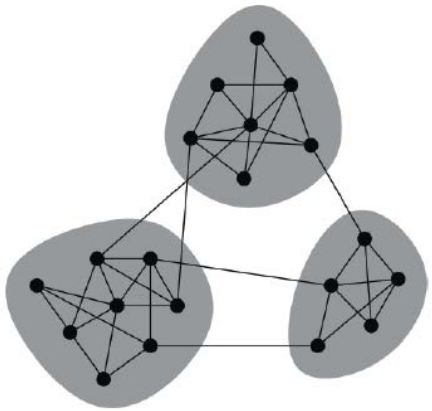
Qu'est ce que c'est une communauté ?

Plan

Cas simple

- Etant donné un réseau social représenté par un graphe $G = (V, E)$ 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 dans une même comunauté soient **fortement connectés**
  - les sommets de communautés différentes soient **peu connectés**

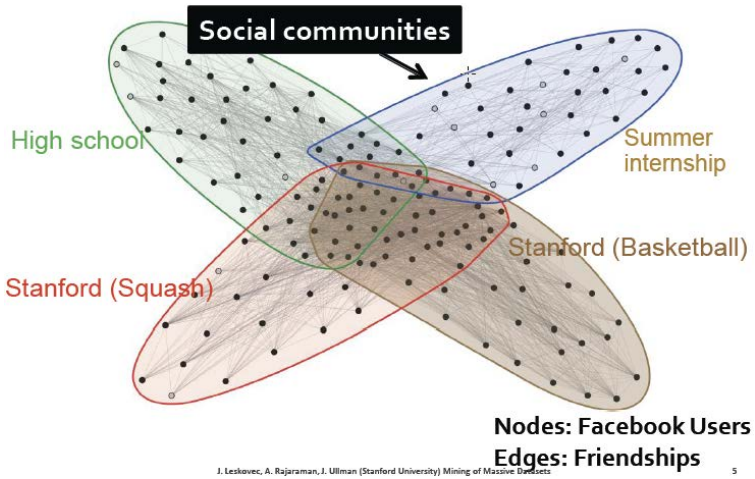


- Définition courante basée sur **les liens** mais qui reste ambigüe et qui n'est pas universelle

Cas complexes

Autres types de structures communautaires

- Communautés recouvrantes [Palla,2005]

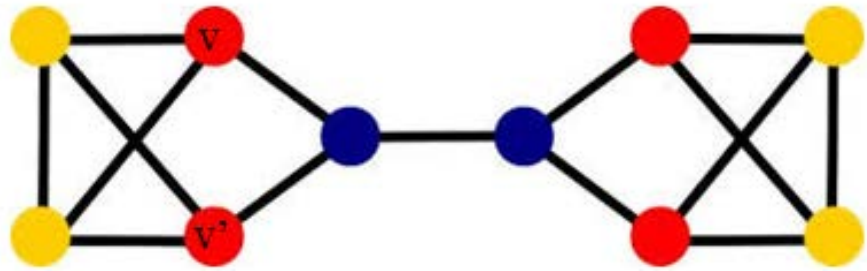


- Non recouvrement de V
- Partition floue
- Ensemble de partitions, éventuellement hiérarchisé

Cas complexes

Autres définitions

- Basée sur l'équivalence structurelle [Wasserman1994]
  Deux sommets appartiennent à la même communauté s'ils ont les mêmes voisins.



v et v' sont dans la même communauté

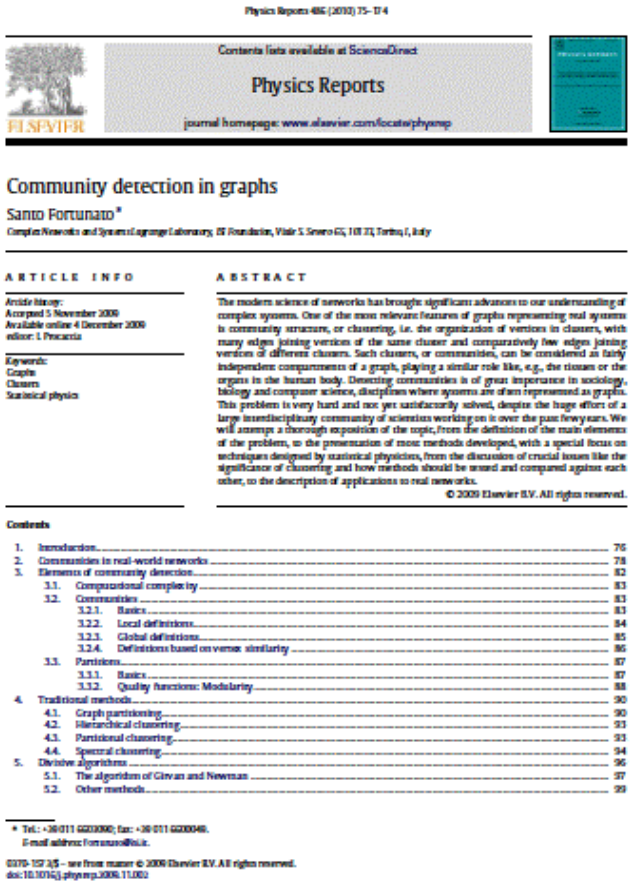
- Basée sur le leadership [Shah2010]
  Deux sommets appartiennent à la même communauté s'ils suivent le même leader.

Méthodes

- Min-cut [Goldberg and Tarjan 1988]
- Spectral clustering [Ng2002]
- Modularité [Newman2004]
- Conductance [Leskovec2009]
- Stochastic Block model [Holland1983]
- Mixed Membership block model [Airoldi2008]
- etc.

Méthodes

Community detection in graphs, Santo Fortunato, Physics Reports 486, 75-174 (2010)



Physics Reports 486 (2010) 75–174

Contents lists available at ScienceDirect

Physics Reports

journal homepage: www.elsevier.com/locate/physrep

Community detection in graphs

Santo Fortunato *

Complex Networks and Systems Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133 Torino, Italy

ARTICLE INFO

Article history:
Accepted 5 November 2009
Available online 4 December 2009
Editor: I. Procaccia

Keywords:
Graphs
Clusters
Statistical physics

ABSTRACT

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, i.e. the organization of vertices in clusters, with many edges joining vertices of the same cluster and comparatively few edges joining vertices of different clusters. Such clusters, or communities, can be considered as fairly independent compartments of a graph, playing a similar role like, e.g., the tissues or the organs in the human body. Detecting communities is of great importance in sociology, biology and computer science, disciplines where systems are often represented as graphs. This problem is very hard and not yet satisfactorily solved, despite the huge effort of a large interdisciplinary community of scientists working on it over the past few years. We will attempt a thorough exposition of the topic, from the definition of the main elements of the problem, to the presentation of most methods developed, with a special focus on techniques designed by statistical physicists, from the discussion of crucial issues like the significance of clustering and how methods should be tested and compared against each other, to the description of applications to real networks.

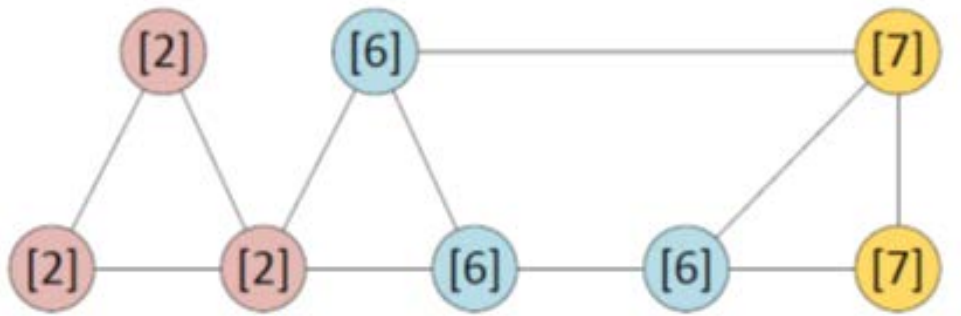
Contents

* Tel.: +39 011 6603060; fax: +39 011 6603049.
E-mail address: fortunato@isi.it.

0370-1573/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.physrep.2009.11.002

Cas du graphe attribué [Zhou2009, Yin2010, Gong2011]

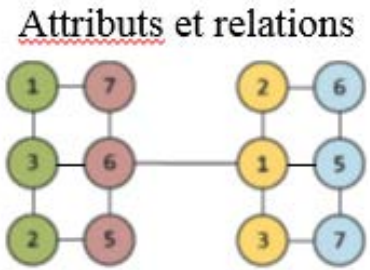
Réseau d'information représenté par un graphe attribué G = (V,E) où un vecteur d'attributs est associé à chaque sommet

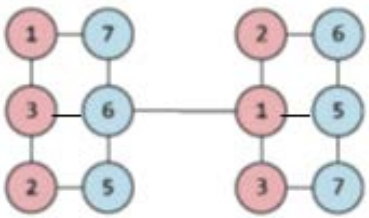


Détection de communautés dans un graphe attribué

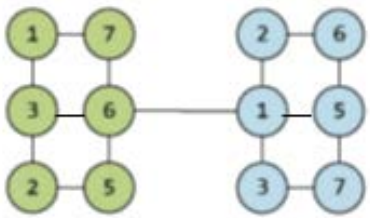
Etant donné un graphe attribué G = (V,E), il s'agit de définir une partition $\mathscr{P} = \{C_1, \ldots, C_r\}$ de V en r classes telle que :

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





Attributs



Relations

Approches méthodologiques

Clustering attributed graphs: models, measures and methods. C. Bothorel et al. (2015)

- Exploitation des attributs puis des relations : enrichissement du graphe
  - Valuation des arêtes à l'aide des attributs [K. Steinhaeuser et al.,2008, Yang,2013 : Codicil]
  - Ajout de sommets et d'arêtes basés sur les attributs [Y.H. Zhou et al., 2009, Li et al. 2011]
- Exploitation des relations puis des attributs
  - Regroupement des communautés en fonction des attributs [Li et al., 2008]

Approches méthodologiques

- Exploitation conjointe des relations et des attributs
  - NetScan, JointClust : K-means avec des contraintes de connexion des classes [M. Ester et al.,2006, F. Moser et al. 2007]
  - Modèles génératifs probabilistes [Phits-Plsa2001, Pcl-dc2009, ppl-dc2013, ppsb-dc2013, CohsMix2010, Bagc2012, Dbagc2014, Cesna2013]
  - Extension de Louvain [V.D. Blondel, J.L. Guillaume, R. Lambiotte, E. Lefevre, 2008]
    - J.D. Cruz Gomez, C. Bothorel, F. Poulet 2011
    - T.A. Dang et E. Viennet,2012
    - ToTeM, **2Mod Louvain**, Combe, 2013-2014

Plan

Définition de la modularité pour données relationnelles [Newman et Girwan, 2004]

Mesure de qualité du partitionnement par rapport aux relations

- Etant donné un graphe $G = (V, E)$ et $\mathscr{P}$ une partition de V

$$Q_{NG}(\mathscr{P}) = \frac{1}{2m}\Sigma_{ii'}\left[(A_{ii'} - \frac{k_i \cdot k_{i'}}{2m})\delta(c_i, c_{i'})\right] \quad (1)$$

où

- A est la matrice d'adjacence de G
- k_i est le degré du sommt $i \in V$
- δ est la fonction de Kronecker.
- $m = Card(V)$

Définition de l'inertie

Mesure de qualité d'un partitionnement par rapport aux attributs réels

- Inertie inter-classe (ou intra-classe)

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

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.

- Adaptée pour la comparaison de partitions de même taille

Modularité pour données vectorielles [Combe2013]

- Mesure de qualité du partitionnement basée sur l'inertie
- Etant donné V un ensemble d'éléments représentés dans R^p et $\mathscr{P}$ une partition de V

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

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

Modularité pour données vectorielles : 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.
- Calcul de la variation de $Q_{inertie}$ induit par le déplacement d'un élément d'une classe vers une autre ne dépend que de l'information locale.

Plan

2Mod-Louvain

- Détection de communautés dans un **graphe à attributs**
- Extension de la méthode de Louvain [Blondel et al.,2008]
- basée sur l'optimisation du critère global $Q_{NG} + Q_{inertie}$
- Calcul incrémental du gain de modularité à partir de l'information locale
- Répétition d'une phase itérative et d'une phase de fusion

Algorithme 2Mod-Louvain

- Initialisation : chaque sommet constitue une communauté

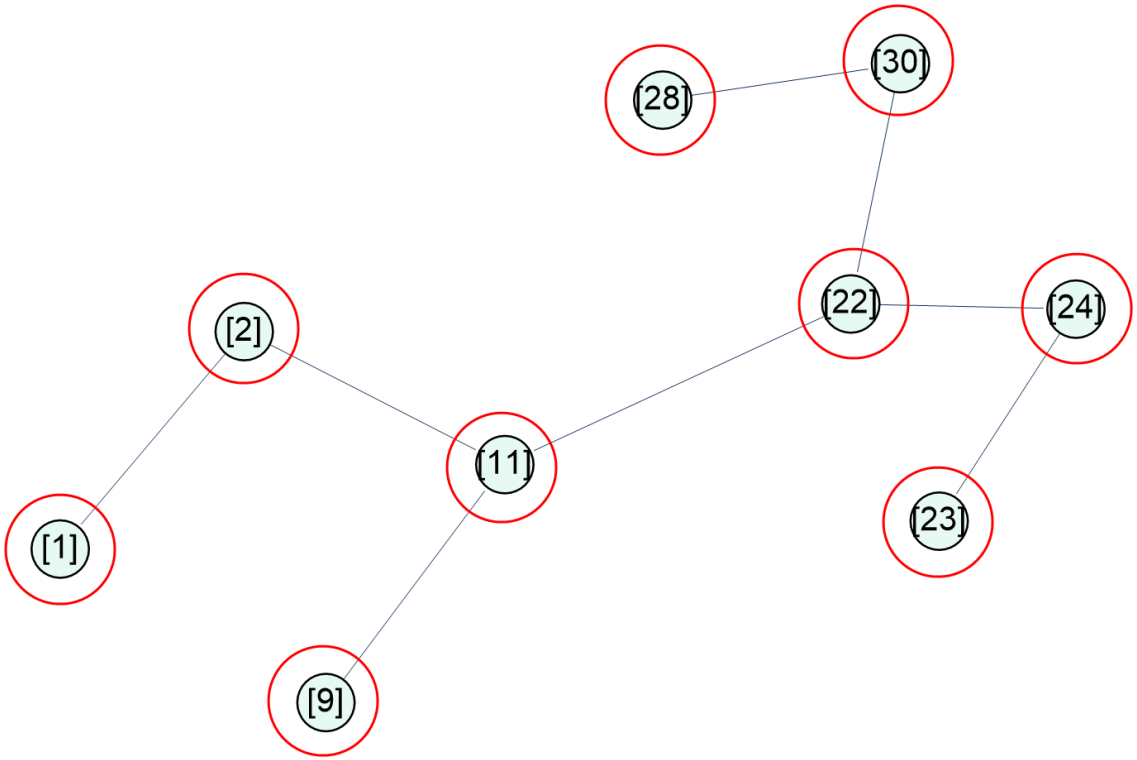


Figure: Initialisation

Algorithme 2Mod-Louvain

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

  jusqu'à ce qu'un maximum local soit atteint

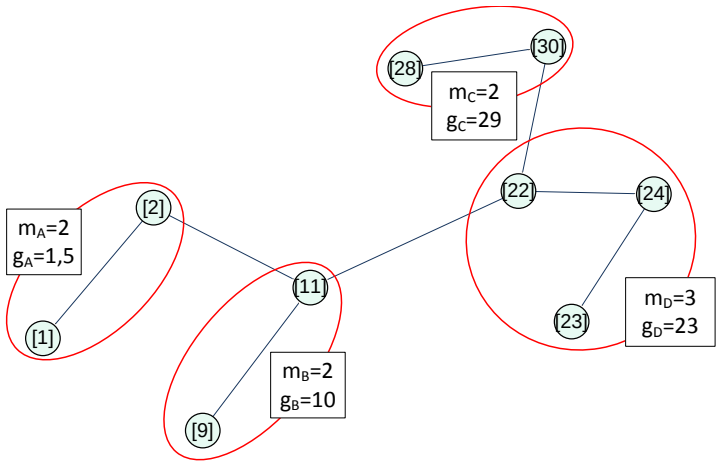


Figure: Fin de phase 1

Algorithme 2Mod-Louvain

- Phase de fusion
  Construction d'un nouveau graphe $G' = (V', E')$ à partir de la partition P'
  - Chaque sommet v de G' correspond à une classe C de P'
  - La valuation entre deux sommets v et v' 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

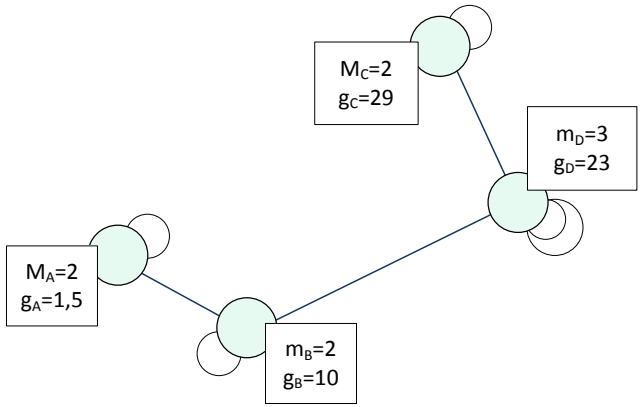


Figure: Fin de phase 2

Plan

Résultats sur données réelles de copublications entre auteurs

$G = (V, E)$ où :

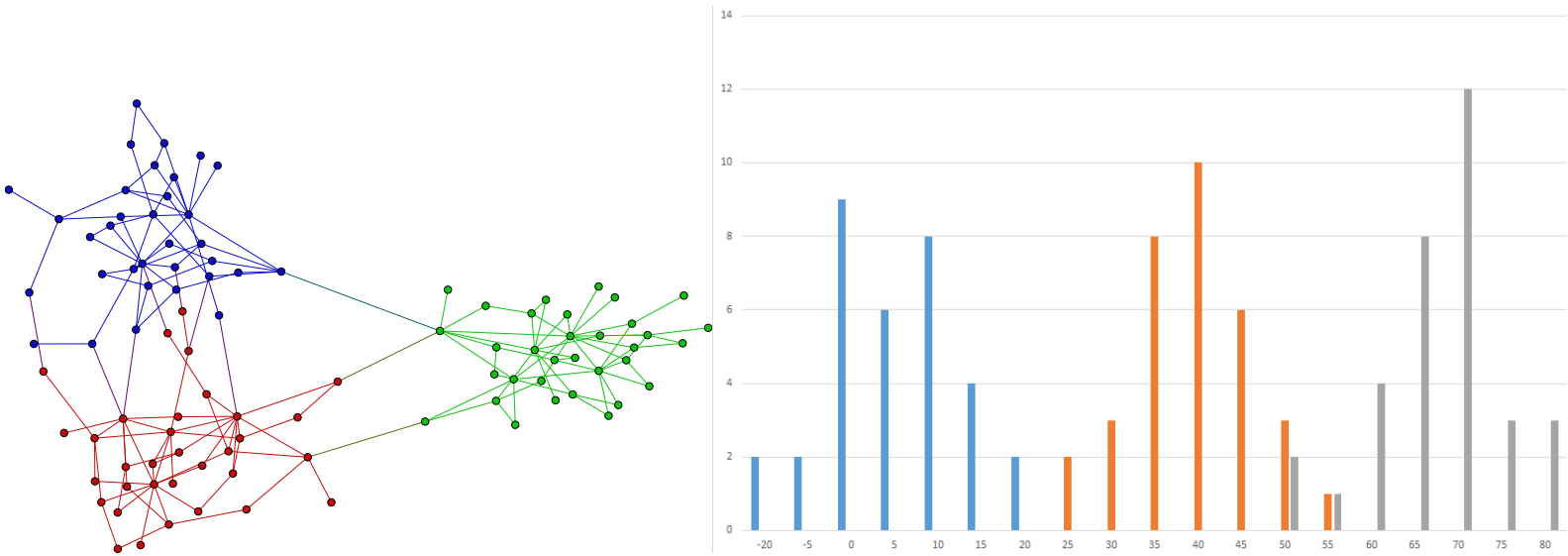
- $|V| = 2515$
- $|E| = 5313$ et il existe un lien entre deux auteurs s'ils ont copubliés au moins un article en informatique dans DBLP (06/18/2014) identifié aussi dans MAS Microsoft Academic Search (02/03/2014).
- 23 attributs correspondant aux nombres de publications par domaine défini Microsoft Academic Search (02/03/2014)
- Communauté réelle : domaine majeur de publication

Table: Evaluation according to the normalized mutual information (NMI)

	Louvain	K-means	ToTeM	I-Louvain
NMI	0.69	0.58	0.69	**0.72**

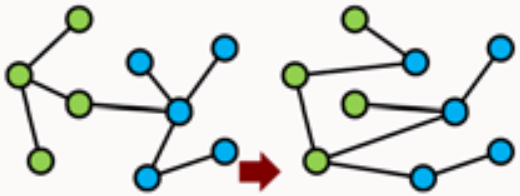
Résultats sur des données artificielles

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



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

Mesure des conséquences de différentes évolutions du réseau

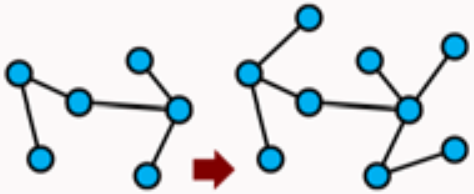


Dégradation de l'information relationnelle

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

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 : taux de bien classés

TBC	Louvain TBC	Louvain #cl.	K-means TBC	ToTeM TBC	ToTeM #cl.	2Mod-Louvain TBC	2Mod-Louvain #cl.
Graphe de référence							
R	84%	4	96%	97%	3	**98%**	3
Dégradation de l'information relationnelle							
$degr_{rel} = 0,25$	33%	8	N/A*	18%	30	**78%**	5
$degr_{rel} = 0,5$	23%	9	N/A*	14%	36	**63%**	6
Étalement des distributions							
$\sigma = 10$	N/A*		90%	95%	3	**96%**	3
$\sigma = 12$	N/A*		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	N/A*	95%	3	94%	3
$\|E\| = 508$	97%	3	N/A*	**98%**	3	**98%**	3

Résultats : NMI

NMI	Louvain	K-means	ToTeM	2mod-Louvain
Graphe de référence				
R	0,784	0,883	0,861	**0,930**
Dégradation de l'information relationnelle				
$degr_{rel} = 0,25$	0,220	N/A*	0,489	**0,603**
$degr_{rel} = 0,5$	0,118	N/A*	**0,377**	0,353
Dégradation des attributs				
$\sigma = 10$	N/A*	0,721	0,819	**0,885**
$\sigma = 12$	N/A*	0,637	0,567	**0,930**
Augmentation de la taille du réseau				
$\|V\| = 999$	0,597	**0,880**	0,854	0,800
$\|V\| = 5001$	0,586	**0,892**	0,376	0,774
Augmentation du nombre d'arêtes				
$\|E\| = 315$	**0,848**	N/A*	0,807	0,816
$\|E\| = 508$	0,876	N/A*	**0,917**	**0,917**

Résultats : temps d'éxécution

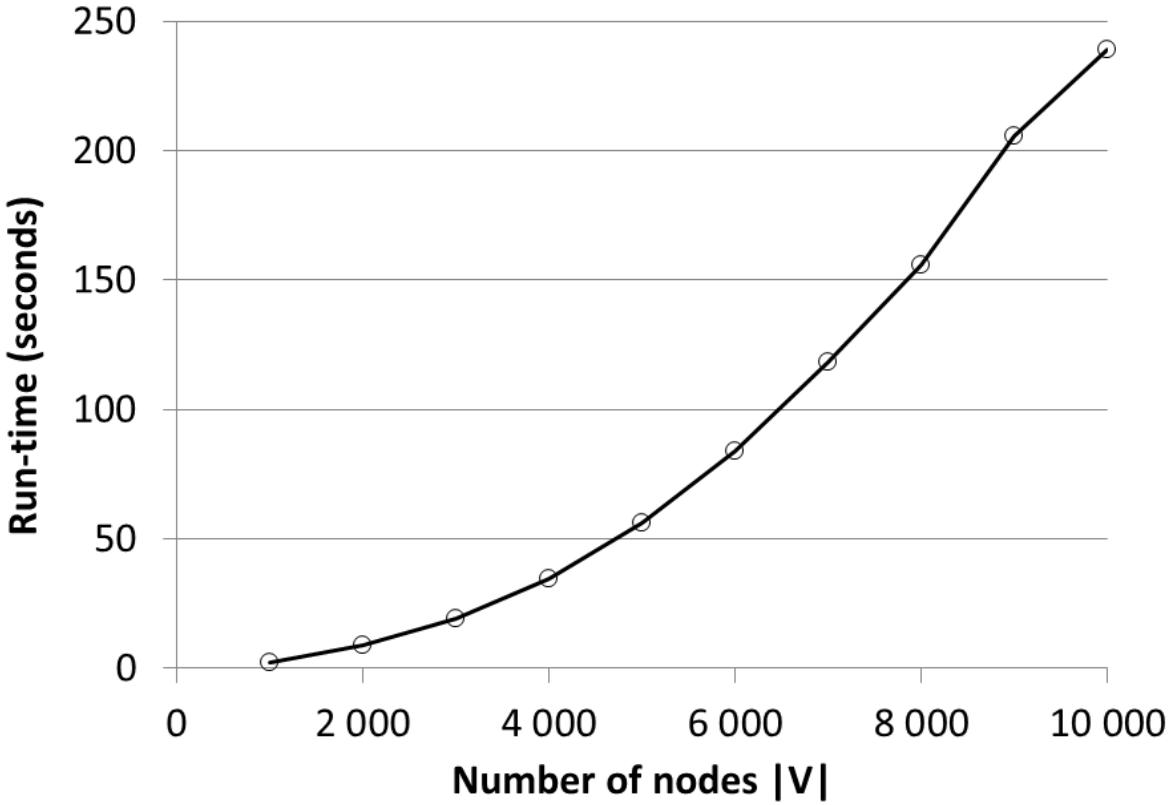


Figure: Run-time of **I-Louvain** on different networks $G = (V, E)$ with $|E| = 3 \times |V|$

Plan

Conclusion

- Détection de communautés dans un graphe à attributs
- Mesure de modularité pour des données vectorielles réelles basée sur l'inertie permettant de comparer des partitions de taille différente
- Méthode 2Mod-Louvain basée sur l'optimisation d'un critère global de modularité
- Résultats encourageants sur jeux de données
- Générateur de réseaux attribués DANC (Largeron, PlosOne2015)
- Intérêt pour traiter d'autres jeux réels

Merci pour votre attention.... des questions ?