Détection de communautés, étude comparative sur graphes réels

Emmanuel Navarro, Rémy Cazabet

IRIT, Université de Toulouse.

11 octobre 2010

Plan

"Communauté" quésako ?

... dans des "grands graphes de terrain" ...

- Ensemble de sommets que l'on peut "distinguer du reste",
- Structure mésoscopique **porteuse de sens**.

graphes lexicaux : **concepts**,
graphes de documents : **thèmes**,
graphes sociaux : **communautés**, **familles**, **clubs**, etc...

Problème de détection de communautés :

- héritier des problèmes de *clustering* (data mining) et de *partitionnement de graphes* (informatique),
- apprentissage non supervisé.
- Article fondateur : [Girvan and Newman, 2002]

"Communauté" quésako ?

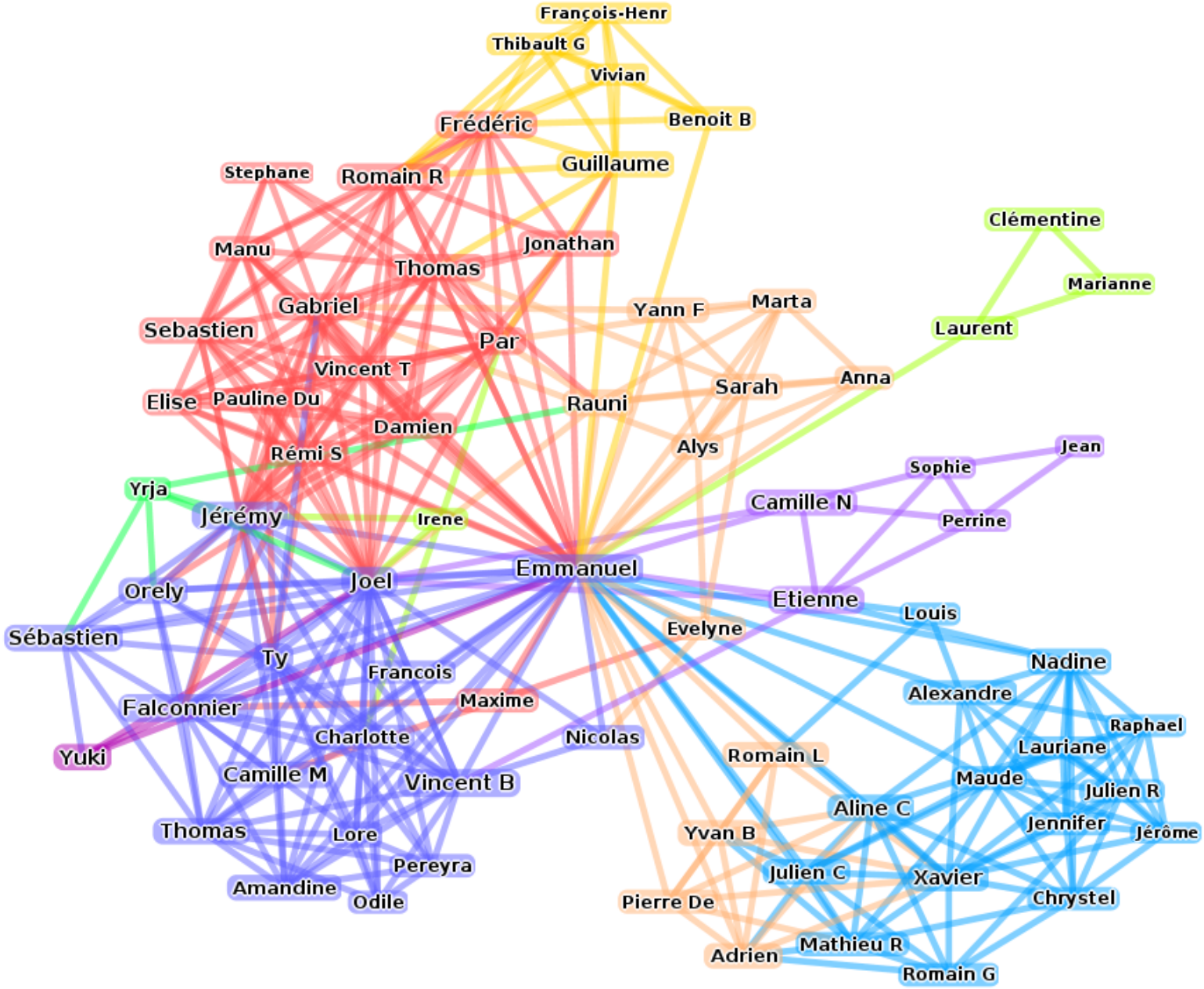
... dans des "grands graphes de terrain" ...

- Ensemble de sommets que l'on peut "distinguer du reste",
- Structure mésoscopique **porteuse de sens**.

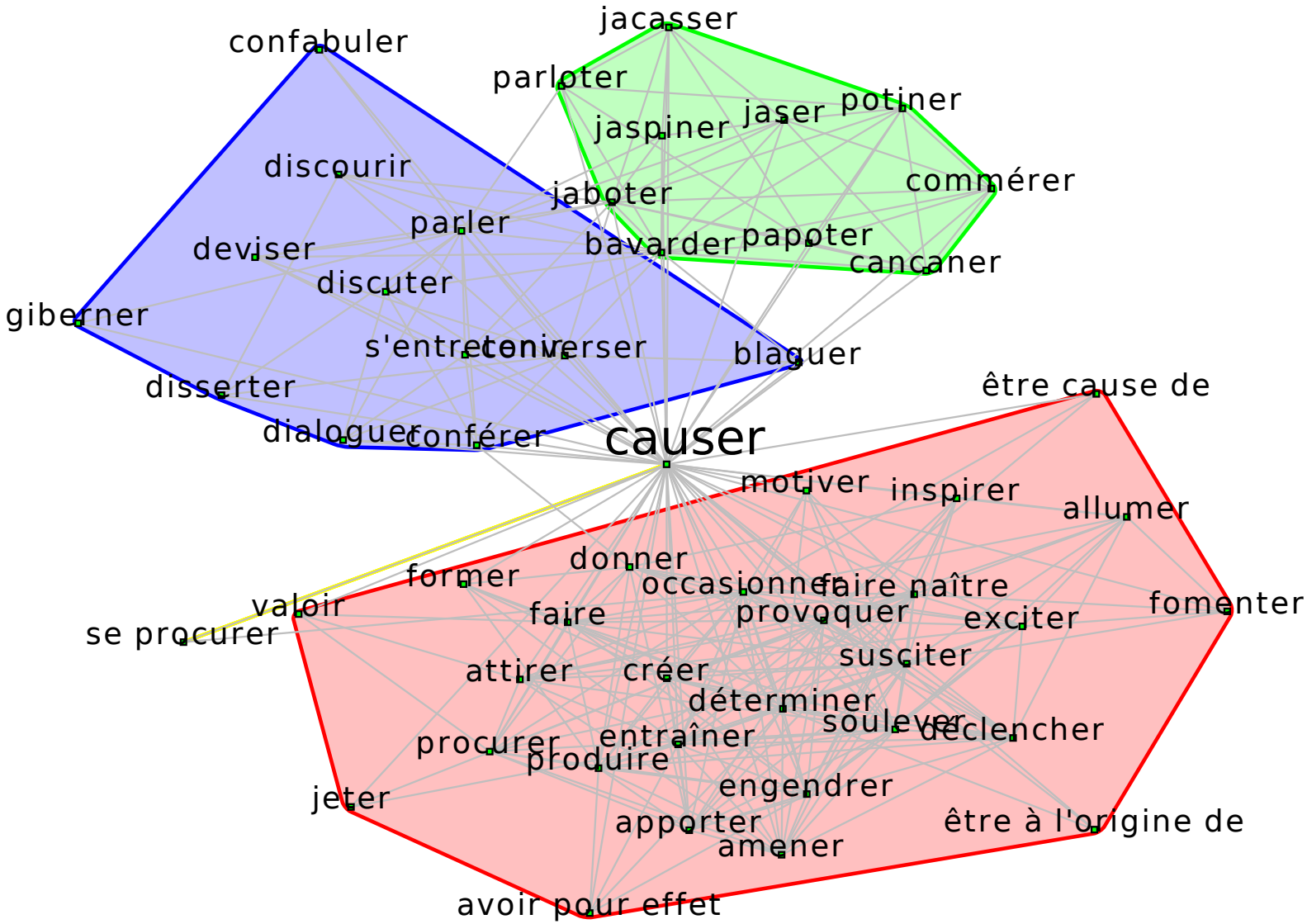
graphes lexicaux : **concepts**,
graphes de documents : **thèmes**,
graphes sociaux : **communautés**, **familles**, **clubs**, etc...

Problème de détection de communautés :

- héritier des problèmes de *clustering* (data mining) et de *partitionnement de graphes* (informatique),
- apprentissage non supervisé.
- Article fondateur : [Girvan and Newman, 2002]



http://apps.facebook.com/touchgraph/



Dicosyn, Walktrap (avec igraph0.6)

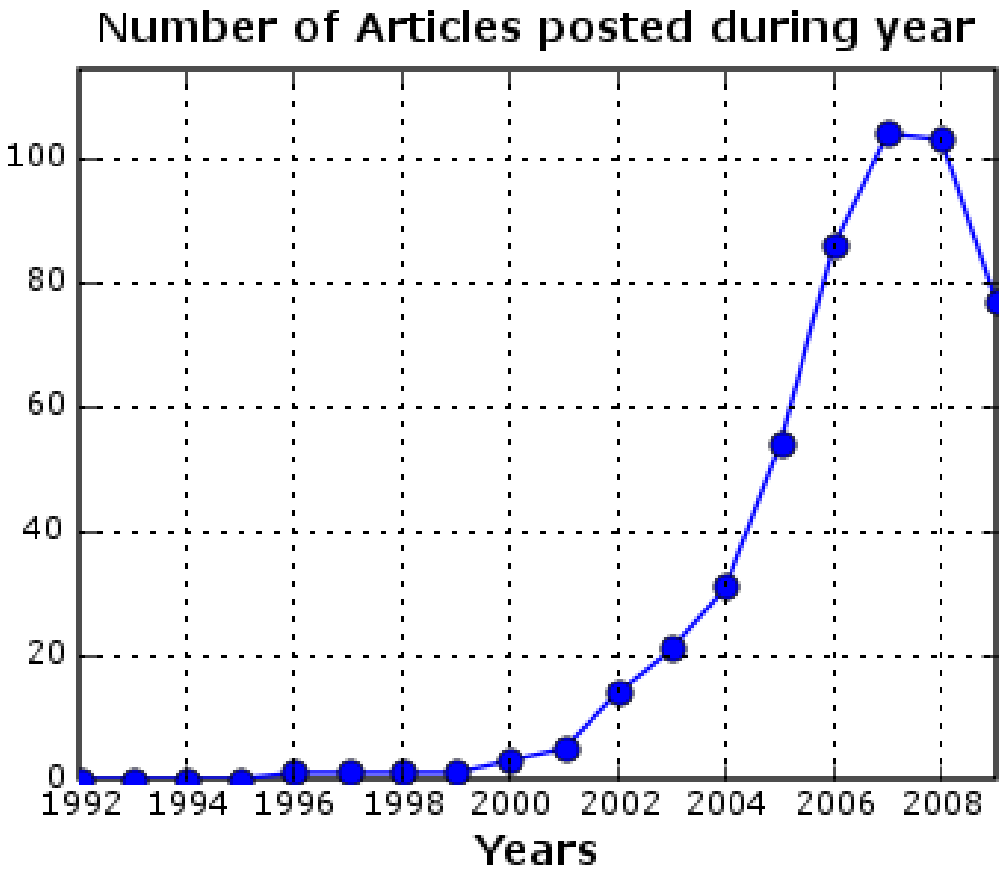
oui mais...

"No definition is universally accepted"
[Fortunato, 2010]

Chaque méthode induit sa propre définition :

- Similarité/distance entre sommets
- Fonction de qualité globale
- Motifs locaux
- Autre ...

"*Community detection*" sur arXiv.org



http://xstructure.inr.ac.ru/x-bin/theme3.py?level=1&index1=447991

Bien des approches...

1. Similarité/distance entre sommets
   - Marches aléatoires : [Pons, 2007]
   - Méthode spectrale : [Donetti and Munoz, 2004]
   - ...
2. Fonction de qualité globale
   - Modularité : [Clauset et al., 2004], [Blondel et al., 2008], ...
   - Marches aléatoires : [Rosvall and Bergstrom, 2008]
   - Modèles statistiques : [Zanghi et al., 2008], ...
   - ...
3. Motifs locaux
   - k-cliques adj. : [Palla et al., 2005]
   - Version "locale" de la modularité : [Clauset, 2005]
   - ...
4. Autre ...
   - Marches aléatoires (MCL) : [van Dongen, 2000]

Bien des approches...

1. Similarité/distance entre sommets
   - Marches aléatoires : [Pons, 2007]
   - Méthode spectrale : [Donetti and Munoz, 2004]
   - ...
2. Fonction de qualité globale
   - Modularité : [Clauset et al., 2004], [Blondel et al., 2008], ...
   - Marches aléatoires : [Rosvall and Bergstrom, 2008]
   - Modèles statistiques : [Zanghi et al., 2008], ...
   - ...
3. Motifs locaux
   - k-cliques adj. : [Palla et al., 2005]
   - Version "locale" de la modularité : [Clauset, 2005]
   - ...
4. Autre ...
   - Marches aléatoires (MCL) : [van Dongen, 2000]

Bien des approches...

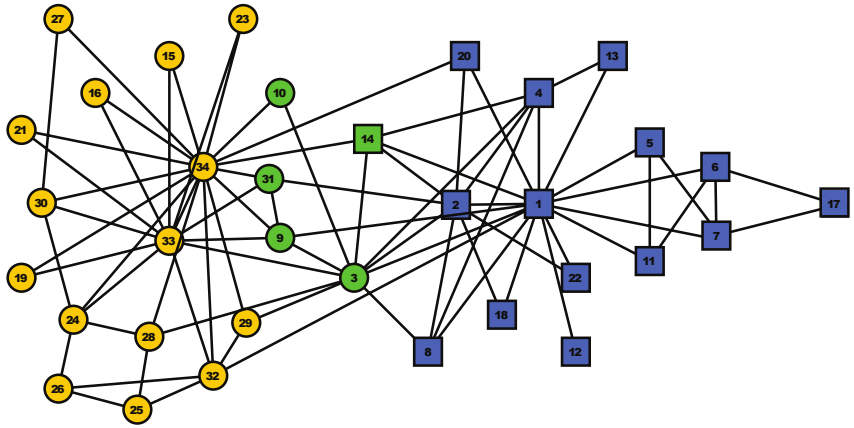
1. Similarité/distance entre sommets
   - Marches aléatoires : [Pons, 2007]
   - Méthode spectrale : [Donetti and Munoz, 2004]
   - ...
2. Fonction de qualité globale
   - Modularité : [Clauset et al., 2004], [Blondel et al., 2008], ...
   - Marches aléatoires : [Rosvall and Bergstrom, 2008]
   - Modèles statistiques : [Zanghi et al., 2008], ...
   - ...
3. Motifs locaux
   - k-cliques adj. : [Palla et al., 2005]
   - Version "locale" de la modularité : [Clauset, 2005]
   - ...
4. Autre ...
   - Marches aléatoires (MCL) : [van Dongen, 2000]

- **Walktrap** [Pons, 2007]
  - Clustering hiérarchique agglomératif,
  - Distance entre sommets basée sur des marches aléatoires.
- **Fastgreedy** [Clauset et al., 2004]
  - Opti. gloutonne de la modularité.
- **Louvain** [Blondel et al., 2008]
  - Autre méthode d'optimisation de la modularité.
- **Infomap** [Rosvall and Bergstrom, 2008]
  - Autre mesure de qualité : basée sur des marches aléatoires,
  - utilise une variante de la méthode d'opti. de Louvain.
- **CFinder** [Palla et al., 2005]
  - Communauté = chaine de k-cliques adj.
  - Recouvrement entre communautés !

- Petits graphes réels,
  - Graphes de *benchmark*.

- n communautés ...
- liens internes $P_{int} = \mu$
- liens externes $P_{ext} = 1 - \mu$
- Capable de retrouver les groupes ?

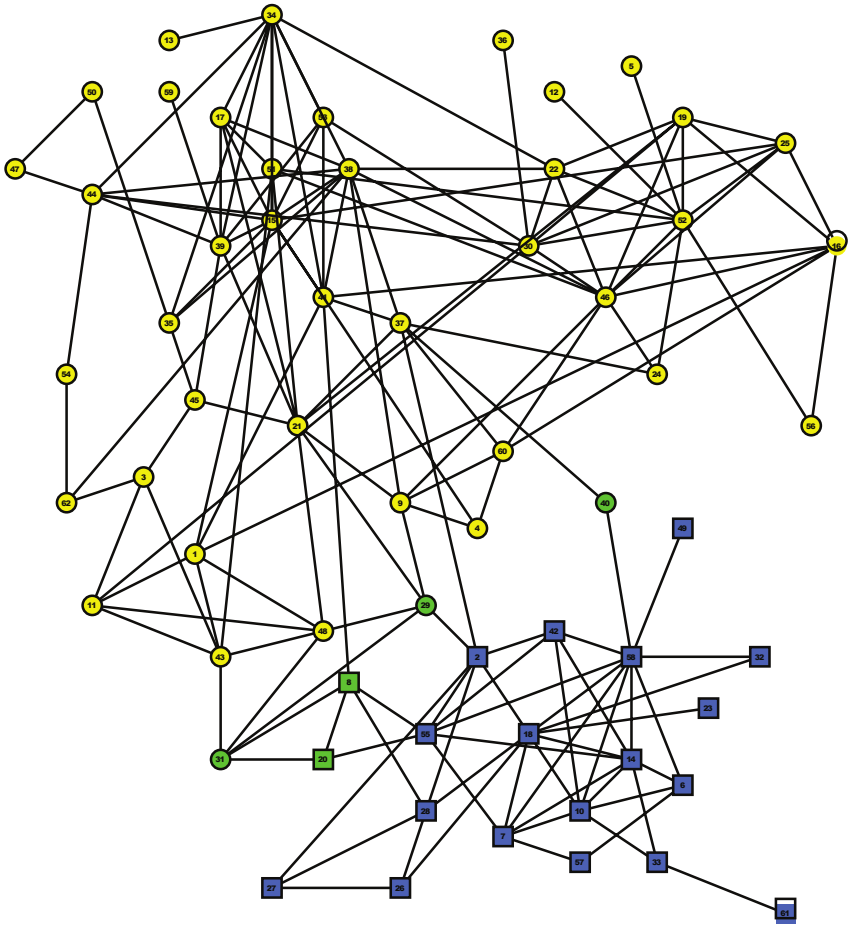
- Petits graphes réels,



karate

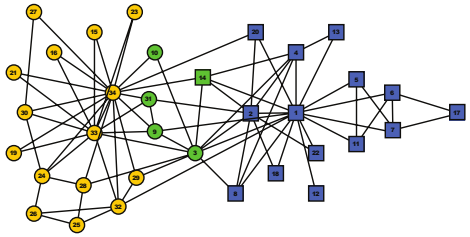
- Graphes de *benchmark*.
  - n communautés ...
  - liens internes $P_{int} = \mu$
  - liens externes $P_{ext} = 1 - \mu$
  - Capable de retrouver les groupes ?

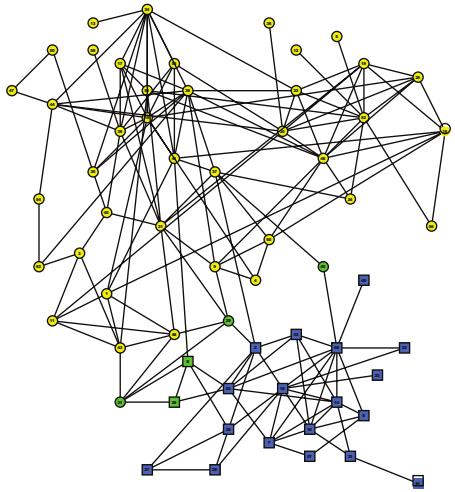
- Petits graphes réels,
- Graphes de *benchmark*.
  - n communautés ...
  - liens internes $P_{int} = \mu$
  - liens externes $P_{ext} = 1 - \mu$
  - Capable de retrouver les groupes ?



Dolphins

- Petits graphes réels,

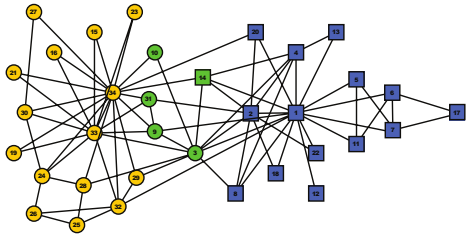


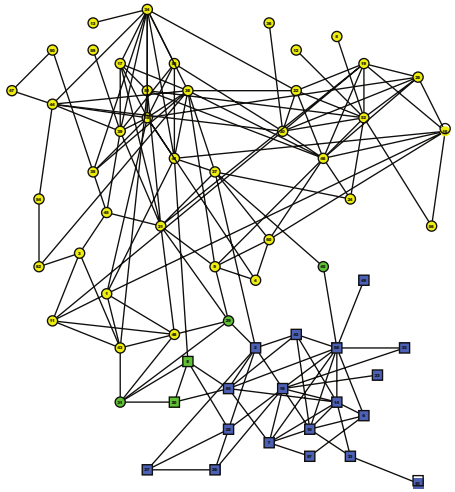


- Graphes de *benchmark*.

- n communautés ...
- liens internes $P_{int} = \mu$
- liens externes $P_{ext} = 1 - \mu$
- Capable de retrouver les groupes ?

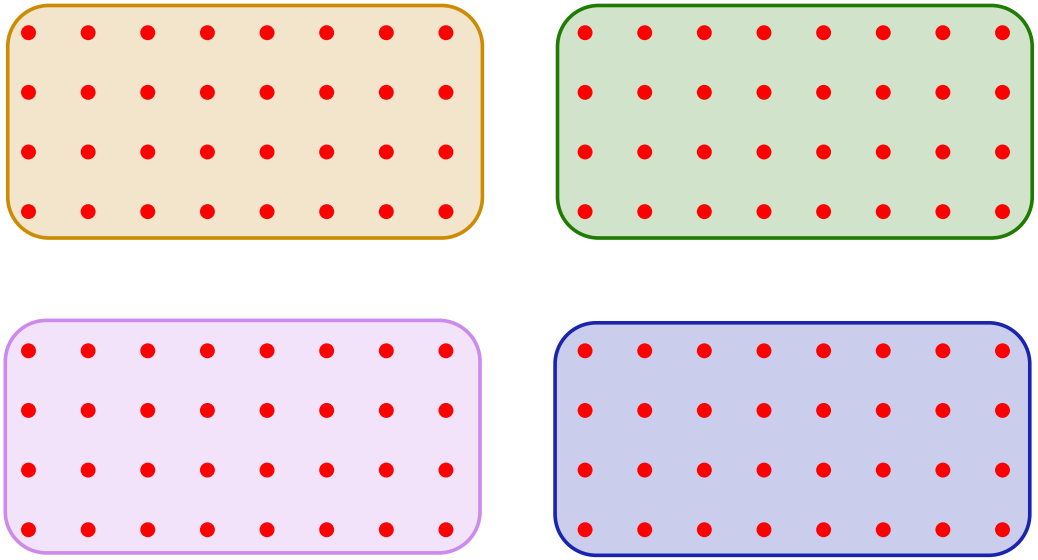
- Petits graphes réels,



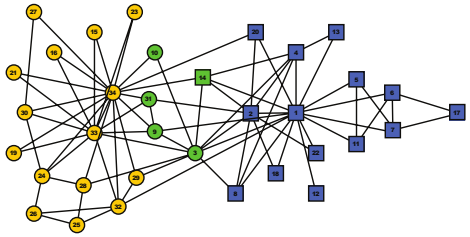


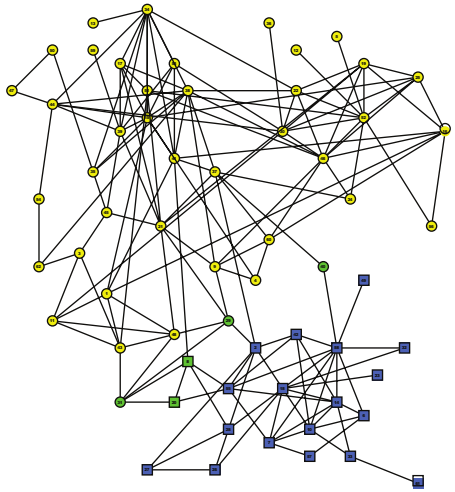
- Graphes de *benchmark*.

- n communautés ...
- liens internes $P_{int} = \mu$
- liens externes $P_{ext} = 1 - \mu$
- Capable de retrouver les groupes ?



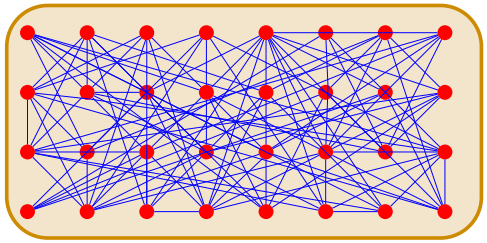
- Petits graphes réels,

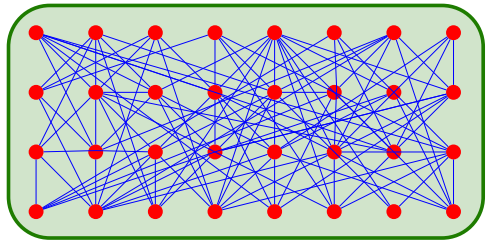


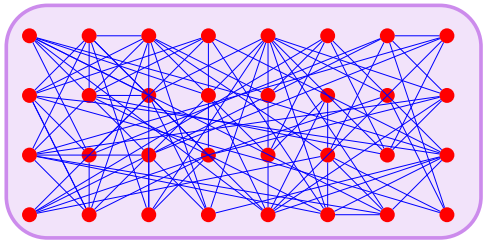


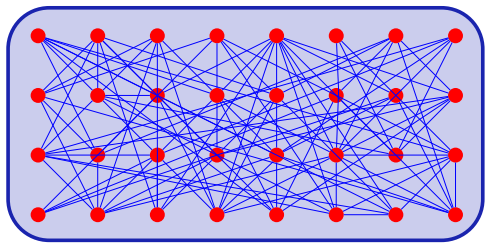
- Graphes de *benchmark*.

- n communautés ...
- liens internes $P_{int} = \mu$
- liens externes $P_{ext} = 1 - \mu$
- Capable de retrouver les groupes ?

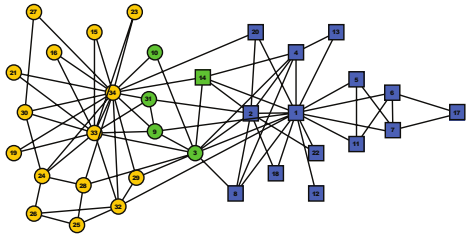


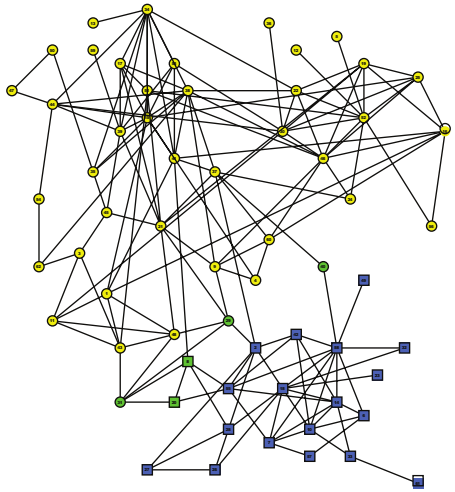






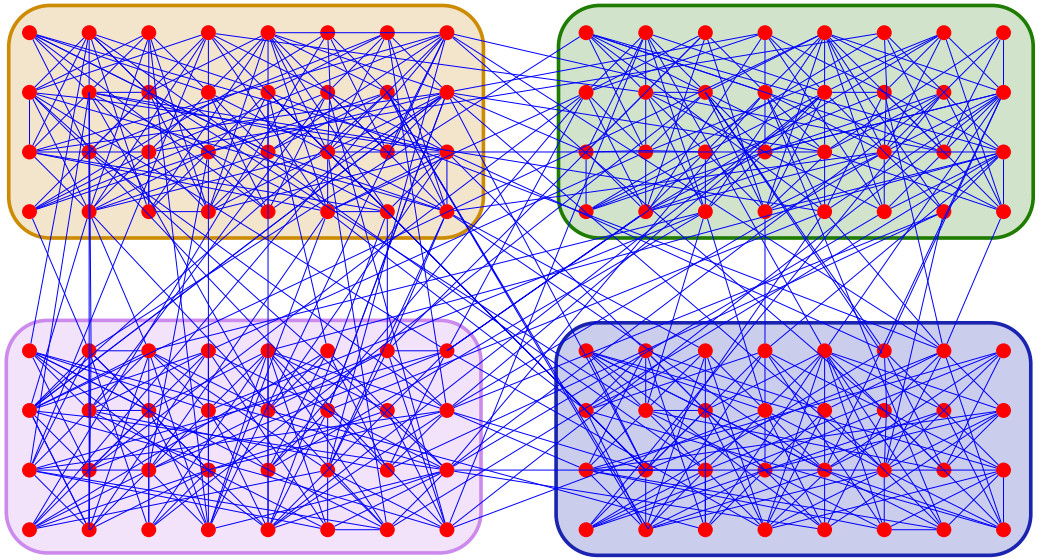
- Petits graphes réels,



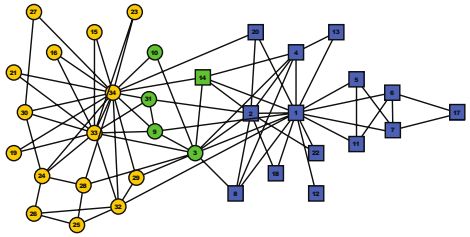


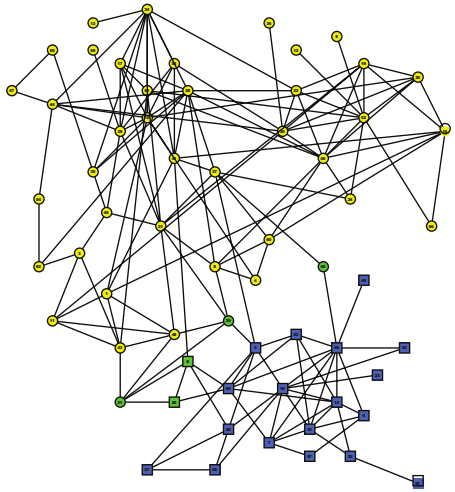
- Graphes de *benchmark*.

- n communautés ...
- liens internes $P_{int} = \mu$
- liens externes $P_{ext} = 1 - \mu$
- Capable de retrouver les groupes ?



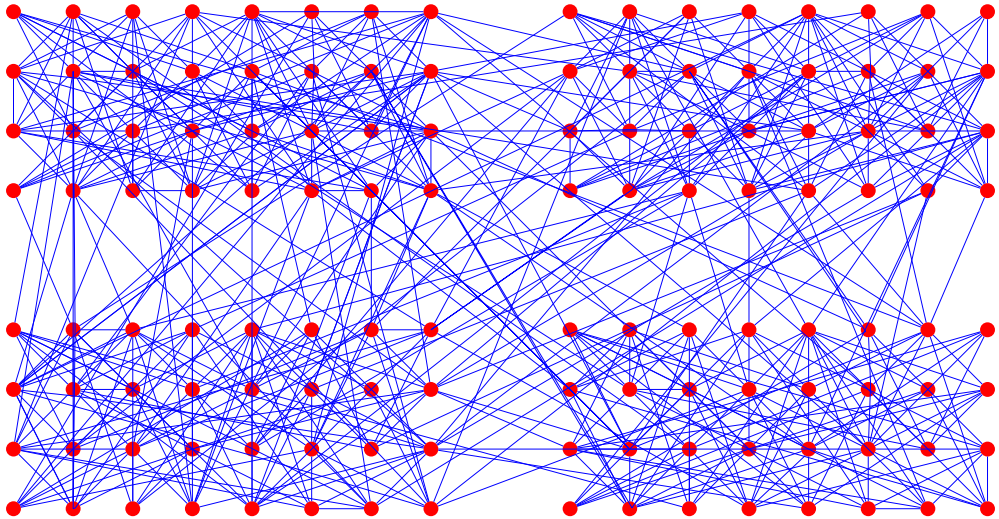
- Petits graphes réels,

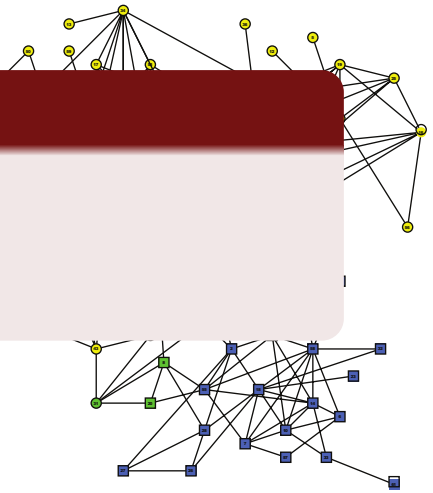




- Graphes de *benchmark*.

- n communautés ...
- liens internes $P_{int} = \mu$
- liens externes $P_{ext} = 1 - \mu$
- Capable de retrouver les groupes ?

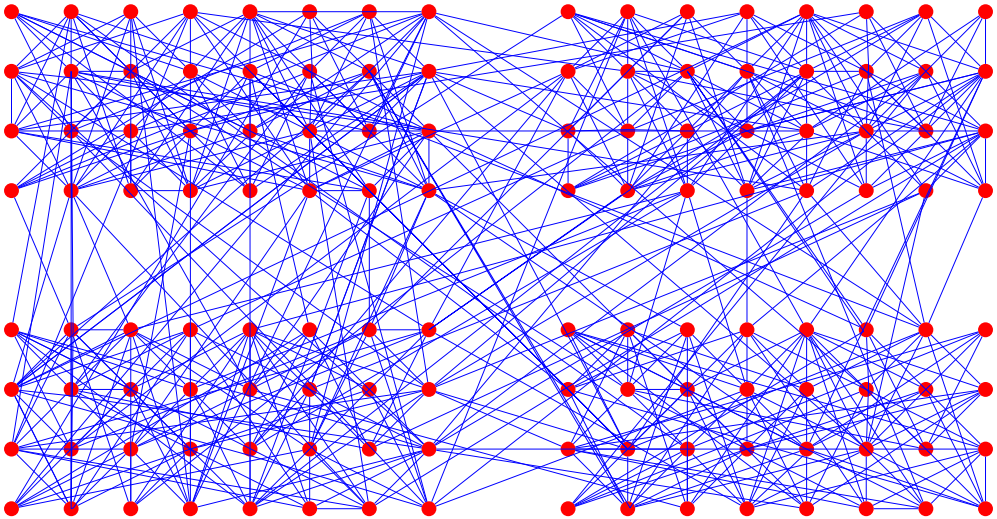




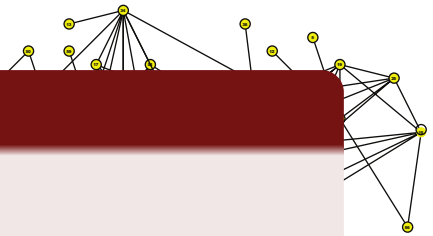
Plusieurs limitations ...

- Graphes réels trop petits !
- Graphes de *benchmark* peu réalistes...

- Graphes de *benchmark*.

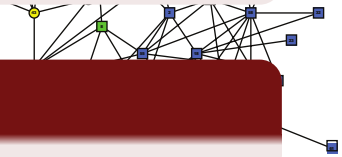


- n communautés ...
- liens internes $P_{int} = \mu$
- liens externes $P_{ext} = 1 - \mu$
- Capable de retrouver les groupes ?



Plusieurs limitations ...

- Graphes réels trop petits !
- Graphes de *benchmark* peu réalistes...

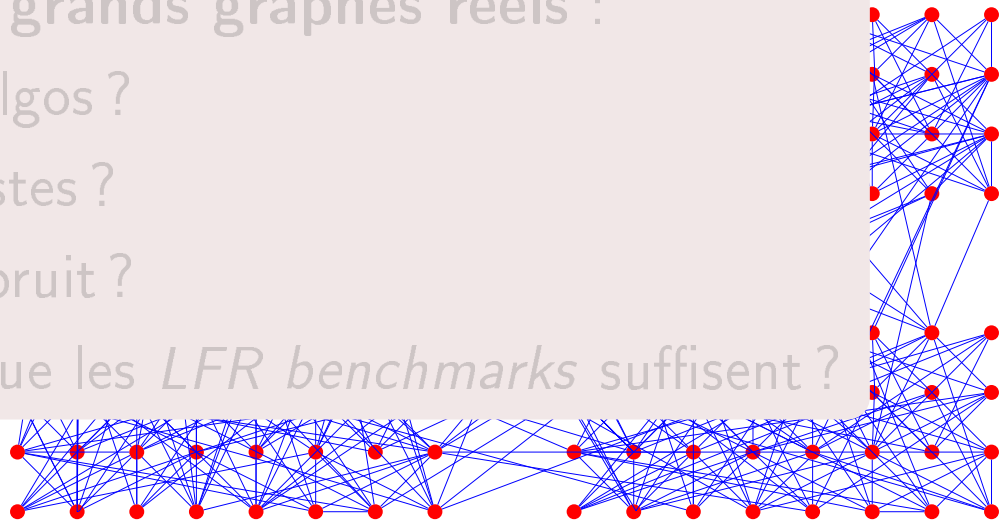


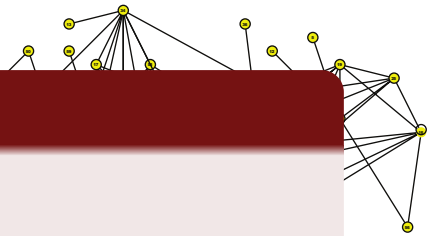
Autres évaluations ?

- *LFR benchmarks* ! [Lancichinetti and Fortunato, 2009]

Ce que l'on propose, sur des **grands graphes réels** :

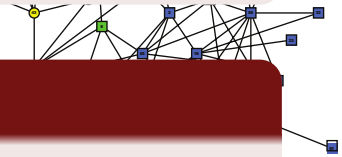
- Expé 1 : Accord entre algos ?
- Expé 2 : Résultats réalistes ?
- Expé 3 : Sensibilité au bruit ?
- Indirectement : est-ce que les *LFR benchmarks* suffisent ?





Plusieurs limitations ...

- Graphes réels trop petits !
- Graphes de *benchmark* peu réalistes...

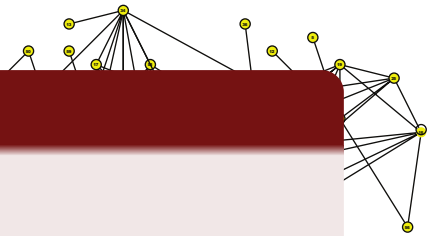


Autres évaluations ?

- *LFR benchmarks* ! [Lancichinetti and Fortunato, 2009]

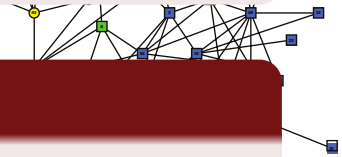
Ce que l'on propose, sur des **grands graphes réels** :

- Expé 1 : Accord entre algos ?
- Expé 2 : Résultats réalistes ?
- Expé 3 : Sensibilité au bruit ?
- Indirectement : est-ce que les *LFR benchmarks* suffisent ?



Plusieurs limitations ...

- Graphes réels trop petits !
- Graphes de *benchmark* peu réalistes...

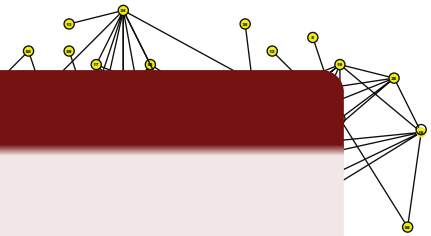


Autres évaluations ?

- *LFR benchmarks* ! [Lancichinetti and Fortunato, 2009]

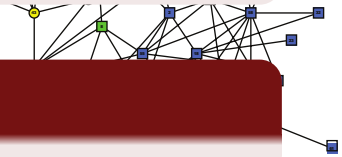
Ce que l'on propose, sur des **grands graphes réels** :

- Expé 1 : Accord entre algos ?
- Expé 2 : Résultats réalistes ?
- Expé 3 : Sensibilité au bruit ?
- Indirectement : est-ce que les *LFR benchmarks* suffisent ?



Plusieurs limitations ...

- Graphes réels trop petits !
- Graphes de *benchmark* peu réalistes...

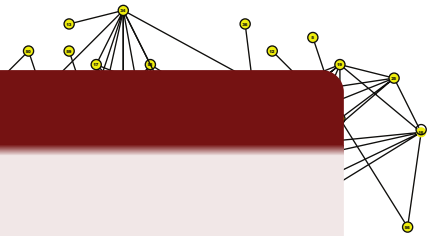


Autres évaluations ?

- *LFR benchmarks* ! [Lancichinetti and Fortunato, 2009]

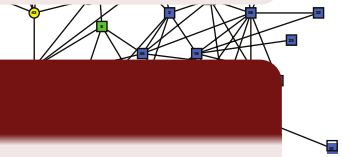
Ce que l'on propose, sur des **grands graphes réels** :

- Expé 1 : Accord entre algos ?
- Expé 2 : Résultats réalistes ?
- Expé 3 : Sensibilité au bruit ?
- Indirectement : est-ce que les *LFR benchmarks* suffisent ?



Plusieurs limitations ...

- Graphes réels trop petits !
- Graphes de *benchmark* peu réalistes...

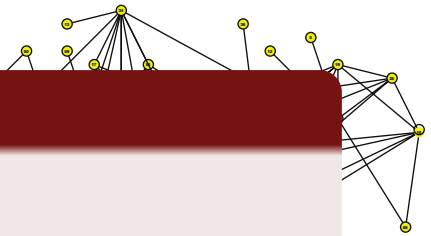


Autres évaluations ?

- *LFR benchmarks* ! [Lancichinetti and Fortunato, 2009]

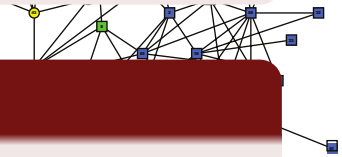
Ce que l'on propose, sur des **grands graphes réels** :

- Expé 1 : Accord entre algos ?
- Expé 2 : Résultats réalistes ?
- Expé 3 : Sensibilité au bruit ?
- Indirectement : est-ce que les *LFR benchmarks* suffisent ?



Plusieurs limitations ...

- Graphes réels trop petits !
- Graphes de *benchmark* peu réalistes...

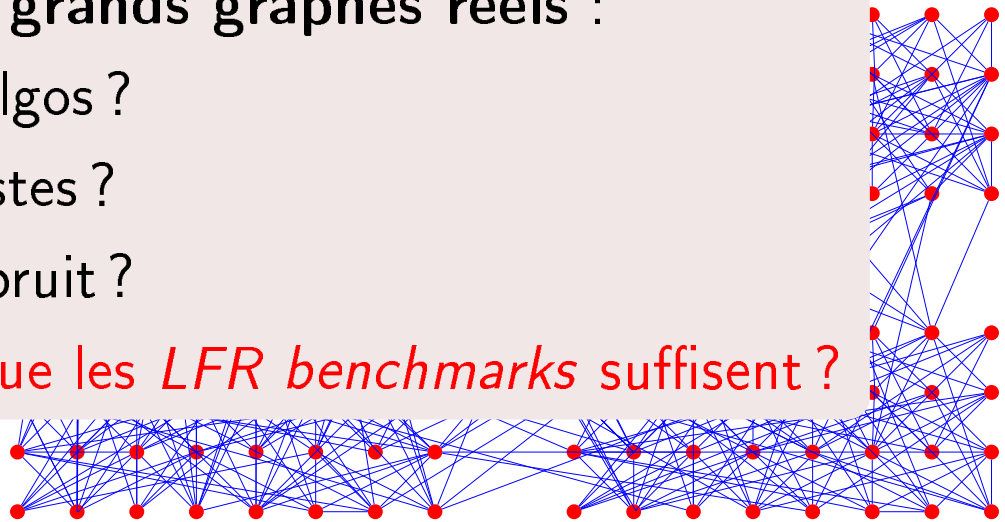


Autres évaluations ?

- *LFR benchmarks* ! [Lancichinetti and Fortunato, 2009]

Ce que l'on propose, sur des **grands graphes réels** :

- Expé 1 : Accord entre algos ?
- Expé 2 : Résultats réalistes ?
- Expé 3 : Sensibilité au bruit ?
- Indirectement : est-ce que les *LFR benchmarks* suffisent ?



Plan

Pour voir, quand même, sur des graphes de bench...

1. Graphe de référence. $n = 5000$, $\langle k \rangle = 15$, $c = [10, 50]$, $\mu = 0,3$
2. Petit graphe. $n = 1000$.
3. Grandes communautés. $c = [30, 100]$,
4. Communautés mal définies. $\mu = 0.5$
5. Graphe très dense $\langle k \rangle = 25$.
6. Graphe peu dense $\langle k \rangle = 7$.

Résultats :

- Globalement, ca marche bien !
- Sauf **FastGreedy** !
- 3, 4, 6 : **CFinder** bof bof... (sur-divise)
- 6 : **Louvain** sur-divise...

Passons à des graphes sérieux...

	n	m	$\langle k \rangle$	n_{lcc}	C	L_{lcc}	λ	r^2
Wiktionary [1]	7 339	8 293	1.13	4 285	0.106	8.98	-2.4	0.94
DicoSyn [2]	9 147	51 423	5.62	8 993	0.142	4.20	-1.9	0.91
CA-HepTh [3]	9 875	25 973	2.63	8 638	0.284	5.95	-2.3	0.91

- Deux réseaux lexicaux, un réseau social,
- Ordres similaires, densités différentes,

1. Synonymie verbe en, [Navarro et al., 2009]
2. Synonymie verbe fr.
3. Co-citations arXiv *"High Energy Physics - Theory"*, [Leskovec et al., 2007].

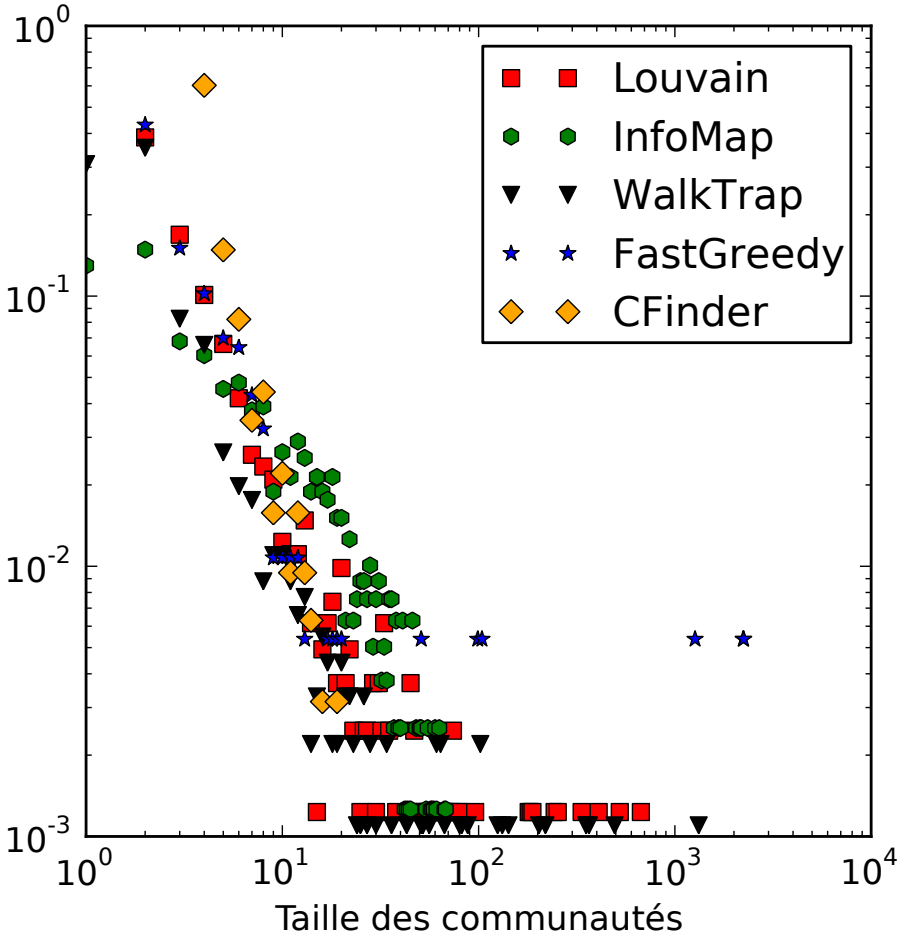
Expé 1 : Accord entre les méthodes ?

	DicoSyn (Verbe, fr)					Wiktionary (Verbe, en)					CA HepTH				
Louvain	.	0.31	0.52	0.50	0.08	.	0.78	0.78	0.70	0.23	.	0.60	0.69	0.56	0.26
InfoMap	0.31	.	0.28	0.44	0.03	0.78	.	0.85	0.80	0.20	0.60	.	0.64	0.62	0.18
WalkTrap	0.52	0.28	.	0.48	0.04	0.78	0.85	.	0.83	0.16	0.69	0.64	.	0.65	0.17
FastGreedy	0.50	0.44	0.48	.	0.02	0.70	0.80	0.83	.	0.09	0.56	0.62	0.65	.	0.09
CFinder	0.08	0.03	0.04	0.02	.	0.23	0.20	0.16	0.09	.	0.26	0.18	0.17	0.09	.

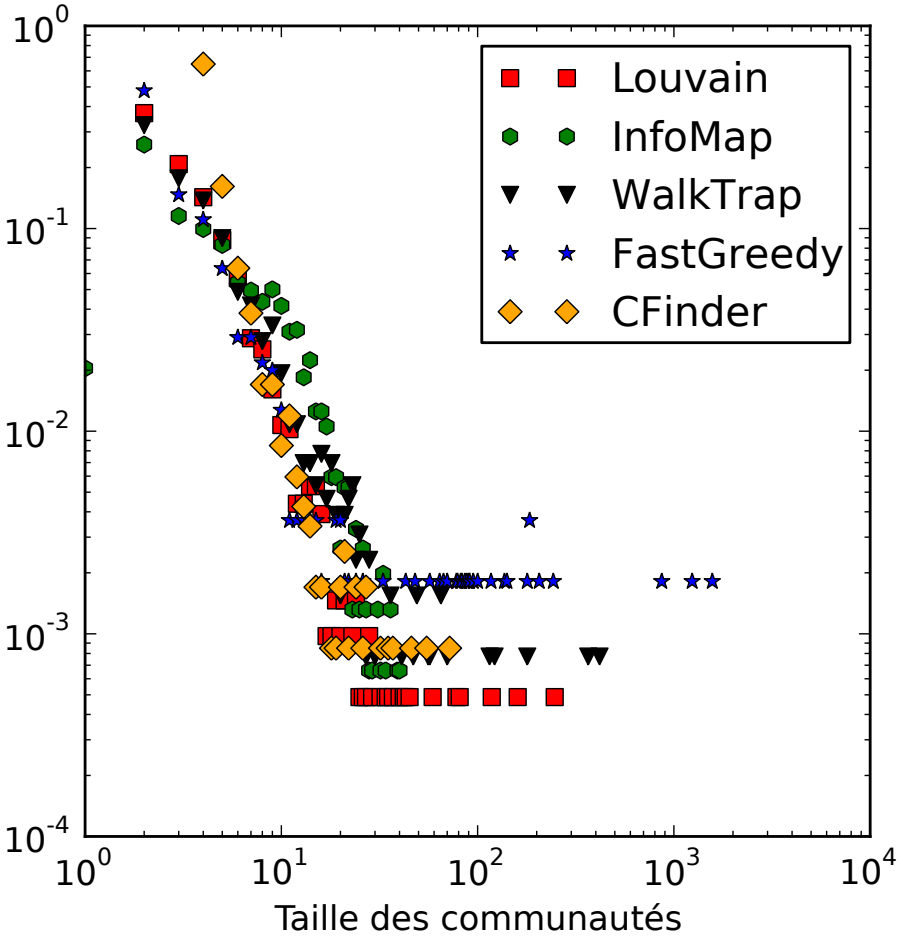
Mesure : variante de l'IMN [Lancichinetti et al., 2009].

- **Accord relativement faible**, *mais* variable en fct des graphes
- **CFinder** diverge (overlaps, nœuds pendants)
- **FastGreedy** catastrophique sur LFR, "pas pire" ici...

Expé 2.1 : Taille des communautés (1/2)



DicoSyn



CA_hepTh

Expé 2.1 : Taille des communautés (2/2)

	DicoSyn		Wiktionary		CA-HepTh	
	N	$\%_{max}$	N	$\%_{max}$	N	$\%_{max}$
Louvain	812	7.7%	2130	0.5%	2052	4.9%
InfoMap	794	1.1%	1609	0.6%	1517	0.8%
WalkTrap	910	18.5%	1490	9.1%	1295	16.9%
FastGreedy	186	26.5%	1075	5.0%	551	15.9%
CFinder	317	46.4%	70	0.2%	1177	3.7%

- Création de "**super-communautés**", peu pertinentes...
- Sauf **InfoMap** :
  - car meilleure fct. de qualité ?
  - ou meilleure méthode d'opti. ? (cherche tjs à re-découper pdt l'opti.)

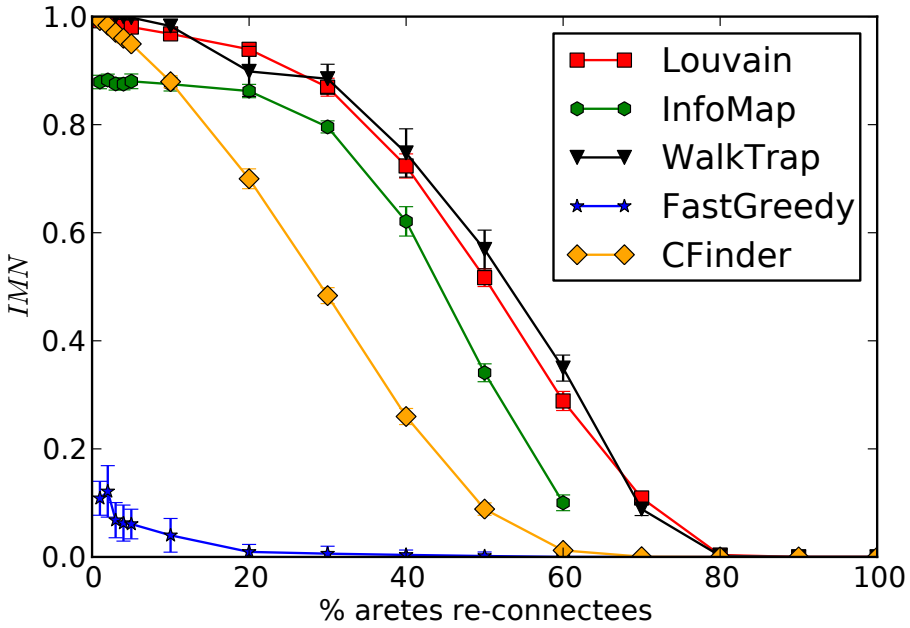
Expé 2.1 : Quel rôle jouent les petites communautés ?

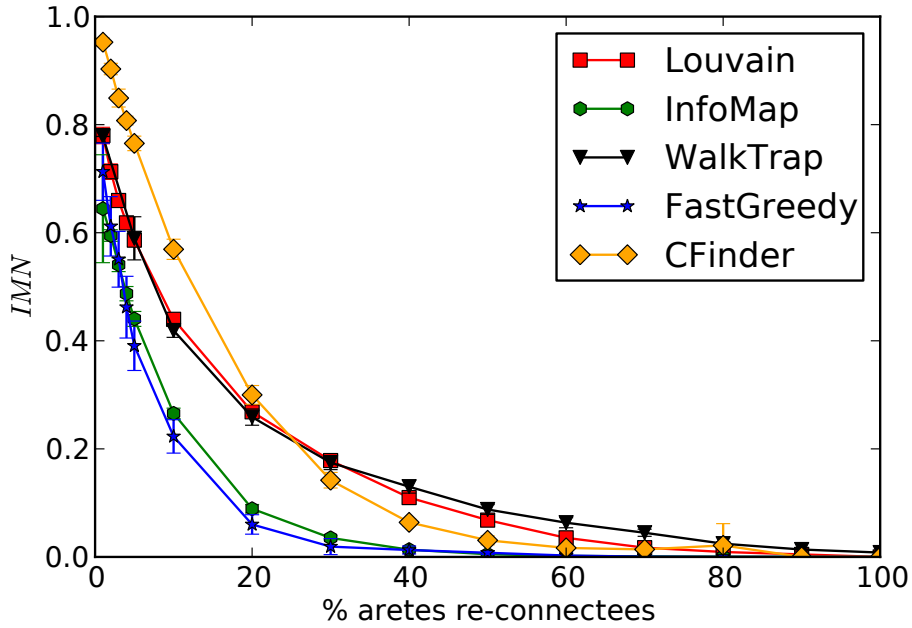
	$\lvert c\rvert > 6$					$\lvert c\rvert \leq 6$				
Louvain	.	0.19	0.27	0.09	0.11	.	0.38	0.57	0.56	0.05
InfoMap	0.19	.	0.20	0.22	0.06	0.38	.	0.34	0.49	0.02
WalkTrap	0.27	0.20	.	0.14	0.06	0.57	0.34	.	0.52	0.03
FastGreedy	0.09	0.22	0.14	.	0.00	0.56	0.49	0.52	.	0.01
CFinder	0.11	0.06	0.06	0.00	.	0.05	0.02	0.03	0.01	.

sur DicoSyn

- Accord plus fort sur les petites communautés
- bcp de petites communautés facile à détecter ?

Expé 3 : Résistance au bruit





LFR puis DicoSyn

Conclusions

- Super-communautés (sauf InfoMap),
- Accord entre algorithmes faible,
- Faible robustesse,
- Comportements différents : LFR vs. graphes réels.

Des questions...

- est-ce juste un pb. d'**overlaps** ?
- des méthodes sont-elles à abandonnées ?

Perspectives

Ce n'est qu'un premier travail...

- Plus de **graphes** et plus d'**algo** !
- Graphes LFR plus "proches" des graphes réels...
- Différents **recablages** (conservant distr. degrés),
- **Méthodes extrêmes** : cliques max, et composantes connexes,
- Résultats sur **graphe complet vs. graphe "nettoyé"**
- **Mesure d'accord** entre partitionnements,
  quelque chose de plus "parlant" ?

Merci !

Questions ?


Blondel, V. D., Guillaume, J., Lambiotte, R., and Lefebvre, E. (2008).
Fast unfolding of communities in large networks.
Journal of Statistical Mechanics : Theory and Experiment, 10.


Clauset, A. (2005).
Finding local community structure in networks.
Physical Review E, 72(2) :026132.


Clauset, A., Newman, M. E. J., and Moore, C. (2004).
Finding community structure in very large networks.
Phys. Rev. E, 70(6).


Donetti, L. and Munoz, M. A. (2004).
Detecting network communities : a new systematic and efficient algorithm.
Journal of Statistical Mechanics : Theory and Experiment, 2004(10) :P10012.


Fortunato, S. (2010).
Community detection in graphs.
Physics Reports, 486(3-5).


Girvan, M. and Newman, M. E. J. (2002).
Community structure in social and biological networks.
Proceedings of the National Academy of Sciences of the United States of America, 99(12) :7821–7826.


Lancichinetti, A. and Fortunato, S. (2009).
Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities.
Phys. Rev. E, 80(1) :016118.


Lancichinetti, A., Fortunato, S., and Kertész, J. (2009).
Detecting the overlapping and hierarchical community structure in complex networks.
New Journal of Physics, 11(3) :033015.


Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007).
Graph evolution : Densification and shrinking diameters.
ACM Trans. Knowl. Discov. Data, 1(1) :2.


Navarro, E., Sajous, F., Gaume, B., Prévot, L., Hsieh, S., Kuo, I., Magistry, P., and Huang, C.-R. (2009).
Wiktionary and NLP : Improving synonymy networks.
In *Proceedings of the 2009 ACL-IJCNLP (workshop)*, pages 19–27, Singapore.


Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005).
Uncovering the overlapping community structure of complex networks in nature and society.
Nature, 435(7043) :814–818.

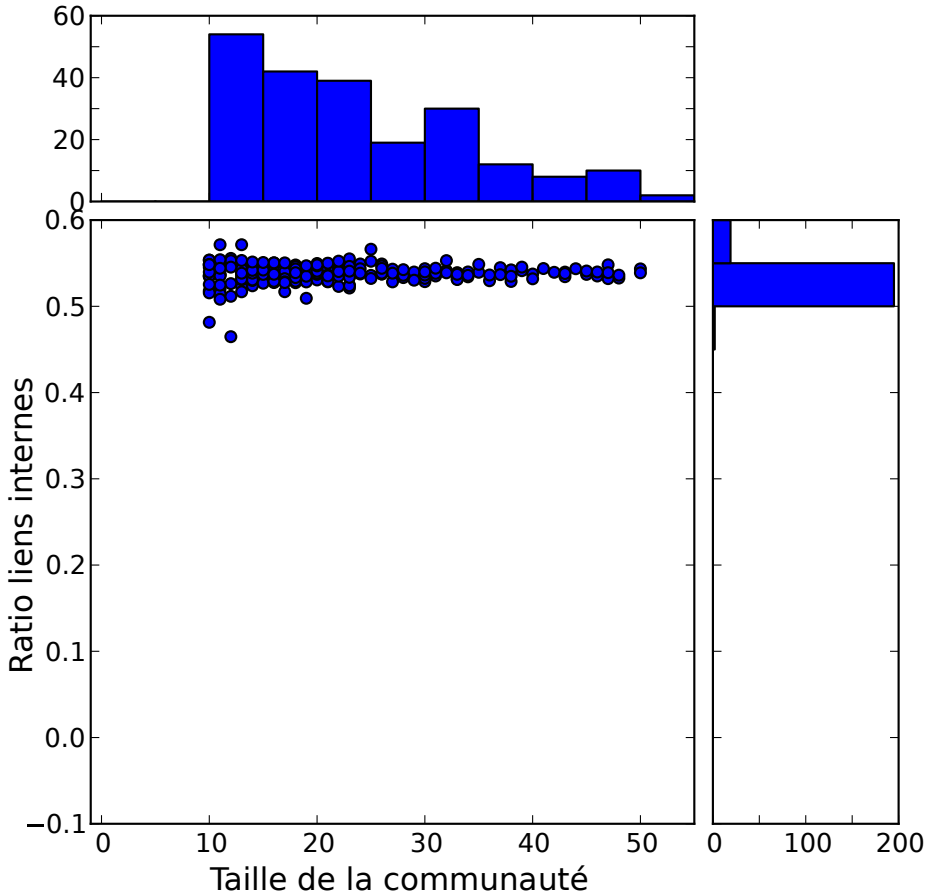

Pons, P. (2007).
Détection de communautés dans les grands graphes de terrain.
PhD thesis.

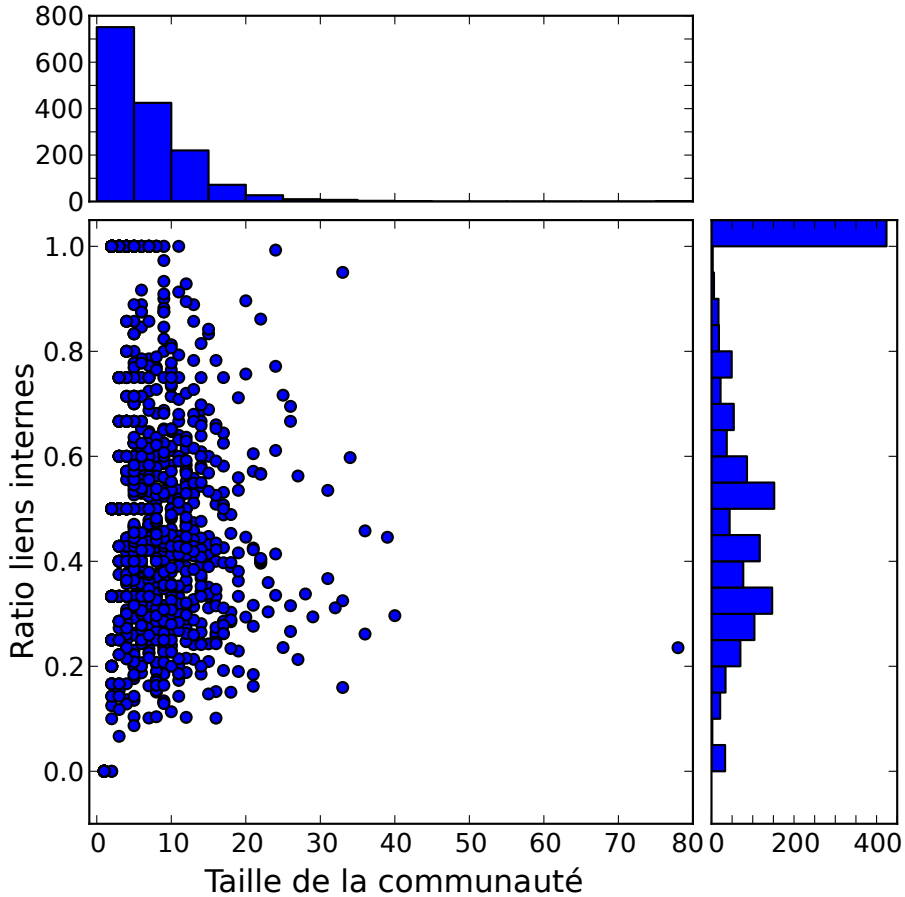

Rosvall, M. and Bergstrom, C. T. (2008).
Maps of random walks on complex networks reveal community structure.
Proceedings of the National Academy of Sciences, 105(4) :1118–1123.

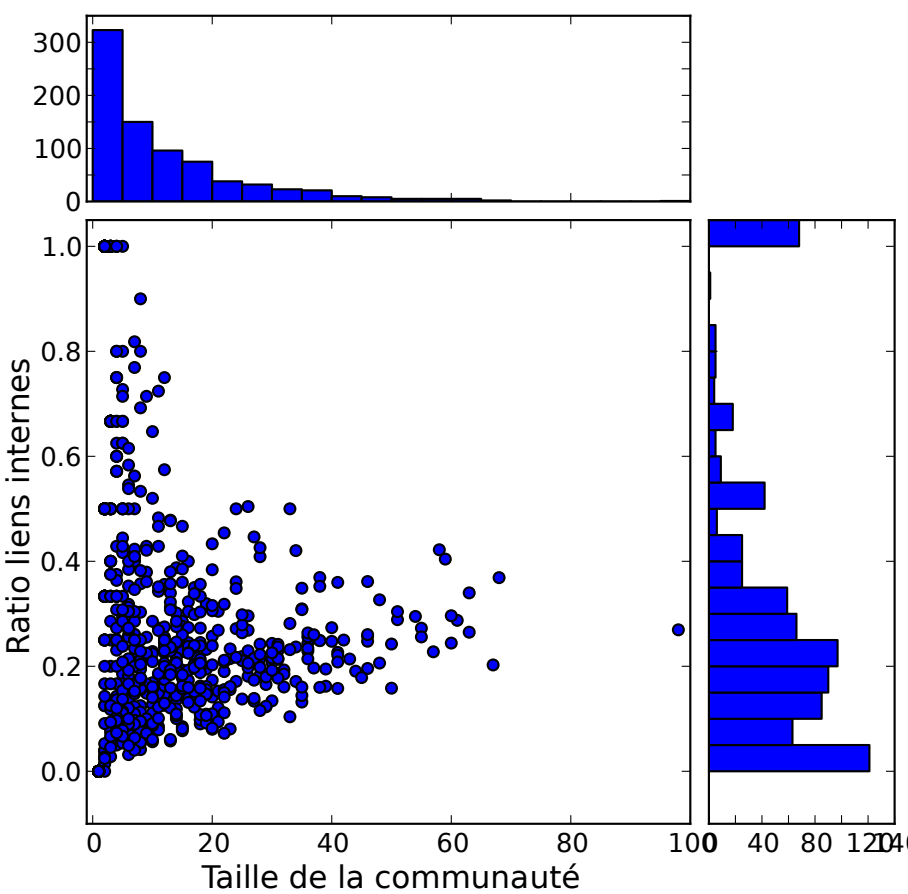

van Dongen, S. (2000).
Graph Clustering by Flow Simulation.
PhD thesis, University of Utrecht.

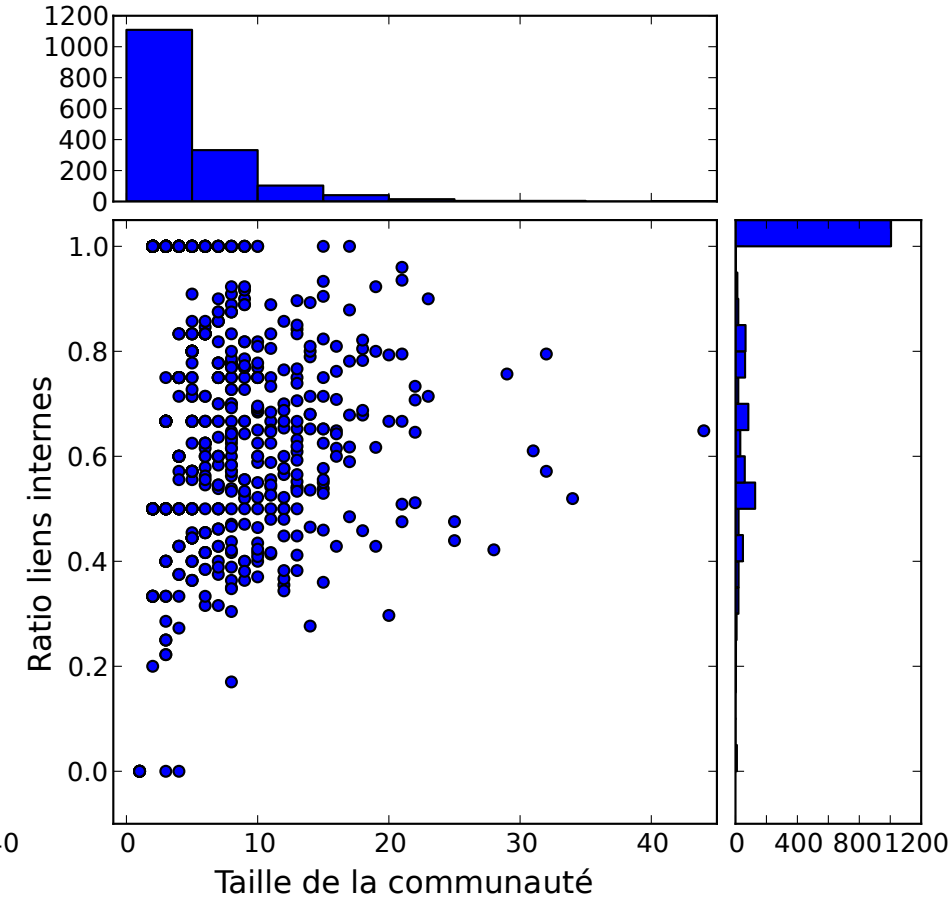

Zanghi, H., Ambroise, C., and Miele, V. (2008).
Fast online graph clustering via Erdos-Rényi mixture.
Pattern Recognition, 41(12) :3592–3599.

Annexe


Ratio liens internes
Taille de la communauté


Ratio liens internes
Taille de la communauté


Ratio liens internes
Taille de la communauté


Ratio liens internes
Taille de la communauté