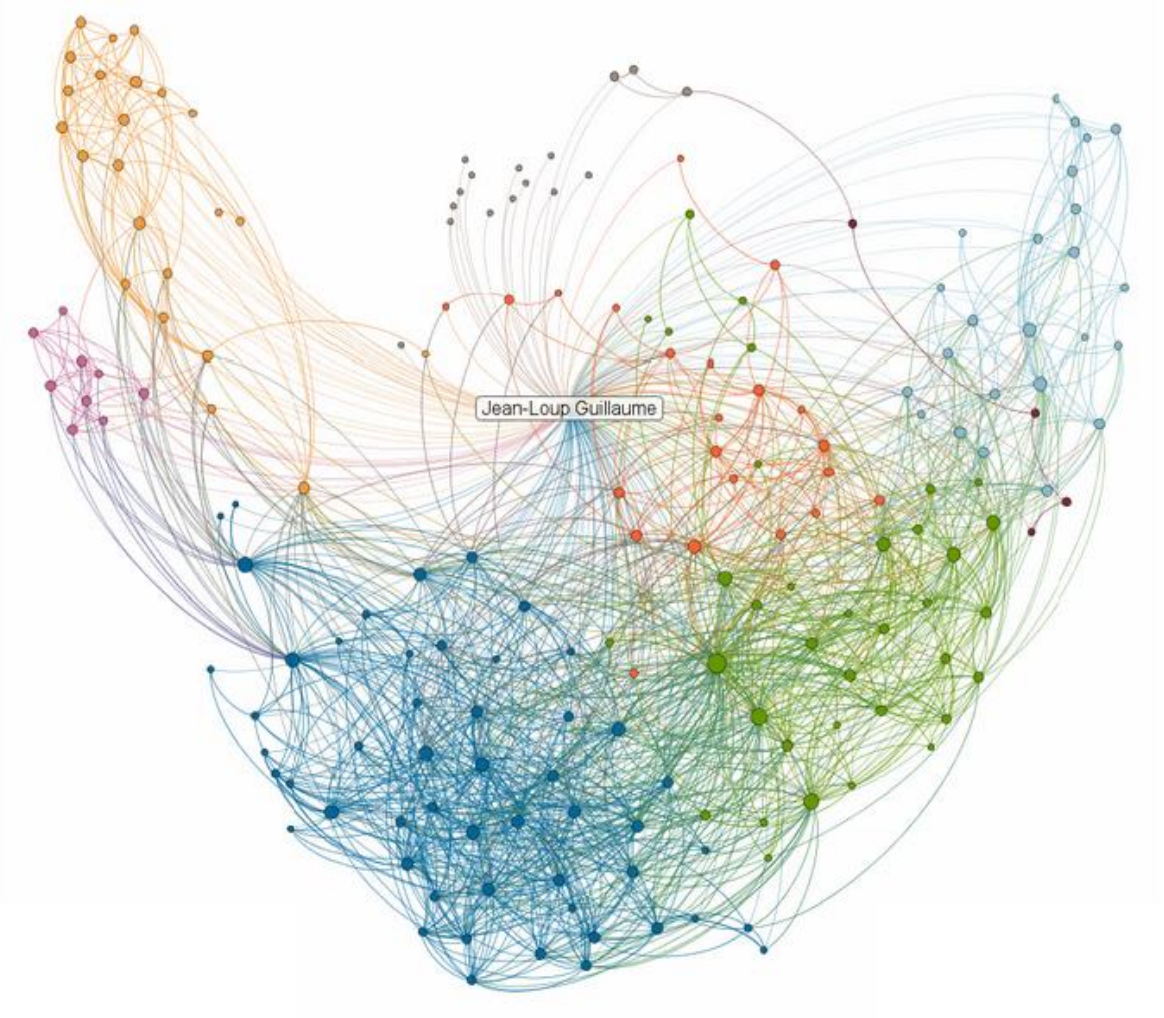
COMMUNAUTÉS TOPOLOGIQUES DANS LES GRAPHES DE TERRAIN
DU PARTITIONNEMENT AU RECOUVREMENT

Jean-Loup Guillaume

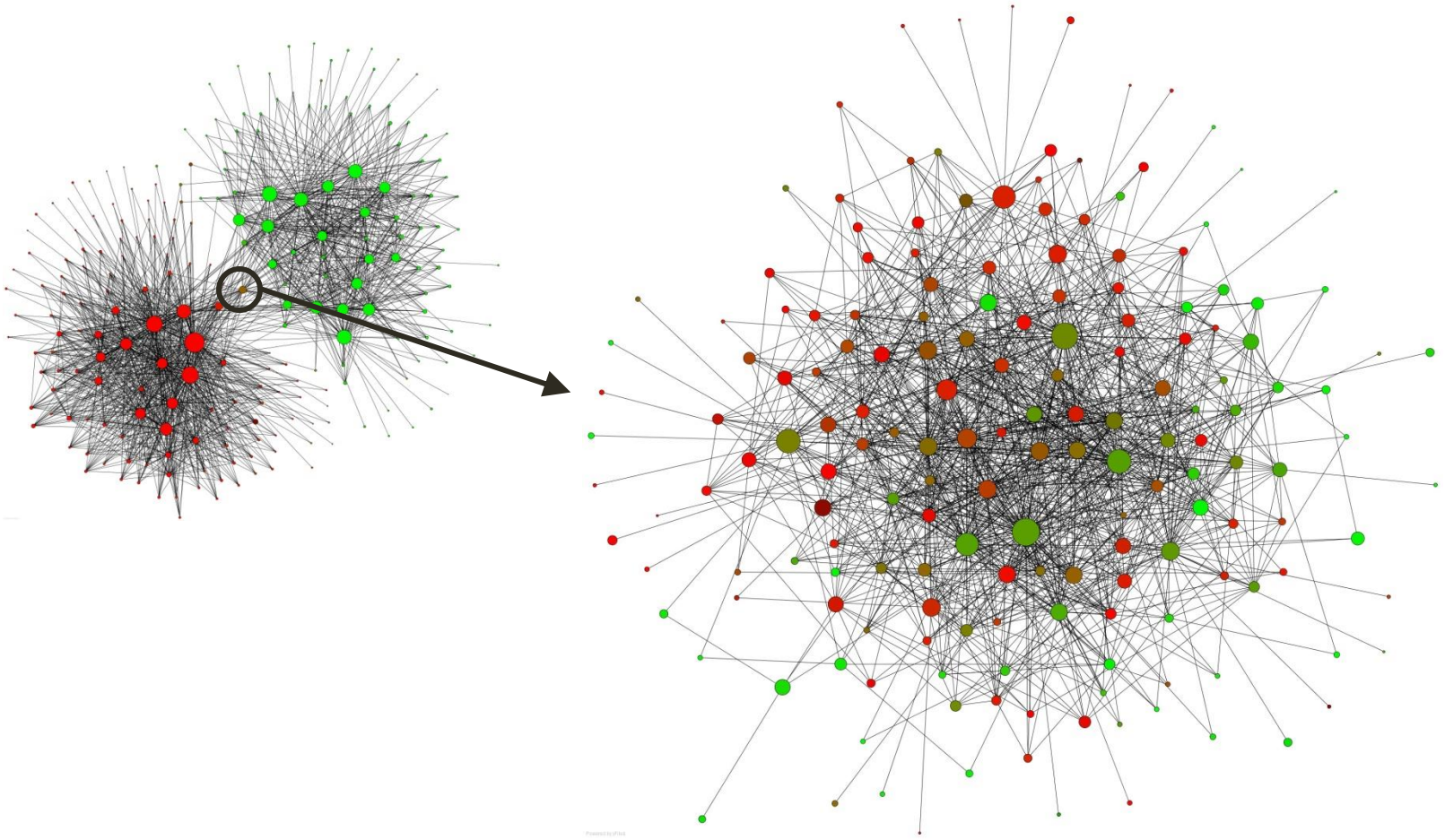
jean-loup.guillaume@univ-lr.fr

Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

RÉSEAU LINKEDIN/INMAPS - 04/2014



BELGIQUE : COMMUNICATIONS TÉLÉPHONIQUES



GRAPHES DE TERRAIN

Données relationnelles de terrain modélisées par des graphes :

- Informatique : web, Internet, email, P2P, …
- Sociologie : réseaux d'amitié, de collaboration, d'appels téléphoniques…
- Biologies : neurones, interactions entre protéines, éthologie…
- Linguistique, transports…

Nombreuses propriétés topologiques communes :

- Faible distance moyenne / effet petit-monde
- Degrés très hétérogènes / réseaux sans-échelle
- Motifs fréquents / triangles ou sous-graphes plus complexes
- Clustering / variation de densité et communautés

DÉTECTION DE COMMUNAUTÉS

Applications :

- Recommandation d'amis, groupes d'amis, classification d'inconnus
- Zones fonctionnelles dans le cerveau, prédiction de fonction de protéines
- Visualisation/navigation dans les graphes…

En informatique (au sens large) :

- Concevoir des algorithmes d'extraction automatique de communautés

Challenges :

- Nombre et taille des communautés inconnus
- Communautés recouvrantes et dynamiques
- Réseaux de grande taille

BESOIN D'ALGORITHMES SPÉCIFIQUES ?

Taille :

- Wikipedia = 4,5 millions de pages + dynamique + contributeurs
- Facebook = 1,4B utilisateurs actifs avec plus de 100 « amis » en moyenne
- Twitter = 302M utilisateurs actifs, 500M tweets/jour
- Web = dizaines de milliards de pages dans l'index Google

Il est non trivial de :

- Stocker le graphe en mémoire
- Faire des calculs sur le graphe

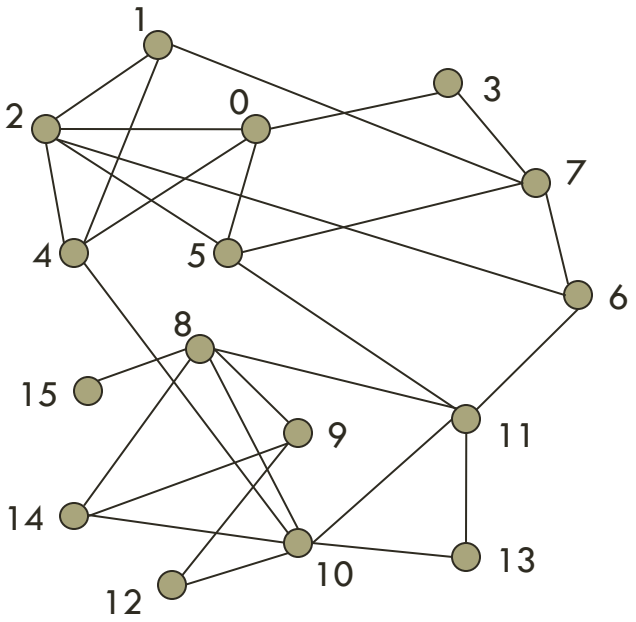
La notion de difficile change (n^2 commence à être difficile)

- Un calcul effectué pour chaque sommet/lien doit être local

UN EXEMPLE

Graphe à 16 sommets + découpages non recouvrants :

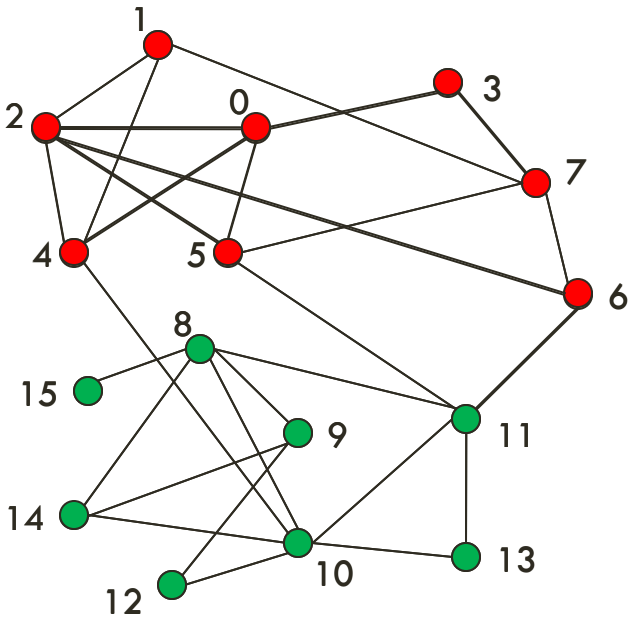
- Combien de découpages ?
- Combien de découpages connexes ?
- Combien de découpages optimaux ?



UN EXEMPLE

Graphe à 16 sommets + découpages non recouvrants :

- Combien de découpages ? ~ 10 milliards (nombre de Bell)
- Combien de découpages connexes ? 44484
- Combien de découpages optimaux ? 1 (pour la modularité)



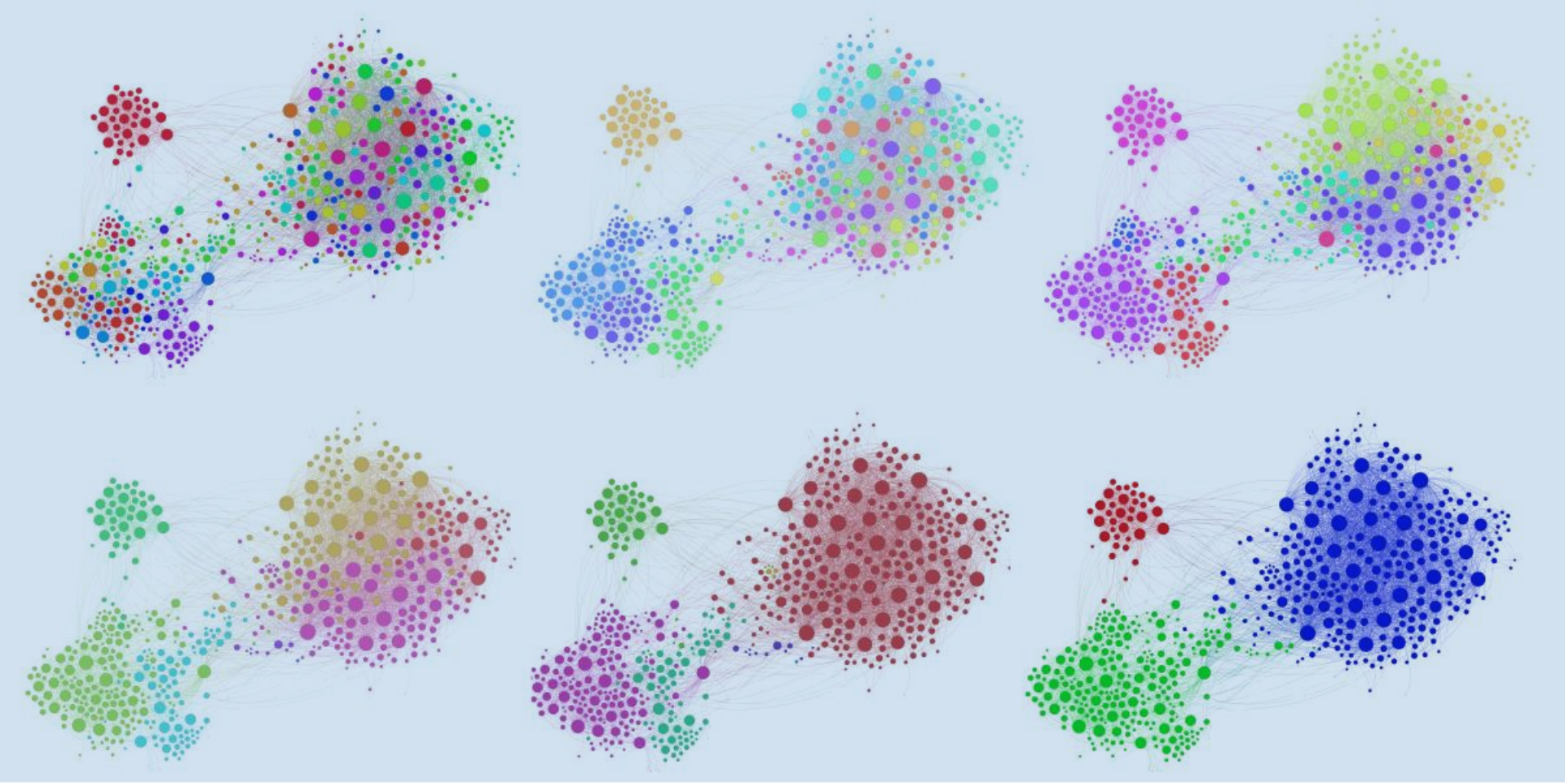
DANS LA SUITE DE L'EXPOSÉ

Quelques définitions topologiques

Algorithmes de partitionnement

Communautés recouvrantes

Quelques problèmes d'actualité



COMMUNAUTÉS
DÉFINITIONS TOPOLOGIQUES

Jean-Loup Guillaume

jean-loup.guillaume@univ-lr.fr

Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

QU'EST-CE QU'UNE COMMUNAUTÉ ?

Ensemble d'entités similaires, proches ou fortement connectées

- Amis, collègues, personnes avec des intérêts similaires
- Pages web avec un même contenu ou sur le même thème
- …

Lien avec la structure du réseau ?

- Similarité / proximité / densité

Questions supplémentaires :

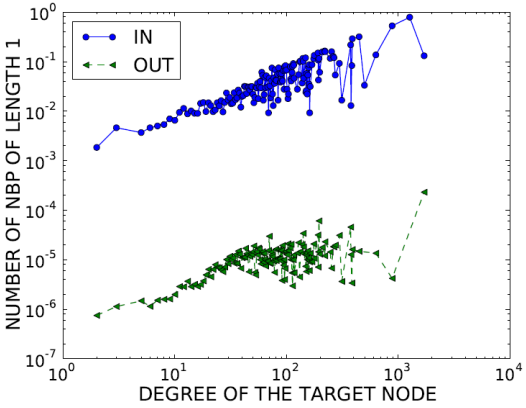
- Qualification des liens : types, pondérations, orientations
- Partitionnement (une communauté par individu) ou recouvrement
- Dynamique des réseaux

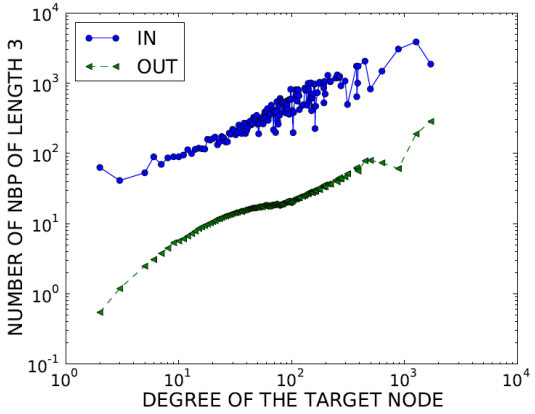
LES COMMUNAUTÉS TOPOLOGIQUES EXISTENT-ELLES ?

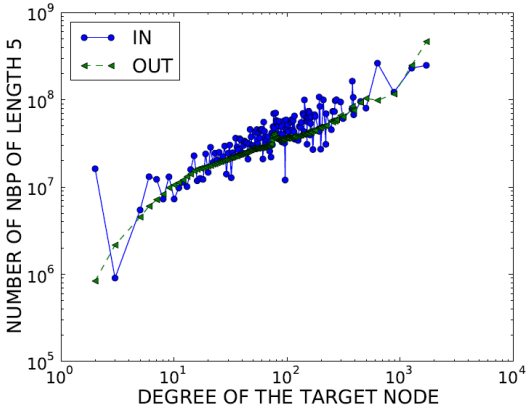
Vérité de terrain = catégories Wikipédia

- Chemins courts connectent dans les communautés
- Chemins longs connectent tout le monde

Lien clair entre proximité et catégories Wikipédia





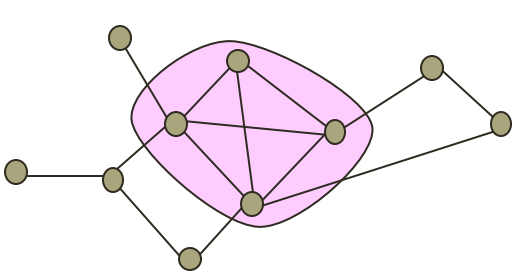


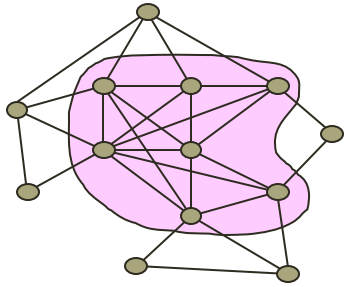
DÉFINITIONS CLASSIQUES

Composante (k-)connexe :

- Au moins un (k) chemin (disjoints) entre chaque paire de sommets

Clique : complètement connecté (ou clique recouvrante)

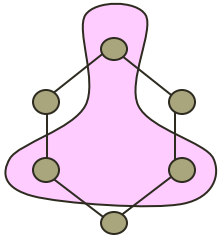


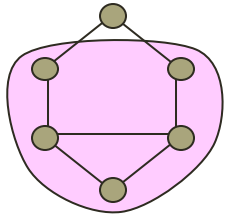


n-clique:

- Sous-graphe G' avec distance inférieure à n dans G

Peut être déconnecté



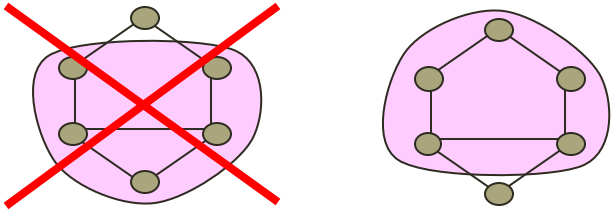


Diamètre peut être>n

DÉFINITIONS CLASSIQUES

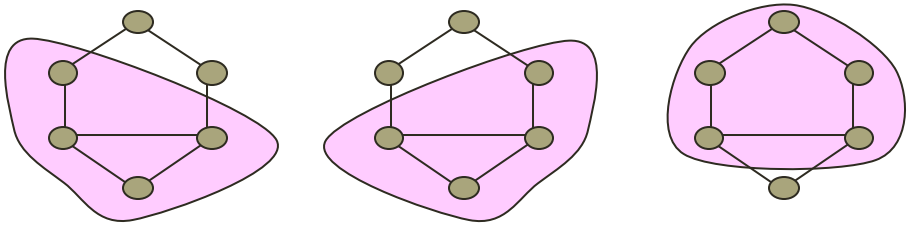
n-clan

- n-clique de diamètre n



n-club:

- Sous-graphe maximal de diamètre n



Et beaucoup d'autres…

LAQUELLE CHOISIR ?

Définitions trop contraintes ou pas assez :

- Composante connexe = il existe un chemin.
- Clique = sommets tous connectés. S'il manque un lien ?

Calculs souvent très couteux

- Souvent NP-complet : Cliques, k-plex…
- Parfois polynomial : LS, Lambda sets (n^4 ou moins)…
- Passage à l'échelle : web ~ milliards de sommets !

Nombre de communautés, tailles, …

Une réponse (à l'heure actuelle) : la modularité

LA MODULARITÉ

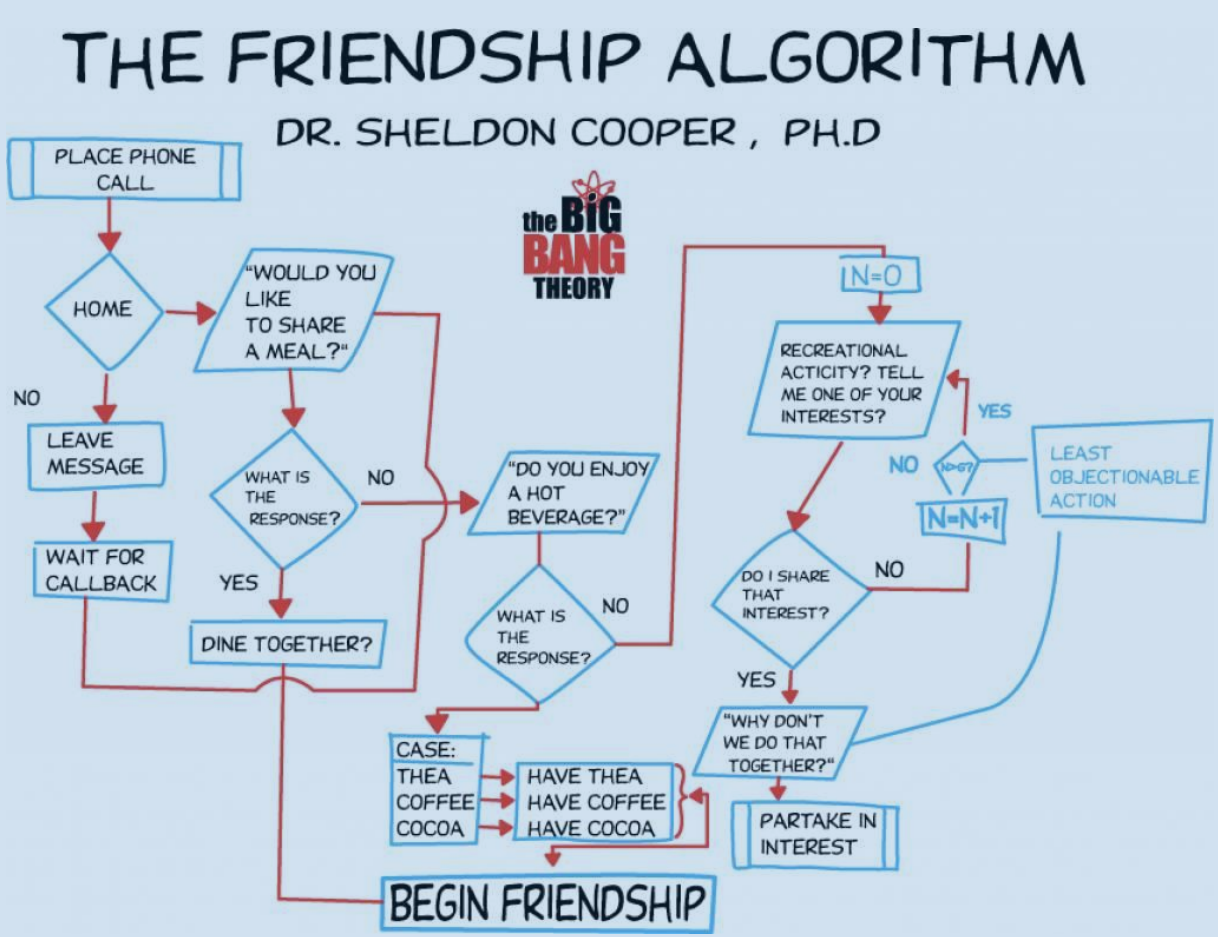
Principe de base :

- Communauté = sous-graphe dense.
- Pas forcément pertinent si tout le graphe est dense.

La modularité : communauté = sous-graphe plus dense que prévu

- Comparaison entre :
  - Le nombre de liens dans une communauté et
  - Le nombre de liens attendu
- Permet de s'abstraire du nombre de connexions

$$Q = \frac{1}{L}\sum_{s=1}^{m}\left(l_s - \frac{d_s^2}{4L}\right) = \sum_{s=1}^{m}\left[\frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2\right]$$



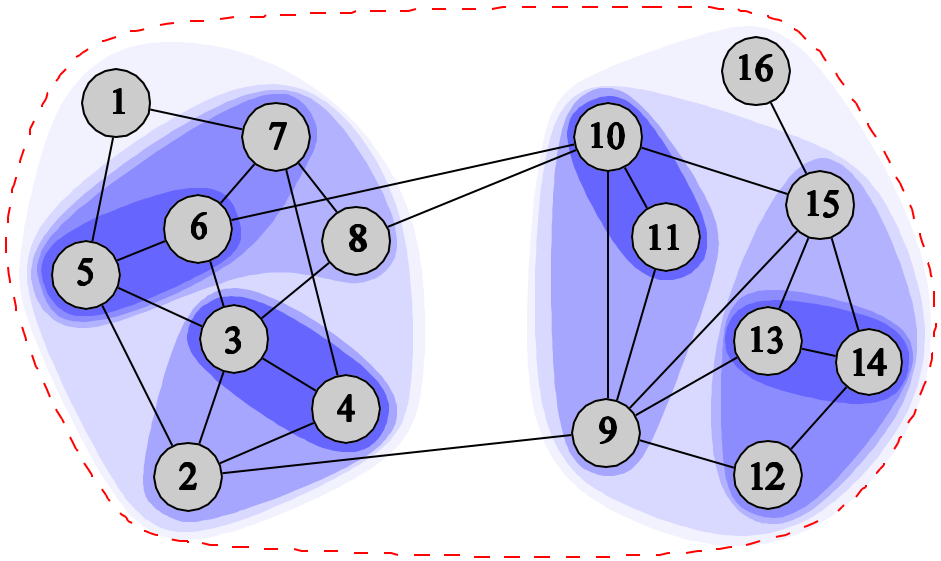
QUELQUES ALGORITHMES DE PARTITIONNEMENT

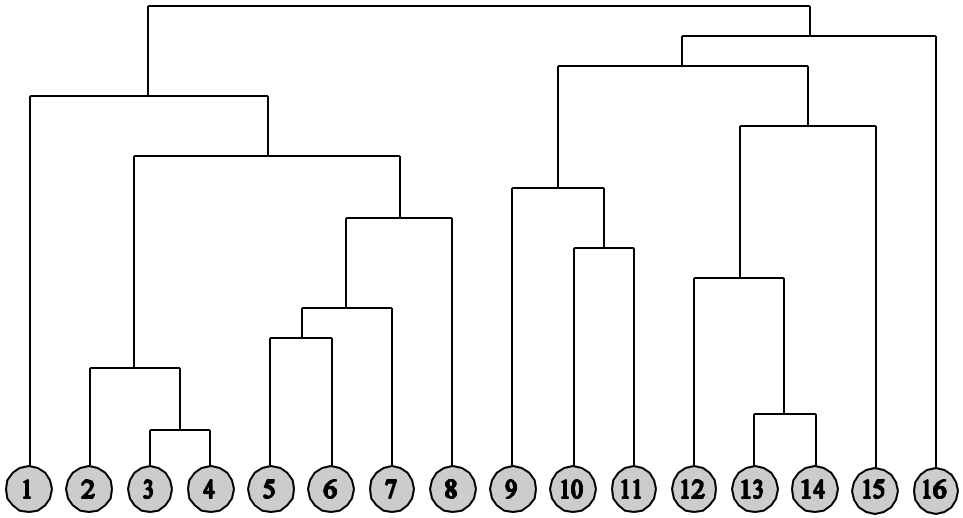
Jean-Loup Guillaume

jean-loup.guillaume@univ-lr.fr

Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

CLUSTERING HIÉRARCHIQUE – FORTE SIMILARITÉ





Adapté d'un slide de Pascal Pons

CLUSTERING HIÉRARCHIQUE (SUITE)

Algorithme générique :

- Chaque sommet est dans une communauté
- Calculer une distance entre chaque paire de communautés
- Fusionner les deux plus proches
- Revenir à l'étape 2

Uniquement besoin d'une distance entre sommets

Distance entre communautés = distance entre sommets +

- Min, max, moyenne, centre de gravité, …

QUELQUES DISTANCES

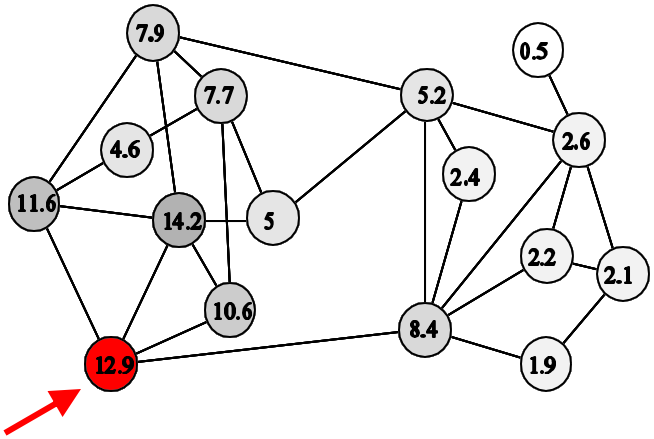
Distance = modularité

- À chaque étape effectuer la fusion qui maximise la modularité

Distance = perception du graphe / marches aléatoires

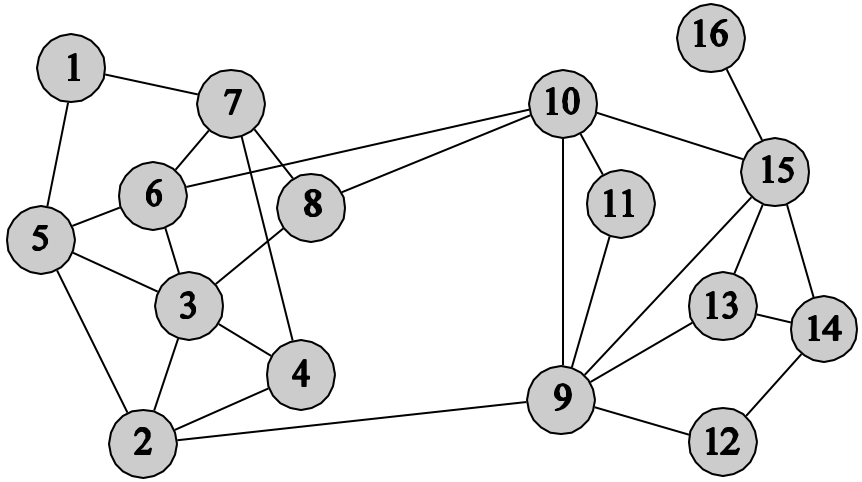
- Chemins courts : pas assez d'information
- Chemins longs : aucune information (proba~degré)

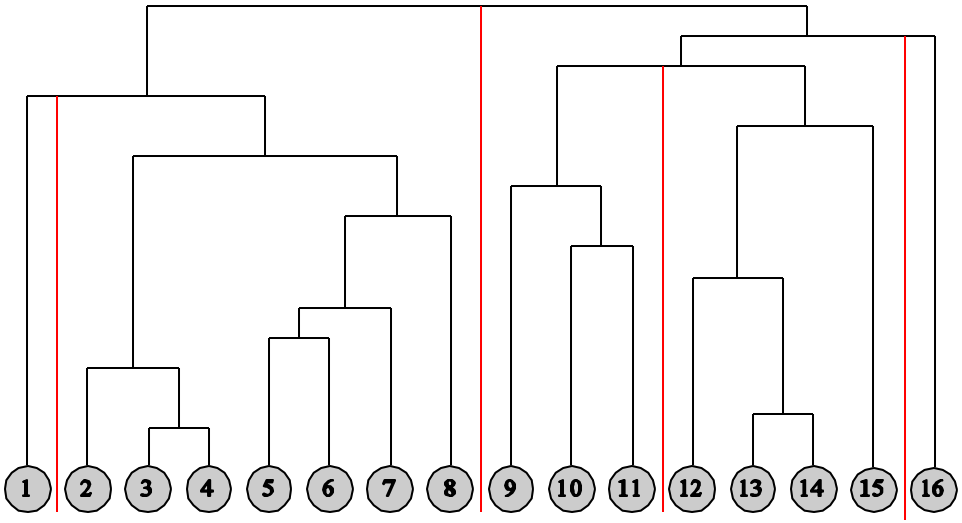
t = 3



Adapté d'un slide de Pascal Pons

APPROCHES DIVISIVES – LIENS FAIBLES





Adapté d'un slide de Pascal Pons

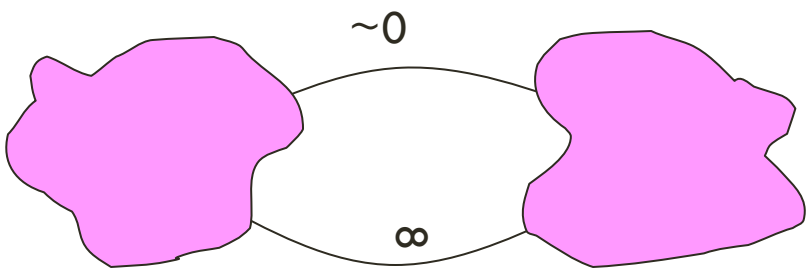
APPROCHES DIVISIVES (SUITE)

Algorithme générique :

- Calculer un score d'inter-communautarisme pour chaque lien
- Supprimer le lien ayant le score le plus élevé
- Revenir à l'étape 1

Uniquement besoin d'un score

Attention : il faut recalculer le score à chaque fois



QUELQUES SCORES DE CENTRALITÉ DE LIENS

Un lien entre deux communautés sera plus certainement utilisé si l'on cherche un chemin entre deux sommets quelconques

Centralité de plus courts chemins :

- Nombre de plus courts chemins utilisant un lien donné

Centralité de chemins aléatoires :

- Nombre de fois qu'un lien est utilisé dans un chemin aléatoire

Modification des distances (E=distance moyenne)

$$c_{\{i,j\}} = \frac{\Delta E_{\{i,j\}}}{E} = \frac{E(G) - E(G \setminus \{i, j\})}{E(G)}$$

ET BEAUCOUP D'AUTRES

Plus de 500 articles cités dans [Fortunato 2009]

- Méthode de Louvain
- Infomap
- Recuit simulé
- Algos génétiques
- Équations de Kirchhoff
- Wu and Huberman, Eur Phys B 38, 2004
- Modèle de Potts
- Reichardt and Bornholdt, Phys Rev Lett 93, 2004
- …

QUELLE(S) MÉTHODE(S) UTILISER ?

Très peu d'algorithmes peuvent traiter des millions de sommets

Author	Ref.	Label	Order
Girvan & Newman	(Girvan and Newman, 2002; Newman and Girvan, 2004)	GN	$O(nm^2)$
Clauset et al.	(Clauset *et al.*, 2004)	Clauset et al.	$O(n \log^2 n)$
Blondel et al.	(Blondel *et al.*, 2008)	Blondel et al.	$O(m)$
Guimerà et al.	(Guimerà and Amaral, 2005; Guimerà *et al.*, 2004)	Sim. Ann.	parameter dependent
Radicchi et al.	(Radicchi *et al.*, 2004)	Radicchi et al.	$O(m^4/n^2)$
Palla et al.	(Palla *et al.*, 2005)	Cfinder	$O(\exp(n))$
Van Dongen	(Dongen, 2000a)	MCL	$O(nk^2)$, $k < n$ parameter
Rosvall & Bergstrom	(Rosvall and Bergstrom, 2007)	Infomod	parameter dependent
Rosvall & Bergstrom	(Rosvall and Bergstrom, 2008)	Infomap	$O(m)$
Donetti & Muñoz	(Donetti and Muñoz, 2004, 2005)	DM	$O(n^3)$
Newman & Leicht	(Newman and Leicht, 2007)	EM	parameter dependent
Ronhovde & Nussinov	(Ronhovde and Nussinov, 2009)	RN	$O(m^\beta \log n)$, $\beta \sim 1.3$

Table extradite de "Community detection in graphs". S. Fortunato.

Meilleurs choix (d'après Santo Fortunato) :

- Infomap : excellents résultats, limité sur les très grands graphes
- Louvain : très bons résultats et gestion de très grands graphes

CONCLUSION

Deux approches génériques :

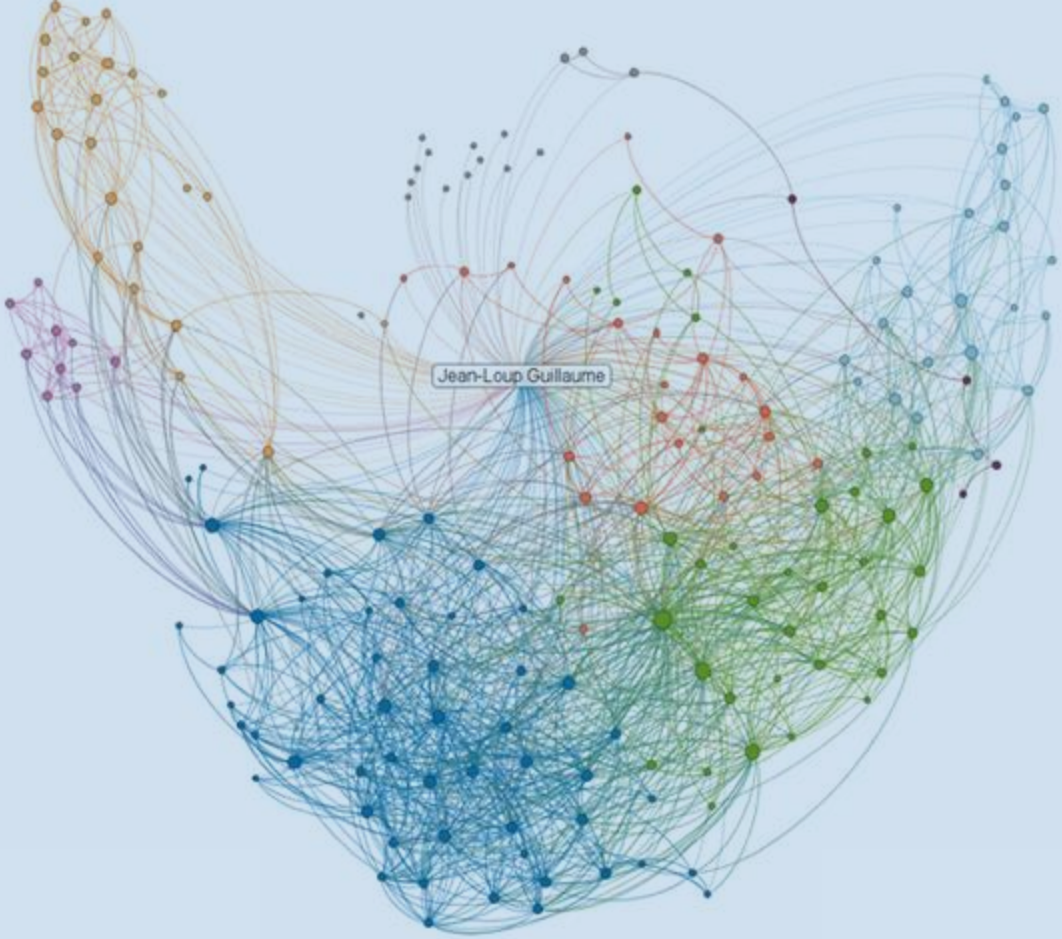
- Similarité entre sommets / liens faibles / et de nombreuses autres

Comment valider les résultats :

- Modularité
- Expertise / vérités de terrain (si disponibles)
- Graphes artificiels avec structure communautaire connue

Généralisation :

- Orientations et pondérations : trivial la plupart du temps
- Dynamique : complexe
- Appartenance à plusieurs communautés : complexe



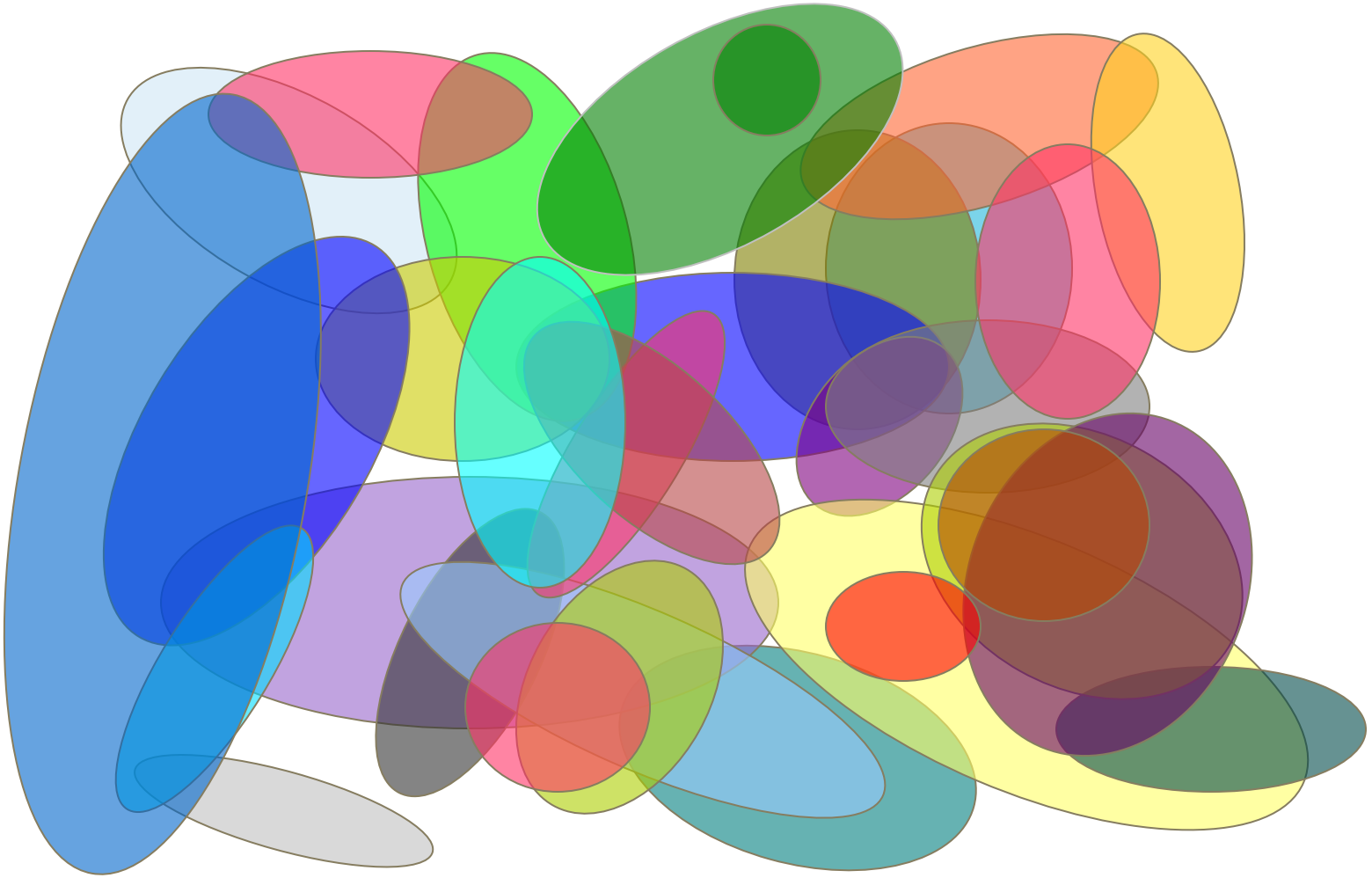
COMMUNAUTÉS ÉGOCENTRÉES

Jean-Loup Guillaume

jean-loup.guillaume@univ-lr.fr

Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

STRUCTURE RECOUVRANTE



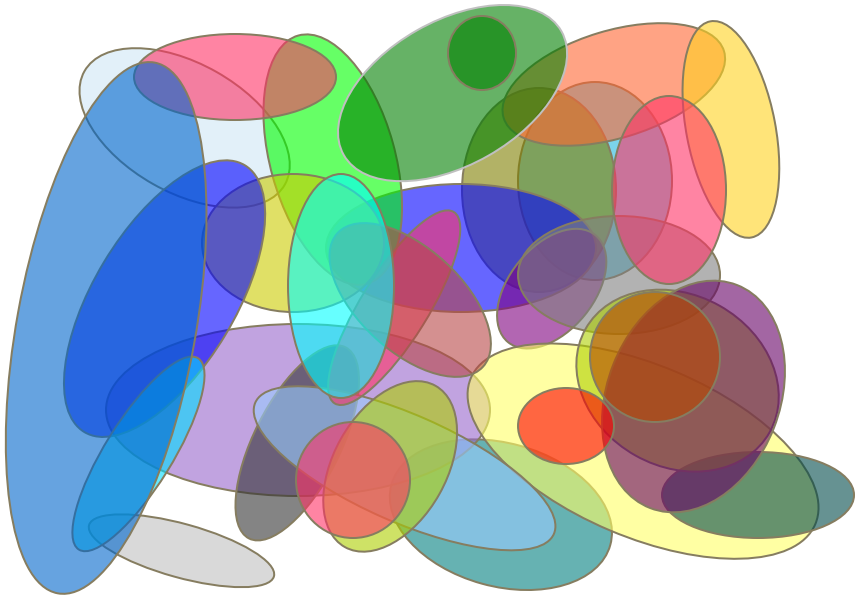
MÉTHODES RECOUVRANTES

Il existe des méthodes recouvrantes globales mais :

- Passent difficilement à l'échelle

Méthodes égocentrées = recherche la (les) communauté(s) d'un sommet

- Local => potentiellement plus efficace / parallélisable
- ! Pas uniquement le voisinage direct du sommet



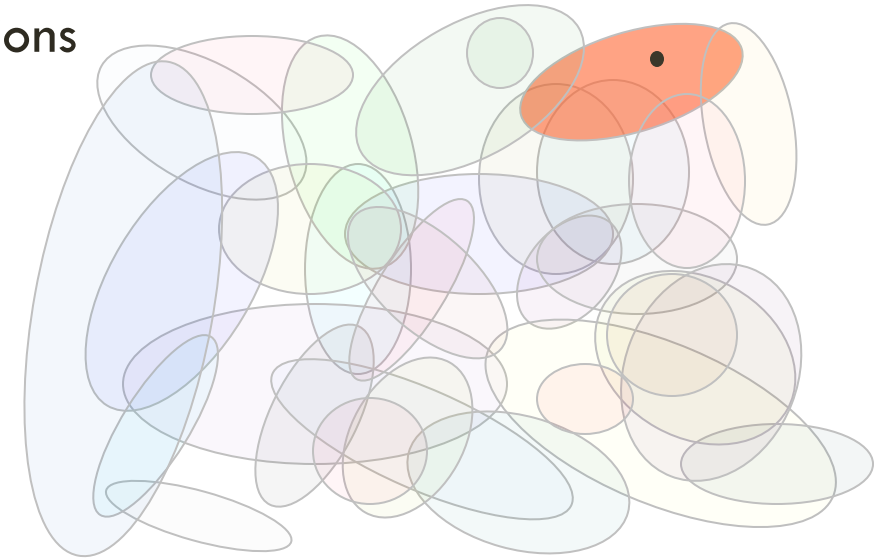
FONCTIONS DE QUALITÉ

Utilisation de fonctions exprimant la qualité d'une communauté :

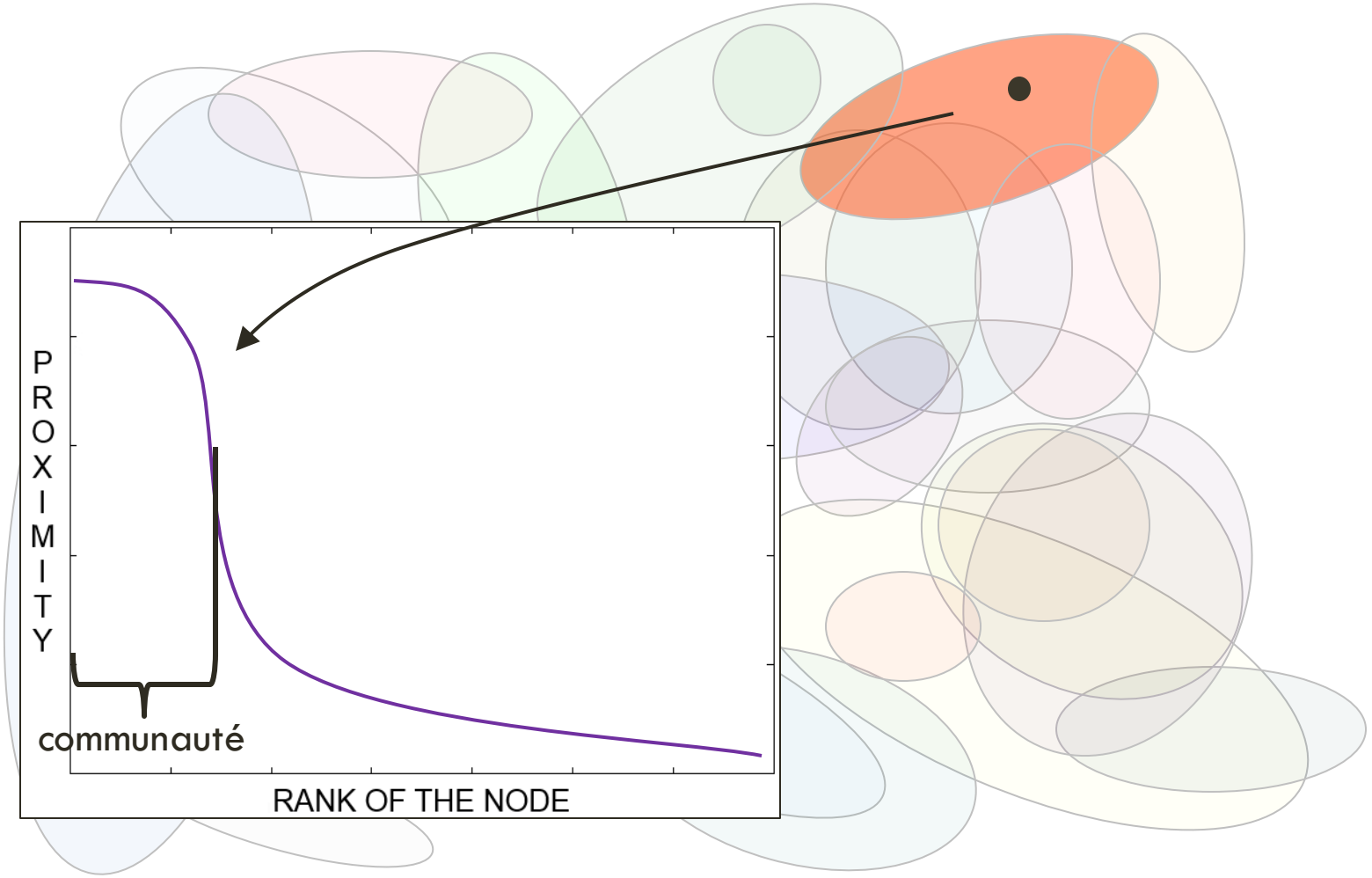
$$rd(G) = \frac{liens_{internes}}{liens_{internes} + liens_{sortants}}$$

$$C(G) = \frac{\Delta_{internes}(G)}{\binom{|G|}{3}} \times \frac{\Delta_{internes}(G)}{\Delta_{internes}(G) + \Delta_{sortants}(G)}$$

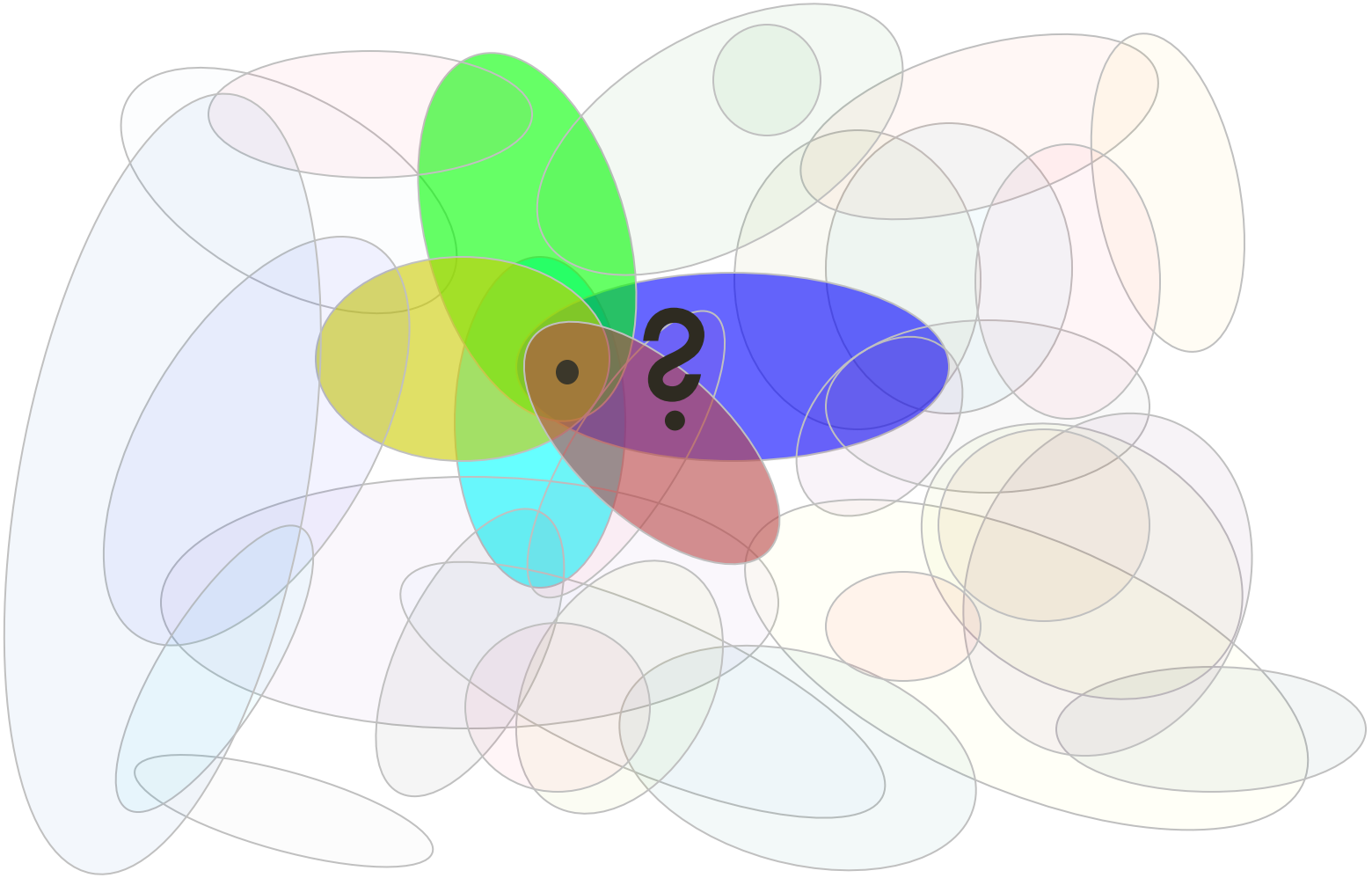
Puis on procède par ajouts/suppressions



MESURES DE PROXIMITÉ



APPARTENANCE À PLUSIEURS COMMUNAUTÉS



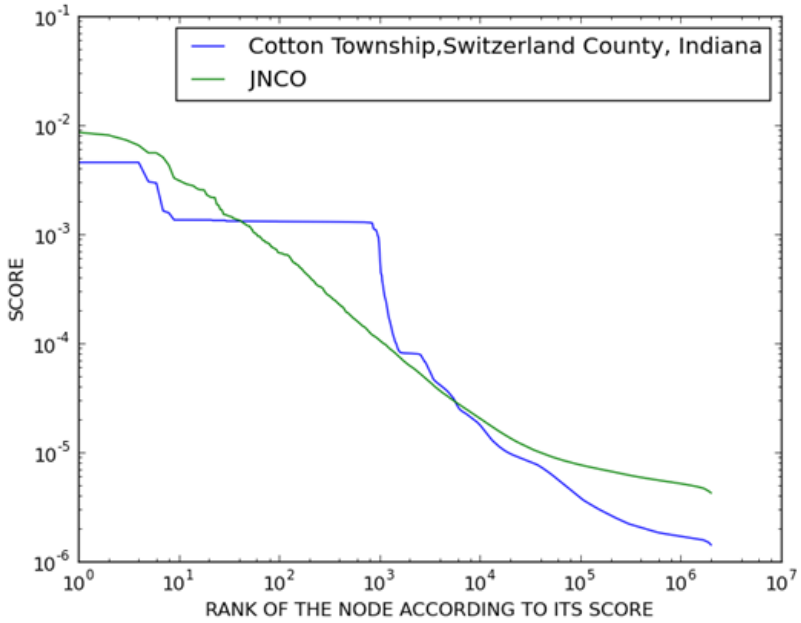
APPARTENANCE À PLUSIEURS COMMUNAUTÉS

Si elles représentent différents niveaux hiérarchiques

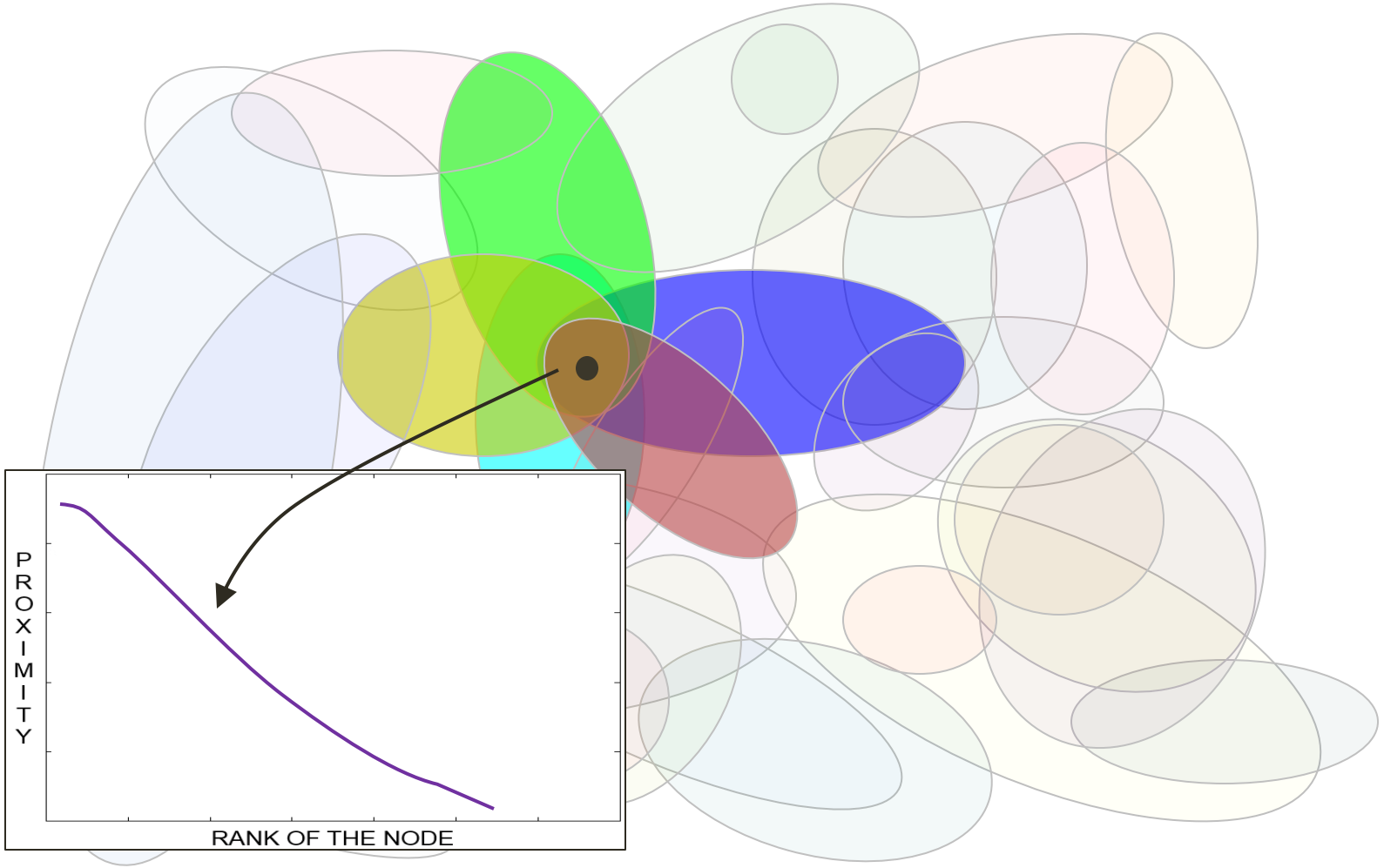
- Plusieurs plateaux

De différentes tailles / mal définies / qui se recouvrent

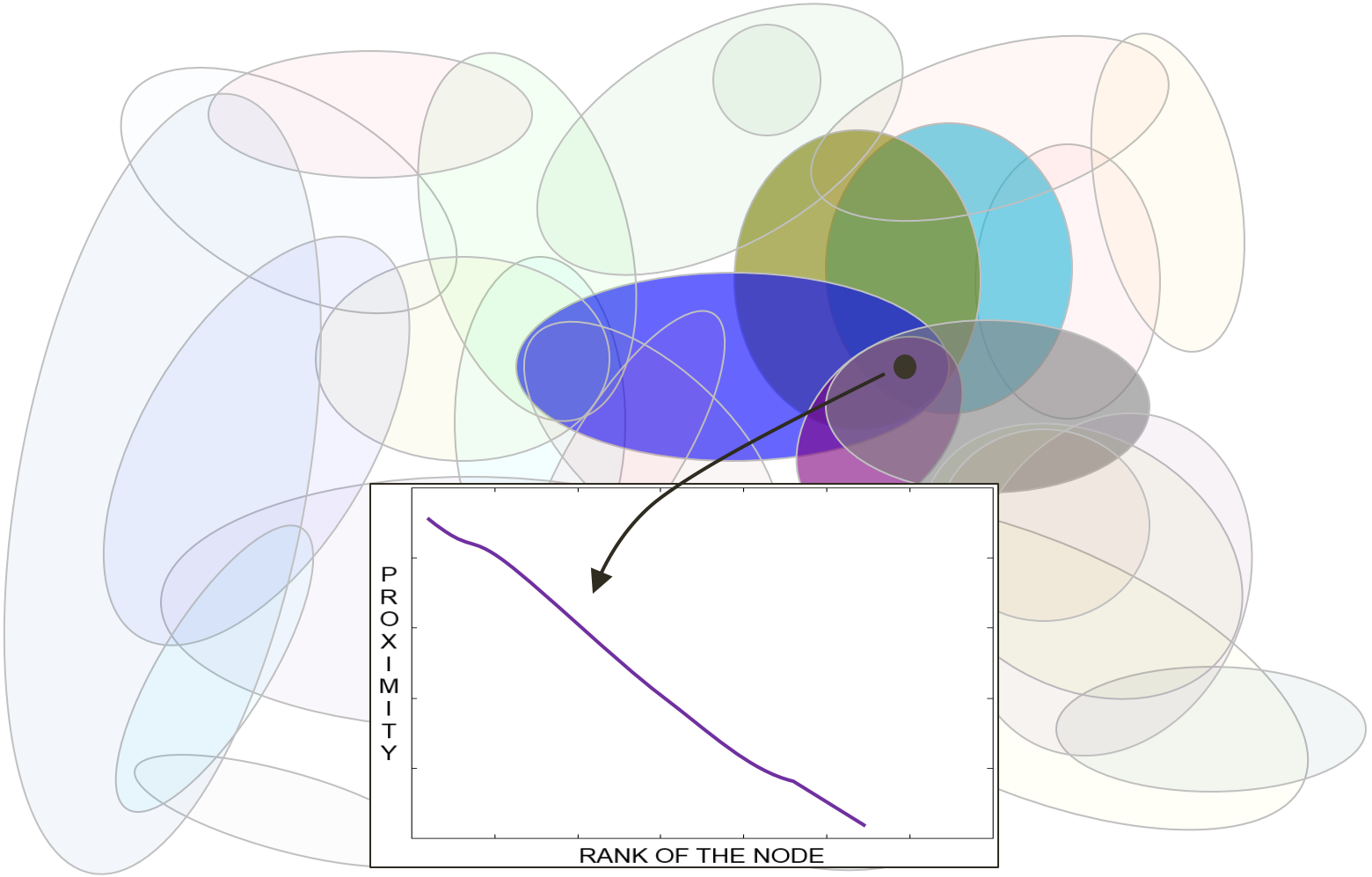
- Pas de plateau…



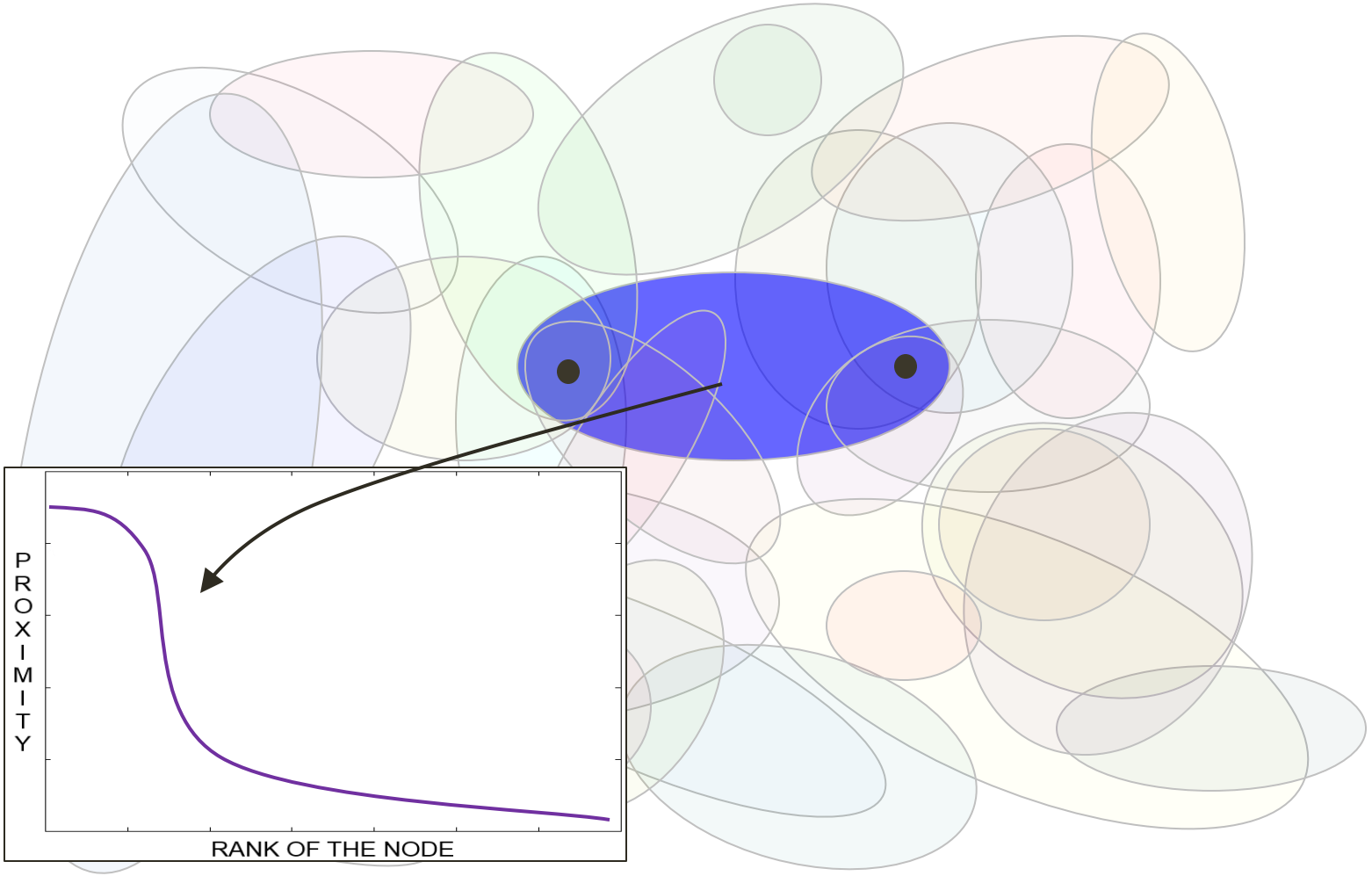
COMMUNAUTÉS ÉGOCENTRÉES



COMMUNAUTÉS ÉGOCENTRÉES

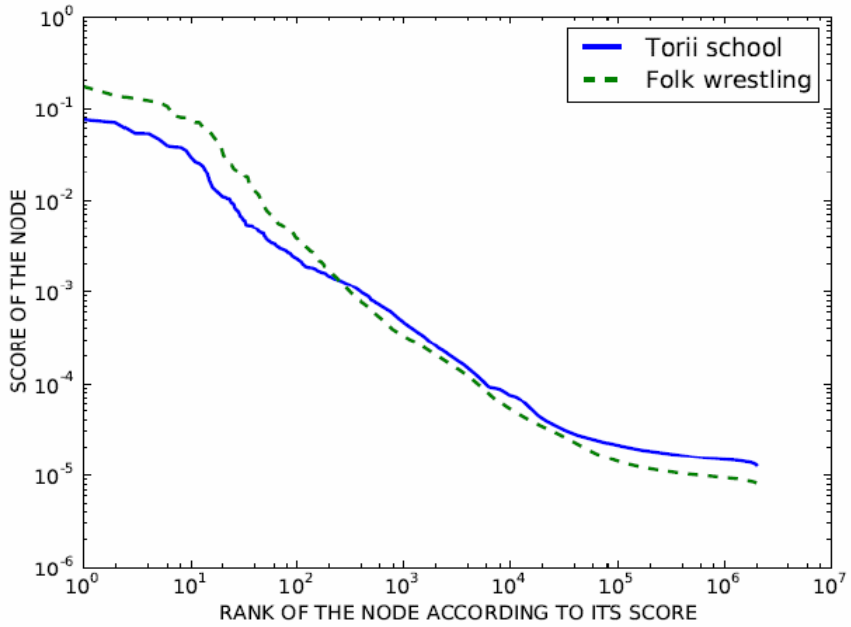


COMMUNAUTÉS BI-ÉGOCENTRÉES



COMMUNAUTÉS BI-ÉGOCENTRÉES

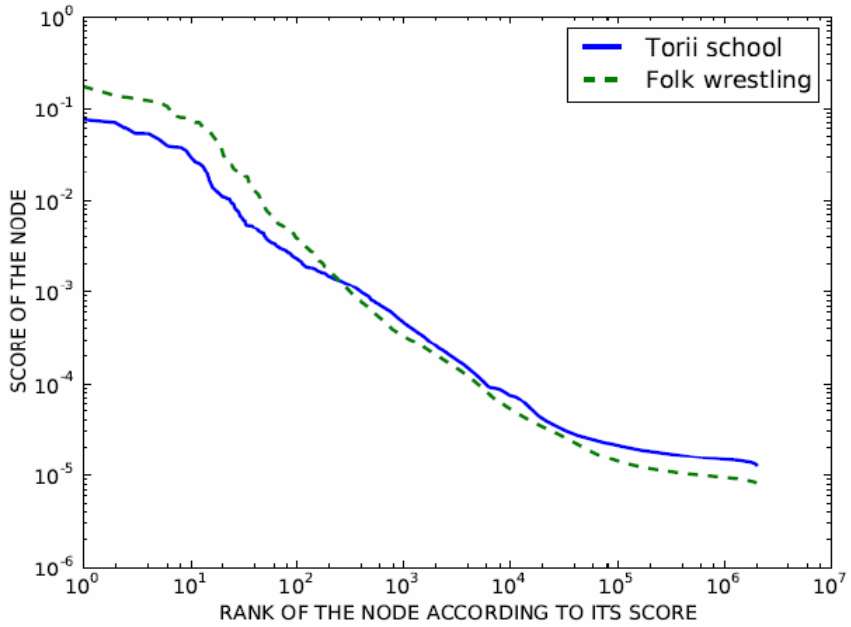
Torii school + Folk wrestling

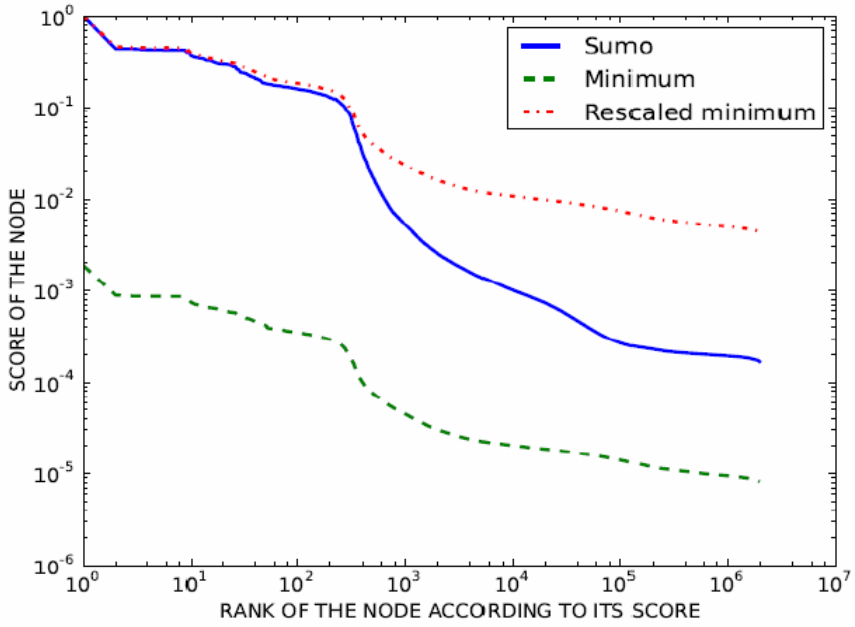


COMMUNAUTÉS BI-ÉGOCENTRÉES

Torii school + Folk wrestling = Sumo

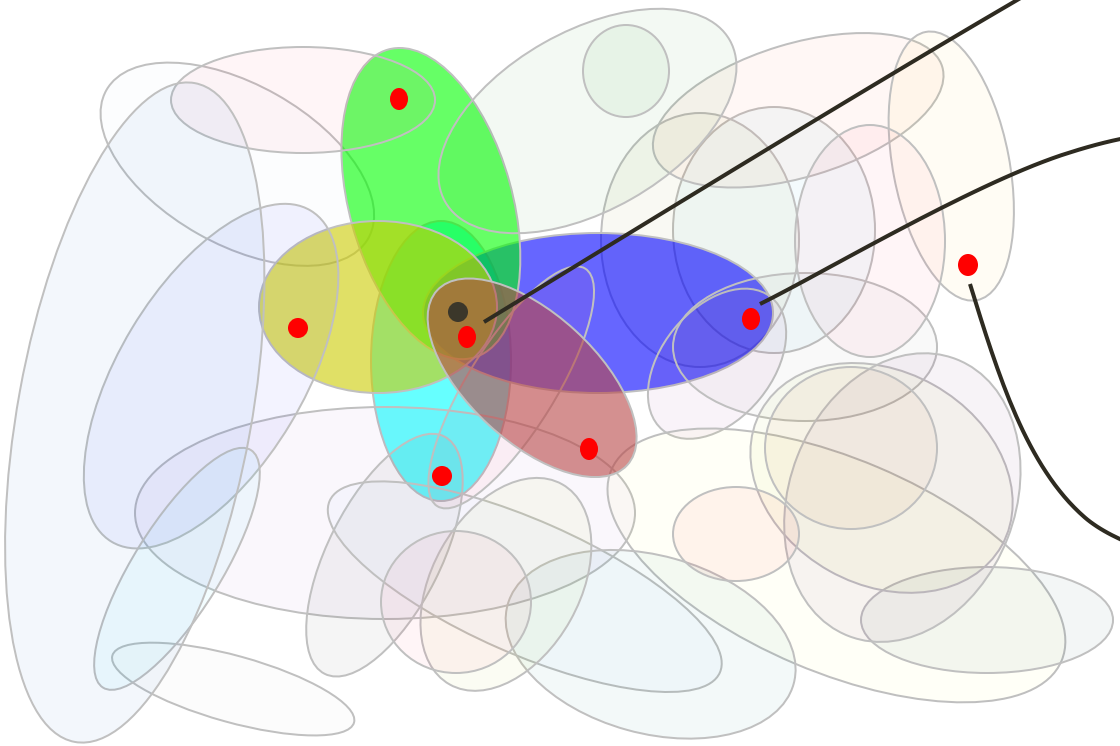
(parmi les 350 premiers sommets de sumo, 337 sont dans le minimum)

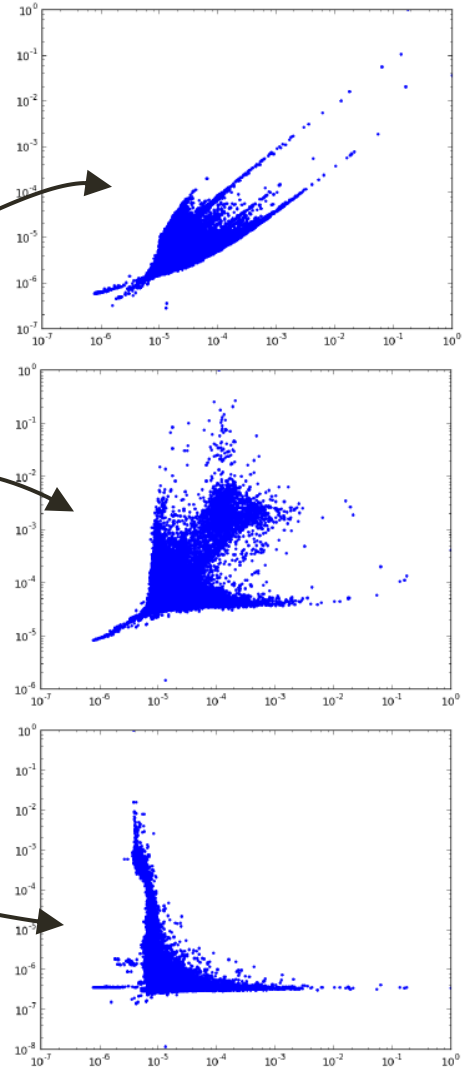




TROUVER TOUTES LES COMMUNAUTÉS

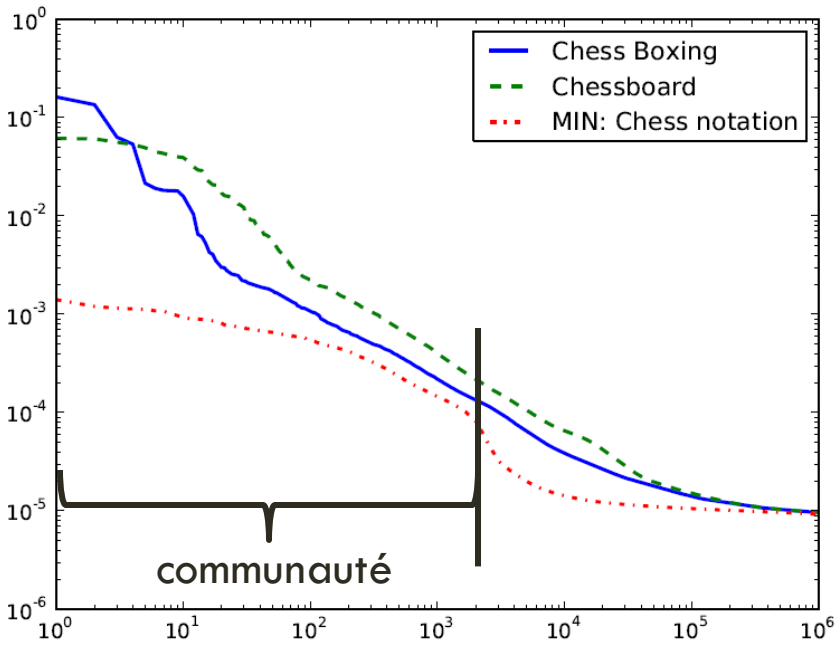
1. Sélectionner des candidats
   - Combien ?
   - Lesquels ?

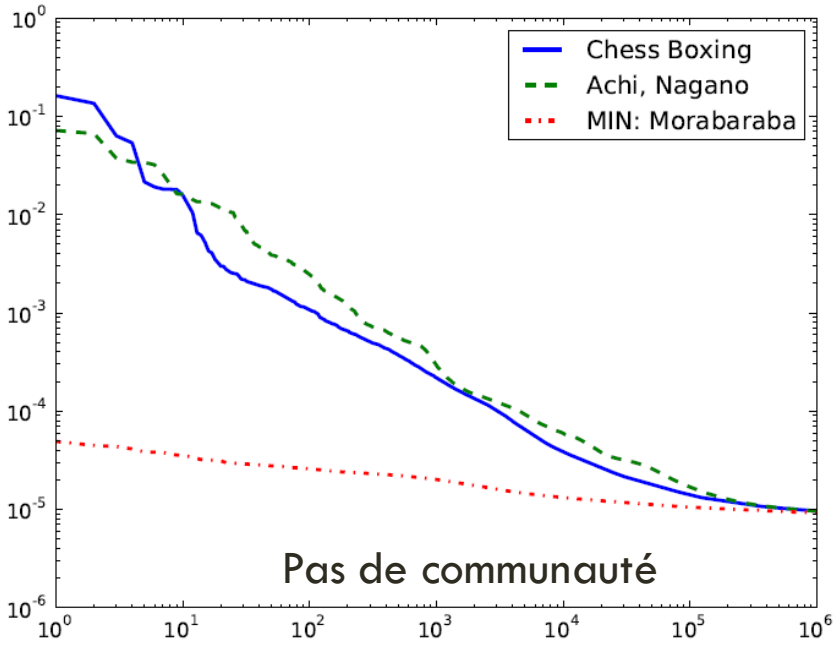




TROUVER TOUTES LES COMMUNAUTÉS

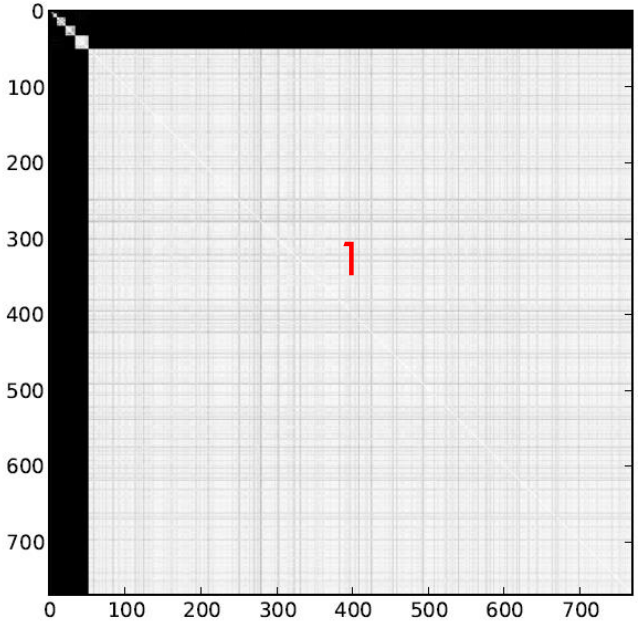
2. Calculer des communautés bi-égocentrées
   - Minimum de deux scores
   - Garder les sommets avant la forte décroissance (s'il y en a une)

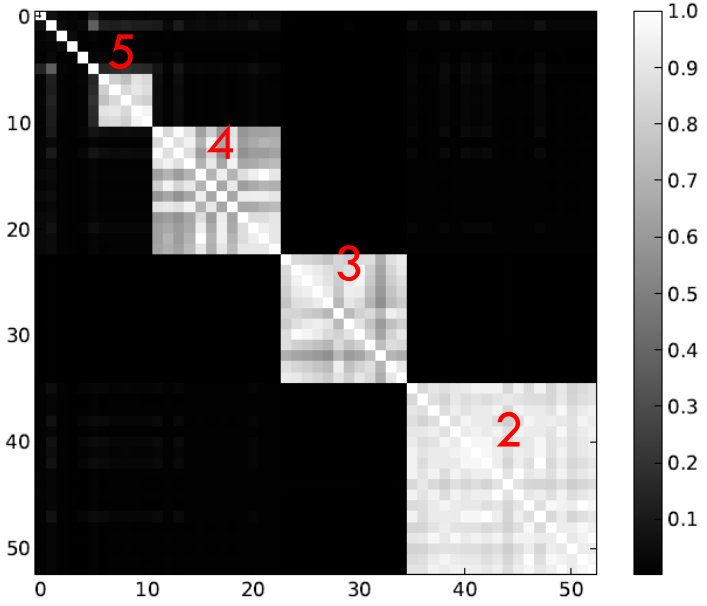




TROUVER TOUTES LES COMMUNAUTÉS

3. Nettoyer les communautés trouvées
   - Certaines sont obtenues par plusieurs candidats : fusion
   - Certaines ne sont trouvées qu'une fois : suppression





CONCLUSION / PERSPECTIVES

Méthode pour trouver des communautés (multi-)egocentrées

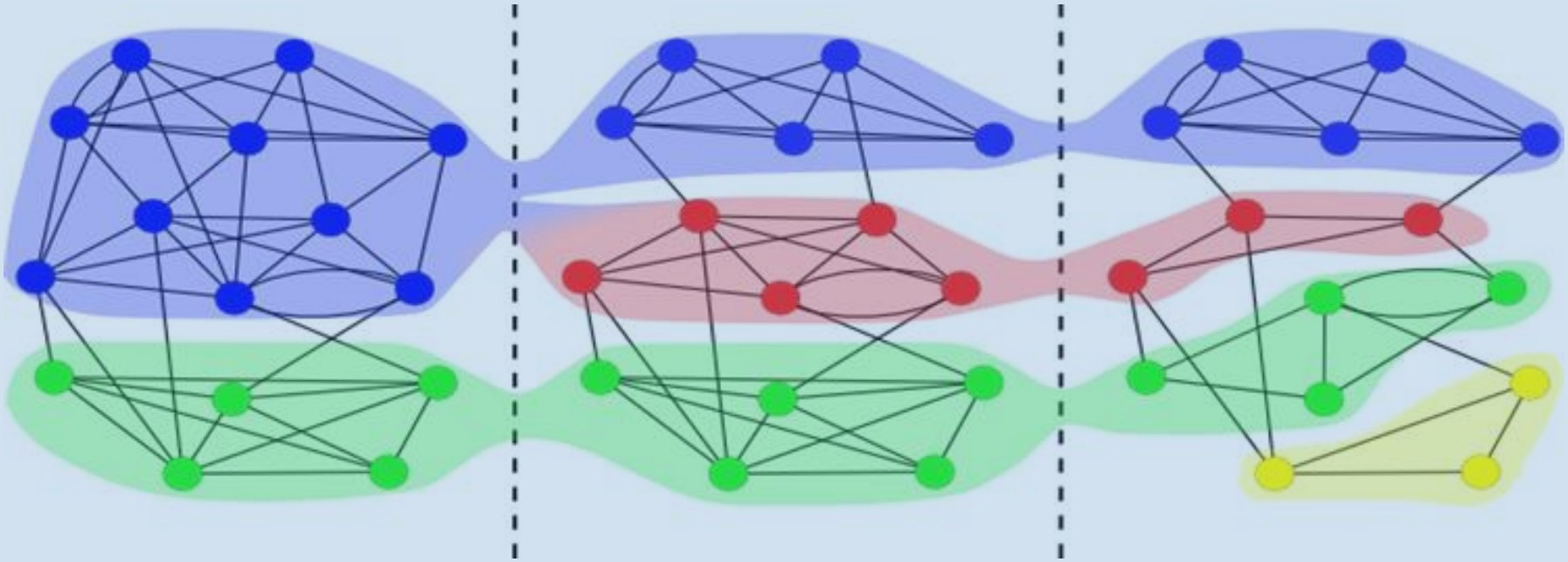
- Rapide à calculer
- Dépend d'une mesure de proximité (à choisir/concevoir)
- Peu efficace pour les pages très connectées
- Généralisable pour trouver toutes les communautés du graphe

Détection uniquement de la plus forte pente

- Autres pentes pertinents ?
- Que dire des sommets sur la pente (frontière des communautés) ?

Restriction à des communautés bi-égocentrées

- Pertinence d'utiliser plus de 2 sommets ?
- Temps de calcul => sélection intelligente des candidats



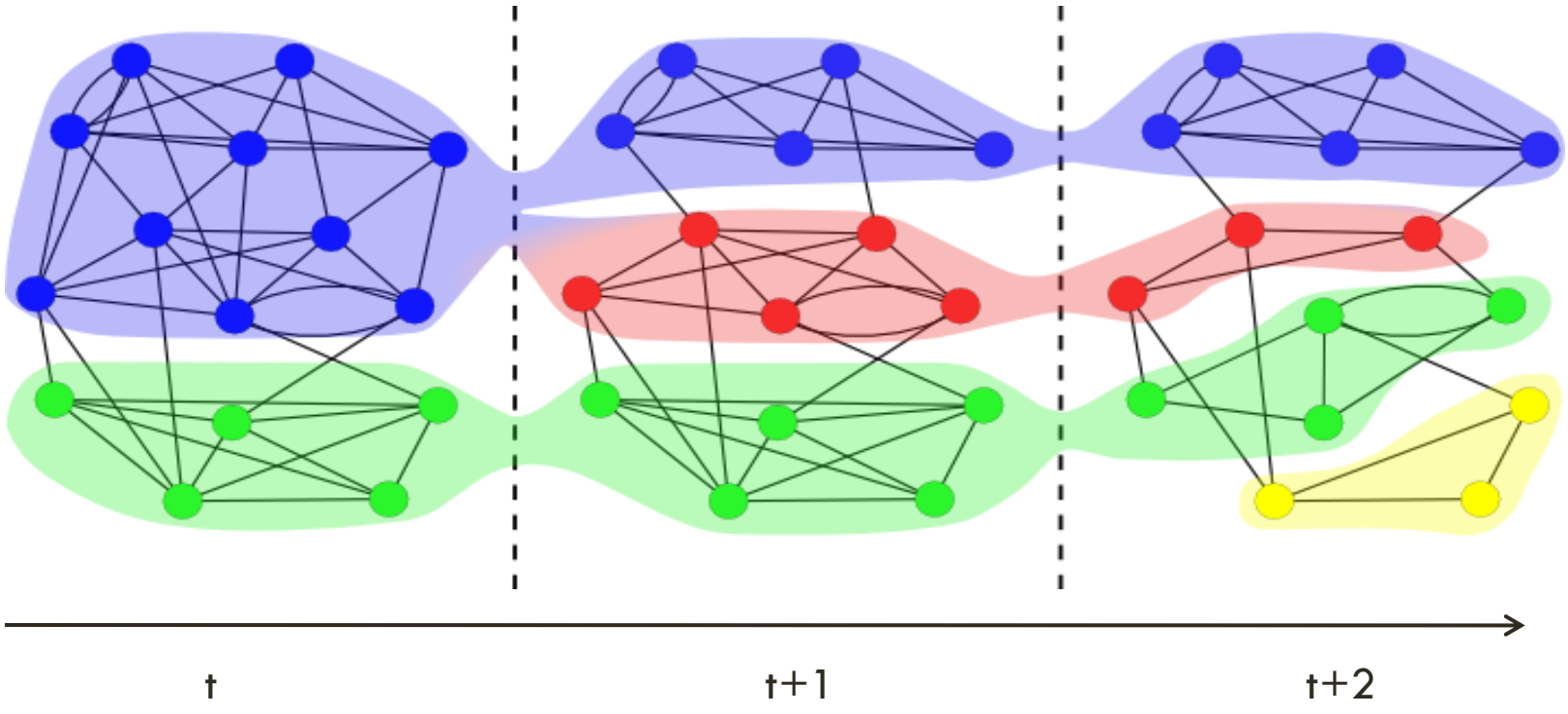
DEUX PROBLÈMES ACTUELS

Jean-Loup Guillaume
jean-loup.guillaume@univ-lr.fr

Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/

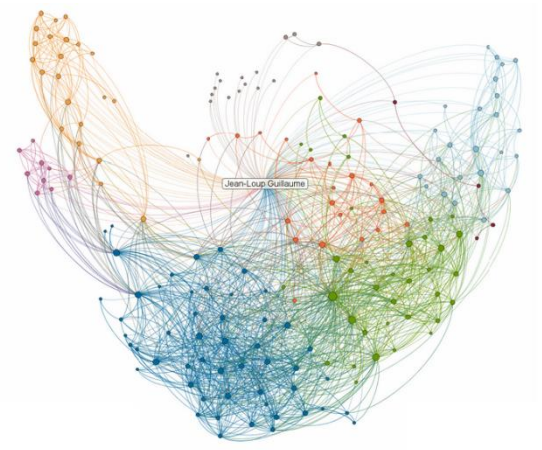
COMMUNAUTÉS DYNAMIQUES

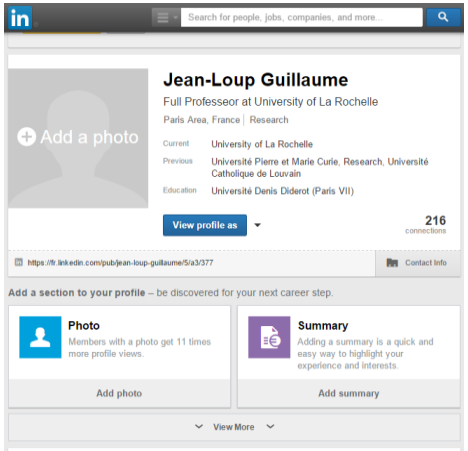
Comment calculer et suivre des communautés dans le temps ?

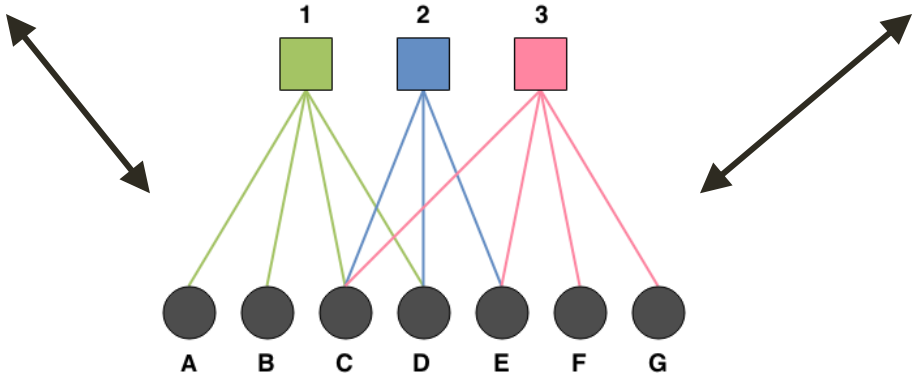


RÉSEAUX HYBRIDES / MULTIPLEXES / MULTIPARTIS

Comment tirer parti de plus que les connexions simples







COMMUNAUTÉS TOPOLOGIQUES DANS LES GRAPHES DE TERRAIN
DU PARTITIONNEMENT AU RECOUVREMENT

Jean-Loup Guillaume

jean-loup.guillaume@univ-lr.fr

Laboratoire Informatique Image Interaction (L3I)
Université de La Rochelle - Pôle Sciences et Technologie - Avenue Michel Crépeau - 17042 LA ROCHELLE CEDEX 1 France
Tél : +33 (0)5 46 45 82 62 – Fax : 05.46.45.82.42 – Site internet : http://l3i.univ-larochelle.fr/