Mesures de centralité

NŒUDS

- L'importance des nœuds se mesure avec des **centralités**.

- Attention: pas forcément être "au centre" du graphe

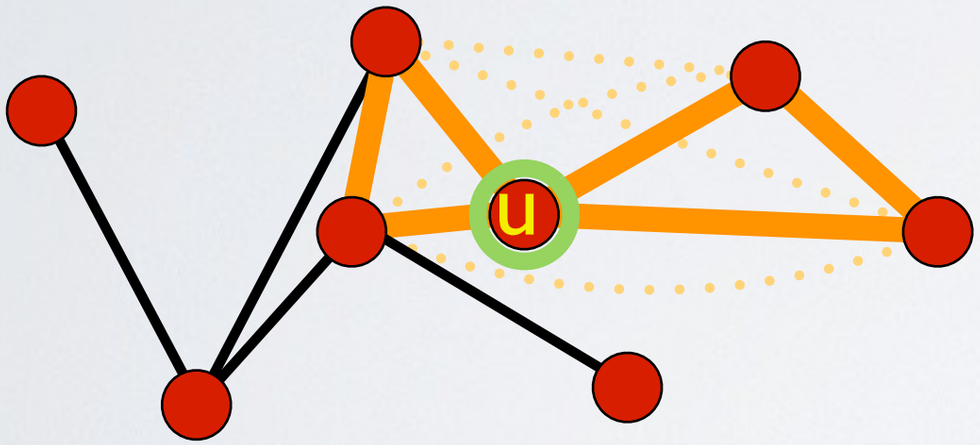
- Usage:
  - La plupart ont une interprétation claire
  - Certaines peuvent être utilisées par exemple pour du *Machine learning* (Prédire la réussite ou l'échec d'un politicien en fonction de sa position dans le réseau…)

DEGRÉ

- **Degré**: Combien de voisins
- Souvent suffisant pour trouver les nœuds importants
  - Les personnages principaux d'une série sont ceux qui parlent le plus
  - Les aéroports les plus importants ont le plus de connexions
  - …
- Mais pas toujours
  - Les utilisateurs de Facebook avec le plus de contacts sont souvent des spammeurs
  - Les pages webs/Wikipedia avec le plus de liens sortants sont souvent de bêtes listes de pages. Les liens entrants sont facile à truquer via d'autres sites.
  - …

CLUSTERING COEFFICIENT

C_u - **Node clustering coefficient:** density of the subgraph induced by the neighborhood of u, $C_u = d(H(N_u)$. Also interpreted as the fraction of all possible triangles in N_u that exist, $\frac{\delta_u}{\delta_u^{\max}}$



Edges: 2

Max edges: 4*3/2=6

C_u=2/6=1/3

Triangles=2

Possible triangles= $\binom{4}{2}$ =6

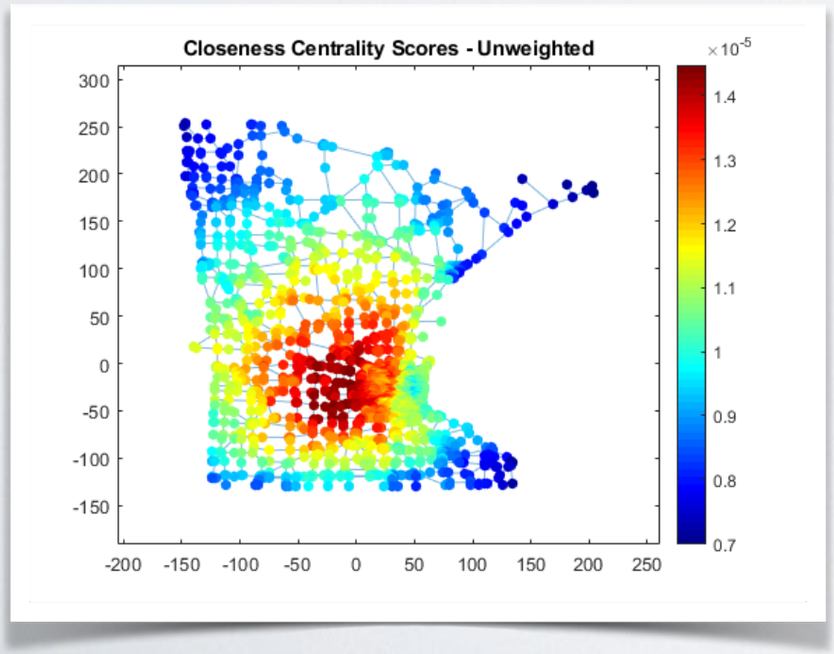
C_u=2/6=1/3

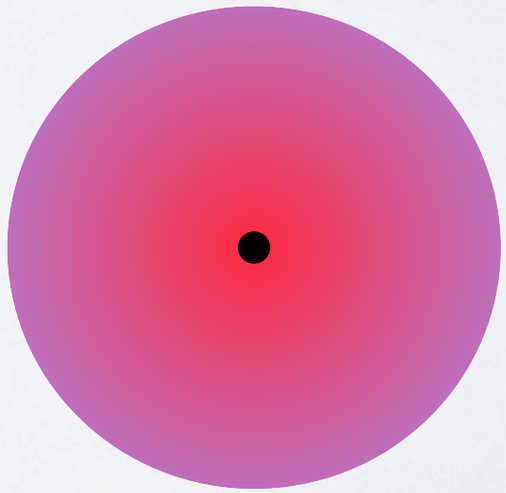
FARNESS, CLOSENESS

HARMONIC CENTRALITY

FARNESS, CLOSENESS

- Est-ce que le nœud est proche des autres, en moyenne.
- Parallèle avec le barycentre d'une figure:
  - Le barycentre d'une figure (centre d'une cercle) est le point le plus proche en moyenne de tous les autres points de la figure.



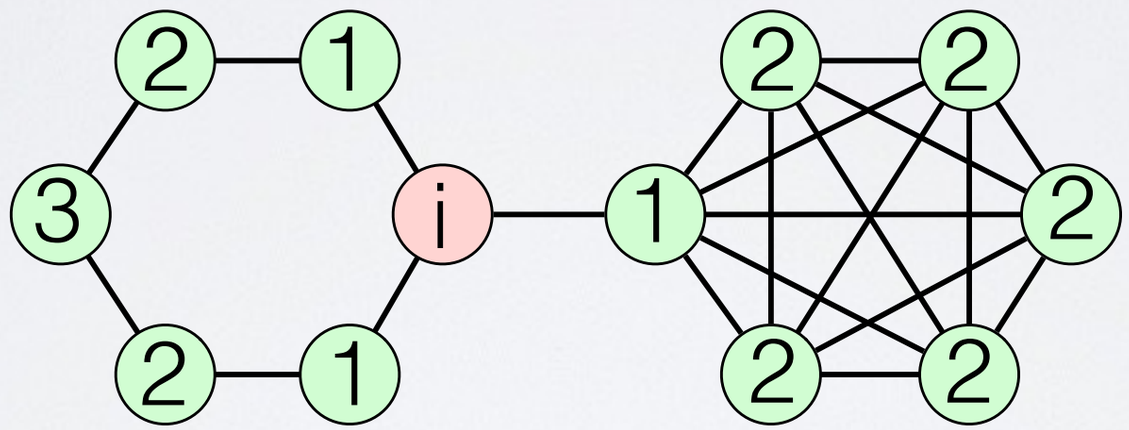


FARNESS, CLOSENESS

Farness: Distance moyenne à tous les nœuds du graphe.

$$\text{Farness}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \ell_{u,v}$$

CLOSENESS CENTRALITY



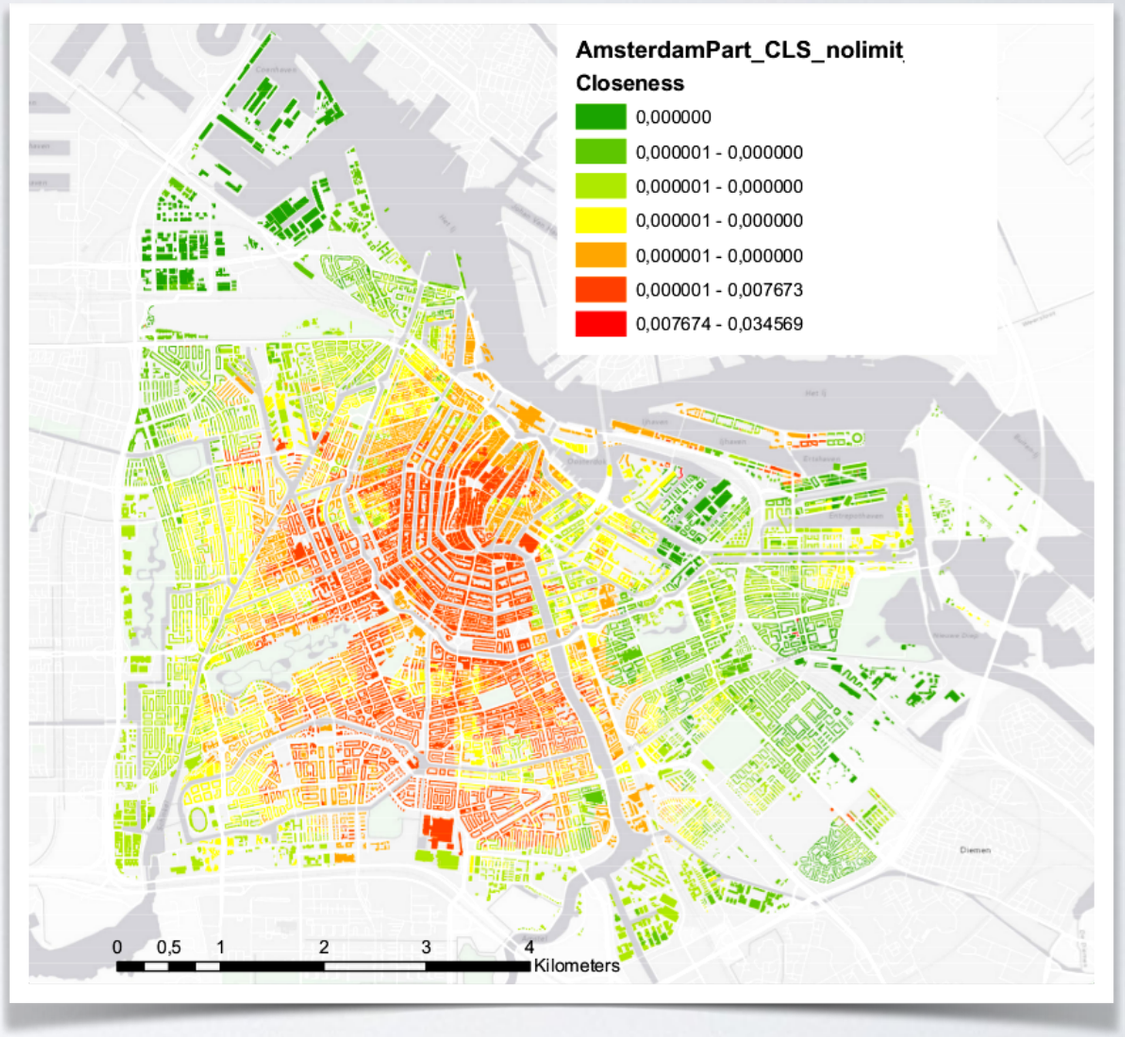
$$C_{cl}(i) = \frac{12 - 1}{(3 \times 1 + 7 \times 2 + 1 \times 3)} = \frac{11}{20} = 0.55$$

CLOSENESS CENTRALITY

Closeness: Inverse de la farness

$$\text{Closeness}(u) = \frac{N-1}{\sum_{v \in V \setminus u} \ell_{u,v}}$$

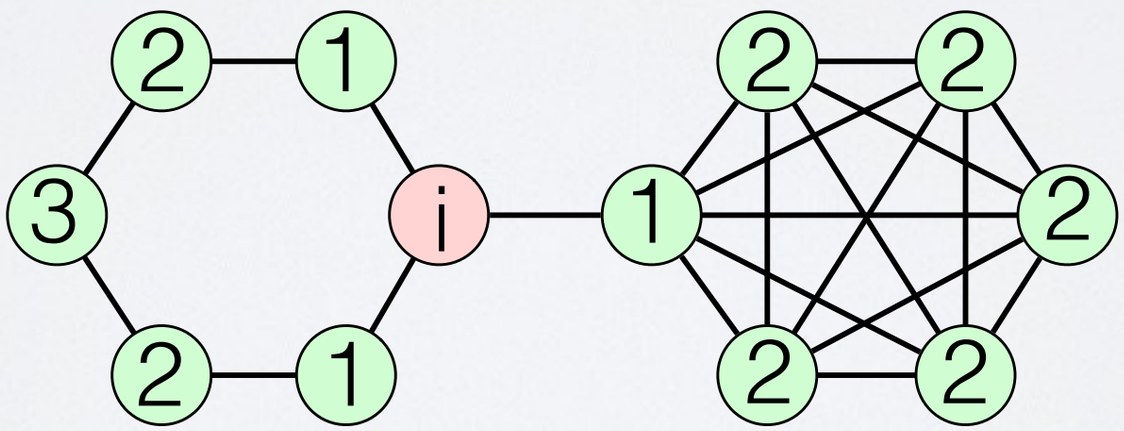
I =tous les nœuds sont à distance I



Harmonic Centrality

Centralité Harmonique: Une variante de la closeness définie comme la moyenne des inverses des distances à tous les autres nœuds (moyenne harmonique). Cette mesure est définie même sur des graphes non connexes, à condition de définir $\frac{1}{\infty} = 0$. Son interprétation est la même que la Closeness.

$$\text{Harmonic}(u) = \frac{1}{N-1} \sum_{v \in V \setminus u} \frac{1}{\ell_{u,v}}$$



$$C_h(i) = \frac{1}{12-1}\left(3 \times \frac{1}{1} + 7 \times \frac{1}{2} + 1 \times \frac{1}{3}\right) = \frac{41}{66} = 0.6212$$

BETWEENNESS CENTRALITY

Centralité d'intermédiarité

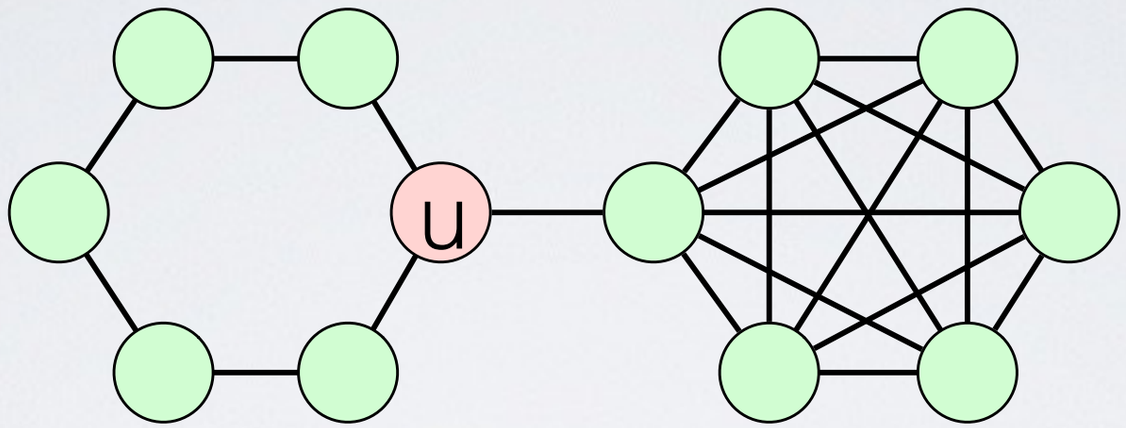
- Mesure à quel point le nœud joue le rôle d'un pont
- Betweenness de *u:* fraction de tous les plus courts chemins entre tous les nœuds qui passent par *u*

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

avec σ_{st} le nombre de plus court chemins entre s et t et $\sigma_{st}(v)$ le nombre de ces chemins qui passent par le nœud v.
La betweenness tend à augmenter avec la taille du graphe. Une version normalisée peut être obtenue en divisant par le nombre de paires de nœuds, pour un graphe dirigé: $C_B^{\text{norm}}(v) = \frac{C_B(v)}{(N-1)(N-2)}$.

Betweenness Centrality

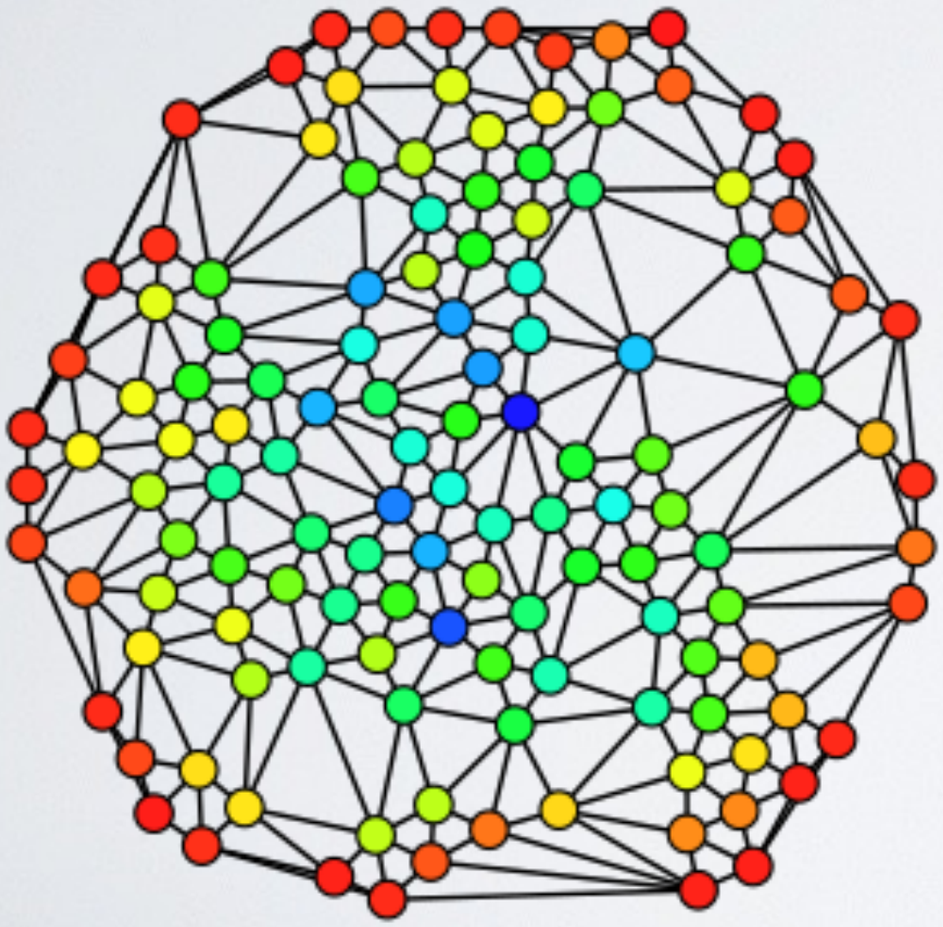
$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$



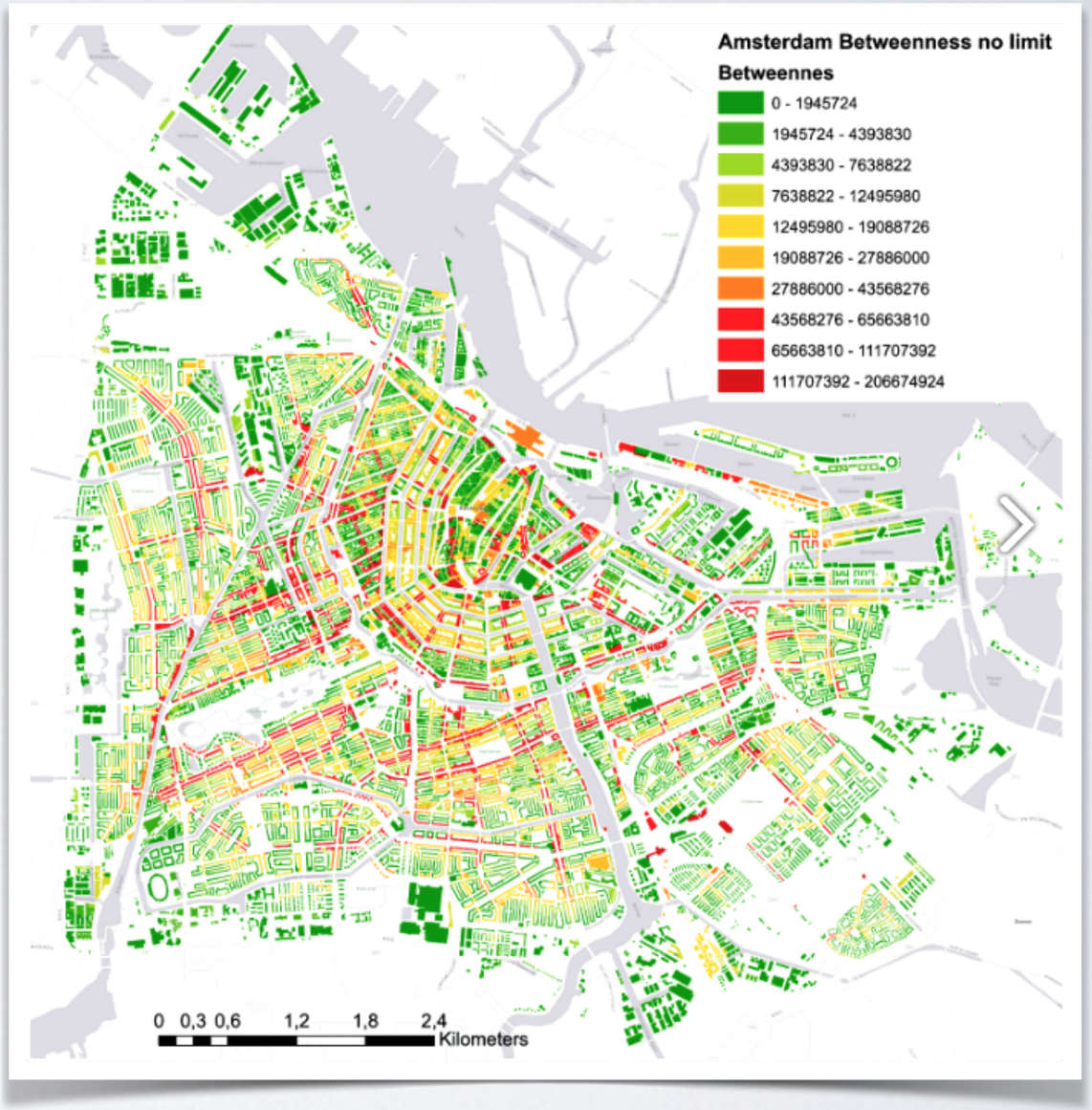
$$C_B(u) = 2\frac{5 * 6 + 1 + \frac{1}{2} + \frac{1}{2}}{11 * 10} = \frac{64}{110}$$

Peut être calculé de manière exacte (très couteux) ou approximative

BETWEENNESS CENTRALITY



Bleu valeur élevées



Rouge élevées

EDGE - BETWEENNESS

Intermédiarité des liens

Même définition que pour les nœuds.

Lien de plus forte betweenness dans le réseau ferroviaire Européen ?

DÉFINITIONS RÉCURSIVES

DÉFINITIONS RÉCURSIVES

- Importance récursive:
  - **Un nœud est important** s'il est connecté à (pointé par) ***des nœuds importants***

- Plusieurs centrales sont basées sur ce principe:
  - Centralité Eigenvectors ("valeurs propres")
  - PageRank
  - Hub et Autorités

DÉFINITIONS RÉCURSIVES

- Définissons l'objectif:
  - Chaque nœud à un score (centrality),
  - Si chaque nœud envoie son score à ses voisins, la somme (normalisée) des scores qu'il a reçu est égale à sa valeur de centralité

$$C_u^{t+1} = \frac{1}{\lambda} \sum_{v \in N_u^{in}} C_v^t \qquad (1)$$

- Avec λ une constante de normalisation

DÉFINITIONS RÉCURSIVES

- Plus qu'à trouver une solution mathématique à ce problème
- Peut être résolue par la *power method (méthode des puissances itérées)*
  - 1) Tous les scores sont initialisés avec des valeurs aléatoires entre 0 et 1
  - 2)On applique la règle définie auparavant jusqu'à atteindre un point stable (les valeurs ne changent plus, donc objectif atteint)
  - Garantie de convergence vers un point stable
- Pourquoi est-ce que ça marche?
  - Théorème de Perron-Frobenius
  - => Vrai pour des graphes non-dirigés avec une seule composante connexe.

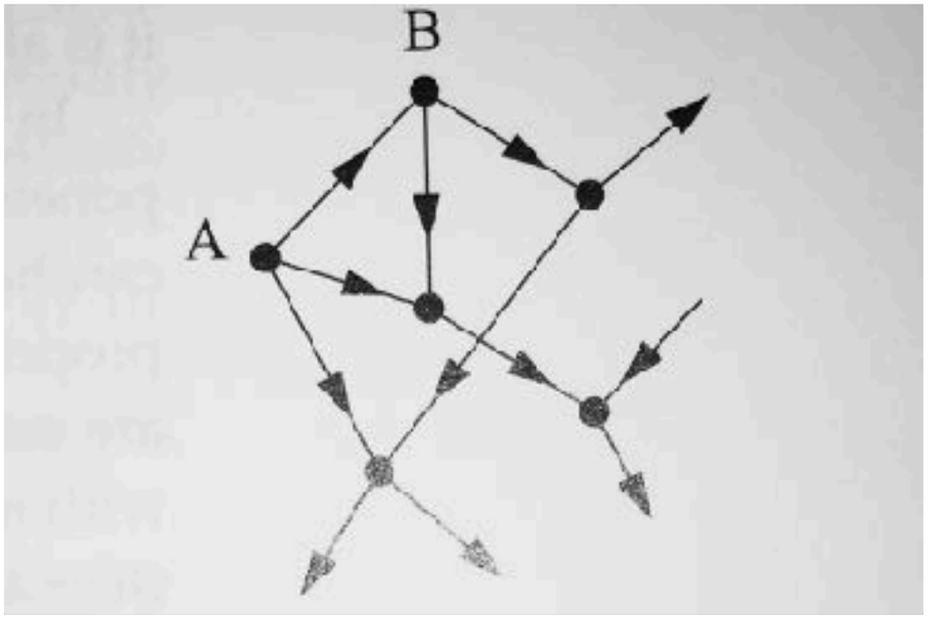
CENTRALITÉ EIGENVECTORS

- Ce que l'on viens de décrire : Centralité Eigenvector
- Un couple vecteur propre (x) et valeur propre (λ) est défini par la relation: $Ax = \lambda x$
  - x est un vecteur colonne de taille *n*, interprété comme les scores de nœuds
- Ce que dit le théorème Perron-Frobenius est que la méthodes des puissances va converger vers *le premier vecteur propre*, i.e., le vecteur propre associé à la valeur propre la plus élevée.

Eigenvector Centrality

Des problèmes avec les graphes dirigés:

- 2 ensembles de vecteurs propres (Gauche et Droit)
- On utilise les vecteurs propres droit : les nœuds envoient leurs poids dans le sens de la flèche.



Mais problème avec les nœuds source (degré entrant=0)
-Nœud A n'a que des liens sortants = sa centralité après la première itération

est 0

-Nœud B a des liens entrants et sortants, mais son lien entrant viens de A = Centralité de 0 au second tour

-etc.

Solution: Calcul seulement dans la plus grande composante connexe forte

Note: Les réseaux acycliques (e.g., réseau de citation) n'ont pas de composante connexe forte

PageRank Centrality

- Centralité Eigenvector généralisée aux graphes dirigés

PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

PageRank Centrality

- Eigenvector centrality generalised for directed networks

PageRank

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full text and hyperlink database of at least 24 million pages is available at http://google.stanford.edu/

PageRank Centrality

(Side notes)

-"We chose our system name, Google, because it
is a common spelling of googol, or 10^{100} and fits well with our goal of building very large-scale search "

-"[…] at the same time, search engines have migrated from the academic domain to the commercial. **Up until now most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art and to be advertising oriented (see Appendix A). With Google, we have a strong goal to push more development and understanding into the academic realm."**

-"[...], we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers."

PageRank Centrality

(Side notes)

Sergey Brin received his B.S. degree in mathematics and computer science from the University of Maryland at College Park in 1993. Currently, he is a Ph.D. candidate in computer science at Stanford University where he received his M.S. in 1995. He is a recipient of a National Science Foundation Graduate Fellowship. His research interests include search engines, information extraction from unstructured sources, and data mining of large text collections and scientific data.

Lawrence Page was born in East Lansing, Michigan, and received a B.S.E. in Computer Engineering at the University of Michigan Ann Arbor in 1995. He is currently a Ph.D. candidate in Computer Science at Stanford University. Some of his research interests include the link structure of the web, human computer interaction, search engines, scalability of information access interfaces, and personal data mining.

PAGERANK

- 2 améliorations principales:
  - Problème des nœuds source
    - => Ajout d'un petit gain constant à tous les nœuds ("téléportation")
  - Les nœuds de centrality forte donnent une centralité forte à tous leurs voisins (Même s'ils en ont énormément, et que certain n'ont pas d'autres entrées)
    - => Ce que chaque nœud "vaut" est divisé entre ses liens sortants (Normalisation par le degré)

$$C_u^{t+1} = \frac{1}{\lambda} \sum_{v \in N_u^{in}} C_v^t \quad \Rightarrow \quad C_u^{t+1} = \alpha \sum_{v \in N_u^{in}} \frac{C_v^t}{k_v^{out}} + \beta$$

With by convention $\beta = 1$ and α a parameter (usually 0.85) controlling the relative importance of β

PageRank - Marche aléatoire

Compréhension intuitive : Interprétation en tant que marche aléatoire

Marche aléatoire : On démarre d'un nœud pris au hasard, puis on "marche" au hasard dans le réseau, en suivant les liens. (On choisit un lien sortant au hasard)

Probabilité de téléportation : le paramètre α correspond à la probabilité de faire ce processus normalement, et $1\text{-}\alpha$ de sauter aléatoirement à n'importe quel nœud à la place.

Pagerank score d'un nœud correspond à la probabilité que le marcheur aléatoire soit sur ce nœud après un nombre infini de déplacements.

PAGERANK

- Comment Google classe les résultats de nos recherches ?

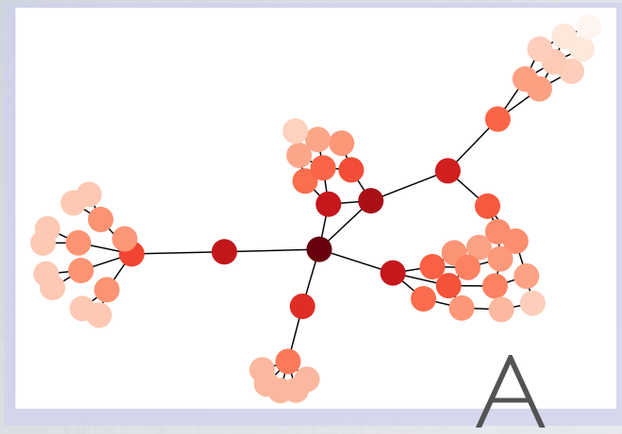
- Calcul de pagerank (Power method)

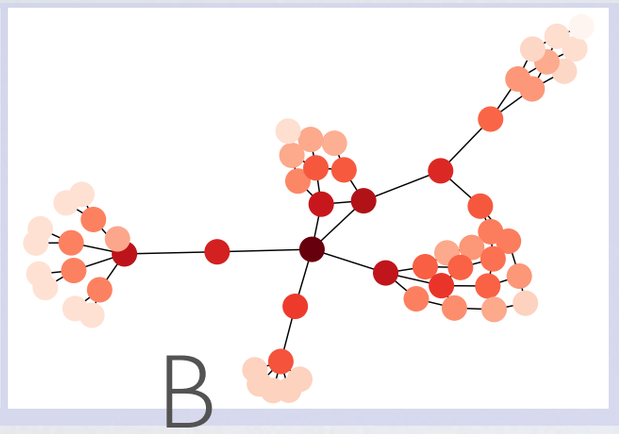
- Filtre les pages contenant les mots recherchés

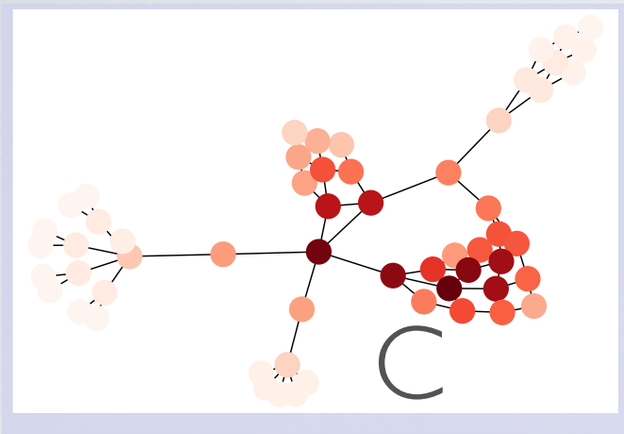
- Bien sûr aujourd'hui les méthodes sont plus complexes, mais non publiques:
  "Most search engine development has gone on at companies with little publication of technical details. This causes search engine technology to remain largely a black art" [Page, Brin, 1997]

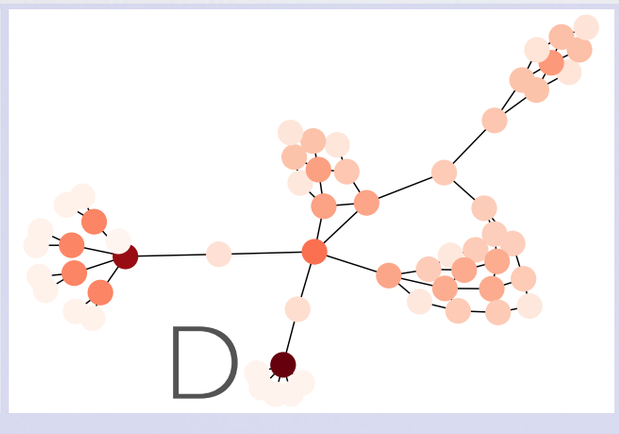
OTHERS

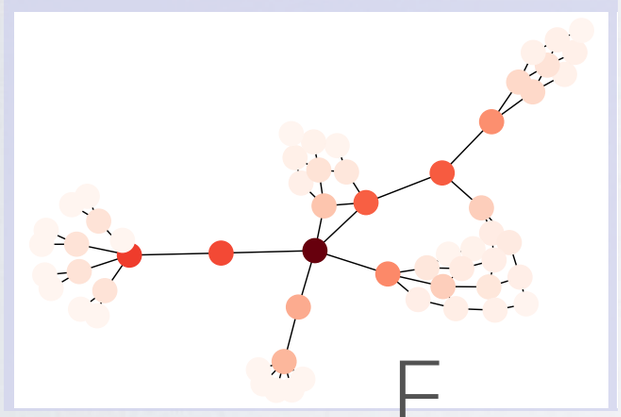
- Beaucoup d'autres centralités ont été proposées
- Le problème est souvent comment les interpréter

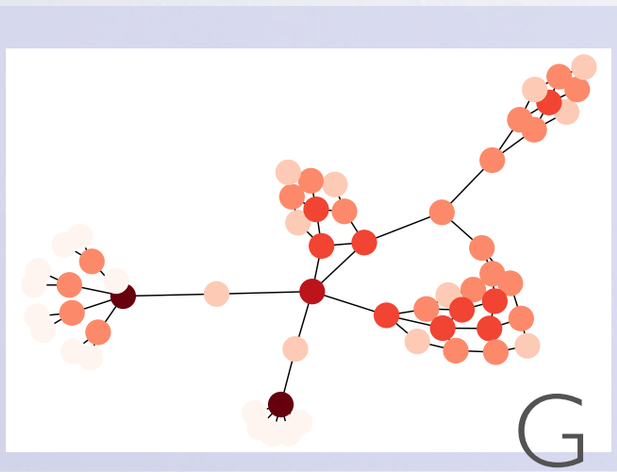


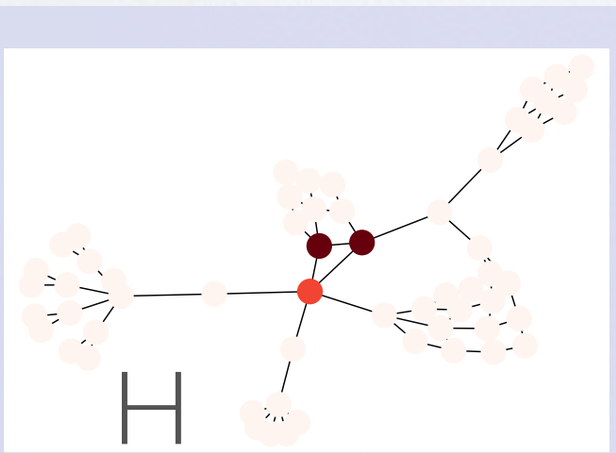








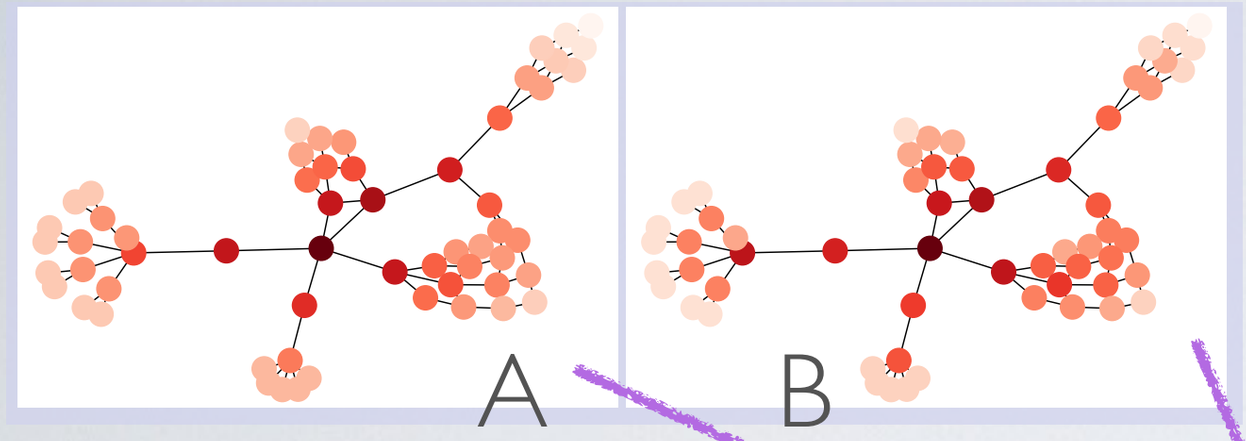


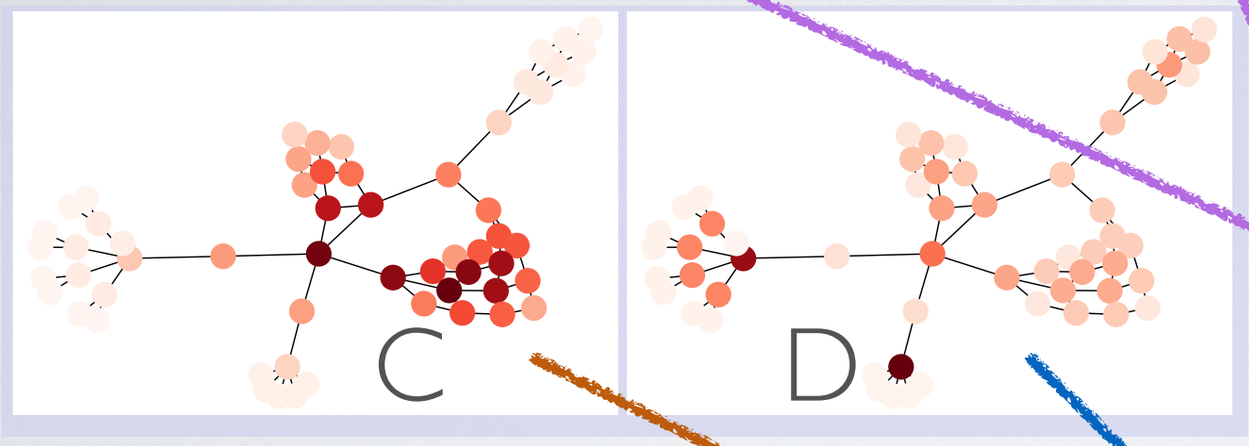


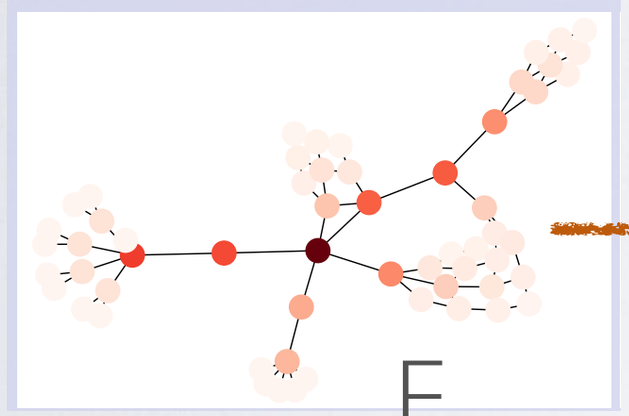
Qui est qui ?

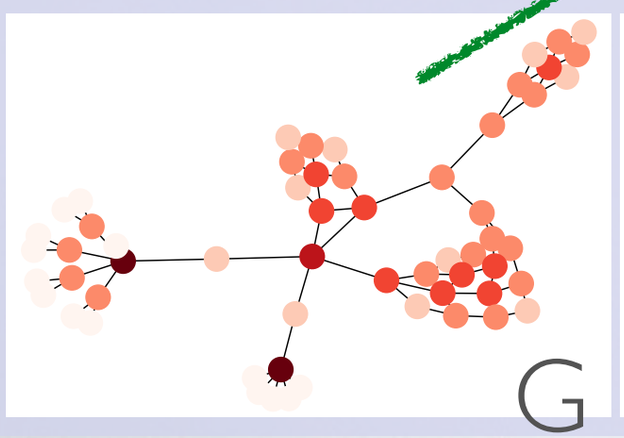
Degré
Clustering coefficient
Closeness
Harmonic Centrality
Betweenness

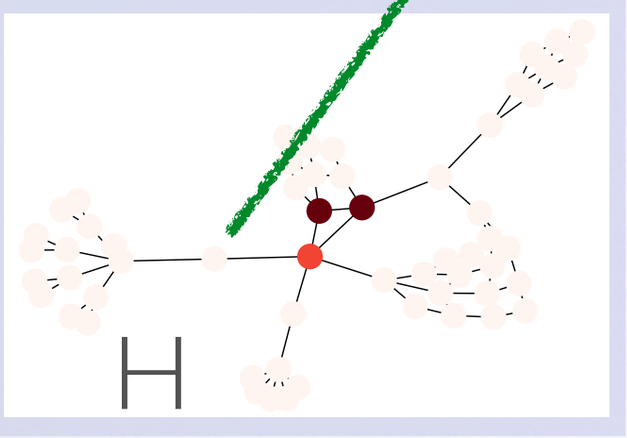
Eigenvector
PageRank











Qui est qui ?

Degré

Clustering coefficient

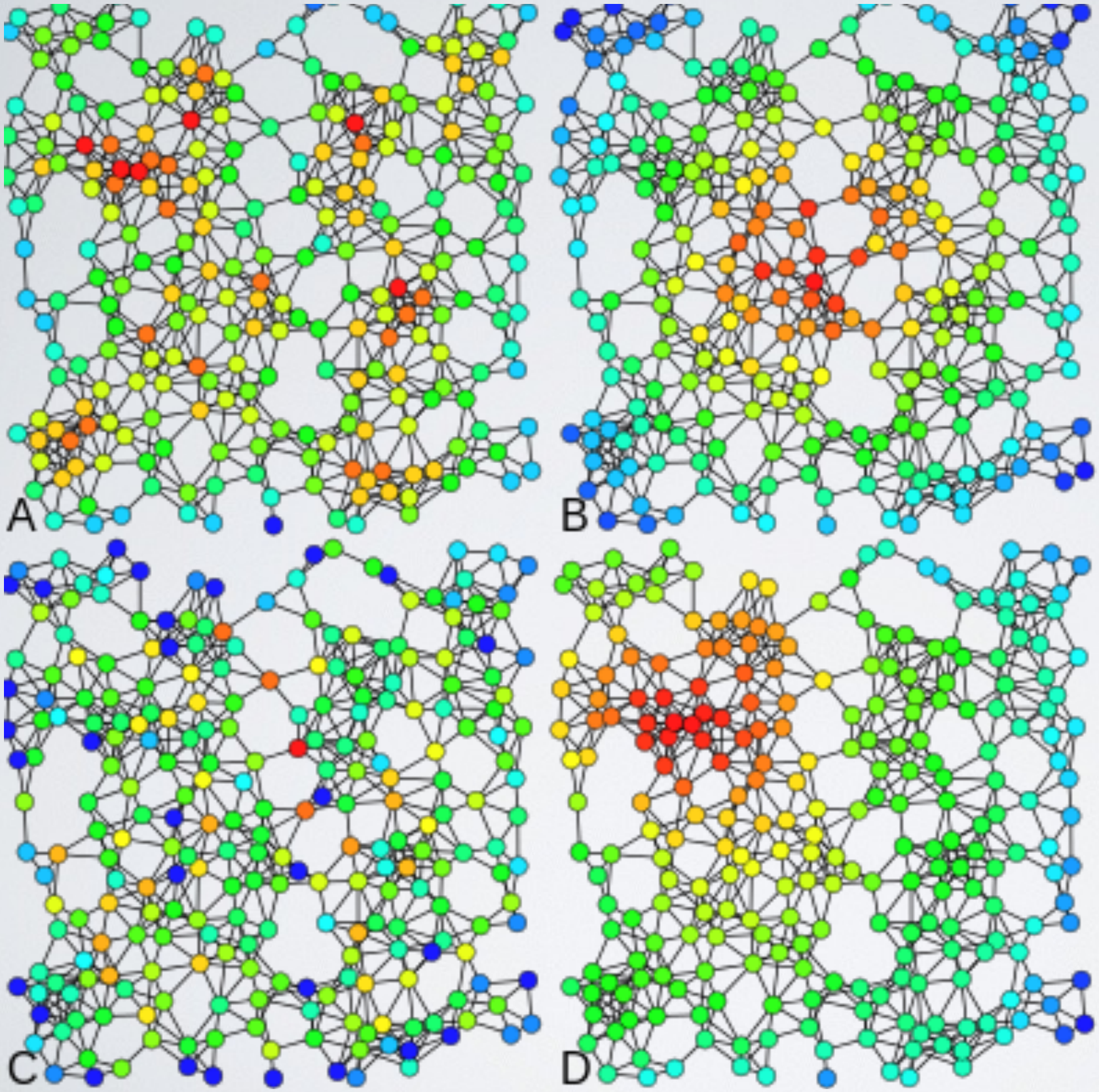
Closeness

Harmonic Centrality

Betweenness

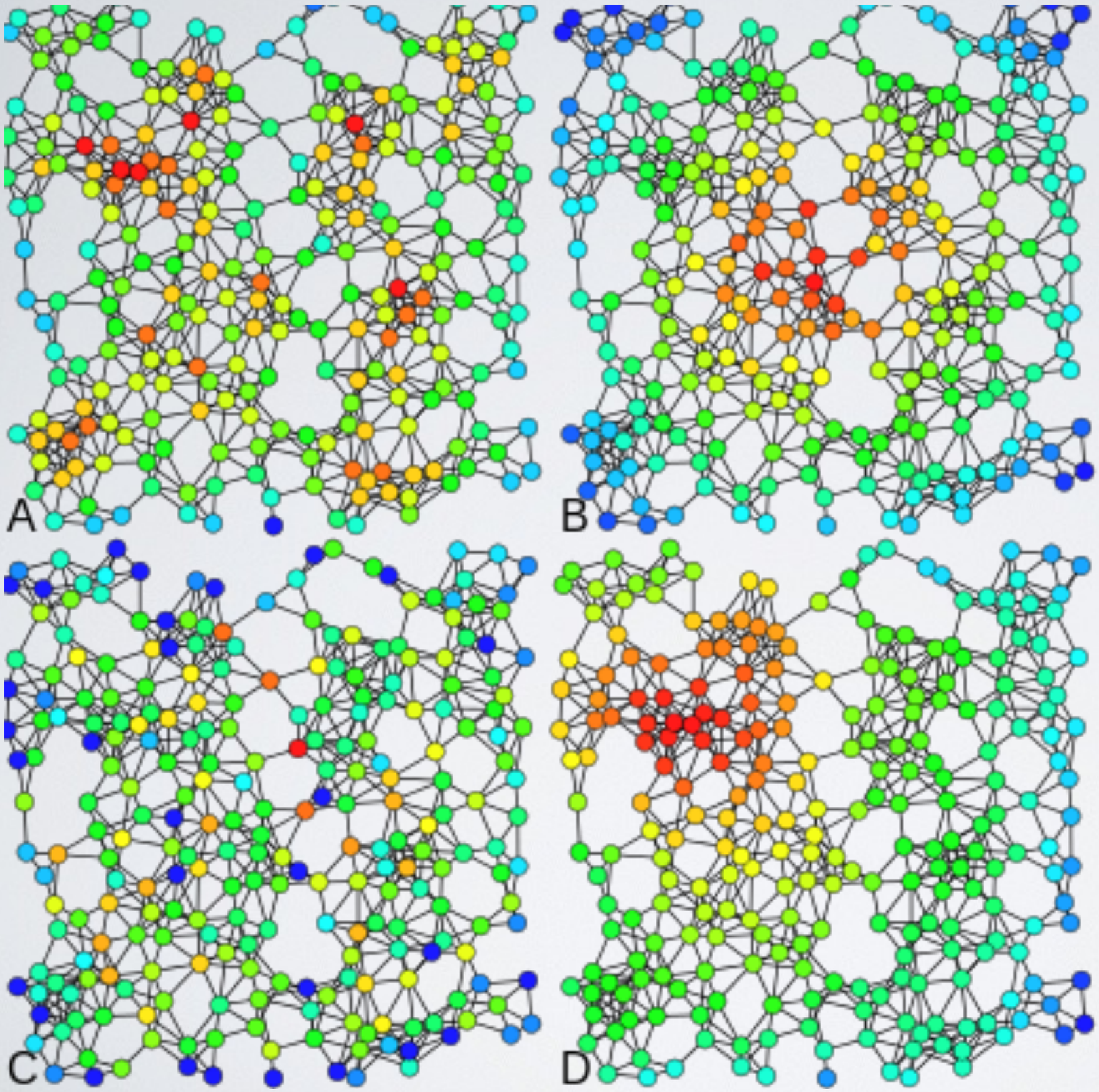
Eigenvector

PageRank



Autre essai :)

Degree
Betweenness
Closeness
Eigenvector

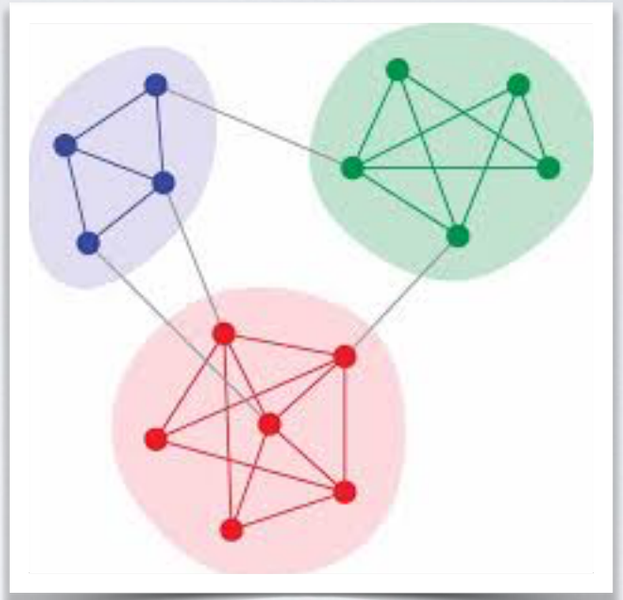


Autre essai :)

A: Degree
B:Closeness
C: Betweenness
D: Eigenvector

DÉTECTION DE COMMUNAUTÉS

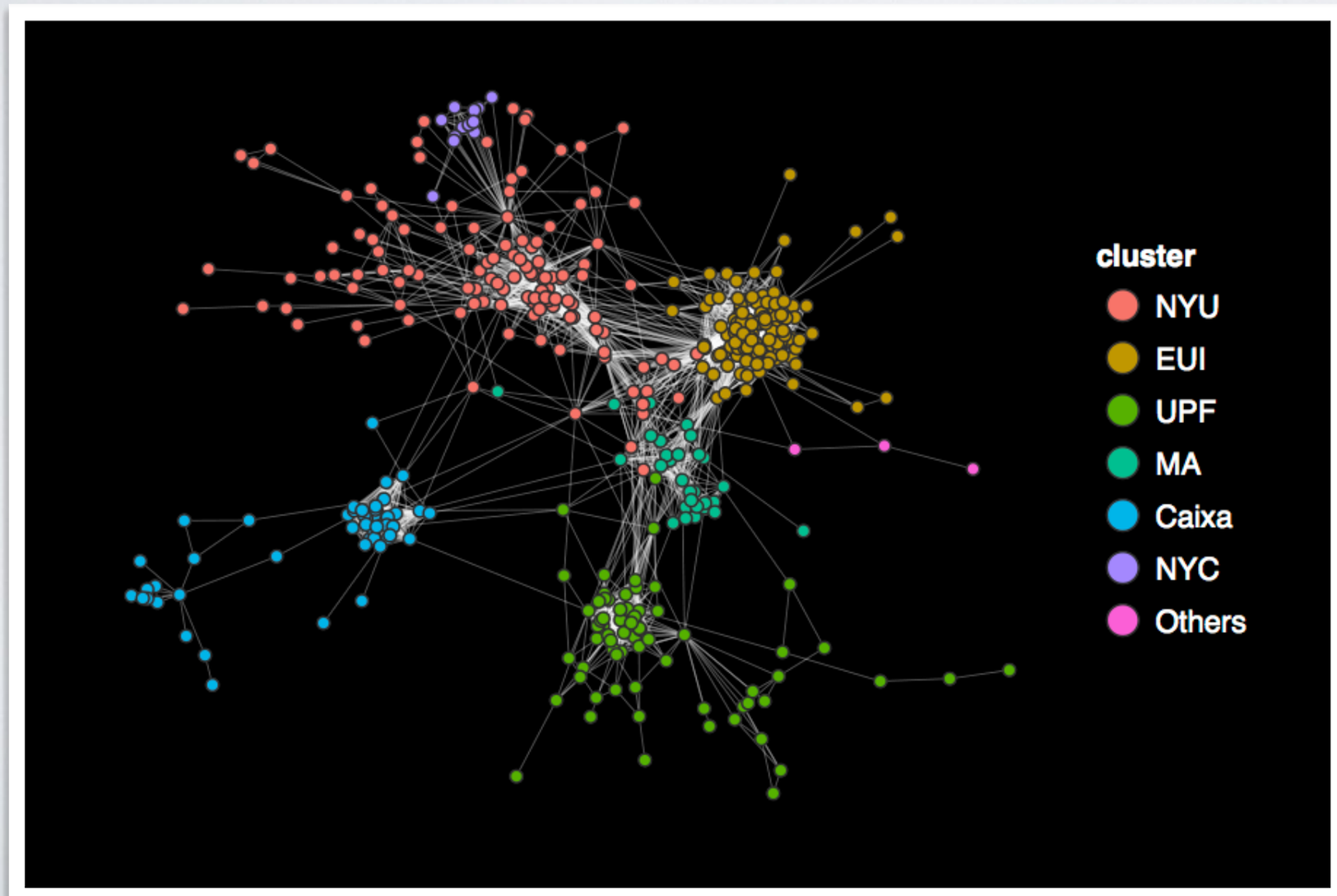
COMMUNAUTÉS



- Détection de communautés
- Découvrir des groupes de nœuds:
  - Fortement connectés entre eux
  - Faiblement connecté au reste du réseau
- Pas de définition mathématique universelle
  - Plusieurs définitions imparfaites existe

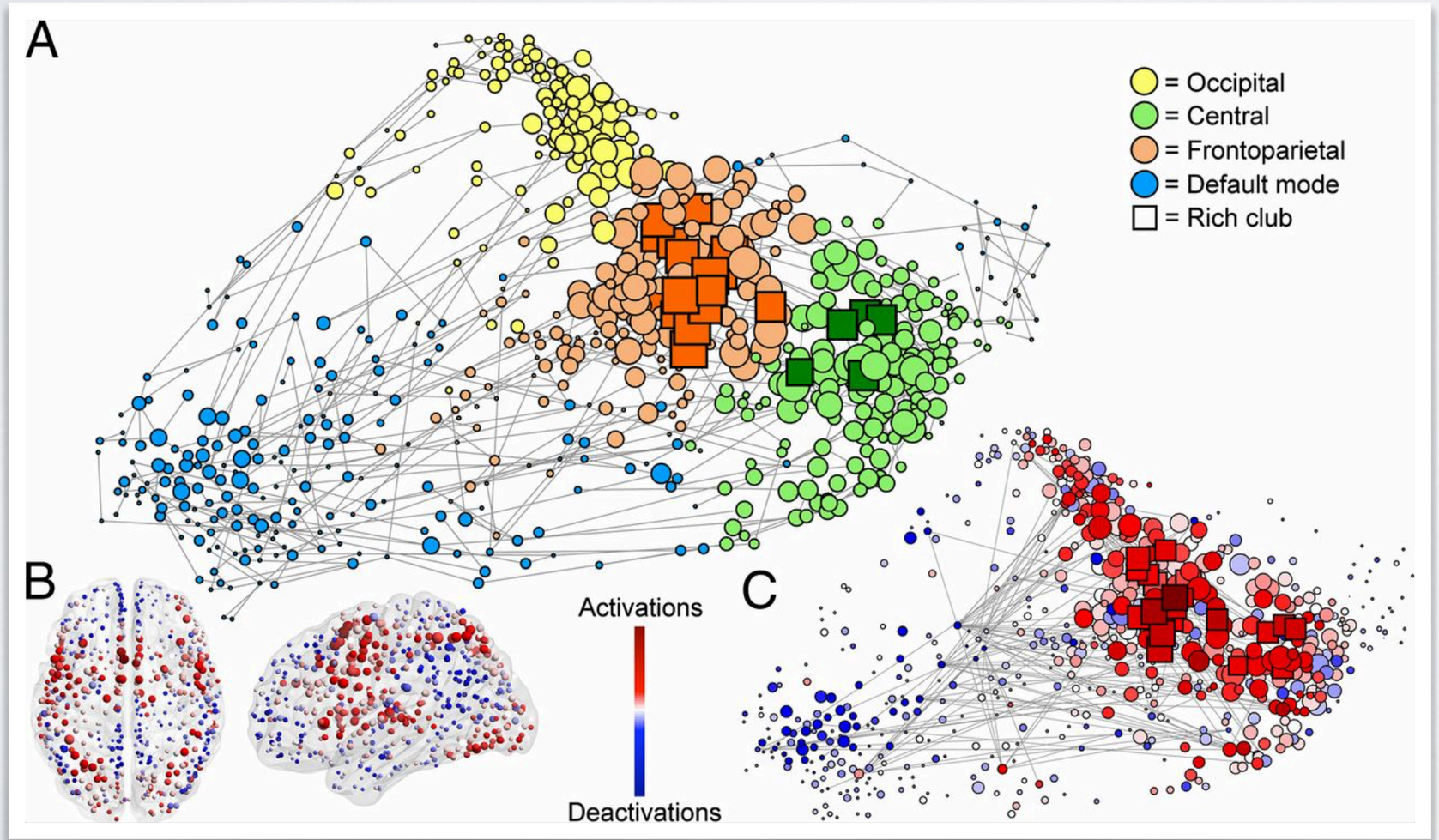
COMMUNAUTÉS DANS DES GRAPHES RÉELS

- Réseau social (Linked-in/Facebook/etc.)



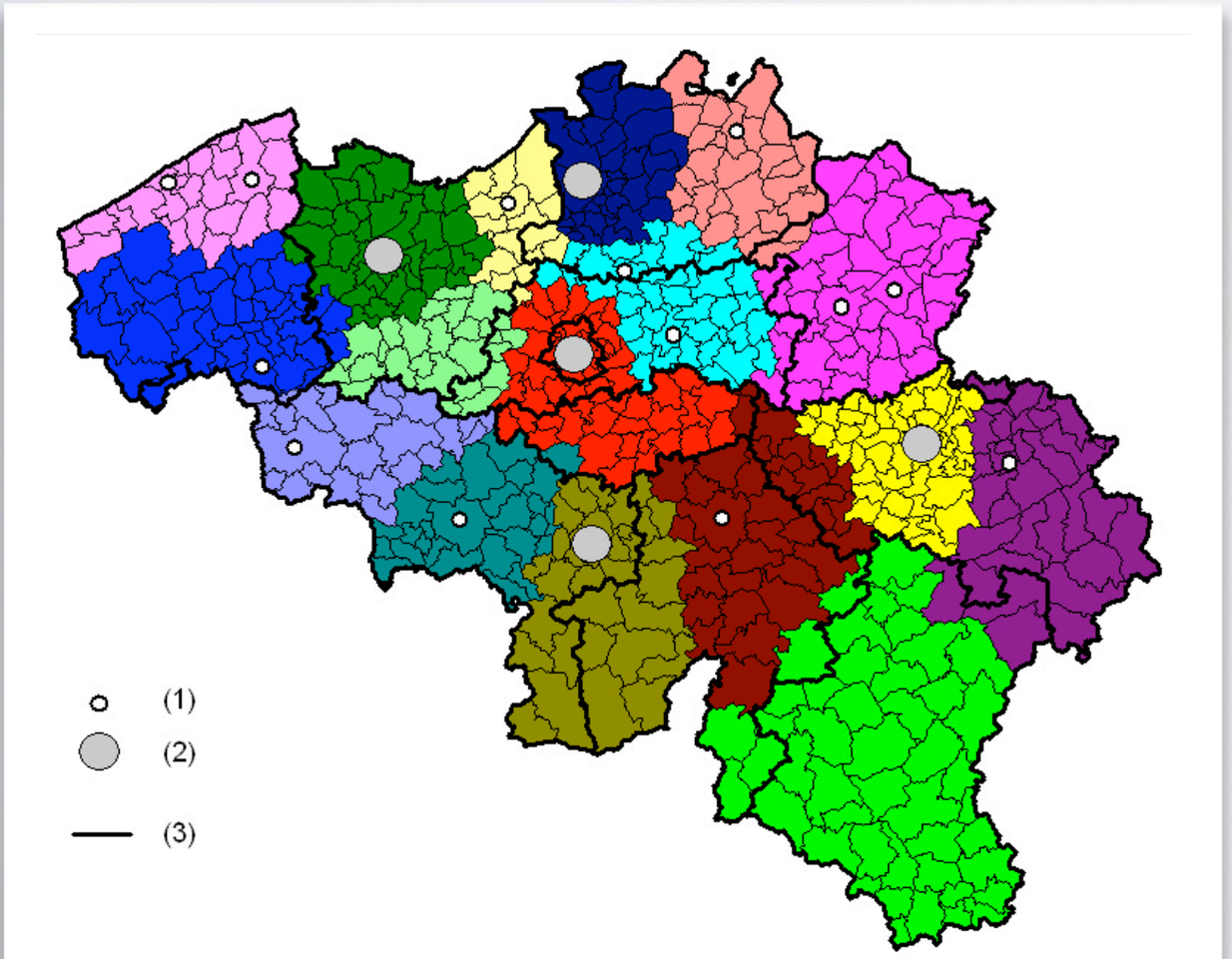
COMMUNAUTÉS DANS DES GRAPHES RÉELS

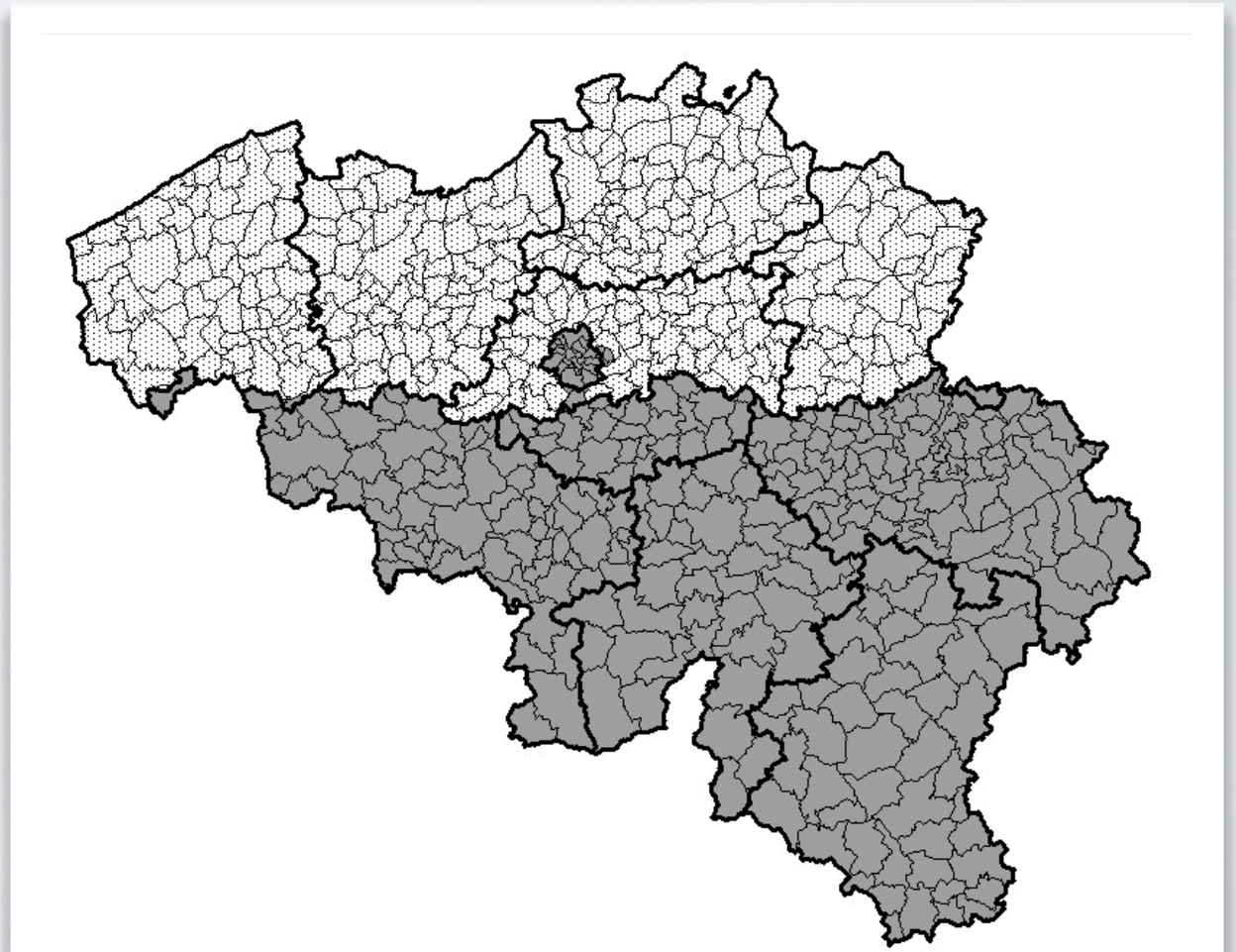
- Connexions dans le cerveau



COMMUNAUTÉS DANS DES GRAPHES RÉELS

- Appels téléphoniques en Belgique ?





ALGORITHME DE LOUVAIN

- Algorithme le plus connu, et celui présent dans Gephi
- Cherche à optimiser la **modularité**
  - **Modularité:** un score de "qualité" des communautés
  - L'algorithme cherche parmi toutes les partitions possibles celle de meilleur score (Approche gloutonne)
- Partitions *non recouvrantes*: chaque nœud appartient à une et une seule communauté.

MODULARITÉ

- Défini comme la différence entre:
  - La fraction des liens **observés** à l'intérieur des communautés
  - La fraction des liens **attendus** à l'intérieur des communautés
    - Attendu si les liens étaient distribués au hasard dans un graphe où l'on conserve le nombre de nœuds, de liens, et les degrés de chaque nœud.

$$Q = \frac{1}{L} \sum_{i=1}^{|C|} (L_i - \frac{1}{2} K_i^2)$$

with $L_i = L(H(c_i))$ the number of edges inside community i and $K_i = \sum_{u \in c_i} k_u$ the sum of degrees of nodes in community i.

ALGORITHME DE LOUVAIN

- On donne un graphe à l'algorithme, il retourne une partition (un ensemble de communautés)

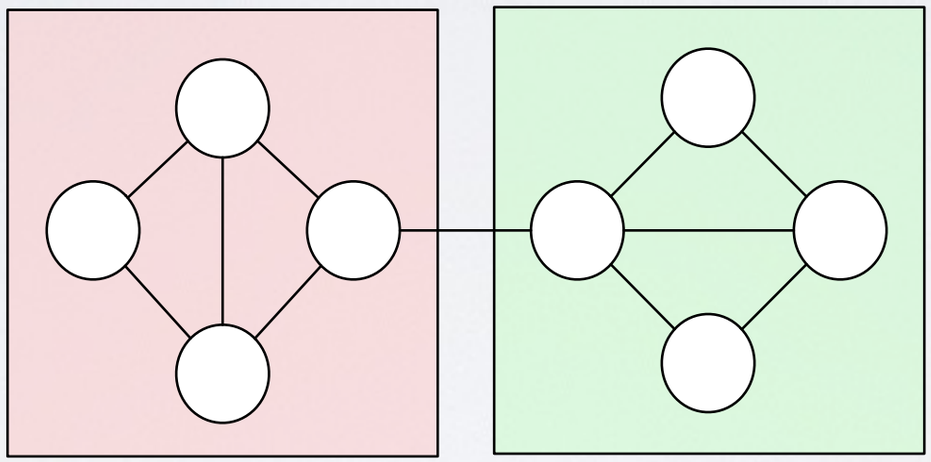
- Attention, définition imparfaite:
  - Limite de modularité=>paramètre de résolution
  - Pas de garantie de trouver la "meilleure" solution
  - Algorithme stochastique: 2 exécutions peuvent renvoyer des résultats différents

- En pratique, fonctionne très bien

MODULARITY INTUITION

$n = 8$

$m = 11$



ER random graph

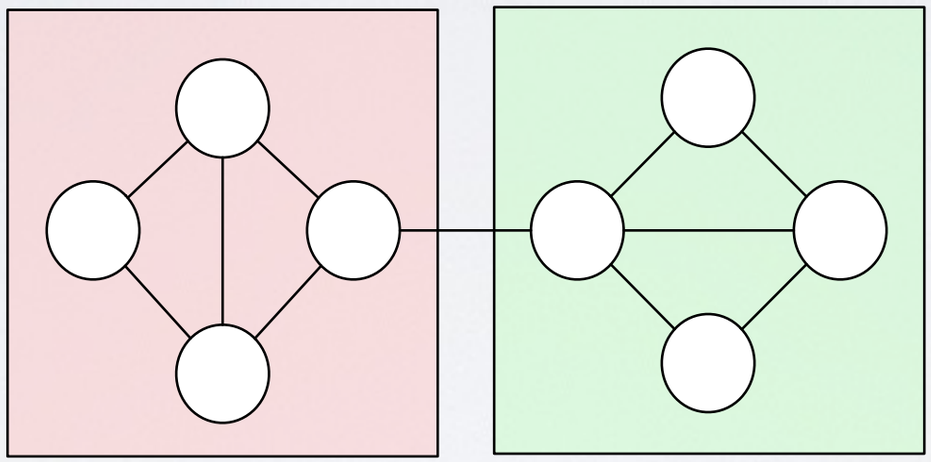
MODULARITY INTUITION

$n = 8$

$m = 11$

$p(u, v) \approx 0.39$

$$d(G) = p(u, v) = \frac{11}{\frac{1}{2}8(8-1)} = \frac{11}{28} \approx 0.39$$

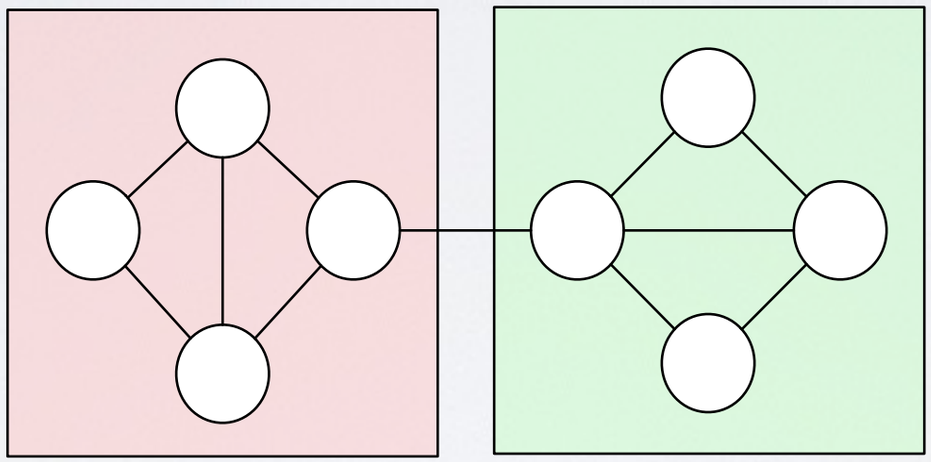


MODULARITY INTUITION

$n = 8$

$m = 11$

$p(u, v) \approx 0.39$



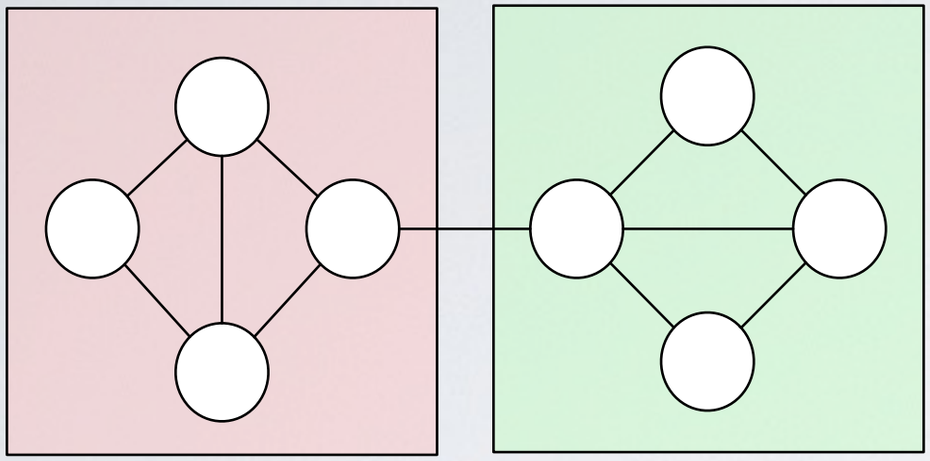
ER random graph

Expected edges inside red (or green)
(#node pairs * prob to observe an edge)

$$\frac{4(4-1)}{2} * p(u, v) = 2.34$$

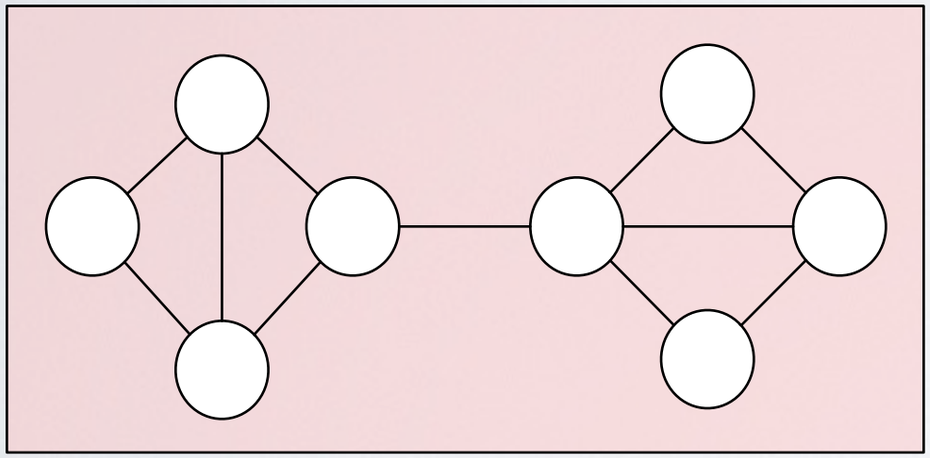
$$\text{Modularity} = \frac{2(5 - 2.34)}{m} = 0.48$$

MODULARITY INTUITION

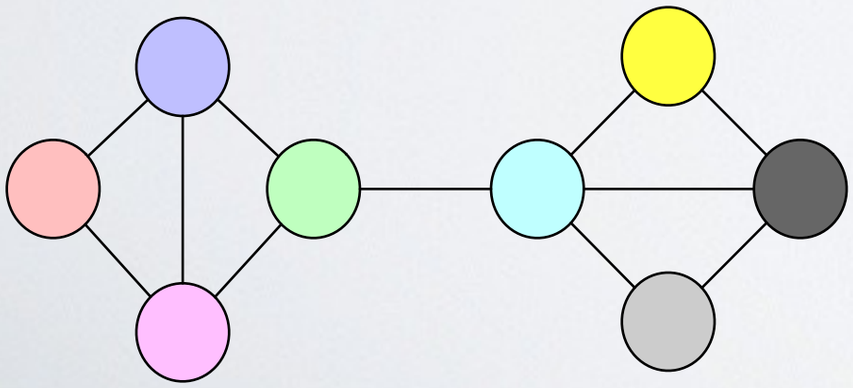


$n = 8$
$m = 11$

Q= 0.48

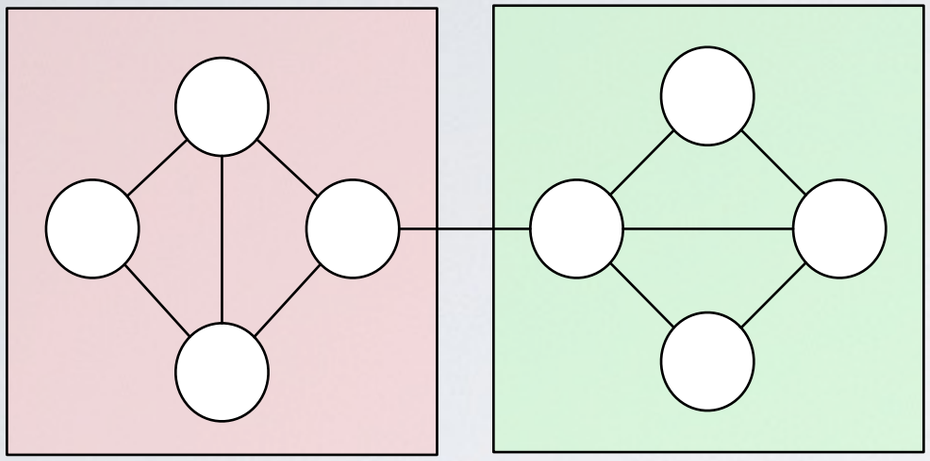


Q= ?



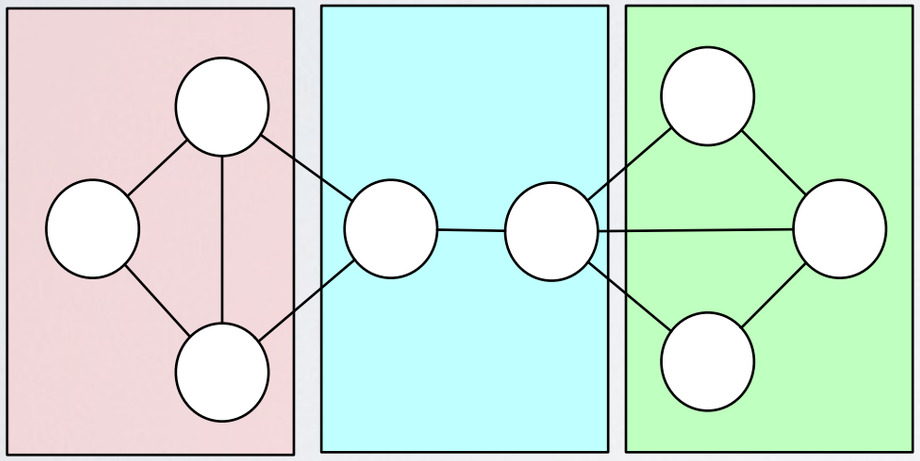
Q= ?

MODULARITY INTUITION

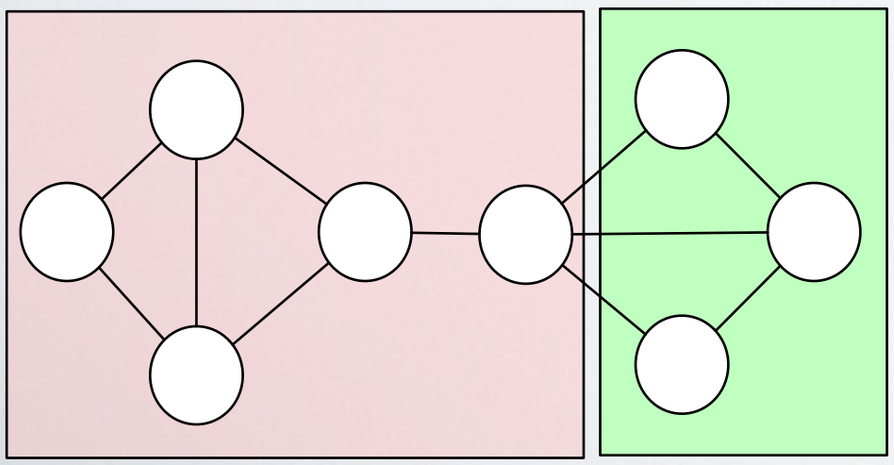


p=0.39

Q= (5-6p)+(5-6p)=10-12p=5.32

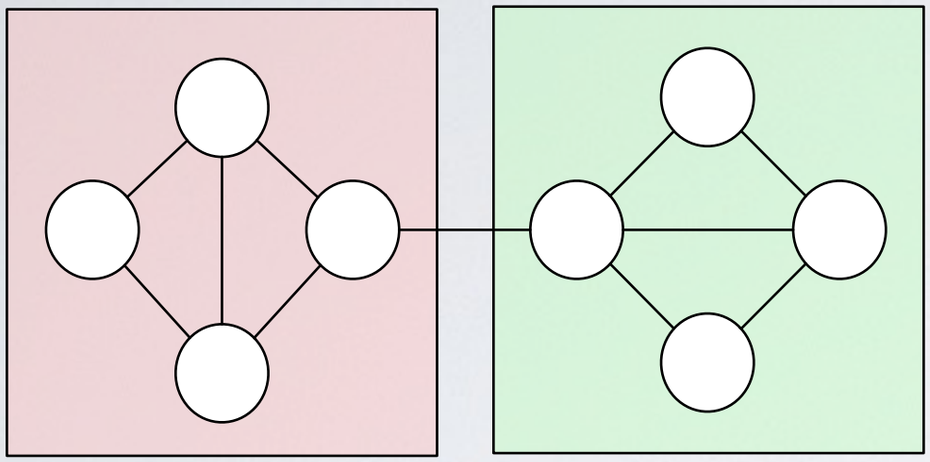


Q=(3-3p)+(1-p)+(2-3p)=7-7p=4.27



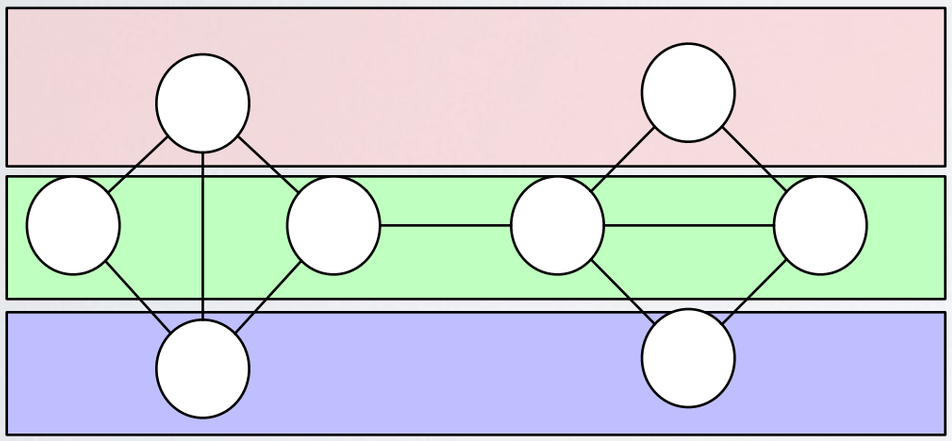
Q=(6-10p)+(2-3p)=8-13p=2.93

MODULARITY INTUITION



p=0.39

Q= (5-6p)+(5-6p)=10-12p=5.32



Q=(0-p)+(2-6p)+0-p=2-8p= - 0.34