Analyse de réseaux sociaux

Daniel Aloise <daniel.aloise@polymtl.ca>

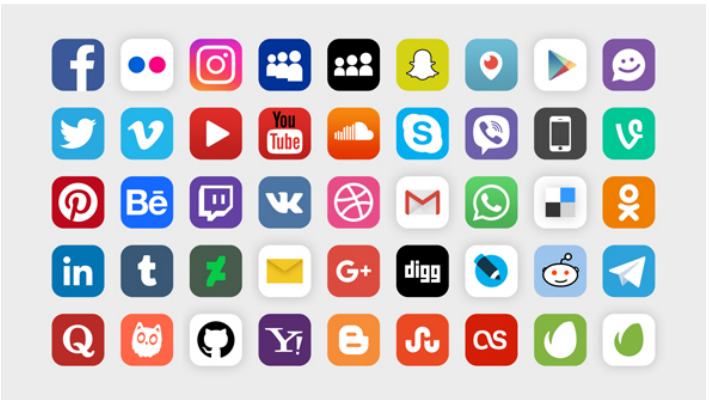
23 juin 2020

Motivation

- La tendance des humains à interagir entre eux précède l'avènement du web.
- Cependant, la popularisation du web a ouvert de nouvelles voies d'interaction.
- Aujourd'hui, ces technologies continuent d'évoluer et de créer une quantité toujours croissante de données.
- Ces données sont un trésor d'informations sur les préférences des usagers, leurs connexions, et leurs influences sur les autres.

Réseaux sociaux

- On imagine souvent les *réseaux sociaux* comme de grands réseaux sur internet (Facebook, Twitter, LinkedIn, etc.).
- Ces réseaux ne représentent qu'une minorité des mécanismes d'interaction possible :
  - La sociologie étudie les réseaux sociaux depuis plus d'un siècle !



Les six degrés de séparation

diametre d'un graphe <= 6

2 personnes dans le globe peuvent etre reliés entre 6 sommets. Ex: Une personne se connait avec une autre, qui connait son amie.... Qq1 qui connait qq1 qui connait qq1...

Définitions

Graphe Non orienté: Fb non orienté, car si je te add comme ami tu dois me add aussi
Graphe Orienté: Twitter, si je vous follow, vous etes pas obliges de me follow. unidirectionnel

- Un réseau social peut être structuré comme un graphe $G = (N, A)$, où N est l'ensemble des sommets et A est l'ensemble des arêtes.
- Chaque entité dans le réseau social est représentée par un sommet dans N.
- Les arêtes de A représentent les connexions entre les différentes entités :
  - Facebook : relation d'amitié (symétrique)
  - Twitter : relation de *follower* (asymétrique)

Propriétés

ceux qui se ressemble s'assemblent

Homophilie les sommets connectés les uns aux autres sont plus susceptibles d'avoir des propriétés similaires.

Fermeture triadique si deux entités d'un réseau social ont un "ami" en commun, alors il est plus probable qu'elles soient connectées ou qu'elles vont éventuellement se connecter dans l'avenir

Si moi et Ming on a un ami en commun, il y a des chances qu'on soit amis aussi ou qu'on sera amis plus tard

Le concept structurel de la fermeture triadique est directement lié au coefficient de regroupement du réseau.

Propriétés

tendance d'un graphe à se rassembler en groupe

- Le coefficient de regroupement peut être considéré comme une mesure de la tendance inhérente d'un réseau à se regrouper.
- Soit $S_i \subseteq N$ l'ensemble de sommets connectés au sommet $i \in N$, et $n_i = |S_i|$ Si: voisins du point i

ni = nb de voisins du point i

- $C_{n_i}^2 = \binom{n_i}{2} = \frac{n_i(n_i-1)}{2}$ arêtes possibles entre les sommets de S_i.

voisins du point i

- Le coefficient de regroupement du réseau est la valeur moyenne de η_i sur tous les sommets du réseau, avec :

EXAMEN enregistrement : minute 10

A: ensemble des arretes Si: voisins du point i

$$\eta_i = \frac{|\{(j,k) \in A : j \in S_i, k \in Si\}|}{C_{n_i}^2}$$

numérateur: nb arretes entre les voisins du pts i

!!! G = moyenne des ni

A: l'ensemble des arrêtes

ni: nb de voisins du sommet étudié i

Dynamique des réseaux sociaux

- En plus des propriétés vues dans la vidéo, les réseaux sociaux présentent d'autres caractéristiques par rapport à leur dynamique :

quand on a un graphe, quand on ajoute des sommets et arretes, le sommet avec le + d'arrets a + de chances d'avoir plus de nouvelles arretes. Ex: qq1 de populaire sur ig, il est plus subeptible de recevoir des followers que qq1 comme moi qui a juste 200 followers

Attachement préférentiel Dans un réseau en croissance, la probabilité qu'un sommet reçoive des nouveaux liens augmente avec son dégrée.

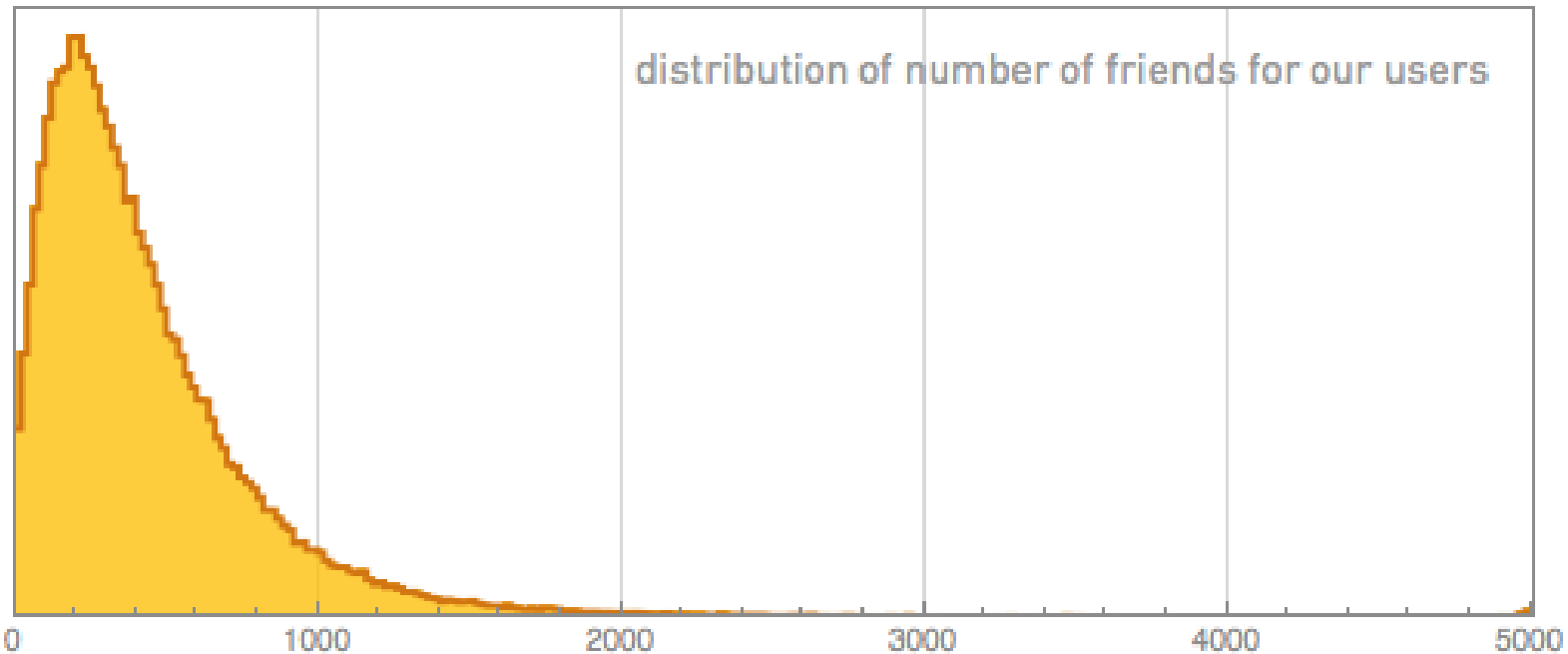
Nouveaux utilisateurs

Composant connecté géant les nouveaux liens sont plus susceptibles de se joindre aux sommets densément connectés et de haut dégrée dans le réseau.

Nouvelles arrêtes

Dynamique des réseaux sociaux

Distrib. de dégrées selon la loi de puissance : ex. Facebook



source : https ://writings.stephenwolfram.com

Mesures de centralité et prestige

- Les sommets centraux du réseau ont un impact significatif sur ses propriétés. du grpahe Ex: célébrités sur ig
- Ces sommets sont souvent plus importants parce qu'ils ont des liens avec de nombreux sommets et sont dans une position de plus grande influence.
- Dans le cas des graphes dirigés, nous parlons du prestige d'un sommet :
  - ex. Twitter

Non orienté: Centralité (Ex: FB)

Prestige: Orienté (Twitter)

Mesures de centralité et prestige

- Centralité de dégrée $C_D(i)$ d'un sommet i :

Il n'est pas possible d'être son propre ami sur Fb ou de se follow nous-mm

nb d'arretes connectés à un sommet

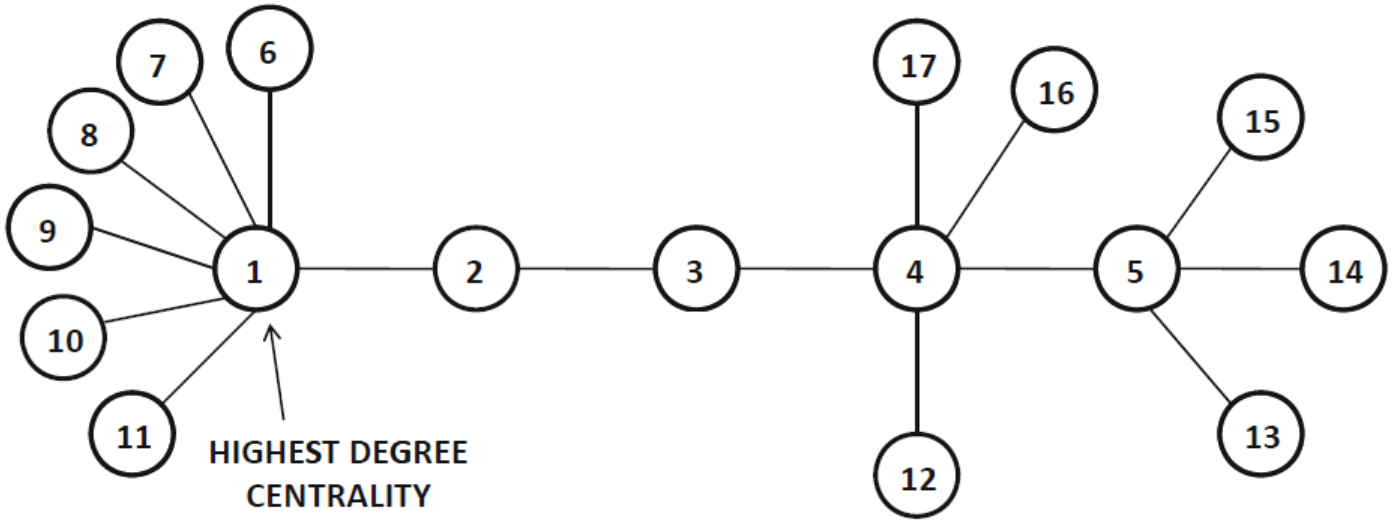
$$C_D(i) = \frac{degree(i)}{n-1}$$

n: nb total total de sommets

- Les sommets avec un dégrée plus élevé sont souvent des sommets centraux qui ont tendance à rapprocher les parties éloignées du réseau.
- Le problème majeur avec la centralité de degré est qu'elle est plutôt myope.

ont tendance à rapprocher les parties éloignées

Exemple



C. Aggarwal, 2015

- Le sommet 1 est pourtant plutôt dans la périphérie du réseau.

  a la centralité la + élevée

Mesures de centralité et prestige

- Le prestige de dégrée est défini pour les réseaux dirigés :

nb arcs entrants au sommet i

$$P_D(i) = \frac{in\_degree(i)}{n-1}$$

où $in\_degree(i)$ est le nombre d'arcs entrants du sommet i.

Mesures de centralité et prestige

- Le plus court chemin moyen d'un sommet i mesuré sur des graphes connectés est donné par :

Somme des distances. les distances les + courtes des sommets se rendant au sommet i

$$AvDist(i) = \frac{\sum_{j=1}^{n} Dist(i,j)}{n-1}$$

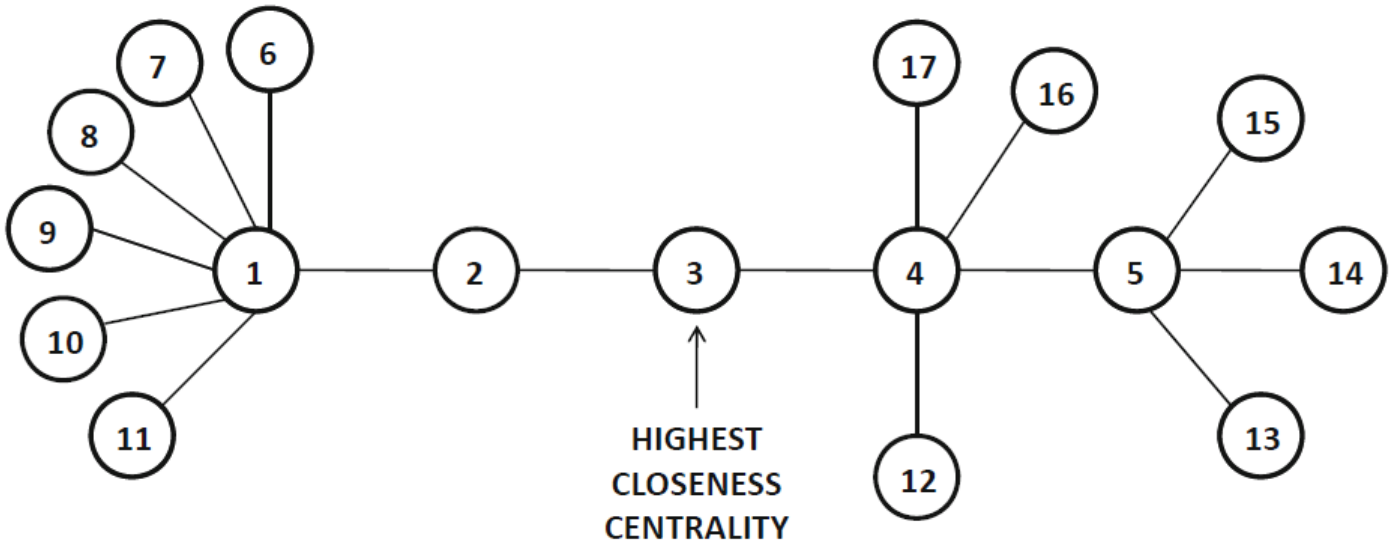
n: nb total de sommets

où $Dist(i,j)$ est la longueur du plus court chemin entre les sommets i et j dans le graphe.

- La centralité de proximité est calculée comme :

$$C_P(i) = \frac{1}{AvDist(i)} \in [0,1]$$

Exemple



Celui qui a l'arrête au milieu du graphe, c-a-d celui qui a l'arrête la + proche de tous les autres en moyenne

C. Aggarwal, 2015

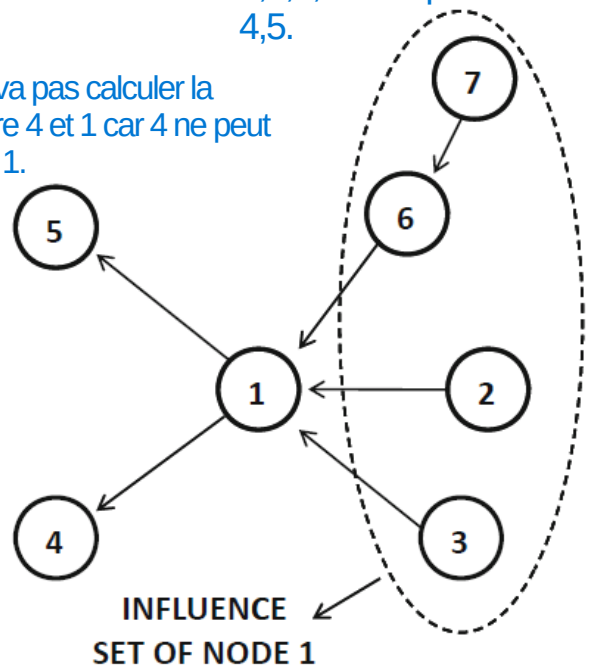
Mesures de centralité et prestige

Écouter minute 27

- Pour le prestige de proximité, il faut considérer que quelques chemins $j \rightsquigarrow i$ sont inexistants.
- Du coup, le premier pas consiste à calculer l'ensemble $Influence(i)$ composé par les sommets qui peuvent atteindre le sommet i.

Noeuds pouvant atteindre le sommet 1 sont les sommets 7,6,2,3 mais pas les sommets 4,5.

Donc, on ne va pas calculer la proximité entre 4 et 1 car 4 ne peut pas atteindre 1.



C. Aggarwal, 2015

La distance = la somme des + courts chemins des sommets j à i uniquement pr les sommets qui peuvent atteindre i
On ne divise pas par n-1 mais par la TAILLE DE L'INFLUENCE

Ex: si on prends le sommet 6
On a 1/1, il n'est pas si prestigieux que ca. Donc pr ca il faut calculer le INFLUENCE FACTOR

Mesures de centralité et prestige

Influence factor(i) = |Influence(i)|/ n-1

- On a maintenant :

Taille de l'ensemble d'influence (donc les sommets se rendant/influencant i) / n-1

On calcule la distance des sommets se rendant au sommet 1 seulement (donc exluant 4 et 5

2 + 1 + 1

$$AvDist(i) = \frac{\sum_{j \in Influence(i)} Dist(j, i)}{|Influence(i)|}$$

Ex: Influence(1) = {2,3,6,7} = 4

- Remarque que $Dist(j, i)$ est calculée du sommet j au sommet i.

- Utiliser l'inverse de la distance moyenne comme dans le cas précédent ne serait pas juste... on ne peut pas faire 1/Avdist

Mesures de centralité et prestige

- On a maintenant :

$$AvDist(i) = \frac{\sum_{j \in Influence(i)} Dist(j, i)}{|Influence(i)|}$$

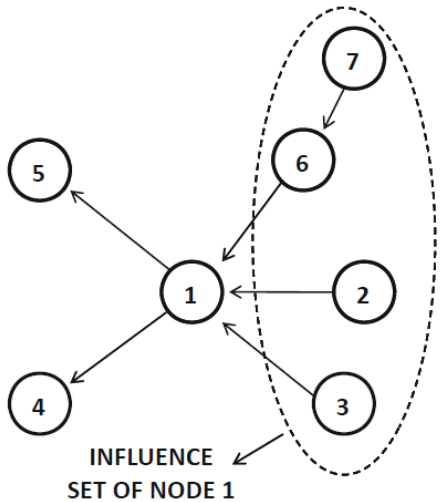
- Remarque que $Dist(j, i)$ est calculée du sommet j au sommet i.

- Utiliser l'inverse de la distance moyenne comme dans le cas précédent ne serait pas juste... Pourquoi ?

on ne tient pas compte de la taille de l'influence

Mesures de centralité et prestige

- Observez le sommet 6 :
  - Son seul sommet influencé est le sommet 7 avec distance de 1.



C. Aggarwal, 2015

- Les sommets avec moins d'influence doivent être pénalisés !

Mesures de centralité et prestige

- Un facteur de pénalité (multiplicatif) est inclus dans la mesure de prestige de proximité.
- Il correspond à la taille relative de l'ensemble d'influence du sommet i :

Noeuds pouvant atteindre sommet i

$$InfluenceFactor(i) = \frac{|Influence(i)|}{n-1}$$

- Le prestige de proximité est finalement donné par :

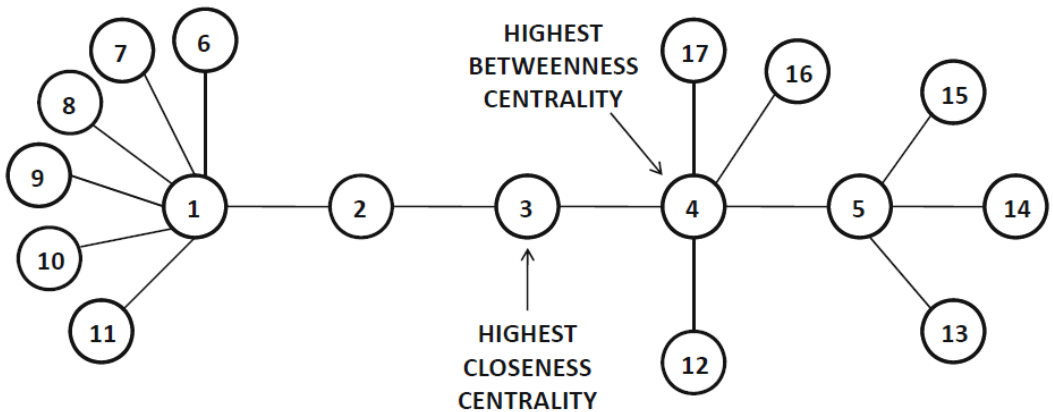
$$P_P(i) = \frac{InfluenceFactor(i)}{AvDist(i)} \in [0,1]$$

AvDist: somme distance noeuds pouvant se rendre à i / nb noeuds se rendant à i

Centralité d'intermédiarité

- Considère la criticité d'un sommet en termes du nombre de chemins les plus courts qui le traversent.

- Crucial pour déterminer les sommets qui contrôlent le plus le flot d'informations entre les autres sommets d'un réseau social.



C. Aggarwal, 2015

Centralité d'intermédiarité

- Soit q_{jk} le nombre de plus courts chemins entre les sommets j et k.

q jk(i) = nb de + courts chemins entre somme i et le sommet k qui contiennent le sommet i

- Soit $q_{jk}(i)$ le nombre de ces chemins qui passent par le sommet i.
- Donc, la fraction des chemins $f_{jk}(i)$ qui passe par le sommet i est donné par $f_{jk}(i) = q_{jk}(i)/q_{jk}$
- Intuitivement, $f_{jk}(i)$ indique le niveau de contrôle que le sommet i a sur les sommets j et k en termes de régulation du flot d'informations entre eux.

EXAMEN: Erreur courante: ne pas considérer les fractions de + court chemin

Centralité d'intermédiarité

- La centralité d'intermédiarité $C_I(i)$ est la valeur moyenne de cette fraction pour toutes les paires de sommets (en excluant le sommet i) :

Somme de tous les éléments f jk dans la matrice supérieure

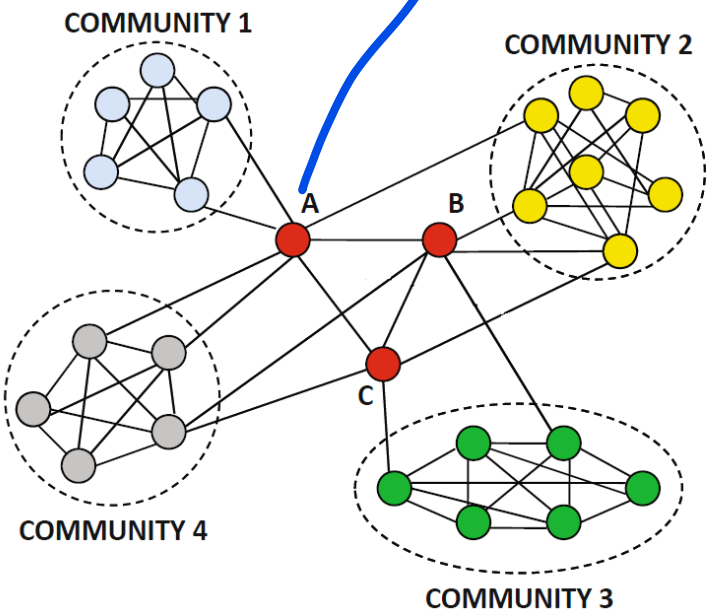
$$C_I(i) = \frac{\sum_{j<k} f_{jk}(i)}{C_{n-1}^2} \in [0, 1]$$

- Contrairement à la centralité de proximité, la centralité d'intermédiarité peut également être définie pour les réseaux déconnectés.
- Peut être généralisé pour les arêtes aussi.

grpahes orientés et non orientés

Exemple

A,B, C ont les + grandes valeurs car controlent



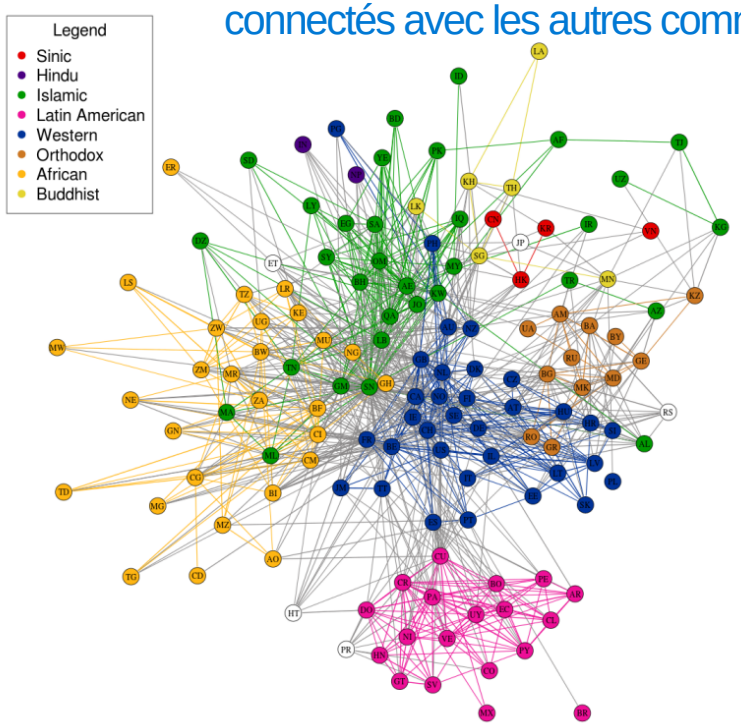
C. Aggarwal, 2015

- Les arêtes qui connectent les sommets rouges (hubs) ont intermédiarité élevée.
- Ces arêtes ont tendance à connecter des sommets appartenant à différents *clusters* du réseau.

Détection de communautés

- *Clustering* pour les réseaux sociaux.
- Pour chaque *cluster*, ses éléments ont plusieurs liens entre eux, et très peu de liens pour le reste des éléments du réseau social.

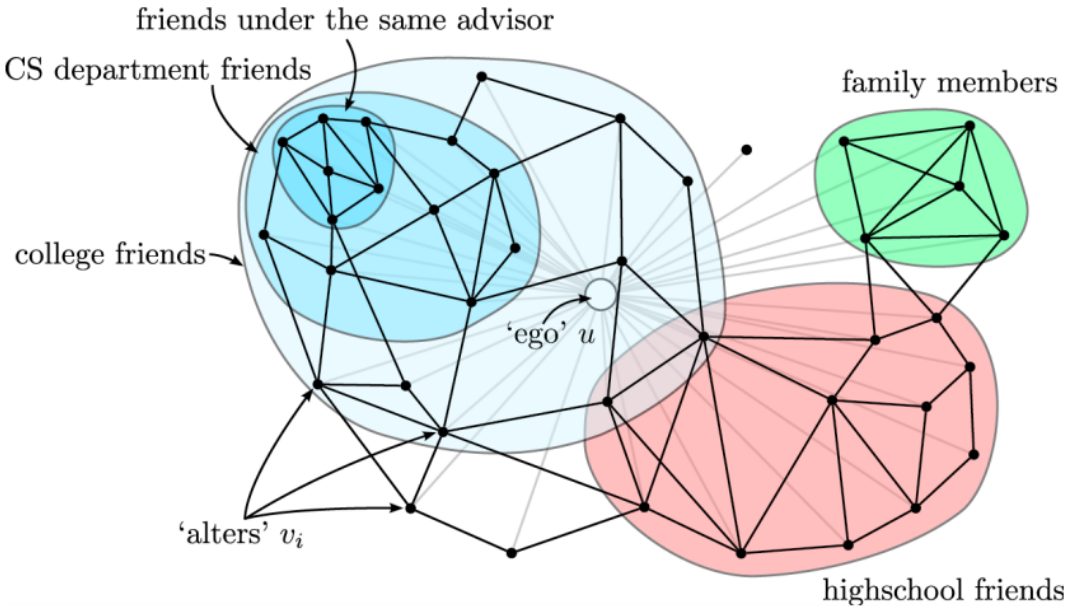
Communautés sont formetement connectés entre elles et peu connectés avec les autres communautés



Source : "The mesh of civilizations and international email flows", B. State et

Détection de communautés

- *Clustering* pour les réseaux sociaux.
- Pour chaque *cluster*, ses éléments ont plusieurs liens entre eux, et très peu de liens pour le reste des éléments du réseau social.



McAuley, Leskovec : Discovering social circles in ego networks, 2012

Algorithme de Girvan-Newman

- Basé sur l'intuition que les arêtes ayant une grande centralité d'intermédiarité C_I ont tendance à connecter les différents *clusters*.

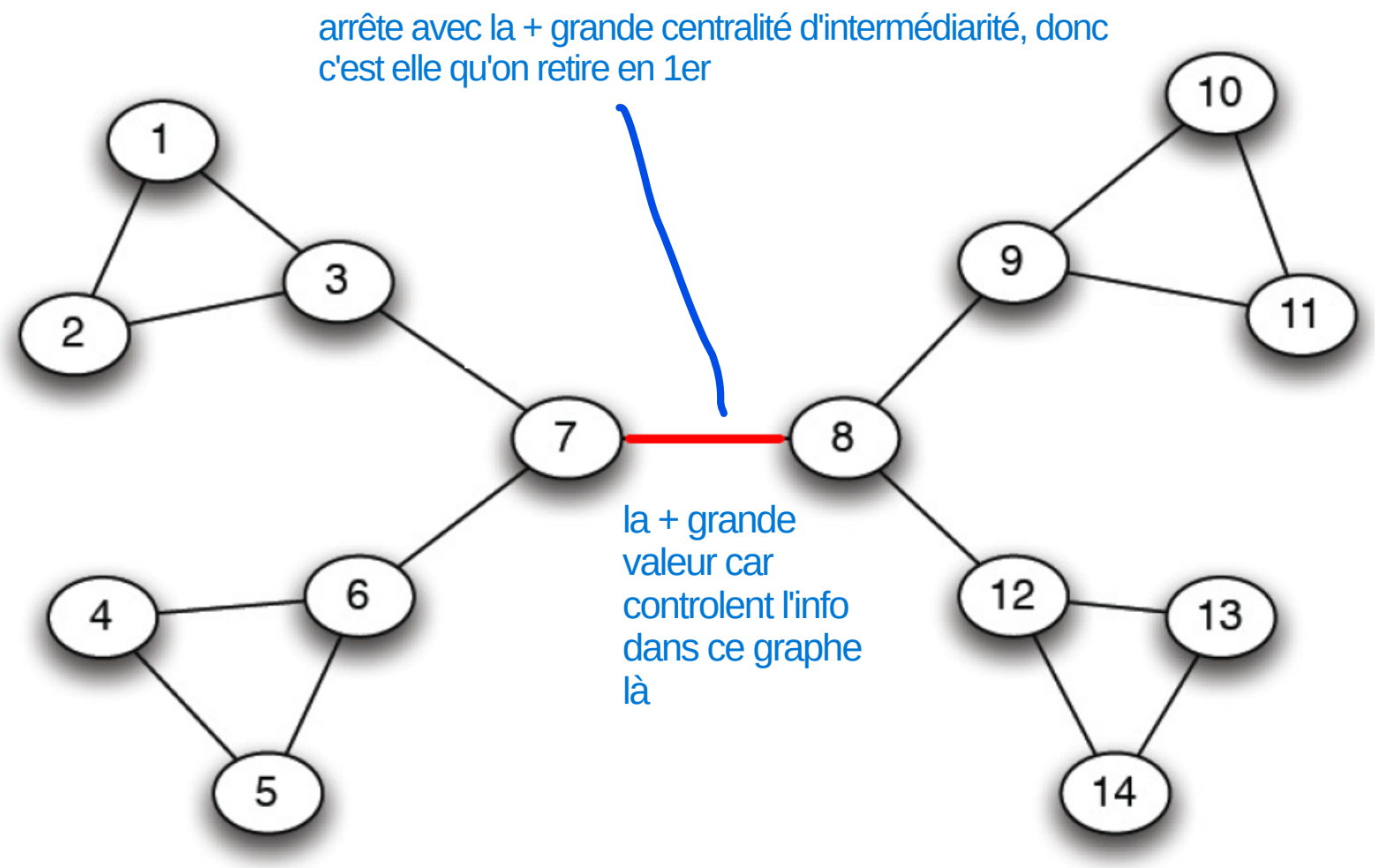
Algorithme de Girvan-Newman

Tant qu'il reste des arêtes :

1. Calculez $C_I(a)$ pour tout $a \in A$. Centralité d'intermédiarité pour toutes les arrêtes
2. Enlevez les arêtes avec les plus grandes valeurs de C_I.

- Algorithme de regroupement hiérarchique descendant (division).
- Les composants connectés sont les communautés.

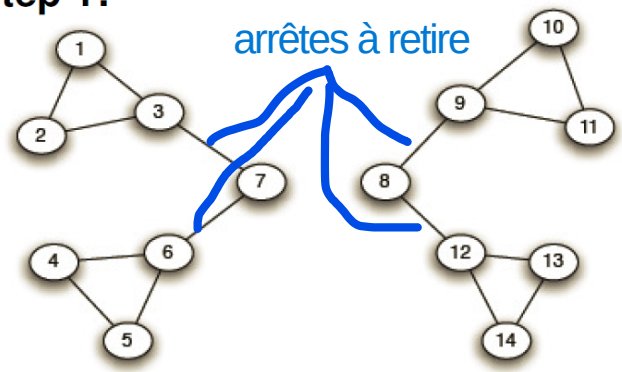
Exemple



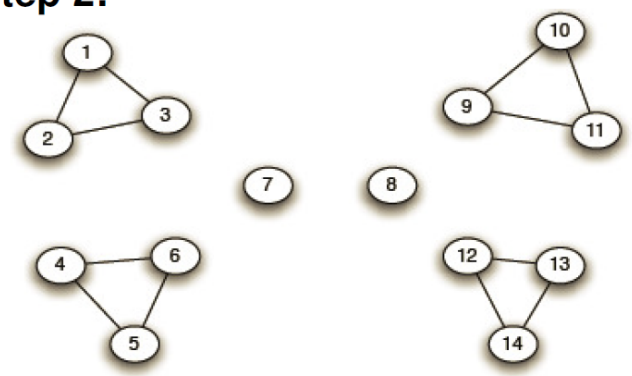
Source : http://www.mmds.org/

Exemple

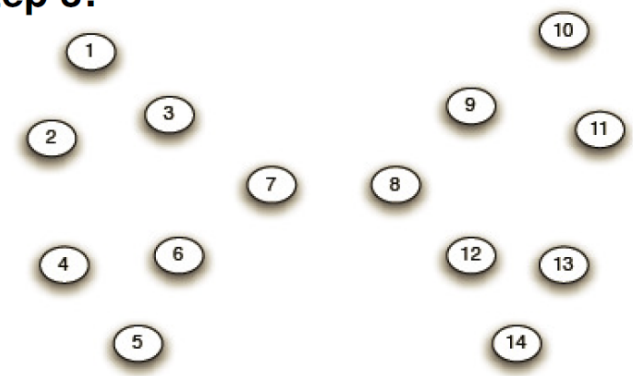
Step 1:



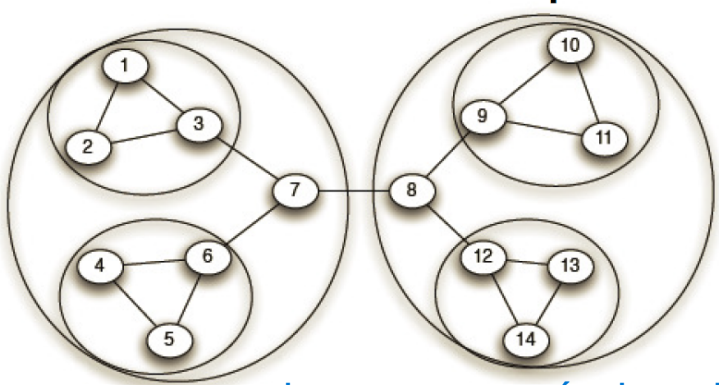
Step 2:



Step 3:



Hierarchical network decomposition:

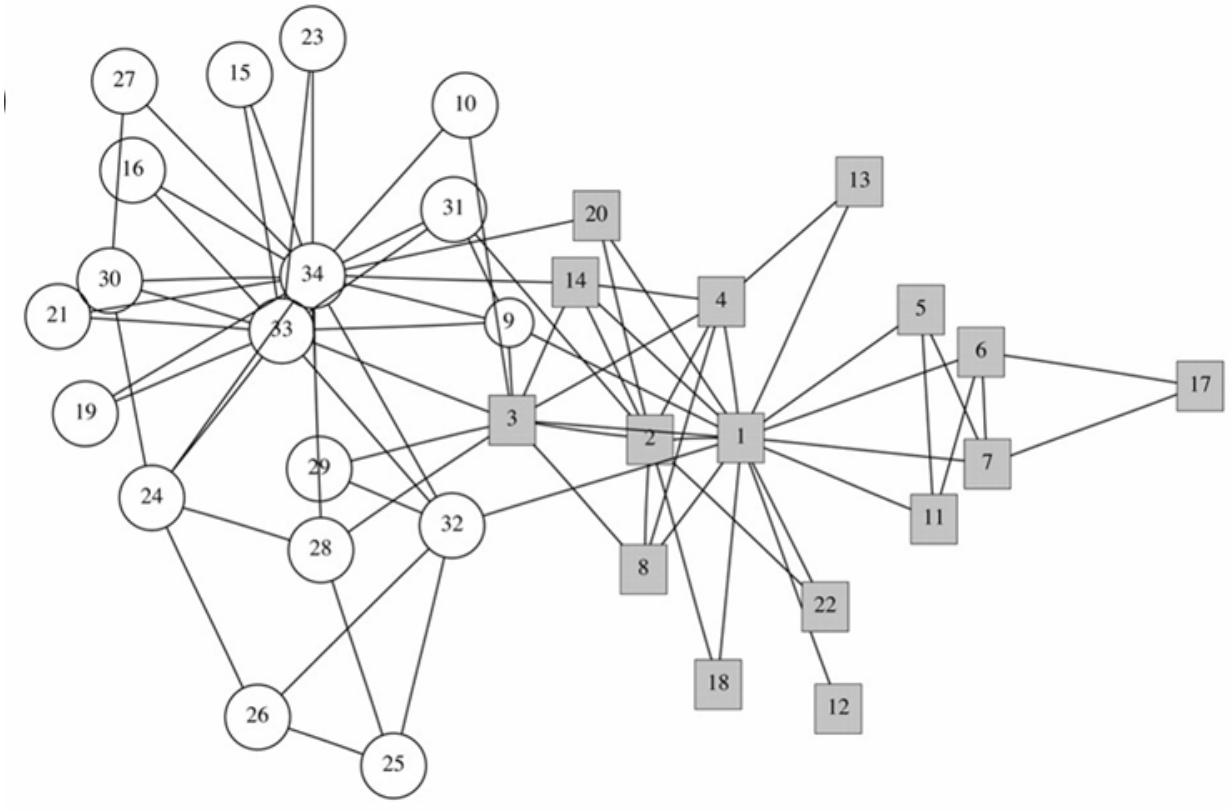


des communautés dans des communautés

Source : `http://www.mmds.org/`

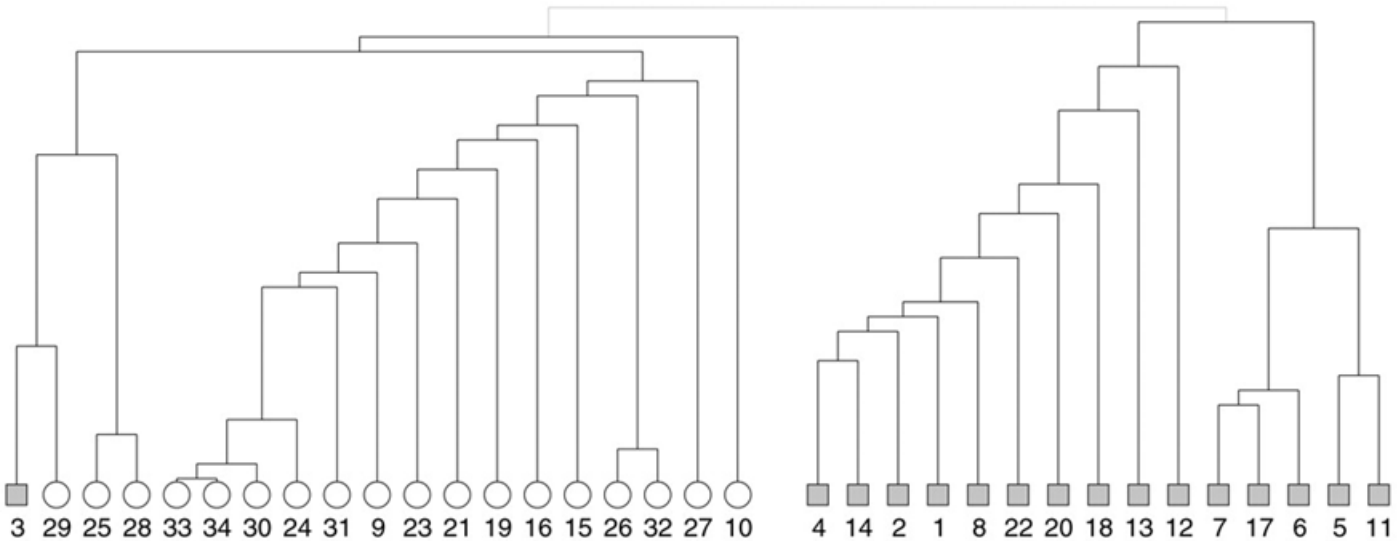
Le club de karaté de Zachary

Réseau social d'un club de karaté étudié par Wayne W. Zachary pendant trois ans de 1970 à 1972.



Le club de karaté de Zachary

- L'algorithme de Girvan-Newman ne commet qu'une seule erreur par rapport au *ground-truth* :



Détection de communautés

- Il nous reste quand même deux défis :

  calculer la Centralité d'interméd. pour les sommets et pour les arrêtes

  - le calcul de la centralité d'intermediarité pour chaque itération de l'algorithme de Girvan-Newman (article sur Moodle).
  - la définition du bon nombre de *clusters*.

Modularité

va nous dire à quel point la partition du graphe est bonne, à quel point quand on sépare les articles fortement connectés

- Critère de *clustering* pour les réseaux.
- Étant donné une partition du réseau en K *clusters*, la modularité Q est directement proportionnelle à :

partitionner le graphe en K

$$\sum_{k=1}^{K} A_k - E_{A_k}$$

où :

- A_k est le nombre d'arêtes présentes dans le *cluster* k
- E_{A_k} la valeur attendue de A_k dans un graphe aléatoire.

nb arretes attendues dans un graphe aleatoire (modele nul)

- E_{A_k} demande la construction d'un modèle nul.

Si on a un graphe avec le nombre d'arretes et le nombre de sommets, et la répartition des sommets,, cmb on s'attends à trouver d'arrtes si notre graphe est aléatoire et on va comparer ca avec cmb d'arretes on trouve réellement. Si on trouve + d'arretes qu'attendu, c'est qu'on a réussi à isoler la communauté, sinon il n'y a pas de reel groupe

Modèle nul

G: Graphe
N: ensemble des ensemble des sommets
A: ensemble des arretes

- Étant $G = (N, A)$, on construit un nouveau réseau aléatoire G' préservant la distribution des dégrées des sommets de G.

la répartition des arrêtes dans le graphe

- Dans ce modèle, le nombre attendu d'arêtes entre les sommets i et j est égal à :

$$\text{nb attendu d'arretes} = \frac{degree(i) \cdot degree(j)}{2|A|}$$

A: nb d'arretes totales dans notre graphe

EXAM:
Démonstration
à noter dans
feuille notes

Modèle nul

- Remarque : le nombre attendu d'arêtes attendues pour tout le graphe G' est :

1/2 car on compte les arretes 2 fois

ex: on va compter l'arrete de 0 vers 1 et ensuite de 1 vers 0

nb arretes attendus pour G' =

$$= \frac{1}{2} \sum_{i \in N} \sum_{j \in N} \frac{degree(i) \cdot degree(j)}{2|A|}$$

$$= \frac{1}{2} \cdot \frac{1}{2|A|} \sum_{i \in N} degree(i) \left(\sum_{j \in N} degree(j) \right)$$

$$= \frac{1}{4|A|} \cdot 2|A| \cdot 2|A|$$

$$= |A|$$

Modularité

- Modularité d'une partition $P = \{C_1, \ldots, C_K\}$ des sommets d'un graphe G :

$$Q(G,P) = \frac{1}{2|A|} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j \in C_k} \left(\delta_{ij} - \frac{degree(i)degree(j)}{2|A|} \right)$$

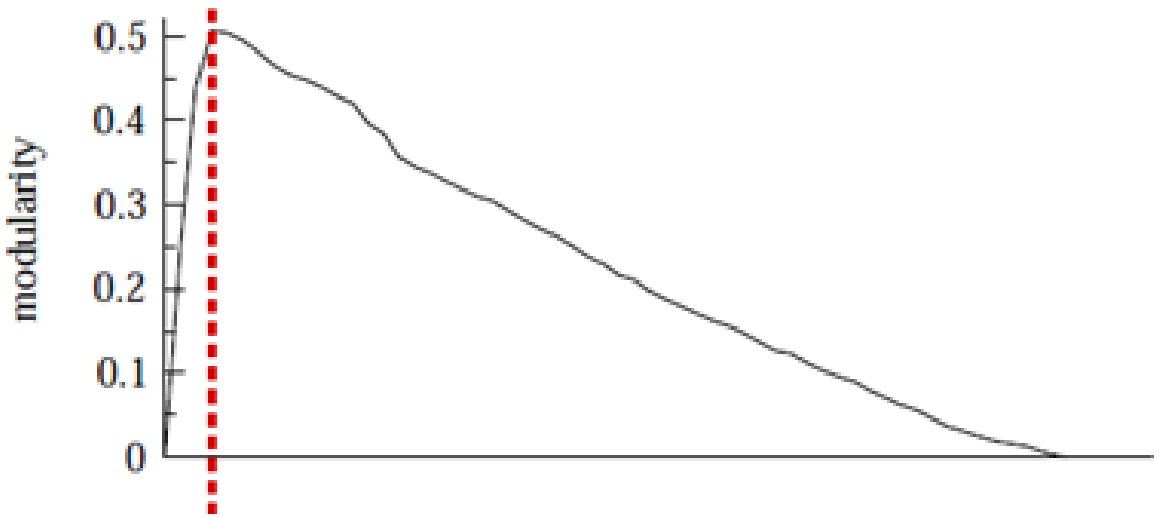
A: nb arretes

où $\delta_{ij} = 1$ si les sommets i et j sont liés, 0 sinon.

- $Q \in [-1, 1]$ modularité est comprise entre -1 et 1
- $Q > 0$ si le nombre d'arêtes dans les *clusters* dépasse le nombre attendu groupe où on a + de connexions que si on avait un graphe aléatoire
- En général, $Q > 0.7$ indique une structure communautaire significative.

Modularité

- On peut utiliser la modularité pour sélectionner le nombre de *clusters* dans l'algorithme de Girvan-Newman.
- On calcule la modularité pour chaque partition de la hiérarchie.
- On garde celle avec la plus grande valeur de Q.



Source : `http://www.mmds.org/`

Modularité

- Pourquoi ne pas optimiser la modularité directement ?
- Sujet du 10e défi DIMACS en 2012 :
  - Des graphes ayant jusqu'à 10 millions de noeuds et 200 millions d'arêtes

Classification collective

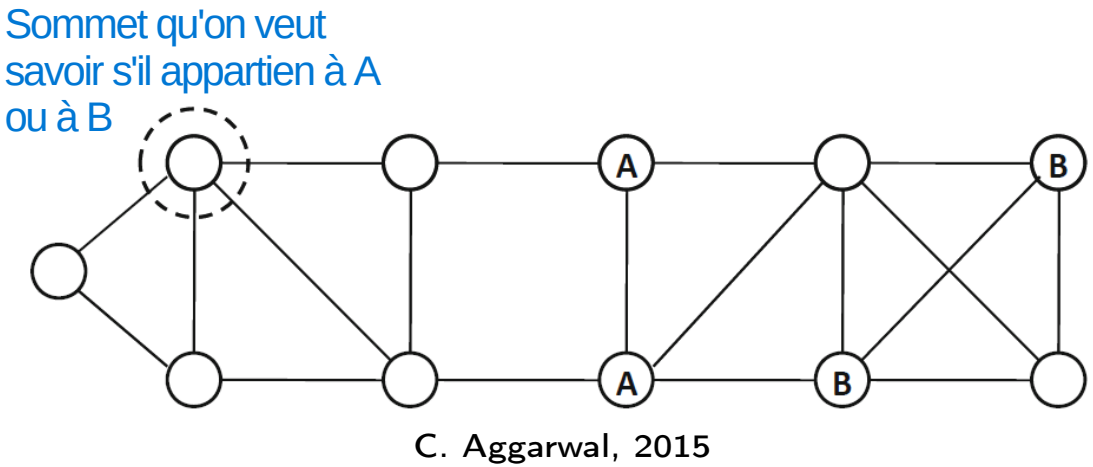
- Dans plusieurs applications, des étiquettes peuvent être associées à des sommets d'un réseau social.
- Étiquettes disponibles $\Rightarrow$ classer les étiquettes inconnues.
- Ex. réseautage social : déterminer les personnes susceptibles de voter pour un certain candidat.

Classification collective

Étiquettes pour les sommets

- Les sommets ayant des propriétés similaires sont généralement connectés. Ceux qui se ressemblent ont tendance à se rassembler. Ex: sur fb ceux qui aiment Trump
- Il est raisonnable de supposer que ceci est également vrai pour les étiquettes. Les sommets avec des propriétés similaires ont tendance à se conncecter
- Solution simple : examiner l'étiquette majoritaire trouvée parmi les voisins d'un sommet (similaire à k-NN).
  prendre relations FB (mes amis) et regarder pour qui ils votent
- Parfois impossible pour la classification collective en raison de la rareté des étiquettes.

Exemple



C. Aggarwal, 2015

Il n'y a pas des sommets étiquetés directement liés au sommet qu'on veut classer

Algorithme de classification itérative

Iterative Classification Algorithm (**ICA**)

- *Input* : Graphe $G(N, A)$, étiquettes pour $N' \subset N$, algorithme de classification $\mathcal{A}$, et nombre d'itérations T
- Tant que tous les sommets ne sont pas étiquetés :
  - Entraîner le classificateur $\mathcal{A}$ en utilisant les attributs de liens et les attributs de contenu des sommets d'entraînement.
  - Prédire les étiquettes des sommets de test.
  - Fixer les étiquettes des sommets les plus "*certains*", et ajouter ces sommets aux données d'entraînement, tout en les retirant des données de test ;

connexions les + fortes

Algorithme de classification itérative

- Le nombre total de itérations dépend du nombre d'étiquettes fixées par itération. Fixer un seuil: je garde l'étiquette slm si la prédiction > une valeur
- Les attributs de contenu sont ceux relatifs aux caractéristiques du sommet, ex. $X_i = \{43\text{ans}, 74\text{kg}, 80000\$\}$. (age, poids, salaire annuel)
- Les attributs de liens sont structurels et mis à jour à chaque itération :
  - ⋆ distribution de classes dans le voisinage d'un sommet
  - ⋆ différentes mesures de centralité, etc.
- L'algorithme $\mathcal{A}$ doit fournir la probabilité qu'un sommet appartienne à une classe.
- Considéré comme un algorithme de classification semi-supervisée.

Marches aléatoires

MÉTHODE IMPORTANTE!!!!!!!

Pour classifier un sommet non étiqueté i, dans notre graphe, suivre les liens qui existent et continuer jusqu'à trouver l'étiquette. On se promène dans le graphe de manière aléatoire jusqu'à ce qu'on trouve une étiquette. Quand on trouve une étiquette, je retourne au sommet et je fais à nouveau une marche aléatoire et j'essaie de trouver une autres étiquette. Tant que je n'ai pas trouvé d'étiquette, jecontinue la marche aléatoire

- Pour classifier un sommet non étiqueté i, des marches aléatoires peuvent être exécutées à partir du sommet i.
- Une marche se termine lorsqu'on trouve le premier sommet étiqueté. suivre les étiquettes
- La classe dont la probabilité de terminaison de la marche aléatoire est la plus élevée est indiquée comme classe prédite du sommet i.
- L'idée est que la marche a plus de chance d'être terminée dans un des sommets étiquetés aux alentours du sommet i.

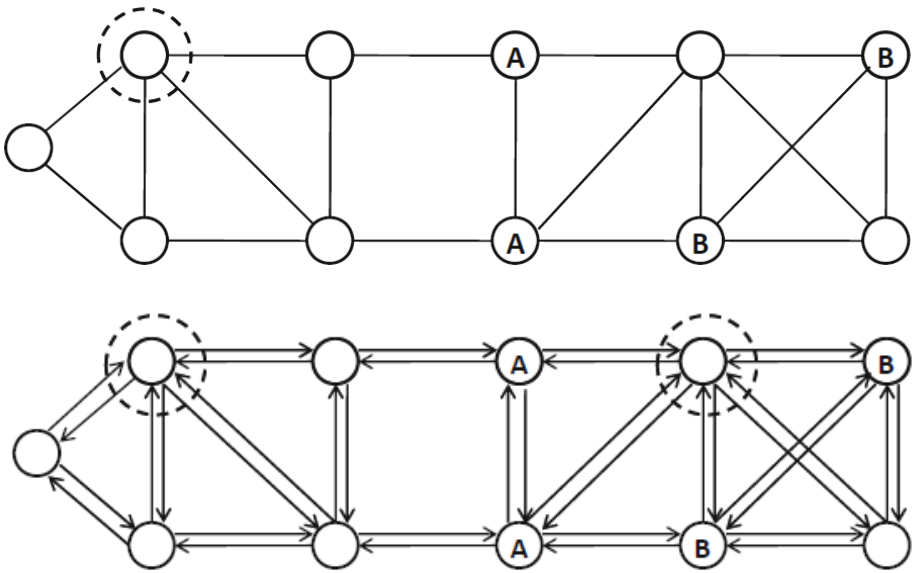
Marches aléatoires

Hypothèse Le graphe est *label connected* : il existe un chemin entre tous sommets non étiquetés et un sommet étiqueté.

Condition Une marche aléatoire se termine toujours aux premiers sommets étiquetés atteints .

Marches aléatoires

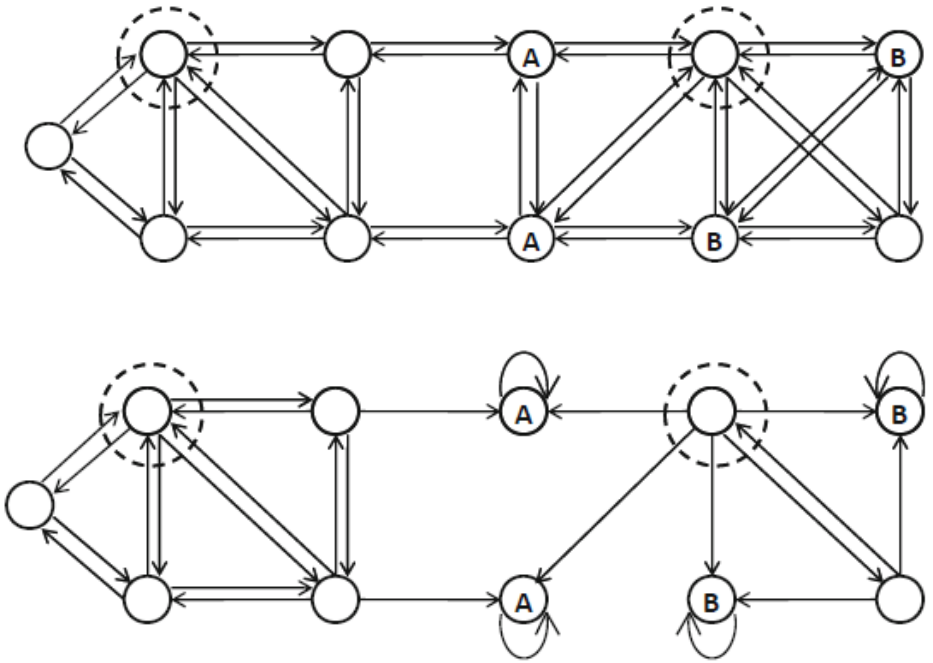
- La dernière condition sera garantie en deux étapes
- Étape 1 : Conversion d'un graphe non dirigé en graphe dirigé



C. Aggarwal, 2015

Marches aléatoires

- Étape 2 : Remplacement des arcs sortants des sommets étiquetés par des *self-loops*



C. Aggarwal, 2015

Marches aléatoires

- Soit M la matrice $(n \times n)$ de transition du graphe après l'étape 2.
- Rappel : M_{ij} est la probabilité de que le prochain sommet parcouru après le sommet j soit le sommet i.
  - Alors $M_{ij} = 1/out\_degree(j)$ (arc sortant), s'il existe l'arc $j \rightarrow i$, $M_{ij} = 0$ sinon.
  - Pour un sommet j tel que j possède un *self-loop*, $M_{jj} = 1$ et $M_{ij} = 0$. (tout les reste = 0)

Marche aléatoire

- Si le point du départ est fixé, il n'y a qu'une distribution de probabilités π qui découle des itérations successives de :

proba de me trouver dans chacun des noeuds apres 1 step aleatoire à partir du sommet de départ

M: matrice de transition
pi: position où je me trouve

$$\pi^k = M \cdot \pi^{k-1}$$

processus markovien

à partir de π^0 avec $(\pi^0)_i = 1$ dans le cas où le sommet de départ est i, 0 sinon

- La classe du sommet i est alors donnée par :

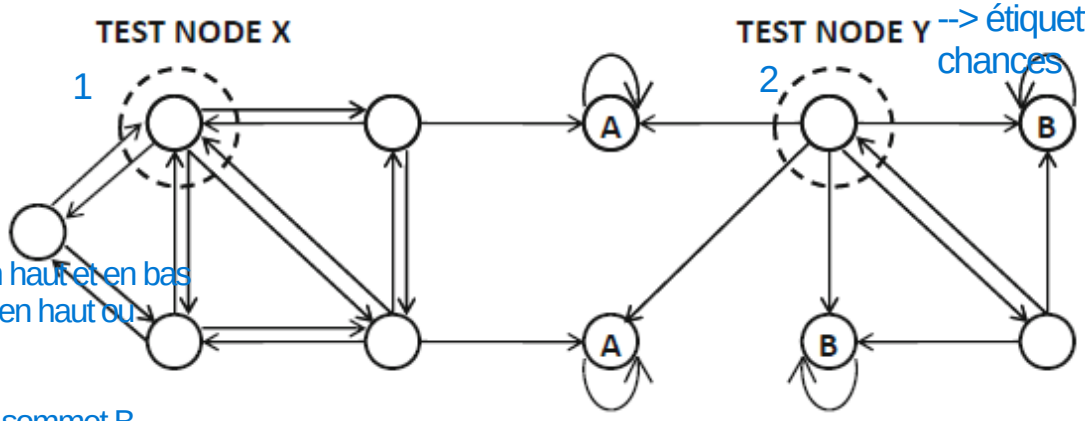
classe avec la proba la + élevée

$$\underset{k}{\text{argmax}} \left\{ \sum_{j:j \in class\ k} \pi_j \right\}$$

Marche aléatoire

J'ai 2 probabilités si on suit la marche aléatoire pr le sommet 1:
- Soit on tombe sur le 1er sommet A en haut, soit sur le sommet A en bas. La proba ici de tomber sur un sommet A est de 1 et proba pour B est 0

Voir explication 1h25



--> étiquette associé est B, car + de chances

Sommet 2:
- Proba pour arriver sommet A en haut et en bas et proba pr soit arriver sommet B en haut ou sommet B en bas

Plus de chances pr arriver sur un sommet B car 1 chance pr aller vers B en haut, 1 chance pr aller vers B en bas et une chance (diagonale vers noeud vide) pr arriver au point de départ et recommencer

C. Aggarwal, 2015

- Une marche aléatoire débutant au sommet X va toujours finir dans un sommet de la classe A.
- Une marche aléatoire débutant au sommet Y peut finir soit dans un sommet de la classe A ou de la classe B.