chapitre 1

Propriétés structurelles des GRI

Section 4

Modularité

et en plus c'est le TP4

Fabien de Montgolfier
fm@irif.fr

25 février 2022

Propriétés communes des réseaux

En plus de la **grande taille** on a (presque toujours) :

1. Distribution des **degrés** : peu de riches beaucoup de pauvres
2. Distribution des **distances** : tout le monde proche de tout le monde (effet *small world*)
3. Existence d'un **cœur** : les riches se connaissent, pas les pauvres
4. Composante connexe **géante**
5. **Transitivité** forte (= *coefficient de clustering*) : les amis de mes amis sont mes amis
6. **Navigabilité** (pas toujours) : on peut atteindre une cible de proche en proche. Ex : routage IP.
7. existence de **communautés** (pas toujours non plus) : sous-graphes denses ayant un sens.

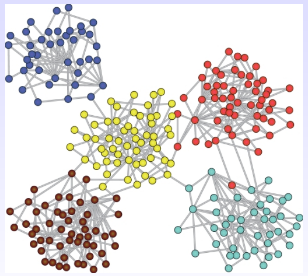
Clustering : intérêt

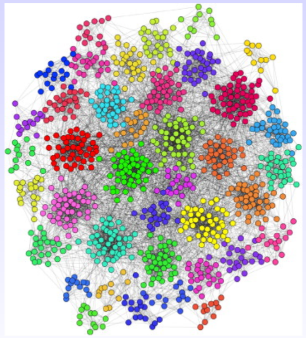
1. Une étude de cas sur un graphe donné (sociologie, biologie,... *ologie)
2. On le découpe (partition) en *clusters*
3. Ces clusters disent des choses sur le graphe et permettent d'appréhender sa compexité *ologique

Questions

- Qu'est-ce qu'un bon clustering ?
- Comment en calculer un ?

On voudrait ça :





Calcul de communautés

Problème (à formulation floue) du calcul des communautés

Trouver une **partition** (ou un recouvrement) des sommets en c sous-graphes (c fourni à l'avance ou pas) tels leur densité (Moyenne ? Minimale ?) soit maximale

unicité très peu garantie ! Dépend fortement de la définition et de l'algorithme qui l'implémente

Définitions

Soit $G = (V, E)$ un graphe non orienté

Partition

$P = V_1..V_k$ tel que chaque sommet appartient à un et un seul ensemble V_i, aucun V_i n'est vide.

Cluster

L'un des V_i

Clustering

Algorithme (et pas coefficient, ici !)

Modularité

Informellement

La proportion des *liens internes* de la partition moins la proportion de liens internes de la même partition mais sur le graphe rebranché aléatoirement.

Donc un calcul en deux parties

- Pourcentage d'arêtes internes
- Soustraction du *null model*

Notations

- $m(i, j)$ le nombre d'arêtes entre le cluster V_i et le cluster V_j
- En particulier $m(i, i)$ est le nombre d'arêtes **internes** au cluster V_i
- On a $\sum_{i,j} m(i, j) = m$.
- $e_{ij} = \frac{m(i,j)}{m}$ proportion d'arêtes entre V_i et V_j

Null Model

Si on recombinait les arêtes du graphe au hasard en respectant les degrés (ce que l'on appelle un *null model*), alors la probabilité qu'il y ait une arête du sommet u au sommet v serait

$$\frac{deg(u)}{2m} \times \frac{deg(v)}{2m}$$

On définit a_{ij} la proportion d'arêtes entre les clusters V_i et V_j dans le *null model* :

$$a_{ij} = \sum_{u \in V_i, v \in V_j} \frac{deg(u) \times deg(v)}{4m^2}$$

On peut remarquer que si $i = j$:

$$a_{ii} = \frac{\left(\sum_{v \in V_i} deg(v)\right)^2}{4m^2}$$

Modularité de Newman

Modularité de la partition P des sommets de G

$$Q(P) = \sum_{i=1}^{i=k} e_{ii} - a_{ii}$$

et en développant

$$\boxed{Q(P) = \sum_{i=1}^{i=k} \left(\frac{m(i,i)}{m} - \frac{\left(\sum_{u \in V_i} deg(u)\right)^2}{4m^2} \right)} \quad (1)$$

Valeurs de Q ?

- Terme de gauche : densité d'arêtes internes $\in [0,1]$
- Terme de droite : densité d'arêtes internes après shuffle $\in [0,1]$
- Donc $Q \in [-1,1]$
- Interprétation :
  - Négatif : mauvais clustering
  - Nul : non signifiant
  - Positif : bon clustering
  - $1 =$ clustering parfait ?
- En fait $Q \in [-0.5, 1[$
- En particulier $Q \leq 1 - \frac{DegréMax^2}{4m^2}$

Modularité d'un graphe

Définition

$$Q(G) = \max_P \{Q(P)\}$$

Tout graphe a modularité ≥ 0

Démonstration : c'est la qualité d'un clustering en un seul cluster

Théorème [Brandes & al 2008]

Calculer $Q(G)$ est NP-complet

Il va donc falloir un algorithme d'approximation, ce qui est appelé une **heuristique**

Remarques diverses

Bornes

- Le clustering trivial $\{V\}$ a modularité 0
- Un clustering en k clusters a modularité $\in [-0.5, 1 - \frac{1}{k}]$
- Un graphe à n sommets a donc modularité $\in [0, 1 - \frac{1}{n}]$

Classes ayant la pire modularité : 0

- Étoiles $K_{1,n}$
- Graphes complet K_n

Conclusion sur la modularité

Il y a deux défintions de la modularité

- Modularité d'une partition (facile à calculer)
- Modularité d'un graphe (difficile à calculer)

Il y a donc deux usages de la modularité

- Comme objectif à maximiser pour un algorithme de clustering
- Comme paramètre de graphe dans un discours *ologique

L'algorithme décrémental

Auteurs

équipes de Vincent Blondel à Louvain-la-neuve et de Matthieu Latapy à Paris

Principe

Cet algorithme est *décrémental* ou encore *aggrégatif*

- on part d'une partition triviale P^1 où chaque cluster est fait d'un unique sommet (il y a donc $k = n$ clusters).
- À chaque étape t (t de 1 à n) on **fusionne deux clusters**.
- On arrive donc à P_n avec un seul cluster
- $Q(G)$ est la meilleure modularité de toutes ces partitions rencontrées

L'algorithme est glouton

Quels clusters et prendre pour les fusionner en un cluster, à chaque étape ?

Ceux, parmis les $O(k^2)$ choix possibles, **les deux qui vont faire le plus augmenter la modularité**.

Notez que la modularité peut diminuer d'une étape à l'autre (quand aucune fusion n'améliore la modularité)

En effet, la modularité initiale est légèrement négative :

$$Q(P^1) = -\frac{\sum_v deg(v)^2}{4m^2} \quad (2)$$

tandis que la modularité finale est $Q(P^n) = 0$. Entre les deux on espère s'être approché de 1.

Complexité

Variantes

Il y a d'autres façons de faire, telles que la "*Méthode de Louvain*". Attention, la page Wikipedia éponyme contient plusieurs erreurs, par exemple la complexité ne peut pas être baissée jusqu'à $O(n \log n)$ (ce qui est plus petit que la taille du graphe dans certains cas ! !).

Complexité naïve

Pour t de 1 à n

- Regarder les $(n+1-t)^2$ paires de cluster
- Choisir les deux qui augementent le plus la modularité
- Les fusionner

Calcul de $Q()$ en $O(m)$ donc en tout $O(n^3 m)$

Améliorons la complexité

$$\boxed{Q(P) = \sum_{i=1}^{i=k} \left(\frac{m(i,i)}{m} - \frac{\left(\sum_{u \in V_i} deg(u)\right)^2}{4m^2} \right)}$$

Amélioration 1 : la somme des degrés

Chaque cluster connaît sa somme des degrés (attribut d'un cluster). Fusionner deux clusters additionne seulement ces sommes

Amélioration 2 : le nombre d'arêtes internes

$m(i,i)$ peut être

- Calculé à la volée en $O(m)$
- Dans une matrice
- Dans une `HashMap`

Incrément de modularité (amélioration 3)

Quand on fusionne les clusters V_a et V_b en V_c :

$$Q(P^{t+1}) - Q(P^t) = (e_{cc} - a_{cc}) - (e_{aa} - a_{aa}) - (e_{bb} - a_{bb})$$

- Ne dépend que de deux clusters
- $e_{cc} - (e_{aa} + e_{bb}) = \frac{m(a,b)}{m}$
- Les a_{ii} dépendent de $somDeg(i) = \sum_{x \in V_i} deg(x)$ stocké
- donc au final calcul en temps constant : $Q(P^{t+1}) - Q(P^t) =$

$$\frac{m(a,b)}{m} - \frac{(somDeg(a) + somDeg(b))^2}{4m^2} + \frac{(somDeg(a))^2}{4m^2} + \frac{(somDeg(b))^2}{4m^2}$$

Usage d'un tas (amélioration 4)

- Enfin au lieu de regarder les k^2 paires à chaque étape on a un **tas de paires** de clusters (file de priorité)
- Permet d'avoir la meilleure paire en temps constant
- chaque cluster V_c crée à l'étape k appartient à $k-1$ nouvelles paires
- inutile de mettre ces $k-1$ paires dans le tas : seule la **meilleure paire** (V_c, V_d) pour un certain V_d peut être mise dans le tas

Publication de [Newman et Givran 2004]

Values approaching $Q = 1$, which is the maximum, indicate strong community structure. In practice, values for such networks typically fall in the range from about 0.3 to 0.7. Higher values are rare.

Donc il est largement admis par les *ologues que plus la modularité est élevée, plus le graphe a intrinsèquement des clusters, cachés, qu'il n'y a plus qu'à « retrouver »par un algorithme de clustering

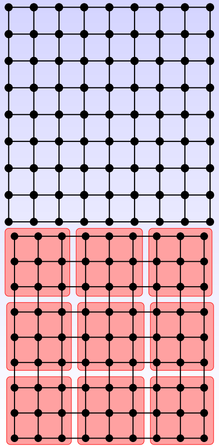
Mais voyons...

Le tore

- Prenons un tore carré de $n = a \times a$ sommets
- Coupons-le en $k = b^2$ clusters carrés
- Terme de gauche $\sum_{i=1}^{i=k} \left(\frac{m(i,i)}{m}\right) = \frac{m - 2ab}{m} = 1 - \frac{b}{a}$
- Somme des degrés d'un cluster $= 4 \times$ sa surface. Terme de droite $=$

$$\frac{\left(\sum_{u \in V_i} deg(u)\right)^2}{4m^2}) = k\frac{(4(a/b)^2)^2}{(4a^2)^2} = \frac{1}{b^2}$$

- Donc $Q(P_{carré}) = 1 - \frac{b}{a} - \frac{1}{b^2}$
- prenons $b = \sqrt[6]{n}$. On a alors $Q(P_{carré}) = 1 - \frac{2}{\sqrt[3]{n}}$



Et en fait [Montgolfier, Soto, Viennot 2011]

- Les graphes de degrés en power law de paramètre $\gamma > 5$ et de degré moyen d ont une modularité minimale de $2/d$
- Des classes **régulières** ont modularité asymptotiquement 1 :
  - Grilles et (hyper)tores,
  - hypercube
  - Arbres de degré $o(n)$
- Cela n'était pas l'intuition de Newman !
- Donc : une modularité élevée d'indique donc pas une *« strong community structure »ni la présence de clusters "naturels".*
- La modularité doit être prise comme un objectif à maximiser, pas comme un paramètre de graphe.