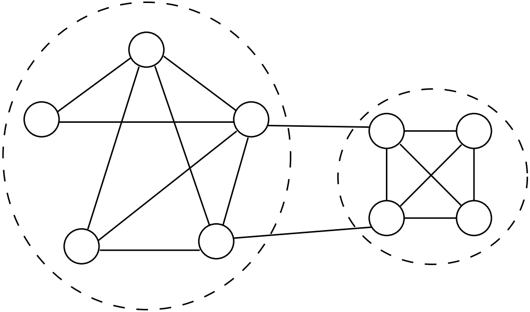
Resolution limit in community detection

Santo Fortunato et Marc Barthélemy

2006

Introduction



Introduction

- Point de départ : un graphe et des sous-graphes.
- But : quantifier le fait que les sous-graphes choisis sont des modules.
- Modularité : compare le nombre d'arcs dans les sous-graphes avec le nombre de liens d'un graphe aléatoire de même taille, avec même répartition des degrés.

Introduction

- Point de départ : un graphe et des sous-graphes.
- But : quantifier le fait que les sous-graphes choisis sont des modules.
- Modularité : compare le nombre d'arcs dans les sous-graphes avec le nombre de liens d'un graphe aléatoire de même taille, avec même répartition des degrés.
- Problème : trouver tous les modules d'un graphe $\stackrel{?}{\Longleftrightarrow}$ maximiser la modularité ?

Résultats

- Démonstration formelle de la non équivalence.
- Limites de la modularité.
- Application à des graphes de terrain.

Plan

1. Préliminaires
   - Module
   - Modularité
2. Résolution limite
   - Problème
   - Méthode
3. Exemples
   - Graphe générique
   - Graphes de terrain

Première partie I

Préliminaires

Définition du module

Un sous-graphe $\mathcal{S}$ est un module si :

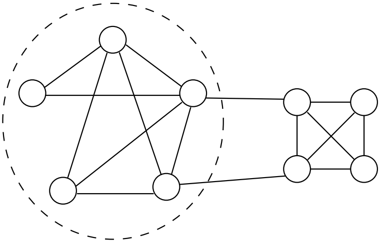
nombre de liens dans S

degre total des noeuds dans S

$$\frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2 > 0$$

nombre de liens dans le reseau

Exemple



$$l_s = 8,\ \ d_s = 18,\ \ L = 16$$

$$\frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2 = \frac{47}{256} > 0$$

Autre expression

On pose $l_s^{out} = al_s,\ a \geqslant 0$. On obtient alors la condition suivante :

$$l_s < \frac{4L}{(a+2)^2}$$

Cette condition est satisfaite si $l_s < L$ et $a < 2$.

Remarque : cela dépend de L, donc de la taille du réseau.

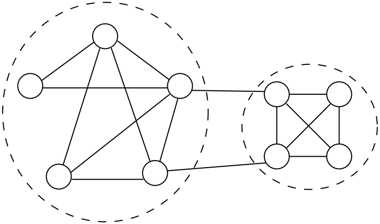
Définition

Rappel : $\mathcal{S}$ module si $\dfrac{l_s}{L} - \left(\dfrac{d_s}{2L}\right)^2 > 0.$

- Point de départ : graphe G et m sous-graphes de G qui forment une partition
- Modularité associée :

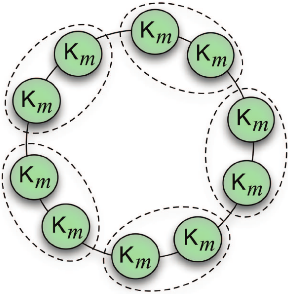
$$Q = \sum_{s=1}^{m} \left(\frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2 \right)$$

Exemple



$$Q = \frac{47}{128}$$

Problème posé



10 cliques de taille 3 ($m = 3$) :

- $Q_{isolées} = 0,650$
- $Q_{appariées} = 0,675$

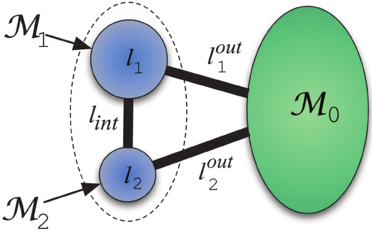
Deuxième partie II

Résolution limite

Rappels

- On a vu sur un contre-exemple que trouver tous les modules d'un graphe $\not\Leftrightarrow$ maximiser la modularité.
- Est-ce généralisable ?

Cas plus général (1/2)



- $l_{int} = a_1 l_1 = a_2 l_2$
- $l_1^{out} = b_1 l_1, \; l_2^{out} = b_2 l_2$
- Cadre de travail :
  $a_1 + b_1 < 2, \; a_2 + b_2 < 2, \; l_1 < L/4, \; l_2 < L/4$

Cas plus général (2/2)

Deux partitions :

- A : $\mathcal{M}_1$ et $\mathcal{M}_2$ sont deux modules séparés ;
- B : $\mathcal{M}_1$ et $\mathcal{M}_2$ forment un seul module.

$$\Delta Q = Q_B - Q_A$$

$$\Delta Q < 0 \;\Leftrightarrow l_2 \; > \frac{2La_1}{(a_1 + b_1 + 2)(a_2 + b_2 + 2)}$$

Rappel : $a_1 + b_1 < 2,\; a_2 + b_2 < 2,\; l_1 < L/4,\; l_2 < L/4$

Résultats

- Dans de nombreux cas, $\Delta Q > 0$.
- Risque de ne pas repérer la structure à petite échelle.
- Maximiser la modularité ne permet pas d'obtenir toute la structure en une seule fois.

Amélioration de l'algorithme

Principe : utiliser l'algorithme d'optimisation de la modularité *récursivement*.

1. appliquer l'algorithme d'optimisation de la modularité ;
2. pour chaque sous-graphe obtenu, regarder si c'est un module ;
3. si ce n'est pas le cas, retourner au point 1

$\hookrightarrow$ Test sur des graphes de terrain.

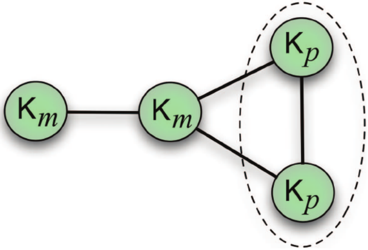
Idées

- Repérer rapidement lorsqu'un sous-graphe est séparable en deux modules (les auteurs donnent une condition suffisante).
- Utiliser d'autres algorithmes sur les petits sous-graphes.

Troisième partie III

Exemples

Illustration



Si $p \ll m$, les deux petites cliques sont réunies en un seul module.
$\hookrightarrow$ Plus généralement, tendance à regrouper les petits modules.

Résultats

Réseau	Après une étape (Q_{max})	Total (Q)
Yeast	9 (0,7396)	57 (0,6770)
E. Coli	27 (0,7519)	76 (0,6615)
Circuit électr.	11 (0,6701)	70 (0,6401)
Social	10 (0,6079)	21 (0,5316)
C. elegans	4 (0,4022)	57 (0,3613)

Conclusion

- Cette définition de la modularité ne convient pas pour résoudre le problème de la détection de communautés.
- Cela est dû aux compromis qui sont faits dans le calcul de la modularité (somme), à l'hétérogénéité des partitions.
- Q dépend de la taille totale du réseau... modularité plutôt locale ?

Remarques

- Mise en évidence des problèmes posés.
- Trop de calculs ?
- Étude sur les graphes de terrain très succincte.

Merci !

Des questions ?