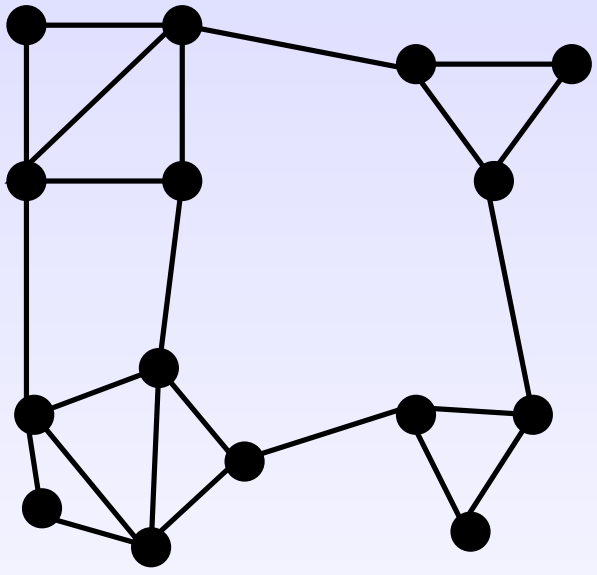
Modularité asymptotique de quelques classes de graphes

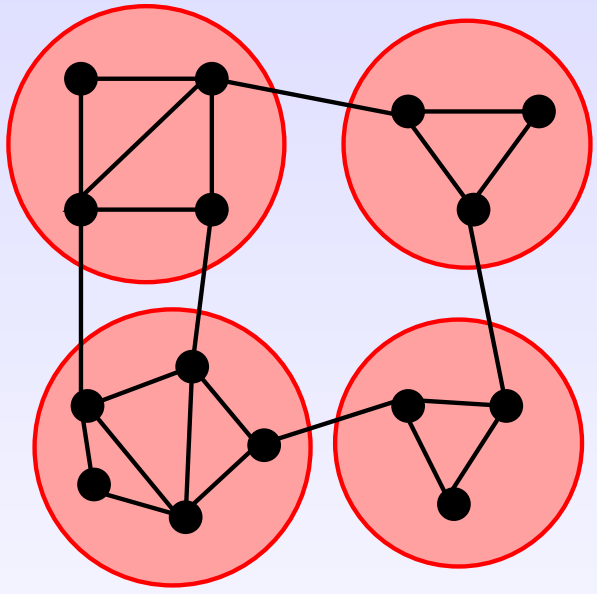
ou la fin de quelques idées reçues sur la modularité en huit minutes chrono

Fabien de Montgolfier, Mauricio Soto et Laurent Viennot

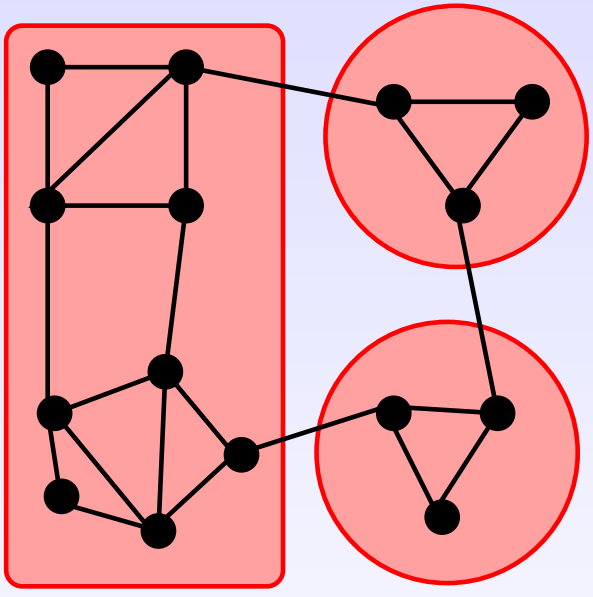
Clustering et modularité



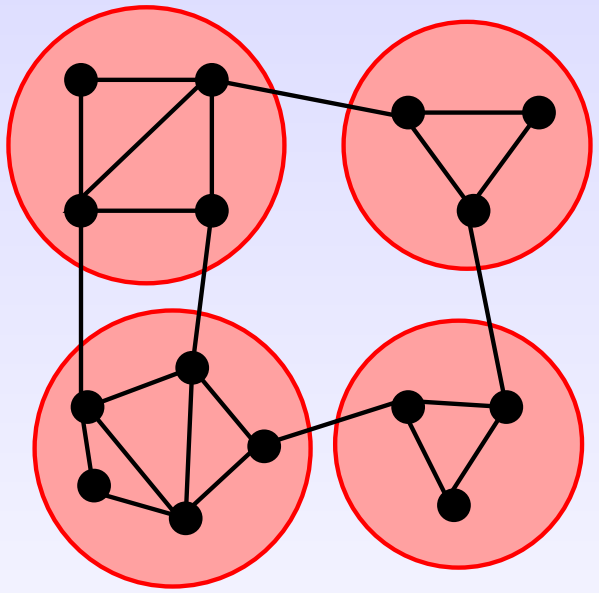
Clustering et modularité



Clustering et modularité

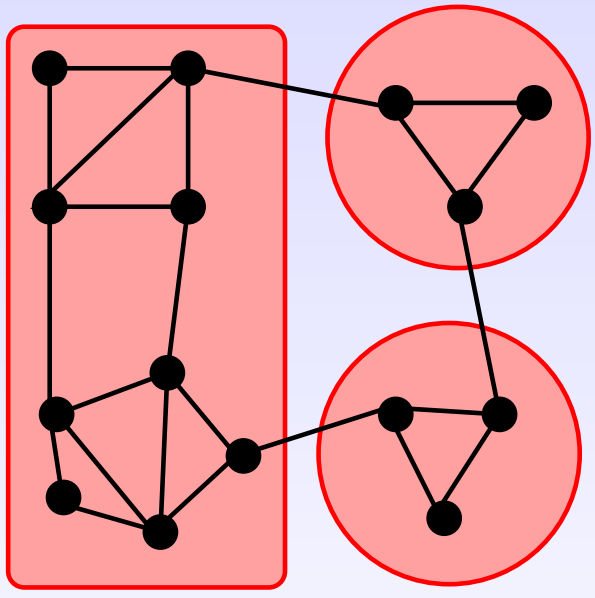


Clustering et modularité



modularité $= 0.68$

Clustering et modularité



modularité $= 0.75$

La modularité

La modularité [Newman & Girvan 2002, 2004]

Valeur entre -1 (mauvais) et 1 (bon).
0 = moyenne attendue pour l'aléatoire.

- Évaluer la qualité **d'un clustering**.

La modularité

La modularité [Newman & Girvan 2002, 2004]

Valeur entre -1 (mauvais) et 1 (bon).
0 = moyenne attendue pour l'aléatoire.

- Évaluer la qualité **d'un clustering**.
- Évaluer la qualité **d'un graphe** (max sur tous les clusterings)
  NP-complet [Brandes et al. 2008]

La modularité

La modularité [Newman & Girvan 2002, 2004]

Valeur entre -1 (mauvais) et 1 (bon).
0 = moyenne attendue pour l'aléatoire.

- Évaluer la qualité **d'un clustering**.
- Évaluer la qualité **d'un graphe** (max sur tous les clusterings) NP-complet [Brandes et al. 2008]

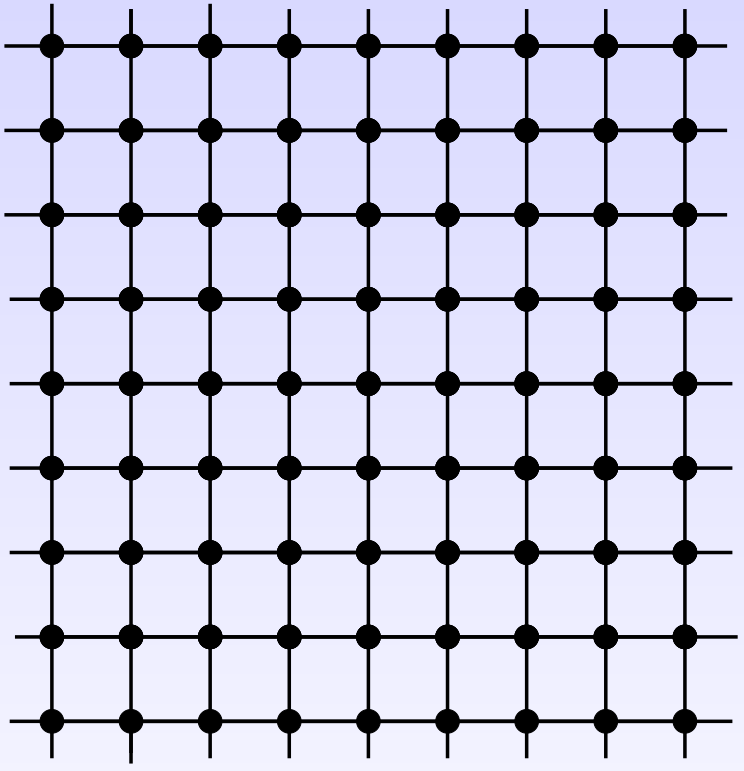
Sens ?

« *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* » [Newman & Girvan 2004].

La plus forte valeur apparaissant dans leurs articles est 0.86.

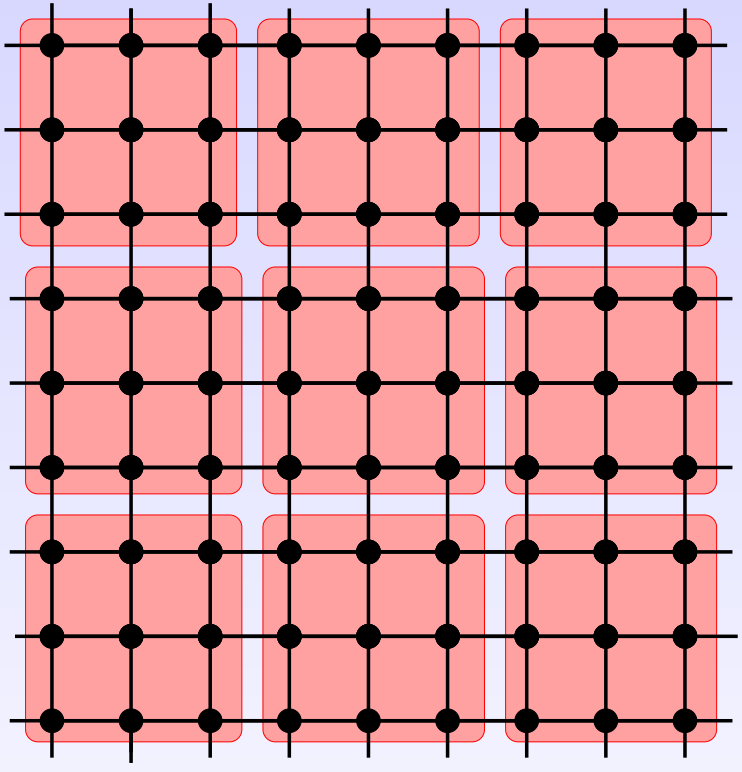
Une *strong community structure*, vraiment ?

- Prenons un tore carré de $n = \sqrt{n} \times \sqrt{n}$ sommets



Une *strong community structure*, vraiment ?

- Prenons un tore carré de $n = \sqrt{n} \times \sqrt{n}$ sommets
- Coupons-le en $k = \sqrt[3]{n}$ clusters carrés, chacun de côté $k = \sqrt[3]{n}$.
- La modularité de ce clustering est $1 - \frac{2}{\sqrt[3]{n}}$
- **Les tores ont donc modularité asymptotiquement 1.**



Ce clustering d'un tore 1000$\times$ 1000 a modularité 0.98

Autres résultats

Classes de modularité asymptotiquement 1

- Tores (même non carré) et grilles : $1 - \Theta(n^{-1/3})$
- Hypertore G de dimension d : $1 - \Theta(n^{-1/2d})$
- Hypercube ($n = 2^{dimension}$) : $1 - \Theta(\frac{\log \log n}{\log n})$

Classes de modularité exactement 0 [folklore]

- Cliques K_n
- Étoiles $K_{1,n}$

2 ème résultat : arbres de degré max $o(n)$

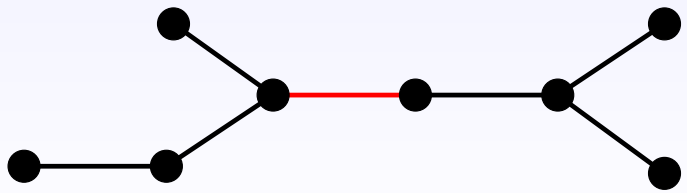
Def : **arête centroïde** d'un arbre T

Sa suppression coupe l'arbre en deux composantes de taille aussi proche que possible.

Def : clustering centroïdal (pour h donné)

F est une forêt. Initialement $F = T$.
Tant qu'il existe un arbre A de F de taille $> h$

- Enlever une arête centroïde de T



2 ème résultat : arbres de degré max $o(n)$

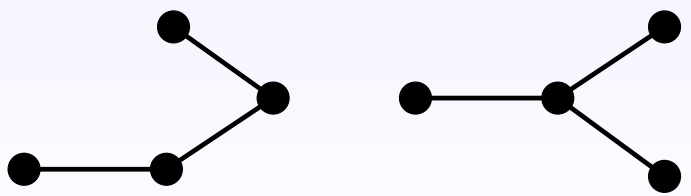
Def : **arête centroïde** d'un arbre T

Sa suppression coupe l'arbre en deux composantes de taille aussi proche que possible.

Def : clustering centroïdal (pour h donné)

F est une forêt. Initialement $F = T$.
Tant qu'il existe un arbre A de F de taille $> h$

- Enlever une arête centroïde de T



2 ème résultat : arbres de degré max $o(n)$

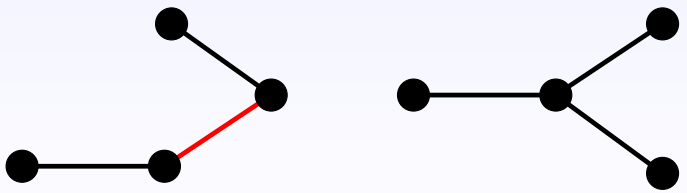
Def : **arête centroïde** d'un arbre T

Sa suppression coupe l'arbre en deux composantes de taille aussi proche que possible.

Def : clustering centroïdal (pour h donné)

F est une forêt. Initialement $F = T$.
Tant qu'il existe un arbre A de F de taille $> h$

- Enlever une arête centroïde de T



2 ème résultat : arbres de degré max $o(n)$

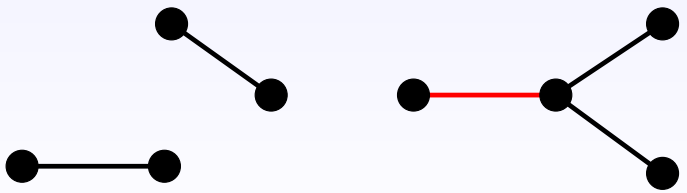
Def : **arête centroïde** d'un arbre T

Sa suppression coupe l'arbre en deux composantes de taille aussi proche que possible.

Def : clustering centroïdal (pour h donné)

F est une forêt. Initialement $F = T$.
Tant qu'il existe un arbre A de F de taille $> h$

- Enlever une arête centroïde de T



2 ème résultat : arbres de degré max $o(n)$

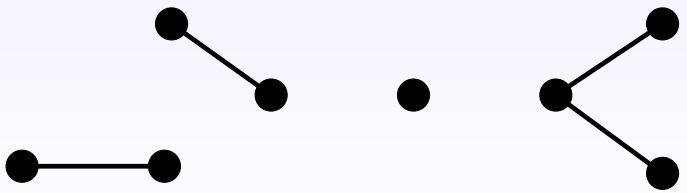
Def : **arête centroïde** d'un arbre T

Sa suppression coupe l'arbre en deux composantes de taille aussi proche que possible.

Def : clustering centroïdal (pour h donné)

F est une forêt. Initialement $F = T$.
Tant qu'il existe un arbre A de F de taille $> h$

- Enlever une arête centroïde de T



2 ème résultat : arbres de degré max $o(n)$

Def : **arête centroïde** d'un arbre T

Sa suppression coupe l'arbre en deux composantes de taille aussi proche que possible.

Def : clustering centroïdal (pour h donné)

F est une forêt. Initialement $F = T$.
Tant qu'il existe un arbre A de F de taille $> h$

- Enlever une arête centroïde de T

Théorème

Si T est un arbre de degré maximum Δ alors en prenant $h = \sqrt{\Delta n}$
la modularité est $\geqslant 1 - \dfrac{5\sqrt{\Delta n} - 5}{2n - 2} = 1 - \Theta(\sqrt{\Delta}/\sqrt{n})$.
Pour les classes d'arbres où $\Delta = o(n)$ la modularité tend vers 1.

Graphes peu denses

Def : graphes peu denses

- graphes connexes
- et de **degré moyen** d (constante)
- et de **degré maximal** $\Delta = o(d\sqrt{n})$

Clustering par arbre couvrant

- Prendre un arbre couvrant T de G
- Lui appliquer le clustering centroïdal
- On suppose toutes les arêtes de $G - T$ sont hors clusters.
- Théorème : modularité $\geqslant \frac{2}{d} - \frac{3\Delta}{d\sqrt{n}}$
- Peu denses $\implies$ modularité $\geqslant \frac{2}{d} - o(1)$

3ème résultat : graphes peu denses et power-law

Théorème

La classe des graphes peu denses (connexes, degré moyen d, degré max $\Delta = o(d\sqrt{n})$) a modularité asymptotique $\geqslant 2/d$

Un algorithme de clustering a garantie de performance

Preuve constructive pour chaque graphe. Algo en $O(n^2)$.

Def : power-law graphs

graphes dont la proportion de sommets de degré au moins k varie en $O(k^{-\alpha})$ (NB : rajouter 1 à α si « degré exactement k »).

Corollaire

Pour $\alpha > 2$, les power-law graphs connexes de degré moyen d ont modularité asymptotique au moins $2/d$

Conclusion

Des classes **régulières** ont modularité asymptotiquement 1

Grilles, (hyper)tores, hypercube... Une modularité élevée n'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.

Un algorithme de clustering à performance garantie

Prendre un arbre couvrant et lui appliquer le clustering centroïdal

$$\text{modularité} \geqslant \frac{2}{d} - \frac{3\Delta}{d\sqrt{n}}$$

Une dichotomie sur les arbres

- Classes de degré max $o(n)$: modularité asymptotique 1
- Degré max non borné : modularité 0 pour les étoiles