Détection automatique des structures communautaires dans les réseaux

Abdoulaye Diop

Octobre 2024

Contents

1 Introduction

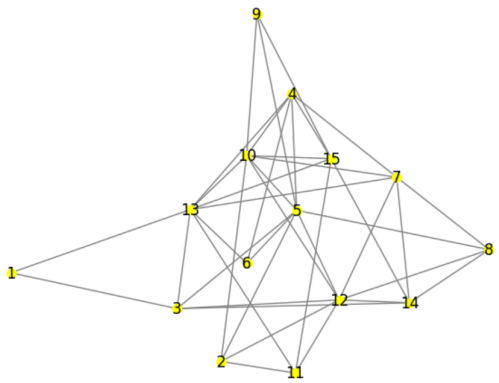
Les systèmes sociaux, technologiques et d'information peuvent être modélisés en **réseaux complexes**, composés de nœuds interconnectés, mêlant **organisation** et **hasard**. Une méthode efficace pour les analyser est de les décomposer en **communautés**, des groupes de nœuds densément connectés. La **détection de communautés** vise à partitionner le réseau en sous-ensembles de nœuds fortement connectés, avec des liens faibles entre communautés, tout en tenant compte de l'**efficacité computationnelle** des algorithmes. On distingue principalement trois types d'algorithmes : 1. Les **algorithmes de division** qui fragmentent le réseau ; 2. Les **algorithmes d'agglomération** qui regroupent les nœuds ; 3. Les **méthodes d'optimisation** qui maximisent la **modularité**.

Ce document se focalise sur les méthodes d'optimisation.

2 Données

Dans ce travail, les données d'entrée sont représentées sous la forme de graphes. Un **graphe** est composé d'un ensemble de nœuds (ou sommets) et de liens (ou arêtes) entre ces nœuds. Les **communautés** correspondent à des sous-ensembles de nœuds connectés plus fortement entre eux qu'avec le reste du graphe.Les caractéristiques fondamentales d'un graphe incluent :

- **Degré** : Le nombre de liens associés à un nœud ; le degré entrant et sortant pour les graphes orientés.
- **Communautés** : Des groupes de nœuds qui sont fortement interconnectés.
- **Matrice d'Adjacence** : Une représentation matricielle qui indique les connexions entre les nœuds.



(a) Graphe de 15 nœuds et 40 liens

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15
N1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
N2	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0
N3	1	0	0	0	1	0	0	0	0	0	0	1	1	1	0
N4	0	0	0	0	1	1	1	0	0	1	0	0	1	0	1
N5	0	1	1	1	0	1	0	1	1	1	0	1	0	0	0
N6	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0
N7	0	0	0	1	0	0	0	1	0	1	0	1	1	1	0
N8	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0
N9	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
N10	0	1	0	1	1	0	1	0	1	0	0	1	1	0	1
N11	0	1	0	0	0	0	0	0	0	0	0	1	1	0	1
N12	0	1	1	0	1	0	1	1	0	1	1	0	0	1	0
N13	1	0	1	1	0	1	1	0	0	1	1	0	0	0	1
N14	0	0	1	0	0	0	1	1	0	0	0	1	0	0	1
N15	0	0	0	1	0	0	0	0	1	1	1	0	1	1	0

(b) Matrice d'adjacence du graphe

Figure 1: Représentation du graphe et de sa matrice d'adjacence.

3 Modularité de la partition

La **qualité d'une partition** est souvent mesurée à l'aide de la **modularité** Q, un indice qui varie entre -1 et 1, et qui évalue dans quelle mesure les nœuds à l'intérieur d'une communauté sont plus densément connectés que ce que l'on pourrait attendre par hasard. La modularité est définie par la formule suivante :

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

où : A_{ij} est le poids de l'arête entre les nœuds i et j. Si i et j ne sont pas directement connectés, alors $A_{ij} = 0$. $k_i = \sum_j A_{ij}$ est la somme des poids des arêtes attachées au sommet i, c'est-à-dire le degré pondéré de i. c_i est la communauté à laquelle appartient le sommet i. $\delta(c_i, c_j)$ est une fonction indicatrice égale à 1 si $c_i = c_j$, c'est-à-dire si les nœuds i et j appartiennent à la même communauté, sinon elle vaut 0. $m = \frac{1}{2} \sum_{i,j} A_{ij}$ est le nombre total d'arêtes pondérées dans le réseau.

4 Les méthodes d'optimisation

Les méthodes d'optimisation dans la détection de communautés sont des approches qui visent à maximiser une fonction objective pour trouver la meilleure partition d'un réseau en communautés. L'une des fonctions les plus couramment utilisées dans ce contexte est la modularité, qui mesure la qualité d'une partition en comparant la densité des liens à l'intérieur des communautés avec la densité attendue des liens dans un réseau aléatoire

4.1 Clauset et al

La méthode de Clauset et al. est un algorithme de détection de communautés dans les réseaux qui vise à maximiser la modularité. Développée par Aaron Clauset, Matthew Newman, et Cristopher Moore en 2004, cette méthode est populaire en raison de sa capacité à identifier efficacement des structures communautaires dans des réseaux complexes.

Algorithme de Clauset et al

1. **Initialisation :** Assignez chaque nœud v_i à sa propre communauté $C_i = \{v_i\}$ pour $i = 1, 2, \ldots, n$, où n est le nombre total de nœuds.
2. **Calcul de la Modularité :** $Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$
3. **Variation de la Modularité :** Pour chaque paire C_a et C_b, $\Delta Q(C_a, C_b)$
4. **Fusion :** Trouvez (C_a, C_b) où ΔQ est maximal et fusionnez-les $C_{new} = C_a \cup C_b$
5. **Répétition :** Répétez les étapes 2 à 4 jusqu'à ce que la modularité Q ne puisse plus être améliorée.
6. **Résultat Final :** À la fin, obtenez une partition maximisant Q.

4.2 Méthode de Louvain

La méthode de Louvain est un algorithme de détection de communautés qui utilise une approche hiérarchique(agglomerative) pour maximiser la modularité d'un réseau. Il divise le réseau en groupes de nœuds, appelés communautés, de manière à ce que les nœuds à l'intérieur d'une communauté soient densément connectés entre eux, tandis que les connexions entre différentes communautés soient rares.

Étapes de l'algorithme

Phase 1 : Optimisation locale

Initialisation : Chaque nœud v_i est dans sa propre communauté $C_i = \{v_i\}, \quad i = 1, 2, \ldots, n$, où n est le nombre total de nœuds.

Modularité : Q

Déplacement : Pour chaque nœud v_i, calculez ΔQ si v_i est déplacé vers C_k :

$$\Delta Q = \left(\frac{P_{\text{in}} + k_{i,\text{in}}}{2m} - \frac{P_{\text{tot}} + k_i}{2m} \right)^2 - \left(\frac{P_{\text{in}}}{2m} - \left(\frac{P_{\text{tot}}}{2m} \right)^2 - \frac{k_i}{2m} \right)^2$$

avec P_{in} somme des poids des arêtes internes, $k_{i,\text{in}}$ degré de v_i vers C_k, et P_{tot} somme des poids de v_i vers tous les autres nœuds.

Répétez jusqu'à ce qu'il n'y ait plus d'amélioration de modularité.

Phase 2 : Agrégation

Nouveau réseau : Chaque communauté devient un nœud dans un nouveau réseau G' avec $A'_{kl} = \sum_{v_i \in C_k, v_j \in C_l} A_{ij}$, où A'_{kl} est le poids entre les communautés C_k et C_l.

Répétez les phases 1 et 2 jusqu'à la convergence.

Résultat Final

À la fin, une partition maximisant la modularité Q est obtenue.

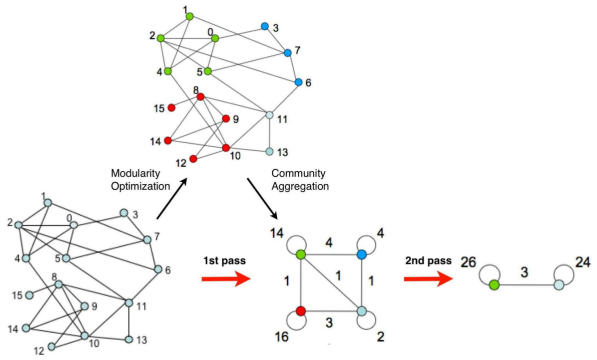


Figure 2: Etapes de l'algorithme

5 Conclusion

La méthode de Louvain est très efficace pour la détection de communautés dans de grands réseaux, traitant des millions de nœuds et d'arêtes, avec une rapidité et une qualité de résultats supérieures. En revanche, la méthode de Clauset, plus simple et directe, convient mieux aux réseaux de taille modérée, bien qu'elle puisse être moins performante pour des réseaux complexes. Le choix entre Louvain et Clauset dépend de la taille et de la complexité du réseau ainsi que de la précision des résultats recherchés.

	Karaté	Arxiv	Internet	Web nd.edu	Téléphone	Web uk-2005	Web WebBase 2001
Nœuds/liens	34/77	9k/24k	70k/351k	325k/1M	2.6M/6.3M	39M/783M	118M/1B
CNM	.38/0s	.772/3.6s	.692/799s	.927/5034s	-/-	-/-	-/-
PL	.42/0s	.757/3.3s	.729/575s	.895/6666s	-/-	-/-	-/-
WT	.42/0s	.761/0.7s	.667/62s	.898/248s	.56/464s	-/-	-/-
Notre algorithme	.42/0s	.813/0s	.781/1s	.935/3s	.769/134s	.979/738s	.984/152mn

Figure 3: Comparaison de la complexite des algorithmes

A Annexe

A.1 Détails Techniques de la méthode de louvain

Fonctions de Base

Calculer **la Modularité** Q pour évaluer la qualité de la partition d'un réseau selon les connexions internes et externes. Évaluer **la Variation de Modularité** ΔQ lors du déplacement d'un nœud vers une nouvelle communauté. Évaluer **la Vraisemblance** L pour mesurer la vraisemblance de Q.

Optimisation par Gradient Descente

Implémenter le Gradient Descente pour minimiser $-\log(L)$ et maximiser Q.

Boucle d'Optimisation

Pour chaque nœud, déplacer vers la communauté C si $\Delta Q(i, C) > 0$. Répéter jusqu'à atteindre un minimum de $-\log(L)$, signalant que la partition optimale est trouvée.

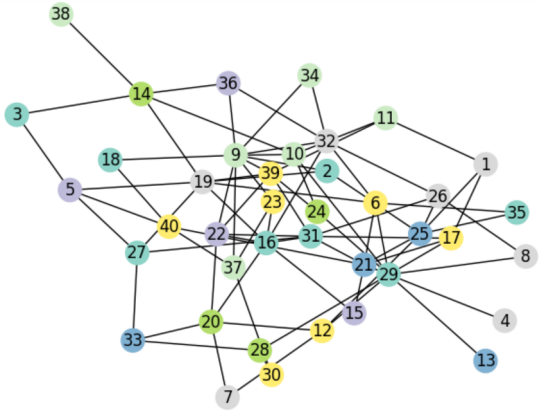
A.2 Glossaire

$$\Delta Q(C_a, C_b) = \frac{1}{2m}\left(2\sum_{i\in C_a, j\in C_b} A_{ij} - \frac{2\sum_{i\in C_a} k_i \sum_{j\in C_b} k_j}{2m}\right)$$

A.3 Applications de la méthode

```
Communauté 1: {32, 6, 11, 12, 19, 27, 28, 29}
Communauté 2: {16, 3, 20, 21, 30, 14, 15}
Communauté 3: {17, 36, 22, 39, 8}
Communauté 4: {0, 38, 23, 9, 10}
Communauté 5: {1, 34, 33, 5, 7, 24, 25, 31}
Communauté 6: {2, 35, 18, 37, 4, 26, 13}
```

(a) Commaunautes formées



(b) Clusters par couleurs pour un reseau de 40 noeuds et 80 liens

Figure 4: Représentation du graphe et de sa matrice d'adjacence.

References

[1] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.