Partition en sous graphes denses pour la détection de communautés

Julien Darlay, Nadia Brauner, Julien Moncel

Laboratoire G-SCOP, 46 avenue Félix Viallet, 38031 Grenoble, France
{julien.darlay, nadia.brauner, julien.moncel}@grenoble-inp.fr

La détection de communautés est une problématique cruciale lors de l'analyse de systèmes d'interactions. Par exemple, on peut souhaiter étudier les interactions entre des individus, entre des protéines ou bien les liens entre différents sites internet. Ces données peuvent se représenter sous forme de graphes où chaque noeud représente un individu et une arête représente une interaction entre deux individus. Ces réseaux (contrairement aux graphes aléatoires) ont la propriété de se diviser en communautés. De manière intuitive, une communauté est un ensemble d'individus ayant de fortes interactions entre eux et peu d'interactions avec l'extérieur. Ce problème est classique et de nombreux modèles ont été proposés [1].

Ce problème nous est apparu lors de nos travaux sur les méthodes de la recherche opérationnelle pour l'analyse combinatoire de données. Nous souhaitions regrouper un grand nombre de caractérisations de population (appelées motifs) en un nombre plus réduit de familles représentatives. Ces motifs sont alors les sommets du graphe, ils sont reliés par une arête s'ils caractérisent la même population et les communautés sont les familles représentatives.

Dans ce travail, nous présentons un nouveau modèle basé sur la densité définie par [2]. Notre approche est théorique, nous donnons des résultats de complexité et d'approximabilité sur les problèmes liés à l'optimisation de ce critère. Nous nous intéresserons au cas particulier des arbres.

1 Formulation mathématique

Soit $G = (V, E)$ un graphe simple et Π l'ensemble des partitions de V sans classe vide. Nous définissons la *densité* d d'une partition $P \in \Pi$ par la formule suivante :

$$d(P) = \sum_{X \in P} d(G[X]) = \sum_{X \in P} \frac{|E(X)|}{|X|}$$

La densité d'un sous-graphe est vue comme le rapport entre le nombre d'arêtes dans le sous-graphe et le nombre de sommets. Le problème classique lié à cette définition de la densité consiste à trouver le sous-graphe de densité maximum. Ce problème peut être résolu en temps polynomial en utilisant la programmation linéaire [3] ou des techniques de flots [2]. Lorsque la taille du sous-graphe fait partie de l'instance, le problème devient difficile [3].

Le problème de partition qui nous intéresse ici est de trouver la partition P qui maximise la densité $d(P)$. Chaque classe de la partition correspond alors à une communauté.

2 Algorithme sur les arbres

Le problème de décision associé à la recherche de la partition de plus grande densité est NP-Complet en effectuant une réduction depuis le problème classique de coloration de graphe. Cette réduction préserve l'approximabilité du problème et par conséquent il a le même facteur d'approximation que coloration. Il n'est donc pas possible de trouver un algorithme d'approximation avec garantie de performance.

Dans le cas particulier des arbres, en s'intéressant à la structure d'une solution optimale, on peut construire un algorithme donnant la partition de densité maximale en temps polynomial. Une première caractérisation des classes d'une solution optimale est donnée par la propriété suivante :

Propriété 1 *Soit $T = (V, E)$ un arbre et P^* une partition de densité maximale. Alors pour toute classe X de P^*, le sous-graphe induit par X est une étoile avec au moins deux sommets.*

Une autre propriété importante vient du lien entre une partition en étoiles et le transversal de cardinalité minimum.

Propriété 2 *Soit T^* un transversal de cardinalité minimum et P^* une partition de densité maximale. Alors $|T^*| = |P^*|$ et de plus chaque sommet du transversal T^* est dans exactement une classe de P^*.*

L'algorithme consiste à enraciner l'arbre puis à résoudre le problème niveau par niveau en commençant par les feuilles. Trouver la meilleure partition à un niveau donné utilise les solutions des deux niveaux précédents et le transversal de cardinalité minimum.

3 Perspectives

Il serait intéressant d'étudier le problème sur les graphes bipartis et plus généralement sur les classes de graphes où la coloration est polynomiale. D'un point de vue plus pratique, il faudrait trouver une heuristique pour résoudre le problème dans le cas général. Une telle heuristique permettrait de comparer les solutions de ce modèle avec les solutions des modèles de la littérature pour le problème de partition en communautés (y compris avec d'autres critères).

Références

[1] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69, 2004.

[2] A. Goldberg. Finding a Maximum Density Subgraph *EECS Department, University of California, Berkeley*, 1984

[3] M. Charikar. Greedy approximation algorithms for finding dense components in a graph *APPROX*, 2000