

Extraction sous Contraintes d'Ensembles de Cliques Homogènes

Pierre-Nicolas Mougel * **, Marc Plantevit* ***
Christophe Rigotti* **, Olivier Gandrillon****, Jean-François Boulicaut* **

*Université de Lyon, CNRS, INRIA

**INSA-Lyon, LIRIS Combining, UMR5205, F-69621, France

***Université Lyon 1, LIRIS Combining, UMR5205, F-69622, France
prénom.nom@liris.cnrs.fr,

****Université Lyon 1, Centre de Génétique Moléculaire et Cellulaire, UMR5534,
F-69622, Villeurbanne, France
gandrillon@cgmc.univ-lyon1.fr

Résumé. Nous proposons une méthode de fouille de données sur des graphes ayant un ensemble d'étiquettes associé à chaque sommet. Une application est, par exemple, d'analyser un réseau social de chercheurs co-auteurs lorsque des étiquettes précisent les conférences dans lesquelles ils publient. Nous définissons l'extraction sous contraintes d'ensembles de cliques tel que chaque sommet des cliques impliquées partage suffisamment d'étiquettes. Nous proposons une méthode pour calculer tous les Ensembles Maximaux de Cliques dits Homogènes qui satisfont une conjonction de contraintes fixée par l'analyste et concernant le nombre de cliques séparées, la taille des cliques ainsi que le nombre d'étiquettes partagées. Les expérimentations montrent que l'approche fonctionne sur de grands graphes construits à partir de données réelles et permet la mise en évidence de structures intéressantes.

1 Introduction

De nombreux problèmes peuvent être étudiés sous l'angle de l'analyse de grands graphes dont les sommets représentent des entités et les arêtes des relations ou interactions entre ces entités. Les contextes d'application sont innombrables et nous traitons ici deux graphes réels issus de l'étude d'un réseau social (coopération de chercheurs) et de réseaux d'interaction protéine/protéine. L'accumulation toujours plus facile de données sur de tels graphes mobilise largement la communauté des chercheurs en fouille de données. Deux approches complémentaires sont principalement étudiées. De nombreux travaux s'intéressent aux propriétés macroscopiques de ces graphes (e.g., valeurs moyennes des diamètres, nombre et taille des composantes connexes, degrés). D'autres se réclament davantage de la découverte de motifs locaux pour identifier des sous-graphes particuliers, par exemple, les cliques maximales.

Nous nous intéressons à des données qui correspondent à un graphe et tel qu'un ensemble d'étiquettes soit associé à chaque sommet. De telles données sont souvent disponibles mais la