Étude théorique des problèmes de γ-Clustering.

Julien Baste[1], Antoine Castillon[1], Clarisse Dhaenens[1],
Mohammed Haddad[2], Hamida Seba[2]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
`{julien.baste,antoine.castillon,clarisse.dhaenens}@univ-lille.fr`
[2] Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR 5205, F-69622 Villeurbanne, France
`{mohammed.haddad,hamida.seba}@univ-lyon1.fr`

Mots-clés : *Graphes, γ-Clustering, Complexité, Complexité paramétrée, Approximations.*

1 Introduction

Les problèmes de Clustering sont bien connus en théorie des graphes et consistent à ajouter et/ou supprimer des arêtes à un graphe jusqu'à obtenir une union de cliques déconnectées. Si seul l'ajout (resp. la suppression) d'arêtes est autorisé on parle alors du problème de COMPLETION (resp. DELETION). Si ces deux opérations sont autorisées on parle alors du problème d'EDITION. Ces problèmes ont déjà été très largement étudiés au travers de multiples aspects : la complexité [7], l'approximabilité [5, 6], la complexité paramétrée et la kernelization [3, 4], etc.

Bien souvent les cliques sont trop restrictives pour bien représenter les sous-graphes denses d'un graphe réel, qui sont en général beaucoup plus irréguliers. Dans le cas des problèmes de Clustering cela se traduit en général par un nombre d'ajouts d'arêtes très supérieur au nombre de suppressions. La Figure 1 présente un graphe où l'on a trois ensembles de sommets denses (de densité d'arêtes proche de 90%). Il est évident que ces trois ensembles correspondront aux cliques finales, cependant la solution optimale nécessitera 124 ajouts d'arêtes pour transformer ces ensembles en cliques. A contrario, il faudra seulement 30 suppressions d'arêtes pour déconnecter ces ensembles.

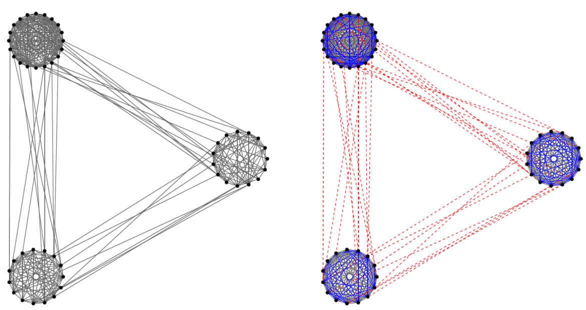

FIG. 1 – Un graphe où la solution optimale de CLUSTER EDITING nécessite 124 ajouts d'arêtes et seulement 30 suppressions.

2 Définitions et Résultats

Afin d'éviter ce problème, nous avons utilisé les quasi-cliques pour représenter les sous-graphes denses.

Définition 1 (Quasi-Clique en degré) *Etant donné $\gamma \in [0,1]$, un graphe $G = (V, E)$ est une γ-quasi-clique en degré, si pour tout $u \in V$, $d(u) \geq \gamma(|V| - 1)$, où $d(u)$ est le degré de u.*

Définition 2 (Quasi-Clique en densité) *Etant donné $\gamma \in [0, 1]$, un graphe $G = (V, E)$ est une γ-quasi-clique en densité, si $2|E| \geq \gamma|V|(|V| - 1)$.*

Nous avons donc introduit deux généralisations, une pour chaque définition des quasi-cliques, pour les problèmes de Completion, de Deletion et d'Editing. La définition formelle du problème de γ-Degree Cluster Editing est donnée Définition 3, les autres problèmes sont définis de façon similaire.

Définition 3 (γ-Degree Cluster Editing) *Etant donné $G = (V, E)$, trouver un ensemble $S \subseteq \binom{V}{2}$ de taille minimale tel que les composantes connexes de $(V, E\Delta S)$ soient des γ-quasi-cliques en degré où $\binom{V}{2}$ l'ensemble des paires de V et Δ la différence symétrique.*

Il est important de noter que même si les quasi-cliques permettent de mieux représenter les sous-graphes denses que les cliques, elles apportent aussi en général beaucoup de complexité. En effet plusieurs problèmes polynomiaux avec des cliques deviennent NP-complets avec des quasi-cliques, et ce pour les deux définitions. Par exemple, il est NP-complet de décider si une quasi-clique donnée est maximale [2] de même pour décider si un ensemble donné est contenu dans une quasi-clique [1].

Nous avons d'abord prouvé que tous ces problèmes sont NP-complets à l'exception du problème de Completion avec des quasi-cliques en degré qui est polynomial. Il est important de noter que le problème classique de Cluster Completion et le problème γ-Degree Cluster Completion sont tous deux polynomiaux, contrairement au problème γ-Density Cluster Completion qui lui est NP-complet. Cela montre encore une fois la difficulté supplémentaire apportée par les quasi-cliques (en densité) par rapport aux cliques, de plus il s'agit du seul problème que nous connaissons où il y a une différence de complexité majeure entre les quasi-cliques en degré et les quasi-cliques en densité. Nous présenterons également plusieurs résultats de non-approximabilité pour certains cas particuliers des problèmes. Enfin, nous présenterons des algorithmes FPT paramétrés par le nombre de modifications pour les problèmes de Deletion et d'Edition avec des quasi-cliques en degré ainsi que pour le problème de Completion avec des quasi-cliques en densité.

Références

[1] Takeaki Uno. An efficient algorithm for enumerating pseudo cliques. In *Proc. of the 18th int. conf. on Algorithms and computation* (ISAAC'07), pp. 402–414. 2007.

[2] Seyed-Vahid Sanei-Mehri, Apurba Das and Srikanta Tirthapura. Enumerating Top-k Quasi-Cliques. *IEEE Int. Conf. on Big Data (Big Data)*, pp. 1107-1112. 2018.

[3] Gramm Jens, Guo Jiong, Hüffner Falk and Niedermeier Rolf. Graph-Modeled Data Clustering : Fixed-Parameter Algorithms for Clique Generation. *Theor Comput Syst*, 38. 2003.

[4] Yixin Cao and Jianer Chen. Cluster Editing : Kernelization Based on Edge Cuts. *Algorithmica* 64, 1, pp. 152–169. 2012.

[5] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation Clustering. In *Proc. of the 43rd Symposium on Foundations of Computer Science (FOCS '02).* IEEE Computer Society, USA, 238. 2002.

[6] Moses Charikar, Venkatesan Guruswami and Anthony Wirth. Clustering with qualitative information, *Journal of Computer and System Sciences*, Volume 71, Issue 3, pp. 360-383. 2005.

[7] Assaf Natanzon, Ron Shamir and Roded Sharan, Complexity classification of some edge modification problems, *Discrete Applied Mathematics*, Volume 113, Issue 1, pp. 109-128. 2001.