Décomposition modulaire de graphes en vue de leur compression

Directeurs : Fabien de Montgolfier et Stefano Zacchiroli

Lieu d'exercice : IRIF

Établissement de rattachement : Université de Paris

Durée : trois ans à partir de septembre 2021

Beaucoup de disciplines scientifiques étudient des *relations*, telles que celles entre différents fichiers et versions d'un logiciel, ou encore les interactions sociales entre individus, les réseaux informatiques, le Web, etc. Ces relations sont représentées par des graphes. A l'heure du 'Big Data', ils atteignent maintenant des tailles telles que réduire l'espace qu'occuppe leur représentation est nécéssaire, tout en conservant la possibilité de les manipuler efficacement. On souhaite donc disposer d'une structure de données qui soit *compressée* (peu de bits par arrête et sommet) tout en permettant de répondre à des requêtes telles que l'adjacence de deux sommets sans tout décompresser. Le but de cette thèse est de developper et d'implémenter des algorithmes de *décomposition modulaire* afin de les appliquer à la compression de graphes.

Un module est un ensemble M de sommets tel que tout sommet $x \notin M$ est soit voisin de tous les sommets de M, soit d'aucun. Compacter un module en un sommet permet de compresser sans perte le graphe. De façon équivalente M est un module s'il n'existe pas de sommet y qui *distingue* M en n'ayant pas la même relation avec tous les sommets de M. Cela permet de définir une *décomposition modulaire* adaptée pour différents objets : graphes orientés, étiquetés sur les arêtes, bipartis, hypergraphes...

Dans cette thèse, on pourra
- *définir* de nouvelles variantes de la décomposition, selon l'objet que l'on veut décomposer
- *concevoir* des algorithmes efficaces de décomposition modulaire de ces objets
- et enfin les *implémenter* dans des bibliothèques existantes de manipulation de graphes

Le cas d'application typique sur lequel on se penchera est celui de Software Heritage—la plus grande archive au monde de code source librement accessible, avec son historique de développement, telle que capturée par les systèmes de contrôle de versions (VCS) modernes (p.ex., Git, Mercurial, etc.)—dont le modèle de données est un graphe sans cycles (DAG) de Merkle, typé, avec des propriétés associées aux noeuds et aux arêtes. Le graphe correspondant à l'archive de Software Heritage contient aujourd'hui ≈ 20 milliard des noeuds et ≈ 200 milliards d'arêtes et est manipulé en format compressé, moyennant des techniques de compression de gros graphes (du Web, de réseaux sociaux, etc.).

Pour appliquer efficacement la décomposition modulaire il faut travailler implicitement (sans la stocker) sur sa fermeture transitive du graphe (ajout d'un arc xy quand il existe un chemin de x à y). Le graphe manipulé ayant une taille considérable il sera intéressant d'essayer de *répartir* les données entre plusieurs machines.

Tout cela définit une compression sans perte. Une extension à regarder est la compression avec perte, en utilisant des approximations de modules (il existe quelques sommets qui les distinguent).

Cette thèse est financée par le projet de recherche ANR Corégraphie :COmpression de RÉseaux et de GRAPHes pour une Informatique Efficace. La doctorante ou le doctorant sera intégré à ce projet ANR.

Compétences requises :
- en algorithmique : bon niveau général, et bases de théorie des graphes
- en programmation : bases de programmation système sous Linux, et maîtrise d'un ou plusieurs langages de programmation entre : C, Java, Python
- un plus serait d'avoir déjà une expérience de codage d'algorithmes efficaces de graphes ou d'usage de frameworks pour l'analyse ou le traitement de gros graphes

Candidature : CV, lettre de motivation, relevés de notes,...etc. à envoyer à Fabien de Montgolfier `<fm@irif.fr>` et Stefano Zacchiroli `<zack@irif.fr>`. Réponse souhaitée avant le 29 juin.