COMPRESSION ADAPTATIVE DE GRANDS GRAPHES

(SUJET DE STAGE, 2019-20)

LUCA CASTELLI ALEARDI (LIX, ECOLE POLYTECHNIQUE)

Résumé. Nous allons considérer le problème de compresser et représenter de manière compacte et efficace la connectivité d'un graphe. On s'intéressera entre autres aux *réseaux complexes* (en font partie les réseaux sociaux, le graphe du web, les réseaux neuronaux), ainsi qu'à d'autres classes de graphes du monde réel (tels que les maillages 3D utilisés en computer graphics). Les graphes issus d'applications du monde réel sont loin du cas aléatoire et partagent des propriétés structurelles assez surprenantes.

Dans se stage on vise à faire une analyse adaptative (à la fois sur le plan théorique ainsi que experimental) des méthodes de compression des graphes, de manière à établir des évaluations rigoureuses et comparaisons précises.

Le but étant de faire intervenir un ou plusieurs paramètres structurels (suite des degrées de sommets, modularity, ...), et pouvoir ainsi concevoir de nouveaux schémas de compression et représentations compactes, qui seraient plus adaptées pour certaines classes de grands graphes.

Thématiques : réseaux sociaux, graphes, maillages 3D, compression.
Laboratoire d'accueil : LIX (Ecole Polytechnique), équipe "Combinatoire".
Responsables du stage : Luca Castelli Aleardi (`amturing@lix.polytechnique.fr`).
Compétences souhaitées : une bonne connaissance de l'algorithmique et des structures de données, maîtrise d'un langage de programmation (Java ou C++, par exemple), quelques notions de math discrètes (rudiments de théorie des graphes et combinatoire)

1. Introduction

Compression de graphes et réseaux. A cause de leur ubiquité dans les applications du monde réel, les graphes constituent un outil privilégié en informatique : il peut s'agir de *maillages 3D* utilisés en computer graphics pour modéliser des formes 3D, ou bien de *réseaux* permettant de décrire les interactions entre les entités d'un système complexe (par exemple les éléments d'un système physique, ou les usagers d'un réseau social).

Compte tenu du coût important de la connectivité des grands graphes utilisés dans la pratique aujourd'hui (ayant plusieurs dizaines, voire centaines de millions de sommets), il est crucial de se munir de représentations qui soient peu gourmandes en mémoire. A titre d'exemple, on peut coder un réseau à l'aide de sa matrice d'adjacence : cette solution est facile à mettre en place et fournit un accès très rapide aux données (pour tester, par exemple, si deux noeuds sont voisins), mais elle est loin d'être adaptée pour représenter des grands graphes (nécessitant de ressources mémoires de taille quadratique en le nombre de noeuds). Une manière plus efficace consiste à se servir de listes d'adjacences (pour chaque sommet on stocke la liste de ses voisins), ce qui conduit à des meilleures représentations dans le cas des graphes creux. Même cette dernière solution est loin d'être adaptée pour faire face à des millions de noeuds (et des dizaines ou centaines de millions de liens) : pour ce faire les travaux existants [1] font appel à des représentations plus sophistiquées, qui visent à réduire la redondance (et donc l'espace mémoire) en tirant profit de plusieurs propriétés structurelles : entre autres la regularité ainsi que la *localité* et la *similarité*.

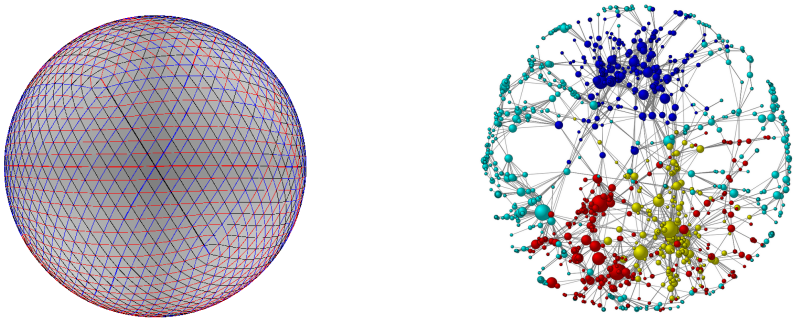

FIGURE 1. Gauche : exemple d'un maillage triangulaire 3D (induit avec une forêt de Schnyder). Droite : layout d'un réseau complexe.

Les travaux existants dans le domaine des structures de données compactes ont montré qu'il est possible de représenter des grands graphes avec des taux de compression très intéressants : peu de bits par lien suffisent pour coder des graphes, permettant aussi de répondre à des requêtes de navigation [1, 2]. Cependant pour la plupart de ces schémas de compression il n'existe pas d'analyses précises permettant d'évaluer de manière rigoureuse leurs performances et obtenir des garanties théoriques : en générale on ne se base que sur une évaluation expérimentale.

2. Objectifs du stage

Analyse adaptative des méthodes de compression pour les graphes. Il s'agit d'étudier de faire appel aux techniques d'analyses des réseaux complexes, et notamment d'utiliser les mesures de complexités (clustering coefficients, modularity, betweeness, centrality, ...) introduites pour identifier leurs propriétés structurelles.

L'objectif primaire du stage est d'analyser les représentations et structures de données existantes à l'aide de ces mesures de complexités : le but étant de fournir une caractérisation fine et rigoureuse de leur performance. On attaquera aussi le problème de concevoir (et éventuellement implémenter) de nouvelles représentations plus performantes, qui seraient plus adaptées pour certaines classes de graphes.

Références

[1] Paolo Boldi and Sebastiano Vigna. The webgraph framework I : compression techniques. In *Proc. of the 13th international conference on World Wide Web (WWW)*, pages 595–602, 2004.

[2] Cecilia Hernández and Gonzalo Navarro. Compressed representations for web and social graphs. *Knowl. Inf. Syst.*, 40(2) :279–313, 2014.