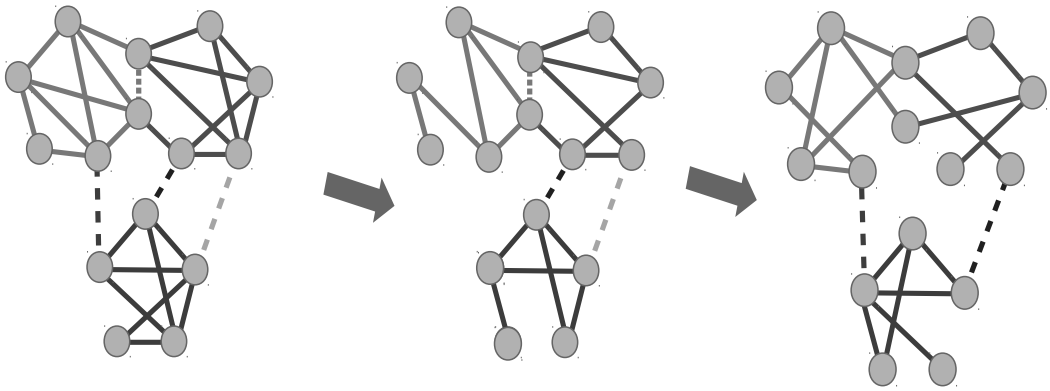
Algorithmique pour les flots de liens

Clémence Magnien, Matthieu Latapy

stages@complexnetworks.fr

http://complexnetworks.fr

LIP6 – CNRS et UPMC – Paris



De nombreux réseaux nous entourent : réseaux sociaux, graphe du web, réseaux d'interactions entre protéines, réseaux de co-occurrences de mots, échanges d'emails, etc. L'état de l'art a commencé à proposer des solutions pour analyser la structure de ces réseaux afin de pouvoir répondre à des questions comme : quels nœuds sont les plus importants ? le réseau peut-il être décomposé en groupes cohérents ?

Dans le cas où les réseaux sont dynamiques, l'approche habituelle consiste à agréger l'information sous forme de graphe. Pourtant, dans de nombreux contextes (trafic réseau, transactions financières, communications entre individus, etc), on est confrontés à des données sous forme de flots de liens : la donnée est essentiellement composée d'une série de triplets (A, B, t) indiquant que A a interagi avec B à l'instant t (par exemple, la machine A a envoyé un paquet à la machine B, ou le compte bancaire A a transféré de l'argent au compte B, la personne A a envoyé un message à la personne B, etc).

L'algorithmique pour les flots de liens a ceci de fondamental qu'il n'est pas question dans ce cas de calculer une propriété à chaque instant, mais bien de prendre en compte la nature intrinsèquement en flot des données. Un exemple type est la connexité : à un instant donné très peu d'emails sont échangés ; il n'y a donc pas de sens à calculer la connexité du réseau à cet instant ; par contre il est possible à une information d'être transférée par une succession d'emails. On voit émerger une notion d'accessibilité, et au-delà de connexité, dynamique.

Le but de cette thèse est de participer à la définition de notions fondamentales pour la description des flots de liens, et de proposer les algorithmes pour les calculer de manière efficace.

Le programme de travail sera organisé selon les directions suivantes :

- étudier les notions de l'algorithmique de graphe qui sont pertinentes pour les flots de liens et proposer une formalisation dans ce cadre (connexité, coupure, arbre couvrant, ...) ; l'équipe a déjà proposé une première définition pour la connexité, qui reste à approfondir ;
- proposer, prouver et implémenter des algorithmes pour calculer ces notions ;
- appliquer ces algorithmes à plusieurs jeux de données provenant de contextes différents, afin de s'assurer de la performance des algorithmes et de la pertinence des notions introduites ;
- une question particulièrement importante est le calcul à la volée (en streaming) des propriétés : est-il possible de faire le calcul en effectuant une seule passe sur les données, sans en stocker l'intégralité en mémoire ?

Ce sujet fait appel à des compétences en algorithmique des graphes, avec une visée empirique forte. Le but est double ; d'une part répondre à des questions théoriques importantes, comme par exemple quel peut être l'équivalent de la connexité dans le cas où la relation d'accessibilité n'est pas transitive ? et d'autre part de proposer et d'implémenter des algorihtmes permettant de traiter de très grands flots de liens (grands tant par le nombre de liens que la durée). Il faudra en particulier porter une attention particulière à l'espace mémoire, aucune approche matricielle n'étant applicable dans le cas de très grands graphes.

Ce projet utilisera des notions de programmation, de manipulation de donnée, de graphes, de statistiques,... On ne s'attend pas à ce que les candidats possèdent toutes ces compétences mais ils s'y formeront au cours du projet.