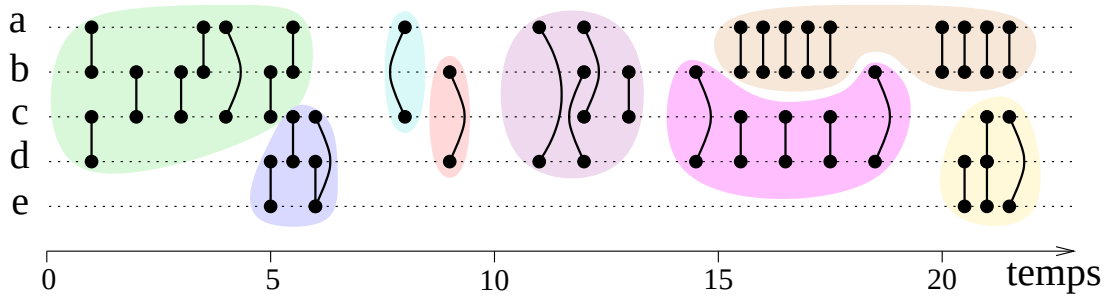
Conversations, Groupes et Communautés dans les Flots de Liens

Matthieu Latapy

stages@complexnetworks.fr

http://complexnetworks.fr

LIP6 – CNRS et UPMC – Paris



Un flot de liens est une suite de triplets (t, u, v) **indiquant que les entités** u **et** v **ont interagi à l'instant** t**.** Il peut s'agir d'échanges de messages (*emails* par exemple) entre individus, de transferts de paquets entre machines sur un réseau, d'achats en ligne par des clients, d'appels téléphoniques, ou encore de contacts entre individus observés par des capteurs. Les contextes pratiques où les données se modélisent naturellement comme des flots de liens sont extrêmement nombreux. Ces objets sont donc cruciaux pour de nombreuses applications, notamment pour un large spectre de questions de sécurité (attaques réseaux, connivences, fraudes, comportements malicieux, etc).

Dans tous ces contextes, **des sous-flots jouent des rôles particuliers**. Par exemple, dans des échanges de messages, des sous-structures de discussions émergent naturellement (comme des fils de discussion sur des listes de diffusion). Dans les contacts entre individus observés par des capteurs, on retrouvera des réunions d'amis ou collègues. Dans des appels téléphoniques ou des transferts de fichiers, on peut identifier une diffusion de rumeur ou d'information. Enfin, dans du trafic réseau, les échanges entre machines participant à une application distribuée (comme un système pair-à-pair ou un *botnet* par exemple) forment de tels sous-flots. La figure ci-dessus illustre la modélisation de ces diverses réalités par des flots de liens.

L'objectif central de ce projet est d'étudier les structures de sous-flots dans les flots de liens, et ce selon plusieurs axes.

En un premier temps, il s'agira d'**observer** de tels sous-flots dans des données où ils sont connus *a priori*. C'est en particulier le cas dans les archives de listes de discussions (*mailing-lists*), où les messages sont regroupés en fils de discussion (*threads*). Intuitivement, un tel fil est initié par un message posté à la liste par un utilisateur, et est constitué de l'ensemble des messages répondant à ce message initial, des réponses à ces messages, etc. Chaque message pouvant être vu comme une interaction entre son auteur et l'auteur

du message auquel il répond, un fil de discussion est bien un sous-flot particulier du flot représentant l'archive complète. L'étude de ces sous-flots permettra d'une part d'introduire les notions pertinentes pour leur description (comme leur densité, mais aussi leurs relations avec les autres sous-flots), et d'autre part de mieux comprendre la structure des sous-flots correspondant à des discussions.

Ceci permettra en un second temps de travailler à la **détection** de sous-flots particuliers dans un flot de liens où de tels sous-flots ne sont pas connus *a priori*. On reposera à cet effet sur les notions introduites dans la phase précédente, qu'on intègrera dans des fonctions de qualité qu'il s'agira d'optimiser afin de partitionner un flot en sous-flots. Ce travail s'inspirera de l'état-de-l'art déjà très riche sur la partition de grands réseaux en *communautés* (ensembles de nœuds densément connectés entre eux et faiblement connectés vers l'extérieur). Les cas dans lesquels les sous-flots sont connus a priori serviront à valider et étalonner l'approche, puis on pourra l'appliquer aux autres sous-flots. On obtiendra ainsi un outil de détection des sous-flots structurant un flot de liens. Nous l'appliquerons notamment aux cas pratiques cités ci-dessus, sur laquelle l'équipe a une expertise reconnue, afin d'obtenir de nouveaux éclairages sur ceux-ci. L'accent sera mis sur les applications sécurité, avec la détection d'attaques coordonnées dans des réseaux, la détection de diffusions de rumeurs, ou la détection de comportements malicieux.

Enfin, l'approche ci-dessus devrait permettre d'identifier des sous-flots dont la structure et la dynamique se démarque des autres, comme par exemple une discussion interrompue pendant un long moment puis redevenant active, des groupes qui se disloquent ou fusionnent, des applications réparties en panne, etc. L'étude de ces **cas déviants** permettra d'obtenir un éclairage plus fin sur les données considérées et les activités sous-jacentes, et aussi de reboucler sur les deux premières étapes afin de les affiner.

Au final, **le programme de travail que nous proposons est structuré comme suit** :

1. Étude d'archives de listes de diffusion afin de caractériser les sous-flots correspondant aux fils de discussions (introduction de nouvelles notions pour ce faire),
2. Conception de plusieurs fonctions de qualité susceptibles de capturer de façon pertinente ces caractéristiques, sous plusieurs points de vue,
3. Conception et implémentation d'algorithmes (au moins un algorithme glouton et une extension de l'algorithme de Louvain) recherchant une partition d'un flot en sous-flots maximisant ces fonctions de qualité,
4. Retour sur la pertinence des fonctions de qualité dans le cas des listes de diffusion (confrontation des résultats des algorithmes aux connaissances *a priori*),
5. Application à d'autres cas d'intérêt, en particulier le trafic réseau, les appels téléphoniques, et les contacts mesurés par capteurs, et interaction avec les applications,
6. Amélioration des fonctions de qualités et notions décrivant les sous-flots particuliers, et itération des points précédents,
7. Etude des cas déviants, c'est-à-dire les sous-flots incorrectement détectés, ou détectés mais ne correspondant pas à une réalité connue *a priori*.