

# Collecte et analyse de traces pair-à-pair pour la lutte contre la pédocriminalité

Matthieu Latapy et Clémence Magnien

Prenom.Nom@lip6.fr

Aujourd'hui, des millions de personnes échangent quotidiennement des millions de données par l'intermédiaire de systèmes pair-à-pair ou assimilés. Le trafic induit représente même la majorité du trafic total de l'internet, dépassant le web. Il est donc essentiel de bien comprendre les mécanismes sous-jacents (notamment pour proposer des protocoles efficaces), en particulier les comportements des utilisateurs et les propriétés des données.

Par exemple, quelle proportion des utilisateurs sont des *free riders* (téléchargent beaucoup de données mais n'en fournissent que très peu)? existe-t-il des corrélations entre les ensembles de fichiers téléchargés et/ou possédés par les utilisateurs? existe-t-il des communautés d'intérêt et peut-on les détecter? existe-t-il de même des ensembles de données *similaires*? des données très populaires? comment la popularité des données évolue-t-elle au cours du temps? ...

Collecter de l'information sur les échanges pair-à-pair est toutefois un défi en soi. En effet, la nature distribuée de ces échanges, la quantité de trafic induit, et l'anonymat de l'internet rendent la capture de traces extrêmement délicate.

Il existe principalement trois façons de collecter de telles traces d'usages. La première consiste en l'écriture d'un *client* pair-à-pair effectuant des requêtes afin de savoir quelles données sont proposées par les autres pairs. La seconde consiste en l'installation d'un pair de grande capacité traitant de nombreux échanges, et les enregistrant. La troisième enfin consiste en la capture de trafic directement sur les routeurs et chez les FAI.

Ce stage se situe dans la première approche. Il s'agit, sur la base de composants logiciels déjà développés, de mettre au point un outil distribué effectuant des requêtes automatisables sur le réseau P2P eDonkey. L'outil devra ensuite être déployé sur une centaine de machines de mesure distribuées dans le monde.

La collecte de données sur les échanges pair-à-pair a pour principale application une meilleure connaissance des propriétés des échanges effectués, notamment pour proposer des protocoles efficaces en tirant parti. Dans le cadre de ce stage, l'objectif applicatif est plus précisément la lutte contre la pédocriminalité.

Une deuxième composante du stage (complémentaire de la première ou prenant le pas sur celle-ci) consiste donc en l'analyse des données collectées pour évaluer précisément et rigoureusement la quantité d'échanges à caractère pédophile observés. Comme les identifiants de pairs sont réalloués (un même utilisateur change d'identifiant, et plusieurs utilisateurs peuvent avoir le même identifiant), c'est une problématique non-triviale. De plus, la taille et la complexité des données à manipuler soulèvent en elles-mêmes de nombreux défis.