Détection d'Attaques dans des Traces de Trafic Réseau

Matthieu Latapy

stages@complexnetworks.fr

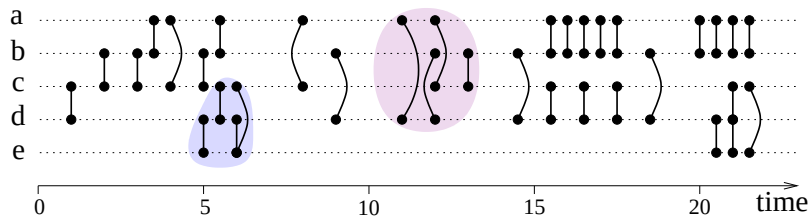
http://complexnetworks.fr

LIP6 – CNRS et UPMC – Paris

Les attaques contre les services en ligne, les réseaux, les systèmes d'information, et les usurpations d'identité ont un coût annuel en milliers de milliards d'Euros et causent de nombreuses faillites. Elles affaiblissent également la confiance des utilisateurs et ralentissent donc les progrès de l'ère numérique.

Un des principaux moyens de lutte contre ces attaques consiste en l'**observation du trafic** acheminé par le réseau et/ou reçu par les serveurs, en la détection du trafic induit par les attaques et en l'analyse de ce trafic (pour le filtrer ou en trouver l'origine).

Les progrès en la matière sont aujourd'hui sévèrement limités par le **manque de méthodes et d'outils pour analyser les interactions au cours du temps**. En effet, le trafic réseau peut être vu comme une suite d'échanges entre machines au cours du temps, et une attaque est alors une séquence d'interactions particulières dans cette suite. Voir figure ci-dessous pour une illustration.



Du trafic réseau représenté comme flot de liens : les machines sont a, b, c, d et e, et chaque lien représente un paquet échangé. Par exemple, c et d ont échangé un paquet à l'instant 1, b et c à l'instant 2, etc. Les zones colorées peuvent indiquer des événements anormaux à détecter (séries d'échanges particulièrement denses, par exemple).

Nous proposons ici de voir le trafic comme une suite de liens (t, u, v) indiquant le fait que les machines u et v ont échangé un paquet à l'instant t. Nous souhaitons développer un ensemble de **notions permettant de décrire ces flots de liens** de façon similaire à ce que la théorie des graphes permet avec les réseaux : densité, degrés, chemins, centralité, ... Le premier défi est de définir sur les flots de liens ces notions classiques en théorie des graphes.

Calculer ces propriétés sur des traces réelles posera des problèmes délicats d'efficacité : les volumes de données sont énormes, typiquement des millions de liens par minute de trafic. Les contraintes en temps et en mémoire nécessaires sont alors cruciales, et des approches originales (comme du calcul en *streaming*) sont à explorer.

L'objectif est d'utiliser les propriétés de flots de liens définies et calculées ci-dessus pour **détecter des anomalies dans du trafic réseau**. Ceci se fera sur le terrain par l'observation manuelle de leurs distributions statistiques, par l'utilisation d'outils de fouille de données (comme *scikit-learn*) et la confrontation à des situations où des événements sont déjà connus.

Ce projet utilisera des notions de réseaux, de programmation, de manipulation de données, 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.

Un travail préliminaire sur ce sujet a été publié récemment : `http://tinyurl.com/netscicom`