Analyse de Malware

Matthieu Latapy

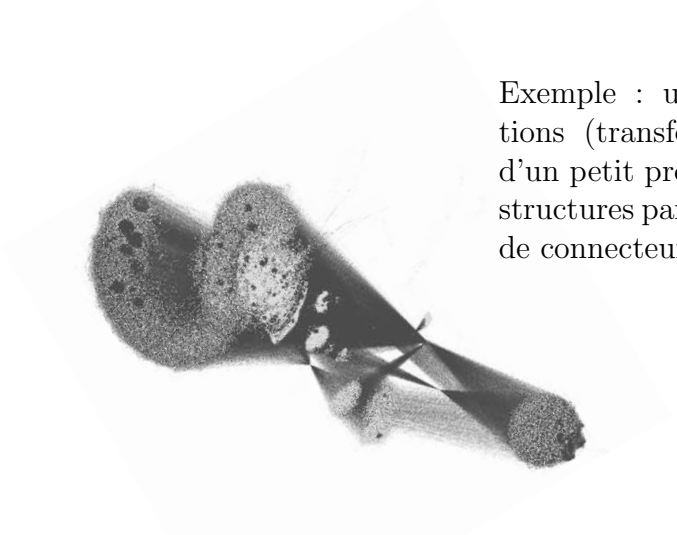
stages@complexnetworks.fr

http://complexnetworks.fr

LIP6 – CNRS et UPMC – Paris

La plupart des actes malveillants perpétrés sur l'internet de nos jours reposent sur l'usage de divers malwares (comme des virus ou des vers). Par conséquent, il est crucial de mieux comprendre ces logiciels très particuliers et de développer des moyens de lutte appropriés. Toutefois, ils intègrent la plupart du temps de nombreuses techniques subtiles pour empêcher leur analyse, comme le cryptage du code, de l'auto-modification, ou encore la détection de leur environnement d'exécution et le lancement d'actions malveillantes seulement en environnement sûr (pour eux).

Face à ces obstacles, les chercheurs développent des méthodes évoluées de capture et d'analyse des malware, voir par exemple http://tinyurl.com/lip6malware On voit notamment dans cet article une vidéo illustrant comment un code en cours d'exécution peut être modélisé par un graphe.



Exemple : un graphe représentant les affectations (transferts mémoires) lors de l'exécution d'un petit programme. On peut repérer des sous-structures particulières et des nœuds jouant le rôle de connecteurs entre ces structures.

Ces représentations ne capturent toutefois pas la dynamique temporelle de l'exécution, qui est essentielle pour comprendre le code sous-jacent et l'activité malveillante du logiciel. Nous proposons ici de représenter l'exécution d'un malware par un **flot de liens,** c'est-à-dire une suite de triplets (t, u, v) indiquant qu'à l'instant t une opération (par exemple un échange de données) a eu lieu entre la partie de code u et la partie v. Cette représentation a l'avantage de capturer à la fois la structure des exécutions (comme les graphes) et leur dynamique temporelle.

Dans ce projet nous utiliserons des traces d'exécution de malware produites par les collègues du LORIA avec lesquels nous travaillons (cf article ci-dessus). La **modélisation de ces traces par des graphes ou des flots de liens** pose de nombreuses questions : quels niveaux de granularité considérer, doit-on représenter les instructions et/ou les espaces mémoire, quelle notion de temps considérer, comment gérer le parallélisme ou le non-déterminisme, ... ? Une fois cette modélisation effectuée, il s'agira de **décrire les flots de liens obtenus d'un point de vue à la fois structurel et temporel**. Ceci reposera sur l'adaptation aux flots de liens de plusieurs notions classiques sur les graphes : degrés, densité, chemins, communautés, centralités, etc. Il s'agira ensuite de calculer ces propriétés sur des traces d'exécution de malware et une interprétation des structures et dynamiques observées. On s'attachera en particulier à **identifier des modules** dans les logiciels sous-jacents, correspondants à des fonctionnalités comme de la cryptologie, des communications, ou diverses actions malveillantes.