Unité et équipe : Loria UMR 7503
Nom de l'encadrant : Jean-Yves Marion, Professeur Université de Lorraine et membre de de l'Institut Universitaire de France.
Courriel : Jean-Yves.Marion@loria.fr

Directeur du laboratoire : Jean-Yves Marion

Titre du stage

Analyse du flot de données dans les codes binaires malveillants. Etude de la cartographie des fonctionnalités et leurs corrélations.

Problématique Générale.

Résumé

L'équipe Carbone du Loria (www.loria.fr) UMR 7503, dans le cadre du Laboratoire de Haute Sécurité (LHS), a développé une méthode originale, dite par ***analyse morphologique***. Cette méthode permet de détecter des similarités dans des codes. Il est ainsi possible de détecter une fonctionnalité particulière dans un code ou de détecter un code malveillant. L'objectif de ce stage est de reconstruire le graphe de flot d'information pour reconstituer la cartographie des fonctionnalités utilisées dans un code malveillant.

Le contexte

Les compagnies d'Anti-Virus sont assez discrètes sur les méthodes de détection employées. Ceci dit, la technique classique de détection de code malveillant est d'attribuer à chaque code malveillant une signature qui caractérise le malware en question. Une signature est une expression régulière, souvent réduite à une simple suite d'octets, qui identifie un malware (voir règle Yara). Tout fichier/binaire ou code exécuté en mémoire contenant cette signature est alors considéré infecté. Etant donné une base de données de signatures, le moteur de détection est alors la partie du logiciel chargée de rechercher une de ces signatures à l'intérieur des fichiers et des programmes. L'avantage de cette approche est sa rapidité et le faible taux de faux-positif. Les inconvénients tiennent essentiellement dans le fait que cette technique n'est pas capable de détecter des variantes ou des mutations d'un code malveillant connu, et donc elle est a fortiori incapable d'identifier une nouvelle attaque. A ce défaut originel, il s'ajoute dorénavant un défaut supplémentaire majeur. Aujourd'hui, *l'industrie du malware* est bien organisée et elle est capable de générer massivement des variantes de codes malveillants par différents procédés d'obfuscation qui échappent aux protections classiques. Ainsi, il faut faire face à des dizaines de millions d'échantillons à analyser régulièrement. Il est clair qu'une analyse manuelle est devenue impossible et qu'il est nécessaire d'avoir des outils automatiques d'analyse et de classification. La première difficulté importante est donc d'avoir des outils suffisamment robustes et efficaces pour traiter de large corpus de programmes à analyser.

Programme de stage

L'objectif principal est d'identifier une fonctionnalité dans un code binaire. L'approche envisagée est de reconstruire le graphe de flot d'information d'un code (protégé/malveillant). La principale difficulté est que les codes sont obfusqués et auto-modifiants. Sans l'accès au code source et en présence d'obfuscation, la reconstruction du graphe de flot d'information passe par une combinaison d'analyse dynamique et d'analyse statique. Ensuite, il s'agira de voir comment un graphe de flot de contrôle permet de discriminer certaine partie du code. Pour cela, il faudra échantillonner les graphes de flot de données en sous graphes enrichis d'information sémantique et voir comment un ensemble de sous graphes peut caractériser une fonctionnalité. La détection pourrait utiliser un modèle prédictif basé sur des outils d'apprentissage.

Déroulement et poursuite du Stage

Les étapes du stage sont les suivantes :

1. Le stagiaire se familiasera aux différents concepts impliqués au sein de l'équipe. Il aura accès à tous les outils de développement nécessaires.
2. Le stagiaire devra réaliser un protype d'extraction de graphe de flot d'information pour des codes malvaillants.
3. Dans la mesure du possible, un aglorpimthe de reconnaissance de fonctionalités sera dévelopé et validé expérimentalement

Ce stage pourra se poursuivre soit en doctorat, soit comme ingénieur dans la start-up Cyber-Detect.