Stage : Définition et évaluation de la qualité des datasets pour l'évaluation d'outils de cybersécurité

Direction de stage : Philippe Owezarski – LAAS-CNRS (owe@laas.fr)

Sujet :
La cybersécurité repose aujourd'hui de plus en plus sur des outils utilisant des techniques d'intelligence artificielle (IA), notamment d'apprentissage automatique. Pour fonctionner, ces techniques ont besoin d'un entrainement sur des données labelisées, les labels indiquant la nature des paquets et des flux du dataset : bénin et son type ou attaque et son type. Pour un bon entrainement des outils d'apprentissage automatique, les datasets doivent donc contenir une grande variété de trafics différents avec des attributs les caractérisant couvrant un large spectre de situations. L'objectif est de disposer de datasets permettant une évaluation complète et pertinente des outils de cybersécurité en évitant des erreurs liées à des phénomènes comme le sous- ou sur-apprentissage, par exemple.
Ce stage a pour objectif, à partir de l'étude des datasets publics actuels et des résultats obtenus par des outils de détection d'attaques sur ces datasets, de les caractériser. Pour cela, il faudra identifier les attributs caractéristiques permettant de distinguer le trafic bénin du trafic des attaques, et des attaques entre elles. Il faudra également définir les métriques permettant de caractériser et quantifier des éléments comme la variété, la variabilité, le niveau de couverture, etc.