Sujet de thèse : Génération de graphes de causalité

Contexte

Cette thèse aura lieu à l'Institut Polytechnique de Paris/Telecom SudParis, Palaiseau, dans le cadre du PEPR « Superviz ».

Introduction

La supervision de sécurité reste aujourd'hui, malgré des avancées remarquables, très difficile à mettre en œuvre. Investiguer des attaques dans des logs reste un processus fastidieux qui nécessite une analyse fine et longue. L'automatisation de ces activités est donc un sujet sur lequel il est nécessaire de s'attarder.

Les SIEMs (Security Information and Event Management) restent les outils les plus utilisés afin de traiter les logs de façon généralement centralisée. Ces outils reposent essentiellement sur des mécanismes de corrélation permettant de détecter automatiquement la présence d'attaques dans des logs. Ces mécanismes reposent sur la génération de règles de corrélation qui expriment comment une attaque se présente sous forme de suite d'événements (réseau, système, applicatifs). Le SIEM génère alors une alerte lorsqu'une de ces suites est détectée.

Toutefois, la génération de ces règles de corrélation reste difficile. Il est facile d'avoir des règles incorrectes ou incomplètes, générant ainsi de nombreuses fausses alertes (faux positifs), ou au contraire loupant des attaques (faux négatifs).

Travaux de thèse

Les travaux vers lesquels se dirige cette thèse visent à automatiser la découverte précise de règles de corrélation. Ces travaux font suite aux travaux de Charles Xosanavongsa [1]. L'idée est de générer dans le système des graphes de causalité ([2], [3]) entre les événements de logs. Ces graphes représentent l'enchainement logique des événements dans le système. Dans le cas d'une attaque, ces graphes contiennent les événements significatifs de l'attaque. Écrire la règle de corrélation consiste à remonter depuis un point d'intérêt (indice de compromission) et retrouver dans le graphe la séquence d'événements qui conduit à la génération de l'attaque.

Construire ces graphes n'est pas un exercice facile : il nécessite soit d'instrumenter le système pour détecter des causalités comme dans [4], afin de pouvoir tagger les événements système, réseaux ou applicatifs, soit de réaliser une approximation à partir des événements de logs eux-mêmes afin de retrouver la cause de leur existence et d'en déduire la causalité entre ces événements. C'est l'approche suivie par [5].

L'objectif de cette thèse est de suivre au contraire la première approche en instrumentant le système afin de construire les graphes de causalité des événements à partir des actions réalisées dans le système. Une fois ces graphes obtenus, des algorithmes pourront être implémentés pour détecter des signatures d'attaques dans les graphes. Ces algorithmes pourront par exemple extraire ou détecter des sous-graphes d'intérêt qui permettront de mettre en évidence l'attaque dans le graphe.

Nous envisageons d'utiliser les techniques récemment développées dans le domaine du « graph learning » [7] (GCN Graph Convolutional Networks) afin de caractériser plus finement les relations causales au niveau des évènements système. Plus précisément, nous souhaitons développer un algorithme se basant sur les réseaux neuronaux convolutif afin d'automatiser la détection de motifs caractéristiques [6] d'attaques dans les graphes de causalité d'évènements système. De plus, en intégrant directement l'aspect temporel des évènements présent dans le graphe de corrélation dans le réseau de neurones [8] nous allons pouvoir plus facilement caractériser/détecter la dynamique de certaines attaques.

Contacts

Directeur de these: Eric Totel : eric.totel@telecom-sudparis.eu
Co-directeur : Vincent Gauthier: vincent.gauthier@telecom-sudparis.eu

Bibliographie

[1] Charles Xosanavongsa, Eric Totel, et Olivier Bettan, « Discovering Correlations: A Formal Definition of Causal Dependency Among Heterogeneous Events », in *Proceedings of the 4th IEEE European Symposium on Security and Privacy*, Stockholm, Sweden, juin 2019.
[2] L. Lamport, « Time, clocks and the ordering of events in a distributed system », *Communications of the ACM*, vol. 21, p. 558-565, juill. 1978.
[3] B. d'Ausbourg, « Implementing secure dependencies over a network by designing a distributed security subsystem », in *Proceedings of the Third European Symposium on Research in Computer Security (ESORICS'94)*, 1994, p. 247-266.
[4] A. M. Bates, D. Tian, K. R. Butler, et T. Moyer, « Trustworthy Whole-System Provenance for the Linux Kernel. », in *USENIX Security Symposium*, 2015, p. 319-334.
[5] C. Xosanavongsa, « Heterogeneous Event Causal Dependency Definition for the Detection and Explanation of Multi-Step Attacks », phdthesis, CentraleSupélec, 2020. Disponible sur: https://theses.hal.science/tel-02947368
[6] Ying, Z., You, J., Morris, C., Ren, X., Hamilton, W., & Leskovec, J. (2018). Hierarchical graph representation learning with differentiable pooling. Advances in neural information processing systems, 31.
[7] Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In International Conference on Learning Representations (ICLR).
[8] Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., & Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs.