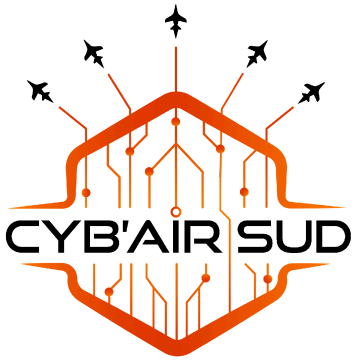


Talk Proposal

For Cyb'Air Sud 2025

Title: 10 Years of Large-Scale Malware Code Comparison: A Retrospective

Speaker: Tristan Pourcelot

Short Description

How to hunt for malware variants across a large corpus?

Abstract

This presentation explores techniques for identifying malware variants within a corpus of several hundred thousand executables. We will begin by introducing the concept of malware code comparison and the challenges encountered by the authors throughout their careers. An overview of existing techniques will follow, with a focus on the Machoc comparison algorithm, based on the condensed control flow graph of an executable. Published at SSTIC in 2016, this algorithm provides an algorithmic fingerprint of a program, enabling fast and resilient code comparison even in the presence of minor changes. We will present the strengths and weaknesses of this algorithm, as well as the challenges faced and solutions developed to scale it to large malware corpora. Finally, we will briefly introduce our internal algorithm, Zubat, which is based on an intermediate representation of executable code. It allows for more precise similarity searches and the ability to identify similar functions individually within the corpus. The presentation will include demonstrations on malware samples linked to advanced threat actors.

Language of the Talk

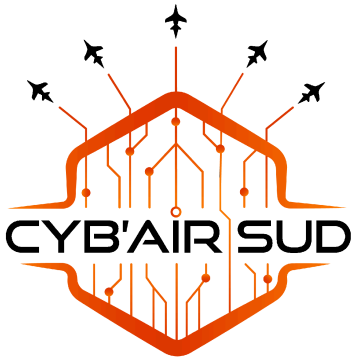
French

TLP Classification

TLP-WHITE

Recording and Replay

Yes



Proposition de conférence

Pour Cyb'Air Sud 2025

Titre : Retour d'expérience sur 10 ans de comparaisons à large échelle de codes malveillants

Intervenant : Tristan Pourcelot

Description concise

Comment chasser des variants de codes malveillants parmi un large corpus ?

Résumé détaillé

Dans cette présentation, nous étudierons des techniques permettant de trouver des variants de codes malveillants parmi un corpus de plusieurs centaines de milliers d'exécutables. Nous introduirons d'abord ce qu'est la comparaison de code malveillants, ainsi que les problématiques liées que les auteurs ont pu rencontrer au cours de leur carrière. Nous effectuerons ensuite un état de l'art des techniques existantes, puis nous ferons un focus sur l'algorithme de comparaison Machoc, basé sur le condensant du graphe de flot de contrôle d'un exécutable. Cet algorithme, publié au SSTIC en 2016, permet d'obtenir une empreinte algorithmique du programme, elle offre une capacité de comparaison de codes très rapide et résistante à des changements mineurs. Nous présenterons les forces et les faiblesses de cet algorithme, puis les problématiques rencontrées et les résolutions de ces verrous pour faire passer cet algorithme sur un large corpus de codes malveillants. Enfin, nous présenterons brièvement l'algorithme que nous utilisons en interne, Zubat, basé sur une représentation intermédiaire du code exécutable, qui permet d'effectuer des recherches de similarité plus précises et de rechercher unitairement des fonctions similaires au sein du corpus. Cette présentation sera illustrée de démonstrations sur des codes malveillants liés à des attaquants avancés.

Langue de la présentation

Français

Classification TLP

TLP-WHITE

Accord concernant l'enregistrement et la rediffusion

Oui