LID-TP1

Chédotal Corentin - Jouvance Alexis

Objectif du projet :

Réaliser un outil capable de détecter les sites de hameçonnage sur les 100 sites les plus visités.

Notre analyse se portera sur :

- Des variations d'url (comparaison avec les URL de la Trust List)
- La date de réservation du nom de domaine
- Les informations whois
- La correspondance avec des termes souvents utilisées pour du phishing.

Le point de départ du projet est la réalisation d'une Trust List des 100 sites les plus visités d'interne$_1$.
Nous allons ensuite comparer différences de nom de domaines entre les domaines de notre Trust List avec des domaines provenant d'un flux d'enregistrement de certificats. Pour obtenir ce flux nous utiliserons le projet **certstream$_2$.** Si un domaine provenant du flux à un nom de domaine assez proche d'un des domaine de notre Trust List nous passons à l'étape suivante de l'analyse.
Si un domaine est retenu par notre premier filtrage, nous allons faire une requête **whois$_3$** sur chacun des domaines afin de pouvoir les comparer.
En parallèle, nous comparerons les domaines provenant du flux avec une liste de termes sensibles$_4$.

Sequence Matcher

Dans un premier temps, il nous faut définir le périmètre des urls analysés. Afin de savoir une url correspond à notre périmètre nous allons nous servir de la librairie SequenceMatcher**.** Cette librairie nous donnes un score de correspondance entre 2 chaînes de caractères.
Notre choix du seuil d'analyse sera :

- Si l'url donné à une correspondance < 70% on ne l'analyse pas
- Si l'url donnée à une correspondance de ≥ 70% on l'analyse.

Suspicious word

Suite au passage dans le sequence matcher, si une url est retenue, on la compare avec une base de mot sensible. Chacun des termes de la base à un score qui lui est appliqué en fonction de la dangerosité potentielle de l'url.
Nous enregistrons aussi la date de capture, afin de la spécifier dans la fiche d'indicateur.

Test des sous-domaines et des tirets

La présence d'un grand nombre de sous domaine dans une url ou d'un grand nombre de tiret va faire augmenter le grade de suspicion d'une URL. En effet, l'ajout de tiret pour remplacer des points, ou l'ajout de sous domaines sont des techniques courante lors de la création d'un site de phishing.

Grade de l'URL

Suite à tous nos test, on calcule la note définitive de notre URL. Si la note de l'URL est supérieure ou égale à 20 on passe à la requête whois.

Whois

Si une url reçoit un score d'au moins 20 on exécute alors une requête whois sur notre url capturée afin d'obtenir des informations telles que :

- Le nom de domaine enregistré,
- La date de création du nom de domaine,
- La date d'expiration du nom de domaine.

Ces informations seront ajoutés à la création de l'indicator STIXv2 de l'URL.

STIXv2

Les informations recueillies seront mises au format STIXv2 afin de créer des indicateurs.

Références

1: Liste des 100 sites les plus visités : https://en.wikipedia.org/wiki/List_of_most_popular_websites *source wikipedia*
2: Projet certstream : https://certstream.calidog.io/
3: Librairie python-whois : https://pypi.org/project/python-whois/
4: Liste de termes sensibles : https://github.com/x0rz/phishing_catcher/blob/master/suspicious.yaml *phishing_catcher de x0rz*