

Projet de Sécurité des réseaux

ETUDE DU CONTENU DE PASTEBIN
MAËL CLESSE

UNIVERSITE DE LORRAINE

Table des matières

Introduction.....	2
Le script	2
Utilisation	3
Analyse du contenu de Pastebin	4
Github	4
Cartes bancaires	5
Emails	7
Films.....	9
Pornographie.....	11
Conclusion	12

Introduction

Un pastebin est une application web qui permet à ses utilisateurs d'y déposer des morceaux de texte pour qu'ils puissent être lus publiquement. Ces textes sont bien souvent des codes sources de logiciels sous licence libre, mais parfois, il peut s'agir de tout autre chose.

Il existe un grand nombre de pastebin à travers le web. Cependant, j'ai décidé de focaliser mon attention sur le plus connu d'entre eux à ce jour, à savoir Pastebin.com.

Ce site anglais, qui a vu le jour en 2002, sert donc en quelque sorte de moyen d'expression public à des millions d'utilisateurs chaque jour. Et étant donnée la variété d'utilisateurs potentiels, le contenu qu'il renferme est, quant à lui, tout aussi varié, bien loin de la publication de simples codes sources.

Pour ce projet, j'ai décidé de me concentrer sur l'exploitation d'archives répertoriant tous les pastes publics du 23 Février 2015. Les informations sur lesquelles se basent ces études étaient donc tout à fait consultables directement sur <http://pastebin.com> le jour de leur parution.

Le script

Pour procéder à l'analyse du contenu de cette journée de pastes, j'ai donc mis au point un script Shell qui permet à l'utilisateur de choisir les informations qu'il souhaite extraire des différentes archives qu'il possède.

Ce script fonctionne donc, dans le cas présent, pour parcourir le contenu de Pastebin, mais il pourrait également servir dans n'importe quel autre contexte, du moment qu'il s'agirait de recherche d'information.

```
#!/bin/sh

#Auteur : Maël CLESSE

#Script shell permettant de rechercher une expression dans un ou plusieurs fichiers

#Pour le projet, le but est de l'utiliser pour essayer de découvrir les données sensibles qui pourraient
#être récupérées sur le site Pastebin au cours d'une journée classique

read -p "Saisir le chemin vers le(s) fichier(s) à analyser " path

while true; do

    read -p "Saisir le fichier dans lequel stocker les résultats " fichier

    read -p "Saisir l'expression à rechercher " exp

    #on crée une en-tête pour le fichier où seront stockés les résultats
```

```
echo " -----" >> $fichier

echo "| Analyseur de fichiers |" >> $fichier

echo "| Créé par Mael Clesse |" >> $fichier

echo " -----" >> $fichier

echo " " >> $fichier

#on recherche l'expression dans tous les fichiers du path et on écrit les résultats dans le
#fichier de destination

zmore $path | grep $exp >> $fichier

echo " " >> $fichier

echo "Nombre d'occurrences de \"$exp\" dans le(s) fichier(s) : " >> $fichier

grep -o $exp $fichier | wc -l >> $fichier

echo "Vous pouvez consulter les resultats de la recherche dans le fichier \"$fichier

done
```

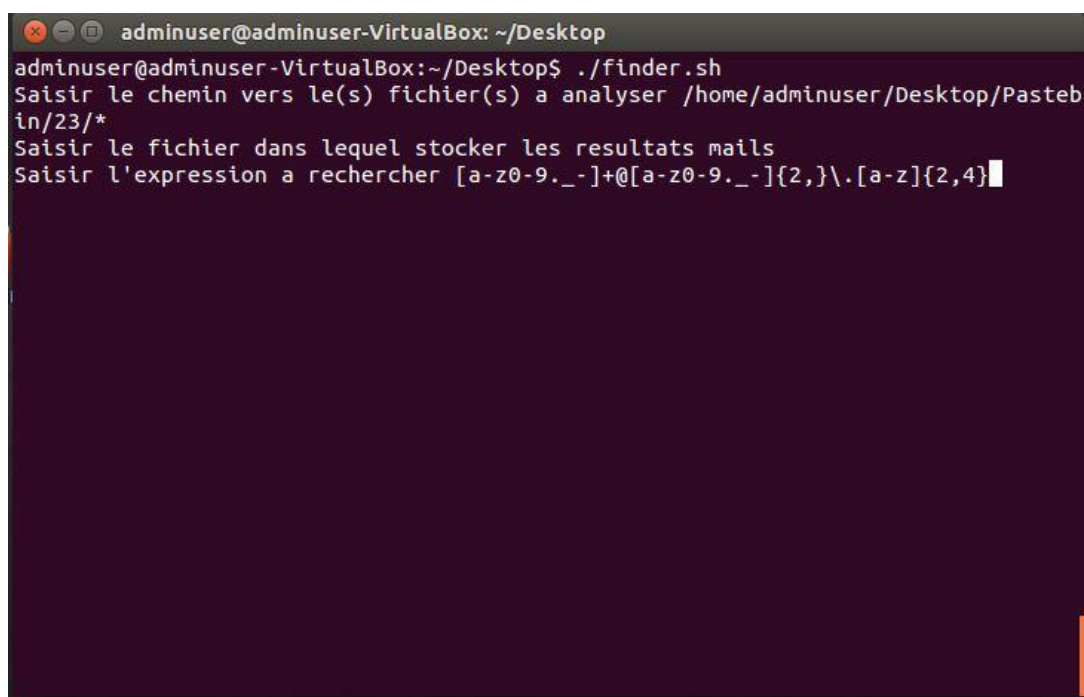
Le script demande tout d'abord à l'utilisateur où se situent les fichiers qu'il va devoir analyser. Ensuite, une boucle se met en place, permettant ainsi d'analyser de plusieurs manières différentes les mêmes fichiers (avec des expressions différentes par exemple).

L'utilisateur va donc maintenant choisir où stocker les résultats de l'analyse, et également les mots qu'il souhaite voir recherchés dans les fichiers fournis en paramètres. Ce sont ces mots qui vont être recherchés par le script.

Utilisation

Cette première version utilise la fonction *grep* pour rechercher l'expression que l'utilisateur entre au clavier. Elle permet donc de chercher des mots, mais pas des expressions régulières. Une autre version du script utilisant la fonction *egrep* a parfois été utilisée pour rechercher des expressions régulières dans les pastes, comme l'expression régulière « `[a-z0-9._-]+@[a-z0-9._-]{2,}\.[a-z]{2,4}` » qui correspond à la recherche de toutes les adresses emails existantes et valides.

Cette expression sert d'ailleurs dans l'exemple ci-dessous



```
adminuser@adminuser-VirtualBox: ~/Desktop
adminuser@adminuser-VirtualBox:~/Desktop$ ./finder.sh
Saisir le chemin vers le(s) fichier(s) a analyser /home/adminuser/Desktop/Pastebin/23/*
Saisir le fichier dans lequel stocker les resultats mails
Saisir l'expression a rechercher [a-z0-9._-]+@[a-z0-9._-]{2,}\.[a-z]{2,4}
```

Capture d'écran de l'utilisation du script Shell pour rechercher des adresses email valides

Une fois le script lancé, il va donc parser tous les fichiers passés en paramètre à la recherche de la dite expression.

Analyse du contenu de Pastebin

Comme dit précédemment, le but de ce projet n'était pas simplement d'écrire ce script, mais bel et bien de s'en servir pour étudier le contenu d'une journée normale sur le site <http://pastebin.com>.

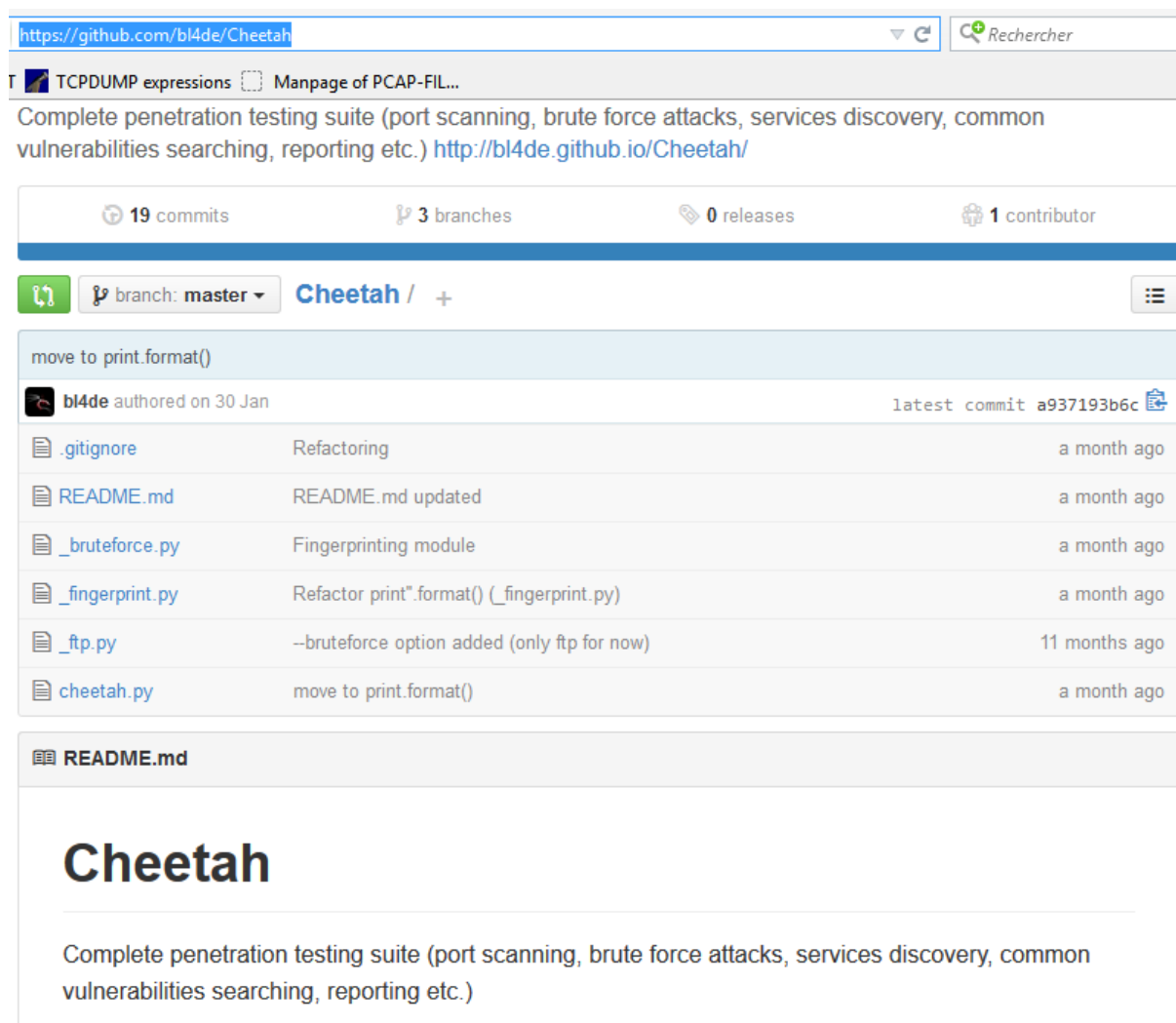
J'ai donc réalisé quelques recherches, d'abord par mot-clé, puis à l'aide d'expressions régulières, afin de mettre à jour des contenus publiés sous le statut public, alors qu'ils contiennent en vérité des informations qui n'auraient jamais dû se retrouver à la vue de tout le monde.

Github

Tout d'abord, la recherche que j'ai effectuée portait sur la découverte des liens Github disponibles. Puisqu'à l'origine, Github et Pastebin avaient, semble-t-il, le même objectif commun, à savoir promouvoir du code libre, il y avait de fortes chances pour que j'obtienne de nombreux résultats.

Cette recherche, qui paraît donc à première vue totalement banale, révèle qu'en effet, des gens partagent des liens Github sur Pastebin. Mais ces liens ramènent essentiellement vers des projets établis dans le but de scanner des applications, sniffer des réseaux, ...

Un exemple parmi tant d'autre se trouve sur [ce Github](https://github.com/bl4de/Cheetah) qui contient un outil permettant de réaliser des tests de pénétration complets (scan de ports, attaques par brute force, scanner de vulnérabilités, ...).



https://github.com/bl4de/Cheetah

Rechercher

TCPDUMP expressions Manpage of PCAP-FIL...

Complete penetration testing suite (port scanning, brute force attacks, services discovery, common vulnerabilities searching, reporting etc.) <http://bl4de.github.io/Cheetah/>

19 commits 3 branches 0 releases 1 contributor

branch: master Cheetah / +

move to print.format()

bl4de authored on 30 Jan latest commit a937193b6c

File	Commit Message	Time
.gitignore	Refactoring	a month ago
README.md	README.md updated	a month ago
_bruteforce.py	Fingerprinting module	a month ago
_fingerprint.py	Refactor print".format() (_fingerprint.py)	a month ago
_ftp.py	--bruteforce option added (only ftp for now)	11 months ago
cheetah.py	move to print.format()	a month ago

README.md

Cheetah

Complete penetration testing suite (port scanning, brute force attacks, services discovery, common vulnerabilities searching, reporting etc.)

Vu de la page d'accueil du Github précédemment cité

Pour trouver cette liste, j'ai simplement recherché le mot-clé « github » dans toutes les archives regroupant les pastes. Le fichier produit pèse 557ko et contient exactement 1026 occurrences du mot. Ces références ne sont cependant pas toutes liées à des gits présents sur le site, mais parfois, il s'agit de dossiers présents sur les PC d'utilisateurs.

Après cette première recherche, j'ai décidé d'en lancer une nouvelle, qui visait un type d'informations un peu plus sensibles.

Cartes bancaires

La deuxième recherche effectuée portait donc sur l'éventualité que des numéros de cartes de crédit aient pu se retrouver postés sur Pastebin, certainement à l'insu de leurs propriétaires. Cette

possibilité m'est venue à l'esprit, du fait que depuis quelques temps déjà, certaines personnes, peu soucieuses de la sécurité de leurs données bancaires, n'hésitent pas à poster des photos de leurs nouvelles cartes bancaires [sur Internet](#).

J'ai donc interrogé les pastes avec le mot clé Visa cette fois-ci. Malheureusement, aucune information sensible ne m'a été retournée. Dommage ! Toutefois, j'ai pu trouver quelques informations que je n'aurai jamais pensé voir sur un site accessible de tous.

En effet, sur un des pastes, j'ai pu découvrir ce tableau.

```
1 us (visa, master) = 5 $  
  
-----master and visa bin-----  
  
sell visa debit us : 120$  
mastercard standart, visa classic - $25  
visa gold|platinum|corporate|signature|business $45  
mastercard, visa classic - $30  
visa gold|platinum|corporate|signature|business $55  
mastercard, visa classic - $90  
visa gold|platinum|corporate|signature|business $130  
mastercard| visa classic - $70  
visa gold|platinum|corporate|signature|business $90  
mastercard standart, visa classic - $30  
visa gold|platinum|corporate|signature|business $50  
mastercard, visa classic - $40  
visa gold|platinum|corporate|signature|business $60  
mastercard, visa classic - $100  
visa gold|platinum|corporate|signature|business $160  
mastercard| visa classic - $100  
visa gold|platinum|corporate|signature|business $120
```

Il s'agit, à première vue, des tarifs que pourrait pratiquer un organisme bancaire lorsqu'il délivre chaque type de cartes bancaires existant à ses clients. Mais en fait, il n'en est rien. Dans ce paste, il s'agit en fait des tarifs auxquels on peut obtenir, illégalement, des numéros de cartes bancaires en fonctionnement sur le Darknet. L'utilisateur de Pastebin qui a publié ces informations semble donc être assez mal intentionné ...

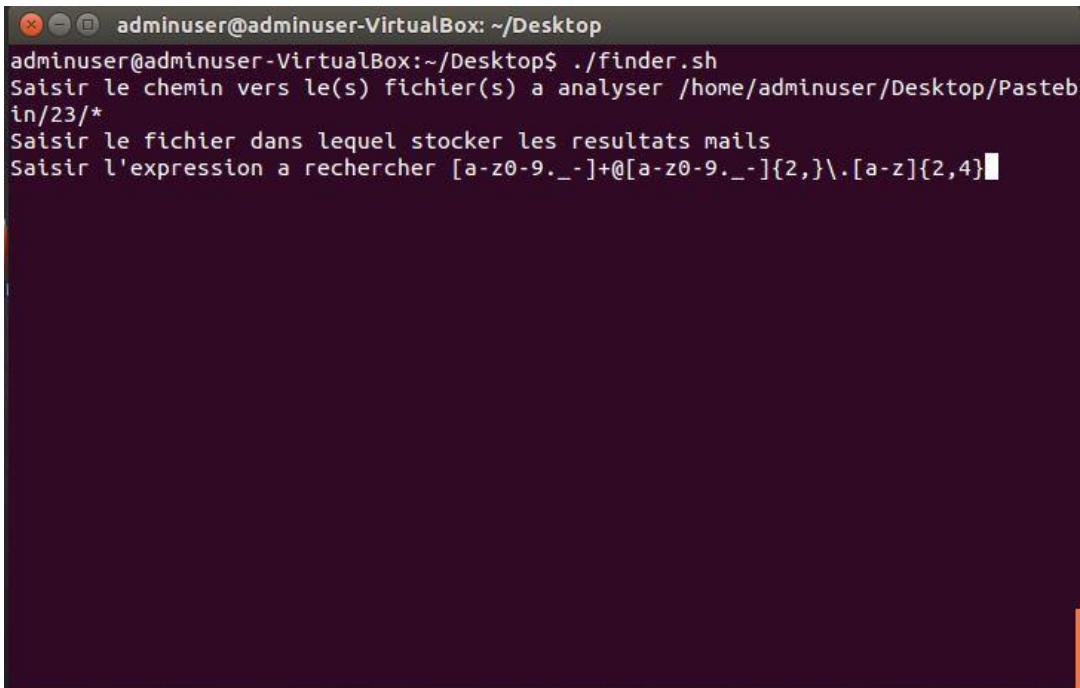
Emails

Pour revenir dans un registre un peu moins illégal, ma troisième recherche portait sur l'obtention d'adresses email. Dans la société actuelle, l'information représente une monnaie d'échange considérable. Une personne possédant des adresses email valides (comprenez qui correspondent réellement à une personne physique) peut les revendre à des personnes dans le milieu de la publicité en ligne par exemple, ou bien à des entreprises tout simplement. Lors d'un stage réalisé au cours de mes études, une personne m'a révélé que les entreprises à qui elle revendait ces adresses email pouvaient les racheter jusqu'à 20 centimes par adresse valide. Il suffit donc de trainer sur Pastebin pour amasser assez rapidement une somme conséquente.

L'analyse que j'ai menée se décompose en 3 temps :

- Dans un premier temps, j'ai utilisé l'expression régulière citée en page 3 pour rechercher n'importe quel type d'adresse email
- Dans un deuxième temps, j'ai spécifiquement cherché des adresses sur le domaine « hotmail.com »
- Enfin, j'ai répété l'étape numéro 2 sur le domaine « gmail.com ».

La première étape consistait donc à fournir à la fonction *egrep* du script une expression régulière capable de ressortir tous les emails qu'elle rencontrait. Après quelques recherches, l'expression « `[a-z0-9._-]+@[a-z0-9._-]{2,}\.[a-z]{2,4}` » fut formée. Elle matche toute expression comportant une ou plusieurs fois les caractères de a à z, de 0 à 9, ainsi que ., -, _ suivis d'un arobase, lui-même suivi de 2 ou plus des caractères de a à z, de 0 à 9, ainsi que ., -, _ pour terminer par un point et deux à quatre caractères alphabétiques.



```
adminuser@adminuser-VirtualBox: ~/Desktop
adminuser@adminuser-VirtualBox:~/Desktop$ ./finder.sh
Saisir le chemin vers le(s) fichier(s) à analyser /home/adminuser/Desktop/Pastebin/23/*
Saisir le fichier dans lequel stocker les résultats mails
Saisir l'expression à rechercher [a-z0-9._-]+@[a-z0-9._-]{2,}\.[a-z]{2,4}
```

Capture d'écran de l'utilisation du script Shell pour rechercher des adresses email valides

Cette recherche m'a permis de recevoir le fichier « mails » qui comportait quelques données intéressantes.

La première information qui ressort est celle-ci : « Contact Us at : thepstop100@gmail.com ». En recherchant l'adresse email sur Google, on ne trouve étonnamment que le paste dont elle provient. Le paste contient un discours de publicité pour un nouveau site Internet qui propose de s'inscrire gratuitement et en échange, d'obtenir des avantages pour votre propre serveur web. Étonnamment, le site n'est déjà plus accessible à l'heure où je rédige ce rapport (c'est-à-dire seulement 2 semaines après qu'il ait créé sa publicité sur Pastebin).

Ensuite, j'ai découvert quelque chose d'assez incroyable. Il s'agit d'une liste d'environ 10 000 noms d'utilisateurs et mots de passe correspondants. Cette liste provient très certainement d'un dump de base de données sur un site non précisé dans le paste, et s'est retrouvée publiée de manière publique. Si la liste s'avère exacte, cela permet de faire un constat concernant la sécurité informatique en aparté de ce rapport, à savoir qu'il existe donc toujours des sites Internet qui stockent les mots de passe de leurs utilisateurs en clair, ou chiffrés à l'aide d'algorithmes réversibles. De plus, étant donné qu'un utilisateur moyen sur Internet ne possède en général que très peu de mots de passe différents, il est fort possible qu'en essayant les combinaisons présentes sur différents sites à la mode (Facebook, Twitter, Instagram, Snapchat, ...), voire même sur les comptes de messagerie eux-mêmes, nous puissions obtenir un accès à nombre de ces comptes. **La démarche étant toutefois illégale, aucun test de ce type n'a été réalisé durant ce projet !**

Et comme une nouvelle de ce genre n'arrive jamais seule, à partir de la ligne 12915 du fichier mails, on peut trouver le contenu de ce qui semble être la table « account » de la base de données d'un site. Cette table servirait apparemment à stocker les utilisateurs possédant un compte sur le site, et donc leurs pseudos, mots de passe, adresses emails, dates de connexion et d'inscription et carrément, leurs adresses IP.

```
INSERT INTO `account` VALUES ('69035817', 'lost1234',  
'*38A900C947909C02BB2E6349F5A2F7483917BBF3', 'tom', '1234567', 'kh_tom.elsner@yahoo.de',  
null, null, null, '', '2014-10-22 00:54:50', '1', 'c7e02e30849e19e182bd2905c4746317', null, null, '0',  
'OK', '', '0', '0', '0', '0000-00-00 00:00:00', '0', '0', '0000-00-00 00:00:00', '0000-00-00 00:00:00', '2015-  
10-22 00:54:50', '2015-10-22 00:54:50', '0000-00-00 00:00:00', '0000-00-00 00:00:00', '0000-00-00  
00:00:00', '0', '0', '', '2014-11-14 03:11:25', '0', '250', '0', '88.68.220.255', '', '0', '0', '0', null,  
'1414369961');
```

Exemple d'une des lignes du dump de base de données

Là encore, si quelqu'un arrive à déchiffrer les mots de passe présents, il y a de fortes chances pour que ces derniers soient réutilisables par une personne mal intentionnée pour détourner l'identité informatique des personnes présentes dans la liste.

Le fichier comporte encore quelques autres contenus du même type que les deux présentés ici (il contient en tout plus de 26 000 lignes !). Passons maintenant à la deuxième étape, qui visait à récupérer spécifiquement des adresses du domaine « hotmail.com ».

Comme pour la liste d'emails générale, le fichier « hotmail » ne contient que deux types de données :

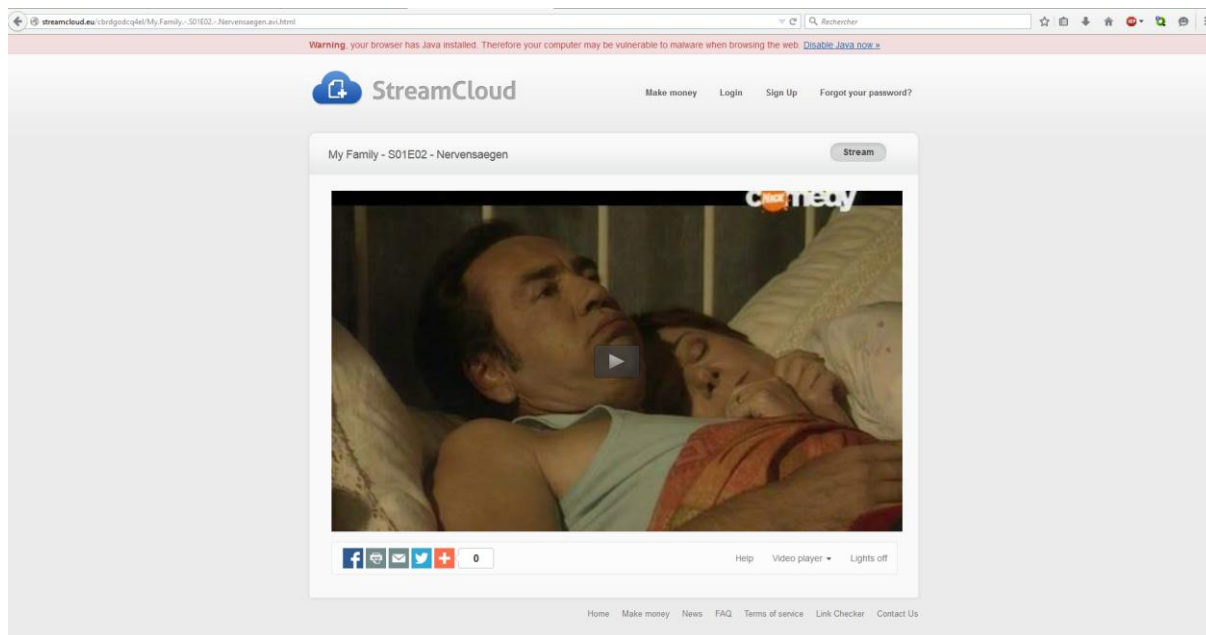
- Des listes d'adresses du domaine, avec ce qui ressemble à une liste de mot de passes en vis-à-vis
- Des dumps de base de données.

Là encore, nous sommes en présence d'un fichier conséquent (8952 lignes) qui contient pas moins de 8950 occurrences de l'expression « hotmail.com ». Imaginez revendre cela à des sociétés qui font de la publicité sur Internet ...

Enfin, avec l'expression « gmail.com », le constat est le même (fichier de « seulement » 3726 lignes), le fichier est composé du même genre de données.

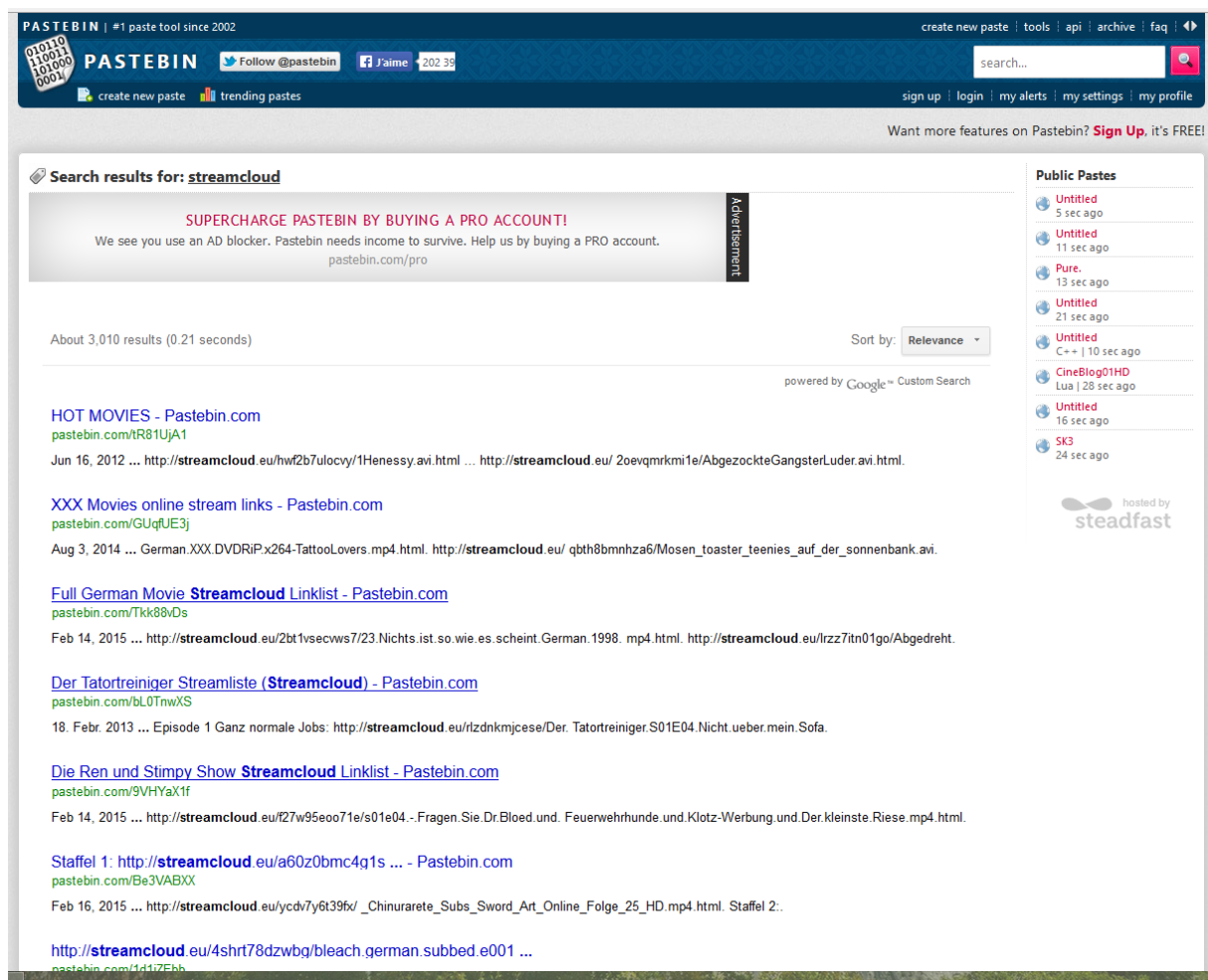
Films

Un autre type de contenu que l'on peut étonnement trouver sur Pastebin réside dans l'échange de liens de téléchargement illégal de films. Rappelons une nouvelle fois qu'à l'origine, les pastes ne contenaient que du code source sous licence libre, pour permettre à d'autres développeurs de forker ces projets et de créer de nouveaux contenus, et l'on comprend aisément que le but premier a une nouvelle fois été détourné. Dans ce paste (numéro 01vKHZFH), on trouve, entre autres, les liens vers les épisodes des 4 premières saisons d'une série appelée « My family » en allemand. Après vérification, les liens sont toujours fonctionnels comme le montre la capture ci-dessous



Capture d'écran d'un site de streaming trouvé dans les pastes analysés

On note que l'hébergeur utilisé est ici le site « StreamCloud ». En me rendant sur le site <http://pastebin.com>, j'ai pu constater que l'on peut tout à fait y chercher du contenu. J'ai donc, par simple curiosité, tenté de chercher le nom de cet hébergeur et voici ce qui m'a été proposé



Capture d'écran d'une recherche effectuée directement sur Pastebin

On peut voir dans cette recherche (qui compte tout de même 10 pages) qu'il existe un grand nombre de pastes dans lesquels on peut trouver des liens vers des films, à l'instar de ce que proposent, à plus grand échelle, les forums de téléchargement direct (direct download). Pastebin héberge donc des liens vers des œuvres protégées par des droits d'auteurs (œuvres cinématographiques, et même musicales), alors qu'encore une fois, son but premier était le libre partage de morceaux de code sous licence libre, donc dépourvus de quelconques droits d'auteur.

Enfin, comme nous pouvons également le remarquer sur la capture précédente, ma dernière recherche porte sur le fait que Pastebin est également sujet à du partage de contenu pornographique.

Pornographie

D'après [Wikipédia](#), « la pornographie sur Internet est la pornographie accessible sur Internet via les sites web, l'échange de fichiers par des réseaux peer-to-peer et les réseaux Usenet. Même si la pornographie était disponible sur Internet dès les années 80, c'est l'explosion du nombre d'internautes au début des années 90 qui a mené à une très forte expansion de la pornographie sur Internet. L'Internet a permis aux à la population d'accéder à la pornographie de façon anonyme, dans le confort de leur foyer, avec une diversité toujours grandissante. »

Grâce à l'analyse des pastes du 23 Février 2015, j'ai pu constater que Pastebin est un site qui pourrait être rattaché, à cause d'une partie de son contenu, à la catégorie des sites proposant de l'échange de fichiers à caractère pornographique décrite par Wikipédia.

En effet, la dernière expression que j'ai cherchée dans les pastes était « porn ». Il m'a été retourné la bagatelle de 425 occurrences de ce mot dans l'ensemble des fichiers analysés (sur la seule journée du 23 Février 2015 encore une fois).

L'analyse du résultat produit nous montre tout d'abord que certaines personnes utilisent ce terme pour former ce qui ressemble à un mot de passe

corinthiano_skate@hotmail.com:porno123
--

Ensuite, une liste répertoriant pas moins de 25 sites à caractère pornographique nous est présentée, ce qui ne sera pas pour déplaire à l'internaute qui surfe anonymement dans le confort de son foyer.

Puis vient une nouvelle liste de sites équivalents à ceux de la précédente liste. Cependant, cette fois-ci, les liens mènent directement vers des vidéos au format mp4 disponibles en ligne. Le dernier lien de la liste renvoie, lui, vers ce qui pourrait passer pour un site tout à fait légal, mais en fait il n'en est rien. Ce dernier propose en effet un lien vers un film du même genre que les précédents, sauf que le site sur lequel il est hébergé n'a rien de légal. Il s'agit en fait d'un site de téléchargement direct (comme ceux évoqués dans le paragraphe précédent) qui héberge, probablement illégalement, du contenu à caractère pornographique.

L'utilisateur qui suivrait ce lien dans le paste utiliserait donc doublement mal le site Pastebin puisqu'il s'en servirait pour de la consommation de contenu pornographique, mais également de la consommation de contenu soumis à des droits d'auteur (oui, cela existe même dans le monde du porno).

Enfin, au milieu de tout ce partage de contenu illégal, on trouve également un texte parlant de pornographie. C'est un texte qui provient de [ce blog](#) et qui parle d'une affaire de « revenge porn » qui aurait conduit les deux parties devant un tribunal. Cette affaire remonte à 2012 mais continue apparemment à faire parler d'elle (pour plus d'informations sur cette affaire, consultez [cette page](#)).

Conclusion

Avec l'augmentation constante du nombre de personnes ayant accès à Internet, le profil des utilisateurs s'est largement diversifié ces dernières années. Internet n'est donc plus du tout un espace de collaboration scientifique et de partage des connaissances. De fait, les sites qui le composent ne dérogent pas à la règle, et certains voient ainsi leur utilisation changer au fil des années.

C'est le cas de Pastebin, comme nous avons pu le voir tout au long de ce rapport. Un site qui autrefois avait été imaginé et conçu pour partager des morceaux de code informatique sur lesquels de nouveaux utilisateurs pouvaient faire naître de nouvelles grandes idées, et qui se voit aujourd'hui réduit à l'état de poubelle informatique, dans laquelle chacun est libre de venir déposer n'importe quel contenu, parfois même illégal, avec l'absence de contrôle la plus totale.