Techniques de collecte d'informations sur Internet

Isaline LAURENT

16 mai 2012

Introduction

Les techniques présentées ici ne constituent pas des attaques en elles-même, mais ont pour but d'obtenir le plus d'informations possibles sur la cible en vue de mener une attaque. Un adage bien connu en sécurité est que, mieux l'on connaît l'adversaire, mieux l'on saura déjouer ses défenses. Et vice-versa : une forteresse ne sert à rien si l'on peut aisément obtenir le mot de passe permettant l'ouverture du pont levis.

Je vais dérouler ici une liste non exhaustive de moyens de collecte d'informations via internet, qui, mis bout à bout peuvent largement aider à infiltrer un réseau protégé. Je parlerais dans un premier temps de ce qu'on appelle le google hacking, ou encore google dorks, qui consiste en la manipulation de requêtes spécifique sur le très connu moteur de recherche Google. Après en avoir expliquer les principaux opérateurs, je l'illustrerais de quelques exemples, puis expliquerais comment s'en protéger au maximum. Ensuite, je parlerais des autres outils que j'ai sélectionné pour cet exposé.

1 Le Google Hacking

Le Google Hacking est l'art et la manière d'utiliser les opérateurs de recherche du célèbre moteur éponyme. La plupart de ces opérateurs ont une utilité première tout à fait anodine. Cependant, utilisé à mauvais escient, ils peuvent apporter des informations cruciales.

1.1 Présentation des opérateurs

Voici un récapitulatif des différents opérateurs de recherche de Google.

Syntaxe du filtre	Description
inurl	Retourne les pages contenant un lien vers un fichier du type mentionné en argument – exemple : `inurl:admin`
filetype	Retourne les pages contenant un lien vers un fichier du type mentionné en argument – exemple : `filetype:pdf`
intext	Recherche un motif dans le contenu – exemple : `intext:mysql _ connect`
site	Permet de filtrer la recherche sur un site passé en argument.
link	Retourne l'ensemble des pages contenant un lien vers le motif passé en argument
cache	Permet d'accéder à la version mis en cache par Google. Cette option est surtout utilisée pour visiter discrètement un site.
define	Fournit une définition au terme passé en argument
intitle	Recherche dans le champ ‹title›‹/title› d'une page HTML
ext	Recherche dans les pages dont l'extension (html, php, etc..) est le motif passé.
[X]..[Y]	Effectue une recherche dans l'intervalle [X, Y]. Par exemple: `page+1...100`
info	Récupère des informations sur le site passé en paramètres. Par exemple, `info:www.kernel.org`
related	Retrouve les sites sémantiquement liés au paramètre. Par exemple, `related:www.kernel.org`

A cela s'ajoute les opérateurs + (AND), - (NOT), et | (OR).

1.2 Exemples

Voici quelques exemples d'utilisation de ces opérateurs.

Site mapping

Ce que l'on appelle Site mapping consiste à la récupération des noms de tous les sous domaines d'un domaine. Cela peut être utile pour plusieurs raisons, comme par exemple trouver l'url vers l'espace administrateur. Il suffit de rechercher : site :<domaine> inurl :<domaine> -www.<domaine>

kickstart filetype :cfg

Kickstart est un outil permettant d'automatiser l'installation du système d'exploitation Fedora. Il est utile par exemple, dans le cas de la migration d'un parc informatique, où la répétition de la même procédure d'installation sur de nombreuses machines peut être fastidieuse. Le fichier de configuration a pour extension .cfg et commence toujours pas #Kickstart.

On peut ainsi faire la requête # kickstart filetype :cfg. En cliquant sur le premier résultat, nous obtenons la page suivante :

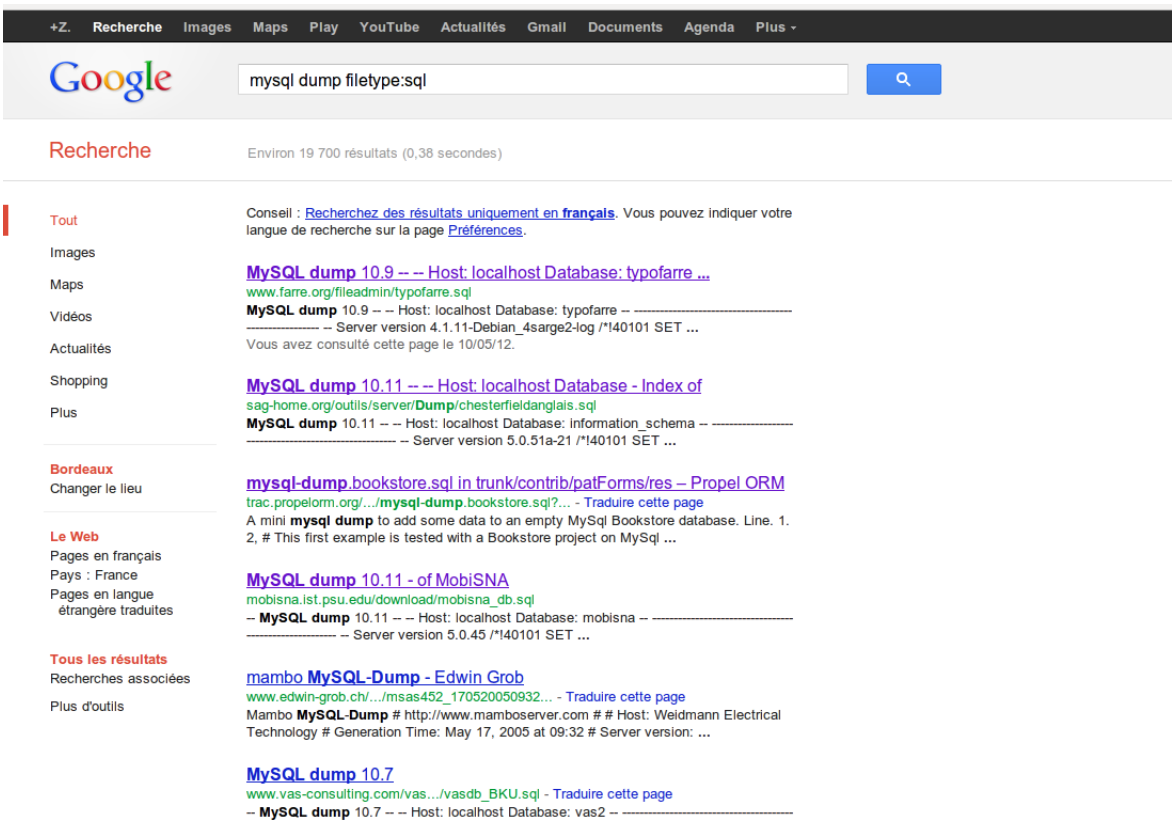
```
# Kickstart file anne.possoz@epfl.ch
# Test d'installation machine virtuelle
#

install
url --url http://elle.epfl.ch/centos-5/os/i386
lang en_US.UTF-8
# Deprecated:langsupport --default en_US.UTF-8 fr_CH.UTF-8 fr_FR.UTF-8
keyboard fr_CH
# Deprecated:mouse
timezone --utc Europe/Zurich
selinux --disabled
skipx
network --device eth0 --bootproto static --ip 128.178.131.5 --netmask 255.255.255.0 --gateway 128.178.131.1 --nameserver 128.178.15.8,128.178.15.7 --hostname ditgepcks
rootpw --iscrypted $1$zEeV8QCa$RnfC0Gsxe8QN1qGQgaCB9.
firewall --enabled --port=22:tcp
authconfig --enableshadow --enablemd5
#
# Deprecated avec parametre:zerombr yes
zerombr
bootloader --location=mbr --driveorder=sda
part / --fstype ext3 --size=9216 --asprimary
part swap --size=1024 --asprimary
# Reboot aprÃ¨s installation
reboot
#
#---------------------------------------------------------------------
#
```

On a donc de nombreuses informations sur l'installation.

mysql dump filetype :sql

MySQL est un système de gestion de base de données très utilisé. Ce que l'on appelle un dump est un fichier de sauvegarde, constitué des requêtes sql qui permettent à qui les chargent d'avoir exactement la même base qu'au moment de la sauvegarde.

En tapant mysql dump filetype :sql dans le champ de recherche, on obtient les résultats suivant :



Puis, en cliquant sur l'un des résultat, en faisant une recherche sur 'insert into 'mos_user', après avoir remarqué que chaque table est préfixé de 'mos' :

```
INSERT INTO mos_users VALUES (62,'Administrator','admin','notreal@nowhere.no',
'21232f297a57a5a743894a0e4a801fc3','Super Administrator',0,1,25,
'2005-02-13 18:53:01','2005-05-13 16:02:07','','','','','','','','','','',
'','','','','','','US','','','','','','','shopper','','','','','',Checking);
```

On a donc le mot de passe crypté du compte Administrateur. La plupart des mots de passe étant crypté en md5 ou en sha1, il suffit d'essayer de les cracker avec ces deux systèmes avec n'importe quel outil en ligne. Celui-ci est en md5 et correspond à 'admin'.

inurl :sitemanager filetype :xml Filezilla

Filezilla est un outil de gestion de connexion à un serveur ftp. L'utilisateur a la possibilité d'enregistrer la configuration de connexion de chacun des

serveur ftp auxquels il a accès. Filezilla enregistre ces informations au format xml dans un fichier nommé "sitemanager.xml".

Normalement, ce fichier n'a rien à faire sur Internet. Cependant, une mauvaise configuration peut mener à la fuite de celui-ci. C'est ce que l'on va exploiter ici.

En tapant la requête citée plus haut dans Google, nous obtenons les résultats suivants à la page 2 :

Ici, c'est le troisième résultat qui nous intéresse. En cliquant sur le lien, on arrive sur une page blanche indiquant que la page n'existe plus. En utilisant l'opérateur cache de Google, nous obtenons les informations suivantes :

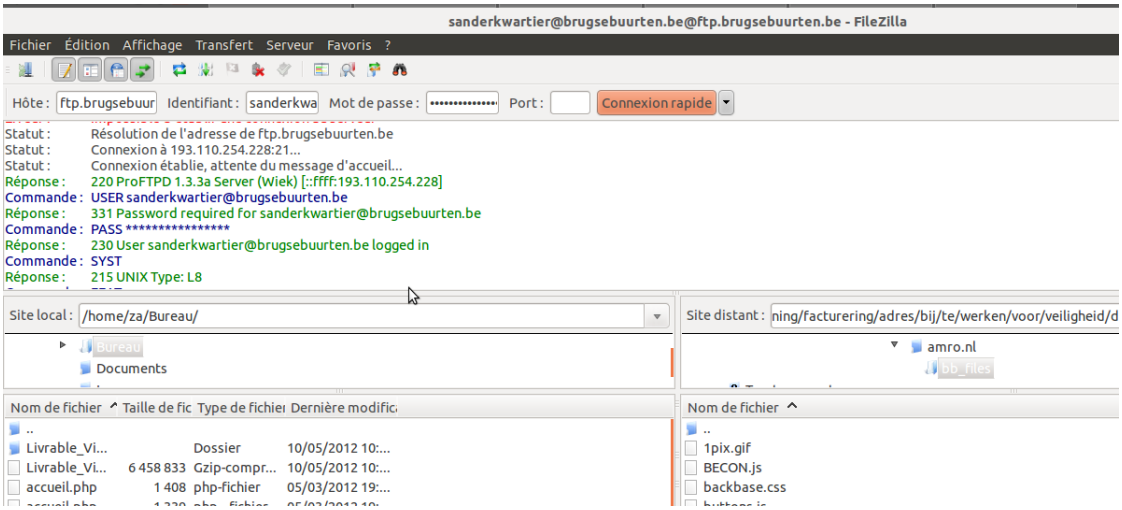
```
ftp.brugsebuurten.be 21 0 0 sanderkwartier@brugsebuurten.be sanderkwartier01
1 0 MODE_DEFAULT 0 Auto 0 Sanderkwartier 0Sanderkwartier
```

Habituellement, la syntaxe du fichier xml est la suivante :

```
<FileZilla3>
<Servers>
        <Server>
            <Host>ic.cz</Host>
            <Port>21</Port>
            <Protocol>0</Protocol>
            <Type>0</Type>
            <Logontype>1</Logontype>
            <User>nahravky</User>
            <Pass>224044702</Pass>
            <TimezoneOffset>0</TimezoneOffset>
            <PasvMode>MODE_DEFAULT</PasvMode>
            <MaximumMultipleConnections>0</MaximumMultipleConnections>
            <EncodingType>Auto</EncodingType>
            <BypassProxy>0</BypassProxy>
            <Name>Nové místo</Name>
            <Comments/>
            <LocalDir/>
            <RemoteDir/>
            Nové místo
        </Server>
    </Servers>
</FileZilla3>
```

On remarque alors que les champs User et Pass sont situés à la suite du champ Port, souvent à 21, et de deux champs régulièrement à 0. En observant le résultat de la page en cache, on devine assez facilement que les identifiants sont s ûrement sanderkwartier@brugsebuurten.be et sanderkwartier01. L'adresse du serveur ftp est aisément reconnaissable, ftp.brugsebuurten.be.

Et bingo, en entrant ceux-ci dans Filezilla, la connexion se fait :



Grâce à une simple recherche Google, nous avons pu nous connecter à un serveur ftp qui ne nous appartient pas.

1.3 Comment s'en protéger

Dans un premier temps, pour éviter qu'une personne mal intentionnée ne devine l'architecture d'un site, il vaut mieux ne pas mettre des chemins conventionnels : éviter les /admin, /sql ... Dans la même idée, la mise en place d'un .htaccess est une bonne défense.

Ensuite, le fichier robots.txt peut être d'une grand aide. C'est lui qui définit le comportement que doivent avoir les robots de crawling en arrivant sur un site. On peut donc y indiquer la portée qu'on les robots, et ainsi limiter l'apparition d'url sensibles dans les résultats d'un moteur de recherche.

Enfin, le plus simple reste encore de vérifier la visibilité de son site par soi-même, et de faire en sorte que seul les documents que l'on veut soit accessible.

2 Autres outils

2.1 Autres moteurs de recherche

Google n'est pas le seul moteur de recherche. On trouve par exemple gigablast.com, qui est très interessant dans le sens où il permet de faire des recherches par adresse ip.

Je vous invite à visiter le site http ://www.searchengineshowdown.com/features/ qui détaille les capacités de différents moteurs de recherche.

2.2 Netcraft

Cette application tout à fait légale permet de rechercher des informations sur des entreprises. Elle peut donc être utile pour obtenir le numéro SIRET ou l'adresse d'une entreprise par exemple. Cependant, on peut trouver de nombreuses informations, comme, dans le cas d'une société disposant d'un nom de domaine, le système d'exploitation sous lequel tournent les serveurs, leur serveur web, leur adresse ip, ou encore leur emplacement physique.

2.3 Whois

Whois est un outil, disponible en console, qui fournit des informations sur une adresse ip ou un nom de domaine. Il n'y aucun format défini pour le résultat d'un whois, ainsi, les informations que l'on peut y trouver varient d'une recherche à une autre. Généralement, Netcraft est plus orienté vers l'entreprise, alors que Whois va nous fournir des informations sur les personnes physiques.

Conclusion

Encore une fois, l'utilisation de ces outils n'est pas obligatoire, mais facilite grandement les attaques. Ainsi, il est important de bien protéger les informations que l'on trouve sensible, pour éviter qu'elles ne se retrouvent sur Internet, disponible à tous. La meilleure solution reste encore de faire soi-même des recherches, avant de déterminer si la visibilité de nos informations nous convient.