

Google Hacking

1 Introduction

Malgré la sobriété et la simplicité de sa page d'accueil, Google constitue un moteur de recherche extrêmement puissant et riche en fonctionnalités. Il permet d'effectuer pour le commun des mortels des recherches dans différentes sources d'informations telles les pages web, les images, les groupes de discussion et plus récemment les blogs ou les vidéos.

Cette puissance de recherche peut tout aussi bien être utilisée par les pirates informatiques pour leur permettre d'obtenir des informations sensibles. On parle alors de « Google Hacking ». Ce terme désigne communément l'utilisation des nombreux opérateurs de recherches disponibles avec des mots clés ou des phrases judicieusement choisies pour obtenir des informations sensibles de toute nature (fichiers de configuration, version des logiciels utilisés, données personnelles, numéro de cartes bleues, ...). La récupération de telles informations peut alors constituer une première étape dans le cas de tests d'intrusion ... où d'attaques réelles !

Même si Google dispose de fonctionnalités uniques telles la mise en cache ou la mise à disposition d'API (interfaces de programmation) permettant d'automatiser les requêtes, de nombreux autres moteurs de recherches peuvent être utilisés à cet effet.

Cet article présente en premier lieu les différents moyens et opérateurs existant pour effectuer une recherche avec Google. Il s'attache ensuite à présenter quelques pistes d'exploitation « détournées » dont peuvent se servir les auditeurs en sécurité ou les personnes malveillantes cherchant à exploiter des failles. Enfin quelques contre-mesures sont présentées afin de faire face à ce phénomène qu'est le Google Hacking.

2 La recherche d'informations avec Google

Rares sont les internautes n'ayant jamais utilisé Google pour effectuer une quelconque recherche comme point de départ pour « surfer » sur le réseau mondial. Il peut sembler inutile de présenter la manière « basique » d'effectuer une recherche à partir de la page d'accueil de Google (saisie d'un ou plusieurs mots clés séparés par des espaces dans le formulaire de saisie unique puis de clic sur le bouton « Recherche Google ») mais l'intérêt est grand de vous présenter les principaux opérateurs avancés mis à votre disposition et de vous en donner quelques exemples d'utilisation :

- **intitle:hacking** : recherche les pages web contenant le mot « **hacking** » dans leur titre (1 230 000 résultats),
- **inurl:login** : recherche les pages contenant l'occurrence « **login** » dans leur url (27 000 000 résultats.),
- **intext:"md5 reverse hash"** : recherche les pages contenant la phrase « **md5 reverse hash** » dans leur corps (2630 résultats),
- **link:www.blackhat.com** : recherche les pages web contenant un lien vers www.blackhat.com (1020 résultats),
- **filetype:log** : recherche les fichiers dont le type ou l'extension est « **log** » (1 030 000 résultats),
- **site:www.sans.org** : fourni l'ensemble des pages indexées du site www.sans.org (260 résultats).

Les moteurs de recherche indexent une quantité d'information phénoménale (plusieurs dizaines de milliards de pages web). Effectuer une recherche avec quelques mots clés peut fournir des centaines de milliers (voir des millions) de pages web en retour. Il est donc indispensable d'utiliser les opérateurs avancés, en les combinant éventuellement, avec des mots clés les plus précis ou spécifiques possible afin d'affiner au maximum la requête. A titre d'exemple, la recherche basique « google hacking » renvoi **15 900 000** résultats, la recherche « google intitle:hacking » **289 000** résultats et « google intitle:hacking filetype:pdf » seulement **296**. Voici quelques exemples de requêtes combinant différents opérateurs permettant d'obtenir des résultats parfois intéressants :



L'illustration d'emploi de ces quelques opérateurs avancés permet d'ores et déjà d'imaginer les multiples possibilités d'exploitation des moteurs de recherche dans le domaine des audits de sécurité, des tests de pénétration ou d'attaques ciblées ... Dans ce dernier cas, il est impératif de récolter un maximum d'informations sur sa « victime » avant de passer à l'offensive tout en assurant un minimum de discrétion. Ce service peut être rendu par des serveurs proxy, des jeux de rebonds sur différents serveurs piratés ou plus simplement par la fonctionnalité de cache offerte par Google.

3 Exploitation de la mise en cache des pages web

Pour la majorité des pages web parcourues, les « robots » mis en œuvre par Google conservent une copie sur leurs serveurs, c'est le principe de mise en cache. De cette manière une personne peut consulter un site web entier en n'utilisant que les pages web copiées sur les serveurs de Google. Aucune connexion n'est alors effectuée avec le site en production à l'exception des images. Il est donc nécessaire de consulter les pages web mode texte uniquement. Dans ce mode de navigation, le site hébergeant les pages web en production n'a pas connaissance de sa consultation. Cette fonctionnalité offerte Google permet donc de consulter des informations de manière « anonyme ». Il est toutefois important de noter que ces copies de pages web ne se font que lorsque les robots de Google les parcourent. Il peut donc à un instant donné exister des différences entre les pages situées sur les serveurs en production et celles copiées sur les serveurs de Google.

Cette synchronisation asynchrone peut être dans certain cas un avantage de taille pour les pirates. En effet, toute page web qui était publiquement accessible et qui ne l'est plus pour différentes raisons (mise en place d'une authentification par login/mot de passe, retrait de la page sur serveur web, indisponibilité du serveur) peut encore être présente sur les serveurs de Google. Dans cette mesure, il peut être possible d'accéder à des informations qui ne sont plus disponibles. Cette fonctionnalité de cache peut tout de même être utilisée à des fins plus « nobles ». A titre d'illustration, nous pouvons rappeler l'incident électrique qu'a connu l'hébergeur Internet Redbus Interhouse en région parisienne rendant ainsi indisponible de nombreux site internet français. La seule solution à ce moment là était le service de cache de Google ...

Google Hacking

Google Hacking : Le Google Hacking consiste à découvrir des informations sensibles et des sites internet vulnérables à l'aide de requêtes Google spécifiques ...
www.dicodunet.com/definitions/google/google-hacking.htm - 52k -

[En cache](#) [Pages similaires](#)










Lien d'une page web en cache

La visualisation de pages en cache peut naturellement aboutir sur des index de répertoires qui étaient, qui sont encore, disponibles. Ceci nous amène donc à un autre type de recherche possible : celui des répertoires de données.

4 Parcours de répertoires et localisation de fichiers

La localisation et l'exploration de répertoires non protégés peut permettre de découvrir des fichiers contenant des informations sensibles (fichiers de mot de passe, de configuration, de journaux d'audit, ...) et permettre de s'introduire sur les serveurs web et les équipements réseaux. Partant de l'observation que de nombreux serveurs web affichent le contenu d'un répertoire de données en affichant en premier lieu « index of » suivi du chemin complet dans la barre de titre de la page web il est opportun d'interroger Google à l'aide de la requête « **intitle:index of** ». Ce dernier fourni alors une liste impressionnante de résultats (plus de 22 millions). Afin d'affiner la recherche, il est conseillé d'ajouter des mots clés dans le chemin du répertoire recherché. La requête « **intitle:"index of" inurl:admin** » permet de rechercher l'ensemble des répertoires visités par Google ayant le mot « admin » dans leur url. Il est alors possible de remonter et d'explorer les différents répertoires en utilisant le lien « Parent directory ». La recherche peut être encore affinée en spécifiant un nom de fichier à rechercher dans la page. Ainsi la requête « **intitle:"index of" router.cfg** » permet de rechercher des répertoires contenant un fichier nommé « router.cfg ». La localisation d'un répertoire peut par ailleurs peut permettre d'exploiter des failles de type directory traversal.

Index of /src/JONADAB/Net-Server-POP3-0.0009

<u>Name</u>	<u>Last modified</u>	<u>Size</u>
 Parent Directory		-
 .bash_history	23-Jun-2004 02:26	85
 .bash_logout	07-Jul-2003 17:01	24
 .bash_profile	07-Jul-2003 17:01	191
 .bashrc	07-Jul-2003 17:01	124
 .emacs-places	23-Jun-2004 02:26	51
 .gimp-1.2/	12-Jun-2004 19:03	-
 .mailcap	15-May-2001 10:02	141
 .screenrc	15-Jun-2003 15:02	3.6K

Parcours d'un répertoire indexé par Google

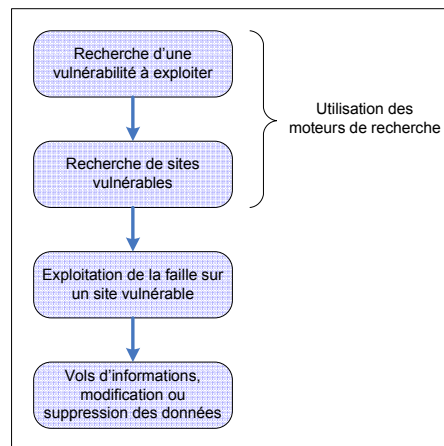
5 Messages et adresses électroniques

Les messages électroniques, qu'ils soient de nature privée ou professionnelle, constituent une mine d'information très importante à propos des individus qui les échangent. Par la lecture de messages ou de boîtes aux lettres complètes, il est possible d'obtenir des renseignements concernant l'entourage, les loisirs, l'environnement de travail, les projets en cours, ... Toutes ces informations récoltées peuvent ensuite servir à des attaques de type « Social Engineering ». De nombreux mails ou boîtes aux lettres sont disponibles sur Internet. Des requêtes couplant l'opérateur « filetype: » avec des mots clés contenus dans les en-têtes SMTP (From, Subject, Received, Message-ID, ...) permettent d'effectuer des recherches relativement efficaces. A titre d'exemple la requête « filetype:eml eml intext:subject intext:From » permet de rechercher des messages sauvegardés au format Outlook. Obtenir des adresses électroniques valides correspondant à des employés d'une société en interrogeant la base des groupes de discussions (ex @societe.com) est aisée. Il est alors possible d'exploiter ces adresses pour envoyer des e-mails contenant des programmes malveillants afin de dérober toute sorte d'informations. L'obtention d'un grand nombre d'adresses valides peut être facilitée par l'utilisation de scripts perl.

Des vers malicieux s'appuyant sur la messagerie électronique peuvent ainsi en effectuant des requêtes automatisées auprès des serveurs de Google obtenir de nouvelles adresses valides et continuer leur propagation. Après avoir infecté leur victime, ces vers peuvent tout à fait renvoyer des informations internes permettant par la suite de mettre en œuvre des exploits.

6 Utilisation d'exploits

Pour pénétrer un système informatique, les pirates s'appuient très souvent sur des failles logicielles. Nombreux sont les sites qui publient des informations permettant d'exploiter ces failles. Il reste toutefois possible de se baser sur les en-tête typiques des fichiers source et de préciser le cas d'emploi pour effectuer des recherches alternatives (par ex. « "#define <stdio.h>" usage exploit »). L'étape suivante consiste à localiser des sites permettant d'exploiter la faille choisie. La méthode la plus simple consiste à exploiter les signatures typiques que laisse souvent les webmasters sur leur site web. En effectuant une requête à partir de la chaîne clé caractérisant l'intitulé et la version du logiciel, il est alors aisé d'obtenir une liste de sites cible. Ainsi la requête « Powered by CuteNews v1.3.1 » permet d'effectuer une recherche de l'ensemble des sites répertoriés par Google qui mettent en œuvre la version 1.3.1 de CuteNews.



Mise en œuvre d'exploits

L'automatisation d'un tel processus peut amener un programme malveillant à se reprendre de manière fulgurante. Cela a été le cas par exemple du ver Santy en décembre 2004. Il se propageait via les sites phpBB en versions égales ou inférieures à 2.0.11 et utilisait de manière autonome Google pour découvrir de nouvelles cibles. Aujourd'hui cette démarche peut être simplifiée via l'utilisation des interfaces de programmation Google ...

7 Les Google API (Application Programming Interface)

Disponibles depuis janvier 2005, les Google API constituent un SDK (Software Development Kit), sont disponibles gratuitement et permettent le développement de nouvelles applications sachant interroger de manière autonome les bases de données de pages web indexées par Google. L'utilisation de ce service est toutefois soumise à la délivrance préalable d'une clé d'authentification validant 1000 requêtes par jour. Plusieurs logiciels ont été développés à partir de ces API et sont utilisables en l'état afin d'auditer des sites web ... Parmi le plus connu, nous pouvons citer SiteDigger, Gooscan, Goolink ou AdvancedDork, SiteDigger permet par exemple de récupérer de multiples informations au travers d'une dizaine de catégories qui sont les fichiers de backup, les consoles d'administration distantes, les fichiers de configurations, les messages d'erreurs, les données confidentielles, les vulnérabilités ou encore les profils d'application. L'objectif officiel de ces types de logiciel est de faire prendre conscience aux webmasters des éventuels problèmes de sécurité sur leur site, il devient toutefois aisé de les utiliser à d'autres fins...

8 Conclusion

La finalité première des moteurs de recherche sur Internet est d'indexer la quantité astronomique d'informations existantes et de la rendre aisément accessible au public. Seulement certaines de ces informations peuvent être utilisées à l'encontre même des personnes les mettant à disposition. Il convient alors de filtrer l'ensemble des informations rendues publiques sur vos serveurs directement ou indirectement connectés à Internet. Pour certains moteurs de recherche, notamment Google, il est possible de placer à la racine des serveurs web un fichier de configuration (souvent nommé Robots.txt) permettant de spécifier aux robots d'indexation de ne pas parcourir certaines pages ou certains répertoires. Il peut par ailleurs être demandé explicitement à certain moteur de recherche le retrait de leur base de données d'informations vous concernant.

Pour finir, nous vous rappelons quelques mesures élémentaires à mettre en œuvre :

- Appliquer régulièrement les mises à jour et les correctifs de sécurité pour l'ensemble des serveurs et des logiciels installés,
- Désactiver la fonction de « directory browsing » sur l'ensemble des serveurs web,
- Supprimer toute information permettant de déterminer les versions des logiciels utilisés,
- Modifier l'ensemble des messages d'erreur spécifiques à certaines versions de logiciels
- Utiliser des moyens de chiffrement avancé pour utiliser les services de messagerie électronique ou de partage de documents.

Les références

Centre d'aide de Google - <http://www.google.fr/support/?hl=fr>
Page de recherche avancé de Google - http://www.google.fr/advanced_search?hl=fr
Page de demande de suppression d'indexation de données auprès de Google - <http://www.google.com/remove.html>
Google API - <http://code.google.com/apis/>
Publications SSTIC - <http://actes.sstic.org>
The Google Hacking database - <http://johnny.ihackstuff.com/ghdb.php>
Fichiers d'exclusion Robots.txt - <http://www.robotstxt.org/wc/robots.html>
Google Hacking for penetration testers – Johnny Long – Ed. Syngress
Google Honey Pot – <http://ghh.souceforge.net>

Mars 2007.

Brice Le Tallec (Consultant Réseau et Sécurité pour l'agence ESEC <https://esec.fr.sogeti.com>).

Association LABO <http://www.labo-asso.com/>