Comment récupérer les 35 millions d'adresses Gmail

Par Sébastien Bilbeau, publié le 20 mai 2011

Une négligence de la part de Google vient d'être découverte. En effet, il est possible de récupérer par le biais d'un Sitemap l'ensemble des adresses des profils Google. Celles-ci, bien que masquées par un ID, redirigent dans la plupart des cas, vers des adresses de la forme https://profiles.google.com/PSEUDOGMAIL. Ce qui permet à n'importe qui d'en déduire l'adresse Gmail rattachée au profil consulté. Voici donc une méthode à coup de commandes bash, pour récupérer l'ensemble des 35 millions d'adresses Gmail.

Pour commencer, on récupère le fameux sitemap de Google :

```
mkdir google-urls-profile && cd google-urls-profile
wget --quiet http://www.gstatic.com/s2/sitemaps/profiles-sitemap.xml
```

Après ouverture, on se rend compte que le fichier contient lui-même des adresses pointant vers d'autres Sitemaps :

```
<sitemap>
  <loc>http://www.gstatic.com/s2/sitemaps/sitemap-000.txt</loc>
  <lastmod>2011-03-16</lastmod>
</sitemap>
<sitemap>
  <loc>http://www.gstatic.com/s2/sitemaps/sitemap-001.txt</loc>
  <lastmod>2011-03-16</lastmod>
</sitemap>
```

Du coup, on parse son contenu pour en extraire uniquement les URLs et on les télécharge :

```
cat profiles-sitemap.xml | grep "<loc>" | tr -d '\t' | sed 's/^<.*>\([^<].*\)<.*>$/\1/'
> list.txt
for url in `grep http list.txt` ; do wget --quiet $url; done;
```

Au final, on se retrouve avec 7104 fichiers Sitemap, pour ma part je les ai regroupés en une seule liste :

```
$ cat sitemap-*.txt > profile-urls.txt
```

```
$ wc -l profile-urls.txt
35 442 725 profile-urls.txt
```

La commande wc nous confirme que nous sommes bien en possession des 35 millions d'URLs pointant vers des profils Google. Si on inspecte leur contenu, les adresses sont de la forme suivante :

- https://profiles.google.com/11713590257193879 36
- https://profiles.google.com/11200695271094933 21

Les données ne sont donc pas exploitables telles quelles, car le profil Google de chaque utilisateur est masqué par un ID (ex: 11713590257193879 36) qui ne nous donne aucune indication. Mais cette adresse redirige, dans la plupart des cas, vers https://profiles.google.com/PSEUDOGMAIL. En fait, cela est vrai quand un utilisateur a activé le mode "user friendly" dans son compte.

Nous allons être obligés de tester l'ensemble des 35 millions d'URLs (ce qui prend pas mal de temps), afin de récupérer l'adresse qui se cache derrière cette redirection. Et on peut le faire assez facilement avec ce script, nommé getmail.sh :

```
#!/bin/bash
for url in `cat $1`
do
        curl --silent --head $url | grep "^Location:" >> getmail.txt
done
```

Que l'on appelle ensuite comme ça :

```
chmod +x getmail.sh
for file in `ls sitemap-*.txt`; do ./getmail.sh $file; rm -f $file; done;
```

L'avantage de cette méthode est la suivante : si une adresse de profil ne possède pas de redirection, le champ "Location:" donné par la commande curl ne sera pas renseigné. Cela permet de filtrer uniquement les adresses avec un pseudo Gmail valide.

Une fois le traitement terminé, il ne nous reste plus qu'à rajouter "@gmail.com" à la fin de chaque pseudo, pour construire une liste valide :

```
cat getmail.txt | awk -F'google.com/' '{print $2"@gmail.com"}'
```

Dans ce cas précis, cette méthode m'a permis de construire une liste de 15 000 adresses Gmail en moins d'une heure. Je me suis arrêté avant la fin, car je n'ai finalement aucun intérêt à collecter ce genre d'informations. Mais n'importe qui peut utiliser cette méthode pour constituer une liste d'adresses mail et la revendre à des spammeurs par exemple.

Source

Autres articles du même sujet

- Analyse du piratage des 20K comptes Hotmail, Gmail et Yahoo
- Linux local root exploit via SUID
- Lulzsec dévoile 62 000 mots de passe
- Analyse rapide du pirate de Sony Pictures par Lulzsec
- Analyse rapide du piratage du groupe Anonymous

44 Commentaires pour "Comment récupérer les 35 millions d'adresses Gmail"

Flux des commentaires de cet article Ajouter un commentaire

- mouais, j'ai lu ça ce matin sur Korben.info, c'est assez flippant, là tu nous montre que si on si intéresse, c'est très simple a mettre ne œuvre juste une question de temps en fait. C'est navrant de la par e Google. On ce retrouve dans 2 situation possible soit Google la fait exprès, soit c'est juste une boulette monumentale.
Et je ne trouve pas en général qu'il fasse tant de boulette que ça.
en tant qu'utilisateur de leur service, je n'ai rien à leur reprocher mais ce genre d'information me pousse chaque jour un peu plus vers l'auto-hébergement de mes service internet, mail, site et pourquoi pas un serveur jabber 😉 .

nithir , le 20 mai 2011 à 14:28

- Dommage google avait simplifié la tache pour toi, tu pourrais checker une dizaine de fichiers des 7000 en parallèle hehe

PiTiLeZarD , le 20 mai 2011 à 14:30

- C'est vrai que de l'auto hébergement c'est tellement plus sécurisé et fiable dans le

temps... haha

Samy , le 20 mai 2011 à 14:37

- il y a une erreur dans la dernière ligne de commande !
  Très jolie démo sinon !

crash , le 20 mai 2011 à 14:46

- **@Samy** : plus fiable dans le sens ou si je fait une boulette dans mes configs je serait responsable. mais ça m'étonnerai beaucoup que volontairement je me mette à balancé mes info perso sur le net comme ça pour le fun.
  Alors oui, google à sûrement une équipe de ninja du fbi en sécurité info que je suis très loin d'égalé en terme de compétence mais 1 petit gas perdu au milieu de l'océan du net a moins de chance de se faire attaquer et de perdre ses infos perso dans la nature, que si il les confis à des grosses boites (cf. sony, facebook, et maintenant google).
  Les derniers soucis le montres, les grosses sociétés ne sont pas plus fiable parce que le fait qu'elle détiennent beaucoup d'info rend les attaques contre elle très rentable, donc attire des criminels plus compétents.
  Bref la seul chose qui me rebute dans l'auto-hébergement, c'est la charge de travaille nécessaire, à maintenir mon système à jour et performant pour mon utilisation.

nithir , le 20 mai 2011 à 14:48

- En l'occurence, ici, la boulette n'impactera pas les utilisateurs, vu l'efficacité de l'antispam Gmail.

  De plus, l'auto hébergement pose d'autres problèmes, comme le cout (matériel, connexion, temps), l'absence de redondance réelle, ou même ne serait ce que le travail nécessaire pour ne pas se faire blacklister son IP.

Samy , le 20 mai 2011 à 14:51

-

Et si je te dis que Google utilise les spam-reports de Gmail pour mieux filter les résultats du moteur de recherche.

¥€$, le 20 mai 2011 à 14:53

-

Parce que le parallélisme c'est bien (mais qu'on veut pas non plus faire une fork bombe) : for url in `cat list.txt`; do if [$(ps aux | grep wget | wc -l) -gt 500]; then sleep 20; fi; (nice wget --quiet $url &); done;

À noter que chez moi avec ta commande le fichier list.txt ne contient pas que des url, j'ai fait un « grep txt » en plus.

erdnaxeli , le 20 mai 2011 à 15:00

-

@Samy : pour ce qui est du temps à passé je suis d'accord ça va etre long (mais c'est une aventure, un passe temps),
pour le cout, je le trouve pas si génant, je compte pas avoir un site avec 10 000 visite par jours, pour tous dire c'est même pas le but d'en avoir, alors pas besoin d'avoir de matos très cher, de la haute disponibilité toussa, bidule qui supporte une bd mysql, apache, php, ça me suffit, un 486 DX 2 turbo avec un gros disque dur de 500Mo et 256Mo me suffirait pour hébergé un site).
Alors quand à la question du blacklistage j'avoue que pour le coup se serait utile pour limité les visites et pas pourrir ma bande passante, quoiqu'il en soit ce serait peut etre plus efficace que ce que j'ai actuellement un pauvre truc tous naz chez free dans lequel je ne poste rien finalement depuis qu'il s'est fait dégommé une fois sûrement par un script kiddy tous ça parce que je peu pas mettre à jours un blog wordpress depuis ma fac parce que le port 80 est bloqué.
en gros tu peu facilement faire ce que je veux avec un nas synologie ou si tu veut joué avec un petit pc pas trop cher.

(PS: Sinon se faire blacklister par qui?? les moteurs de recherche?)

nithir , le 20 mai 2011 à 15:11

-

C'est marrant j'ai fait le meme genre de test ce matin en lisant le post de korben et j'arrive aux même nombres...

On récupère environ 50% d'adresses mails dans les 5000 urls d'un fichier sitemap.

bartounet , le 20 mai 2011 à 15:24

-

Erdnaxeli : j'ai des besoins trop importants pour un 486 DX2 avec 500 Mo de disque

Je parlais ici de blacklistage de serveur de mail (et d'auto hébergement de serveur POP/SMTP)

Samy , le 20 mai 2011 à 15:27

-

@¥€$: Tu sous-entends que Google a laissé volontairement cette bêche ouverte pour que les attaques de spams améliorent leur moteur de recherche. Malin !

crash , le 20 mai 2011 à 15:28

-

@All : le problème de base est que Google doit faire en sorte que vos profiles soient correctement indexés dans les moteur de recherche, afin que l'on trouve tout sur vous rien qu'avec votre nom. A partir de là, difficile de concevoir une méthode qui ne révèle pas des informations sur vous. Bref, à mon avis la méthode indiqué ici risque de fonctionner encore longtemps.

@crash : qu'elle erreur ?

@erdnaxeli : bien vu pour le list.txt. J'ai amélioré la commande du coup. Sinon la parallélisation est très intéressante dans ce genre de cas, car on traite du gros volume. Ici j'ai préféré me concentrer sur la méthode afin de publier un POC.

@bartounet : intéressant ton pourcentage. Donc en gros on pourrais récupérer la moitié des 35 millions d'adresses Gmail en vrai.

pti-seb , le 20 mai 2011 à 18:00

- En réalité c'est pas réellement un faille de google mais plutôt des utilisateur qui ne vérifie pas le paramètre de sécurité de leur donnée personnelle et des fois c'est le but d'avoir une adresse public.

xeros , le 20 mai 2011 à 19:32

- j'ai commencé à flipper à la vue de la simplicité du code mais après vérification c'est OK de mon coté.

  je sais pas pour plus pour quelles raisons mais j'étais déjà tombé sur ces paramètres et avais désactivé les redirections URL …

  merci pour cet article, c'est assez "magique" au premier abord

  et puis je vois pas comment on peut contourner ce type de "faille" défaut à partir du moment ou c'est volontaire …

  Sebastien

informatique Grenoble , le 20 mai 2011 à 20:00

- Bonjour Bonsoir,

  J'avais déjà lu la news sur korben.info
  mais la avec les commande toutes faite j'ai pas pu m'empecher de tester ^^

  par contre j'ai ta 3 ieme commande qui ne fait pas ce quelle doit faire alors je l'ai arranger sauvagement :
  cat profiles-sitemap.xml | grep "" | tr -d '\t' | sed 's/^\([^<].*\)$/\1/' | grep "^http" > list.txt
  je comprend pas bien a quoi sert le grep "" ??

  la ca mouline !!

Daguette , le 20 mai 2011 à 22:38

- Bonjour,
Je ne suis pas méga calé en informatique et je ne comprend pas comment faire pour les récupérer.
Je ne sais pas à quoi ça correspond les trucs du genre :
#!/bin/bash
for url in `cat $1`
do
curl --silent --head $url | grep "^Location:" >> getmail.txt
done

Quelqu'un pourrait-il soit m'expliquer soit me passer un lien ou les données finales sont téléchargeables s.v.p ?
Merci.

wth4 , le 21 mai 2011 à 12:33

- Je suis un n00b et je me touche la nouille.
Donnez moi de l'argent aussi pendant que vous y etes.

wth4 , le 21 mai 2011 à 15:08

- C'est pour mon apprentissage personnel. Tu crois que j'ai que ça à foutre de vendre des adresses mails ?
J'aimerai savoir un peu comment ça marche c'est tout.

wth4 , le 22 mai 2011 à 15:40

- **@wth4** : man bash

Fugitif , le 22 mai 2011 à 16:32

- **@Fugitif** : Ça a l'air plutôt compliqué... :/ J'en suis au XHTML et au CSS mais là ça a l'air d'être un niveau plus haut...
Tant pis ce sera pas pour aujourd'hui. Mais merci du renseignement, ma curiosité aura au moins pu être assouvie.

wth4 , le 22 mai 2011 à 22:18

- for url in `cat list.txt | grep http`; do wget --quiet $url; done;
vs
for url in `grep http list.txt`; do wget --quiet $url; done;

?

none , le 23 mai 2011 à 08:28

- **@Daguette** : il manquait le <loc> dans la commande grep, c'est corrigé.

@none : j'ai amélioré la boucle for comme tu nous le montre.

@wth4 : c'est des commandes bash, rien à voir avec le HTML et CSS. Il faut connaître linux assez bien pour comprendre ce qui est dit dans cet article.

pti-seb , le 23 mai 2011 à 08:37

- **@none** : +1
Beaucoup oublient que grep peu chercher tout seul dans un fichier, inutile de faire un cat avant.
Une commande que j'utilise souvent pour rechercher un bout de code dans une tripoter de répertoire.

grep -R 'mon_bout_de_code' *

Avec un alias dans le .bashrc pour coloré le grep et zgrep

alias grep='grep --color=auto'
alias zgrep='zgrep --color=auto'

Fugitif , le 23 mai 2011 à 13:36

-

à raison d'environ 1 seconde par requête curl+ajout dans le fichier d'adresses, 35 millions d'adresses avec une petite bécane, c'est environ 440 jours non stop. Sans compter que c'est 35 millions d'adresses dans le monde. Vas-y toi de trouver les adresses de tel ou tel pays 🙂

En résumé, la manip est super sympa à connaitre et permet d'apprendre des trucs, et en plus, elle n'est certes pas destinée aux "vrais" voleurs 🙂

hornetbzz , le 23 mai 2011 à 15:35

-

@hornetbzz : si tu parallélise la commande bash finale, tu peux réduire considérablement ce temps.

pti-seb , le 23 mai 2011 à 19:32

-

Pour le parallélisme, je vous conseil la conseil la commande 'parallel' de moreutils, et pour chercher du texte entre deux bout de text (ici et), j'ai bien la commande pcregrep :

parallel -j500 wget --quiet -- `pcregrep -o '(?<=(<loc>)).*(?=(</loc>))' profiles-sitemap.xml`

ravomavain , le 23 mai 2011 à 19:49

-

c'est pas dutout pro de la part de google,

j'ai fait un script de dl des adresse avec parallel (20 mail décodé /seconde)

je le laisse tourner jusqua que google sécurise sa plateforme

il faudra 30jours pour avoir les 35millions d'email mais j'espère que google va vite sécurisé ce truc

aliel , le 24 mai 2011 à 12:53

-

Seb, super ton tuto,

J'ai testé (juste 100 fichiers xml) et ça marche sans problème.

NimaX , le 24 mai 2011 à 12:54

-

Statistique intéressante, sur les 250 000 premiers profiles,
j'ai pu extraire 111 013 mail soit 44.052 % !!

ce qui fait potentiellement 14,2 millions de mail

jieff , le 29 mai 2011 à 10:16

-

Les apprentis spammeurs en profitent lol

Fugitif , le 29 mai 2011 à 17:28

-

oula :/, j'ose pas imaginer les dégât que sa va engendré quand chromeOS sera livré au grand public!, puisque l'O.S est exclusivement et entièrement dépendant d'un compte google

vincentpsp2 , le 6 juin 2011 à 08:39

-

@ravomavain : certes avec parallel tu pourra lancer 500 wget simultanément si tu le souhaite, mais à moins d'être en fibre optique avec 100Mb/s de débit, tu n'ira pas loin

avec une simple ligne adsl.
Avec un seul process de wget j'en suis déjà à 650ko/s de download.

Fugitif , le 21 juin 2011 à 15:54

-

Salut,

Moi aussi quand je suis tombé sur la news de Korben, je me suis lancé un script à ma sauce. Bon ce n'est pas aussi optimisé mais en lançant 10 terminaux j'arrivais à atteindre une bonne vitesse.
Je me suis areté hier à 12 millions de mails valides (reboot machine) et j'ai pas relancé, ça tourne depuis la news d'il y a 4 semaines ou plus.
Je doute qu'il y en ai plus de 15 millions ou alors je me suis trompé quelquepart.
Durant ce temps je me suis posé pas mal de questions et j'en ai déduit que les sitesmaps sont générable assez facilement en fait :
Il suffit d'aller sur Picasa, pointer sur une photo, voir le profil associé et l'email est dans l'adresse du profil. Quand l'utilisateur de la photo n'a pas de profil, le lien est en gris.
Si un robot scrute periodiquement les photos des gens il peut recreer la sitemap des picasseurs donc la sitemap serait bloquée depuis longtemps si il n'etait pas aussi simple de la recreer.
Ils doivent donc etre conscients du problème mais cette faille touchant la plupart de leurs applications ne doit pas etre simple à corriger.
Maintenant, en faisant ce que l'on fait : on diminue la valeur d'un mail et on augmente le spam chez eux.

Enfin, ce fut amusant de se mettre dans la peau d'un des premiers spameurs, à attendre que le robot recupere les sous : 0.001 centime, 0.002 centimes... 1 mois plus tard : 12000.001€, 12000.002€
lol

Bon qui veut des mails ? Je vous en donne 1000/allopass 😀
C'est ça oui ! je les gardes dans un coin, ca me servira peut etre un jour ?

Jack , le 4 juillet 2011 à 23:41

-

@Jack : Tes mails te serviront a quoi ? Tu passera surement pas l'anti spam de google

très longtemps. C'est un des meilleurs anti spam que j'ai vu.

Fugitif , le 6 juillet 2011 à 16:31

- Je peux toujours faire du spam manuel en ciblant très très fort

jack , le 11 juillet 2011 à 00:47

- **@jack** : avec un spam manuel tu gagnera pas grand chose.

Fugitif , le 13 juillet 2011 à 04:29

- C est effectivement hallucinant, je viens d essayer et je suis arrive a 59 445 900 profiles. Je vais lancer la recherche d emails valides.

cyberlord , le 19 août 2011 à 12:07

- salut tous le monde merci sebastien pour tes nouveauté, moi je suis pas fort en linux et j'aimme beaucoup informatique et tous ca je suis vraimen posseder par la connaissance, j'ai suivie le tuto mais a la fin quand j'introduit for file in `ls sitemap-*.txt` ; do ./getmail.sh $file; rm -f $file; done; dans terminal alors je recois des millier de ce message ./getmail.sh: Ligne 4: curl :commande introuvable

  quelqu'un pourrai m'aider a coprendre merci d'avance

apprenti , le 27 août 2011 à 19:48

- **@apprenti** : aptitude install curl

Fugitif , le 29 août 2011 à 18:06

- salut salut merci fugitif mon chapeau bas je vient d'installer curl et ca marche maintenant je suis bloquer dans la derniere command mon fichier getmail.txt de 2.6 Kio ya dedans location: /despseudo

  quand je met cat getmail.txt | awk -F'google.com/' '{print $2"**@gmail.com"}'** je verifier ca change rien pour quoi ?? crach a dit qu'il ya une erreur dans la derniere command ?? a oui sans oublier que tous ca je l'est fait en plus au moins 10 minute j'aretai le copiage juste pour avoire 12 a 13 fichier insi de suite
  pouvez vous me dire pour quoi la derniere command ne marche pas ? merci

derniere command , le 29 août 2011 à 23:33

- personne 😟 ?

derniere command , le 2 septembre 2011 à 19:41

- coucou les tuxiens toujour po de repense ?

derniere command , le 1 octobre 2011 à 22:06

- Ouvre ton fichier getmail.txt et regarde se que curl ta mis dedans et adapte ta commande.
  Je vois pas pourquoi ca t'intéresse autant, sauf pour pouvoir utiliser ces millions d'adresses pour spammer ?

Fugitif , le 2 octobre 2011 à 16:03

Ajouter un commentaire

A propos

 Créé en 2005, Tux-planet est un site qui a pour ambition de regrouper des articles sur Linux et le monde des logiciels libres. Lire la suite ...

Dernières réactions

(S'abonner ...)

- Anonymous: @clotaire : Bonjour. Permettez-moi ces deux questions : - Avez-bien bien installé...
- kakou: Est ce que quelqu'un sais comment consulter son forfait ?? Et avec un forfait à 2 euro est que les...
- pti-seb: @Felix_Faure @Vigor : ces distributions ne doivent pas encore faire partie du projet lsb. @Ellendhel :...
- clotaire: J'ai un probleme , lorsque je veut installer mes « paquets &raq uo; gnome tweak tool...
- fran6t: @makc : théoriquement la seule limite est celle de l'utilisateur d'un bon père de famille, 100...

En direct du Forum

(S'abonner ...)

- disparition du wifi
- OilRush

- bibliothèque a la windows 7 sur ubuntu 11.04
- Wifi neuf box?
- Plantage du serveur Tux-planet le 01 janvier 2012

Derniers Twitts

(S'abonner ...)

- Tux-planet : l'Empire Samsung - Un Oeil sur la Planete: http://t.co/qEvXrFDa
- Tux-planet : Je peux toujours pas consulter le solde de mon forfait 2€ chez Free. C'est prévu pour quand ?
- Tux-planet : Oil Rush est disponible dans la logithèque Ubuntu. http://t.co/KlvPRFfU

Des liens au hasard

- Korben
- Astuces facebook
- Application iPhone

- Trouvez toutes les offres d'emploi linux sur le moteur de recherche JobiJoba

actualité apache apple astuce astuces bash bilboblog blog boot chrome clavier commande commandes conky crack date debian Desktop développement exploit faille fedora firefox flash gimp gnome google graphique Graphisme hack hacking Hardware intel internet iphone jackalope jailbreak jaunty Jeux Kde kernel libre Linux log logiciels Logiciels Libres lucid lynx maemo mail maquette microblog microsoft mini-blog mobile mockup monitoring mozilla multi-touch musique mysql n900 nautilus nokia noyau openoffice open source password photos php Planet plugins publicité redhat red hat rpm réseau screenshot script serveurs shell sql ssh statistiques system sécurité thème tux-planet tv twitter ubuntu unity vidéo vidéos vlc voyage wallpaper windows wordpress yum