Groupe Avera : 4 rue du nouvelet – 94310 ORLY
Technapol S.A. au capital de 40 000 € - RCS Créteil 448 797 886
www.technapol.com - Tél : +33.148922244 - Fax : +33.148842312

Récupération distante d'informations

Introduction

La recherche d'information distante, préalable à toute attaque sur des réseaux ou des systèmes permet, dans le cadre d'un audit d'intrusion externe, de savoir comment on est 'vu' de l'extérieur (d'un point de vu informatique).
Cela permet ainsi de définir la surface d'attaque d'une cible et donc les axes qui seront potentiellement envisagés par un Hacker.

Cette méthode peut également permettre de collecter des informations précieuses dans le cadre d'une expertise informatique (Forensic). En effet, après avoir constaté l'adresse IP à l'origine d'une attaque, il est intéressant de pouvoir identifier des traces et remonter à sa source (notamment pour les applications juridiques qui peuvent en découler).

Pré requis:

- URL du site ou email original à étudier
- Récupération d'une adresse IP grâce à une analyse Forensic du système/réseau (suite à une intrusion)

N°TVA Intra : FR11448797886

Groupe Avera : 4 rue du nouvelet – 94310 ORLY
Technapol S.A. au capital de 40 000 € - RCS Créteil 448 797 886
www.technapol.com - Tél : +33.148922244 - Fax : +33.148842312

Il est courant d'entendre dire qu'il faut connaître les méthodes de l'attaquant afin de pouvoir se protéger efficacement, mais cette affirmation peut également s'appliquer dans l'autre sens. En effet, il est important lors d'un audit de savoir comment les personnes gérant le système d'information pensent et les méthodes qu'elles utilisent (par exemple, la convention de nommage des login utilisateurs ou des serveurs). Ces particularités pourront être exploitées lors de la phase de recherche d'information.

Cette phase peut suivre différents modèles selon les données que l'on souhaite obtenir et/ou l'audit à effectuer. Dans cet article, le modèle décrit sera destiné à l'audit orienté web.

Ce modèle va s'appuyer sur différentes étapes de dégrossissement. Tout d'abord, une recherche sera faite pour lister un maximum de serveurs susceptibles de contenir des plates-formes web appartenant à une cible donnée. Puis, une concentration toute particulière devra être portée sur l'ensemble des informations provenant de ces plates-formes. Enfin, certaines données sur le(s) utilisateur(s) / développeur(s) / administrateur(s) pourront être récupérées à partir des sites web cibles.

Bien sûr, si l'on ne dispose que d'un « nom prénom » on peut commencer ce processus par l'étape 3 et reprendre ensuite à l'étape 1 si l'on découvre, par exemple, que cette personne est liée à un site internet.

I Découverte topologique des cibles

Internet permettant d'interconnecter différents réseaux, il est important de ne négliger aucun détail susceptible de faciliter une attaque par rebond conduisant à la cible principale.
Ainsi il n'est pas toujours nécessaire que le site web principal de la cible soit faillible pour s'y introduire. Il est en effet fréquent de découvrir des sites/serveurs annexes qui permettent de rebondir sur le site/serveur principal. Devant ce constat, cette étape doit permettre d'énumérer les plages d'IP utilisées, les sous-domaines de la cible et les serveurs potentiellement intéressants.

Un audit débute généralement à partir de l'URL d'un site cible. La démonstration qui sera faite ici, s'articulera autour d'un site réel rebaptisé « blah.com » afin de préserver son anonymat. La première étape est donc

N°TVA Intra : FR11448797886

de résoudre le nom de domaine en une adresse IP, ceci s'effectue par l'intermédiaire de « dig » sous UNIX:

```
$ dig blah.com +short
217.118.84.30 (l'IP a été modifiée)
```

La seconde étape consiste en un « whois », il s'agit d'interroger les bases de registres (des noms de domaines et des plages d'IP) afin d'en connaître le propriétaire. Cette étape est totalement passive puisqu'elle s'effectue sur des organismes externes à la cible, les NIC (Network Information Center).

Lors d'un « whois » sur blah.com:

```
$ whois blah.com
```

Il ressort que le nom de domaine à été enregistré chez Gandi, avec les coordonnées de la société propriétaire dont le siège se situe à Lille (l'adresse du siège figure dans le whois). Il y figure également le numéro de téléphone du propriétaire (la société blah) ainsi que celui du service technique. Une recherche supplémentaire sur le site « pagejaunes.fr » avec comme nom de société « blah », indique en plus le numéro de téléphone du service client.

Un « whois » sur l'adresse IP du site permet quant à lui d'obtenir la plage IP appartenant à BLAH:

……

```
inetnum: 217.118.84.0 – 217.118.84.255.
```

……

Depuis cette plage IP, il est possible d'effectuer du « ping sweeping », afin de savoir quels sont les différentes machines « up ». Cela s'effectue depuis une commande « nmap »:

```
$ nmap -sP 217.118.84.*
```

Une fois la liste des machines actives récupérée, il est intéressant de tester la présence d'un serveur web sur celle-ci. En effet, si un tel serveur était découvert, cela procurerait un point de rebond intéressant puisque cette machine appartiendrait à la plage IP de blah. Une autre utilisation possible de cette plage IP, consiste à procéder à un « reverse lookup dns » sur chaque IP appartenant à cette plage afin de retrouver à quel nom de domaine celle ci est associée. La commande « nmap » sera utilisée une nouvelle fois pour effectuer cette étape :

```
$ nmap -sL 217.118.84.* |grep \( | grep Host | awk '{print $2,$3}'
```

Les résultats renvoyés par cette commande, permettent d'identifier les machines suivantes :

`gw.blah.com , mail01.blah.com , mailing.blah.com` et `beta.blah.com`

Ce sont des données très intéressantes! En effet, on peut deviner l'utilité de chacune de ces machines. « gw » à des chances d'être le diminutif de gateway. « mail01 » est très certainement le serveur mail avec 01 comme identifiant du serveur (peut être en existe-t-il plusieurs, il serait judicieux de tester l'existence d'autres serveurs mail en incrémentant l'identifiant: mail02, mail03, …). « mailing » peut correspondre à la partie gestion des mailing-lists. « beta » est peut-être un serveur de test où de nouvelles fonctionnalités sont mises à l'épreuve, et la sécurité est parfois moindre que sur le serveur principal car il n'est pas censé être connu du public.

Bien sûr, il ne s'agit pour le moment que de suppositions, mais si elles s'avèrent exactes, ces machines peuvent être intéressantes. Il faut aussi se rappeler que certaines de ces machines ne sont pas censées être connues du grand public, ainsi parfois leur accès n'est pas réglementé (sécurité par l'obscurité quand tu nous tiens..) et des informations confidentielles peuvent s'y trouver!

Il faut à présent vérifier la théorie selon laquelle « mail01 » serait le serveur mail de blah.com. Pour ce faire, il faut à nouveau utiliser la commande « dig », mais cette fois-ci en lui spécifiant que seules les entrées « MX » sont nécessaires. Cette commande permettra de récupérer les adresses des serveurs mails associés au nom de domaine:

```
$ dig -t MX blah.com
……
blah.com.   3600  IN    MX    5 mail01.blah.com.
blah.com.   3600  IN    MX    50 mail02.blah.com.
……
```

Les résultats montre bien la présence de mail01 en tant que serveur mail, mais également celle de mail02. La théorie évoquée précédemment semble être exacte.

Une autre utilisation intéressante de « dig », est l'extraction des entrées NS qui correspondent aux serveurs DNS du domaine:

```
$ dig -t NS blah.com
……
blah.com.         3402  IN    NS    ns.blah.com.
blah.com.         3402  IN    NS    ns1.blah.com.
……
```

Depuis ces serveurs NS, il est possible d'exploiter un « transfert de zone » consistant à abuser de la mauvaise configuration de certains serveurs NS afin d'obtenir leur liste d'entrées pour un hôte donné. Normalement ce

mécanisme est utilisé entre plusieurs serveurs NS afin de répliquer la configuration de l'un vers les autres. En toute logique, seuls les serveurs devant recevoir la configuration du serveur « maitre » devraient être capables d'effectuer un transfert de zone, malgré tout il n'est pas rare d'avoir une configuration laxiste qui permette à n'importe qui de demander un transfert de zone à un serveur NS. Une tentative de transfert de zone s'effectue par la commande « host »:

```
$ host -l blah.com ns.blah.com
```

Mais celle ci échoue pour le serveur ns.blah.com, il semblerait que celui ci soit bien configuré, ce qui est souvent le cas des serveurs NS primaire, mais heureusement pas celui des serveurs NS secondaire. Il faut donc maintenant tester ns1.blah.com:

```
$ host -l blah.com ns1.blah.com
……
blah.com name server ns.blah.com.
blah.com name server ns1.blah.com.
blah.com has address 217.118.84.45
mrtg.blah.com has address 217.118.84.45
smtp.blah.com has address 217.118.84.45
webmail.blah.com has address 217.118.84.45
stock.blah.com has address 217.118.5.120
pop.blah.com has address 217.118.84.45
mail.blah.com has address 217.118.84.45
beta.blah.com has address 217.118.84.246
stats.blah.com has address 217.118.5.121
*.blah.com has address 217.118.84.45
```

Bingo! Le serveur NS secondaire n'est pas protéger et affiche la configuration mise en place sur celui-ci.

Cette configuration permet de noter différentes choses. Tout d'abord, la récupération d'une liste de sous-domaines qui permettra peut être d'exploiter quelque chose. Ensuite, elle révèle également que les sous-domaines « stats » et « stock » ne sont pas situés sur la même plage IP que celle découverte précédemment. Il s'agit peut être d'une autre plage IP appartenant à BLAH. Il suffit de procéder comme précédemment pour le vérifier.

Un dernier point, celui-ci moins sympathique, concerne la présence d'une entrée « *.blah.com », indiquant que tout les sous-domaines non déclarés explicitement pointent sur l'IP du site principal de blah.com. Ceci limitera donc l'étape suivante qui consiste en l'énumération des sous-domaines de blah.com. Pour preuve, si un faux sous-domaine est testé, au lieu de ne pointer sur rien, celui-ci pointera sur blah.com. Il est ainsi difficile de distinguer un sous-domaine valide d'un faux.

La dernière étape de la première partie de recherche d'information est primordiale de par son efficacité. En effet, il n'est pas rare de trouver des vulnérabilités bien plus facilement sur un sous-domaine de la cible

que sur le site principal. De plus, l'utilisation d'un sous-domaine lors d'une attaque permet parfois d'être plus discret (potentiellement moins surveillé que le domaine principal).

C'est même souvent un bon moyen de rebond : le sous-domaine peut être hébergé sur le même serveur et/ou partager le même serveur SQL. Dans certains cas, le serveur SQL est lui aussi mal configuré et ceci permet de consulter l'ensemble des bases de données qu'il contient (et donc parfois celle du site principal ou même des bases de données réservées à l'usage interne à l'entreprise cible).

Note : Il arrive, lors de l'énumération des sous-domaines, de tomber sur un sous-domaine n'appartenant pas à une des plages IP découvertes jusqu'ici. Il est bien entendu nécessaire dans ce cas de recommencer la recherche d'information à partir de la nouvelle adresse IP découverte afin, encore une fois, d'étendre le périmètre d'investigation.

Pour énumérer ces sous-domaines plusieurs techniques existent, plus ou moins discrètes, elles sont aussi plus ou moins efficaces. Bien sûr, il est possible d'utiliser divers services web afin de rester passif dans la récolte d'informations. Dans la liste de ces différents services, on peut citer le site netcraft.com qui, de part son moteur de recherche, apporte énormément d'informations.
Le test de *.blah.com sur netcraft.com, permet de retourner plusieurs résultats:

```
ak.blah.com
clients.blah.com
secure.blah.com
www2.blah.com
```

Parmi les autres informations importantes que stocke netcraft, il y a la version du serveur web ainsi qu'un historique des IP utilisées par ces serveurs. Ceci peut permettre de détecter d'anciennes plages IP utilisées publiquement par la cible. Ces plages IP peuvent contenir encore des informations intéressantes, voire même des points d'entrées pour divers rebonds dans le cas où elles seraient toujours utilisées par la cible (mais pour un usage privé). On retrouve majoritairement ici des serveurs sous distribution Debian et utilisant le couple apache2/php5.

En dehors de cela, il est également possible de se servir de netcraft afin d'énumérer les noms de domaines dans lesquels blah figure. Ainsi, les résultats obtenus peuvent potentiellement appartenir à la cible. Dans le cas présent, netcraft ne retournera que www.blah-site.com.

Une autre méthode consiste à utiliser Google et ses opérateurs tel que 'site:' afin de lui demander d'afficher tout les sous-domaines étant la propriété de la cible:

site:blah.com -site:www.blah.com

Malheureusement, pour le cas de blah.com, aucune page en dehors de celles appartenant au domaine principal n'a été indexée.

Une autre technique, bien que moins discrètes consiste à effectuer un brute force à partir d'un dictionnaire de sous-domaines possibles (admin, team, www2, backup, ...). Cette technique renvoie des résultats souvent utiles, mais ici elle n'est pas applicable du fait de la présence de *.blah.com dans les entrées dns comme expliqué ci-dessus.

Pour contourner cette protection, il est possible de repérer un « motif » apparaissant dans la page de redirection des noms de domaines non existant. Puis d'effectuer un brute force (avec une très petite liste de sous-domaine possible) en vérifiant que ce « motif » n'apparait pas dans les domaines testés. Si c'est le cas, le domaine est bien un sous-domaine. Cette technique n'est pas très fiable, assez lourde et de ce fait quasiment jamais utilisée. Elle ne sera donc pas traitée dans cet article.

En utilisant ces techniques d'énumération des sous-domaines et à partir d'une liste de résultats, une convention de nommage pour les sous-domaines peut être identifiée. Par exemple, si un sous-domaine team01 est détecté, il est possible qu'un sous-domaine team02 existe. Il est également possible de tomber sur des sous-domaines faisant références à un domaine en particulier comme des noms de dieu grecs ou d'atomes. Il est donc nécessaire d'établir une 'wordlist' basée sur ce thème, afin de tester tout un ensemble de nouvelles possibilités. Il est aussi important d'analyser correctement les résultats renvoyés par nos outils de récupérations de sous-domaines afin d'avoir une vision la plus exhaustive possible. Ce processus d'analyse peut être automatisé en incrémentant les valeurs numériques et en utilisant http://labs.google.com/sets, comme le fait « DNSPredict ».
Cet outil permet de créer divers « ensembles » en y injectant les paramètres des différents sous-domaines ciblés. Si pour le cas présent un test est effectué avec les paramètres suivants:
`beta`, `stats`, `pop`, `webmail` et `stock`. La liste retournée contient `imap`, `email`, `ftp` et `server` représentant des sous-domaines existants chez blah.com

Il est parfois possible que le site audité soit difficilement pénétrable. Dans ce cas, il est préférable de tenter d'énumérer les différents sites hébergés sur le même serveur. Ils pourront peut-être contenir des pages faillibles permettant par la suite, de prendre la main sur tout le serveur. Il n'est en effet pas rare de voir des serveurs web héberger plusieurs sites du même groupe. Cette étape va donc fournir une liste de nouveaux sites cibles qui pourront subir les étapes précédentes et être audités.

La récupération de ces informations se faisant par l'intermédiaire de service en ligne, cette phase est totalement passive, ce qui est d'autant plus intéressant en termes d'anonymat.

Parmi les services web proposant cette fonctionnalité, il est à noter la présence du moteur de recherche « live » de Microsoft, qui de part sa directive « IP » permet de lister l'ensemble des sites référencés possédant l'IP fournie. Il existe aussi d'autres sites comme serversniff.net ou encore myipneighbors.com. Bien sûr, il en existe d'autres. Mais certains requièrent une authentification sur le site, d'autres nécessites la validation d'un captcha afin d'accéder aux résultats et d'autres encore sont payants, ce qui limite la liste des sites pouvant être utilisés par des outils d'automatisation.

Après l'utilisation de ces différents sites et l'analyse de ces informations de nouveaux sous-domaines ont été obtenus: v2 et jeu. Mais également l'existence d'un blah.fr et d'autres sites reliés a la société blah, concernant la vente de sac d'aspirateur et de recherche de cadeau. Il est évident qu'il faut reprendre les étapes précédentes pour tous ces nouveaux domaines.

II Approfondissement des résultats

Dès qu'un ensemble d'hôtes possédant un serveur web a été repéré, il faudra se concentrer sur les informations récupérées à partir des sites web hébergés. En effet, une grande quantité d'informations peut être récupérée de façon totalement passive en effectuant uniquement des requêtes à partir de sites en ligne proposant des services de recherche (Google, live, Yahoo,...) ou SEO (Search Engine Optimisation).

Le cas des sites de SEO sera traité un peu plus tard. Comment, à partir de quelques exemples, récupérer des informations sur un site grâce à Google.

Google possède encore une fois tout un ensemble d'opérateurs pouvant être très utiles, leur utilisation dans le cadre de la sécurité informatique est souvent connue sous le nom de « Google hacking ». Parmi eux, il y a l'opérateur « site: » qui restreint l'affichage des résultats aux pages appartenant à un domaine spécifié. Cet opérateur nous permet d'effectuer plusieurs actions: l'énumération de l'ensemble des pages indexées sur le moteur de recherche (qu'il est possible de consulter par le cache), à partir de cet ensemble, une partie de la structure du site peut être récupéré(les dossiers utilisés pour stocker les fichiers inclus par exemple, ou encore de discerner le style de nommage des dossiers ou des

fichiers, ce qui permettrait d'effectuer ensuite un brute force localisé si besoin).

Par exemple, le site blah.com contient des pages web en php, l'affichage d'erreurs sur la page s'effectue par l'utilisation d'un mot clé « Warning », « Parse Error », ou encore « Fatal Error ». En utilisant cette particularité et l'opérateur « site: » il est possible de lister l'ensemble des erreurs du site qui ont été indexées par Google. Grâce à ces erreurs, le fullpath de la page web ou encore la fonction et la ligne les ayant causées peuvent être récupérés. Ces informations permettront de déduire la structure du code de la page (pour, par exemple, retracer les actions effectuées sur des variables). Elles permettront également de connaître la méthode utilisée par le webmaster pour traiter certaines entrées ou actions (utilisation ou non de fonction permettant l'échappement des caractères spéciaux dans une requête sql, ou encore la sécurisation avant l'affichage de donnée sur le site).

Ici une requête « `site:blah.com 'Warning: ' php` » ne révèle rien, par contre la même requête en remplaçant cette fois `'Warning: '` par `'Fatal Error'` renvoie des résultats intéressants.

Il apparaît en effet que le dossier contenant les pages web est « /var/www/v2/ ». Il est également possible, grâce à ces erreurs, de lister certains fichiers situés dans des dossiers comme « obj », « inc » ou « skin_c » (les extensions des fichiers situés dans skin_c laissent présager qu'il s'agit de fichier templates). Une requête avec `'Parse Error'` renvoie également le même type de résultats.

De façon plus générale, des motifs courants peuvent être utilisés pour détecter les erreurs d'un site qui utiliserait un affichage personnalisé de celles-ci. Par exemple « `Erreur sql:` » suivis parfois de la requête sql peut permette de cibler d'éventuelles failles mais aussi de récupérer des noms de tables et de champs utilisés dans la base de données.

Un autre opérateur intéressant est « `filetype:` » qui filtre les résultats pour n'afficher que ceux dont le type de fichier correspond au paramètre. Cela permet de lister des fichiers potentiellement sensibles comme ceux dont l'extension est .inc, .sql, .zip (peut être un backup), .txt, .php~, .bak, .old, ... Mais ici, rien ne sera trouvé pour blah.com.

Un outil existe, permettant la récupération d'un certains nombre de « fichiers bureautiques » (ppt, pdf, doc, xls, ...) sur le site cible en passant par une recherche Google, il s'agit de metagoofil. Cet outil permet aussi d'extraire les métadonnées des fichiers récupérés, afin de lister les informations comme le path du fichier sur le disque local de la personne l'ayant créé (pc interne a la société), le nom et le prénom du rédacteur ou

créateur, le nom d'utilisateur (permet parfois de déduire la convention de nommage utilisé dans l'entreprise pour les logins). Ici metagoofil ne renvoie rien d'intéressant. En revanche, les fichiers pdf qu'il a trouvé sont stockés dans des dossiers dont les noms sont intéressants « img_upload » et « docs ». Il est possible qu'il existe une zone permettant l'upload de fichier (à garder en mémoire !).

Précédemment une partie de l'architecture du site à pu être récupérée grâce à l'utilisation de Google, il est possible de poursuivre cette reconstitution en utilisant d'autres services web afin de rester passif vis-à-vis de la cible. Certains sites, ayant pour thème l'optimisation du référencement des sites web, proposent des pseudos crawler en ligne. Le service permet de simuler le passage d'un robot indexeur et retourne la liste des URL présentes dans l'adresse de la page fournie. Ainsi en passant l'URL racine du site, puis en faisant appel de façon récursive aux différents liens énumérés, il est possible de crawler de façon passive le site.

Malgré tout, la majorité des services de ce genre présentent des inconvénients, comme l'obligation de remplir un captcha avant chaque recherche, le non affichage des liens vers les images/css/js ainsi que le contenu du site, ...

Le site webuildpages.com propose l'un des meilleurs « fake crawler » existant sur le net. En effet, il affiche pour une URL donné, la liste des liens trouvés (y compris les images, css, js, …). Il affiche également le contenu de la page ainsi que l'intégralité du code source (html) de celle-ci! Ceci permet de naviguer et même d'aspirer le site de façon passive en appliquant des filtres supplémentaires à ceux déjà effectués par ce service.

Maintenant que des informations concernant la cible ont été récupérées de façon passive, il est temps de s'y connecter afin de pouvoir continuer notre analyse. Durant cette phase, quelques indications pour scripter un petit crawler orienté information gathering seront données.

Pour commencer, il est conseillé d'aspirer le site en entier ou en partie en fonction des besoins et de sa taille. Bien sûr l'aspiration du site devra être réalisée de façon intelligente et pourra se faire par l'intermédiaire de proxy. Elle devra également être faite à différentes périodes afin de rester assez furtif.

Le but de cette récupération des pages est de pouvoir, si nécessaire, effectuer ensuite des analyses « off line » à n'importe quel moment de l'audit.

Le crawler qui aura pour rôle le parcours et l'enregistrement des différentes pages, peut utiliser une base de donnée afin d'enregistrer l'ensemble des URL parcourues, ceci dans plusieurs buts:

- Pouvoir reconstruire l'arborescence du site facilement

Groupe Avera : 4 rue du nouvelet – 94310 ORLY
Technapol S.A. au capital de 40 000 € - RCS Créteil 448 797 886
www.technapol.com - Tél : +33.148922244 - Fax : +33.148842312

- Permettre l'étude du format des noms de pages ou des dossiers afin de pouvoir constituer une liste probable de dossiers existants et non listés lors du crawling
- Permettre de détecter si chaque URL contient des formulaires et/ou des données passées en GET (indiquant alors de potentiels points d'entrée)
- Pouvoir suivre l'avancement de la vérification des pages « sensibles » (cf. ci-dessus).

Il est également intéressant d'intégrer une détection des données telles qu'un numéro de téléphone, une adresse email…, dans la phase d'analyse du crawler, afin de pouvoir s'en servir pour différents usages ultérieurs :

- Social Engineering
- Étude du format des adresses emails de la cible (nom.prénom, p.nom,...)
- Point d'entrées possibles

A noter également qu'à chaque dossier découvert dans une URL parcourue par le crawler, celui-ci devra également tenter d'accéder à ce dossier seul. Ces tentatives d'accès pourront peut être permettre de découvrir des « index of/ » pouvant contenir des informations sensibles. Malheureusement, dans le cas de blah.com les « directory listing » ont été désactivés.

Un fichier important que notre crawler doit aussi prendre en compte est le « robots.txt » à la racine du site, donnant les règles que doivent suivre les crawlers des moteurs de recherches. En effet, ce fichier spécifie les dossiers étant autorisés à être parcourus par le crawler dans le but de les indexer. Il spécifie également les dossiers non autorisés, c'est-à-dire ceux qui ne doivent pas être indexés (ce sont à priori ces dossiers qui vont être intéressants). En effet, il est parfois possible qu'un webmaster indique des dossiers sensibles (comme le dossier ou se situe la console d'administration du site) afin que ceux ci ne se trouvent pas indexés sur l'Internet. Or, ils ne sont censés être parcourus que si un lien pointe vers ce dossier. Dans le cas d'un dossier sensible il est évident que l'administrateur ne fera pas pointer de lien depuis les zones publiques du site vers celui-ci.

Ainsi, le crawler se devra de parcourir tout les dossiers étant considéré comme « disallow ». Attention, il est possible que la cible intègre un système « anti-aspirateur » ou « anti crawler trop curieux », qui bloque l'IP de la personne ayant tenté d'accéder à un faux dossier marqué comme « disallow ».
Le fichier robots.txt du site blah.com contient:

```
User-agent: *
```

```
Disallow: /js/
Disallow: /adm/
Disallow: /bat/
Disallow: /comparo/
Disallow: /pop/
Disallow: /obj/
Disallow: /mod/
Disallow: /inc/
Disallow: /skin/
Disallow: /skin_adm/
Disallow: /skin_c/
```

Une chose amusante pour le cas du site blah.com, est le dossier /pop/ que l'on retrouve dans son robots.txt. Celui-ci contient une page de type popup ayant pour but d'afficher une animation flash, mais il contient aussi un code javascript malicieux qui une fois exécuté se « décrypte » et affiche une iframe vers d'autre javascripts malicieux. Ce code est le signe que le site est infecté et il y a donc de fortes probabilités que celui-ci soit faillible. Cette piste ne sera pourtant pas détaillée dans cet article dont ce n'est pas l'objet.

Hormis cela, nous retrouvons certains dossiers listés de façon passive précédemment (comme obj, ou inc). Mais, le nom de certains dossiers comme « adm » sont très attirant. En effet, ce dossier tend à correspondre à administration, ce qui sera confirmé plus tard lors de la tentative d'accès, un .htaccess demande une authentification pour accéder à une zone de « statistiques ».

Il est aussi possible de fournir au crawler une liste de dossiers dont le nom paraît sensible et pouvant exister sur le système cible. Il devra également tenter de les parcourir tout en restant furtif : admin/, config/, includes/, lib/, membres/, logs/…

Voici une liste non exhaustive de fonctionnalités qui peuvent être intéressante d'intégrer au crawler:

La détection de présence d'erreurs sur la page (comme « warning: » du langage php). En effet, celles-ci sont éloquentes et donnent beaucoup d'informations à un attaquant. Cependant, les erreurs indexées par Google ne sont plus forcément à jour, et une page ayant été rajoutée récemment peut ne pas avoir été indexée. Il est donc nécessaire de faire cette recherche en « live » et ne pas compter uniquement sur le contenu des moteurs de recherche.

La détection de commentaires dans la page.

Parfois, ces commentaires peuvent être des notes laissées par les webmasters, ou du code html affichant des liens vers certaines pages du site internet étant en maintenance ou non utilisées à l'heure actuelle. Ces pages sensibles n'étant pas considérées comme visible par un visiteur, il n'est pas exclus que le niveau de sécurité y soit moindre. Il se pourrait également qu'une de ces pages comporte des bugs rendant son utilisation impossible pour un visiteur classique mais pouvant nous indiquer des informations pertinentes. A noter que le site archive.org ou le cache de Google, peuvent contenir des images « passées » du site cible. La récupération de ces « backup » au regard des pages actuelles peut révéler le même genre d'informations que les mises en commentaire de certaines parties de code HTML.

Il existe également différentes tactiques qui permettent de rendre le site cible plus 'bavard'. On peut par exemple évoquer la méthode du bug du tableau en php. Cette méthode consiste, si le site cible utilise le langage php, à faire passer des tableaux dans les variables de l'URL de la façon suivante :
`http://site.com/page.php?bla[]=valeur`
au lieu de
`http://site.com/page.php?bla=valeur.`

Cela aura pour effet, dans de nombreux cas, de faire afficher des warning sur la page puisque les fonctions ne s'attendent généralement pas à recevoir des tableaux en variables.

Nous pouvons enfin évoquer l'utilisation de scripts « open source » tels que joomla, ou spip utiles pour la génération des sites. Ces détections peuvent être ajoutées à la liste des opérations effectuées par notre crawler puisque ces types de scripts utilisent souvent des « signatures ».

Ces signatures peuvent revêtir plusieurs aspects : l'affichage en clair d'une mention sous la forme « fait par *nom_du_script* » ou plus discret avec des balises et propriétés html classiques du script:

```
<a href="http://www.spip.net/" title="Site realise avec
SPIP"><img src="dist/spip.png" alt="SPIP" width="48"
height="16" class="format_png" /></a>
```
Ou encore,
`propulsé par fluxbb,...`

Ce genre de signature permet de cibler un ensemble d'attaques possibles basé sur de l'audit white box (à partir du code source) ou encore à partir de vulnérabilités déjà découvertes et publiées sur l'Internet.

De plus, il est possible de détecter « précisément » la version des scripts utilisés, soit parce que ceux-ci intègrent la version dans leur « signature », soit parce que le package contenant le script, fournis parfois des fichiers lisible (txt, html,..) indiquant la version utilisé (ex: un changelog) : « dotclear/CHANGELOG ».

A ce stade, nous devrions avoir obtenu suffisamment d'informations sur la cible, mais il ne faut pas omettre de naviguer sur le site 'à la main' et de rechercher des pages telles que « recrutement ». Dans le cas de blah.com, il est indiqué que la société recherche un chef de projet php/mysql. Grâce à cette information, l'utilisation du php par la société Blah est confirmée et le système de base de données utilisé est découvert. L'étape suivante sera alors la récupération d'informations sur les « utilisateurs ».

III Récupération d'informations sur les utilisateurs

Lors des phases précédentes, il a normalement été possible de récupérer des listes d'adresses emails, de noms et prénoms, numéros de téléphone ou autres informations appartenant à des employés de la cible. Ces employés peuvent être une source d'information très intéressante ou même un vecteur d'entrée sur la cible si l'attaquant est en mesure de les exploiter. C'est le but de cette partie qui donnera quelques pistes sur la recherche d'information mais cette fois spécialisée sur les utilisateurs.

En ce qui concerne les emails, il est possible de récupérer des emails appartenant à des employés de blah.com en utilisant Google. Ce qui permettra de ne pas se focaliser sur celles uniquement présente sur le site de blah. Pour ce faire, il faut utiliser la requête suivante sur Google « *mail *@blah.com », celle-ci renvoie plusieurs résultats comme: sav@blah.com, service-client@blah.com et le mail de deux employés.

Un des sites intéressants pour cette dernière étape de la recherche d'information, est le site « societe.com » qui, en plus de fournir l'état financier d'une société, indique également les noms de ses dirigeants.

Parmi la liste des noms récupérés, il va falloir relever un maximum d'informations sur les personnes les plus influentes dans la société. Ceci pourra permettre de consolider un scénario dans une phase de « Social Engineering » ou d'extrapoler des identifiants (cas de mots de passe faibles).

Ces mots de passes faibles pourront être déduits à partir des informations récupérées, d'où l'utilité de cette étape.

Groupe Avera : 4 rue du nouvelet – 94310 ORLY
Technapol S.A. au capital de 40 000 € - RCS Créteil 448 797 886
www.technapol.com - Tél : +33.148922244 - Fax : +33.148842312

A noter également, qu'il peut être utile de générer une liste possible de logins basés sur les noms/prénoms découvert. Cela pourra peut être permettre de découvrir des sites fréquentés par la personne cible.
Pour les générer il faut utiliser les formats qu'on retrouve le plus souvent: dupont.albert, a.dupont, duponta, dupont, … dans le cas ou notre employé se nomme albert dupont.

Une fiche d'identité de chaque personne potentiellement intéressante devra être établie pendant l'audit. Pour cela, il est bon d'utiliser les réseaux sociaux, ainsi que quelques sites de renseignements. Parmi les sites de renseignement, les pages jaunes serviront parfois à récupérer l'adresse postale de l'employé ainsi que son numéro de téléphone ou encore à retrouver le propriétaire d'un numéro de téléphone grâce a son service d'annuaire inversé. Dans ce cas, nous pouvons également citer le site de Free qui, pour vérifier l'éligibilité d'une ligne, transmet certaines informations utiles.

Afin de tester la présence de certains autres employés de blah.com encore non identifiés, le site de Google peut une nouvelle fois être mis à contribution en utilisant la requête suivant: « blah.com site:copainsdavant.linternaute.com ».
Cette recherche révèle plusieurs employés actuels non découverts précédemment. Mais également la présence d'un groupe « blah.com » constitué d'employés inscrits sur copains d'avant.
Une simple inscription sur copains d'avant permet ensuite d'accéder aux profils de ces employés et de récupérer des informations privées.
La rubrique « Mon parcours » est ici importante, puisqu'elle indique le poste occupé par cette personne dans la société blah.com. Il est également possible de récupérer des informations telles que la ville de résidence, l'âge, les passions mais également parfois une photo de cette personne.
Parmi les employés inscrits sur ce site, il est à remarquer la présence d'un responsable du service informatique, deux webmasters, une commerciale ainsi que le pdg de blah.com.

D'autres recherches du même genre peuvent être effectuées sur facebook, en utilisant par exemple une photo d'un employé trouvé précédemment, afin de créer un faux compte et de devenir « ami » des employés de blah.com

Un autre site pouvant être utile ici est 123people.com qui recherche de façon automatisé sur l'Internet des informations sur une personne à partir de son nom et de son prénom. On retrouve dans ces informations les adresses emails, page web où ce nom est cité, son employeur…

Groupe Avera : 4 rue du nouvelet – 94310 ORLY
Technapol S.A. au capital de 40 000 € - RCS Créteil 448 797 886
www.technapol.com - Tél : +33.148922244 - Fax : +33.148842312

Enfin pour linkedin.com qui se spécialise sur le parcours professionnel des utilisateurs inscrits, la même procédure que pour copains d'avant doit être suivie. La requête « site:linkedin.com blah.com » apporte également de nouveaux employés.

Au final la méthode est souvent la même, il suffit juste de l'appliquer sur les différents sites cités ci-dessus et d'éplucher les profils afin de récupérer une quantité non négligeable de données privées.

Conclusion

Grâce à ces trois grandes étapes, nous avons pu collecter suffisamment d'informations pour débuter l'audit de blah.com dans les meilleures conditions.
La recherche d'information est vraiment l'étape incontournable et préalable à tout audit. Celle-ci peut s'effectuer en majorité 'par rebond' via des sites dont la puissance est souvent méconnue. Il est ainsi possible de rester plus furtif vis-à-vis de la cible et peut être intéressant pour tester la réactivité d'une équipe informatique face à une attaque non planifiée (mais dans un cadre légal) de son Système d'Information.
L'utilisation de ces méthodes peut également permettre d'obtenir rapidement des informations utiles au traitement juridique d'un litige (intrusions informatiques, vol d'informations confidentielles, concurrence déloyale…).

André MOULU

N°TVA Intra : FR11448797886