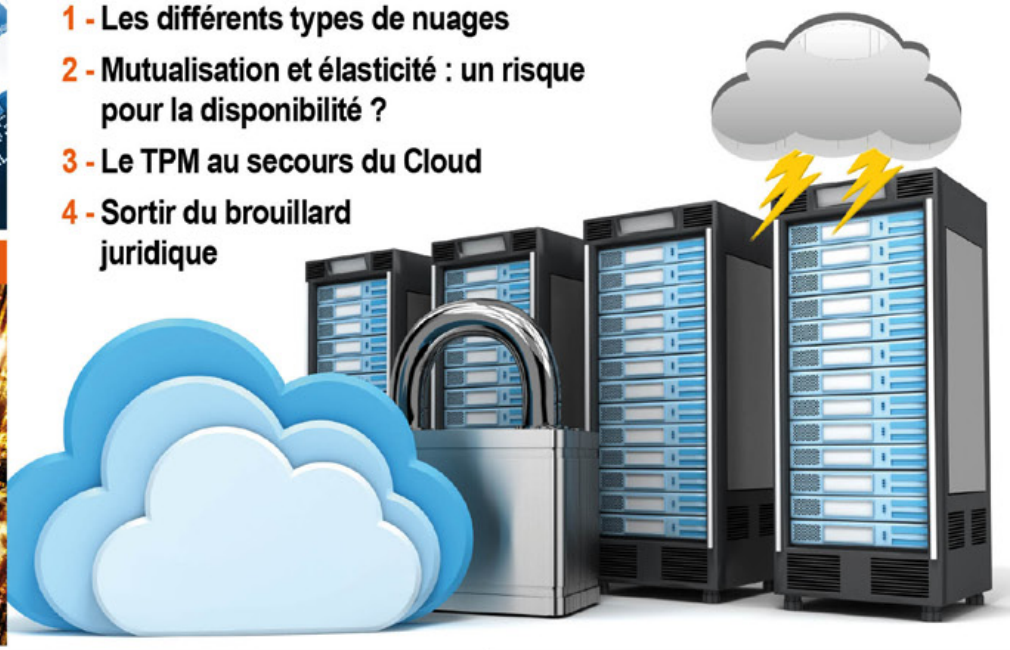
MISC
Multi-System & Internet Security Cookbook
100 % SÉCURITÉ INFORMATIQUE
L 19018 - 76 - F: 8,90 € - RD
N° 76 NOV./DÉC. 2014
France METRO : 8,90 € - CH : 15 CHF - BE/PORT CONT : 9,90 € - DOM TOM : 9,50 € - CAN : 16 $ cad - Maroc : 110 MAD - Tunisie 19 TND



ÉDITO

Mammouthification vs coopération (qui aime bien châtie bien)

Commençons par une banalité : l'affaire Snowden, qui nous envoie de *bons baisers de Russie*, a été un électrochoc. Ceci dit, quelques mois auparavant, l'intrusion de Bercy aussi. Tout comme celles de l'Élysée, de Belgacom, d'Areva, bref, la routine est faite d'*Opérations Tonnerre*. Elles ont toutes en commun d'être récentes. Difficile de trouver trace d'une grosse intrusion en France avant 2010.

Depuis, la gestion de crise se professionnalise, s'organise comme LE truc à la mode où tout le monde est expert (que tous les CERT en France lèvent le doigt), mais où concrètement, seule l'ANSSI chez nous, et 2-3 boîtes non françaises ont géré des crises de bout en bout. Si on s'y prend mal, on tape sur un malware, mais il revient et *meurt un autre jour*... quand il meurt.

Et justement, pas mal de choses bougent sous l'impulsion de l'ANSSI, bien que dans une position ambiguë :

- réglementation... qui conduit globalement les gens (même en interne) à se plaindre de la lourdeur administrative bien de chez nous. Vive la centralisation et cette habitude d'empiler des strates de hiérarchies plus ou moins compétentes pour assouplir le processus #ironie. S'installe alors le rythme administratif : *vivre et laisser mûrir*.
- les « clients » de l'ANSSI : l'État et les OIV (Organisme d'Importance Vitale). Ça couvre du monde, mais hors de ça, que se passe-t-il : *pwner* **n'est pas jouer** ? Sans oublier les nouvelles prérogatives liées à la LPM comme la contre-attaque.
- les actions de l'ANSSI vers l'extérieur : les avis du CERT FR (merci pour la planète, on ne risque pas la déforestation si on les imprime), des guides #nocomment et de temps en temps, un outil (oups, 6 sur le GitHub) ou CLIP, un Linux durci... mais qui n'est pas disponible à tous. N'oublions pas non plus ce site web so De Gaulle. Heureusement, il y a de très bonnes conférences #sincère (cf. ANSSTIC il y a 2 ans), je rage de voir tant de compétences et si peu de résultats (ce qui ne veut pas dire qu'on y glande).

Pour résumer, l'ANSSI cherche à contrôler tout ce qui se passe en sécurité en France, mais les retours vers la communauté de la sécurité ne sont pas à la hauteur des espoirs engendrés par l'Agence #Frustration.

En sécurité, on mesure couramment la légitimité de quelqu'un à ce qu'il produit. Or là, on ne voit quasiment rien de ce que crée l'ANSSI, car il n'y a pratiquement rien de public (*for your eyes only*). Du coup, après avoir recruté des juniors à tour de bras, l'ANSSI impose sa légitimité par la Loi. Cette légitimité législative donne une petite arrogance à certains (très peu pour être honnête) alors qu'ils ont encore du lait qui leur coule au coin des lèvres en termes de sécurité.

Il est quasi-sûr que, sans ce passage en force, les grandes entreprises et OIV ne bougeraient pas autant qu'il faudrait (la sécurité n'est que rarement la priorité, ça coûte de la money avec un budget d'un penny). Quand le sage montre le doigt, le fou regarde la Lune et raque (*moon raker* en anglais, évidemment). Alors, rien que pour ça, merci l'ANSSI.

Cette frénésie réglementaire est aussi liée à une obligation de neutralité : l'ANSSI, au service de l'État, ne peut (veut) pas dire que tel produit est pourri, ou telle société incompétente, elle n'a pas de *permis de tuer* en tant que juge. Pour compenser, la réglementation a vocation à éliminer les « mauvais », mais regardez les 8 CSPN et 13 PASSI, pas de surprise ! Du coup, à quoi servent ces usines à gaz ? Les clients continuent à jouer à la roulette pour choisir produits et prestataires comme au *casino royale*. Tous les CSPN/CESTI et PASSI ne sont pourtant pas équivalents. Pourquoi ne pas mettre une sorte de graduation par domaines d'expertise par exemple ?

L'Agence pourrait aussi révolutionner ses règles de diffusion (articles et outils) guidées par le politiquement correct. Les bulletins de sécurité fourniraient différents niveaux d'analyses des failles et une signature pour détecter les attaques utilisant ce vecteur. Offrez du vrai contenu ! Aujourd'hui, on récupère un bulletin, un rapport, et *on ne lit que deux fois* la prose de l'Agence.

La sécurité est un domaine avant tout humain. Pas technique, organisationnel, juridique ou je ne sais quoi : humain ! Et la sécurité marche parce que des personnes collaborent, discutent, se font confiance. Et ça, pas une réglementation ou un contrat ne pourra le forcer. Tout le monde se connaît. Les labos de l'ANSSI font des trucs impressionnants : ouvrez-les à la coopération, simple, sans cadre administratif, juste parce que des échanges naît le progrès.

La France est un petit pays avec de très nombreux talents, entre autres avec ses ingénieurs. Il serait peut-être temps de les libérer, de les laisser briller : *les diamants sont éternels*. Faites-leur confiance : *la France ne suffit pas*. Mais pour ça, il faudrait aussi arrêter de sans cesse se reposer sur l'État, et prendre ses responsabilités : c'est aux acteurs d'agir, pas d'attendre que ça vienne à eux.

N'en doutez pas, je suis fan de l'ANSSI. Ou plus exactement d'un paquet de monde qui y travaille et qui essaye de faire bouger la bête administrative et hiérarchique de l'intérieur. Mais je suis vraiment frustré de voir la sous-exploitation, et plus encore des jeunes déprimer devant la lourdeur interne et me sortir comme excuse « on n'y peut rien, c'est l'administration ». Des choses bougent, certes, mais tellement lentement si on suit le système.

Conclusion : pour faire un *Bond* (*James Bond*) en avant, bougez-vous le *Q* ! Ou pour être plus souverain, on ne manque pas d'*os, sans disette*, dans ce nid d'APT.

God save l'ANSSI

Fred RAYNAL (idéaliste)
@fredraynal
@MISCRedac

SOMMAIRE

www.miscmag.com

MISC est édité par **Les Éditions Diamond**
B.P. 20142 / 67603 Sélestat Cedex
Tél. : 03 67 10 00 20 - **Fax** : 03 67 10 00 21
E-mail : cial@ed-diamond.com
Service commercial : abo@ed-diamond.com
Sites : www.miscmag.com
boutique.ed-diamond.com
IMPRIMÉ en Allemagne - PRINTED in Germany
Dépôt légal : A parution
N° ISSN : 1631-9036
Commission Paritaire : K 81190
Périodicité : Bimestrielle
Prix de vente : 8,90 Euros

Directeur de publication : Arnaud Metzler
Chef des rédactions : Denis Bodor
Rédacteur en chef : Cédric Foll
Secrétaire de rédaction : Aline Hof
Conception graphique : Kathrin Scali
Responsable publicité :
Black Mouse Communication
Tél. : 03 67 10 00 27
Service abonnement : Tél. : 03 67 10 00 20
Illustrations : www.fotolia.com
Impression : pva, Druck und Medien-Dienstleistungen GmbH, Landau, Allemagne
Distribution France : (uniquement pour les dépositaires de presse)
MLP Réassort :
Plate-forme de Saint-Barthélemy-d'Anjou. Tél. : 02 41 27 53 12
Plate-forme de Saint-Quentin-Fallavier. Tél. : 04 74 82 63 04

Charte de MISC

MISC est un magazine consacré à la sécurité informatique sous tous ses aspects (comme le système, le réseau ou encore la programmation) et où les perspectives techniques et scientifiques occupent une place prépondérante. Toutefois, les questions connexes (modalités juridiques, menaces informationnelles) sont également considérées, ce qui fait de MISC une revue capable d'appréhender la complexité croissante des systèmes d'information, et les problèmes de sécurité qui l'accompagnent. MISC vise un large public de personnes souhaitant élargir ses connaissances en se tenant informées des dernières techniques et des outils utilisés afin de mettre en place une défense adéquate.
MISC propose des articles complets et pédagogiques afin d'anticiper au mieux les risques liés au piratage et les solutions pour y remédier, présentant pour cela des techniques offensives autant que défensives, leurs avantages et leurs limites, des facettes indissociables pour considérer tous les enjeux de la sécurité informatique.

ANALYSE D'UNE INSCRIPTION EN LIGNE : COMMENT VOS DONNÉES FUITENT SUR INTERNET...

Stéphane Labarthe - slabarthe@cnil.fr - misc@labarthe.es

Benjamin Vialle - bvialle@cnil.fr - misc@vialle.io

Auditeurs des systèmes d'information au service des contrôles de la CNIL

mots-clés : WEB / CDN / HTTPS / CHIFFREMENT / TRACKING / REFERER / CNIL

Lorsqu'un internaute visite un site web ou s'y inscrit, il laisse certaines de ses données personnelles au site. Mais des acteurs tiers peuvent aussi en être destinataires... Peu visibles de l'internaute, ils agissent parfois à l'insu du site web lui-même. Ce sont d'abord des acteurs situés au niveau applicatif, dont le code source s'exécute directement dans la page web (régies publicitaires, services de mesure d'audience, plateformes RTB). Ce sont aussi les CDN, qui se positionnent comme des intermédiaires au niveau réseau. Nous vous proposons l'analyse de deux méthodes utilisées pour détecter ces fuites de données. Les exemples présentés dans cet article sont fictifs, mais fondés sur des situations réelles rencontrées lors de contrôles de la CNIL.

L'essentiel est invisible pour les yeux. (Antoine de Saint-Exupéry)

1 Méthodologie et configuration de l'environnement proposées

1.1 Description générale de la méthodologie

L'un des objectifs de cet article est de proposer au lecteur de MISC une méthodologie reproductible afin de lui permettre d'analyser une partie des flux de données mis en jeu lors d'une inscription sur un site web (de e-commerce, de jeu en ligne, service de messagerie électronique, réseau social...). Le lecteur pourra alors se rendre compte par lui-même que, dans de nombreux cas, certaines des données personnelles transmises de son navigateur vers les serveurs du site sont également envoyées, souvent à son insu, vers des tiers qu'il ne connaît pas forcément.

La méthodologie se veut simple : elle consiste à réaliser une inscription sur le site web choisi puis, éventuellement, un achat avec paiement. On capturera alors d'une part les flux HTTP/HTTPS entre le navigateur et les serveurs web distants et d'autre part les flux réseaux entre la carte réseau de l'ordinateur de l'internaute et ces mêmes serveurs. Par ailleurs, l'examen du code source de la page web (y compris les éventuelles fonctions JavaScript externes) pourra permettre d'aborder le problème sous un autre angle.

1.2 Une configuration basée sur des logiciels libres

La configuration proposée pour l'analyse repose sur un ordinateur directement connecté à Internet (c'est-à-dire sans proxy) afin de visualiser immédiatement, dans le flux réseau, l'adresse IP de l'hôte destinataire et de s'affranchir de toute modification par le proxy des réponses renvoyées par les serveurs web.

Ensuite, afin de faciliter la reproductibilité de l'expérience, notre article s'appuie sur des logiciels libres :

- Pour la navigation : Mozilla Firefox augmenté de Live HTTP Headers **[LIVE_HTTP_HEADERS]**. Cette extension capture les entêtes HTTP.
- Pour la capture du réseau, WireShark : Ce logiciel permet la capture des flux émis et reçus par la carte réseau de l'ordinateur de l'internaute.

À noter que lorsque la communication est chiffrée (HTTPS), par son positionnement, Live HTTP Headers permet d'accéder au contenu des données échangées, ce qui n'est pas possible avec Wireshark (qui ne permet alors même pas d'accéder aux URL).

Un autre point important concerne le « nettoyage » de la configuration du navigateur, afin que celui-ci ne bloque ni ne produise de requêtes parasites. Pour cela, nous recommandons notamment de ne pas avoir d'extension produisant ou bloquant des requêtes (par exemple, les outils tels que NoScript qui bloquent l'exécution de code JavaScript), de supprimer tout l'historique, de désactiver certains paramètres pouvant générer des requêtes automatiques du navigateur (comme les mises à jour de sécurité) et de désactiver le blocage des cookies tiers.

À cette fin, une machine virtuelle ou un profil Firefox dédié (via la commande « `firefox -p` ») pourrait être une aide utile.

Enfin, il est indispensable d'avoir une horloge interne synchronisée sur un serveur de temps.

Remarque : Une autre approche est possible : celle du proxy local en rupture de chiffrement qui permet une recherche directe des contenus dans les flux chiffrés. Des solutions libres de ce type existent également **[ZAP]**. Cependant, nous avons préféré ici l'usage de WireShark, que le lecteur connaîtra probablement et qui réalise une écoute totalement passive.

2 Ces acteurs qui agissent au cœur de la page web (analyse de la capture HTTP/HTTPS)

2.1 « Cachons ces partenaires que vous ne sauriez voir » : comment les pages web communiquent avec des tiers

Lorsque l'on navigue sur un site web, notre navigateur communique évidemment avec les serveurs web de ce site mais également, dans la majorité des cas, avec une multitude d'autres serveurs web, dits « tiers » ou « partenaires ». Certains peuvent se matérialiser de manière concrète, par exemple par le biais d'un bouton Facebook ou d'un bandeau publicitaire, mais la plupart sont invisibles « à l'œil nu ». Parmi les plus répandus, on peut citer les acteurs publicitaires (régies publicitaires, plateformes d'enchères publicitaires ou RTB **[RTB]**, etc.), les réseaux sociaux, les plugins vidéos et les solutions de mesure d'audience (« analytics »).

Ces tiers ou partenaires ont fourni à l'éditeur du site web un morceau de code source (appelé *tag*) que l'éditeur du site a inclus dans le code HTML de la page web.

Ce *tag* fait en général appel à une ressource externe (image ou fonction JavaScript) située sur le serveur web du tiers. C'est cet appel qui initie la communication avec le serveur web tiers.

Un examen du code source de la page pourra faire apparaître ce type de *tags*. Dans notre cas :

```
<script type="text/javascript" src="http://www.ma-regie-pub.com/
pagead/conversion.js">
```

On peut remarquer que ce type de fonction n'est pas directement intégré dans le code HTML de la page renvoyée par le serveur : elle peut donc être modifiée par le « partenaire » sans que l'éditeur du site web ne s'en aperçoive. En s'exécutant à l'intérieur de la page web, cette fonction a accès à l'intégralité de son contenu. Des questions en termes de sécurité se posent donc. De plus, même un examen du code source « appelé » ne révélera que très difficilement ce qui se passe réellement à l'exécution, car il a souvent été « obfusqué » — c'est l'une des raisons pour lesquelles nous avons privilégié une analyse de capture réseau plutôt que du code source.

Alors comment garantir la sécurité d'un site web si l'on n'est pas capable de savoir ce que fait le code source qui s'y exécute ? (La même question se pose de manière accrue s'il est inconnu ou susceptible d'être modifié de manière invisible et sans préavis.) Nous y reviendrons dans la suite.

2.2 Inscription en ligne

Imaginons une inscription en deux temps sur **ma-boutique-en-ligne.org** : un formulaire de saisie de notre adresse électronique, suivi d'un autre formulaire permettant de saisir les autres données personnelles nécessaires à notre inscription. Nous le complétons avec les données suivantes (et le mot de passe « azerty78 »), voir Figure, page suivante.

Dans notre cas, les deux formulaires sont situés sur des pages ayant une URL de type HTTPS. Nous en déduisons donc que les données sont transmises directement aux serveurs de **ma-boutique-en-ligne.org** de manière chiffrée. L'authentification de l'unique destinataire et la confidentialité de la transmission semblent donc assurées.

2.3 Analyse HTTP de la requête d'envoi de données : jusqu'ici tout va bien…

La première recherche que nous allons effectuer dans notre fichier de captures HTTP/HTTPS est la requête qui a servi à envoyer les données du formulaire d'inscription. Pour cela, nous recherchons des chaînes de caractères complétées dans le formulaire, par exemple avec le mot de passe « azerty78 ».

La recherche nous renvoie une seule réponse qui nous amène à la requête HTTPS suivante :

```
https://www.ma-boutique-en-ligne.org/fr/secure/subscription_s.html
POST /fr/secure/subscription_s.html HTTP/1.1
Host: www.ma-boutique-en-ligne.org
[...]
Referer: https://www.ma-boutique-en-ligne.org/fr/secure/
inscription_client/index.html?Email=slabarthe@cnil.fr
[...]
Email=slabarthe@cnil.fr&password=azerty78&gender=M&name=St%E9phane&
lastname=LABARTHE &birth_day=06&birth_month=01&birth_year=1978
```

Cette requête montre que l'ensemble des données saisies est envoyé en HTTPS et avec la méthode POST vers une URL de **ma-boutique-en-ligne.org** (ces données apparaissent ci-dessus en rouge). C'est donc un bon point.

Un champ de l'en-tête HTTPS attire cependant notre attention : il s'agit du champ *Referer*, mentionné plus haut. Il contient ici l'URL principale de la page précédente. On peut observer que l'adresse électronique y est passée en paramètre à travers la chaîne : «**?Email=slabarthe@cnil.fr** ».

2.4 Mais où va donc mon adresse électronique ? (Transmission directe)

Mais la suite nous réserve une surprise. La recherche de « slabarthe@cnil.fr » dans le fichier de capture fait apparaître beaucoup d'autres requêtes dont certaines destinées à des URL externes au site web. La première est la suivante :

```
https://ssl.mesuredaudience.com/image.gif?version=5.4.0&id=127283
4227&site=www.ma-boutique-en-ligne.org&mid=42572054&zmp=%2Ffr%2Fs
ecure%2Finscription_client%2Findex.html%3FEmail%3Dslabarthe@cnil.
fr&idclient=2545080-2
```

L'URL contient de nombreuses informations en paramètres dont notre adresse électronique. Elle est à destination du service de mesure d'audience du site et son envoi a été déclenché par une fonction JavaScript de celui-ci (un *tag*), qui s'est exécutée dans le contexte de la page.

On peut également remarquer l'utilisation par le service de mesure d'audience du site du protocole HTTPS afin de ne pas soulever d'alerte de sécurité au niveau du navigateur.

Les fonctions JavaScript des prestataires qui s'exécutent dans une page HTTP ou HTTPS ont accès au contenu et à l'URL de la page et peuvent communiquer de manière invisible avec les serveurs du prestataire. L'usage du protocole HTTPS n'offre donc, ici, aucune protection contre ce type de fuite de données.

2.5 Mais où va donc mon adresse électronique ? (Transmission par Referer)

Indépendamment des fonctions JavaScript « partenaires » qui peuvent aller chercher la donnée dans l'URL, la transmission de données personnelles par la méthode GET, dans le cas du protocole HTTP comme HTTPS, pose un autre problème, celui du *Referer*.

Lorsqu'une requête HTTP(S) à destination d'une URL externe est initiée par un *tag* tiers, le champ *Referer* de cette requête contiendra l'URL de la page web.

Or, si cette URL contient, comme dans notre exemple, une adresse électronique (ou d'autres données personnelles) dans ses paramètres, alors toutes les requêtes partant de la page web vers des tiers auront cette donnée dans leur champ *Referer*. Ainsi, les données personnelles seront effectivement transmises à tous les partenaires dont des *tags* s'exécutent dans la page. Dans notre cas, on s'aperçoit que l'adresse électronique d'inscription a été transmise par le biais du *Referer* à une dizaine de tiers, eux-mêmes susceptibles de la retransmettre en cascade (cas du *Real Time Bidding* notamment). L'exemple suivant l'illustre :

```
https://ma-regie-pub.com/ots/js-3.2/313349/9iE1AeLO80a9AUke0nTadZ1u
ap0c3DmgNKH7qOf2JN0iGQGvlE2mCFLHqkasZd4Dx1wm4vC-q684Rbvwm6hCwbI8PnL
ZbATqC399jcKM229ANxVYBquNnstVrkOYLNzn/1364290796051-132/1/2
GET /ots/js-3.2/313349/9iE1AeLO80a9AUke0nTadZ1uap0c3DmgNKH7qOf2JN0i
GQGvlE2mCFLHqkasZd4Dx1wm4vC--q684Rbvwm6hCwbI8PnLZbATqC399jcKM229ANx
VYBquNnstVrkOYLNzn/1364290796051-132/1/2 HTTP/1.1
Host: ma-regie-pub.com [...]
Referer: https://www.ma-boutique-en-ligne.org/fr/secure/
inscription_client/index.html?Email=slabarthe@cnil.fr
```

2.6 Une conséquence : la rupture d' « anonymat »

Une des conséquences de ce type de transmission est la rupture d'« anonymat » (en réalité de pseudonymat **[PSEUDO]**) par les régies publicitaires et autres acteurs tiers. En effet, ces acteurs reconnaissent en général les internautes grâce à des numéros identifiants (qu'ils disent « anonymes ») contenus dans des cookies. La transmission conjointe de l'identifiant du cookie et d'une donnée personnelle directement identifiante dans une même requête HTTP ou HTTPS (cf. la requête précédente) peut lui permettre d'associer ces données au numéro d'identification contenu dans le cookie. Le pseudonymat est ainsi rompu.

La tentation d'exploiter ces masses de données devient d'autant plus grande pour ces acteurs, même dans le cas des pure players (qui ne font par exemple que régie publicitaire ou que service de mesure d'audience). L'heure est aujourd'hui en effet à l'orientation vers le *cross device*, qui consiste à reconnaître un même utilisateur sur ses différents terminaux. Or, un traçage par cookies ou même par *fingerprinting* ne permet pas à une régie publicitaire de savoir que c'est le même internaute qui se connecte depuis son ordinateur personnel, son ordinateur professionnel, son smartphone… ; la régie verra trois cookies ou trois empreintes distinctes. Dès lors, utiliser l'adresse électronique discrètement récoltée pour lier ces trois identifiants devient tentant…

2.7 Aspect organisationnel de la problématique : pourquoi le service informatique ne devrait pas obéir au service marketing

Dans la pratique, ces *tags partenaires*, qui sont insérés dans le code source de la page d'un site, sont souvent fournis par le service marketing de la société éditrice au service informatique qui se contente alors de les positionner dans la page.

De plus, les prestations associées ne font en général pas l'objet d'un vrai contrat, avec des clauses de confidentialité **[CLAUSES]**, mais d'une simple facture, non connue du service informatique. Il en résulte souvent que, même lorsque la prestation est terminée avec le partenaire, le *tag* n'est pas retiré et continue de s'exécuter dans la page et donc de générer le dépôt de cookies aux utilisateurs du site et la transmission des données au prestataire, cela sans finalité et en dehors de tout encadrement contractuel. Ce problème, très répandu, a une expression pour le désigner : les *tags obsolètes*.

Lors d'un contrôle de la CNIL sur un important site français d'e-commerce, le même type de transmission que celui décrit dans la partie 2.4 (transmission directe de l'adresse électronique en paramètre de

l'URL) avait été identifié par l'analyse de la capture HTTP. Parmi les prestataires destinataires de l'adresse électronique, il y en avait un qui était totalement inconnu du responsable *front-office* de la société ainsi que du responsable développement. Finalement, il aura fallu interroger l'un des plus anciens développeurs de la société pour se rendre compte que le prestataire en question était une société de *A/B testing* dont le contrat était échu depuis plusieurs années. Le responsable *front-office* et le directeur informatique nous indiqueront même qu'avec le grand nombre de *tags* partenaires (certains obsolètes) s'exécutant dans le code source du site, son fonctionnement en était considérablement ralenti !

Il est vrai que le nombre de serveurs tiers (et donc de noms de domaine) appelés depuis un site de e-commerce sur un simple parcours « inscription-achat » dépasse régulièrement la vingtaine…

2.8 Premières recommandations à destination des éditeurs de sites

Si le risque de fuite de données depuis un site web ne peut être réduit à zéro dès lors que celui-ci inclut des *tags* partenaires, on peut néanmoins effectuer les recommandations suivantes :

1. Le service juridique et le service informatique doivent être impliqués du début à la fin de toute nouvelle prestation impliquant l'insertion d'un *tag* dans le code source de la page (publicité, mesure d'audience, plugin vidéo ou de réseau social, A/B testing, etc.). Cette collaboration devra conduire en particulier à la suppression des *tags* obsolètes en fin de prestation.

2. Pour chacun des *tags* partenaires, le rapport risque/bénéfice doit être évalué en prenant en compte son positionnement (page d'accueil, formulaire de saisie de données, etc.).

3. L'envoi des données de l'ensemble des formulaires (pré-inscription, inscription, authentification, commande, paiement avec saisie de données bancaires, etc.) doit se faire en utilisant le protocole HTTPS et la méthode POST.

4. Des tests sur la façon dont fonctionne le site web et les destinataires des données doivent être régulièrement effectués. La méthode décrite dans cet article est facilement reproductible.

Pour aller plus loin et ne pas se contenter d'analyser uniquement les échanges HTTP/HTTPS entre le navigateur de l'internaute et les serveurs web distants, on peut étendre cette analyse et s'intéresser plus précisément aux serveurs destinataires des données à caractère personnel. On s'aperçoit alors que d'autres acteurs en sont destinataires.

3 Quand les CDN entrent en jeu (analyse WireShark)

3.1 Mais à qui est cette adresse IP ?

Reprenons l'inscription réalisée précédemment sur le site de e-commerce **ma-boutique-en-ligne.org** et observons la capture réseau effectuée à l'aide du logiciel WireShark.

Celle-ci a été faite en parallèle de la capture des en-têtes HTTP/HTTPS. Par souci de simplification, nous allons traiter le cas d'une requête HTTP (donc lisible en clair dans WireShark). Le cas d'une requête HTTPS est en effet un peu plus complexe, puisqu'il faut faire correspondre les en-têtes capturés d'une part et les captures réseaux WireShark d'autre part, ou bien utiliser un outil tel que OWASP — ce que nous ne faisons pas dans ce type de constat.

Cherchons d'abord à récupérer la (ou les) trame(s) ayant transmis notre adresse électronique :

```
tshark -r capture.pcap -T text -Tfields -e frame.number -e ip.proto
-e ip.dst -e http.host -e text | grep cnil
```

```
9       6       203.0.113.62            [...]Timestamps:
TSval 5587, TSecr 1204208032,POST fr/secure/subscription_s.
html  HTTP/1.1\r\n,Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7\
r\n,Keep-Alive: 115\r\n,\r\n,HTTP request 1/1,[truncated]
Email=slabarthe@cnil.fr&password=azerty78&gender=M&name=St%E9phane&
lastname=LABARTHE &birth_day=06&birth_month=01&birth_year=1978
```

La trame n° 9 de la capture nous indique que le serveur contacté pour la requête contenant l'adresse électronique et les autres données d'inscription a pour adresse IP 203.0.113.62.

Une requête *whois* sur cette adresse IP renvoie vers un FAI étranger. Le site **ma-boutique-en-ligne.org** a pourtant tous ses serveurs hébergés en France.

L'explication est la suivante : cette adresse IP est celle d'un serveur du CDN (Content Delivery Network) situé chez un FAI étranger. Il s'avère en effet que les sociétés fournissant des services de CDN possèdent des serveurs le plus souvent placés dans les datacenters des FAI afin d'être au plus près des utilisateurs finaux.

3.2 Mais par où transitent nos coordonnées bancaires ?

Le site **ma-boutique-en-ligne.org** propose également de stocker les coordonnées bancaires de ses clients, pour ne plus avoir à les leur demander lors de leur achat suivant.

En se rendant sur la page d'enregistrement des coordonnées bancaires, on peut vérifier que l'on se trouve bien sur une page HTTPS : les échanges de données entre le premier serveur de **ma-boutique-en-ligne.org** et le

navigateur sont protégés (chiffrement) et le serveur du site ma-boutique-en-ligne.org est bien celui qu'il prétend être si l'on en croit le certificat (authentification).

Nous allons de nouveau rechercher la requête ayant soumis le formulaire. Contrairement à l'exemple précédent, les flux sont cette fois chiffrés (HTTPS).

Voici la requête telle que capturée par Live HTTP Header, directement dans Firefox :

```
https://secure.ma-boutique-en-ligne.org/fr/secure/enregCoordBancaires
POST /fr/secure/enregCoordBancaires/ HTTP/1.1

Host: secure.ma-boutique-en-ligne.org
[...]
Referer: https://www.ma-boutique-en-ligne.org/fr/secure/
envoiCoordBancaires
[...]
Date: Wed, 09 Oct 2013 13:27:45 GMT
Card_number=0123456789012345&ExpDate=06/01/78&CCV=007
```

On constate que la méthode utilisée est POST. Nos coordonnées bancaires seront donc envoyées en POST à l'adresse https://secure.ma-boutique-en-ligne.org. Un envoi bel et bien sécurisé, donc : authentification, chiffrement et envoi direct vers ma-boutique-en-ligne.org, seule à pouvoir déchiffrer les données.

Comme les flux sont chiffrés, nous ne pouvons pas utiliser WireShark pour effectuer des recherches de texte et ainsi retrouver la requête TLS correspondante. Procédons autrement : cherchons, dans la capture réseau, la résolution DNS. Celle-ci a dû avoir lieu au même moment puisque c'est la première fois que nous nous rendons sur le domaine secure.ma-boutique-en-ligne.org :

```
tshark -r capture.pcap  -T text -Tfields -e frame.number -e
ip.proto -e ip.dst  -e dns -e text -e frame.time | grep 13:27
```

```
1002    17      198.51.100.1         Queries,secure.ma-boutique-en-ligne.org:
type A, class IN,Answers,secure.ma-boutique-en-ligne.org: type CNAME, class
IN, cname secure.ma-boutique-en-ligne.org.mon-cdn.net,secure.ma-boutique-en-
ligne.org.moncdn.net: type CNAME, class IN, cname e678.mon-cdn.net: type A,
class IN, addr 203.0.113.62    Oct  9, 2013 13:27:14.526743000
```

La résolution DNS nous indique ici que secure.ma-boutique-en-ligne.org est un alias qui pointe vers secure.ma-boutique-en-ligne.org.mon-cdn.net qui, lui, pointe vers e678.mon-cdn.net !

Le reste de la capture montre qu'une connexion TLS s'établit entre l'adresse IP du CDN (203.0.113.62) et celle de l'utilisateur : nos coordonnées bancaires sont donc envoyées, dans un premier temps, au CDN.

Par ailleurs, on constate que, pour le site de ma-boutique-en-ligne.org, c'est l'intégralité des requêtes qui transite par l'adresse IP 203.0.113.62. Ainsi, le CDN est destinataire de l'ensemble des informations.

3.3 L'intermédiaire (presque) invisible

Dans la pratique, les serveurs de CDN sont très souvent destinataires du contenu des requêtes HTTP ou HTTPS des internautes, requêtes susceptibles de contenir des données à caractère personnel. Autrement dit, le CDN peut-être vu comme un intermédiaire quasi invisible, par lequel transitent non seulement les contenus statiques des pages internet, mais également les données personnelles échangées entre le serveur web et l'internaute.

Notons que, dans presque tous les cas rencontrés par la CNIL lors des contrôles, l'intégralité des flux HTTP ou HTTPS transitait par les serveurs du CDN.

3.4 La question des certificats cryptographiques

3.4.1 Explications

Les CDN proposent deux options dans le cas de flux chiffrés :

- En général, le CDN se voit fournir les certificats et clefs privées des sites web. C'est le cas dans l'exemple précédent. Les contenus peuvent être ainsi directement délivrés en étant chiffrés au moyen du certificat du site web sur lequel l'internaute navigue.
- Le CDN peut également utiliser ses propres certificats, soit sur l'ensemble des pages, soit pour certains contenus.

Dans les deux cas, les flux sont accessibles en clair pour les CDN. Lors des contrôles, ces faits étaient le plus souvent ignorés du DSI et, s'il y en avait un, du RSSI.

3.4.2 L'accès aux certificats cryptographiques par le CDN : aspect contractuel

La société *mon-cdn*, dans le cadre de son service de CDN, se place dans le premier cas mentionné dans la partie précédente. La société a accès au certificat SSL de ses clients, cela faisant partie intégrante du contrat. Ce dernier peut prévoir que le CDN est autorisé à héberger et utiliser les certificats au nom du site éditeur sur ses propres serveurs. Par exemple, la société AKAMAI, leader du marché, précise dans un document accessible sur son site web : « *If an Akamai customer wants to deliver entire SSL pages, they do so by CNAMEing their Web site over to the Akamai network. The Akamai network serves SSL pages over a secure connection on behalf of the customer using a customer-provided SSL certificate.* » **[AKAMAI_CNAME]**.

Si nous revenons à notre exemple, un examen de la requête suivant la requête « SERVER HELLO » du protocole TLS fait d'ailleurs apparaître le certificat, dans lequel nous pouvons observer un champ « *Common Name* » et un champ « *Organisational Unit Name* » faisant respectivement référence à ma-boutique-en-ligne.org et à *Mon-CDN*.

3.5 Une contre-mesure : le déroutage des flux porteurs de données sensibles

Il existe un moyen de s'assurer que les flux porteurs de données « sensibles » ne sont pas accessibles par le CDN : ces flux doivent être « déroutés » et aller directement au serveur web depuis le navigateur du client, sans transiter par un CDN.

S'agissant de l'exemple précédent, ma-boutique-en-ligne.org pourrait utiliser, *a minima*, deux noms de domaine supplémentaires :

- images.ma-boutique-en-ligne, redirigeant (CNAME) vers les serveurs du CDN, pour les contenus statiques, ou du moins ne contenant pas de données sensibles ;
- secure.ma-boutique-en-ligne.org, pointant directement vers les serveurs web de ma-boutique-en-ligne.org, pour les flux porteurs de données sensibles.

Dans tous les cas, il faut également prendre en compte les remarques de la partie précédente s'agissant du *Referer*. En effet, cette technique du déroutage des flux porteurs de données sensibles n'a été rencontrée qu'une seule fois lors des contrôles de la CNIL. Malheureusement, parmi les requêtes qui transitaient par le CDN, certaines avaient pour *Referer* une URL contenant l'adresse électronique de l'internaute renseignée à l'inscription.

Attention !

Attention ici aux termes « données sensibles » : l'usage du terme « données sensibles » n'est pas employé dans cet article exclusivement au sens de l'article 8 de la loi Informatique et Libertés, qui concerne les données relatives aux origines raciales ou ethniques, aux opinions politiques, philosophiques ou religieuses ou à l'appartenance syndicale des personnes, ou qui sont relatives à la santé ou à la vie sexuelle. Ici, le terme « données sensibles », à prendre dans un sens très large, regroupe toute donnée que le lecteur de MISC juge comme ne devant pas être communiquée à des tiers (les secrets commerciaux et industriels, et les données à caractère personnel en général, dont les données bancaires et données de santé, etc.).

3.6 Les CDN et l'informatique dans les nuages : exemple des réseaux sociaux

Le recours à des CDN peut occasionner des fuites de données lorsque les serveurs hébergent des données personnelles non publiques : c'est le cas de certains réseaux sociaux. Ainsi, les photos ou vidéos privées servies par un réseau social peuvent être stockées sur les serveurs de son CDN. Lorsque l'internaute affiche

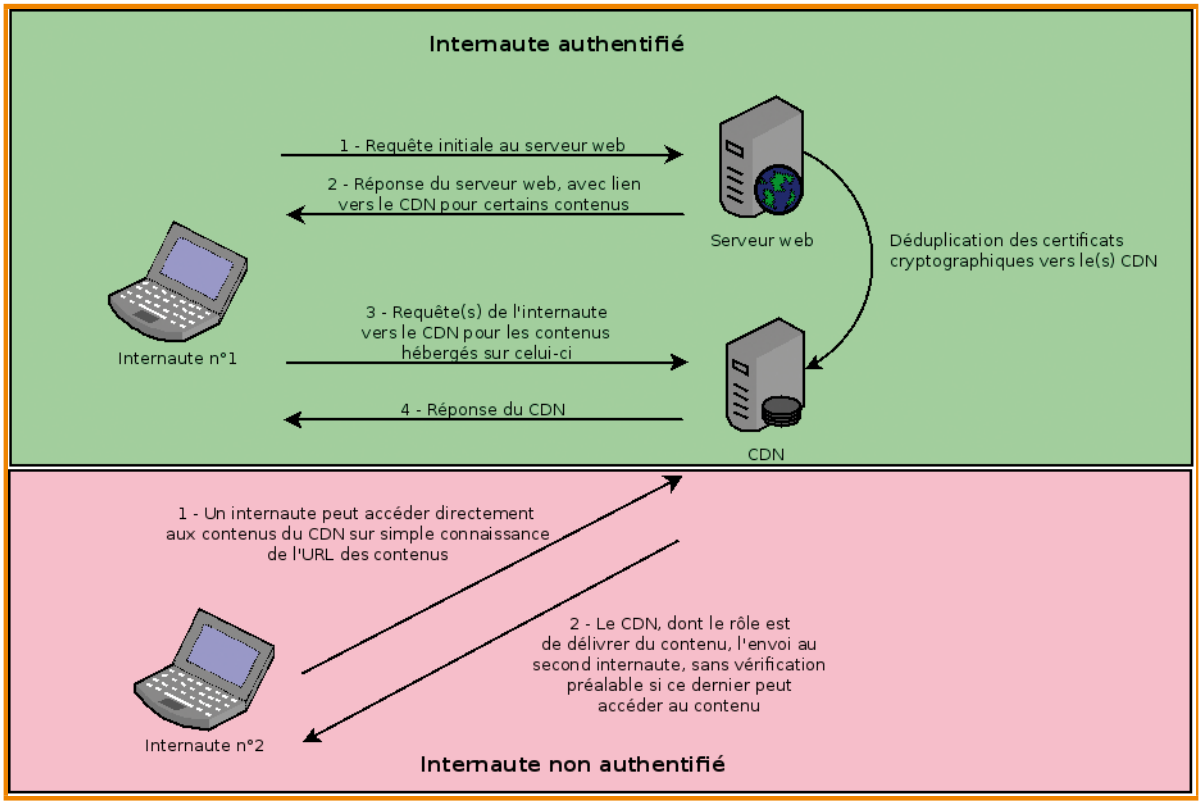


Fig. 2 : Schéma simplifié illustrant le cas donné en exemple.

une page du site web avec ces contenus, ceux-ci sont récupérés depuis le CDN et intégrés dans la page web. Or, on constate que ces contenus peuvent être accessibles directement depuis les serveurs du CDN, via leurs URL, sans authentification. Ainsi, il est donc possible à un utilisateur d'accéder à du contenu non public du site sans être authentifié, directement sur les serveurs du CDN, sur simple connaissance de l'URL dudit contenu, comme l'illustre la figure n° 2, page précédente.

3.7 Les CDN et l'informatique dans les nuages : exemple d'une application SaaS professionnelle

Voici un exemple de cas réel rencontré lors d'un contrôle de la CNIL : la société contrôlée fournissait une application en ligne (*cloud computing* de type *SaaS*) à destination de professionnels qui pouvaient y charger leurs bases de données. Les données transitant par le CDN étaient déchiffrées par celui-ci après un éventuel ajout de contenu statique. Lorsque la société a compris que toutes les communications entre ses clients et ses serveurs étaient accessibles par le CDN et qu'elle allait donc devoir, sur le fondement de la loi informatique et libertés, en informer ses clients, elle a préféré rompre le contrat avec le CDN en question...

4 CDN, Big Data et NSA

À l'ère du Big Data, les CDN s'intéressent au contenu des myriades de données qui transitent par leurs réseaux.

Aujourd'hui, 50 % du trafic Internet mondial transiterait par les réseaux des CDN **[TRAFIC_CDN]** dont 20 % à 30 % pour le seul acteur AKAMAI, leader du marché et qui posséderait pour cela plus de 125 000 serveurs répartis dans environ 80 pays **[AKAMAI]**.

Par conséquent, au vu de la formidable masse de données accessibles déchiffrées par les CDN, dont les acteurs majeurs sont américains (AKAMAI, AMAZON CLOUDFRONT, CLOUDFLARE, etc.), on ne peut pas exclure le risque d'un espionnage massif par ce biais, en complément du risque d'usage détourné à des fins publicitaires.

La vulnérabilité évoquée dans la partie 3.6 a d'ailleurs effectivement été exploitée par un programme d'espionnage opéré par le GCHQ dénommé « SPRING BISHOP » et permettant d'accéder aux photos privées de Facebook, sans authentification, directement sur les serveurs d'AKAMAI **[SPRING BISHOP]**.

5 Les limites de notre analyse

Les exemples mis en évidence dans la partie 2 de cet article ne montrent pas les méthodes de traçage indirect que peuvent mettre en place les sites web. Certaines sont connues et utilisées depuis maintenant de nombreuses années, comme les cookies, tandis que d'autres font encore l'objet de travaux de recherche et de développement, comme le *fingerprinting*.

De plus, les exemples qui illustrent cet article présentent systématiquement des données à caractère personnel « lisibles » au niveau du protocole HTTP. Or, des contrôles ont mis en évidence certaines techniques de camouflage par encodage, hachage ou chiffrement des données lors de leur transmission à un tiers. Ainsi, l'internaute n'est pas toujours en mesure de savoir que son adresse électronique ou toute autre donnée est transmise et utilisée pour l'identifier sur d'autres sites que celui sur lequel il s'est inscrit.

Enfin, notre analyse ne permet pas de mettre en évidence les communications de serveurs à serveurs, notamment s'agissant des données transmises pour la publicité ciblée, entre acteurs publicitaires.

Conclusion : à vous de jouer !

Aujourd'hui, l'internaute n'est pas clairement et suffisamment informé de la transmission à des tiers d'une partie de ses données à caractère personnel. Cependant beaucoup de ces transmissions, parfois qualifiables de « fuites de données », sont facilement détectables. En cas de détection, il est d'ailleurs possible de saisir la CNIL qui s'est vue octroyée, depuis cette année, le droit d'effectuer des contrôles en ligne. ■

Note

Le rapport annuel de la CNIL contient, chaque année, la liste des sociétés contrôlées. Les rapports annuels des années 2012 et 2013, contiennent un certain nombre d'acteurs importants de l'Internet français et mondial. Ils sont disponibles en ligne : http://www.cnil.fr/fileadmin/documents/La_CNIL/publications/CNIL_RA2012_web.pdf et http://www.cnil.fr/linstitution/actualite/article/article/bilan-2013-la-protection-des-donnees-une-preoccupation-croissante-des-particuliers/.

Remerciements

Nous tenons à remercier notre institution et plus particulièrement son Secrétaire Général, Édouard GEFFRAY, ainsi que Florence FOURETS et Christophe VIVENT de nous avoir soutenus dans la rédaction de cet article. Merci également à Thierry CARDONA pour ses relectures actives.

Retrouvez les références de cet article sur le blog de MISC : www.miscmag.com