Les données que récolte Google – Ch.4

Voici déjà la traduction du quatrième chapitre de Google Data Collection, l'étude élaborée par l'équipe du professeur Douglas C. Schmidt, spécialiste des systèmes logiciels, chercheur et enseignant à l'Université Vanderbilt. Si vous les avez manqués, retrouvez les chapitres précédents déjà publiés.
Il s'agit cette fois d'explorer les stratégies des régies publicitaires qui opèrent en arrière-plan : des opérations fort discrètes mais terriblement efficaces…

Traduction Framalang : Côme, goofy, Khrys,Obny, Penguin, Piup, serici.

IV. Collecte de données par les outils des annonceurs et des diffuseurs

29. Une source majeure de collecte des données d'activité des utilisateurs provient des outils destinés au annonceurs et aux éditeurs tels que Google Analytics, DoubleClick, AdSense, AdWords et AdMob. Ces outils ont une portée énorme ; par exemple, plus d'un million d'applications mobiles utilisent AdMob[1], plus d'un million d'annonceurs utilisent AdWords[2], plus de 15 millions de sites internet utilisent AdSense[3] et plus de 30 millions de sites utilisent Google Analytics[4].

30. Au moment de la rédaction du présent rapport, Google a rebaptisé AdWords « *Google Ads* » et DoubleClick « *Google Ad Manager*« , mais aucune modification n'a été apportée aux fonctionnalités principales des produits, y compris la collecte d'informations par ces produits[5]. Par conséquent, pour les besoins du présent rapport, les premiers noms ont été conservés afin d'éviter toute confusion avec des noms de domaine connexes (tels que doubleclick.net).

31. Voici deux principaux groupes d'utilisateurs des outils de Google axés sur l'édition — et les annonces publicitaires :

- Les éditeurs de sites web et d'applications, qui sont des organisations qui

possèdent des sites web et créent des applications mobiles. Ces entités utilisent les outils de Google pour (1) gagner de l'argent en permettant l'affichage d'annonces aux visiteurs sur leurs sites web ou applications, et (2) mieux suivre et comprendre qui visite leurs sites et utilise leurs applications. Les outils de Google placent des cookies et exécutent des scripts dans les navigateurs des visiteurs du site web pour aider à déterminer l'identité d'un utilisateur et suivre son intérêt pour le contenu et son comportement en ligne. Les bibliothèques d'applications mobiles de Google suivent l'utilisation des applications sur les téléphones mobiles.

- Les annonceurs, qui sont des organisations qui paient pour que des bannières, des vidéos ou d'autres publicités soient diffusées aux utilisateurs lorsqu'ils naviguent sur Internet ou utilisent des applications. Ces entités utilisent les outils de Google pour cibler des profils spécifiques de personnes pour que les publicités augmentent le retour sur leurs investissements marketing (les publicités mieux ciblées génèrent généralement des taux de clics et de conversion plus élevés). De tels outils permettent également aux annonceurs d'analyser leurs audiences et de mesurer l'efficacité de leur publicité numérique en regardant sur quelles annonces les utilisateurs cliquent et à quelle fréquence, et en donnant un aperçu du profil des personnes qui ont cliqué sur les annonces.

32. Ensemble, ces outils recueillent des informations sur les activités des utilisateurs sur les sites web et dans les applications, comme le contenu visité et les annonces cliquées. Ils travaillent en arrière-plan — en général imperceptibles par des utilisateurs. La figure 7 montre certains de ces outils clés, avec des flèches indiquant les données recueillies auprès des utilisateurs et les publicités qui leur sont diffusées.

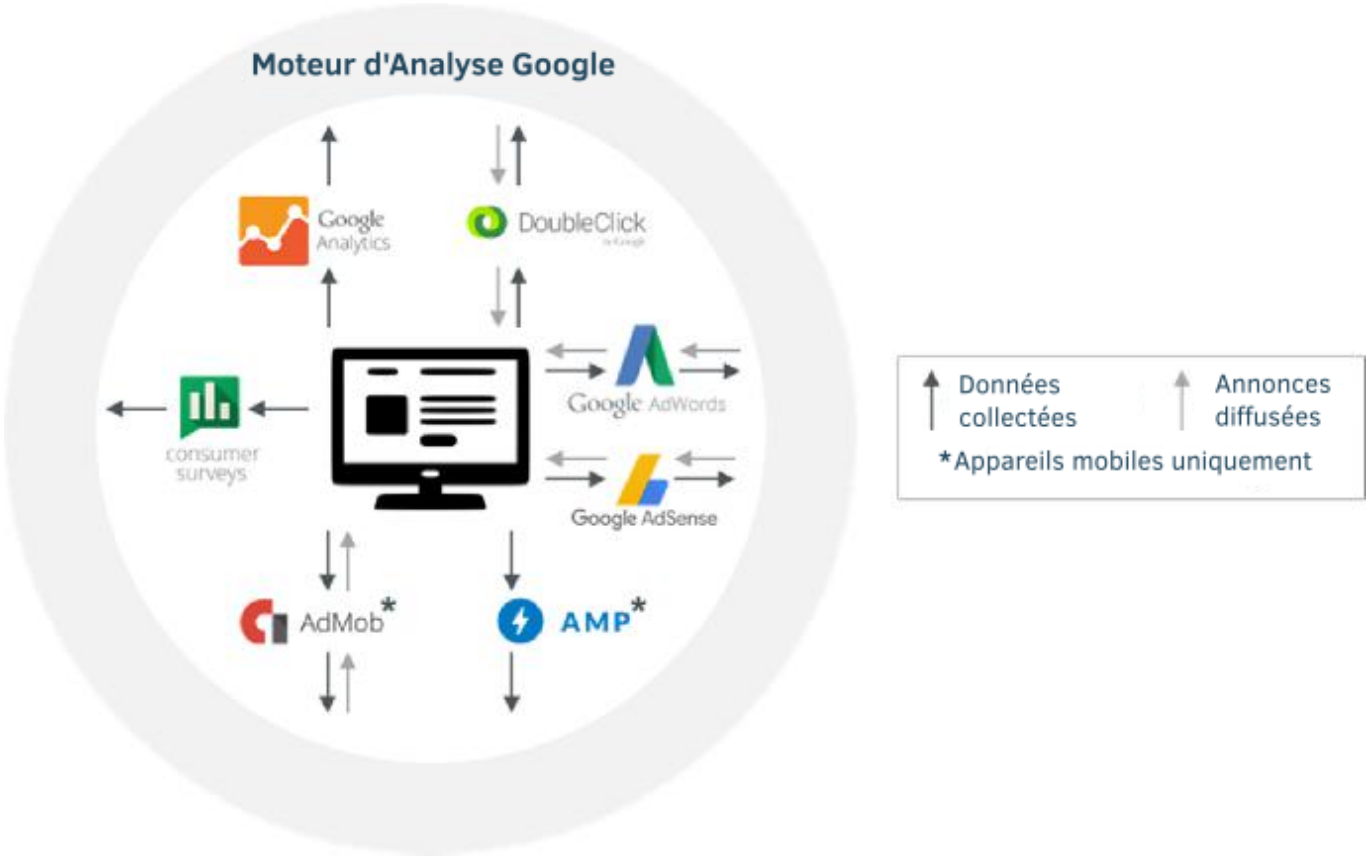

Figure 7 : Produits Google destinés aux éditeurs et annonceurs[6]

33. Les informations recueillies par ces outils comprennent un identifiant non personnel que Google peut utiliser pour envoyer des publicités ciblées sans identifier les informations personnelles de la personne concernée. Ces identificateurs peuvent être spécifiques à l'appareil ou à la session, ainsi que permanents ou semi-permanents. Le tableau 1 liste un ensemble de ces identificateurs. Afin d'offrir aux utilisateurs un plus grand anonymat lors de la collecte d'informations pour le ciblage publicitaire, Google s'est récemment tourné vers l'utilisation d'identifiants uniques semi-permanents (par exemple, les GAID)[7]. Des sections ultérieures décrivent en détail la façon dont ces outils recueillent les données des utilisateurs et l'utilisation de ces identificateurs au cours du processus de collecte des données.

Tableau 1: Identificateurs transmis à Google

Identificateur	Type	Description

GAID/IDFA	Semi-permanent	Chaine de caractères alphanumériques pour appareils Android et iOS, pour permettre les publicités ciblées sur mobile. Réinitialisable par l'utilisateur.
ID client	Semi-permanent	ID créé la première fois qu'un cookie est stocké sur le navigateur. Utilisé pour relier les sessions de navigations. Réinitialisé lorsque les cookies du navigateur sont effacés.
Adresse IP	Semi-permanent	Une unique suite de nombre qui identifie le réseau par lequel un appareil accède à internet.
ID appareil Android	Semi-permanent	Nombre généré aléatoirement au premier démarrage d'un appareil. Utilisé pour identifier l'appareil. En retrait progressif pour la publicité. Réinitialisé lors d'une remise à zéro de l'appareil.
Google Services Framework (GSF)	Semi-permanent	Nombre assigné aléatoirement lorsqu'un utilisateur s'enregistre pour la première fois dans les services Google sur un appareil. Utilisé pour identifier un appareil unique. Réinitialisé lors d'une remise à zéro de l'appareil.
IEMI / MEID	Permanent	Identificateur utilisé dans les standards de communication mobile. Unique pour chaque téléphone portable.
Adresse MAC	Permanent	Identificateur unique de 12 caractères pour un élément matériel (ex. : routeur).

Numéro de série	Permanent	Chaine de caractères alphanumériques utilisée pour identifier un appareil.

A. Google Analytics et DoubleClick

34. DoubleClick et Google Analytics (GA) sont les produits phares de Google en matière de suivi du comportement des utilisateurs et d'analyse du trafic des pages Web sur les périphériques de bureau et mobiles. GA est utilisé par environ 75 % des 100 000 sites Web les plus visités[8]. Les cookies DoubleClick sont associés à plus de 1,6 million de sites Web[9].

35. GA utilise de petits segments de code de traçage (appelés « balises de page ») intégrés dans le code HTML d'un site Web[10]. Après le chargement d'une page Web à la demande d'un utilisateur, le code GA appelle un fichier *analytics.js* qui se trouve sur les serveurs de Google. Ce programme transfère un instantané « par défaut » des données de l'utilisateur à ce moment, qui comprend l'adresse de la page web visitée, le titre de la page, les informations du navigateur, l'emplacement actuel (déduit de l'adresse IP), et les paramètres de langue de l'utilisateur. Les scripts de GA utilisent des cookies pour suivre le comportement des utilisateurs.

36. Le script de GA, la première fois qu'il est exécuté, génère et stocke un cookie spécifique au navigateur sur l'ordinateur de l'utilisateur. Ce cookie a un identificateur de client unique (voir le tableau 1 pour plus de détails)[11] Google utilise l'identificateur unique pour lier les cookies précédemment stockés, qui capturent l'activité d'un utilisateur sur un domaine particulier tant que le cookie n'expire pas ou que l'utilisateur n'efface pas les cookies mis en cache dans son navigateur[12]

37. Alors qu'un cookie GA est spécifique au domaine particulier du site Web que l'utilisateur visite (appelé « cookie de première partie »), un cookie DoubleClick est généralement associé à un domaine tiers commun (tel que doubleclick.net). Google utilise de tels cookies pour suivre l'interaction de l'utilisateur sur plusieurs sites web tiers[13] Lorsqu'un utilisateur interagit avec une publicité sur un

site web, les outils de suivi de conversion de DoubleClick (par exemple, Floodlight) placent des cookies sur l'ordinateur de l'utilisateur et génèrent un identifiant client unique[14] Par la suite, si l'utilisateur visite le site web annoncé, le serveur DoubleClick accède aux informations stockées dans le cookie, enregistrant ainsi la visite comme une conversion valide.

B. AdSense, AdWords et AdMob

38. AdSense et AdWords sont des outils de Google qui diffusent des annonces sur les sites Web et dans les résultats de recherche Google, respectivement. Plus de 15 millions de sites Web ont installé AdSense pour afficher des annonces sponsorisées[15] De même, plus de 2 millions de sites web et applications, qui constituent le réseau Google Display Network (GDN) et touchent plus de 90 % des internautes[16] affichent des annonces AdWords.

39. AdSense collecte des informations indiquant si une annonce a été affichée ou non sur la page web de l'éditeur. Il recueille également la façon dont l'utilisateur a interagi avec l'annonce, par exemple en cliquant sur l'annonce ou en suivant le mouvement du curseur sur l'annonce[17]. AdWords permet aux annonceurs de diffuser des annonces de recherche sur Google Search, d'afficher des annonces sur les pages des éditeurs et de superposer des annonces sur des vidéos YouTube. Pour suivre les taux de clics et de conversion des utilisateurs, les publicités AdWords placent un cookie sur les navigateurs des utilisateurs pour identifier l'utilisateur s'il visite par la suite le site web de l'annonceur ou s'il effectue un achat[18].

40. Bien qu'AdSense et AdWords recueillent également des données sur les appareils mobiles, leur capacité d'obtenir des renseignements sur les utilisateurs des appareils mobiles est limitée puisque les applications mobiles ne partagent pas de cookies entre elles, une technique d'isolement appelée « bac à sable »[19] qui rend difficile pour les annonceurs de suivre le comportement des utilisateurs entre différentes applications mobiles.

41 Pour résoudre ce problème, Google et d'autres entreprises utilisent des « bibliothèques d'annonces » mobiles (comme AdMob) qui sont intégrées dans les applications par leurs développeurs pour diffuser des annonces dans les

applications mobiles. Ces bibliothèques compilent et s'exécutent avec les applications et envoient à Google des données spécifiques à l'application à laquelle elles sont intégrées, y compris les emplacements GPS, la marque de l'appareil et le modèle de l'appareil lorsque les applications ont les autorisations appropriées. Comme on peut le voir dans les analyses de trafic de données (Figure 8), et comme on peut trouver confirmation sur les propres pages web des développeurs de Google[20], de telles bibliothèques peuvent également envoyer des données personnelles de l'utilisateur, telles que l'âge et le genre, tout cela va vers Google à chaque fois que les développeurs d'applications envoient explicitement leurs valeurs numériques vers la bibliothèque.

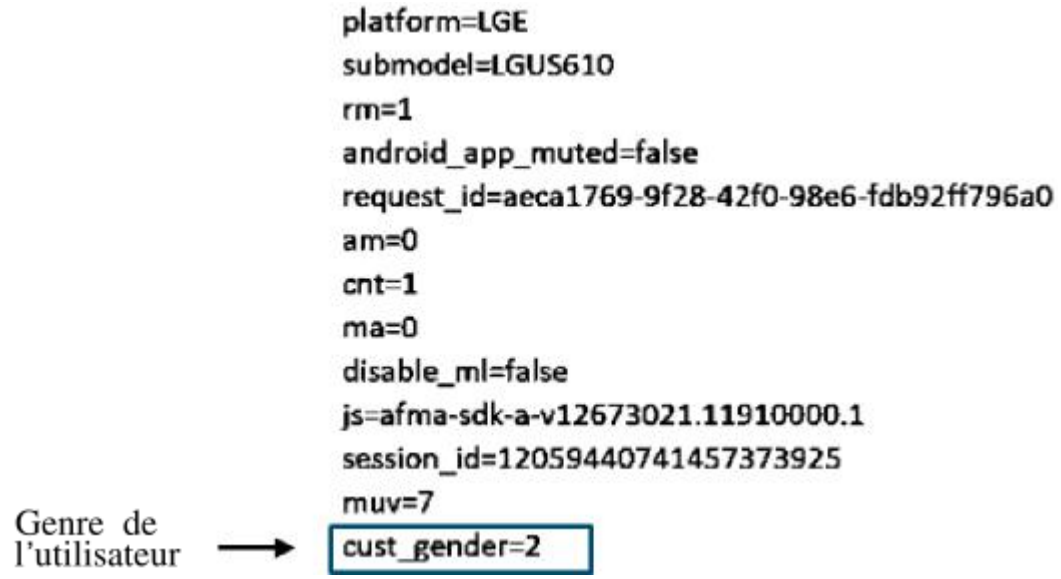

Figure 8 : Aperçu des informations renvoyées à Google lorsqu'une application est lancée

C. Association de données recueillies passivement et d'informations à caractère personnel

42. Comme nous l'avons vu plus haut, Google recueille des données par l'intermédiaire de produits pour éditeurs et annonceurs, et associe ces données à une variété d'identificateurs semi-permanents et anonymes. Google a toutefois la possibilité d'associer ces identifiants aux informations personnelles d'un utilisateur. C'est ce qu'insinuent les déclarations faites dans la politique de confidentialité de Google, dont des extraits sont présentés à la figure 9. La zone

de texte à gauche indique clairement que Google peut associer des données provenant de services publicitaires et d'outils d'analyse aux informations personnelles d'un utilisateur, en fonction des paramètres du compte de l'utilisateur. Cette disposition est activée par défaut, comme indiqué dans la zone de texte à droite.

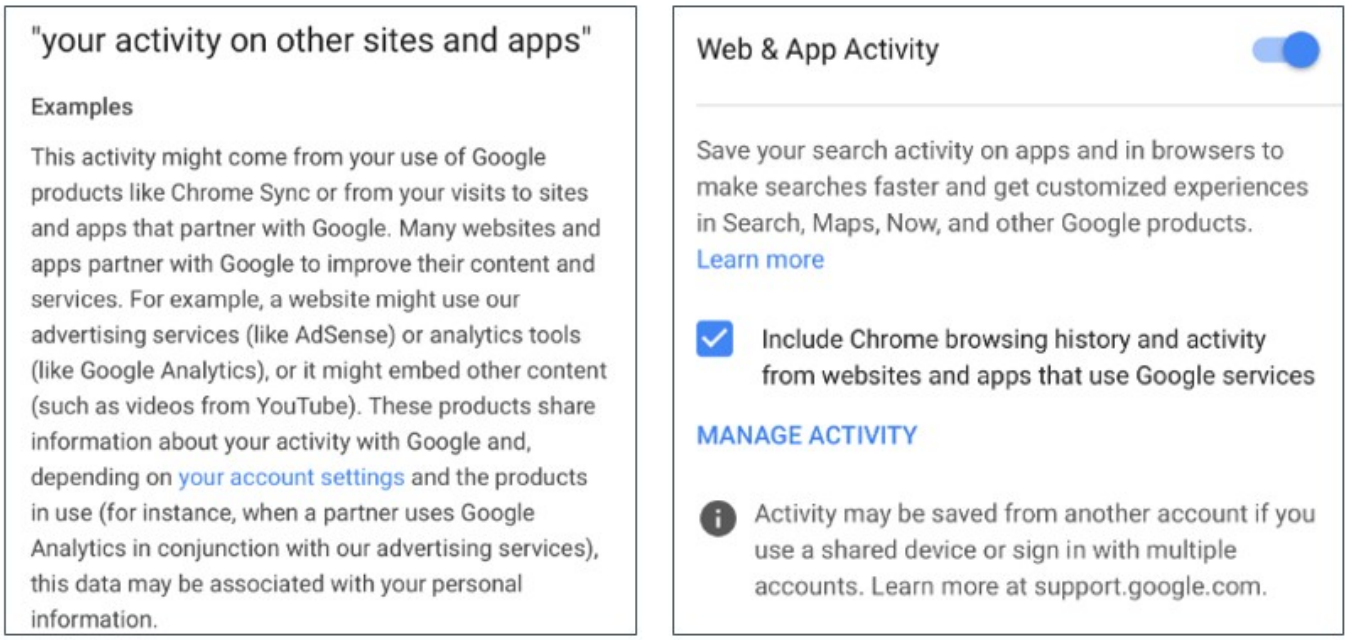

Figure 9 : Page de confidentialité de Google pour la collecte de sites web tiers et l'association avec des informations personnelles[2122].

43. De plus, une analyse du trafic de données échangé avec les serveurs de Google (résumée ci-dessous) a permis d'identifier deux exemples clés (l'un sur Android et l'autre sur Chrome) qui montrent la capacité de Google à corréler les données recueillies de façon anonyme avec les renseignements personnels des utilisateurs.

1) L'identificateur de publicité mobile peut être *désanonymé* grâce aux données envoyées à Google par Android.

44. Les analyses du trafic de données communiqué entre un téléphone Android et les domaines de serveur Google suggèrent un moyen possible par lequel des identifiants anonymes (GAID dans ce cas) peuvent être associés au compte Google d'un utilisateur. La figure 10 décrit ce processus en une série de trois étapes clés.

45. Dans l'étape 1, une donnée de *check-in* est envoyée à l'URL android.clients.google.com/checkin. Cette communication particulière fournit une synchronisation de données Android aux serveurs Google et contient des informations du journal Android (par exemple, du journal de récupération), des messages du noyau, des crash dumps, et d'autres identifiants liés au périphérique. Un instantané d'une demande d'enregistrement partiellement décodée envoyée au serveur de Google à partir d'Android est montré en figure 10.

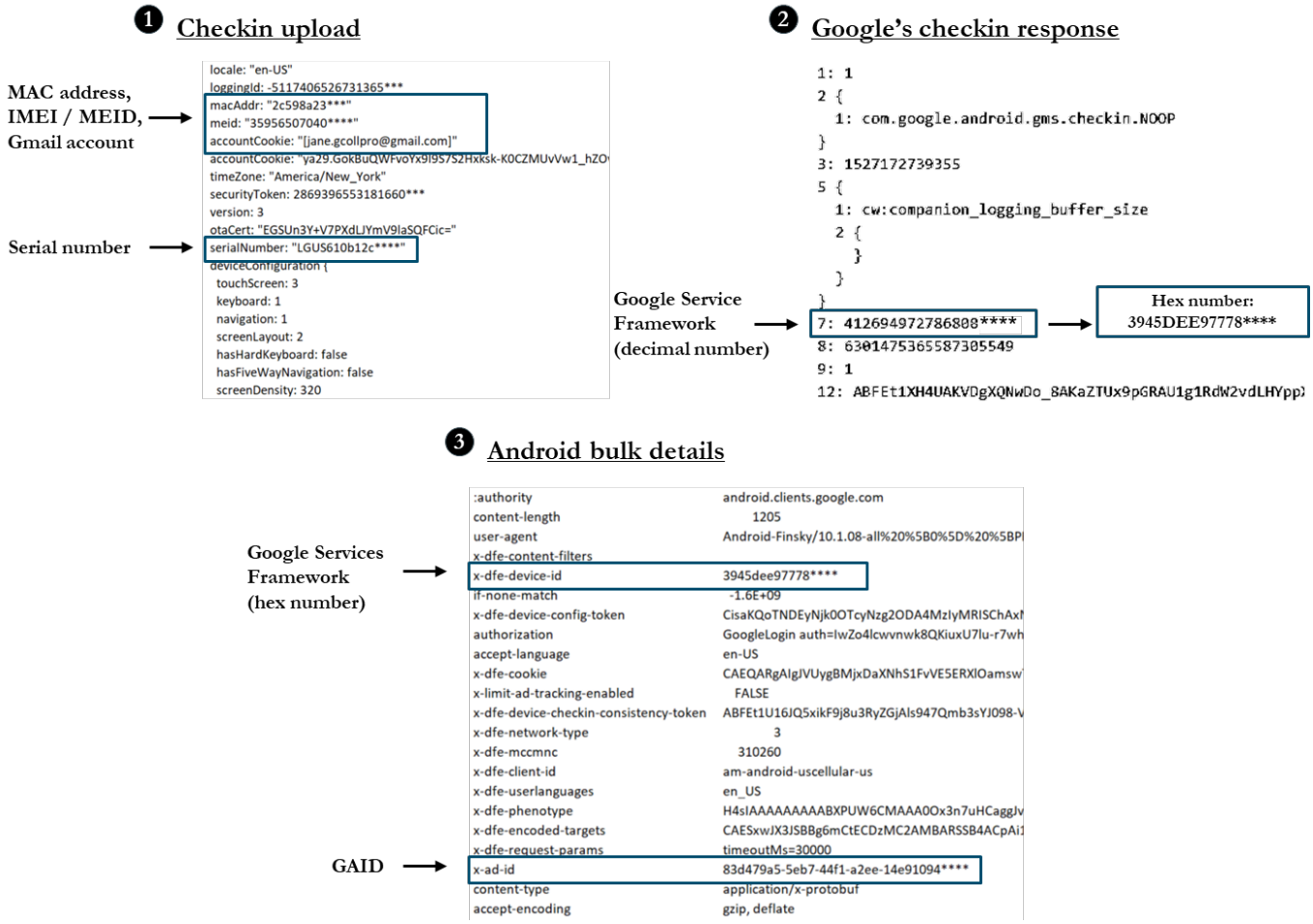

Figure 10 : Les identifiants d'appareil sont envoyés avec les informations de compte dans les requêtes de vérification Android.

46. Comme l'indiquent les zones pointées, Android envoie à Google, au cours du processus d'enregistrement, une variété d'identifiants permanents importants liés à l'appareil, y compris l'adresse MAC de l'appareil, l'IMEI /MEID et le numéro de série du dispositif. En outre, ces demandes contiennent également l'identifiant Gmail de l'utilisateur Android, ce qui permet à Google de relier les informations personnelles d'un utilisateur aux identifiants permanents des appareils Android.

47. À l'étape 2, le serveur de Google répond à la demande d'enregistrement. Ce

message contient un identifiant de cadre de services Google (GSF ID)[23] qui est similaire à l'« Android ID »[24] (voir le tableau 1 pour les descriptions).

48. L'étape 3 implique un autre cas de communication où le même identifiant GSF (de l'étape 2) est envoyé à Google en même temps que le GAID. La figure 10 montre l'une de ces transmissions de données à android.clients.google.com/fdfe/bulkDetails?au=1.

49. Grâce aux trois échanges de données susmentionnés, Google reçoit les informations nécessaires pour connecter un GAID avec des identifiants d'appareil permanents ainsi que les identifiants de compte Google des utilisateurs.

50. Ces échanges de données interceptés avec les serveurs de Google à partir d'un téléphone Android montrent comment Google peut connecter les informations anonymisées collectées sur un appareil mobile Android via les outils DoubleClick, Analytics ou AdMob avec l'identité personnelle de l'utilisateur. Au cours de la collecte de données sur 24 heures à partir d'un téléphone Android sans mouvement ni activité, deux cas de communications d'enregistrement avec des serveurs Google ont été observés. Une analyse supplémentaire est toutefois nécessaire pour déterminer si un tel échange d'informations a lieu avec une certaine périodicité ou s'il est déclenché par des activités spécifiques sur les téléphones.

2) L'ID du cookie DoubleClick est relié aux informations personnelles de l'utilisateur sur le compte Google.

51. La section précédente expliquait comment Google peut désanonymiser l'identité de l'utilisateur via les données passives et anonymisées qu'il collecte à partir d'un appareil mobile Android. Cette section montre comment une telle désanonymisation peut également se produire sur un ordinateur de bureau/ordinateur portable.

52. Les données anonymisées sur les ordinateurs de bureau et portables sont collectées par l'intermédiaire d'identifiants basés sur des cookies (par ex. Cookie ID), qui sont typiquement générés par les produits de publicité et d'édition de Google (par ex. DoubleClick) et stockés sur le disque dur local de l'utilisateur. L'expérience présentée ci-dessous a permis d'évaluer si Google peut établir un lien entre ces identificateurs (et donc les renseignements qui y sont associés) et

les informations personnelles d'un utilisateur.
Cette expérience comportait les étapes ordonnées suivantes :

1. Ouverture d'une nouvelle session de navigation (Chrome ou autre) (pas de cookies enregistrés, par exemple navigation privée ou incognito) ;
2. Visite d'un site Web tiers qui utilisait le réseau publicitaire DoubleClick de Google ;
3. Visite du site Web d'un service Google largement utilisé (Gmail dans ce cas) ;
4. Connexion à Gmail.

53. Au terme des étapes 1 et 2, dans le cadre du processus de chargement des pages, le serveur DoubleClick a reçu une demande lorsque l'utilisateur a visité pour la première fois le site Web tiers. Cette demande faisait partie d'une série de reqêtes comprenant le processus d'initialisation DoubleClick lancé par le site Web de l'éditeur, qui a conduit le navigateur Chrome à installer un cookie pour le domaine DoubleClick. Ce cookie est resté sur l'ordinateur de l'utilisateur jusqu'à son expiration ou jusqu'à ce que l'utilisateur efface manuellement les cookies via les paramètres du navigateur.

54. Ensuite, à l'étape 3, lorsque l'utilisateur visite Gmail, il est invité à se connecter avec ses identifiants Google. Google gère l'identité à l'aide d'une architecture single sign on (SSO) [NdT : authentification unique], dans laquelle les identifiants sont fournis à un service de compte (ici accounts.google.com) en échange d'un « jeton d'authentification », qui peut ensuite être présenté à d'autres services Google pour identifier les utilisateurs. À l'étape 4, lorsqu'un utilisateur accède à son compte Gmail, il se connecte effectivement à son compte Google, qui fournit alors à Gmail un jeton d'autorisation pour vérifier l'identité de l'utilisateur.[25] Ce processus est décrit à la figure 24 de la section IX.E de l'annexe.

55. Dans la dernière étape de ce processus de connexion, une requête est envoyée au domaine DoubleClick. Cette requête contient à la fois le jeton d'authentification fourni par Google et le cookie de suivi défini lorsque l'utilisateur a visité le site web tiers à l'étape 2 (cette communication est indiquée à la figure 11). Cela permet à Google de relier les informations d'identification Google de l'utilisateur à un cookie DoubleClick. Par conséquent, si les utilisateurs n'effacent pas régulièrement les cookies de leur navigateur, leurs informations de navigation sur les pages Web de tiers qui utilisent les services DoubleClick

pourraient être associées à leurs informations personnelles sur Google Account.

Figure 11 : La requête à DoubleClick.net inclut le jeton d'authentification Google et les cookies passés.

56. Il est donc établi à présent que Google recueille une grande variété de données sur les utilisateurs par l'intermédiaire de ses outils d'éditeur et d'annonceur, sans que l'utilisateur en ait une connaissance directe. Bien que ces données soient collectées à l'aide d'identifiants anonymes, Google a la possibilité de relier ces informations collectées aux identifiants personnels de l'utilisateur stockés sur son compte Google.

57. Il convient de souligner que la collecte passive de données d'utilisateurs de Google à partir de pages web tierces ne peut être empêchée à l'aide d'outils populaires de blocage de publicité[26], car ces outils sont conçus principalement pour empêcher la présence de publicités pendant que les utilisateurs naviguent sur des pages web tierces[27]. La section suivante examine de plus près l'ampleur de cette collecte de données.