Analisi tecnica di Google Analytics

Indice

Introduzione

1. Google Analytics è un servizio di analisi fornito da Google LLC (Google) che traccia e riporta le attività dei visitatori di un sito web fornendo agli amministratori dello stesso statistiche aggregate rispetto a innumerevoli dimensioni: demografiche, tecniche, economiche e comportamentali.
2. Tramite tale software gli amministratori di un sito web possono sfruttare sofisticate analisi sui dati dei diversi visitatori senza però possedere la capacità di accesso a quelli "grezzi", che sono effettivamente registrati da Google a fronte di ogni singola azione effettuata dall'utente [1].
3. L'introduzione di Google Analytics su un sito web comporta l'inserimento all'interno del codice HTML che sarà inviato al browser del visitatore di un riferimento a un file o un frammento di codice di programma JavaScript fornito da Google. Tale frammento include sempre un identificativo di tracciamento (chiamato Tracking ID o Measurement ID, a seconda della versione del software) attribuito da Google al sito.

[1] https://support.google.com/analytics/answer/7029846

4. Esistono molti modi di effettuare tale inclusione nelle pagine HTML inviate al browser dei visitatori, fra cui:
    - l'inclusione diretta in un tag `<script>` di un riferimento a www.google-analytics.com/analytics.js
    - l'introduzione di piccoli "snippet" (frammenti) di codice JavaScript forniti da Google stessa
    - l'abilitazione del servizio di Google Analytics in altri servizi forniti da Google stessa, come per esempio Google Tag Manager.
5. In tutti i casi, l'inclusione di Google Analytics nel codice HTML delle pagine inviate al browser dell'utente determina automaticamente l'invio di una richiesta HTTP ai server di Google.

L'identificazione del visitatore

6. La prima richiesta inviata ai server di Google per ottenere il JavaScript da eseguire causa un primo trasferimento di diversi dati personali del visitatore a favore di Google LLC, fra cui:
    - indirizzo IP
    - data, ora e timezone della navigazione sul sito
    - stringa identificativa dello User Agent
    - referer, ovvero l'indirizzo della pagina che il visitatore sta visualizzando
    - eventuali cookie precedentemente impostati
7. Questi dati iniziali, il cui trasferimento automatico (in presenza del codice di attivazione di Google Analytics) è determinato dalle specifiche dei protocolli TCP/IP e HTTP, sono sufficienti per Google a identificare il visitatore nella maggioranza dei casi.
8. Per esempio, ogni qualvolta l'utente utilizzi in contemporanea al sito web su cui è installato Google Analytics uno dei servizi o delle App gratuite offerte da Google che prevedono l'autenticazione, sarà sufficiente confrontare l'IP o il cookie identificativo inviati alla prima richiesta con quelli degli utenti correntemente autenticati per stabilirne l'identità.
9. Una volta avviato il proprio programma JavaScript nel "runtime" (ambiente di esecuzione) del browser, Google potrà identificare il visitatore attraverso numerosi parametri offerti dal "runtime" stesso che, fornendo un numero di bit di entropia[2] nettamente superiore all'IP, rendono identificabile l'utente, permettendo il cosiddetto *browser fingerprinting*[3].

[2] https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf
[3] https://www.amiunique.org/fp#table_js_wrapper

10. La straordinaria quantità di informazioni raccolte all'insaputa degli utenti su milioni di applicazioni mobili e siti web rende estremamente facile, per Google, l'utilizzo di un qualsiasi insieme di dati personali anche non identificativi, con funzione di identificativi di fatto[4], deanonimizzando i dati ad essi associati[5] per svelare l'identità del soggetto da cui sono stati emessi e perfezionarne la profilazione.

11. Dunque per Google è sempre possibile identificare i cittadini europei e tracciarne l'attività, le opinioni e gli interessi su tutti i siti web (e tutte le App) che utilizzino Google Analytics, per ricondurne poi le registrazioni all'identità personale. In altri termini, anche in totale assenza dell'IP, tutti i dati raccolti attraverso Google Analytics rimangono riconducibili all'interessato, restando dati personali soggetti al GDPR.

La compromissione della riservatezza nel rapporto con l'Ente

12. Nella sua configurazione predefinita, Google Analytics registra automaticamente 43 tipi diversi di eventi[6] per tracciare ogni micro-interazione che il visitatore effettua con il sito web, raccogliendo per ogni tipologia di evento diversi parametri specifici. A questi eventi predefiniti si aggiungono eventi "su misura" e peculiari di ogni sito web, introdotti dagli amministratori secondo le raccomandazioni di Google[7], nonché eventi personalizzati registrati a discrezione dell'amministratore del sito web[8].

13. Oltre a raccogliere i dati necessari all'identificazione del visitatore, il programma JavaScript di Google Analytics può infatti annotare i vari eventi registrati dal browser a valle di ogni azione dell'utente: ogni click, ogni selezione del testo, ogni zoom, ogni tasto premuto, ogni scroll della pagina potrà essere utilizzato per arricchire il profilo comportamentale dell'utente.

14. Ciascuno di questi eventi registrati da Google Analytics include inoltre informazioni tecniche (device, sistema operativo, versione del browser, dimensione dello schermo e molte altre), identificativi (advertising_id, user_id, user_pseudo_id), coordinate geografiche e informazioni storico-comportamentali sul visitatore[9].

15. Infine, ai dati personali derivati dalle micro-interazioni dell'utente si possono aggiungere quelli deducibili dai contenuti che visualizza, tutti

[4] https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
[5] https://www.csee.umbc.edu/~kunliu1/p3dm08/proceedings/2.pdf
[6] https://support.google.com/analytics/answer/9234069
[7] https://support.google.com/analytics/answer/9267735
[8] https://support.google.com/analytics/answer/12229021
[9] https://support.google.com/analytics/answer/7029846

accessibili al JavaScript di Google Analytics, inclusi quelli presenti nelle eventuali pagine di profilo dell'utente.

16. Di conseguenza, tutte le comunicazioni fra i cittadini e gli Enti pubblici veicolate da siti web sorvegliati da Google Analytics, subiscono una grave compromissione della riservatezza che le dovrebbe caratterizzare.

Inefficacia delle misure di anonimizzazione offerte da Google

17. Google Analytics fornisce da tempo la possibilità per gli amministratori dei siti web di abilitare la cosiddetta IP Anonymization[10].

18. Si tratta sostanzialmente di una configurazione a fronte della quale Google promette contrattualmente di scartare "appena tecnicamente possibile" gli ultimi 8 bit dell'indirizzo IP del visitatore (costituito, nella versione 4 del protocollo IP, da una sequenza di 32 bit).

19. Analogamente, nella versione 4 di Google Analytics, Google promette contrattualmente di scartare l'intero IP dell'utente "appena tecnicamente possibile".

20. Tale rimozione **risulta però del tutto insufficiente a costituire un'efficace misura tecnica supplementare a protezione dei dati personali dell'utente**, per diverse ragioni:

    - anzitutto perché è Google stessa a scartare tali dati e dopo averli ricevuti. Subito prima di scartarli, potrebbe essere costretta a inviarli e all'insaputa del Titolare del trattamento, verso agenzie governative USA nei termini previsti dalle norme statunitensi applicabili;
    - il numero di bit di entropia forniti dall'IP del visitatore a cui Google promette contrattualmente di rinunciare è nettamente inferiore al numero di bit di entropia forniti, in media, dal runtime di esecuzione del browser;
    - anche con Google Analytics 4, Google sostanzialmente promette di scartare dati ampiamente ridondanti di cui non ha comunque bisogno per identificare, tracciare e profilare l'utente. Esattamente come avviene con la versione precedente peraltro, in cui gli 8 bit dell'IP scartati "appena tecnicamente possibile" erano compensati dai 10 bit di entropia forniti in media dalla stringa identificativa del browser[11] che viene comunque trasferita.

[10] https://support.google.com/analytics/answer/2763052
[11] https://www.eff.org/deeplinks/2010/01/tracking-by-user-agent

Reverse proxying Google Analytics

21. In astratto, una delle possibili misure tecniche supplementari che i Titolari del Trattamento potrebbero adottare per proteggere i dati personali dei visitatori pur continuando a utilizzare Google Analytics, consiste nel mediare attraverso un reverse proxy specificatamente programmato, le comunicazioni fra i visitatori e i server di Google.

22. Tale proxy dovrebbe intercettare tutte le comunicazioni fra il browser del visitatore e i server di Google; effettuare una ispezione profonda dei pacchetti (*deep packet inspection*) e rimuovere qualsiasi dato personale che possa permettere a Google di identificare il visitatore o arricchirne la profilazione, prima di inoltrare tali pacchetti ridotti ai server di Google.

23. Teoricamente, attraverso una simile intermediazione, il Titolare del trattamento potrebbe efficacemente nascondere a Google l'IP del visitatore, le sue coordinate satellitari, l'identificativo dello User Agent, la data e l'ora delle richieste (conservando per un tempo variabile i dati raccolti prima di inviarli) e tutti i parametri del runtime in cui il programma JavaScript di Google Analytics verrebbe eseguito.

24. In pratica, questa soluzione soffre di gravi problemi tecnici che, oltre a renderla estremamente costosa, ne minano l'efficacia e l'affidabilità nel lungo periodo:

    - l'efficacia della rimozione dei dati personali dipende dalla specifica versione di Google Analytics in esecuzione: il filtro operato dal reverse proxy dovrebbe essere continuamente aggiornato da un'organizzazione indipendente e dovrebbe impedire il transito a qualsiasi dato non specificatamente ed esplicitamente autorizzato a priori dal Titolare del trattamento;
    - ciò comporterebbe un degrado inevitabile della qualità delle statistiche a fronte di ogni minimo aggiornamento del sistema;
    - l'introduzione e il continuo monitoraggio di tale intermediazione obbligherebbe alla messa in opera di datacenter dedicati sotto il controllo di terze parti indipendenti. Ciò comporterebbe costi fissi e ricorrenti molto difficili da stimare a priori, ma nettamente superiori alle soluzioni alternative disponibili, prima fra tutte il sistema Web Analytics Italia che è raccomandato e fornito gratuitamente dall'Agenzia per l'Italia Digitale;
    - la presenza di "reverse proxy" verso Google sotto domini specifici di ciascuna PA, impedirebbe ai visitatori di proteggere la propria privacy tramite strumenti come AdAway, Pi-hole, uMatrix o uBlock Origin che impediscono al browser di inviare dati personali a Google;

25. L'efficacia teorica di tale misura tecnica supplementare, si scontra inoltre con la già menzionata possibilità, da parte di Google, di utilizzare dati personali descrittivi non trasferiti a fini di identificazione con funzione di

identificativi di fatto, per deanonimizzare efficacemente l'intera sessione applicativa del visitatore.

La collocazione geografica dei data center

26. Google dichiara di elaborare presso i suoi data center europei i dati dei cittadini europei.
27. Tuttavia il software utilizzato per tale elaborazione è controllato centralmente dagli USA come dimostrato dal blackout globale del dicembre 2020 e dalla sua simultanea risoluzione in tutto il pianeta.
28. Il controllo sul software che elabora i dati dei cittadini europei implica ovviamente la possibilità di accedere ai dati stessi. Ad esempio Google potrebbe essere costretta, nei termini delle già citate normative USA, a inviare presso uno o più dei suoi data center europei aggiornamenti in grado di prelevare i dati di interesse di una agenzia governativa e rimuovere ogni evidenza di tale "data breach" prima del riavvio dei servizi.

XXXXXXXXXXXXX, 22 giugno 2022

Firme

Giacomo Tesio

GIACOMO TESIO

Massimo Maria Ghisalberti

MASSIMO MARIA GHISALBERTI