Allegato Tecnico

Indice

Il presente allegato tecnico descrive i trasferimenti sistematici avviati automaticamente dai siti web che utilizzano Google Analytics (in una qualsiasi versione) o le API di Google Fonts, nonché le problematiche per la privacy e la sicurezza dei visitatori che tali strumenti determinano.

L'allegato descrive inoltre l'output dell'osservatorio automatico distribuito di Monitora PA, che include le evidenze raccolte dei trasferimenti in corso.

Google Analytics

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.

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.

6. A valle della risposta dal server, l'esecuzione dello script JavaScript restituito determina la creazione di alcuni cookie utilizzati da Google Analytics per identificare e tracciare l'utente. Il nome di tali cookie può variare in funzione della configurazione del servizio, della versione dello stesso o di ragioni tecniche come collisioni con i nomi dei cookie già presenti, ma tipicamente includono cookie denominati `_ga` e `_ga_<container-id>` usati in entrambe le versioni per distinguere gli utenti fra loro identificandoli per la durata massima di due anni [2] [3] [4]. Tali cookie partecipano alla definizione del Client ID [5] che Google utilizza per identificare il visitatore.

[1] https://support.google.com/analytics/answer/7029846
[2] https://support.google.com/analytics/answer/11397207?hl=it
[3] https://developers.google.com/analytics/devguides/collection/gtagjs/cookie-usage
[4] https://developers.google.com/analytics/devguides/collection/analyticsjs/cookie-usage
[5] https://developers.google.com/analytics/devguides/collection/analyticsjs/cookies-user-id

Google Fonts

7. Google Fonts è un servizio per webmaster fornito gratuitamente da Google e composto di:
    - una libreria di font con licenza libera
    - un sito web interattivo per navigare tale libreria
    - un servizio HTTP che fornisce un'API di tipo ReST [6] per incorporare tali font direttamente dai server di Google .
8. Sebbene le licenze con cui i font sono distribuiti ne permettano l'installazione sul server del sito che ne fa uso, l'utilizzo dell'API di Google viene spesso preferito per scaricare su Google il consumo della banda necessaria alla loro distribuzione, nonché la responsabilità del loro aggiornamento.
9. Esistono molti modi di effettuare tale inclusione nelle pagine HTML inviate al browser dei visitatori, fra cui:
    - l'inclusione diretta in un tag `<link rel="stylesheet">` di un riferimento a https://fonts.googleapis.com/css specificando in query string i font, gli effetti e i parametri desiderati [7]
    - l'introduzione all'interno di CSS locali del sito di direttive `@import` [8] che puntino a indirizzi web costruiti secondo la stessa specifica
    - l'introduzione all'interno di CSS locali del sito di direttive `@font-face` [9] che facciano riferimento a file TTF, WOFF o WOFF2 serviti dal server https://fonts.gstatic.com
10. Durante lo scaricamento di suddette risorse, Google istruisce il browser affinché mantenga la risposta in una cache locale, dedicata allo specifico sito. [10] Nel caso dei CSS, la direttiva `Cache-Control` impone al browser di mantenere il dato in cache per ventiquattro ore (`max-age=86400`) dopo le quali la direttiva `stale-while-revalidate=604800` impone al browser di rivalidare la cache, richiedendo nuovamente la risorsa in una richiesta asincrona nei sei giorni successivi.
11. Inoltre Google riconosce esplicitamente di includere i dati personali raccolti attraverso l'utilizzo di Google Fonts nelle analisi effettuate tramite Google Analytics [11], rendendoli anche parzialmente accessibili al pubblico in una pagina dedicata[12].

[6] https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm
[7] https://developers.google.com/fonts/docs/getting_started
[8] https://developer.mozilla.org/en-US/docs/Web/CSS/@import
[9] https://developer.mozilla.org/en-US/docs/Web/CSS/@font-face
[10] https://developer.chrome.com/blog/http-cache-partitioning/#how-will-cache-partitioning-affect-chromes-http-cache
[11] https://developers.google.com/fonts/faq#what_does_using_the_google_fonts_api_mean_for_the_privacy_of_my_users
[12] https://fonts.google.com/analytics

L'identificazione del visitatore

12. La prima richiesta inviata ai server di Google per ottenere il JavaScript, i CSS o i font da utilizzare durante l'elaborazione della pagina, nonché tutte le richieste inviate per convalidare o aggiornare la versione in cache delle stesse risorse una volta al giorno, causano il 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
    - lingue note all'utente, attraverso l'header HTTP Accept-Language
    - data dell'ultima visita, attraverso l'header HTTP If-Modified-Since
    - l'indirizzo della pagina che incorpora la risorsa, attraverso l'header HTTP Referer
    - eventuali cookie precedentemente impostati

13. Questi dati iniziali, il cui trasferimento automatico è determinato dalle specifiche dei protocolli TCP/IP e HTTP, sono sufficienti per Google a identificare il visitatore nella maggioranza dei casi.

14. Per esempio, ogni qualvolta l'utente utilizzi, in contemporanea al sito web su cui è incorporata una risorsa fornita da Google, 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.

15. Nel caso di Google Analytics o altri prodotti basati su programmi JavaScript scaricati ed eseguiti dai suoi server, Google potrà identificare il visitatore attraverso numerosi parametri offerti dal "runtime" stesso che, fornendo un numero di bit di entropia[13] nettamente superiore all'IP, rendono identificabile l'utente, permettendo il cosiddetto *browser fingerprinting*[14].

16. 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[15], deanonimizzando i dati ad essi associati[16] per svelare l'identità del soggetto da cui sono stati emessi e perfezionarne la profilazione.

17. 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 i suoi servizi, per ricondurne poi le registrazioni all'identità personale. In altri termini, anche in totale assenza dell'IP,

[13] https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf
[14] https://www.amiunique.org/fp#table_js_wrapper
[15] https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf
[16] https://www.csee.umbc.edu/~kunliu1/p3dm08/proceedings/2.pdf

tutti i dati raccolti attraverso le risorse incorporate da Google rimangono riconducibili all'interessato, restando dati personali soggetti al GDPR.

La compromissione della riservatezza nel rapporto con i Partiti

18. Lo scaricamento dei font o dei CSS da server di Google Fonts comporta il trasferimento sistematico di dati personali sensibili come sono gli interessi e le opinioni politiche dei cittadini che Google potrà dedurre dalla frequenza delle visite al sito di ciascun partito.

19. Google Analytics, nella sua configurazione predefinita **in tutte le versioni individuate presso i siti analizzati** (Urchin, Universal Analytics e Google Analytics 4), registra automaticamente oltre 40 tipi diversi di eventi[17] 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[18], nonché eventi personalizzati registrati a discrezione dell'amministratore del sito web[19].

20. 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.

21. 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[20].

22. Infine, ai dati personali derivati dalle micro-interazioni dell'utente si possono aggiungere quelli deducibili dai contenuti che visualizza, tutti accessibili al JavaScript di Google Analytics, inclusi quelli presenti nelle eventuali pagine di profilo dell'utente.

23. Di conseguenza, tutte le comunicazioni fra i cittadini e i partiti politici veicolate da siti web sorvegliati da Google Analytics e che usano Google Fonts, subiscono una grave compromissione della riservatezza che le dovrebbe caratterizzare.

[17] https://support.google.com/analytics/answer/9234069
[18] https://support.google.com/analytics/answer/9267735
[19] https://support.google.com/analytics/answer/12229021
[20] https://support.google.com/analytics/answer/7029846

Inefficacia delle misure di anonimizzazione offerte da Google

24. Google Analytics fornisce da tempo la possibilità per gli amministratori dei siti web di abilitare la cosiddetta IP Anonymization[21].

25. 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).

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

27. Analogamente, Google promette contrattualmente di non registrare l'IP dei visitatori che richiedono CSS o font ai server di Google Fonts, ma riconosce esplicitamente di "registra i dettagli della richiesta HTTP, inclusi il timestamp, l'URL richiesto e tutte le intestazioni HTTP (incluse la stringa referrer e user agent) fornite in relazione all'utilizzo dell'API CSS". [22]

28. 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[23] che viene comunque trasferita.

[21] https://support.google.com/analytics/answer/2763052
[22] https://developers.google.com/fonts/faq#what_does_using_the_google_fonts_api_mean_for_the_privacy_of_my_users
[23] https://www.eff.org/deeplinks/2010/01/tracking-by-user-agent

Reverse proxying Google Analytics

29. 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.

30. 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.

31. 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.

32. 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;
    - 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;

33. 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

34. Google dichiara di elaborare presso i suoi data center europei i dati dei cittadini europei.

35. 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.

36. 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.

Le evidenze raccolte dall'osservatorio di Monitora PA

37. L'osservatorio automatico distribuito di Monitora PA analizza una serie di siti web per identificare problemi di non conformità alla normativa, producendo una serie di file TSV (tab-separated values) contenenti le evidenze delle problematiche raccolte durante l'analisi. Copia di tale output è disponibile nel archivio MonitoraPA_AnalisiPartiti_2022-09-12.zip allegato alla presente segnalazione.

38. L'output di tale analisi consiste in un archivio contenente una copia dell'elenco di siti internet alla data di esecuzione (nel caso attuale, `PartitiPolitici.tsv`) nonché i diversi risultati delle verifiche eseguite, ciascuno in un file dedicato nella directory `check/` o in una sottocartella.

39. I nomi dei file in output, ordinati alfabeticamente, riproducono l'ordine in cui le diverse verifiche vengono effettuate. In particolare

    - `check\browsing\000-actual-url_PartitiPolitici.tsv` identifica la URL effettiva della verifica, a valle di eventuali redirezioni
    - `check\browsing\000-cookies_PartitiPolitici.tsv` registra i cookie impostati da ogni sito internet **prima** del consenso al cookie banner
    - `check\browsing\100-google-analytics-clientID_PartitiPolitici.tsv` registra il Client ID associato da Google al visitatore **prima** del consenso al trasferimento di dati personali
    - `check\browsing\103-google-fonts_PartitiPolitici.tsv` registra le evidenze raccolte sull'utilizzo delle API di Google Fonts o dello scaricamento di font dai suoi server (con il conseguente trasferimento sistematico di dati personali ad esso conseguente)

- `check\browsing\300-consent_PartitiPolitici.tsv` registra il pulsante di consenso cliccato (fra `<` e `>`) seguito dal testo completo del cookie banner proposto al visitatore
- `check\browsing\301-google-analytics-trackingID_PartitiPolitici.tsv` registra il Tracking ID / Measurement ID assegnato da Google al sito web
- `check\browsing\900-google-analytics-clientID_PartitiPolitici.tsv` registra il Client ID assegnato da Google al visitatore **dopo** il consenso al trasferimento di dati personali
- `check\browsing\999-cookies_PartitiPolitici.tsv` registra i cookie impostati da ogni sito internet **dopo** il consenso al cookie banner

40. Il formato dei TSV, attentamente studiato per massimizzare la possibilità dei cittadini di utilizzare personalmente l'osservatorio di Monitora PA [24], è costituito da sei colonne dalla semantica posizionale. Nell'ordine da sinistra a destra le colonne contengono:

    1. Identificativo del titolare del trasferimento (come presente nella prima colonna del file `PartitiPolitici.tsv`)
    2. Tipologia di automatismo verificato (nel caso in questione, Web)
    3. Indirizzo dell'automatismo da verificare
    4. Data e ora della verifica
    5. Stato di completamento della verifica
       `"1"` significa che la verifica è stata completata con successo
       `"0"` significa che non è stato possibile completare la verifica
    6. Evidenze riscontrate della problematica oggetto di verifica:
       - se la colonna è vuota, il problema oggetto di analisi non è stato individuato sul sito dall'osservatorio automatico
       - se la colonna non è vuota, contiene le evidenze del problema individuate dall'osservatorio (o dati sull'errore che ha impedito la verifica, quando la colonna precedente contiene `"0"`)

[24] https://github.com/MonitoraPA/monitorapa/blob/main/ARCHITETTURA.md

XXXX XXX XXXXXXX, 13 settembre 2022

Firme

Giacomo Tesio

GIACOMO TESIO

Massimo Maria Ghisalberti

MASSIMO MARIA GHISALBERTI