For Context / High level RoadMap: https://docs.google.com/document/d/1Cqvx3umg5_4s3tcRneVFIffzyc4OpWqwue1P-_CFGhA/edit?usp=sharing

Contributors: Shashi Gharti, Maddie Shang

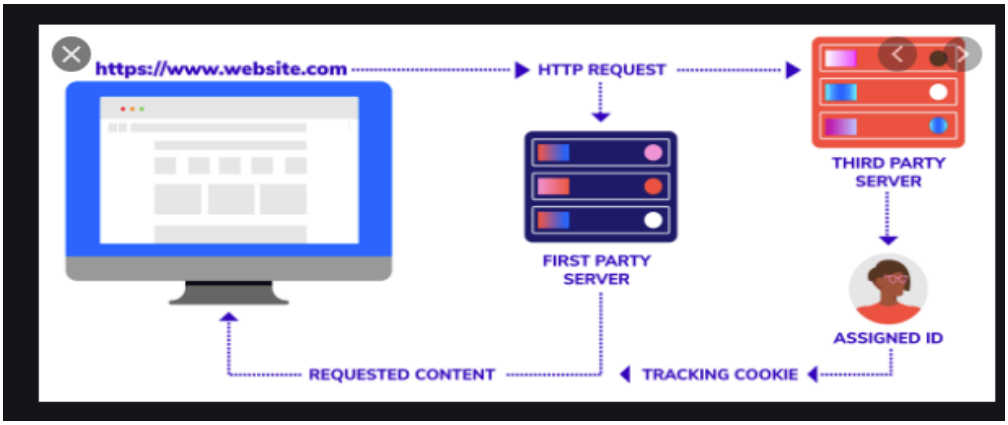
Topics

1. How ad exchanges work in the context of web apps / websites / the browser
    a. how are users tracked / how good is the continuity
    b. how are user choices recorded and communicated (to the ad exchange, to the buyers of the ad impression)

Date Jan 4, 2021

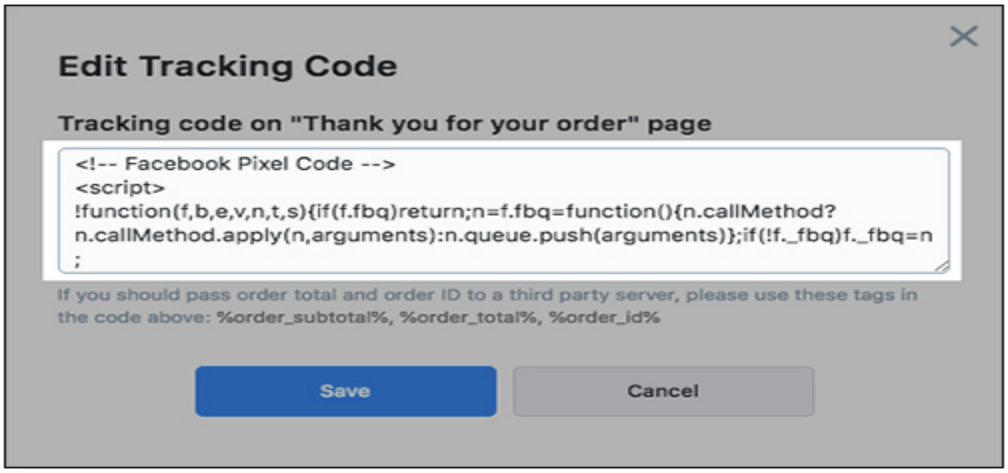
Week I

a. How are users tracked?



1. **Cookies**
    a. **First Party Cookies (**stores information such settings and necessary information that improves browsing experience, it can be used to identify and track).
    b. Third Party Cookies
        i. **Cross-site tracking:** Collecting browsing data from numerous sources(devices, websites browsing history etc).
        ii. **Retargeting:** Collect data based on search history.
        iii. Ad-serving
        iv. Example: Facebook Pixel, Google Analytics, Hubspot.
2. **Super Cookies**
3. **Http Referrer**
4. **IP Information**
5. **User Agents**
6. **Browser Fingerprinting** (timezone, location, default languages etc)

7. **Tracking Pixel**



b. References

- https://www.digitalmarketer.com/blog/what-is-tracking-pixel/
- https://www.bounteous.com/insights/2020/04/09/google-chrome-third-party-cookies/
- https://us.epsilon.com/blog/1st-party-vs-3rd-party-cookies-whats-the-difference#:~:text=A%20first%2Dparty%20cookie%20is,provide%20a%20good%20user%20experience.

Next Week Study Plan:

- Detailed information about Browser Fingerprinting.
    - Level of danger; how much does it isolate the identity.
    - Not allow people to track information about us(even browser fingerprinting).
- How information of the same user collected from different devices and websites are merged and used to identify and track the user?
    - Advertising network work.
- Detailed information about Advertising Networks and how they collect and store data. How are these datas used to target users for ads?
- Super Cookies.
- Tracking pixel.
    - Emails.
    - Browsing.

Date Jan 10, 2021

Week II

Super Cookies

Super cookies are worse than regular cookies used for tracking. But unlike regular cookies, deleting and clearing browsing data does not help to prevent being tracked. Even though it is named super cookie it is not similar to http cookie because it is not stored in our browser.

How are they used to track information:

1. These cookies are injected by ISPs into HTTP header and the information injected(id) is unique to the user.
2. Using this unique identity, it can be used to collect wide range of data such as:
    a. Website visited by the user to understand browsing habits
    b. Information stored by traditional cookies such as login information, cache files etc

Properties:

- They do not use local storage instead it is done by ISP by injecting a unique identifier to the HTTP header.
- Users might not be aware of it and it is difficult to detect. ISP might be secretly collecting the data.
- ISP after collecting the data can sell it to third parties.
- Super cookies can't be blocked or deleted. Super cookies can store the information even after the normal cookies are deleted. Super cookies can even restore the data of normal cookies after they are deleted.
- Even adblocking app can't block super cookies .
- Changing browsers doesn't help and users can be tracked across different websites.

> *[S]uppose an ad network assigned you a cookie with the unique value "cookie1," and Verizon assigned you the X-UIDH header "old_uid." When Verizon changes your X-UIDH header to a new value, say "new_uid," the ad network can connect "new_uid" and "old_uid" to the same cookie value "cookie1" and see that they all three values represent the same person. Similarly, if you subsequently clear cookies, the ad network will assign a new cookie value "cookie2." Since your X-UIDH value is the same (say, "new_uid") before and after clearing cookies, the ad network can connect "cookie1" and "cookie2" to the same X-UIDH value "new_uid." The back-and-forth bootstrapping of identity makes it impossible to truly clear your tracking history while the X-UIDH header is enabled.*

Ref: https://tinyurl.com/yy6bsdsd

How to protect from super cookies:

To protect tracking from super cookies, users should use an encrypted connection https or VPN.

References:

https://tinyurl.com/yxq2bsna

https://tinyurl.com/yy6bsdsd

Browser Fingerprinting.

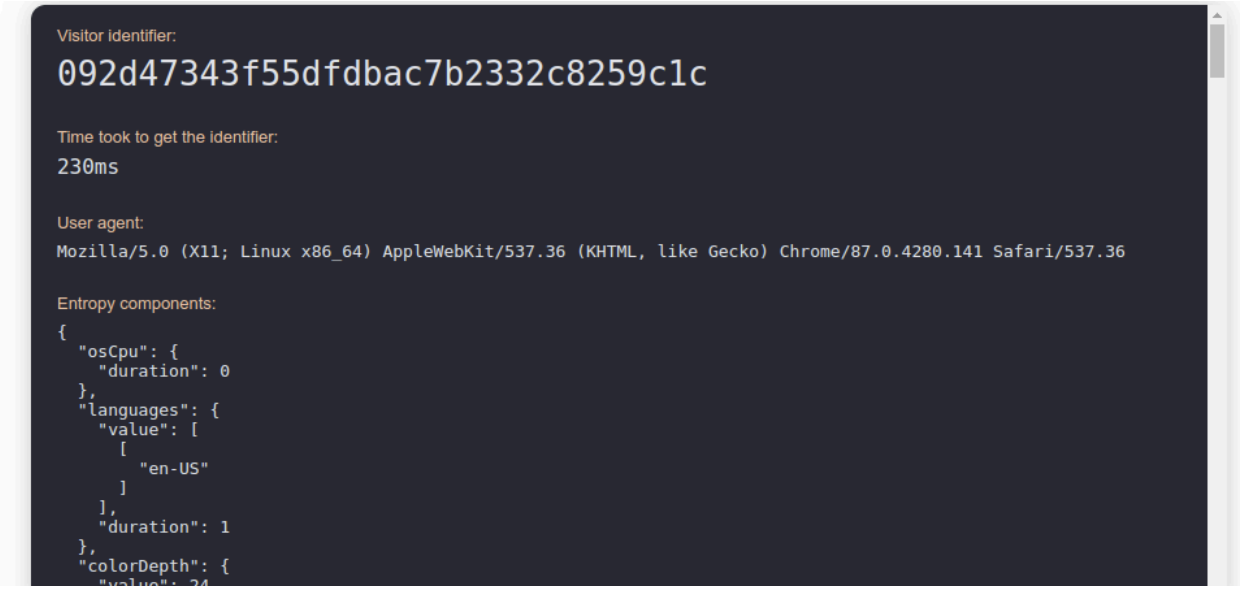
How is it collected?

Browser fingerprint is collected by the websites by reading the device details while the user is surfing the website. They can track us without knowing our ip or without adding cookies to our browser. The more data the browser sends, the more easier it will be for the websites(trackers) to identify us.
Sites that helps identify the digital fingerprint of our browser: https://coveryourtracks.eff.org/ , https://amiunique.org/

Usage:

Ad and personalization(highly used, advertisers), Fraud protection(banks)

Example of browser fingerprinting using https://fingerprintjs.github.io/fingerprintjs/ library. Using the information a unique id is created.



What is the level of danger?. How much does it isolate the identity?

Level of danger:

No Transparency and Control:
The data collection happens without the consent of users and even clearing cookies and browsing history doesn't help.The user doesn't know what level of data is being collected. Unlike cookies, users do not have the control to block the data that is being collected.

No way to reset the data once collected:
Once the data is collected it can't be cleared or reset.

Identity Revealed:
Companies can use this fingerprint information to reveal the identity of the user by correlating the data with identifying information such as email, surname etc.

Reference: https://www.w3.org/TR/fingerprinting-guidance/#privacy_threat_models

Accuracy:

Even though the information can not directly reveal the name but they can uniquely identify our behavior and track us across websites. The accuracy depends on the number of the data points that are collected. Here is an article that explains the relationship of data points and the accuracy.

Information Collected:

Information such as ***plugin details, timezone, os details, ip, language and various other settings*** are collected and it can uniquely identify the user. Because these kinds of details are not changed on a regular basis. These informations are the same for a user so the user can be identified across sites as well.

Level of Isolation:

It is not able to reveal the name or other identity but still it can uniquely identify a user. And if it is somehow mapped with other identifying information from another dataset, it can reveal other specific details as well.

Can it be prevented?

It is difficult to hide such information as these are the basic information that are shared while loading a website. Apple, Mozilla and google are planning to limit browser fingerprinting within their browser.

Mozilla

- Mozilla is working together with disconnect to tackle browser fingerprinting. Check the official article of mozilla.
- Disconnect maintains a list of companies that participate in cross site tracking and lists those that finger print users. Using this information mozilla blocks those companies from collecting the browser fingerprint data.
- They also use measurement techniques from past research and help disconnect identify new domains that track information
- Reference URL for above information.
  https://blog.mozilla.org/security/2020/01/07/firefox-72-fingerprinting/
- Reference URL for previous academic research.
  https://webtransparency.cs.princeton.edu/webcensus/
- Enhanced Tracking Protection(ETP)
  https://blog.mozilla.org/blog/2019/06/04/firefox-now-available-with-enhanced-tracking-protection-by-default/

Apple

Safari

1. Only sharing little information to the trackers so that it is difficult to single one out.
2. Doesn't add any custom tracking headers or unique identifiers like other browsers.
3. Some plugins are no longer supported and this can reduce the information.
4. For detailed information, here is the document which explains about it under 'Fingerprinting defense' heading.

Google

1. Google is proposing the idea of a privacy budget where browser allows websites to make limited API calls to collect the information.
2. Article with detailed information can be found here.

Some Companies that collect fingerprint data with code samples

This information is listed by Disconnect.

- https://github.com/disconnectme/disconnect-tracking-protection/blob/master/descriptions.md

Useful References

1. https://pixelprivacy.com/resources/browser-fingerprinting/#:~:text=Browser%20fingerprinting%20is%20a%20powerful,and%20various%20other%20active%20settings.
2. https://digiday.com/marketing/what-is-device-fingerprinting/#:~:text=Device%20fingerprinting%20can%2C%20in%20some,to%20definitively%20track%20individual%20devices.
3. https://thenextweb.com/syndication/2020/04/24/how-browser-fingerprints-identify-you-even-when-you-have-cookies-turned-off/
4. https://www.comparitech.com/blog/vpn-privacy/what-is-browser-fingerprinting-how-to-protect-yourself/
5. https://medium.com/slido-dev-blog/we-collected-500-000-browser-fingerprints-here-is-what-we-found-82c319464dc9
6. https://gizmodo.com/apple-declares-war-on-browser-fingerprinting-the-sneak-1826549108
7. https://arxiv.org/pdf/1905.01051.pdf
8. https://www.secureauth.com/blog/why-browser-fingerprinting-is-creating-challenges-for-identity-security/

How information of the same user collected from different devices and websites are merged and used to identify and track the user?

Some third party advertisers and data collection companies:

Google Analytics and Google Tag Manager

By default GA can't identify users across different domains and devices. Various ways of tracking user activities are:

Cookie:

clientid - Client id is generated for a user by google analytics, whenever a user visits a site. It is stored in the cookie as clientid. It can track users across multiple sessions. If the cookie is deleted, a new cookie id is generated and it will be identified as a new user.

Cross Platform and Devices:

userid - (GA Universal analytics). It is similar to client id but the unique id is generated using user login details. It allows companies to create a unique user id using user login details. Using this user id, GA can track users ***across different platforms and devices***. It gives power to accurately identify users.

Cross domain(subdomain) and different domains:

It is done by using a cross-domain link by passing client id in all the links that point to another domain.

```
ga('create', 'UA-XXXXX-YY', 'auto',{'allowLinker': true }); // allow source domain to pass cookie information
```

Different types of tracking done by google analytics: Email tracking(opened email, unsubscribed etc), User tracking. Email tracking is done by creating a trackable url but it can't gather as much information as pixel tracking.

Hubspot

User Tracking

1. Hubspot provides features to create forms and allows that form to be included in websites. When a user fills the form a new contact is created.
2. Hubspot also creates a unique id to identify users by using cookies on the browser.
3. Cookies are linked to the contact using email id.
4. Similar to google analytics tracking, if cookies are deleted then the user is identified as a new user.
5. Different types of tracking:
    a. User activities on the website.
    b. Email tracking: Email tracking is done using Pixel tracking. Various user activities can be tracked: opening of email, clicking of link (using link) etc.

Pixels Tracking

Properties:

1. Pixels are similar to cookies but they can't be blocked.
2. 1x1 pixel graphics are included in the email.

What is it used for ?

1. User Tracking:
   They are similar to the cookies but are more capable than cookies to track users across devices and various platforms. Cookies can only track users across multiple sessions. The difference between them is where the

data is stored. In cookies data are stored in the browser whereas for tracking pixels, send data directly to the server. That is why the users can be tracked across devices and platforms. Like cookies, user can't disable pixels.

2. Email Tracking: Tracking user behavior after the user opens the email. A trackable URL is created and hidden in the email as an image pixel. When recipients open the email, it tracks user behavior. Image pixel can track following information:
    a. Whether the user lands on the brand's website and actions taken by the user on the actual site.
    b. Os, mailbox type, screen resolution, ip address, time spent on email etc.

Next Week Study Plan:

a. Tracking Pixels for user tracking
b. how are user choices recorded and communicated (to the ad exchange, to the buyers of the ad impression)
    i. How to get information and what will be the scope?
        1. Select 2,3 advertising agencies and study all the data they collect.
        2. Study using mozilla and chrome platform.
c. What are some ways we can achieve milestone 1?
    i. Make a browser extension that scrambles user choice (the base case)
        1. Part1 : See other existing examples of scrambling user choices if there are any and share it in the next meeting for discussion.

d. ...Understanding OpenWPM
maybe consider exploring OpenWPM to understand its capabilities and functionalities + if / how we can leverage this and similar tools for b.?

e. (optional) look into brave browser and BAT token (brave attention token)

Date Jan 17, 2021

Week III

Tracking Pixel

Tracking pixel is a way of tracking users by adding a script to the website provided by third party ad agencies such as google or facebook.

1. Facebook

   Based on the available information, under facebook tracking pixel domain, there are two ways of tracking the user. One is adding a pixel code to the page and another is using a 1x1 pixel.

   How is data collected and sent to the server?

   1. *Tracking Pixel - Javascript Code*

      This is the javascript code that is added in the head section of the page on which the user's activities are being tracked. This code sends data to the facebook server on events triggered by user activities. Such as button click, page scroll etc. The information is sent using the function 'fbq' as shown in the image.

```
<!-- Facebook Pixel Code -->
<script>
!function(f,b,e,v,n,t,s)
{if(f.fbq)return;n=f.fbq=function(){n.callMethod?
n.callMethod.apply(n,arguments):n.queue.push(arguments)};
if(!f._fbq)f._fbq=n;n.push=n;n.loaded=!0;n.version='2.0';
n.queue=[];t=b.createElement(e);t.async=!0;
t.src=v;s=b.getElementsByTagName(e)[0];
s.parentNode.insertBefore(t,s)}(window, document,'script',
'https://connect.facebook.net/en_US/fbevents.js');
fbq('init', '208091314281684');
fbq('track', 'PageView');
fbq('track', 'Donate');
</script>
```

   2. *Tracking Pixel - 1x1 Image*

      This comes under no script tag and is used when the javascript is disabled. Using this we can send the data to facebook server only once when the page loads. The data is sent as the additional parameters on the src tag as shown in the image below:

```
<noscript>
 <img height="1" width="1" style="display:none" src="https://www.facebook.com/tr?id=208091314281684&ev=PageView&noscript=1"/>
 <img height="1" width="1" style="display:none" src="https://www.facebook.com/tr?id=208091314281684&ev=Donate&noscript=1"/>
</noscript>
```

Using javascript, website owners can collect lots of data on various events triggered by user activities and send it to facebook server.

3. *Matching Users*
    - Article related to advance matching: https://developers.facebook.com/docs/facebook-pixel/advanced/advanced-matching
    - Article related to automatic advance matching https://www.facebook.com/business/help/1993001664341800
4. *References*

    https://developers.facebook.com/docs/facebook-pixel/advanced/

What information of the online users can be collected?

Marketers/Website owners can send any information about you which they collect

- Email, phone etc. These information are willingly shared by the users while filling forms.
- IP address, device details, user activities(search, links clicked, area of interest, button clicks, etc). These informations are sent without the consent of the user as they are collected in the background.

How does Facebook match users across websites?

- Information provided by marketers such as email, phone numbers can be used to match the users:

    To use advanced matching, format the visitor's data as a JSON object and include it in your pixel's base code `fbq('init')` function call as a third parameter. For example, if your pixel ID was `283859598862258`, you could do this:

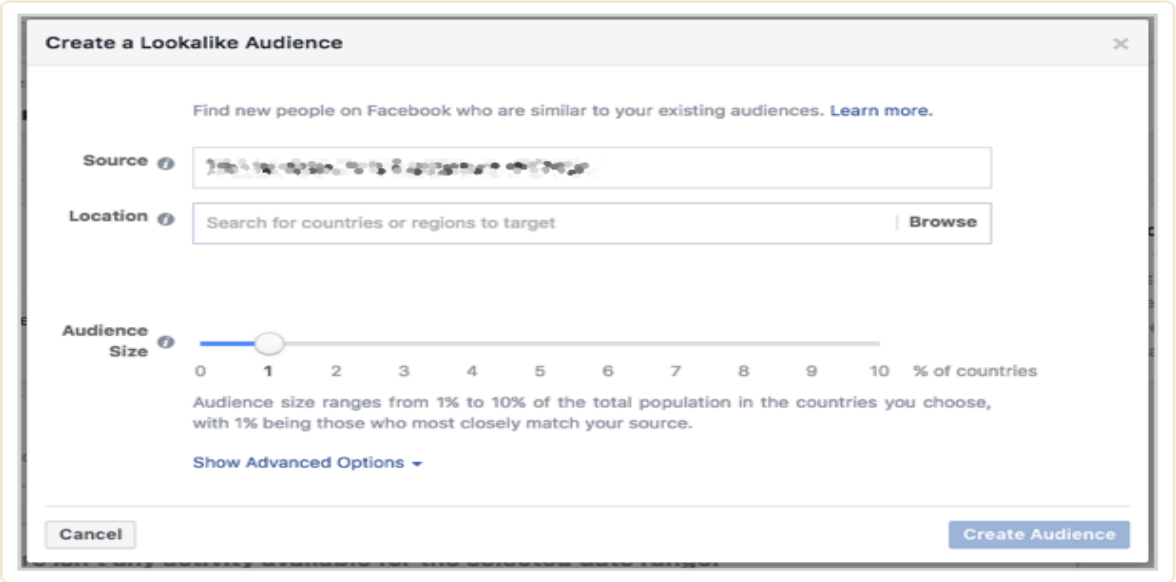
    ```
    fbq('init', '{facebook-pixel-id}', {
      em: '{user-email}',         // Values will be hashed
      fn: '{user-first-name}',    // automatically by the pixel
      ln: '{user-last-name}'      // using SHA-256
      ...
    });
    ```

- Segmenting user groups by location, age etc
- Facebook allows marketers to upload a customer list to match with facebook user profile. Reference article (https://www.facebook.com/business/help/170456843145568?id=2469097953376494)
- Reference: https://www.eff.org/deeplinks/2019/01/guided-tour-data-facebook-uses-target-ads

What does Facebook use it for?

- Retargeting:

- Segmenting Users:Showing ads to the relevant users based on what they visited in the marketers website. The users can be grouped based on :
  - which pages they visited
  - which product was clicked
  - time segment(visited within 7 days)
  - age, gender, location etc.
  - interest, behavior etc
- Creating lookalike customers: Marketers upload email addresses of their clients and facebook uses their personal information to create lookalike customers.



2. Google

How is data collected and sent to the server?

1. *Tracking Code*

   This is a javascript code which is added in the head section of the page in which users' activity will be tracked. The data is sent to the google server when events are triggered due to various user activities(similar to facebook pixel).

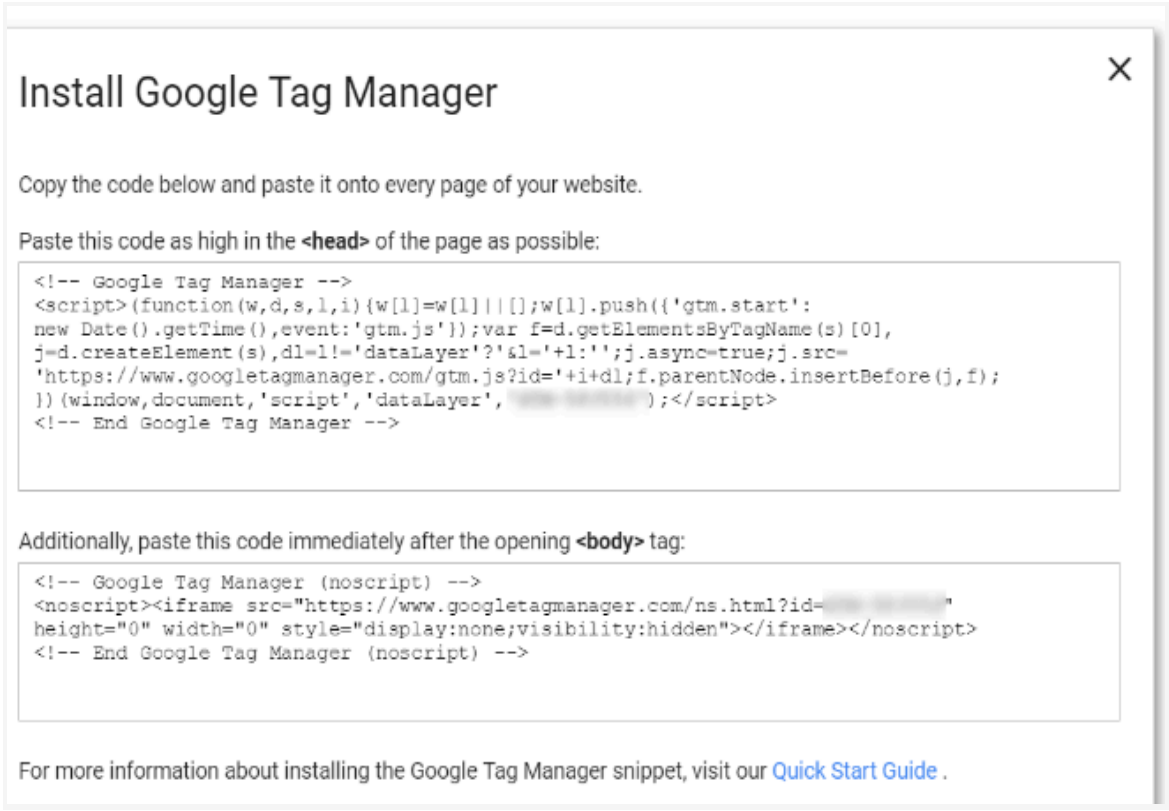


image reference: https://www.analyticsmania.com/post/google-tag-manager-noscript/

What information of the online users can be collected?

- User activities such as button clicks, link click, downloads, scrolls etc.
- IP address is not collected. More information about data collected by GTM can be found here.

How does Facebook match users across websites?

Refer to Week II (Google Analytics and Google Tag Manager)

What does Google use it for?

- Remarketing

Privacy Policies

What are the privacy policies that help to protect user privacy?

- GDPR.
- California Privacy Right Act.
- The Lei Geral de Proteção de Dados (LGPD)

How does facebook comply with the law?

Off Facebook Activity: They allow users to see what advertising agencies are sending data to facebook, what data is being collected etc. by providing a feature 'off-facebook' activity. Detailed explanation of this feature is here .

How does Google comply with the law?

Granular data retention control: Google has introduced a 'Granular data retention control' feature which limits the retention time of data that is collected using cookies and events(google tag manager). All the collected data from users' will be deleted if it is above retention period. Marketers can set the time line for data retention. More details about this feature is here .

Why are you seeing an ad: User can see details about why a particular ad is being served to him/her. More details can be found here

Compliance with data protection laws: Detailed information of how google complies with data protection laws is here

How are advertising companies finding new ways of collecting data?

- Facebook: Conversion API (from website server to facebook server using API).
- Google: Server side tagging (Relevant article is here)

References:

1. GDPR and Facebook Marketers: https://leadsbridge.com/blog/guides/gdpr-and-facebook-all-you-need-to-know-to-keep-advertising-safely/
2. Why are you seeing an ad: https://support.google.com/accounts/answer/1634057?hl=en
3. Granular Data Retention: https://datadrivendesignnashville.medium.com/google-analytics-releases-granular-data-retention-controls-14cc1398c66c
4. Googles' compliance with data protection laws: https://privacy.google.com/businesses/compliance/
5. What data GTM Collects: https://support.google.com/tagmanager/answer/9323295?hl=en#:~:text=Data%20collected%20by%20Google%20Tag,associated%20with%20a%20particular%20individual.

What are some ways we can achieve milestone 1?

Make a browser extension that scrambles user choice (the base case).

1. Part1 : See other existing examples of scrambling user choices if there are any and share it in the next meeting for discussion.
    a. Obscure user data: Track Me Not
        i. Github: https://github.com/vtoubiana/TrackMeNot
    b. Obscure use data: AdNauseam
        i. Github: https://github.com/dhowe/AdNauseam

Papers

1. Engineering Privacy and Protest: a Case Study of AdNauseam

Tools: Monitor how your data is being tracked

- Visualize all http requests made by web application: https://github.com/mozilla/lightbeam-we
- Block trackers: https://github.com/disconnectme/disconnect

References:

- Privacy preserving tools: https://www.privacytoolbox.gppi.net/obscure-me/
- MIT review: https://rednoise.org/AdNauseamVsGoogle.pdf

Next Week Study Plan:

- **How to use ad nauseam to integrate with ML algorithms**
    1. Classes and interfaces.
    2. Hook ML algorithm.
    3. Architecture diagram and technical document.
- **Study Code in Detail**
    - https://github.com/dhowe/AdNauseam
- **Papers:**
    - Trackmenot: Resisting Surveillance in Web Search.
    -