Web Content
CS249i
Modern Websites

Third Party Resources

• Modern websites rely on many different types of *third-party resources* to provide services to keep their websites functional

• Third party resources are ones served by external parties

  • If you are on cnn.com, any resource served from a domain that is NOT cnn.com (e.g., doubleclick.com, google-analytics.com)

• These resources could be anything from static images to JavaScript libraries to analytics, advertising, ...
Trump escalates January 6 cover-up

The former President is trying to keep the House select committee probing January 6 from seeing a list of documents as he ramps up his political comeback.

Brian Steil’s ominous prediction: Imagine it’s 2022 and...

January 6 committee is losing patience with Trump’s former chief of staff Mark Meadows as it seeks his testimony.

Washington Post report rebuts the January 6 alt-reality that Tucker Carlson promotes.

Biden says US ‘continuing to suffer’ from Trump’s decision to pull out of Iran nuclear deal.

Astros top Braves 9-5 in World Series Game 5

- Trivia: Can you name the only player to play in all 3 cities that the Braves have called home?
- Analysis: The Braves may win the World Series. But they’re striking out with some fans.

Students are fed up with raging adults at school board meetings

- A Texas lawmaker is investigating 850 books on race and gender that could cause ‘discomfort’ to students.
- Opinion: When parents scream at school board meetings, how can I teach their children?

Southwest launches investigation into pilot reportedly using anti-Biden phrase on flight

- Reporter reveals what Lindsey Graham said during January 6 riot.
- White House press secretary tests positive for Covid, last saw Elderly.

BREAKING: Japan’s Fumio Kishida defies expectations as ruling party keeps majority.

Aurora borealis puts on a gorgeous show.

- ‘Step up or step out’: Lawmaker calls out attorney general.
- Police investigating desecration of Torah scroll at fraternity.

COP26 climate talks off to an ominous start after weak G20 leaders’ meeting.

Video shows passengers fleeing knife attack on train.
"Twilio’s $3.2B Segment acquisition is about helping developers build data-fueled apps"
Many websites rely on user analytics to improve their services

For example, Google Analytics, appears on an estimated 70% of the top websites

As an analytics user, you can see where your clients are connecting from, you can see how long they spent on the page, what devices they’re connecting from, and a ton of other interesting details

These are typically scoped to a single request, but in recent years, companies have been expanding the scope of what they know about users...
Draw insights from data you can trust

Click maps and scroll maps give a crystal-clear picture of user engagement without the confounding factors of screen resolution, dynamic page content, pop-ups, or confusing "sitoclines."

Inspect specific user segments

Answer questions like "What do visitors from my ad campaign click on first?", "How do returning customers interact with our navigation?", or "Are users clicking on the latest product update?"

Complete data for a complete understanding

Easy implementation and intuitive UI empowers your whole team to jump in right away. *Tagless autocapture* means you never have to worry about data points slipping through the cracks.

In App Analytics
(From FullStory)
Web Tracking
Cookies and Code

- Major companies typically use *cookies* to offer extended functionality for websites (e.g., keeping you logged in, keeping certain settings stored in your browser, etc.)

```
POST /login
```
Web Tracking
Cookies and Code

• Major companies typically use *cookies* to offer extended functionality for websites (e.g., keeping you logged in, keeping certain settings stored in your browser, etc.)
Web Tracking
Cookies and Code

• Major companies typically use cookies to offer extended functionality for websites (e.g., keeping you logged in, keeping certain settings stored in your browser, etc.)

• Once a cookie is set, the browser attaches a cookie to every subsequent request sent out for that particular domain
  • Cookies are by default scoped to the first-party domain that set the cookie
  • No other domains can read the cookie value!

• …then how does web tracking work?
Web Tracking
Cookies and Code

(CNN)
GET / HTTP/3

GET /facebook-like.js HTTP/3
Cookie: User=Deepak, Referer=cnn.com
(Facebook)

- With this request, companies can link your cookie to your browsing data (e.g., through Referer header, Host headers, Origin, or just JavaScript)
Web Tracking
Cookies and Code

• What exactly is sent in the referer?

HTTP Request:
GET https://site-two.example/mars.jpg
Referer https://site-one.example/stuff/detail?tag=red

https://site-one.example/stuff/detail?tag=red
Referer-Policy: no-referrer-when-downgrade

https://site-two.example/mars.jpg
### Referer sent in an request, depending on the Referrer-Policy and the request

<table>
<thead>
<tr>
<th>Policy:</th>
<th>Referer: No data</th>
<th>Referer: Origin only</th>
<th>Referer: Full URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>no-referrer</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>origin</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unsafe-url</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>strict-origin</td>
<td>request from HTTPS to HTTP</td>
<td>request from HTTPS to HTTPS</td>
<td>request from HTTPS to HTTPS or HTTP to HTTP</td>
</tr>
<tr>
<td>no-referrer-when-downgrade</td>
<td>request from HTTPS to HTTP</td>
<td></td>
<td>request from HTTPS to HTTPS or HTTP to HTTP</td>
</tr>
<tr>
<td>origin-when-cross-origin</td>
<td>cross-origin request</td>
<td></td>
<td>same-origin request</td>
</tr>
<tr>
<td>same-origin</td>
<td>cross-origin request</td>
<td></td>
<td>same-origin request</td>
</tr>
</tbody>
</table>

#### Policies that don’t take the context of the request into account.
- Policies focused on whether the request is made to a destination that is less secure than where the request came from. These help reach security goals.
- Policies focused on whether the request is cross- or same-origin. These help reach privacy goals.

#### Policy focused on the request security AND whether it is cross- or same-origin.
- It helps reach privacy and security goals.
Web Tracking
Browser Fingerprinting

- Websites can also fingerprint you effectively with browser fingerprinting, which is a technique that leverages all your settings to identify you, and stores this in a cookie on your browser

- https://amiunique.org/

- So long as JavaScript can run (by third-parties), you run the risk of being “followed” on the web
My browser fingerprint

Are you unique?

Yes! You are unique among the 2382170 fingerprints in our entire dataset.

The following information reveals your OS, browser, browser version as well as your timezone and preferred language. Moreover, we show the proportion of users sharing the same elements.

OS: macOS
Browser: Chrome
Language: en
Timezone: UTC-08:00

HTTP headers attributes

- User-Agent: 6.15%
  Mozilla/5.0 (Macintosh; Intel Mac OS X 10.15;7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/71.0.3578.98 Safari/537.36

- Accept-Language: 1.92%
  en, de, fr, es, pt, it, zh, zh-Hans, zh-Hant, zh-CHS, zh-CHT, zh-HK, zh-MO, zh-TW, zh-CN, zh-HK, zh-MO, zh-TW, zh-CN

- Content-Language: 16.51%
  en

- Upgrade-Insecure-Requests: 17.01%
  1

- Permissions: 5.34%
  - Not supported
    - ambient-light-sensor
    - camera
    - clipboard-read
    - clipboard-write
    - geolocation
    - background-sync
    - magnetometer
    - microphone
    - mtd
    - notifications
    - payment-handler
    - persistent-storage
    - push
    - sync

- WebGL Vendor: 1.83%
  - Google Inc. (Apple)

- WebGL Renderer: 0.01%
  - ANGLE (Apple, ANGLE Metal Render Engine: Apple M1, Unspecified Version)

- WebGL Data: 0.13%
  - 35 different extensions
  - 25 different general parameters analyzed
  - 36 different shaders precision analyzed
Web Tracking
Prevalence of Major Companies

- Major companies have large presences on the web, and as a result, can see the majority of websites that you visit

- Google appears on 82.2% of the Top 1M (by AS), because of analytics and advertising services

- Facebook appears on 34.1%, to enable social sharing + tracking

<table>
<thead>
<tr>
<th>Company</th>
<th>Prevalence on Top 1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>82.2%</td>
</tr>
<tr>
<td>Facebook</td>
<td>34.1%</td>
</tr>
<tr>
<td>Amazon</td>
<td>32.6%</td>
</tr>
<tr>
<td>Cloudflare</td>
<td>30.7%</td>
</tr>
<tr>
<td>Akamai</td>
<td>20.3%</td>
</tr>
<tr>
<td>MaxCDN</td>
<td>19.0%</td>
</tr>
<tr>
<td>Edgecast</td>
<td>17.9%</td>
</tr>
<tr>
<td>Fastly</td>
<td>15.5%</td>
</tr>
<tr>
<td>SoftLayer</td>
<td>11.8%</td>
</tr>
<tr>
<td>Twitter</td>
<td>11.2%</td>
</tr>
</tbody>
</table>
Web Tracking

Cookie Syncing

• Even if a company is not available on every website, companies often times *share* cookie information

• “Cookie Synchronization: Everything You Always Wanted to know but were afraid to ask” – WebConf 2019

• Core idea is simple: If you have a collaboration agreement with another third-party, you *redirect* requests to them upon receiving requests
Web Tracking

Cookie Syncing

GET tracker.com/pixel.jpg

Response, Set-Cookie: User=user123
Web Tracking

Cookie Syncing

GET advertiser.com/pixel.jpg

Response, Set-Cookie: User=userABC
Web Tracking
Cookie Syncing

GET tracker.com/pixel.jpg, cookie=user123
GET tracker.com/pixel.jpg, cookie=user123

REDIRECT, advertiser.com?syncID=user123&publisher=nytimes.com
Web Tracking

Cookie Syncing

GET tracker.com/pixel.jpg, cookie=user123

REDIRECT, advertiser.com?syncID=user123&publisher=nytimes.com

GET syncID=user123, cookie=userABC
Web Tracking
Cookie Syncing

GET tracker.com/pixel.jpg, cookie=user123

REDIRECT, advertiser.com?syncID=user123&publisher=nytimes.com

GET syncID=user123, cookie=userABC

- Third-parties with cookie syncing are enabled on 78% of modern websites
Web Tracking
Cookie Ghostwriting

- Not all first-party cookies *should* be treated the same!
Web Tracking

Cookie Ghostwriting

• Not all first-party cookies should be treated the same!

GET tracker.com/script.js

tracker.com
Web Tracking

Cookie Ghostwriting

• Not all first-party cookies should be treated the same!

GET tracker.com/script.js

document.cookie = "user=userABC"
Web Tracking

Cookie Ghostwriting

- 42% of identifier cookies are *ghostwritten* in modern websites

```
GET tracker.com/script.js

script.js

document.cookie = "user=userABC"
```

tracker.com

advertiser.com
Online Advertising Ecosystem
Online Advertising
The Best Thing Since Sliced Bread!
Available for $4.99 at your local Costco.

• Companies typically track you around the web to build profiles for targeted advertising

• The more targeted your advertising, the more revenue you can make from advertisers who are potentially willing to give you more money to sell the ad spot

• Useful for advertisers to know if people with your browsing habits, your properties, your whatever are browsing on the web
Online Advertising
The Many Internet Players in Advertising

Publisher Content Server -> Publisher Ad Server -> Supply side platform -> Demand Side Platform

Publisher Ad Server
- OpenX, PubMatic, AppNexus, ShareThrough, 33Across, Google Ad Manager

Supply side platform
- Data Management Platform

Ad Exchange
- Google AdX, Magnite, OpenX, PubMatic, ConnectAd, Amazon Publisher Services
- Affirm, Criteo, Demandbase, Google Marketing Platform, Quantcast

Agency Ad Server
- Google Ad Manager

Brand
Online Advertising
The Many Internet Players in Advertising

Publisher Content Server
Publisher Ad Server
Supply Side Platform
Ad Exchange

Agency Ad Server

OpenX, PubMatic, AppNexus, ShareThrough, 33Across, Google Ad Manager

Google AdX, Magnite, OpenX, PubMatic, ConnectAd, Amazon Publisher Services

Demand Side Platform
Affirm, Criteo, Demandbase, Google Marketing Platform, Quantcast
Online Advertising

Publishers

- Publishers (e.g., nytimes.com, cnn.com, other websites) often have advertising space that they are hoping to make revenue off of

- In some cases, publishers have explicit agreements with specific companies and can sell their space directly that way too
Online Advertising
Supply Side Platforms

• If a publisher wants to place the ad spot on the open advertising market, they typically go through an intermediary called a Supply Side Platform (SSP)

  • Examples: Pubmatic, Rubicon Project, Verizon Media, etc.

  • This aggregates information about the client (through a DMP) and participates in ad exchange

• Multiple Types of Data in DMP:
  • "First Party", e.g., CRM Data. This can include data from customer behaviors, actions, purchases or interest
  • "Second party" — statistics related to cookie pools on external publications and platforms. The
  • "Third party" — sourced from external providers and aggregated across websites. Businesses sell third-party data
Online Advertising
Demand Side Platforms

• On the other end of the pipeline, you have advertisers

• There are analogous entities called demand side platforms, which participate in Real-Time Bidding, which is a real-time auction for ad space (examples: Google DoubleClick, QuantCast, Criteo, Adform)
  • Typically happens in < 100ms
Online Advertising

Ad Exchanges

- Advertising exchanges receive spots from supply side, and facilitate real time bidding from the demand side based on properties of the ad spot.
- Examples: Google DoubleClick, Facebook Exchange, PubMatic, Microsoft Advertising
Online Advertising
Bid Requests

```
"site": {
  "id": "1234",
  "name": "Example Site",
  "domain": "examplesitedomain.com",
  "mobile": 1,
  "amp": 0,
  "pub": {
    "id": "9876",
    "name": "Example Publisher, Inc.",
    "domain": "examplepubdomain.com"
  }
},

"user": {
  "id": "e8af45c77890045d0ec100ac88443baff57c",
  "consent": "ihdknhkhqk8y",
  "buyeruid": "fcd4262456238256034abcde226d39a5892",
  "yob": 1990,
  "gender": "m",
  "ext": {
    "consented_providers_settings": {
      "consented_providers": [
        1,
        52,
        45,
        23
      ]
    }
  }
},

"device": {
  "type": 4,
  "ifa": "8846d6fa1000bceaaaf322908dfeb221",
  "ip": "1.2.3.4",
  "ua": "Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.8; en-US; rv:1.9.2.16) Gecko/20110319 Firefox/3.6.16",
  "make": "Apple",
  "model": "iPhone",
  "hwv": "6s",
  "os": "13",
  "osv": "11.4.1",
  "mccmnc": "310-005",
  "geo": {
    "country": "US",
    "region": "CA",
    "city": "San Francisco"
  }
}
```
Online Advertising
Bid Response

```json
{
    "id": "d7d3e107-987h",
    "cur": "usd",
    "ext": {
        "protocol": "6.0"
    },
    "seatbid": [
        {
            "id": "qwerty-998765",
            "item": "asdf-7000",
            "price": 1.45,
            "cid": "app-mraid-campaign-3442",
            "url": "https://adserver.com/winnotice?impid=102&winprice=${AUCTION_PRICE}"
        }
    ],
    "ext": {
        "agencies_id": "agency_123",
        "advertiser_name": "example advertiser"
    },
    "media": {
        "ad": {
            "id": "creative_id_1234",
            "domain": {
                "example.com",
                "example.io"
            },
            "cat": {
                "cat_1",
                "cat_2"
            },
            "lang": "en",
            "attr": {
                3,
                7
            }
        }
    }
}
```

https://protocol.bidswitch.com/rtb/response-examples.html
Online Advertising
Bidding for Ad Spots

• Real-time bidding is an auction process that is kicked off when a publisher tells an advertising network that they have an open ad-spot with certain properties

• Two most widely used methods of auctioning
  • Waterfall bidding
  • Header bidding
Online Advertising

Waterfall Bidding

- Publishers would pre-define a hierarchy of advertising networks that they wanted to ask in order (e.g., in a waterfall) about any given advertising spot
- Publishers would then set a floor bid rate that they needed for the ad spot
  - The first network to fulfill the floor would win the spot, but floor price goes down with lower priority
- Problems:
  - Slow (serial computation)
  - Anti-competitive!
    - Google had both an SSP and a DSP, which often meant they got first pick at ad spots

Image stolen from: https://www.youtube.com/watch?v=FqESUjO1Y
Online Advertising
Header Bidding

• Every DSP is offered the auction at the same time, and DSPs are incentivized to provide their true value for the advertising spot (theoretically)

• This typically happens in under 100 ms

• Two options:
  • Client-side header bidding (happens in JavaScript), potentially makes the page slower, but have finer grained access to cookies
  • Server-side header bidding (happens in the SSP), can be faster, but requires cookie syncing, could make things slower
2024 — The Year of Linux on the Desktop

2024 — The Year of the End of Third Party Cookies?

• Firefox:
  • Third-Party Cookies from known trackers are dropped
  • Third-party cookies use separate cookie jar per site, so they can't be used to track users across sites

• Safari: Blocks third-party cookies

• IE: blocks some third-party cookies based on baked-in blacklist

• Edge does not block third-party cookies by default

• Chrome announced that they will drop support for third party cookies by the end of 2024
Google Topics

Topics are selected from a taxonomy consisting of hierarchical categories such as /Arts & Entertainment/Music & Audio/Soul & R&B and /Business & Industrial/Agriculture & Forestry.

The (maximum) three topics returned for a user are chosen at random from the top five for the past three epochs (with a 5% chance of getting a random topic).
## Google Topics

<table>
<thead>
<tr>
<th>Site</th>
<th>Topics</th>
<th>API callers on site</th>
</tr>
</thead>
<tbody>
<tr>
<td>yoga.example</td>
<td>Fitness</td>
<td>adtech1.example adtech2.example</td>
</tr>
<tr>
<td>knitting.example</td>
<td>Crafts</td>
<td>adtech1.example</td>
</tr>
<tr>
<td>hiking-holiday.example</td>
<td>Fitness, Travel &amp; Transportation</td>
<td>adtech2.example</td>
</tr>
<tr>
<td>diy-clothing.example</td>
<td>Crafts, Fashion &amp; Style</td>
<td>[none]</td>
</tr>
</tbody>
</table>

At the end of the epoch (currently one week) the Topics API generates the browser's top topics for the week.

- adtech1.example is now eligible to receive the "Fitness" and "Crafts" topics, since it observed them on yoga.example and also on knitting.example.
- adtech1.example is not eligible to receive the "Travel & Transportation" topic for this user as it is not present on any sites the user visited recently that are associated with that topic.
- adtech2.example has seen the "Fitness" and "Travel & Transportation" topics, but has not seen the "Crafts" topic.

The user visited diy-clothing.example, which has the "Fashion & Style" topic, but there were no calls to the Topics API on that site. At this point, this means the "Fashion & Style" topic would not be returned by the API for any caller.

In week two, the user visits another site:

<table>
<thead>
<tr>
<th>Site</th>
<th>Topics</th>
<th>API callers on site</th>
</tr>
</thead>
<tbody>
<tr>
<td>sewing.example</td>
<td>Crafts</td>
<td>adtech2.example</td>
</tr>
</tbody>
</table>

In addition, code from adtech2.example is added to diy-clothing.example:

<table>
<thead>
<tr>
<th>Site</th>
<th>Topics</th>
<th>API callers on site</th>
</tr>
</thead>
<tbody>
<tr>
<td>diy-clothing.example</td>
<td>Crafts, Fashion &amp; Style</td>
<td>adtech2.example</td>
</tr>
</tbody>
</table>

As well as "Fitness" and "Travel & Transportation" from week 1, this means that adtech2.example will now be able to receive the "Crafts" and "Fashion & Style" topic — but not until the following epoch, week 3. This ensures that third parties can't learn more about a user's past (in this case, an interest in fashion) than they could with cookies.
Studying the Web
Headless Chrome

- Headless Chromium allows running Chromium in a headless/server environment.

- “Expected use cases include loading web pages, extracting metadata (e.g., the DOM) and generating bitmaps from page contents — using all the modern web platform features provided by Chromium and Blink.”

- Let’s you load and manipulate websites DOM’s after Chromium has loaded them using Node.js
Top Lists

- Most studies about the web use “Top Lists” — one million domain long lists of who thinks the most popular sites on the web are.

- Most popular list was “Alexa Top Million”, which went away in 2022!

- Where does this data come from??! 🤔

Example Top Sites

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>google.com</td>
</tr>
<tr>
<td>2</td>
<td>amazonaws.com</td>
</tr>
<tr>
<td>3</td>
<td>facebook.com</td>
</tr>
<tr>
<td>4</td>
<td>a-msedge.net</td>
</tr>
<tr>
<td>5</td>
<td>microsoft.com</td>
</tr>
<tr>
<td>6</td>
<td>apple.com</td>
</tr>
<tr>
<td>7</td>
<td>googleapis.com</td>
</tr>
<tr>
<td>8</td>
<td>youtube.com</td>
</tr>
<tr>
<td>9</td>
<td>akamaiedge.net</td>
</tr>
<tr>
<td>10</td>
<td>akamai.net</td>
</tr>
<tr>
<td>11</td>
<td>azure.com</td>
</tr>
<tr>
<td>12</td>
<td>twitter.com</td>
</tr>
<tr>
<td>13</td>
<td>instagram.com</td>
</tr>
<tr>
<td>14</td>
<td>googlevideo.com</td>
</tr>
</tbody>
</table>
Top Lists — Up Until Recently

- Alexa: approximates site popularity by tracking the browsing behavior of several million users through partnerships with a reported 25K browser extensions as well as through websites that install Alexa Certify code.

- Cisco Umbrella 1 Million is a list of the most popular names (e.g., .com is ranked #1) looked up using Cisco Umbrella’s DNS service.

- Majestic Million [20] is a list of popular websites maintained by Majestic SEO, which is calculated based on the number of backlinks that each site has.

- Secrank list is a researcher-built list that aggregates several features of DNS data from a major resolver in China.

- Tranco aggregates data from the Alexa, Umbrella, and Majestic lists over a 30 day window to form a ranking that is more temporally stable and resistant to adversarial manipulation.
The Tale of Website Popularity Rankings: An Extensive Analysis (2019)
Tajalizadehkhoob et al.
Top Lists

- In 2022, Google Chrome agreed to start publishing a list of top sites as seen by anonymous telemetry.
- Gathered from Chrome Users who have explicitly enabled sharing URLs with Google and who have usage statistic reporting enabled.
- Part of Chrome User Experience Report (CrUX) dataset.
Top Lists

- In 2022, Google Chrome agreed to start publishing a list of top sites
  - Gathered from Chrome Users who have explicitly enabled sharing URLs with Google and who have usage statistic reporting enabled.
  - Part of Chrome User Experience Report (CrUX) dataset
  - Can be downloaded from https://github.com/zakird/crux-top-lists

The Chrome User Experience Report and Cloudflare Radar rankings have been integrated into the default Tranco list, starting from the daily updated list of August 1, 2023. The Alexa ranking has been removed from the default Tranco list, as it is no longer available.

(a) Jaccard Index
Shape of the Web