TEU00311

What is the Internet doing to me? (witidtm)

Stephen Farrell
stephen.farrell@cs.tcd.ie

https://github.com/sftcd/witidtm
https://down.dsg.cs.tcd.ie/witidtm

Online Advertising (what I want you to ponder...)

- What's your attitude to online advertising?
- What do you know about how it works?
  - What do you want to know?
- What concerns do you have about online ads?
- Are you ok with being the product when using "free" services?
  - Always or just sometimes?
  - What would you do to avoid being the product?

Overview

- A bit of IP and cookie background
- A very quick overview of web ads
- Real-Time Bidding for your eyeballs
- (Your) conclusions?

Proxying to hide client IP address

- One of the most basic ways advertisers can track you is to record your client IP address, which may or may not be relatively stable
- There are a number of proxy-based technolgies being developed that could hide the client's IP address from (web) servers
  - masque, odoh, ohttp, ...
  - These are a bit like lightweight VPNs
- Apple have deployed their "private relay" service for paid subscribers
- Seems like a reasonable trend, but some caution warranted (who operates proxies and for what reasons?)
  - Good overview: https://blog.apnic.net/2023/03/23/hiding-behind-masques/

Cookie Resources

- Fairly simple overview
  - https://www.cloudflare.com/learning/privacy/what-are-cookies/
- Wikipedia: lots of (probably too much) detail
  - https://en.wikipedia.org/wiki/HTTP_cookie
- Dabrowski, Adrian, et al. "Measuring Cookies and Web Privacy in a Post-GDPR World." International Conference on Passive and Active Network Measurement. Springer, Cham, 2019.
  - https://eprints.cs.univie.ac.at/6632/1/201903%20-%20ADabrowski%20-%20Measuring%20Cookies.pdf
- Gotze, Matthias, et al. "Measuring web cookies in governmental websites." Proceedings of the 14th ACM Web Science Conference 2022. 2022.
  - https://dl.acm.org/doi/fullHtml/10.1145/3501247.3531545

Cookies

- When your browser/app contacts a web site, the HTTP response may attempt to “set-cookie”
- If cookies are turned on (the default) then your browser/app will store the accompanying name/value pair in some long term storage (e.g. disk) for the amount of time requested by the web server
- When your browser re-visits (another URL at) that same web site, it will send the cookie name/value pair in the HTTP request
- That allows you to login and not have to keep presenting your password
- That also allows horrendous tracking that’s a large part of the web advertising model

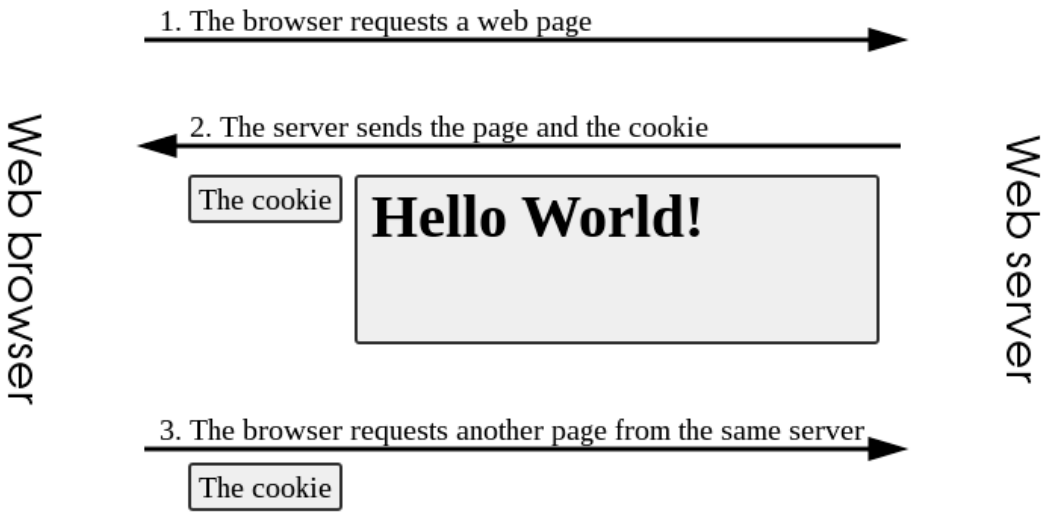


Image from: https://en.wikipedia.org/wiki/HTTP_cookie

3rd Party Cookies

- 1st party: set by the site you're "visiting" (what appears on the URL bar)
- 3rd party: set by other sites from which resources (e.g. images) are retrieved while rendering the web page
- 1st party can load a resource (e.g. Javascript that'll eventually lead to an ad being rendered) that includes a URL parameter that identifies user in some way
- 3rd party resource can also do all this again before/while being rendered

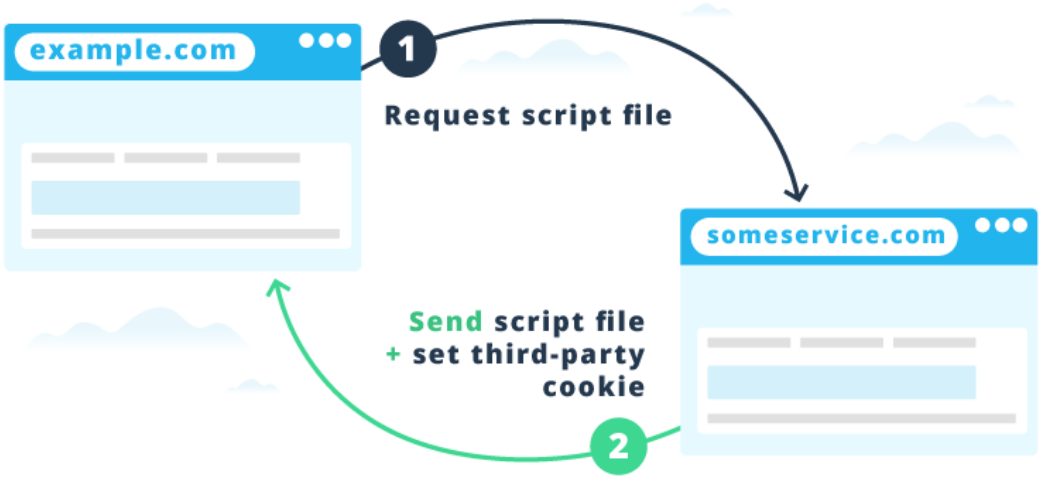


Image from: https://cookie-script.com/all-you-need-to-know-about-third-party-cookies.html

Browser Security Model and CNAME Tracking

- Browser security model requires only sending cookies back to where they've come from (Same Origin Policy aka SOP)
- Browsers are now implementing more controls over 3rd party cookies (blocking/protection)
- Trackers can instead use DNS CNAMEs to achieve a similar effect
  - Dimova, et al, "The CNAME of the Game: Large-scale Analysis of DNS-based Tracking Evasion" PETS 2021
    - https://arxiv.org/abs/2102.09301
    - Image to the right from the paper
- https://www.theregister.com/2021/02/24/dns_cname_tracking/
- Mitigation in brave: check for CNAME and if found treat requests as if they were being sent to x.tracker.com and not track.example.com
  - https://brave.com/privacy-updates/6-cname-trickery/

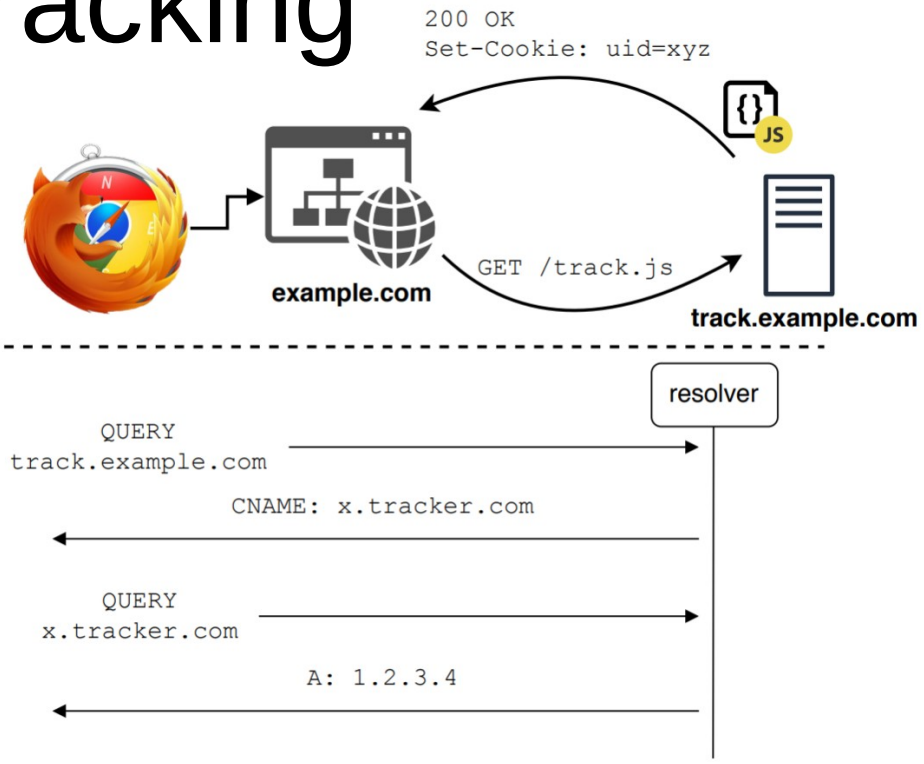


Fig. 1. Overview of CNAME-based tracking.

Facebook advertising...

- Chouaki, Salim, et al. "Exploring the Online Micro-targeting Practices of Small, Medium, and Large Businesses." Proceedings of the ACM on Human-Computer Interaction 6.CSCW2 (2022): 1-23.
  - https://arxiv.org/pdf/2207.09286.pdf
- Points:
  - Good overview of FB ads in first few pages
  - Lots of “micro targeting” happens, perhaps increasingly done by FB now vs. via advertiser specification
  - FB pixel on other web sites for tracking: 81% of small businesses seen have done that; 69% or larger businesses

Two More Documents

- Estrada-Jiménez, José, et al. "Online advertising: Analysis of privacy threats and protection approaches." Computer Communications 100 (2017): 32-51.
  - https://upcommons.upc.edu/bitstream/handle/2117/99742/Online%2Badvertising%2Bprivacy%2Bthreats%2Band%2Bsolutions.pdf

  Some of the slides here are based on that (tables or diagrams without a reference are from there)

  That's based on work done in or before 2016
- UK Information Commissioner's Office report from June 2019
  - https://ico.org.uk/media/about-the-ico/documents/2615156/adtech-real-time-bidding-report-201906-dl191220.pdf

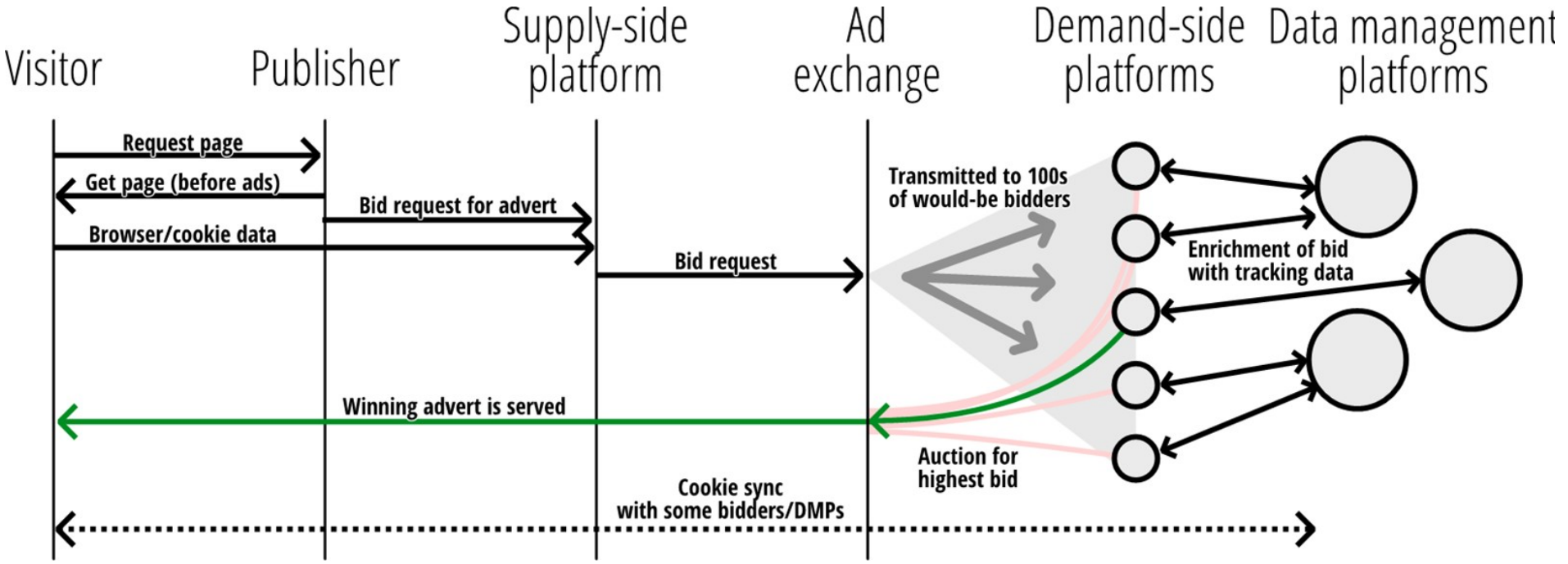
Real-Time Bidding (RTB) ...as presented by fans

- “When the page starts loading, an ad request is generated by the header bidding wrapper and sent to ad exchange(s), ad networks, ad exchanges, and ad networks. The bid requests includes information such as page URL, location, age, gender, etc., so advertisers know whether the user is relevant.”
  - https://headerbidding.co/real-time-bidding/
- “Let's say Silk is a UK-based beauty brand that just launched a new brow line and is running a campaign. They set up their campaign on a Demand-Side Platform (DSP) and are targeting users who regularly shop for makeup products, are located in the Manchester area, and are between 18 to 30 years of age. The brand also wants its ads to only show on sites related to beauty and lifestyle.”
  - https://blog.hubspot.com/marketing/real-time-bidding

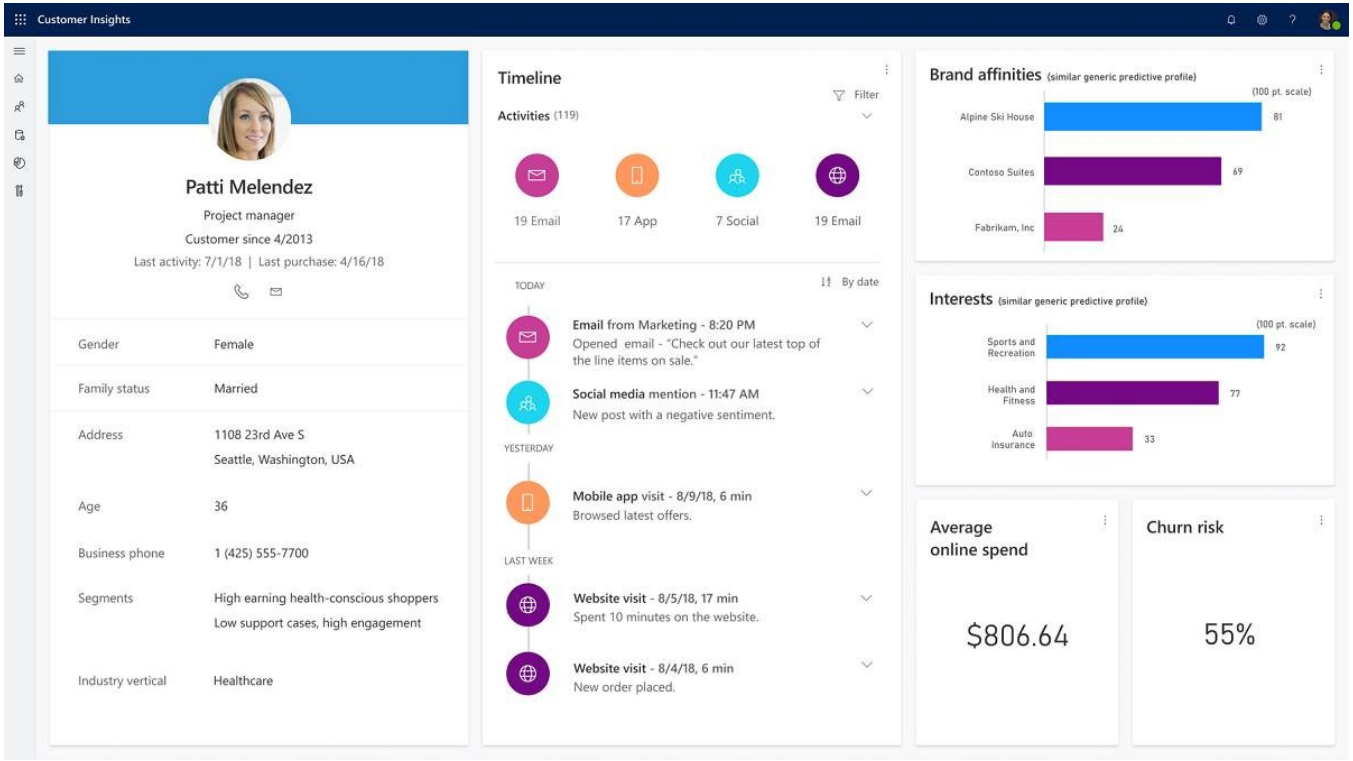
Real-Time Bidding (RTB) ...as presented by detractors

- “The online advertising's Real-Time Bidding (RTB)is the biggest data breach ever recorded. It tracks and shares what people view online and their real-world location with countless companies. This happens 178 Trillion times every year in U.S. & Europe.”
  - https://www.iccl.ie/rtb/
- “We show, first, that the GDPR requires prior consent of the internet user for RTB, as other legal bases are not appropriate. Second, we show that it is difficult—and perhaps impossible—for website publishers and RTB companies to meet the GDPR’s transparency requirements.”
  - Veale et al, “Adtech and real-time bidding under European data protection law." German Law Journal 23.2 (2022): 226-256.
  - https://www.cambridge.org/core/journals/german-law-journal/article/adtech-and-realtime-bidding-under-european-data-protection-law/017F027B4E78EBCAE1DCBC1E12B93B9D

RTB Diagram (from Veale et al)



What I imagine advertisers want...



https://www.theregister.co.uk/2019/09/23/microsofts_connected_store_dynamics_365_announcement_connects_online_and_physical_retail/
From an el Reg article about a Microsoft product to support retailers, so not quite the same, but a nice/ickky illustration.
And it's not clear to me how that'd scale (but maybe they don't care?).

Some Actors

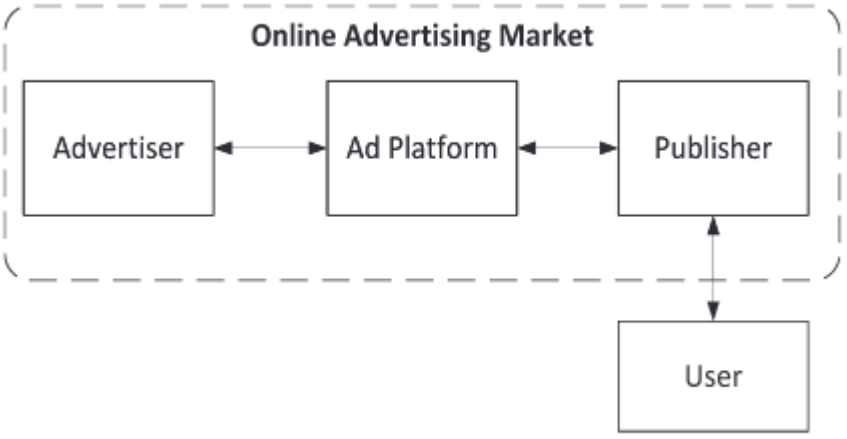


Fig. 2. Main components of the online advertising ecosystem.

User: you and your browser(s)
Publisher: gets paid for display of ad - web site (e.g. google search, CNN, rte.ie)
Ad platform: intermediaries who help advertiser target ads - Google, FB, ...
Advertiser: pay for display of ads - company selling widgets, travel, ...

Moar Actors

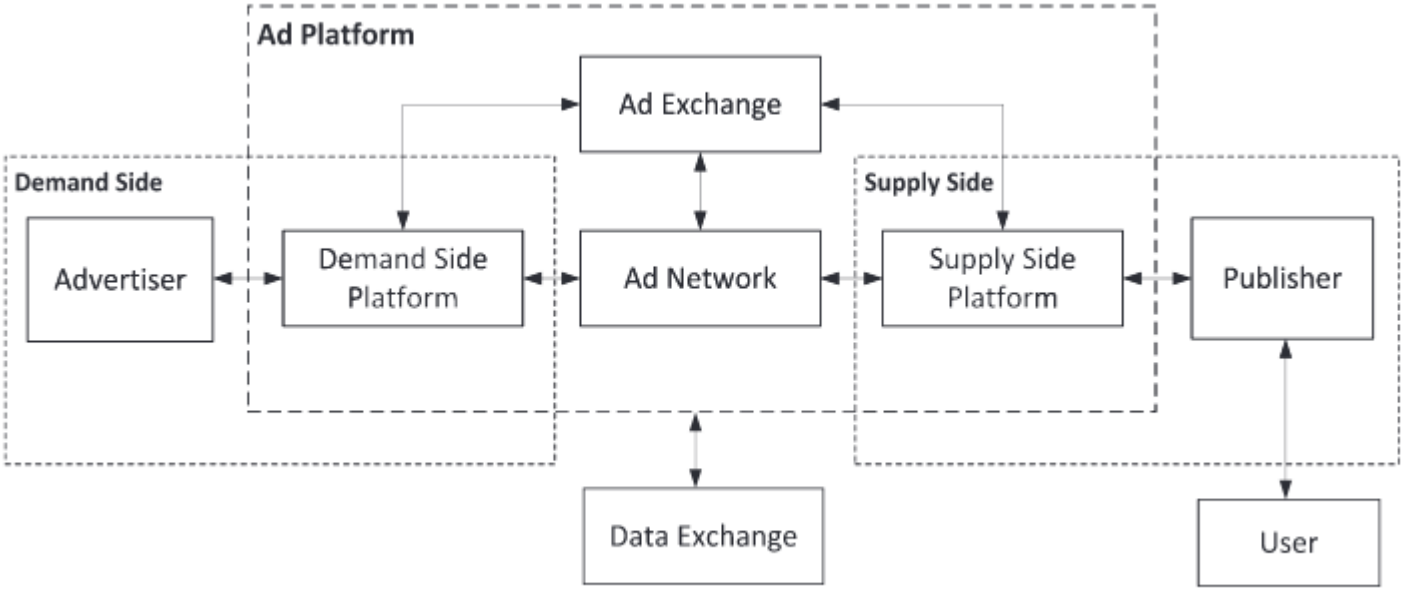
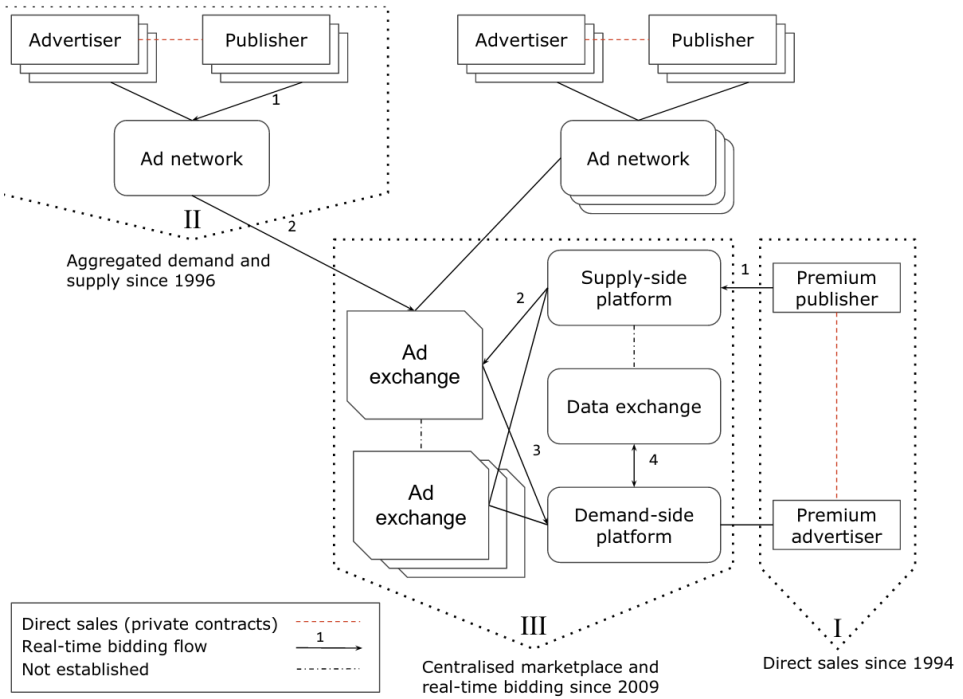


Fig. 3. Disaggregated ad platform scheme and interactions between players.

Real time bidding (RTB):
Publisher -> Ad platform: “I have <this> inventory (of display space)”
Ad platform -> Advertiser**s**: “how much will you pay for <this>?”

Another view

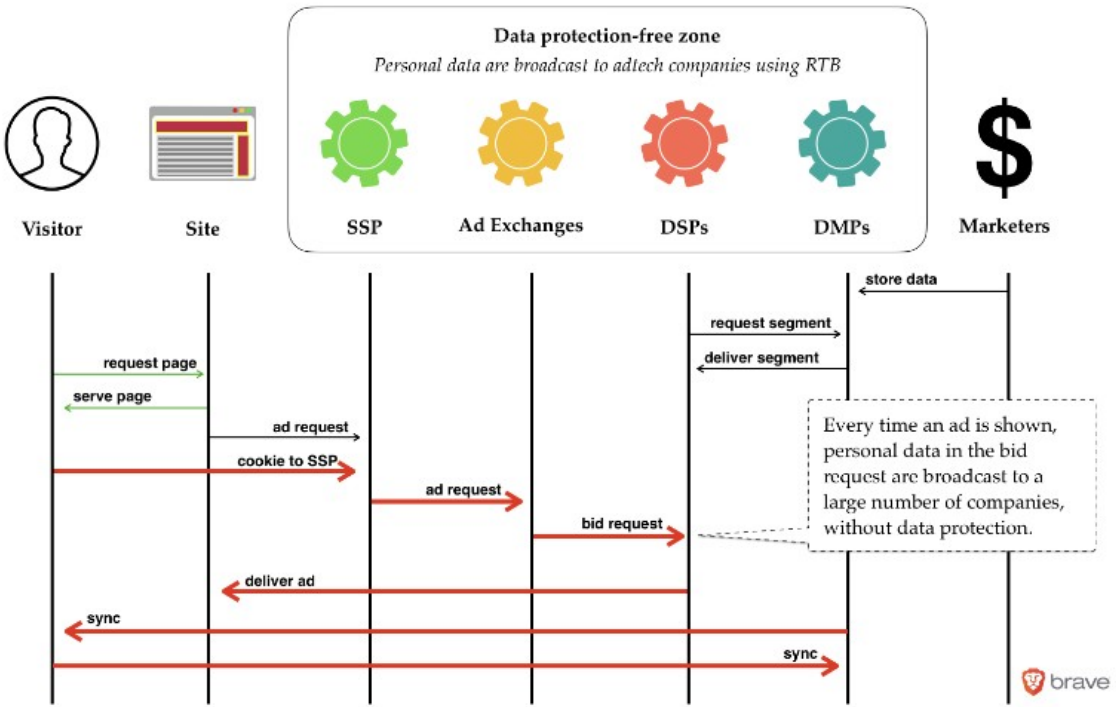


Yuan, Shuai, Jun Wang, and Xiaoxue Zhao. "Real-time bidding for online advertising: measurement and analysis." Proceedings of the Seventh International Workshop on Data Mining for Online Advertising. ACM, 2013. https://arxiv.org/pdf/1306.6542.pdf

Brave’s view...

- Brave is a browser competing with others (chrome, FF, safari...)
  - AFAIK, they have negligible market share at the moment
- As a company, Brave describe themselves as being privacy focused
  - They are trying to promote an alternative to current advertising models
- In 2018-2019, Brave (the company) lodged compaints against real-time-bidding (RTB) in general and e.g., Google’s advertising behaviour with various European data protection agencies
  - https://brave.com/wp-content/uploads/2018/09/Behavioural-advertising-and-personal-data.pdf
- Bid request examples:
  - https://brave.com/wp-content/uploads/2019/02/3-bid-request-examples.pdf
  - Note: These are from Google API samples, not clear to me what’s deployed in the wild now
- The relevant Brave employee is now with ICCL and continues work to challenge RTB
  - https://www.iccl.ie/rtb/

Brave’s view of RTB...



https://brave.com/wp-content/uploads/2018/09/Behavioural-advertising-and-personal-data.pdf

Example OpenRTB bid request 1.

Source: "Sample bid requests: display mobile web request, OpenRTB 2.5", in Configuring an Exchange Bidding Integration, Google Authorized Buyers (URL: https://developers.google.com/authorized-buyers/rtb/exchange-bidding).

```
id: "BIDREQUEST_ID"
imp {
  id: "1"
  banner {
    w: 728
    h: 90
    pos: BELOW_THE_FOLD
    expdir: LEFT
    expdir: RIGHT
    expdir: UP
    expdir: DOWN
    format {
      w: 728
      h: 90
    }
  }
  tagid: "TAG_ID"
  bidfloor: 0.61
  bidfloorcur: "USD"
  secure: true
  metric {
    type: "click_through_rate"
    value: 0
    vendor: "EXCHANGE"
  }
  metric {
    type: "viewability"
    value: 0
    vendor: "EXCHANGE"
  }
  metric {
    type: "session_depth"
    value: 86
    vendor: "EXCHANGE"
  }
  [com.google.doubleclick.imp] {
    billing_id: "BILLING_ID"
    dfp_ad_unit_code: "/DFP_NETWORK_CODE/AD/UNIT/
PATH"
    ampad: AMP_AD_ALLOWED_AND_NOT_EARLY_RENDERED
  }
}
site {
  page: "PAGE_URL"
  publisher {
    id: "SELLER_NETWORK_ID"
    [com.google.doubleclick.publisher] {
      country: "GB"
    }
  }
  content {
    contentrating "DV-G"
    language "en"
  }
  mobile: true
  [com.google.doubleclick.site] {
    amp: DIALECT_HTML
  }
}
device {
  ua: "Mozilla/5.0 (Linux; Android 4.4.4; SM-T560
Build/KTU84P) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/63.0.3239.111 Safari/537.36"
  ip: "IP_ADDRESS"
  geo {
    lat: 42.6495361328125
    lon: 23.35913848876953
    country: "BGR"
    city: "Sofia"
    utcoffset: 120
  }
  make: "samsung"
  model: "sm-t560"
  os: "android"
  osv: "4.4.4"
  devicetype: TABLET
  w: 1280
  h: 800
  pxratio: 1
}
user {
  id: "GOOGLE_USER_ID"
  buyeruid: "HOSTED_MATCH_USER_DATA"
  customdata: "HOSTED_MATCH_USER_DATA"
  data {
    id: "DetectedVerticals"
    name: "DoubleClick"
    segment {
      id: "5444"
      value: "0.3"
    }
    segment {
      id: "1080"
      value: "0.2"
    }
    segment {
      id: "1710"
      value: "0.1"
    }
    segment {
      id: "1715"
      value: "0"
    }
    segment {
      id: "96"
      value: "0"
    }
  }
}
tmax: 162
cur: "USD"
```

What this specific person is reading right now

Distinctive information about this specific person's device

This specific person's IP address

This specific person's GPS coordinates

Various ID codes identifying this specific person, facilitating re-identification and tying to existing profiles

This specific person's inferred interests. This could include highly sensitive special category data such as 571 eating disorders, 410 left-wing politics, 202 male impotence, 862 Buddhism, 625 AIDS & HIV, 547 African-Americans, etc. See Goolge's "publisher verticals" list.

https://brave.com/wp-content/uploads/2019/02/3-bid-request-examples.pdf

Header Bidding (1)

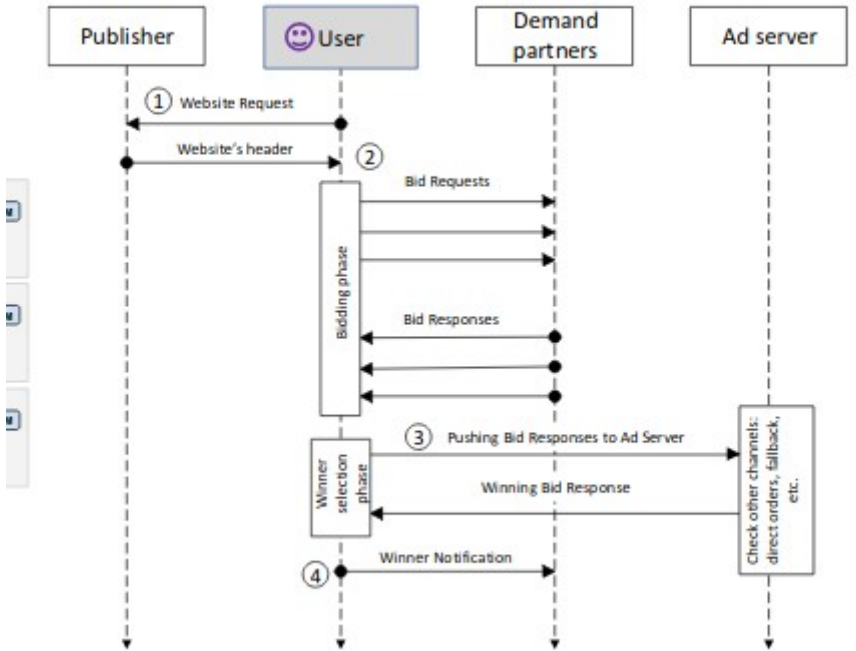


Figure 2: Flow chart of the Header Bidding protocol.

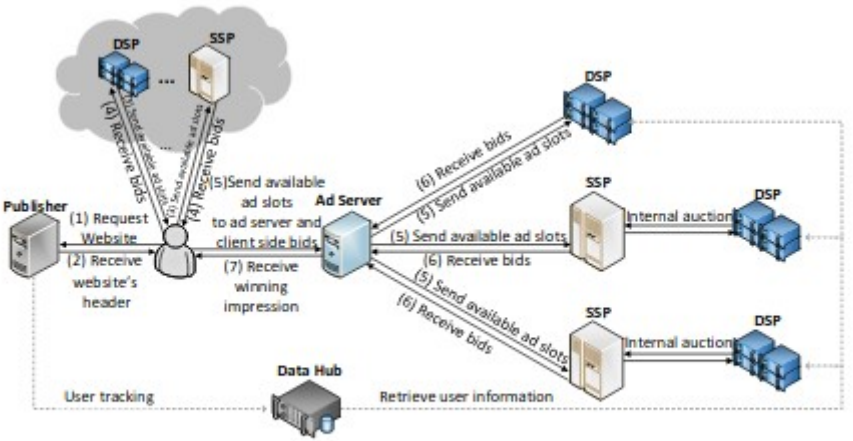
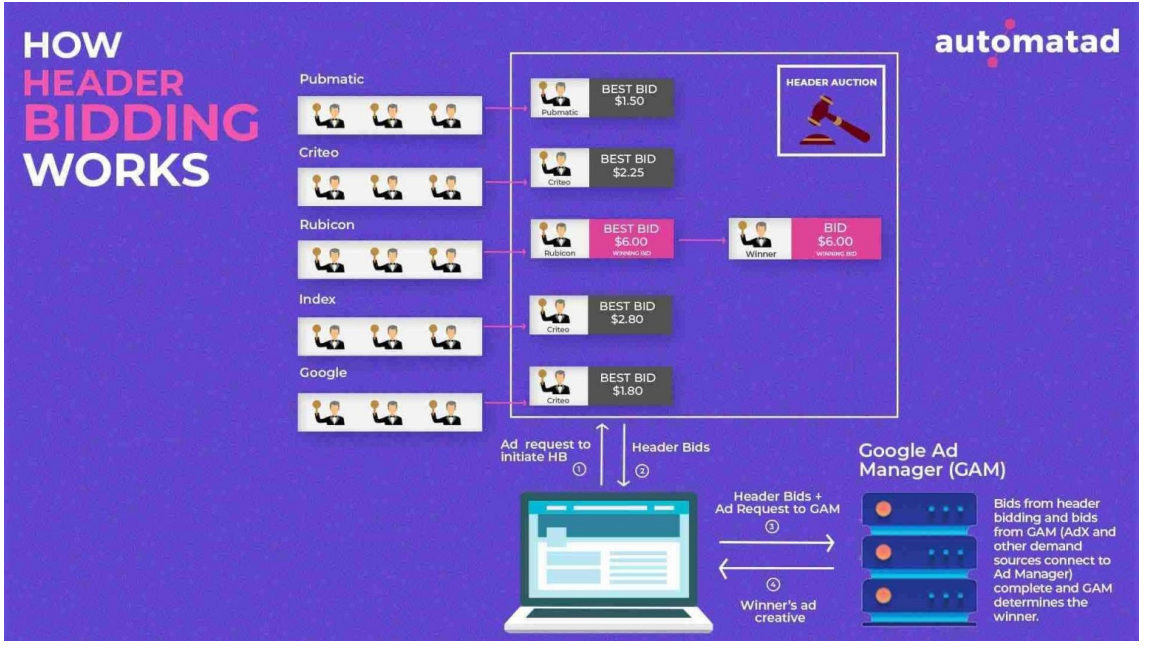


Figure 7: Hybrid HB overview and steps followed.

Pachilakis, Michalis, et al. "No more chasing waterfalls: a measurement study of the header bidding ad-ecosystem." Proceedings of the Internet Measurement Conference. 2019. https://arxiv.org/pdf/1907.12649.pdf

Header Bidding (2)

- Implemented as Javascript on web site that interacts with bidders
- Your CPU/network resources are expended for that
- Structure of bids has similar privacy problems
- Yet another reason to not/selectively enable Javascript?



Source: https://headerbidding.co/header-bidding/

RTB Waterfall vs. Header Bidding

- RTB Waterfall model is (apparently) where SSP tries 1st DSP:
  - If auction won, then render Ad
  - If not, move to next DSP
  - Repeat until done (with possible fallbacks to non DSP Ad sources)
- Header bidding model has much of the action happen in the user's browser via Javascript
- Newer (mostly since 2016) than waterfall model
- Header bidding may represent an adtech attempt to work around Google's exchange
- One claim: in 2022, 16% of top 100k US sites were using header bidding as were 70% of online publishers
  - https://headerbidding.co/header-bidding/

Twitter/X Backend Sharing (1)

- Companies engaged in advertising may say that they do or do not share/sell data but humans are very good at apparently not recognising when they breach/avoid such policies in an entirely self-serving manner
- Twitter advertising example from Oct 2019
  - https://help.twitter.com/en/information-and-ads
  - Still there in Oct 2024
- Advertising partner uploads database including identifiers
- Twitter match that with their user database, sometimes based on phone numbers supplied to twitter for 2-factor authentication
  - Presumably: someone then sends targetted ads to twitter users

X/Twitter Backend Sharing (2)

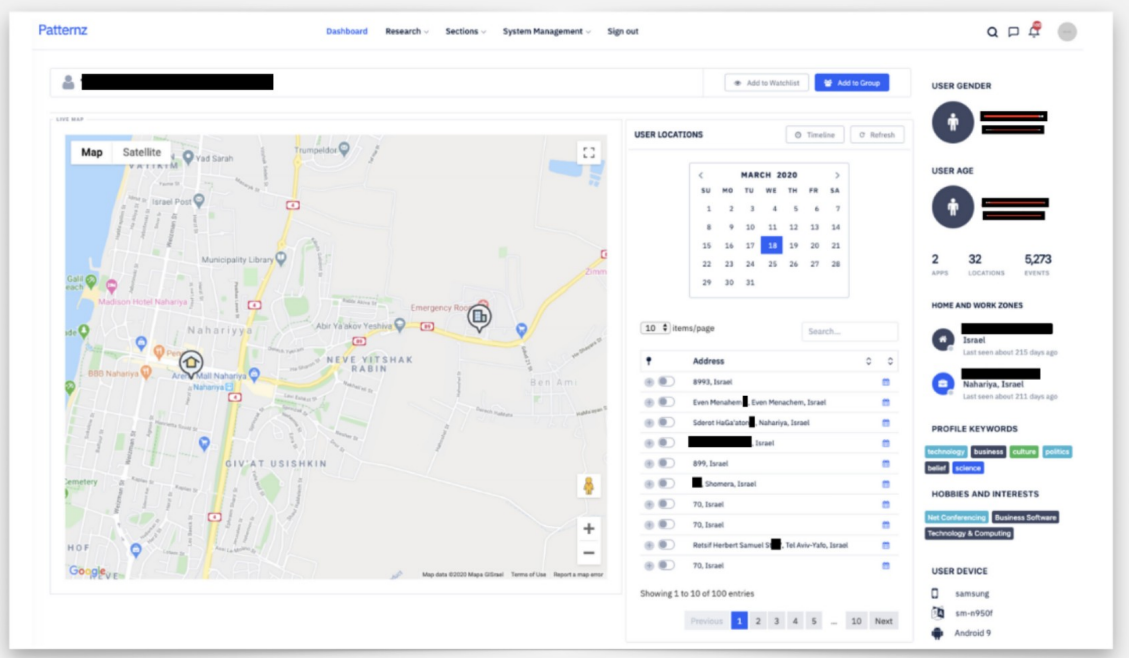
- Are twitter correct in saying “No personal data was ever shared externally with our partners or any other third parties.” ?
- IMO no. Ads may have contained web bugs (1x1 pixel images) allowing “partners or other third parties” to track matching twitter users.
- I’d characterise the above as twitter selling trackable access to their user database.
- I would be extremely surprised if twitter were alone in acting like this. It makes money. Seemingly without harming anyone.
- All of the above is extremely non-transparent.

Who gets to see what?

- In principle: anyone who signs up to an Ad exchange gets request data
  - Could be nation state actors as well as real commercial entities
- “Cookie matching” correlates over time and multiple properties
  - Kind of a collusion between Ad platform and advertisers
  - https://developers.google.com/authorized-buyers/rtb/cookie-guide#examples
    - Includes explanation of how they recover if user clears cookies! (Thanks, google_user_id!)
- Same kind of thing happens with Google user ID and Apple advertising ID
- While most of the above refers to web browsers as the client, all the same things happen with mobile apps (e.g. via Google firebase) and with far less control for the user
- And location, device identifiers, user agent string/application IDs...
- Independent data brokers also exist (more in the US perhaps) that may be able to match non-web data items, e.g. if SSN in both data sets somehow

ICCL's RTB/security Report

- Outlines how RTB data can be abused by "security" entities
  - https://www.iccl.ie/digital-data/europes-hidden-security-crisis/
- Not 100% clear what's real, but pretty horrible IMO



Note: redactions applied to image

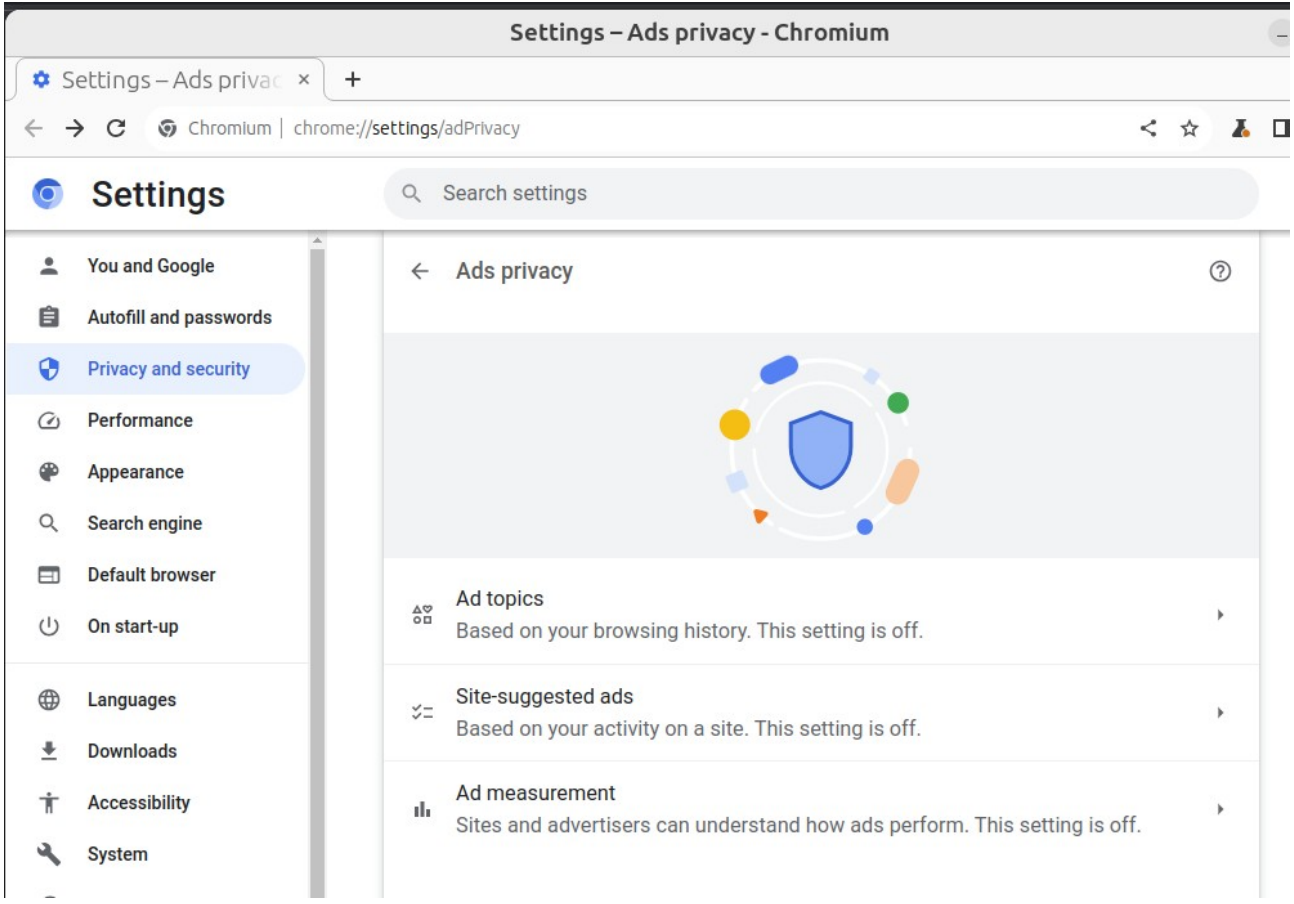
Scale (again)

- To render the Ad, the auction must be speedy
- Speed of light means needing a presence near the auctioneer, e.g. within 120ms => (nearly) the same city as wherever the auctioneer’s data centre
  - 120ms is a Google number, and hey, they’ll also sell you cloudiness so you can meet that number;-)
- Implication:
  - To deal with big Ad platforms, SSP’s and DSP’s need to be big
    => centralisation++

Google “Privacy Sandbox”

- Chrome has repeatedly announced an intent to phase out 3rd party cookies (in ~Q3 2024), this is their replacement, presumably aimed at maintaining their advertising revenue
  - http://privacysandbox.com/timeline
- Most of this is controversial and confusing (the change the names of things quite a bit)
- Topics API – browser infers “topics” based on ~weekly analysis of browsing
  - https://developer.chrome.com/docs/privacy-sandbox/topics/
- Protected Audience API – on-device ad auctions
  - https://developer.chrome.com/docs/privacy-sandbox/protected-audience/
- **BUT**: July 2024: “Google now says it will not deprecate third-party cookies after all”
  - https://martech.org/googles-privacy-sandbox-what-you-need-to-know/

Turning that off



Who benefits?

- User: Sees “relevant” Ads, fewer repeat Ads
  - Cost: privacy, tracking, bandwidth, latency, creepiness
- Publisher: gets revenue
    - +4% for cookies? https://www.eff.org/fa/deeplinks/2019/06/research-shows-publishers-benefit-little-tracking-ads
  - Cost: control -> others (AdX, SSP...), dependency , GDPR costs, technology costs
  - New control issue: Web packaging (AMP etc.)
- Advertiser: presumably gets more sales (or just clicks?)
  - Cost: revenue share with exchanges, technology costs
- Ad platform: YES YES YES
  - Cost: technology costs, so far as I know, little cost due to privacy
- That said, I have not researched (and have no interest in researching) the money flows, I’d prefer it just didn’t!
  - There is a LOT of money flowing though

Online Advertising

- What’s your attitude to online advertising?
- What do you know about how it works?
  - What do you want to know?
- What concerns do you have about online ads?
- Are you ok with being the product when using “free” services?
  - Always or just sometimes?
  - What would you do to avoid being the product?