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’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 kinds of thing would you do to avoid being the product?

A survey paper

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

- Any tables or diagrams without a reference are from there

That's based on work in or before 2016

- So changes will have occurred since then, and will continue

Some Actors

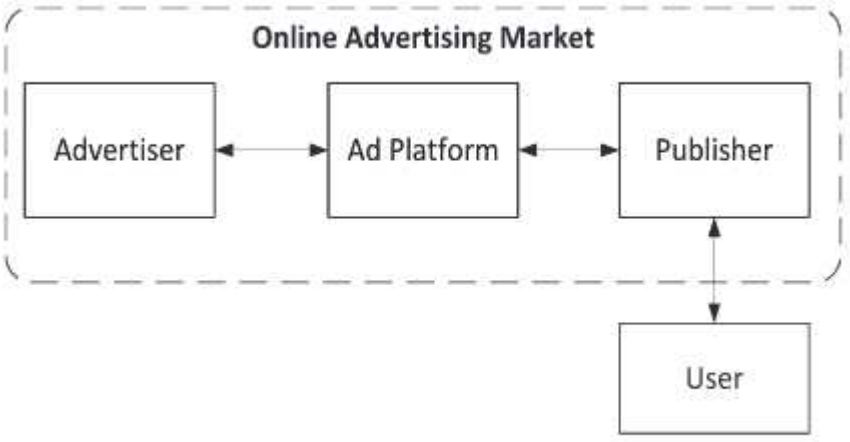
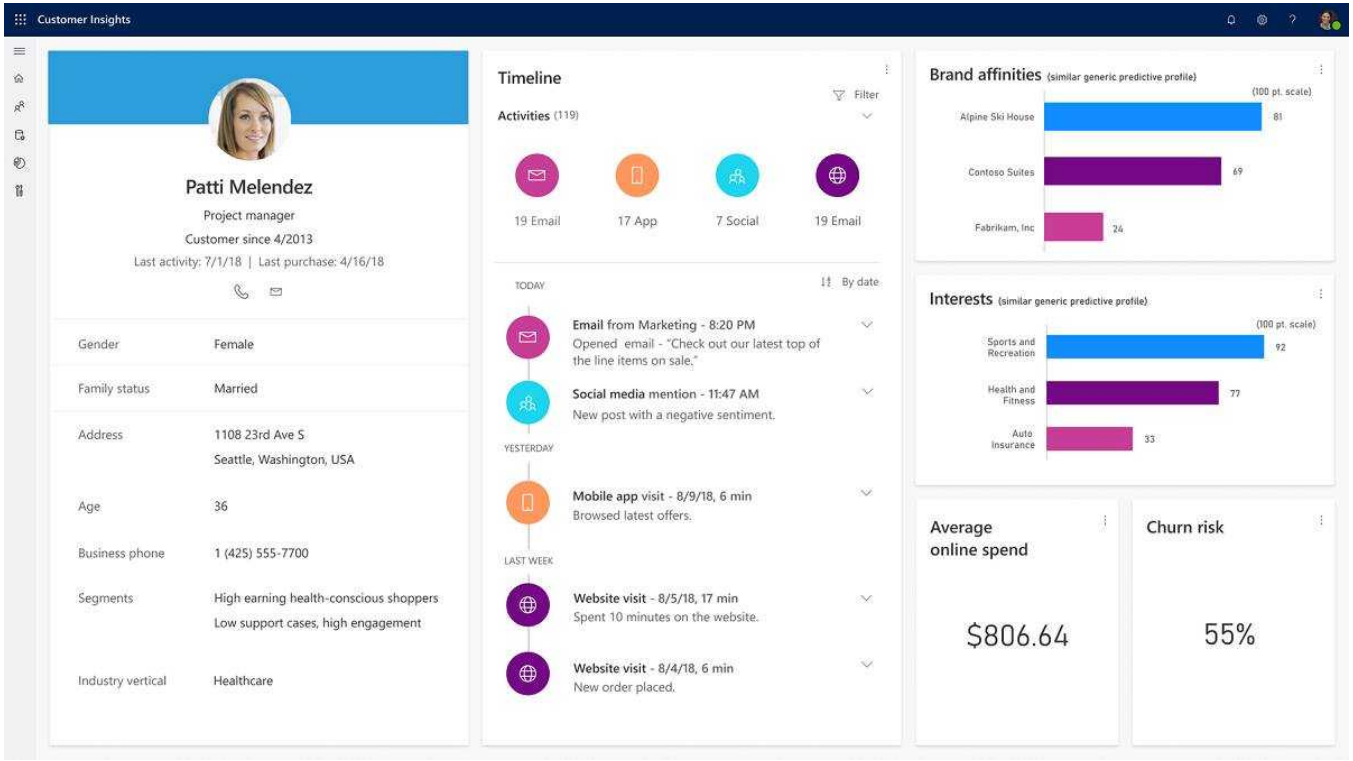


Fig. 2. Main components of the online advertising ecosystem.

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, ...

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?).

Moar Actors

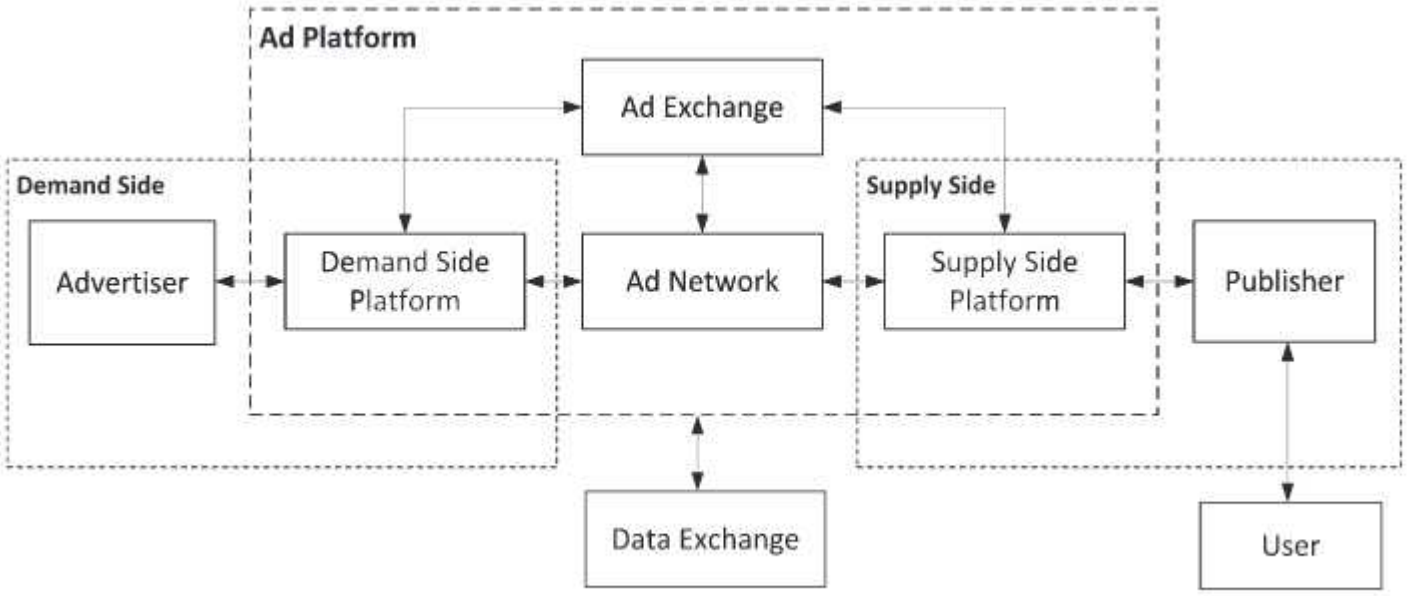
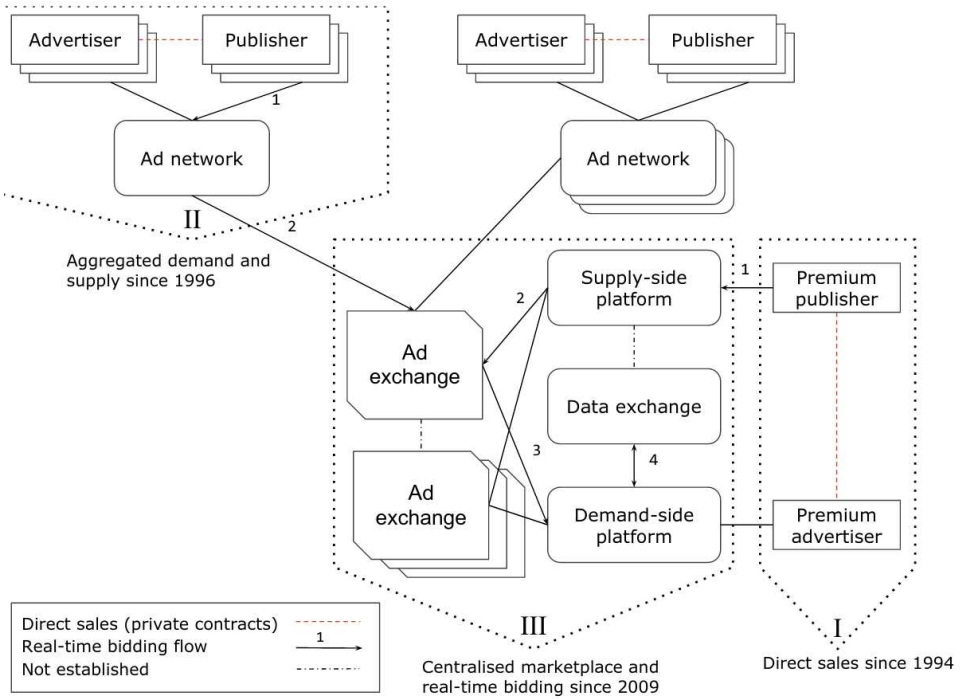


Fig. 3. Disaggregated ad platform scheme and interactions between players.

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

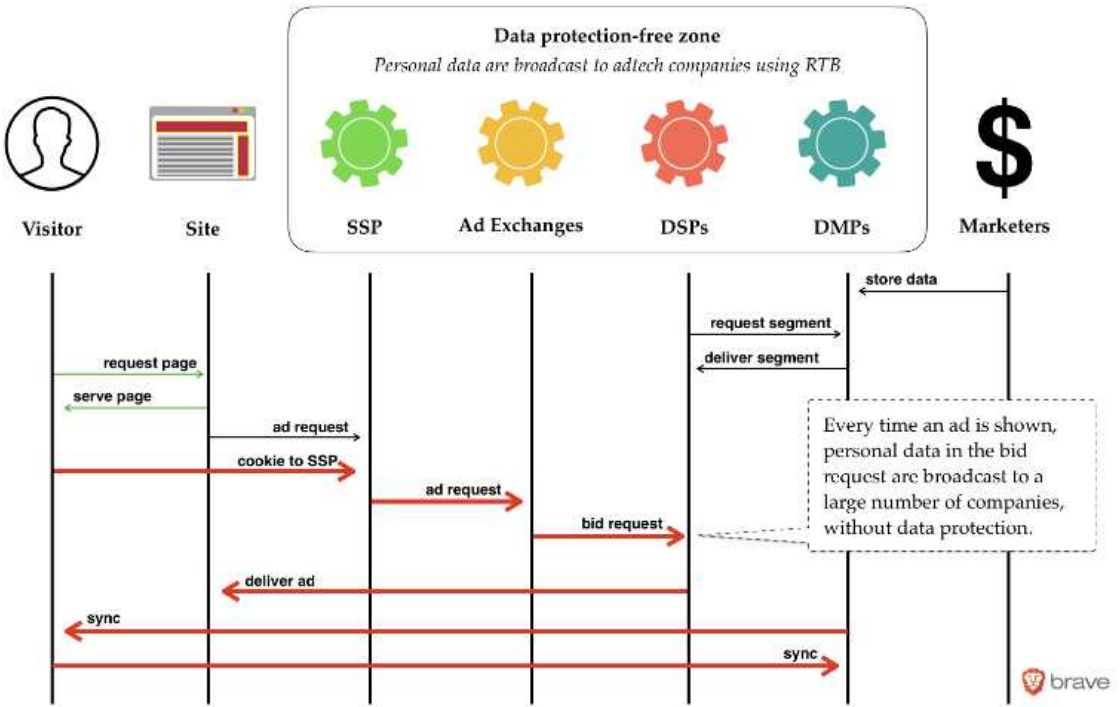


Yuan, Shuai, Jun Wang, and Xiaoxue Zhao. "Real-time bidding for online advertising: measurement and analysis." Proceedings of the Seventh International Workshop on Data Mining for Online Advertising. ACM, 2013. https://arxiv.org/pdf/1306.6542.pdf

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
- Over the last year, Brave (the company) have 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
  - https://brave.com/rtb-updates/
- 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

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
```

What this specific person is reading right now

```
  [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"
```

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

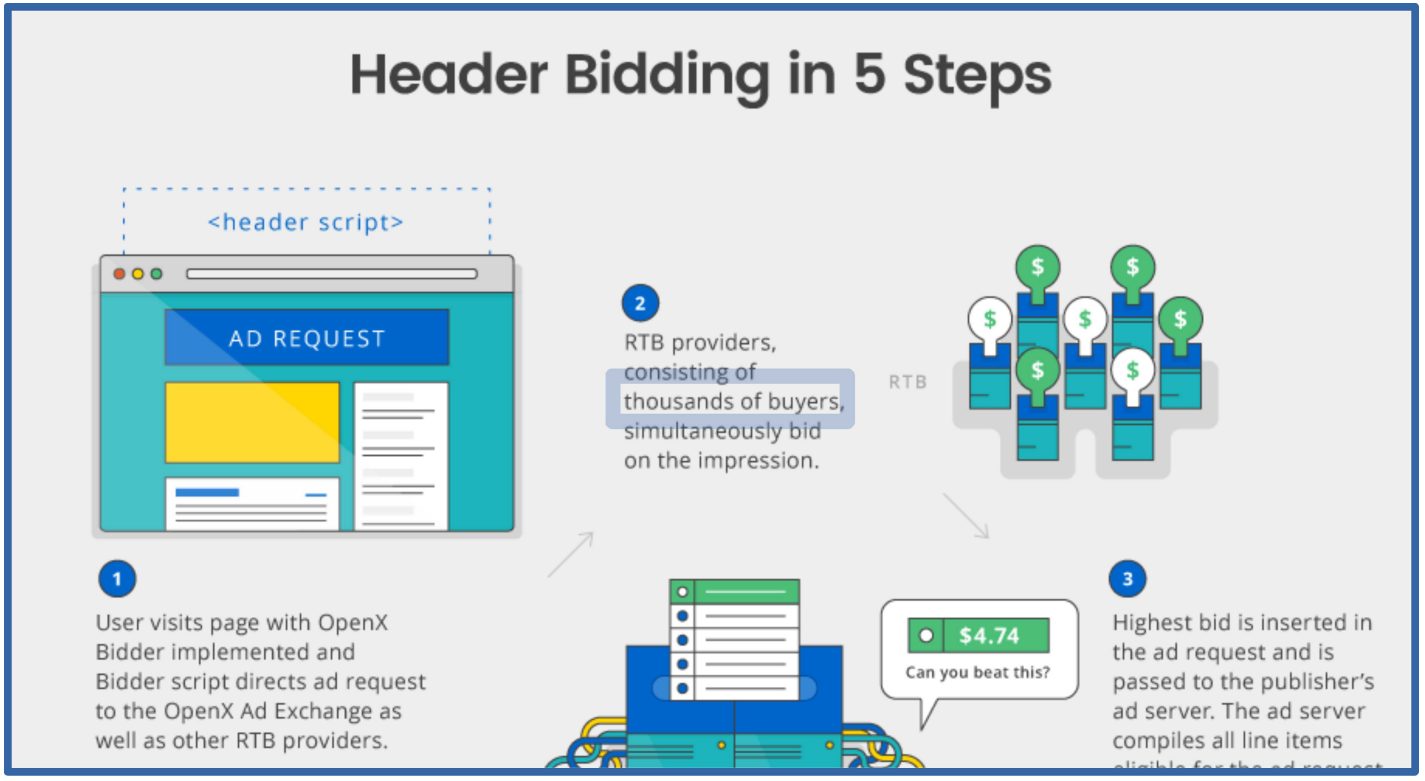
This specific person's inferred interests. This could include highly sensitive special category data such as 571 eating disorders, 410 left-wing politics, 202 male impotence, 862 Buddhism, 625 AIDS & HIV, 547 African-Americans, etc. See Goolge's "publisher verticals" list.

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

RTB Waterfall vs. Header Bidding

- 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 some of the action happen in the user's browser, but is newer and still in flux (also apparently)
- One claim I've seen: in 2019, 14% of top 50k sites using header bidding, 70+% (presumably) using waterfall RTB

OpenX (an Ad exchange)



http://welcome.openx.com/rs/745-BUQ-779/images/Header%20Bidding%20Infographic.png

Header bidding

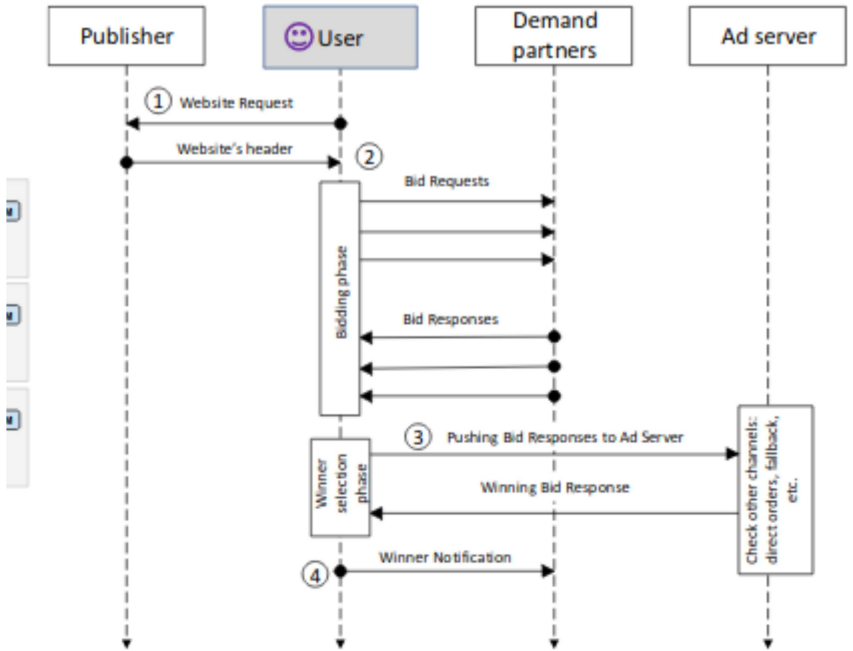


Figure 2: Flow chart of the Header Bidding protocol.

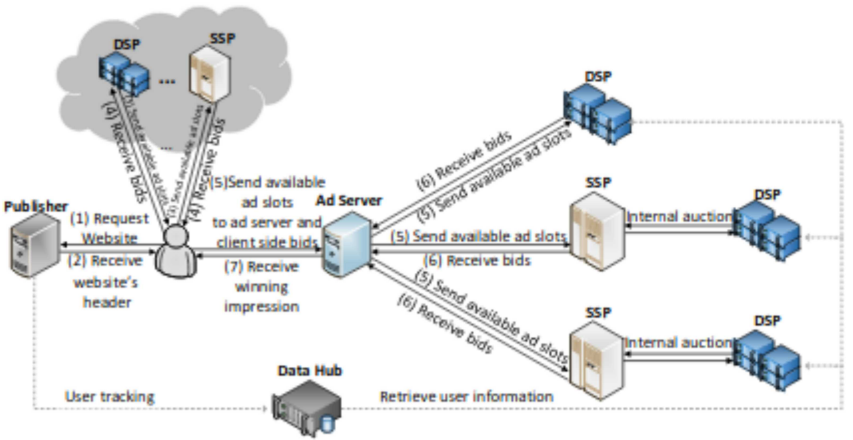


Figure 7: Hybrid HB overview and steps followed.

https://arxiv.org/pdf/1907.12649.pdf

Twitter 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
- Advertising partner uploads database incl. 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

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 well be nation state actors as well as real commercial entities
  - Some researchers use these advertising platforms to do experiments too
- “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
- 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

Some More Papers

- Olejnik, Lukasz, and Claude Castelluccia. "To bid or not to bid? Measuring the value of privacy in RTB." (2015).
  - https://www.inrialpes.fr/planete/people/lukasz/rtb2.pdf
- Pachilakis, Michalis, et al. "No More Chasing Waterfalls: A Measurement Study of the Header Bidding Ad-Ecosystem." arXiv preprint arXiv:1907.12649 (2019).
  - https://arxiv.org/pdf/1907.12649.pdf

Scale (again)

- To render the Ad, the auction must be done asap
  - To win the auction, a speedy response is needed
- 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++

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 (and have no interest in) chasing 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 kinds of thing would you do to avoid being the product?

Assignment 3 (1)

- Investigate and write-up one of:
  - How some web site(s) you use cause(s) cookies to land in your browser, or,
  - How the set of cookies stored in your browser evolves with your normal web usage
- Marks: 15%
- Due date: whacha want? Maybe end of reading week?
  - I'll create that in blackboard sometime soonish:-)
- If there's time today we can make a start now...

Assignment 3 (2)

- How might you do it? Up to you!
- One option:
  - Install a browser (e.g. chromium, opera) that's new for you
  - Explore it's settings (advanced/privacy), zap all it's cookies etc.
  - Visit site(s) of interest, in various states (e.g. logged into to FB/Gmail first or not different browser privacy settings...)
  - Record state of cookies at various stages and with various options
  - Analyse cookie origins, names and content (some extensions/plugins may make that easier)
- Another option: use dev-mode and HAR files
- Caution:
  - Be careful if you zap cookies in a browser you use day-to-day – you might break some login for which you've forgotten the password!
  - If you do install some new addon/extension, check it out first, and consider keeping or deleting it later

Assignment 3 (3)

- Background on cookies:
  - fairly simple overview
    - https://www.gohacking.com/how-browser-cookies-work/
  - 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.
    - http://seclab.tuwien.ac.at/people/atrox/pam19-postgdprcookies.pdf
    - Oddly, https doesn’t work for that one;-(

Assignment 3 (4)

- Report should state: system, browser, (incl. OS/versions) site(s), states, tools-used, process used (incl. tidy-up), results found, your interpetation of those and actions you may (or may not) take
  - Basic idea: if you handed your report to someone else, they should be able to attempt to re-produce your results
  - Tables are fine for the more tech bits, and info about origins/cookies, text for interpretation/conclusions – total amount of prose likely only needs be ~1 page, maybe a couple of pages of tables
  - Co-operation with one another is ok – just say with whom you've worked
- For better marks:
  - Explain the mechanisms involved (e.g. perhaps differences due to logged in gmail or FB account, settings) as well as the consequences of those and cover what you might (or might not) do about what you find
  - Anonymise your text – write text that could be published online without you feeling any additional loss of privacy with which you're uncomfortable
  - For super-duper marks:
    - Try delve into cookie data content (a web search may help decoding, it's not that hard sometimes)
    - Consider non-cookie based tracking mechanisms between web sites and browsers