Marketing and Data Science in the Post-tracking Age

By Martin Anderson

First published ***May 28th, 2020*** *at:*
https://www.iflexion.com/blog/marketing-data-science
Web-archived version

In the wake of the Cambridge Analytica era, the volume of organized public complaint over third-party tracking has outgrown governmental complacence on the topic, with new anti-tracking measures set to all but eviscerate user-based tracking. Let's take a look at the alternative frameworks currently in development, how FAANG has been drawn into both sides of the debate, and what available user data may be left on the table in a post-tracking era.

User-Profiling Across Domains

Cross-domain tracking is the practice of collating information about a specific user from a variety of different websites, apps and interfaces, in order to build up a profile of the user that's more complete than any one site or app would be likely to yield.

Historically this has occurred through the use of 'third party' cookies – long-term, hidden browser preferences that follow the user across a number of websites, and report back interaction data from each domain to a central processing framework that also receives information from other sources.

The method is arguably *too* effective: headlines over the last ten years have decried the intrusiveness of ads that exploited gender issues[1]; disclosed pregnancies[2]; advertised funeral services to the just-bereaved[3]; and that apparently eavesdrop on users' conversations in order to serve them apposite ads[4] (against denials[5] from the software creators), and even track their offline activity[6].

The Fall of Cross-Domain Tracking

In June of 2020 Apple announced that from 'early spring'[7] of 2021, iOS-installed apps would need to explicitly seek permission to track user activity across other apps and websites[8] via its Identifier for Advertisers (IDFA) functionality.

Until now, apps have enjoyed access to a far wider range of user activity data than is indicated in the new default iOS app-tracking permissions dialogue, and the change has led to an aggressive and sustained campaign[9] of protest and resistance from Facebook, among other industry leaders in online advertising.

In March 2020, the WebKit browser engine that powers Apple's Safari browser announced[10] that it would finalize its long campaign against third-party cookies by blocking them outright, stealing a march on Google's Chrome browser, which will not block third-party cookies until 'some time' in 2022[11].

Meanwhile, in February 2021 Firefox implemented 'Total Cookie Protection'[12], preventing any level of cookies (not just third-party cookies) from being tracked across domains, following up on earlier changes[13] in its network architecture designed to completely prevent cross-domain tracking by other methods[14].

The Advertising Giants Fight Back

With its advertising business model threatened, Google has proposed an alternative system called FLOC[15] (Federated Learning of Cohorts[16], see below), that aggregates users into demographic categories without targeting them directly.

TikTok, among many other data-hungry social media ecostructures (particularly in Asia[17]), is also looking into[18] subverting the iOS app-tracking blockade, though details are scarce.

Besides urging demographically valuable iOS users to opt in to tracking[19], and attempting to frame these innovations as an attack on small businesses[20], it's unclear yet if Facebook will likewise respond with alternative tracking approaches.

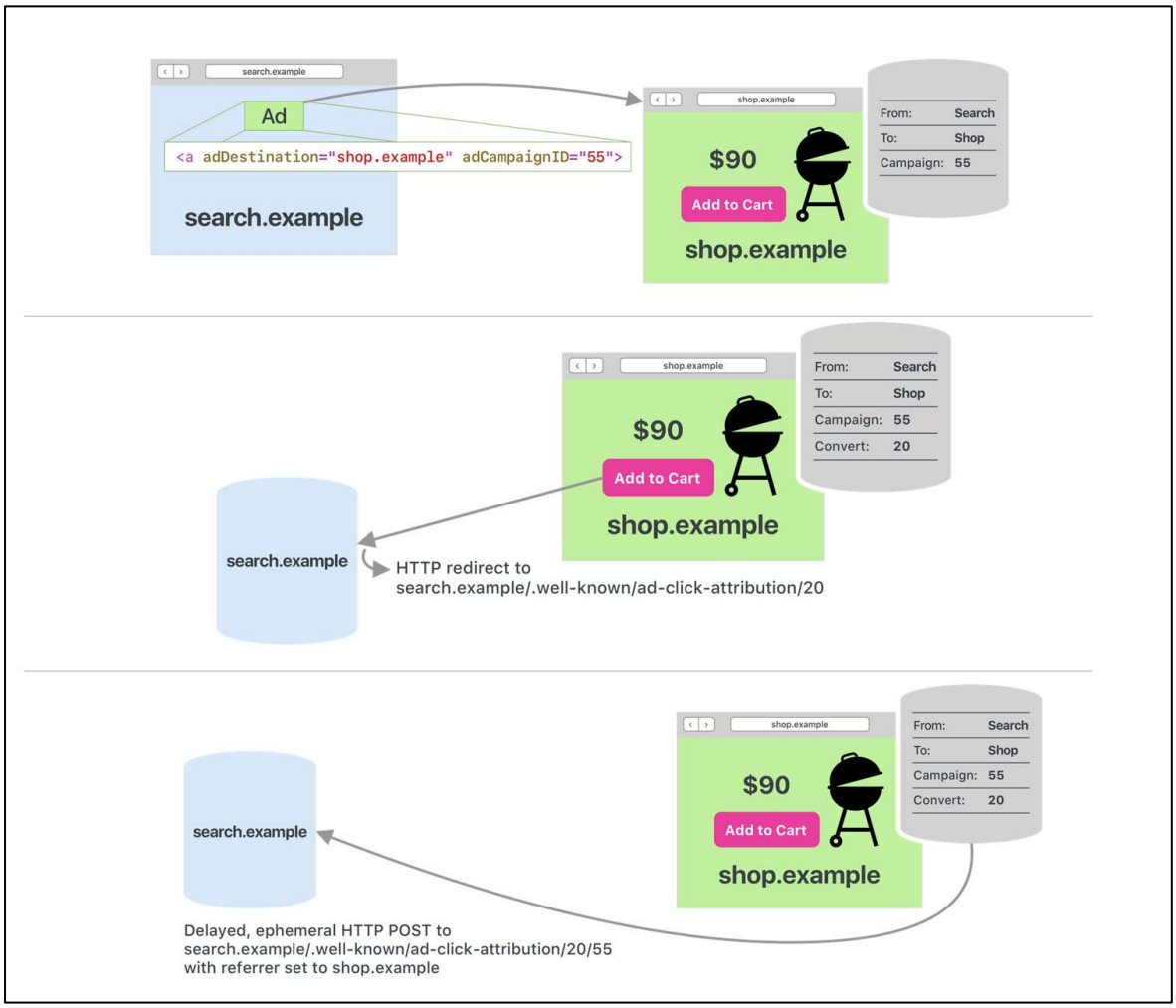
SKAdNetwork: Apple's Cross-Platform Ad Click Tracking

When Apple began to prohibit or limit the use of third-party cookies with Intelligent Tracking Prevention in 2019, WebKit, the upstream open source browser for Apple's Safari, began to implement an alternative tracking method, Ad Click Attribution[21], via the SKAdNetwork SDK.

Ad Click Attribution is effectively a 'firewall' for user data, wherein a limited amount of data reporting on the user is allowed, and semi-anonymized cross-tracking permitted so long as the nodes in the event path are sites that the user actually visits.

SKAdNetwork Architecture

A rudimentary conversion funnel can be delivered to advertisers, without specifying the user:



Source: https://webkit.org/blog/8943/privacy-preserving-ad-click-attribution-for-the-web/

First (top), the user's ad click sends general information to the store about the origin of the click, together with a campaign ID with a maximum ID of #64, so that complex ID-based tracking is theoretically impossible.

Secondly (middle), the browser (essentially Safari[22]) uses a common campaign ID linked to the domain mapping to identify one of four 'conversion events':

- The addition of an item to a shopping cart.
- Subscription to a service.
- Entering of shipping and payment information.
- Purchase of an item.

Finally (bottom of image), as the user converts, the browser schedules a POST request, timed to trigger anywhere between 24 and 48 hours later, that notifies the ad-bearing site that a conversion has occurred. This programmed delay prevents the originating site/s from identifying users that click on ads and then convert rapidly.

Limit on User Data For SKAdNetwork

The domains involved are subject to a 'privacy budget' (though this term was coined by Google's later adaptation, FLOC, as we'll see) which limits the number of data points about the user that can be passed in requests.

Thus the advertiser may not have enough 'budget' left to request, for instance, both the user's geolocation and the time of day that the purchase was made.

There are many more caveats[23] to this process than we have space to list here, but they include:

- The information is sent in Private Browsing Mode, avoiding persistent storage, even where the user is not browsing in Private Mode.
- No cookies, client certificates or potentially semi-permanent data exchange are supported.
- The initial ad click data is only stored for a week, so that no long-term monitoring or 're-identification' through campaign entropy is theoretically possible.
- Neither the website where the ad click event occurs nor the site where the conversion occurs have any knowledge of the interstitial storage or processes. Instead, the originating ad site receives conversion information as a kind of 'blind event', allowing some basic monitoring of the efficacy of ad campaigns.

Emerging Platforms for SKAdNetwork

Though Apple provides documentation[24] for setting up Ad Click Attribution, it does not provide any platform equivalent to DoubleClick or existing simplified frameworks. It has been argued[25] in marketing forums that small publications are unlikely to have the development acumen to set up campaign workflows under this new regime.

However, this does leave a potential gap in the market for service providers who wish to develop new platforms to streamline the process affordably and at scale. A slowly growing number of providers offer dedicated support for the WebKit/Safari SKAdNetwork system, including Singular[26] and Kochava[27].

With Safari occupying approximately 20% of the general browser market[28], and not quite 25% of mobile browser share[29], heavy investment in Ad Click Attribution would currently need to be justified with campaigns that can effectively target the superior spending power[30] of Apple users.

However, if SKAdNetwork ultimately receives better public support than FLOC over the next two years, this may strengthen its long-term viability as a platform for data science consulting.

Is SKAdNetwork Fit for Purpose?

Both Ad Click Attribution and App Tracking Transparency (ATT[31], which will require explicit user consent to enable or re-enable app-based tracking) are currently set to go live 'sometime in spring' of 2021, though the launch[32] of iOS 14.5, which is able to fully activate these systems, has been confirmed for April 2021.

SKAdNetwork can be implemented in Facebook, and the social giant provides some information[33] for businesses about integrating it into Facebook events management, with the caveat that a maximum of 63 events can be programmed per app.

Uptake of SKAdNetwork

A report from mobile ad platform Moloco in March 2021 declared[34] that the percentage of bid requests over the SKAdNetwork rose from 14.5% to 20% in the last two weeks of February 2021 (though this is likely due to mounting pressure over the vague 'spring' deadline).

Anurag Agrawal, Moloco's VP of product, also observed that in some countries bid requests tripled in a matter of days in that period, and that usage will likely rise as launch nears.

SKAdNetwork's privacy budget is forcing developers to choose between salient data points for conversion funnels. This is a trickier decision to make than it appears, since not every app that supports the system is obliged to make all the possible SKAdNetwork signals available.

It seems likely that the entire summer and fall of 2021 will prove an arduous testing ground for the framework, and that clarification on practical usage will be added to the documentation[35] eventually, based on user feedback.

Google's Privacy Sandbox and FLOC

With a 64% share[36] of the browser market and the largest single share (31%)[37] of the online advertising sector, Google's response to the anti-tracking lobby will likely define the near-future of digital marketing.

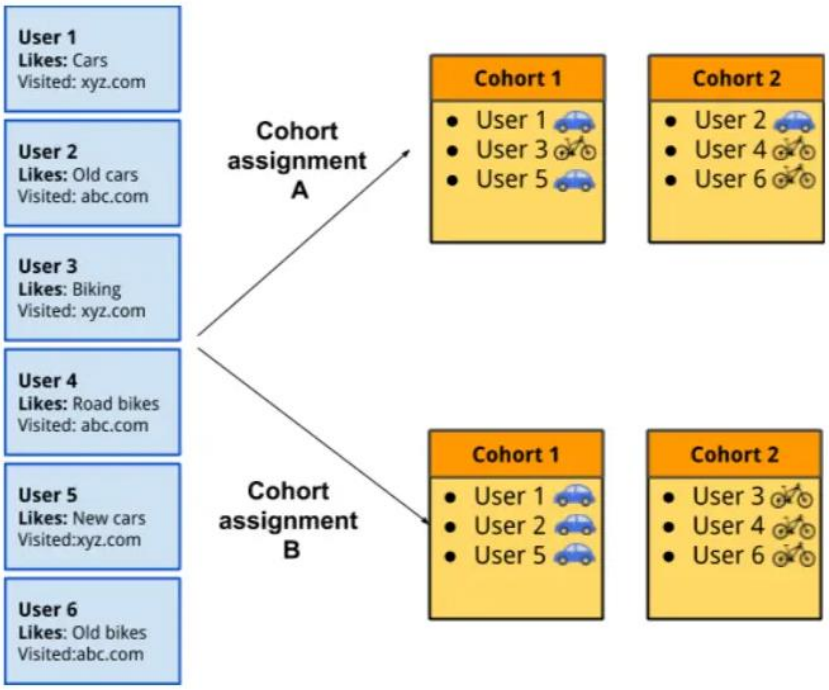
To date, the measures Google has taken, and plans to take, include:

- The removal of third-party cookies (tracking cookies) in Chrome by 2022[38], at least a year behind competing browsers, which will break tracking, targeting and profiling systems that are currently dependent on third-party cookies.
- The wide deployment of a 'Privacy Sandbox' in Chrome – an architecture released as 'Google Signals' in 2018, and which effectively offers even more cross-tracking functionality than third-party cookies, given the right circumstances, and transfers greater control of ad data infrastructure to Google's own platform.
- The implementation of FLOC within the Privacy Sandbox architecture over the course of 2021, with trials currently taking place in countries not covered by GDPR[39].

How FLOC Works

FLOC operates as an embedded client-side technology inside Chrome V89+. It has a more complicated architecture[40] than SKAdNetwork, though it adheres to many of the same concepts.

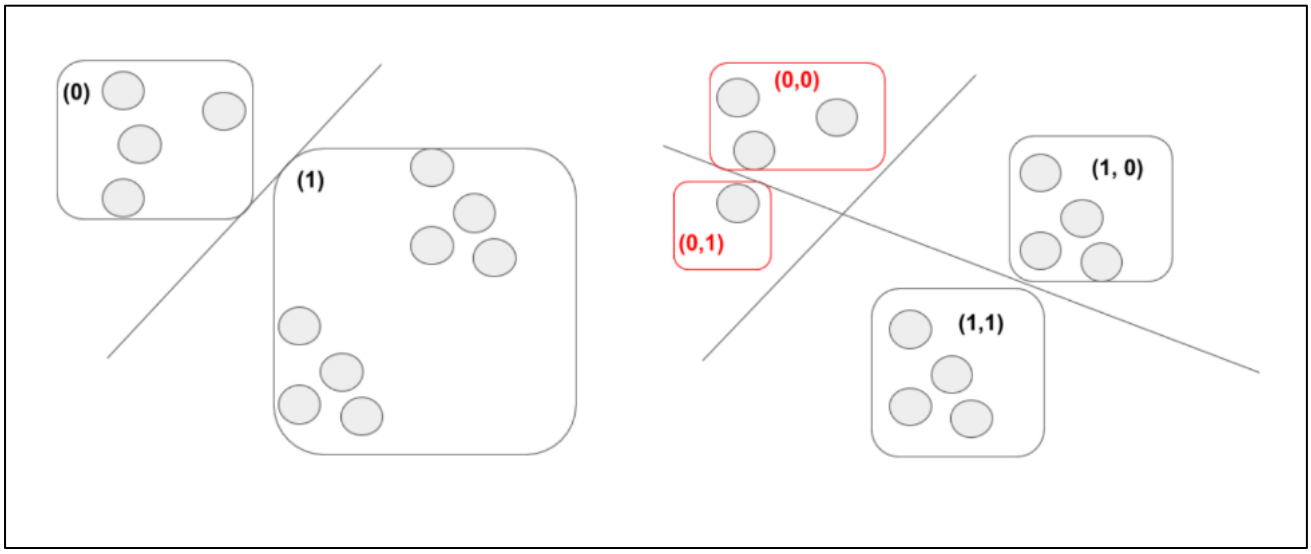
Chrome users are assigned 'Cohort IDs' based on domain visits (and possibly other factors — see below), and are then included in a group (Cohort) with other users with similar interests.



Source: https://github.com/google/ads-privacy/blob/master/proposals/FLoC/FLOC-Whitepaper-Google.pdf

Interactions from a user's Cohort ID will later be correlated to ad campaigns, and eventually reported to the ad-originating site. In theory, advertisers will be able to 'target' the interests of the cohort group, but not the specific profile of anyone within that group.

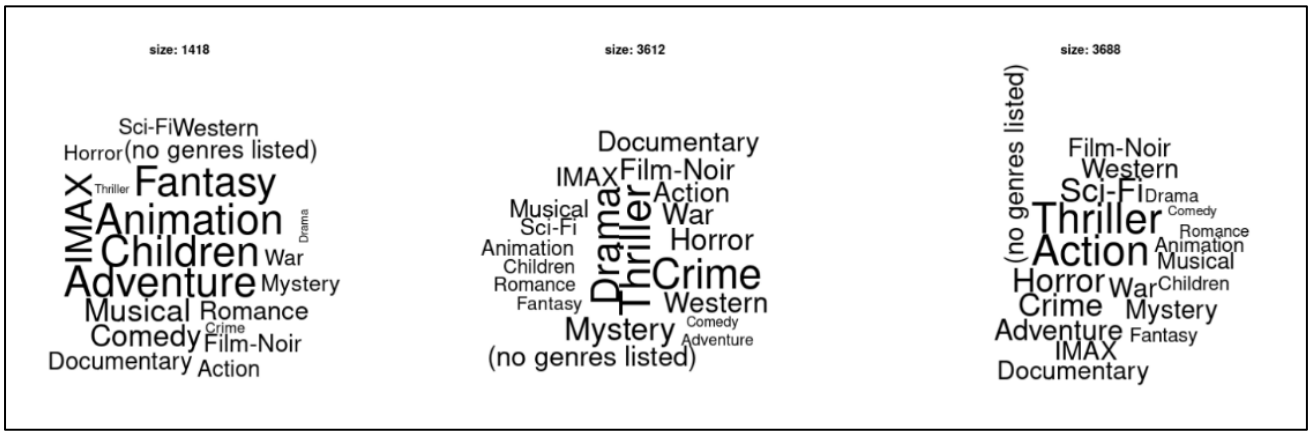
If the cohort group is too small (i.e., the defining topics that generated it are marginal), the users inside it risk exposure. Therefore additional algorithms are run on the cohort to ensure a minimum size:



Red-highlighted cohorts are too small to preserve user anonymity, and will be automatically added to other cohorts with nearest-aligned interest categories.

Since FLOC groups need to be a minimum size in order to semi-anonymize users, smaller groups will be combined with others in order to preserve an adequate 'crowd' to obscure any one user ID; but there is no guarantee that FLOC's final implementation will limit itself in this way.

Contextual feature generation occurs for each Cohort ID via clustering, and affinity clustering eventually siphons matched Cohorts into appropriate groups:



SimHash word clouds in a FLOC test case on a movie dataset, showing how user interest will lead to inclusion in a Cohort group.

As of April 2021, FLOC is being trialed in various locales around the world (with the exception of Europe[41], for purposes of compliance with GDPR).

Re-Enabling Cross-Tracking Under FLOC

FLOC uses the SimHash[42] algorithm to create Cohort groups, with an individual user's Cohort ID initially generated based on their visits to FLOC-enabled domains.

However, the Electronic Frontier Foundation (EFF) has noted[43] that there is nothing to stop Google widening this criteria to page content and other facets, which might add granular detail to a user's profile and group placement, presenting a wider 'attack surface' for companies seeking to circumvent anti-tracking.

There is no documented, finalized bit-length yet for a Cohort ID, and if 16-bit IDs are used instead of 8-bit, this extra length will tend to make a user more identifiable as time passes. Since IDs are calculated weekly, this potentially gives advertising frameworks a week at a time to exploit a rapidly growing level of detail about a user.

Piecing Together Consistent User Profiles From Transient Cohort IDs

In April 2021 web security engineer John Wilander[44], one of the engineers[45] behind Apple's Intelligent Tracking Prevention, commented[46] on FLOC's GitHub issues thread that since multiple sites can monitor the Cohort ID over time, a hash of observed cohorts will become increasingly unique, effectively re-enabling cross-tracking.

The EFF has characterised[47] FLOC tracking as a 'behavioral credit score', and contends that the semi-anonymous nature of the scheme does not prevent third-party companies from adding a user's available FLOC

data to a more complete, long-term internal profile. In this way it could become possible to quickly and systematically 're-acquire' a specific user via the limited and transient information that FLOC offers.

Potential Availability of FLOC Data

Before anticipating a new era of FLOC-based analytics, it's important to consider how available the technology is actually going to be, and the extent to which Google is likely to make any concessions against complaints that FLOC is invasive, and little different from the regime it replaces.

A number of downstream Chromium-based browsers have already committed to actively blocking FLOC functionality, offering Chrome addicts an alternative to Google's brand of the open source project. The Vivaldi browser has announced[48] that it will modify the Chromium engine to remove FLOC, while Brave has also promised[49] to block the technology in its releases. It has also been reported[50] that Microsoft has little interest in implementing FLOC technology in its Chromium-based Edge browser.

FLOC requires client-side code, and cannot be implemented in Firefox, or in any other browser that does not use the Chromium engine, including Safari – which has, perhaps, the least motivation to support FLOC.

Disabling FLOC Programmatically

Users can disable FLOC in their Chrome browser by disabling third-party cookies from preferences or with the `chrome://settings` flag, though this is a sweeping and well-buried solution that will ultimately make no difference to FLOC's functionality, since Chrome plans to disable third-party cookies in 2022 while enabling FLOC by other means.

The popular privacy-preserving search engine DuckDuckGo offers[51] a Chrome extension to block all FLOC interactions for Chrome users, while site owners can opt out of FLOC by adding[52] a new HTML header to their content, either to content page templates, or by including the directive in an .htaccess file[53]:

```
Permissions-Policy: interest-cohort=()
```

At the time of writing, users on free CloudFlare plans may not be able to add this to their cached CDN content unless they are on a paid enterprise plan. Though the header can be injected[54] via CloudFlare Workers, there are limits[55] on requests, CPU and memory usage on free accounts.

Will Popular Sites Support FLOC?

Whether FLOC prospers or falls may depend as much on industry adoption as consumer acceptance. If the tech community fails to convince the average Chrome user of the evils of FLOC, and ad-supported media sees FLOC as the only route out of an existential crisis, the current furor may not achieve its goals.

According to an HTML code search[56] in NerdyData, as of mid-April 2021, only six .com websites in the US are disabling FLOC — and two of those are DuckDuckGo and Brave, each with a stake in the issue. In the same period, the code search engine at publicwww.com identifies[57] just 20 web pages with any domain suffix, (out of 522 million indexed pages) that actively block FLOC.

Obviously, the media is likely awaiting some kind of public consensus around FLOC as the trial progresses into the summer, and in the meantime can enjoy a neutral stance.

At this time, the architecture of FLOC is subject to change as trial results emerge, and headlines continue to harangue the technology. Therefore FLOC lacks the same platform support that is emerging for

SKAdNetwork, not least because all of Apple's privacy initiatives are being finalized and implemented in the spring and summer of 2021, whereas Chrome will not be disabling third party cookies until 2022.

No Road Back to the 'Golden Age' of Cross-Tracking

Though rumors of various solutions to Apple's pro-privacy initiatives have circulated over the last year, there are no obvious 'drop-in' technologies to replace the full capabilities of cross-domain tracking. In the meantime the future of FLOC, currently a volatile and under-documented technology, is in the balance.

Even if cross-tracking *could* be re-established, scandals of recent years have set the public will hard against it: the re-appearance of the 'psychic' ad campaigns that characterised the Cambridge Analytica years would simply reveal a new 'zero day' cross-tracking architecture, which would then come under popular attack in almost the same way as a computer virus.

If the consumer climate is so hostile to targeted ads that an advertiser cannot leverage detailed user data to create them without revealing their hand, there is less benefit in having the information in the first place.

A Level Playing Field?

This leaves marketing companies with SKAdNetwork in the immediate future, and whatever FLOC transitions into in the longer term. Though Apple's system offers fewer users, those users are demographically more valuable, and the system itself is nearer usable deployment than FLOC.

If nothing else, SKAdNetwork may become a proving ground in 2021-22 for the central concept of Cohorts-based advertising, generating insights that could inform the development of FLOC, which has already drawn heavily on Apple's initiative.

Besides this downgraded version of cross-tracking, marketing companies may revive the long-abandoned practice of contextual advertising, where ad categories are defined not by the user, but by the content that they're consuming. Likewise for demographic advertising, long relegated to second position during the cross-tracking years, but a method nonetheless that has produced results for centuries.

Finally, the revived importance of first-party information may lead to the re-emergence of domains and consumer environments where the user, paid or free, must log in via local systems of authentication, instead of 0Auth, Facebook and Google tokens, since those methods co-opt a great deal of the user's data for themselves.

However, this would likely lead to a resurgence of data breaches, as the domains would need to implement their own security frameworks once again.

[1] https://chupadados.codingrights.org/en/gendered-targeted-ads/
[2] https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

[3] https://www.brandwatch.com/blog/react-creepy-marketing-personalisation-goes-far/
[4] https://www.youtube.com/watch?v=U0SOxb_Lfps
[5] https://about.fb.com/news/h/facebook-does-not-use-your-phones-microphone-for-ads-or-news-feed-stories/
[6] https://www.washingtonpost.com/technology/2020/01/28/off-facebook-activity-page/
https://www.washingtonpost.com/technology/2020/01/30/help-desk-if-facebook-is-stalking-me-what-about-instagram/
[7] https://www.cultofmac.com/733548/ios-14-app-tracking-transparency-will-go-live-in-early-spring/
[8] https://www.apple.com/privacy/docs/A_Day_in_the_Life_of_Your_Data.pdf
[9] https://www.wsj.com/articles/facebook-fb-4q-earnings-report-2020-11611717463
[10] https://webkit.org/blog/10218/full-third-party-cookie-blocking-and-more/
[11] https://www.theverge.com/2020/3/24/21192830/apple-safari-intelligent-tracking-privacy-full-third-party-cookie-blocking
[12] https://blog.mozilla.org/security/2021/02/23/total-cookie-protection/
[13] https://blog.mozilla.org/security/2021/01/26/supercookie-protections/
[14] https://www.ftc.gov/system/files/documents/public_comments/2015/10/00064-98109.pdf
[15] https://www.theverge.com/2021/3/30/22358287/privacy-ads-google-chrome-floc-cookies-cookiepocalypse-finger-printing
[16] https://blog.google/products/ads-commerce/2021-01-privacy-sandbox/
[17] https://www.ft.com/content/520ccdae-202f-45f9-a516-5cbe08361c34
[18] https://arstechnica.com/gadgets/2021/03/chinas-tech-giants-test-way-around-apples-new-privacy-rules/?comments=1
[19] https://www.cnbc.com/2021/02/01/facebook-strikes-back-against-apple-ios-14-idfa-privacy-change.html
[20] https://about.fb.com/news/2020/12/speaking-up-for-small-businesses/
[21] https://webkit.org/blog/8943/privacy-preserving-ad-click-attribution-for-the-web/
[22] https://www.dummies.com/web-design-development/site-development/common-webkit-browsers/
[23] https://webkit.org/blog/8943/privacy-preserving-ad-click-attribution-for-the-web/
[24] https://developer.apple.com/documentation/storekit/skadnetwork
[25] https://old.reddit.com/r/adops/comments/buukif/privacypreserving_ad_click_attribution/
[26] https://www.singular.net/skadnetwork/
[27] https://www.kochava.com/skadnetwork-solutions/
[28] https://gs.statcounter.com/browser-market-share
[29] https://gs.statcounter.com/browser-market-share/mobile/worldwide
[30] https://www.marketingdive.com/news/survey-iphone-owners-spend-more-have-higher-incomes-than-android-users/541008/
[31] https://developer.apple.com/documentation/apptrackingtransparency
[32] https://www.cnet.com/how-to/ios-14-5s-release-date-major-new-features-and-more-everything-we-know/
[33] https://www.facebook.com/business/help/188126096313109
[34] https://www.adexchanger.com/privacy/skadnetwork-adoption-rises-but-still-too-low-for-devs-to-avoid-disruption/
[35] https://developer.apple.com/documentation/storekit/skadnetwork
[36] https://gs.statcounter.com/browser-market-share
[37] https://www.t4.ai/industry/internet-advertising-market-share
[38] https://theconversation.com/googles-scrapping-third-party-cookies-but-invasive-targeted-advertising-will-live-on-156530
[39] https://www.adexchanger.com/platforms/google-will-not-run-floc-origin-tests-in-europe-due-to-gdpr-concerns/
[40] https://github.com/google/ads-privacy/blob/master/proposals/FLoC/FLOC-Whitepaper-Google.pdf
[41] https://techcrunch.com/2021/03/30/google-starts-trialling-its-floc-cookie-alternative-in-chrome/
[42] http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33026.pdf
[43] https://www.eff.org/deeplinks/2021/03/googles-floc-terrible-idea
[44] https://github.com/johnwilander
[45] https://webkit.org/blog/7675/intelligent-tracking-prevention/
[46] https://github.com/WICG/floc/issues/100
[47] https://www.eff.org/deeplinks/2019/08/dont-play-googles-privacy-sandbox-1
[48] https://vivaldi.com/blog/no-google-vivaldi-users-will-not-get-floced/

[49] https://brave.com/why-brave-disables-floc/
[50] https://www.theverge.com/2021/4/16/22387492/google-floc-ad-tech-privacy-browsers-brave-vivaldi-edge-mozilla-chrome-safari
[51] https://www.ghacks.net/2021/04/10/duckduckgo-extension-blocks-google-floc-in-latest-update/
[52] https://community.cloudflare.com/t/blocking-google-flocs/257043/3
[53] https://htaccessbook.com/add-custom-headers-htaccess/
[54] https://paramdeo.com/blog/opting-your-website-out-of-googles-floc-network
[55] https://developers.cloudflare.com/workers/platform/limits#worker-limits
[56] https://archive.is/mzq1a
[57] https://archive.is/huBLk