Appendix F: the role of data in digital advertising

Introduction

1. This appendix describes types and sources of data collected by online platforms and how these are used to provide digital advertising in both search and social media platforms.

2. We have sought to understand whether access to data or certain types of data may confer a competitive advantage to large platforms and inhibit entry and expansion by smaller platforms on both sides of the market – the user and advertising sides. The role of data on the user side is discussed in Chapter 3.

3. This appendix draws on academic literature, submissions and documents from market participants to assess the following topics:

    - types of data;
    - sources of data;
    - the role of data in digital advertising; and
    - the value of data.

4. Our view on the role of data in digital advertising is summarised below:

    - Data gives platforms a competitive advantage in the provision of digital advertising. Platforms provide targeting capabilities which allow advertisers to retarget their current customers and also to target potential new customers. For these purposes, detailed data on consumers' demographics, interests, preferences and behaviours is most valuable in terms of profiling consumers, predicting consumers' potential response to advertising and tailoring advertising messages.

    - Platforms also provide verification and attribution services to advertisers. For this purpose, platforms' ability to collect data, beyond their own consumer-facing services, from third-party sites and apps, and to combine it with analytics data to present a unified view of campaign performance to advertisers, is very important to demonstrate their effectiveness in digital advertising.

    - Google and Facebook have a competitive advantage in terms of being able to carry out attribution accurately for campaigns that advertisers run,

at least in part, on their own ‘walled garden’ platforms. At the same time, Google and Facebook's tracking solutions are widely distributed across many other websites. Rival platforms such as Microsoft and Amazon have access to some detailed high-quality data about consumers and to other types of data, but this does not extend widely to the rest of the internet. This more limited access to data in terms of quantity and/or quality of analytics data on Google and Facebook’s properties constitutes a barrier to entry and expansion (although the extent to which may vary by sector) for smaller rivals in the provision of personalised advertising. In particular, lack of access inhibits independent providers of attribution services, and it could make it more difficult for advertisers to compare the relative performance of ads on Google and Facebook against ads on other websites and apps. Therefore, restrictions on third-party access to granular analytics data on Google and Facebook’s properties give Google and Facebook a competitive advantage.

- This finding has several implications for the role of data in digital advertising that we discuss and assess throughout the report. In Chapter 5, we discuss the extent to which data, coupled with other barriers to entry and expansion, impedes effective competition between smaller platforms and Google and Facebook. Chapter 6 discusses how platform’s data advantages may lead to weaker competition and poor returns to consumers. In Chapter 10, we set out the next steps the CMA will be taking in relation, among other things, to data availability and data protection. In Appendix K, we review the choices available to consumers to control their data and in Appendix X we evaluate potential interventions to allow consumers a choice over whether to receive personalised advertising. Lastly, in Appendix Z we present remedy options aimed at reducing or eliminating the competitive advantage that data confers to large platforms.

Types of data

5. A large amount and variety of data is collected online by a wide range of market participants, including platforms, advertisers, publishers and data brokers. This data is not homogeneous, but diverse in content and nature, and its usefulness and value depend on the type of data and the use to which it is put. In order to understand whether and to what extent data is a barrier to competition and the various privacy-related concerns that may arise, we have assessed the different types (in this section) and sources of data (in the next section) that platforms have access to.

6. Data can be classified along various dimensions. For example, it is possible to distinguish data based on its source, the type of information it conveys or the way it is produced. These dimensions can also be combined to identify different subcategories within broader categories.

7. We have drawn from existing literature, stakeholder views and previous reports to identify four broad categories of data for our purposes, namely:

   - user data;
   - contextual data;
   - analytics data; and
   - search data.

8. These categories are not mutually exclusive or exhaustive, and there are many grey areas in which any given kind of data may be used for different purposes and could be included in several of these categories.

9. We consider this is the most useful categorisation to assess the roles that specific types of data play in platforms' market power in the provision of digital advertising. We also consider that this classification is helpful to frame and assess concerns related to consumer privacy.

User data

10. User data refers to all the data that conveys information about consumers' behaviours and their attributes. This includes consumers' age and gender, search queries, and various types of content they share on social media platforms.

11. Market participants collect user data in various ways, which can be grouped into three sub-categories:

    - Volunteered data – information which is intentionally provided by the data subject. For example, in a social media platform context, this includes information that consumers provide when creating or updating their profiles (eg date of birth, gender, email address, mobile phone number, declared interests), but also their posts, photos, comments, etc. In search, it includes users' search queries.
    - Observed data – information which is recorded about the person and what they do. Examples include consumers' browsing history, time

spent and clicks performed on a webpage, time of the day of log-in and log-out, groups joined and friendships on social networking platforms. Observed data also includes data derived from users' devices (so-called 'device data'), such as type of device (eg desktop vs mobile), operating system and its version, browser and IP address. Market participants can also collect observed data when consumers are not actively using their services. Depending on the privacy permissions set by the consumer, mobile applications, for example, can be set to record and send to the platform the device's location at regular time intervals even if the application is running in the background.

- Inferred data – refers to additional information about the person, not directly provided by or observed from the person, but which is derived or deduced from this information.[1] This process combines volunteered and observed data about one consumer and about other consumers to infer additional information about that one consumer. For example, a user's IP address can be used to infer their location. In turn, this can be combined with census demographic information to infer characteristics such as education, income and ethnicity. Empirical research shows that it is possible to infer a large number of user attributes with satisfactory levels of accuracy, including some complex ones such as personality traits.[2]

12. Another relevant categorisation for our purposes distinguishes user data between personal and non-personal data. Personal data is a wide concept under relevant data protection legislation (such as GDPR and DPA 2018) that includes any information about natural persons who can be identified, either directly from the information, or indirectly from using that information in combination with other information.[3] A person can be identified if they can be distinguished from other individuals. Online identifiers such as IP addresses, cookie IDs, advertising IDs, email addresses, user accounts, and device fingerprints (ie highly specific information about the combination of versions

[1] This distinction is relevant to the right to data portability under GDPR, which allows individuals to obtain and reuse their personal data for their own purposes across different services. It only applies to information that an individual has provided (volunteered) or data resulting from observation of an individual's activities (observed). It does not include any additional personal data that has been created from volunteered or observed data. ICO, Guide to the general data protection regulation, pp.128-129.

[2] See Kosinski, M, Bachrach, Y, Kohli, P, Stillwell, D and Graepel, T (2014), 'Manifestations of user personality in website choice and behaviour on online social networks', Machine learning, pp357-380; Matz, SC, Menges, JI, Stillwell, DJ and Schwartz, HA (2019), 'Predicting individual-level income from Facebook profiles', PloS one, p.e0214369; Volkova, S, Bachrach, Y, Armstrong, M and Sharma, V (2015), 'Inferring latent user properties from texts published in social media', Twenty-Ninth AAAI Conference on Artificial Intelligence.

[3] ICO, What is personal data?

and settings on a person's electronic device) can all be personal data in certain contexts.[4]

13. Another useful way to categorise user data distinguishes between demographic and behavioural data. Demographic data refers to information about the consumer such as age and gender, which is usually voluntarily provided by consumers when, for example, creating an account. Behavioural data includes information about consumers' interests, preferences and behaviours. This may be volunteered data in the form of eg declared interests, observed data when platforms collect data about users' search history and clicks on websites, or inferred data when derived from information about other consumers.

14. This latter classification is most useful when considering the relative competitive positions of platforms in relation to the amount of data they collect and use (as described in the 'Concentration and quality of data' section below) and the role of data in targeting advertising.

Contextual data

15. Contextual data refers to data on the context in which an impression is served or a consumer is making a query. For instance, it can relate to the content of the webpage on which the impression is shown, the natural meaning of the keywords the consumer inputs in a query, or information about external factors such as weather conditions. It can also refer to the context of a consumer search such as the consumer's location and their search history (particularly their immediate prior searches).

16. As for user data, some contextual data can be personal data, if it is associated with an identifiable person. For instance, search queries and histories and location data recorded against specific users' profiles may be considered personal data within the meaning of relevant data protection legislation. However, in general contextual data involves a much more limited use of personal data, if any at all, and therefore raises fewer privacy concerns.

17. Contextual data, alongside user data, can be used to target results and advertisements to the consumer.

[4] ICO, What are identifiers and related factors?

Analytics data

18. Analytics data refers to information on the advertising campaign, such as statistics on the number of users who have seen an impression, the actions taken after seeing the impression and verification checks.

19. Some analytics data can also be personal data, if ad views, clicks, conversions and other subsequent behaviours are associated with specific identifiable individuals.

20. This data is valuable to advertisers, as described further below and in Appendix O, because it allows them to understand whether their advertisement is served to the intended audience (verification) and assess and measure the reach and success of their advertising (attribution and evaluation of effectiveness).

Search data

21. Search providers employ specific data to deliver search results that are relevant to users' search queries in several ways.

22. This includes non-user data and metadata about websites and webpages, links to other webpages on each page and the pattern or network of links on the internet, the contents of each page, and the reputation or reliability of webpages (which may include the judgements of human reviewers). Search engines also use data feeds from third parties to supplement their data from crawling and indexing, and to provide better answers to certain queries such those relating to sport scores, exchange rates and weather forecasts.

23. Search data also includes user data, such as what consumers search for, and which results, if any, they click on from a results page (click-and-query data), which is used to refine search engines' algorithms to select and order relevant results.

24. We consider the role of this data in more detail in Appendix I.

Summary of relevant types of data

25. In summary, we consider that data used by market participants can be grouped into four categories. These are summarised in Table F.1 below.

Table F.1: Data categorisation

Category	*Subcategory*	*Examples*
User data	Volunteered data	Name, email address, date of birth, declared interests, posts, photos, comments, likes.
	Observed data	Click-and-query data, clickstream data, time spent on a webpage, device/browser fingerprint.
	Inferred data	Inferred demographics, inferred interests.
Contextual data		Content of a webpage/app, location, weather conditions.
Analytics data		Number of consumers that click on an ad, actions taken after seeing the ad.
Search data	Web-crawling and indexing	
	Click-and-query data	Searches on search engines
	Data feeds	Data about webpages

Source: CMA.

Sources of data

26. Large platforms are able to gather different types of data from a wide range of sources. Understanding these sources is important to assess whether and to what extent rival platforms might be able to access the same or similar data, and the extent to which consumers understand what data is being collected about them.

27. There are many different sources and many possible ways to categorise them. We distinguish between two broad sources that platforms use to collect data: (i) data gathered from the platforms' own consumer-facing services and products, and (ii) data collected from third parties, notably those that use the platforms' services, such as advertisers and publishers.

28. The subsections below describe each of these sources in turn, focussing on Google and Facebook consumer-facing services, and identifying four specific sources of data within the broad category of data collected from third parties. Finally, the last subsection presents and discusses evidence on the concentration and quality of data of different platforms. In doing so, we draw on the academic literature and on platforms' and other parties' submissions.

Consumer-facing services and products

29. Platforms collect a wide range of data from the services and products they provide to consumers. This is first-party data that platforms collect directly from their own audiences.[5]

30. Many platforms collect data on: (i) consumer characteristics such as demographics; (ii) consumer activities such as search history, clicks, content created and shared; and (iii) location through users' device information. The

[5] First-party data is information that a business collects directly from its audience. Therefore, when the business is an online platform, data on the interactions of consumers with the online platform is defined as first-party data. Advertisers collect first-party data as well, ie data about the advertiser's audience.

amount and types of data may vary based on the context in which consumers access platforms' services and products, such as whether they are logged-in, whether they are using an app or a browser, and whether they are using a mobile or other device.

31. Major platforms such as Google and Facebook can collect large datasets from the high number of consumers that are both logged-in and not logged-in onto their array of services and the multiplicity of devices these are offered on. Being able to collect user data from different devices grants platforms access to larger quantities of volunteered and observed user data as platforms can observe a wider range of behavioural data by capturing a larger part of time spent online by multi-device users.[6] Platforms can then create more accurate user profiles by using a richer array of users' behaviours as well as provide more accurate attribution services.

32. Below we describe in more detail the data gathered by Google and Facebook through their consumer-facing services and products.

Google

33. Google collects data directly from its audience through Google consumer services and Android mobile devices.[7]

34. Google provides more than 53 consumer-facing services and products in the UK, including Google Search, and gathers data through them. This data includes:

    - User information. This data is collected only from consumers who have a Google Account and are logged-in at the time of the interaction with the service (Authenticated Users). In 2018 in the UK there were on average [30-40] million active logged-in users of Google Search on mobile/tablet.[8] This user data includes information voluntarily provided by a consumer when creating a Google Account, such as name, contact details, account authentication data (eg username and password), demographic information (eg gender and date of birth), and payment information and associated details (used for Google Pay or identity verification).

[6] Non-mobile devices (such as desktops) are often used by multiple consumers, and so the activity data on those devices may be an amalgam of different people. By contrast, mobile devices are more often used only by a single individual and therefore the data collected from mobile devices is more accurate.
[7] Google collects data also from the Internet of Things but in this appendix, we have not focussed on these devices.
[8] Monthly active users are defined as the 28-day active users as of the 28th day of a given month. This figure relates to users logged-in into a mobile device. In the same period there were on average [10-20] million monthly active users logged-in into a desktop or laptop.

- Information about the apps, browsers, and devices used to access or interact with Google services. For example, when Google services are accessed using a web browser, Google collects data on device and browser type and settings, operating system version, device event information (eg crashes, system activity, hardware settings), IP addresses, URLs (including referral terms), timestamps and cookie data. When Google services are accessed using a mobile app, data may be collected about hardware and operating system version, device event information, unique device identifiers, network operator and unique advertising identifiers, such as the Android Advertising ID (AAID) or iOS Identifier for Advertisers (IDFA).

- Information about a user's activity on Google services. For example, as consumers interact with Google services, Google collects data about their preferences, settings, interaction data (eg clicks and mouse hovers), content of a user's shopping basket, offline transactions (eg those made via Google Pay), search history, advertisements served, pages visited and YouTube watch history. In addition, Google can observe and collect more granular information about Authenticated Users such as content that a consumer creates, uploads or receives from others when using account-based Google services. This content includes emails written and received, photos and videos saved, Docs and Sheets created, and comments made on YouTube videos.

- Information about a user's location when they are using Google services, depending in part on their device settings: Google relies on various technologies to determine a consumer's location, including IP address, GPS and sensors such as accelerometers and gyroscopes. These may, for example, provide Google with information on nearby devices, Wi-Fi access points and cell towers. If Authenticated Users have Web App and Activity setting enabled, Google will save information about their activity on Google sites and apps, including associated information such as location. Google can also fetch useful information about events from other services such as Gmail and Calendar.

35. Google collects data also from mobile devices running Android, Google's own operating system, and from pre-installed apps on Android phones. Table F.2 shows an example of the detailed information collected.

Table F.2: Examples of data collected by Android

[✂]

Source: [✂]

Facebook

36. Facebook owns and operates three main services in the UK (Facebook, Instagram and WhatsApp) from which it gathers user information, users' activity and device data:[9]

- Consumers provide information in a number of ways. To join the Facebook community, consumers need to provide four basic pieces of information: name, email address or phone number, gender and date of birth. However, they can also provide other information about their residence, language, education, employment, hobbies, and favourite movies, books, and music.

- Facebook also receives information about a user's engagement with the service as a whole. This includes, for example, the Facebook Pages a consumer has liked, Facebook Groups the consumer has joined, content like posts, comments or photos that the consumer shares on the services, ads the consumer has interacted with, and location data (depending on the mobile device permissions the consumer has granted to Facebook). Facebook also receives information provided by other people about a consumer, such as when a friend of the consumer shares a photo in which they tag the consumer. In addition to the information Facebook receives regarding consumers' engagement with ads (including whether an ad was viewed, clicked, or dismissed), Facebook may also receive consumer feedback on ads regarding whether an ad was inappropriate, repetitive, or not relevant when consumers choose to provide such feedback.

- Device data collected includes device attributes (eg operating system, hardware, software versions, etc), device operations (such as whether a window is foregrounded or backgrounded, etc), identifiers (UI, device IDs and other identifiers from games, apps and accounts a consumer uses), device signals (Bluetooth signals, etc), network and connections (ISP, language, time zone, mobile phone number) and cookie data (cookie IDs and settings).

[9] Facebook said that it does not use WhatsApp account information in the European Region to improve consumers' Facebook product experiences or provide a more relevant Facebook ad experience.

Data collected from third parties

37. Online platforms also gather data about consumers and their interactions with third-party sites and apps. There are several ways in which this can occur, but we understand that the main are the following:

- data is actively shared by third parties;
- data is collected directly from third-party sites or apps through technology;
- through sign-in functionality on third-party sites or apps; and
- through advertising services on publishers' sites or apps.

Data actively shared by third parties

38. The main types of partners that provide data to platforms are:

- Advertisers – They can collect their own first-party volunteered and observed data (eg through their websites, loyalty programs, etc) to share with platforms that run their advertisement campaigns; or they can feed platforms user data they source from other agents such as data management platforms (more detail in the 'Targeting in digital advertising' section below). Many advertisers that responded to our information requests indicated that they do collect and consider most valuable the data they gather about their own customers. They also confirmed that they upload this information onto platforms in order to better target consumers and extend their reach by finding similar consumers (more detail in the 'Similar audience' section below).
- Data brokers – They mostly provide inferred data generated through their own inference processes, which draw on their own sources of volunteered, observed and inferred data. This data can be fed to platforms either directly or indirectly (eg through the data imported by advertisers). For example, Amazon procures pseudonymous demographic data from a provider on a monthly basis. This data is used to improve interest-based advertising profiles in order to assist with matching specific audiences to more relevant features, products, and services.
- Publishers – Similar to advertisers, publishers can collect their own first-party volunteered and observed data that they can then feed to online platforms (and data brokers).

39. The section on ‘Targeting in digital advertising’ below describes how this data is used to target digital advertising to consumers.

Data collected directly from third-party sites and apps through technology

40. Platforms also provide a range of services and tools that third-party providers may use on their websites and apps. These include, among others, analytics tools such as Google or Facebook Analytics, advertising services such as Google AdSense and social products.[10] Through these tools platforms can collect data relating to consumers’ activities on third-party sites and apps such as existing user or device identifiers or their interactions with their sites.

41. Advertisers and publishers can allow platforms to collect observed and volunteered data directly from their own online services through technologies such as Software Development Kits (SDKs), pixel tags and cookies. For example, Facebook partners can install such code on their websites or apps, in order to better assess the effectiveness of existing advertising campaigns, to target potential customers with future ads more accurately, and to obtain other insights about their user base. The code installed by partners provides information about consumers’ activities on their website or app – including information about device, websites visited, etc – whether or not the consumer has a Facebook account or are logged into Facebook. In a similar way, advertiser and publisher websites can also install Google Analytics, which provides measurement data on how consumers are engaging with content and ads. Through Google Analytics and other tools, Google collects a wide range of data about consumers and how they interact with third-party sites and apps.

42. In addition, many websites and apps make use of platforms’ SDKs to provide social sharing buttons, such as Facebook’s ‘Like’ and ‘Share’ buttons and Twitter’s ‘Tweet’ button, to encourage existing consumers to share on platforms and attract new consumers. Through these buttons, websites and apps send additional data concerning those users’ activities on that website or app to the platform through SDKs.

43. The evidence set out in Appendix G shows that the reach of Google and Facebook tools on third-party sites and apps is very extensive. We have found that Google is the leader in terms of coverage of websites (prevalence) but even more so if we take into account sites’ popularity (prominence). In other words, its reach is wider on most popular websites. After Google,

[10] Google also collects data from third-party sites and apps including with the following products: Google Analytics, Google Tag Manager, Google Ads, Floodlight, Google Ad Manager, AdSense, AdMob, Authorized Buyers, social products such as the +1 button and Waze.

Facebook's tools are the second most widespread trackers on the internet both on desktop and mobile. In particular, multiple studies have found that Google tags (eg Google Analytics, Google Ads and Floodlight tags) are found on over 80% of the most popular websites, Facebook has the second highest prevalence of tags, and it covers between 40-50% of the most popular websites. On mobile apps, Google has SDKs in over 85% of the most popular apps on the Play Store, and Facebook has again the second highest prevalence with SDKs in over 40% of the same. This is supported by submissions we have received during the course of the market study.

44. For example, Oracle Moat said that it is difficult, if not impossible, to use the internet without encountering Google Analytics as approximately 75% of the top 100,000 websites on the internet use Google Analytics. Channel 4 noted that Facebook's attribution tool allows advertisers to track views of their ads on Facebook users' feeds and then link this to behaviour on the advertiser's site. Channel 4 claimed that these kinds of tools place the digital giants at a huge commercial advantage as they can collect and analyse viewing data from content providers such as Channel 4 but then do not provide this data to the content provider.

Through sign-in functionality on third-party sites or apps

45. Platforms collect data when consumers sign into an app or website using their sign-in functionality, whereby consumers can securely sign into third-party apps or websites without having to create, authenticate and remember new usernames and passwords.

46. Google said that the use of this functionality does not result in Google collecting additional data about the consumer's activity in that app, but that Google stores the context under which the user authenticates, like information about the device, IP address and identifiers for the app to which the consumer has authenticated.[11] However, if consumers choose to connect their account with the third-party app to, for example, improve their experience on the app, then Google will collect data on the users held by the third-party service.

47. Equally, when a consumer accesses a third-party site or app through Facebook Login, Facebook receives data from the browser or mobile SDK (such as the IP address of the browser, the date and time the HTTP request was made, the browser type and version, etc.); a cookie file (comprised of a random series of letters and numbers that is associated with the browser); additional data that pertains to the use and functionality of the cookie (eg the

[11] Once the consumer selects the account, the app will be able to access the consumer's name, email and profile photo.

date/time the cookie was installed); and, for mobile apps specifically, a unique app or device ID. In addition, third-party websites or mobile apps may also choose to send Facebook additional data about the consumer's activities on that site or app (such as the fact that a purchase was made on their website). Facebook said that the primary purpose of this functionality is to ask users whether they want to provide their data to third-party websites and apps.

48. Google provided data showing that [0-5]% of UK websites and [10-20]% of apps accessed by UK users on Android used the Google Sign-In functionality.[12] Facebook estimated that in 2019 approximately [0-5]% of UK sites and apps are covered by its Sign-in functionality.[13] However, these figures may underestimate the reach of this functionality because, for example, Google figures exclude apps not used on Android mobile devices. Moreover, and more importantly, these figures do not take into account the popularity of sites and apps. Few studies presented in Appendix G that consider the popularity of sites found that Google and Facebook have the most reach with respect to users online on both web and mobile.

Through advertising services to publishers

49. Platforms can collect data through the advertising services they provide to other websites and apps.[14] In this way, they usually collect user data and contextual data, which can be disseminated to a large number of intermediaries and advertisers in bid requests if advertising is being sold programmatically. Platforms also collect analytics data and additional user data when providing verification, attribution and measurement of advertising effectiveness (more detail in the 'Role of data in digital advertising' section below).

50. For example, Google automatically collects certain user data when its advertising servers receive a request from a user's device. This request may be triggered by the consumer interacting either with a Google advertising service or with a third-party website or app that uses a Google advertising service. Google collects data from Google Ads, Google Ad Manager, Authorized Buyer, AdSense, AdMob, DV360, Campaign Manager and SA360. Although this may vary by Google service, publisher's settings, consumer's preferences and device used, the collected data generally includes:

[12] Google defined UK websites as ".uk" top-level domain country code and the estimate of the total number of UK websites includes websites active as of 19 March 2020 as recorded by Zonefiles.
[13] Facebook used the total number of UK businesses at the start of 2019 as a proxy of the total number of UK sites and apps. This result does not change significantly if we use the total number of UK sites and apps as recorded by Zonefiles.
[14] The role of intermediation in digital advertising is discussed in Appendix M.

- The ad request itself, such as the browser's request for an ad to be served on a non-Google website and the ad slot to be filled, including the date and time of the request;

- System and device information, such as the device, browser version, operating system version, default language and screen size, including IP address and GPS location;

- In the case of web browsers, the full URL of the page being visited together with the referrer URL. In the case of mobile devices, mobile network information. In the case of mobile applications, an identifier for the application and a resettable mobile advertising ID (such as IDFA for iOS or AAID for Android). In the case of web browsers, any cookie IDs that Google has previously set on the user's device; and

- Event data such as impression, click or conversion data.

51. Amazon also receives information from third-party publisher sites where a publisher monetises its ad inventory through Amazon Publisher Service or Amazon ad exchange. This includes information such as campaign information, ad placement information (eg placement on page, size, above/below fold), bid information (such as bid floor and CPM) collected as part of bids, impressions or clicks. Amazon also collects cookie IDs when customers use a web browser and mobile advertising IDs when using mobile devices.

Concentration and quality of data

52. This section draws on the description of the types and sources of data above and platforms' position in consumer-facing and digital advertising services set out in Chapter 3 and Chapter 5 to assess the relative competitive positions of platforms in relation to the amount of data they collect and use.

53. Google and Facebook have significant and enduring market power in consumer-facing services in search and social media, respectively. As a result, Google is the platform with the largest dataset collected, in addition to Google Search, from its leading consumer-facing services such as YouTube, Google Maps, Gmail, Android, Google Chrome and from partner sites using Google pixel tags, analytical and advertising services. A Google internal document recognises this advantage saying that 'Google has more data, of more types, from more sources than anyone else'. It then continues saying that 'Google is a big part of this scaling machine with massive reach across

the internet. [✂] Advertisers and media agencies agreed with this view and said that Google has access to vast and high-quality data.

54. Facebook has a very large audience with over 43 million unique monthly active users in the UK across its three main services, Facebook, Instagram and WhatsApp,[15] from which it collects very granular user data. Facebook said that there are many sources of high-quality user data available to advertisers beyond Google and Facebook and other large platforms, including data brokers, data management platforms and credit references agencies. We discuss some of these in Appendix G, where looking at the evidence in the round we found that large platforms have greater opportunity to track users and collect data.

55. Moreover, as shown above, the reach of Google and Facebook tools on third-party sites and apps is extensive and far greater than that of other platforms.

56. Google and Facebook also have an advantage in the collection of certain types of data:

- Through their extensive reach on third-party sites and apps, Google and Facebook collect a large amount of analytics data that, as described below, they use to provide evaluation and attribution services.

- Through its search engine, Google is able to collect a large quantity of search data, including users' search and click history. Since Google has been and is by far the largest player in search, with more than 90% of the estimated UK share of advertising revenue and of UK shares of supply by page referral in 2019, it has a significant advantage in getting access to this data (more detail in Appendix C). This data is very valuable because, as described in the 'In-market audience' section below, it allows advertisers to target consumers who are actively looking for specific products and services, which is considered a very valuable targeting tool.

- Google has also a significant advantage in relation to a specific type of user data, that is location data, which it gathers systematically and to a great level of detail from mobile devices running Android.

57. Overall, Google and Facebook collect many types of high-quality data from across the web and other sources at scale and use it to provide precise

[15] Comscore MMX MP, Total Digital Audience, Desktop aged 6+, Mobile aged 13+, February 2020, UK. See further Appendix C.

targeting capabilities and attribution services to advertisers.[16,17] As set out in Chapter 5, because of their scale and unique position in search and display advertising they have market power in digital advertising, earning significant revenues and reinforcing their market power on the user side. This in turn reinforces their advantage in data.

58. Compared with Google and Facebook, we consider that other platforms' data, targeting, and attribution capabilities are relatively limited to user data from their own services and limited reach on third-party sites and apps. This was supported by responses to our interim report, most of which agreed with our assessment that Google and Facebook have exclusive access to large amount of data, which gives them a significant competitive advantage over other market participants (see Appendix B).

59. Amazon collects a large high-quality dataset from consumers of its owned and operated services (such as its online shopping, Prime Video, Kindle, Amazon Music, etc.). However, we are of the view that this data is more limited in breadth compared to Google and Facebook, as it relates to consumers interactions in a retail environment.

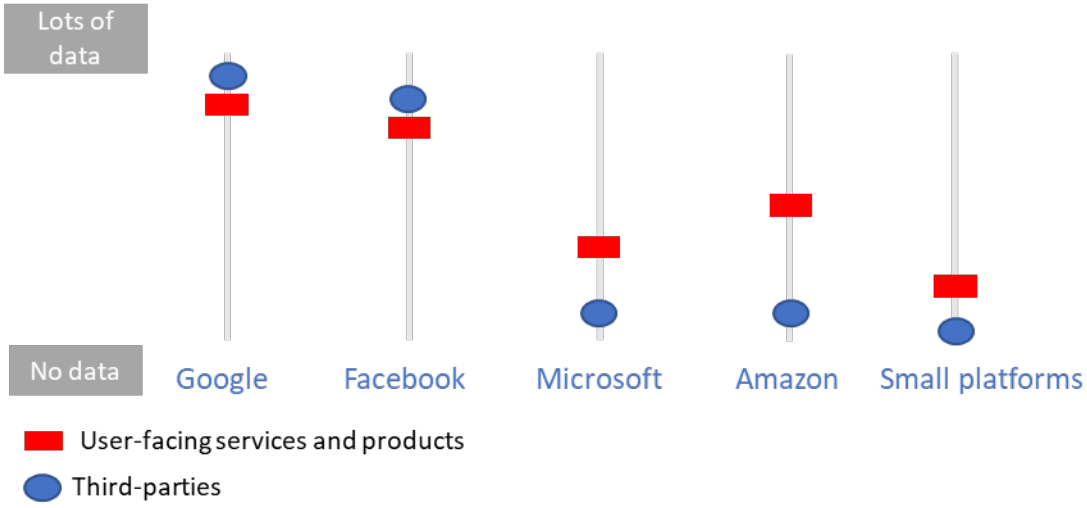
60. Microsoft collects data from consumers of its services such as Bing, LinkedIn, MSN, Xbox, and the Windows 10 operating system. However, we understand that the information it can gather from third-party sites is limited because of the limited coverage of its pixel across the internet. Although Microsoft also gathers search and contextual data through its search engine, the amount of this data it can collect is significantly limited given that in 2019 it accounted for less than 10% of the estimated UK share of search revenue and 5% of shares of supply by page referral (more detail in Appendix C).

61. Although most recent entrant platforms such as Twitter, TikTok, Pinterest and Snap possess high-accuracy data about their users, this is limited by the reach of their consumer-facing services and tags on third-party sites and apps.

62. Figure F.1 illustrates our understanding on the volume of data that certain large platforms and a group of smaller platforms possess.

[16] Since 2012 Google has pooled data it processes about individuals across its services. In June 2016 Google started to combine data from its DoubleClick business and all other Google businesses. For some of services, Google is restricted from merging data for ads.
[17] Google said that data collected from third-party sites are not used for its own purposes. However, customers of Google Analytics can choose to enable data sharing and, in this case, Google may use this data for various purposes including sales and improving Google's own products.

Figure F.1: Illustration of the scale of data collection by certain platforms, split by two broad data sources



Source: CMA.
Note: Small platforms include Twitter, Snap, TikTok and Pinterest.

63. In the next section we discuss the implications of data concentration for competition in the provision of advertising and ad verification and attribution services.

Role of data in digital advertising

64. Data has a key role in digital advertising as it is an essential input used to provide targeted digital advertising to consumers and attribution services to advertisers. Large datasets are useful in both search and display advertising, although in different ways and reflecting different advertiser objectives. Below we draw out any differences, where relevant, between search and display advertising.

65. The following subsections discuss the role of data in targeted digital advertising and in verification, attribution and measurement of effectiveness drawing on the academic literature and parties' submissions.

Targeting in digital advertising

66. Digital advertising is aimed at reaching the right consumer, at the right time and in the right context so that advertisers can achieve their campaign's objectives, such as raising brand awareness or driving specific consumer actions (eg purchase).

67. There are many types of targeting, which can be broadly grouped in two categories, contextual and personalised advertising, according to the degree of targeting and the use of user-specific data. At one end of the spectrum

contextual advertising requires relatively little data about consumers, whereas at the other end personalised advertising uses very specific user data to target advertising to each individual.

Contextual advertising

68. In contextual advertising, ads are selected on the basis of the content of the webpage or app (ie the 'context'), identified by specific keywords or topics, that the consumer is viewing, and are targeted to consumers based on aggregate demographic information about the users of those webpages or on the assumption that consumers are more likely to find ads related to the content they are viewing to be relevant.

69. Contextual advertising is applicable for both search and display advertising. For example, a consumer viewing a search engine results page or a specific webpage about running shoes may be shown ads for footwear, clothing, equipment and accessories that are relevant to running.

70. Although some search engines allow advertisers to target, alongside the context of the page, audiences (eg demographics, affinity, in-market, similar audiences, etc), content and devices (computer, mobile, etc), they primarily rely on the context (ie the search term) to target ads.[18] An experiment conducted by Google shows that the user's search query is essentially what drives Google's targeting together with information about the rough location of the user. This experiment shows that the importance of other data for targeting in search is marginal. By disabling the use of demographics, in-market and affinity and remarketing audiences monetisation decreased by [0-5]%, [0-5]% and [0-5]% respectively compared to the control group.

71. Therefore, search advertising can be thought as one of the most valuable form of contextual advertising. Contextual advertising also applies to social media platforms where, for example, a photo uploaded by a user can be considered the context within which to serve ads. In this example, platforms select ads on the basis of the photo's background, without using any user-specific data to target ads. Although this is the case and social media platforms use contextual advertising, the vast majority of advertising on these platforms is personalised advertising.

72. Contextual advertising typically uses limited user data such as search terms, device, location and language that is collected in real-time, rather than historic

[18] One media agency told us that search is not only about the keyword, but that the intent behind it, the audience associated with it are relevant as well. Another media agency said that the decision to use audience targeting in addition to keywords is normally made on more generic, top of funnel keywords as these can become very costly without additional targeting to restrict the number of less relevant impressions and clicks.

data that has been stored over time. This is one of the key features of contextual advertising and the main difference with personalised advertising that, as explained below, uses detailed historic and real-time user data.

Personalised advertising

73. Personalised advertising uses data about each specific user to display ads that might be of interest to the consumer. In order to do so, advertisers, publishers, and platforms combine multiple data collected over time from a variety of sources into profiles about consumers, which often include information about demographics, interests, home and work addresses, online and offline behaviours.

74. To build these profiles, market participants use various techniques and technologies to identify consumers, assign identifiers for them (such as cookie ID or mobile advertising ID), match these (if necessary) with the identifiers used by other participants, and share these identifiers with each other so that there is a common and mutually understood way to refer to each individual consumer. Volunteered, observed, and inferred data are recorded in user profiles, and market participants may enhance their first-party data about consumers by buying and selling data from third parties. There are significant transfers of personal data across the advertising ecosystem, in order to build up a more complete picture of individuals which helps target advertising and measure the effects of advertising (see Appendix G).

75. Platforms typically group these user profiles into ‘audiences’ characterised by a specific intent, demographic characteristics and interests, and these audience segments are then offered to advertisers as bases for targeted advertising. Any given individual can be a member of multiple audience segments. There are very many audience segments, some of which can be very granular, and advertisers can use combinations of segments to achieve highly targeted advertising.

76. The most common audience segments offered by advertising platforms are demographic, such as ‘Female’, ‘25-34 years old’, ‘Education Status: Bachelor’s Degree’, ‘Homeownership Status: Homeowners’, ‘Marital Status: In a Relationship’, and ‘Parental Status: Not A Parent’, and a large variety of interest-based segments, such as ‘Home Improvement’, ‘Pets’, and ‘Computer Hardware’. Search advertising platforms offer ‘in-market’ audience segments based on the user’s recent search queries, which are particularly valuable to advertisers, as they signal that a consumer is actively considering (or ‘in the market for’) a purchase. Some platforms also offer advertisers the ability to create ‘custom audiences’ using their own first-party data that they supply to

the platform (also known as 'retargeting'), and some platforms additionally offer to find individuals who are similar to the advertisers' existing customers (also known as 'similar audiences' and 'lookalike targeting').[19] Each is described in more detail below.[20]

Demographic audience

77. Demographic audiences enable advertisers to target segments of the population that share common traits such as age, gender and education. For example, Facebook demographic targeting targets ads to audiences based on:

- a consumer's stated location, IP address, mobile location data and / or comparative location data across different time zones. That information facilitates the targeting of ads to audiences in specific locations. Advertisers can opt to target consumers residing in a location currently, or consumers who are simply visiting a given location, or users whose home or workplace is in a given location;
- a consumer's stated or inferred age to enable age-based targeting; and
- a consumer's stated or inferred gender, education and language to enable targeting on this basis.

78. Google has recently launched a detailed demographic audience that groups people on the basis of long-term statuses such as education level, marital status, homeownership and parental status.[21,22] These details allow advertisers to refine their bidding strategies and improve efficiencies.

79. Demographic audiences are typically used when advertisers are interested in broad campaign objectives such as increasing brand awareness or brand consideration. However, they are also used to predict consumer's preferences and interests, when there is a lack of direct information on consumer's interests and behaviours.

[19] Terms such as 'custom audience' and 'lookalike audience' are not used consistently across the industry. In this appendix we define custom audience as retargeting.
[20] If users can be re-identified by market participants and associated with a user profile or their browsing history, then it is possible to show ads which are relevant to them, regardless of the content of the website or app that they are currently viewing. For this reason, personalised targeting is sometimes known as 'context-agnostic' or 'content-agnostic' targeting.
[21] About audience targeting – Google Ads Help.
[22] This audience can be used only for advertising on certain properties, namely Gmail, Discovery, Search, Shopping and Video.

Interest-based audience

80. Interest-based audiences are generated by adding people to different interest groups on the basis of data that platforms gather and infer. This includes data about consumers' characteristics as well as data on their interests and behaviours.

81. Facebook generates interest categories using [✂] taxonomy to which consumers are added based on engagement on Facebook or Instagram, including page engagement (such as likes), ad clicks and signals.[23] Google offers Affinity and Custom Affinity tools that allow advertisers to reach consumers based on a holistic picture of their lifestyles, passions and habits. Custom Affinity audiences are more tailored audiences compared to broader affinity audiences. For example, with Custom Affinity, rather than reaching a sport fan audience, an advertiser can reach avid marathon runners instead.[24]

82. In order to create interest-based segments, Amazon uses information such as searches for products or services, order history, configuration and use of settings on a device, location information, IP address, content downloaded, streamed and/or viewed, information on detail page views and account information.[25] We understand that Microsoft uses primarily data entered by consumers in a logged-in environment and infers consumer's age, gender and interests to build targeting segments.

83. These audiences are richer than demographic audiences and can predict with more accuracy consumers' interests and their likely response to advertisement. However, these are still imperfect, because consumers might not be currently in the market for a specific product or service.

In-market audience

84. Advertisers can also target consumers who are actively looking for specific products and services. Platforms use data they collect on consumers to identify patterns of behaviours in order to differentiate their interests from intents.

[23] Signals are data points used to inform ranking decisions in relation to content presented to consumers. As described above in the section on sources of data, signals are created through a user's conduct on one of Facebook's platforms (so-called 'onsite' signals), and can also be generated by consumer conduct on external platforms, for example on third-party apps (apps signals), websites (website signals) or physical stores (offline signals). The last category of signals concerns partner signals, which are generated through the integration of third-party partners with Facebook (eg Shopify).
[24] About audience targeting – Google Ads Help.
[25] Amazon said that it only obtains and processes personal data in accordance with its privacy notice.

85. This is a powerful tool that allows for very valuable targeting. In an internal document Google states:

[✂]

86. We note that this does not appear entirely consistent with the results of the experiment conducted by Google and described in an internal document at paragraph 69 above.

87. Some advertisers also indicated that in-market audiences are among the most valuable targeting tools. For example, Confused.com said that in-market audiences are very valuable as they enable it to target customers who are actively searching for its products, which results in relevant and more efficient marketing. One large advertiser said that it generally uses Google's proprietary data (such as its in-market segments) over characteristic targeting (such as demographic) as these are better indicators of interest, or its products and services.

Retargeting

88. Retargeting is a specific form of personalised advertising, which is aimed at identifying and serving targeted ads to specific individuals whom advertisers identify as customers or potential customers. Retargeting works in the following way:

- Advertisers provide platforms with hashed customer data consisting of contact lists, email identifiers or other identifiers that the advertiser has previously obtained through its own customer relationships.[26] Advertisers may have collected this data from their websites, apps, physical stores, or other situations where customers have shared this information directly. Alternatively, platforms can collect data on advertisers' customers directly from their websites through SDKs, cookies and pixel tags enabling advertisers to target these consumers.

- Platforms seek to match this customer data with information they hold about these consumers and reveal to the advertiser the number of successful matches, without revealing to the advertiser the specific individuals that have been matched.

[26] Hashing involves representing the data in characters, effectively anonymising the data by turning it into short 'fingerprints' that cannot be reversed by a third party, which protects the privacy and security of the original data.

- Advertisers can then target differently (eg display a particular version of an ad or bid a different price to show their ads) or exclude these consumers from their targeted advertising on the platforms.

- Advertisers can also use this group of customers as their 'seed' audience and expand their reach by targeting consumers who share similarities with the original seed group of customers (more details in the section below).

89. Most platforms offer this retargeting tool. For example, through Google Ads, advertisers can match their customer lists with Google accounts and retarget consumers on Search, Gmail and YouTube campaigns. Alongside the ability to target consumers on the basis of their interaction with advertisers' websites and apps, Facebook also enables advertisers to target specific audience on the basis of their on-platform behaviour such as likes to a specific Facebook Page. Amazon said that it may receive data directly from advertisers or from data management platforms at the request of an advertiser customer, in which case the segments are used only by that advertiser.

Similar audience

90. Platforms also provide a service to advertisers to help them find consumers that are similar to their existing customers. These services are sometimes referred to as 'audience extension' or 'audience expansion'. There are many techniques and methods to do so, but the basic idea for all these methods is to analyse a 'seed' group of existing customers and identify features or combinations of features that are common to many or most of the members of the seed group, and then construct a model to predict and identify which other consumers are similar to the seed group.[27]

91. For example, Facebook launched 'Lookalike Audiences' in 2013 to allow advertisers to run ad campaigns that are directed at Facebook users with characteristics similar to their existing customers or to those users who have liked an advertiser's Facebook Page. Advertisers can select a Custom Audience as their seed audience and ask Facebook to find a broader set of consumers that match the characteristics of the seed audience. Facebook will then run an analysis based on the attributes of the seed audience and, using the user data available to it, create a 'Lookalike Audience' comprising Facebook users whose attributes are most highly correlated with those of the seed audience.

[27] See, for example, this 2010 US patent for systems and methods for generating expanded user segments.

92. Google's similar audience tool finds consumers that are similar to an original remarketing 'seed' list (or other compatible list). It finds consumers that are similar in profile based on the seed list users' recent browsing interests, search queries, and videos watched on YouTube. Google 'scores' consumers based on how similar they are to consumers on the original seed list, with similarity defined as interested in same categories, topics and/or products.

Conclusions on targeting digital advertising

93. There are many types of targeting, which exploit different types and volumes of data as well as level of granularity. As described above, in general platforms have a range of targeting capabilities and advertisers choose the best according to their campaign's objectives and KPIs. Although all platforms seem to be capable of targeting consumers on the basis of high-level information such as demographic characteristics, their ability to target more specific and narrow audiences differs.

94. Several advertisers told us that Google and Facebook offer more granular and higher quality personalised targeting tools compared to other platforms. In search, many advertisers and media agencies are of the view that Google offers more in-depth targeting options, driven by its unique and vast data, compared to Microsoft. The targeting capabilities that Google offers in search are also extended to display advertising and YouTube in particular. In display advertising, Facebook has the advantage of offering the ability to target specific audiences based on demographic, interests and location. Some advertisers also singled out Facebook's remarketing capability, which has a strong match rate with advertisers' first-party data and therefore allows them to reach a large proportion of advertiser's known customers. Alongside these platforms, Amazon is also recognised as having rich and high-quality data for targeting audiences along the customer journey and in particular for driving sales. Other platforms were hardly mentioned by respondents, with the exception of Twitter, which some respondents indicated offers the possibility to reach niche and highly relevant audiences through keyword targeting and a range of ad solutions that are different to others.

Advertising verification, attribution and measurement of effectiveness

95. The second main purpose of data in digital advertising is to provide verification, attribution and measurement of effectiveness. These are discussed in detail in Appendix O.

96. Assessing and evaluating the quality of digital advertising is a process which involves a number of different stages:

- Verification: checking the viewability of the advertising, the context in which it was displayed, and identifying the potential for ad fraud. Was the advert displayed on a webpage in such a way that consumers could actually view it?

- Attribution: tracking what actions the consumer took after being exposed to the advert. For instance, did the consumer click through to the advertiser's website and buy the product after exposure to an advert?

- Measuring effectiveness: did the advertising meet the campaign objectives the advertiser had set eg did the advertising produce an incremental uplift in sales?

97. Verification and measurement of effectiveness allow advertisers, publishers and platforms to confirm whether and the extent to which ads were shown to the right number and kinds of people.

98. Verification involves the authentication of the placing of an advert and is a key starting point in being able to measure the effectiveness of online advertising. For instance, to be able to measure the Return on Investment ('RoI') of an advertising campaign, there is a need first to be able to establish that the advert has been viewed by a potential customer before moving on to evaluating what action they took as a result and what the impact on profits was.

99. The verification of digital advertising is sometimes portrayed as something which is just of concern to advertisers. However, publishers also have an interest in being able to confirm the integrity of their advertising inventory as a means of building and maintaining trust in the quality of their advertising inventory.

100. As set out in Appendix O we found that although Google and Facebook allow third-party verification of their own inventory, they place restrictions on the ability of advertisers to carry out their own independent verification and, as a result, advertisers have to rely on data that has been collated by Google and Facebook.[28]

101. Measurement of ad effectiveness and attribution of digital advertising is not straight-forward. Accurate measurement requires consistent definitions of metrics and methodologies across different advertising platforms and a number of responses from advertisers and agencies argued that a lack of

[28] Google said that its approach to ad verification and attribution is driven by Google's obligations under the GDPR as any initiative to improve the ability of third parties to measure the performance of their ads must not conflict with the requirements of data protection legislation.

standard approaches across platforms made it difficult to measure the impact of advertising on a consistent basis.

102. Attribution is aimed at identifying a set of consumers' actions often across websites and devices that contribute in some way to a desired advertising outcome and then assigning value to each of these actions. For this reason, attribution often requires the matching of data on consumers' exposure to adverts with data on the subsequent consumers' actions. The consumers' actions that are most often monitored are customer purchases, but such actions can also be defined more broadly depending on advertisers' objectives, eg spending a specific amount of time on a website, a specific action taken on a webpage, or a store visit.

103. Advertisers may also be interested in measuring the impact of ads on other things, like brand awareness and positive brand sentiment. However, it is often more difficult to conduct attribution analyses for these outcomes because these are not directly observable actions by consumers and require the use of techniques such as consumer surveys. To measure conversions, advertisers need to be able to track consumer actions online (and to some extent offline). 'Walled garden' platforms are able to combine the ability to accurately monitor conversions with the ability to track users across different devices and sessions and so attribute consumers' actions more accurately. In the case of Google, it is able to track users across more than 53 consumer-facing services which it owns. Its access to mobile data from Android also gives it some ability to monitor the actions of consumers offline (eg to identify store visits).

104. In the case of Facebook, its single-user login feature[29] gives it a significant measurement advantage over more standard cookie-based approaches in that it is able to associate all exposures and conversions across devices and sessions with a specific user as opposed to a browser on a laptop which could be shared between different people in a family. Facebook also receives data from consumers activities across a wide range of other websites and apps when those services are using Facebook Business Tools.

105. The ability to measure the effectiveness of advertising is an important driver of advertisers' decisions on how to allocate their advertising spend across publishers and platforms.[30] Google and Facebook have an advantage in terms of being able to track consumers across their own walled garden 'ecosystem' and across a large number of third-party sites and apps. As a

[29] Facebook requires users to log-in to Facebook each time they access the service on any device and browser.
[30] Measuring the effectiveness of advertising is aimed at assessing whether advertising met the campaign objectives.

result, they are better able to demonstrate the effectiveness of using their platforms relative to others. This finding is supported by advertisers' submissions and responses to our interim report. For example, Beeswax, a DSP, submitted that Google had an advantage in measuring conversions from the data (both ad and non-ad data) it collected from its consumer products (see Appendix B).

106. We noted in our interim report Google's decision to prevent the DoubleClick user IDs being accessed by advertisers and agencies. In 2018, Google restricted access to its User IDs (the DoubleClick ID) by removing it from its Campaign Manager and DSPs log files and curtailed the availability user-level exposure data from ad campaigns. This meant that ad buyers could no longer extract data from the DoubleClick Campaign Manager for reporting on ad performance and ad attribution.

107. Google indicated that the DoubleClick ID could be tied to sensitive information like user search histories and could violate the strict data privacy requirements of GDPR.

108. Stakeholders on the buyer side suggested that stripping out the DoubleClick ID removed visibility about user activity within the DoubleClick ecosystem and made it almost impossible to compare ad performance between ads purchased through the Google adtech stack and ads purchased through other intermediaries. It was also suggested that the change made independent ad attribution much more difficult.

The value of data

109. The extent to which data is important and a driver of advertisers' and publishers' choices of platforms and intermediaries depends on the value of data. This is positive if advertisers can use data to improve the efficiency of their advertising and affect publishers' and intermediaries' revenues. The value of data may be different for different types or sources of data and for different types of advertising, and, if there is differential data access, may have competition implications.

110. There are three main reasons why it is helpful to measure how valuable or useful data is:

- Firstly, in order to understand the extent of competitive advantage that access to data or to certain types of data may confer to platforms and the extent to which this constitutes a barrier to effective competition. This may depend on whether similar data is available from other sources (see

Appendix M for a description of data management platforms), how data flows and whether it is easily shared between market participants (see Appendix G), as well as data attributes such as freshness and velocity.[31]

- Second, understanding the value of data will also help us assess the impact of potential future changes to data protection regulation on the revenue that can be earned through digital advertising. This is particularly relevant, for example, to understanding the impact of Google's announcement that Chrome browsers will stop support for third-party cookies in the future, restricting the ability of publishers to sell personalised advertising.

- Third, in order to make informed decisions about remedies options that may change the availability of data to market participants and their ability to provide and improve user-facing and digital advertising services. The value of personalised advertising is, for example, relevant to our assessment of potential intervention to allow consumers a choice over whether to receive personalised advertising, as discussed in Appendix X. To the extent that certain categories of data are an important input in targeting digital advertising, to which platforms and intermediaries have unequal access, the objective of data-related remedies would be to level the playing field between market participants. Remedy options are discussed in Appendix Z.

111. This section discusses the evidence on the value of data and, in particular: (i) the value of personalised advertising relative to contextual in display advertising, (ii) the value of different types of data, (iii) the value of measurement and attribution services, and (iv) the incremental value of additional data. In doing this assessment, this section draws on the academic literature and on parties' quantitative and qualitative submissions.

Value of personalised advertising in display

112. Despite measurement and attribution challenges, the academic literature seems to concur that personalised advertising is effective and useful to advertisers (see Appendix O).[32] This is also supported by Google internal

[31] In its statement of scope response, Google submitted that the role of data in digital advertising is indeed a fundamental question. It submitted that the value of a particular type of data may depend on its usefulness (measured against criteria such as variety, velocity, volume, and value); whether similar data are available from other sources; whether consumers can port their data between services; how the data is used; and restrictions on data use.
[32] See Marotta, V, Vibhanshu, A and Acquisiti A (2019), 'Online Tracking and Publishers' Revenues: An Empirical Analysis'.

documents, one of which, used to pitch YouTube advertising services to advertisers, says, [✂]

113. At an aggregate level, recent empirical evidence consistently finds that publisher revenues from display advertising increase as a result of personalised advertising as opposed to contextual ads. However, the magnitude of this impact in recent academic work is unclear.

114. For example, a recent paper from Marotta et al found that publishers' revenues increased by a small margin (4%) when user-specific data was used compared to when consumers could not be identified and targeted due to their cookie settings [33] However, the paper is based on a non-randomised observational design and relies on a relatively narrow set of user characteristics to account for selection bias. We expect that evidence from randomised experiments would be better able to address bias deriving from the comparison of users navigating with and without cookies. Google has recently run its own experiment to test this result and we discuss it below.

Google randomised controlled trial

115. In the summer of 2019, Google ran a randomised controlled trial (RCT) on display ads served by Google's adtech services on non-Google sites.[34] The primary purpose of the RCT was to assess the impact of disabling cookie information (ie as if the user's browser had disabled third-party cookies) on publisher revenues from display ads.[35]

116. Google made the UK data from the experiment available to the CMA.[36] We have replicated the original analysis and expanded it to account for potential selection issues.[37] More details can be found in the annex at the end of this appendix.

---

[33] See Marotta et al (2019).
[34] The experiment is described here and the methodology is described here.
[35] Blocking access to cookie information was achieved in two ways. For bid requests going to non-Google DSPs, the publisher cookie ID was simply removed. For bid requests going to Google as a DSP, [✂]. In both cases, the affected user visit was de facto treated as if it was a brand-new cookie that had just surfaced and had never been seen before. Non-cookie traffic was processed as traffic with no cookie, and this affected the treatment and control arms in the same way.
[36] Google run the RCT between May and August 2019. At the CMA's request, Google preserved and submitted the data associated with the experiment that had not already been deleted in Google's ordinary course of business prior to the CMA's request. The data submitted to the CMA covers the impressions of a random sample of UK-based web-users in the period from 21 June to 23 September 2019 that were served display ads via inventory that was either a) sold by publishers using Google's supply-side platform (SSP) solutions (either Google AdSense or Google AdX), or b) bought by advertisers using Google's demand side platform (DSP) systems (either Google Ads, Google Dv360, or Google Consumer Surveys), or c) both.
[37] The CMA has identified a number of selection issues that might bias the estimates of the effect of blocking cookies on publisher revenue: (1) The treatment blinds Google DSP selectively; (2) the treatment reduces number of impressions per query; (3) the treatment increases queries without impressions and (4) the treatment

117. The results of our analysis show that blocking third-party cookies decreases short-run publisher revenue by 70% of the average revenue per query in the control group, which approximates business as usual during the study period. Revenue from iOS users, and users browsing on Safari and Firefox, are less impacted by the blocking of cookies than Chrome users, which is unsurprising since Safari and Firefox both currently block third-party cookies already.

118. However, these results need to be interpreted with caution as there are some important effects that the analysis cannot capture. In particular, the analysis is unable to answer the question of what the long-run, market-wide effects of the removal of third-party cookies throughout the entire ecosystem would be. This is because advertisers, platforms and publishers would be expected to respond to this change in ways that are difficult to predict. For example, over time privacy-enhancing technologies, as discussed in Appendix G, could become an effective substitute for personalised advertising and attribution using user data.[38] The change could also lead to the development of more sophisticated approaches to contextual advertising as a substitute for personalised advertising. Further it could also increase reliance on first-party data for targeting.

119. However, this experiment clearly indicates that, in the short run, unequal access to third-party cookies and the detailed user information associated with them has a significant negative impact on those publishers who cannot sell personalised advertising when competing with those who can.

Publishers' analyses

120. Some publishers have also carried out analysis that sheds light on the value of personalised advertising. These compare the revenue publishers generate from ads on browsers where third-party cookies have been removed (Safari and Firefox) to revenue generated from browsers where third-party cookies are still enabled. In particular:

- News UK analysis shows that, since Firefox removed third-party cookies in September 2019, News UK's monthly revenue generated through Firefox was down by [50-60]%. The impact of the removal of third-party cookies from Safari, that occurred in September 2017, has been gradual over a long period of time, but the difference in News UK's revenue from

increases impressions served by non-Google SSPs. For most of these, the CMA implemented econometric solutions to better estimate the effect of blocking cookies access from the perspective of publishers. More details in the annex below.

[38] Indeed, as discussed in Appendix G, Google has indicated that it would modify its approach (and may delay or suspend) its deprecation of third-party cookies on Chrome if it is not confident that effective privacy-enhancing substitutes for personalised targeting and attribution can be found.

earlier versions of Safari (where the third-party cookie remains intact) compared with the latest versions paints a similar picture to that in respect of Firefox. News UK analysis shows that the value of inventory on Safari, after the removal of third-party cookies, is much lower than on Chrome.

- Similarly, The Telegraph Media Group has seen a significant decline in the value of its inventory following the removal of third-party cookies from Safari, with a difference in CPM between Safari and Chrome of [50-60]%.

- DMG Media finds a [70-80]% lower revenue per page across Safari and Firefox compared to other browsers where third-party cookies are still enabled. This is consistent with the results of another analysis with which DMG Media looks at the difference between personalised inventory and non-personalised inventory (as a result of users not giving their consent to the placements of cookies). The results are similar to the previous analysis and show that DMG Media earned [60-70]% lower revenue per page for non-personalised inventory compared to personalised inventory.

121. These publishers are concerned that the announced removal of third-party cookies from Chrome could have a similar negative impact on their revenue.

122. As mentioned above in relation to the Google RCT, these studies only capture short-run effects and cannot account for long-term equilibrium changes, such as the greater development and use of contextual advertising. For this reason, we would expect that the long-run effect of restricting access to user data for personalised advertising would be lower than these short run estimates.

123. Overall, the academic literature, evidence from stakeholders and our own analysis of the Google RCT data indicate that personalised advertising increase publisher's revenue as opposed to contextual advertising, when both are available.

Value of different types of data

124. Evidence suggests that different types of advertising and targeting, which rely on different types of data, vary in their impact on the outcomes advertisers are interested in.

125. However, consumer-specific data appears less valuable in search advertising than in display. Google said that many search queries are not affected by personalisation signals, even if historic data is available. Similarly, one large advertiser said that first-party data is less useful, and they rely more heavily on third-party data, for example they use characteristics such as location to ensure they serve ads of relevant products to the UK. Nonetheless, data can

still be very useful as noted by WPP, which indicated that while search advertising is driven by intent (the keyword), it normally needs to be augmented with audience targeting. WPP said that it can leverage native targeting signals such as demographic and location data, and their client's own first-party data (eg visits to key pages on their website), to target specific audience groups. This is supported by some Google research showing that the use of remarketing lists for search ads audience has on average a [✂] higher click-through-rate (CTR) and a [✂] higher completed-view-rate (CVR) when compared to non-audience targeting.

126. The value of consumer-specific data in display advertising appears to be much higher. This is not surprising as display advertising reaches consumers who are not 'in-market' – ie consumers who are not looking for specific products/services but are looking for different online experiences (eg connect with friends on Facebook, watch videos on YouTube). Data allows platforms to construct and update rich user profiles in real time (see section on 'Targeting in digital advertising' above) and find people who are most likely to respond positively to ads. This is supported by some empirical research which shows that targeted impressions present significantly higher click-through and conversion rates than non-targeted impressions, with consistently higher costs per impression for advertisers.[39] For example, the results of the Google research mentioned in the paragraph above show that the use of retargeting leads on average to [✂] higher CTR and [✂] higher CVR when compared to non-audience targeting and that Similar Audience leads on average to a [✂] higher CTR and [✂] higher CVR when compared to non-audience targeting.[40]

127. There are certain categories of data that are considered more valuable than others, but this may vary by sector. For example, we have heard that in the insurance sector the most valuable data is the renewal date because this indicates when customers are in-market. Many advertisers said that data about their own audiences (advertisers' first-party data) is the most important as it is unique to them and their proposition. Several respondents mentioned that age, gender, location and interests are valuable. For example, McDonald's view is that age, interests/passion points and gender data can be mapped onto the intended target audience of a particular campaign, whereas location allows targeting based on proximity to a restaurant. Generally, a mix of data points are used across all campaigns with demographic targeting the most important. We have also heard that the value of data also depends on

[39] See Beales, H (2010), 'The value of behavioral targeting', Network Advertising Initiative; Yan, J, Liu, N, Wang, G, Zhang, W, Jiang, Y and Chen, Z (2009), 'How much can behavioral targeting help online advertising?', Proceedings of the 18th international conference on World wide web, pp261-270.
[40] These are global statistics that do not refer solely to the UK.

the position along the 'marketing funnel'. Although this is also affected by the campaign's objective that advertisers want to meet, types of data may be more valuable the closer they are to the bottom of the funnel:[41]

- Data indicating consumers' purchase behaviour is very desirable for those advertisers targeting conversions. In this sense, search data is the most accurate. As explained by Verizon Media in response to our interim report, search data is useful to advertisers as a source of purchase intent and, as a result, it is the most valuable data in the advertising market as a whole. Nonetheless, previous purchases combined with current intent signals result in high-level intent data indicating whether a consumer is close to a conversion (ie a purchase or other desired action). This data type is near the bottom of the 'marketing funnel' and is highly valued.

- Slightly more removed from data related to immediate purchases are data points that indicate consumers who are in-market, ie who have demonstrated a strong intent towards a product by navigating to a product page, adding a product to their cart, or filling out a quote request.

- Even further away in the marketing funnel is interest-based targeting, ie consumers who have demonstrated some level of interest in a product or idea but not strong enough to assign them to the in-market category. Examples of this behaviour are consumers who are reading blogs, articles or product reviews, who are surfing a hobby or fan site, who are reading industry news, etc.

- Demographics data related to a consumer's general income, region (eg rural or urban), or industry type is of similar value as low-level interest data.

- The value of location data may vary significantly. Broad-based location data, such as a postal code, is helpful to narrow down the gap of desired consumers. However, location can also be very specific (eg Wi-Fi-triangulated data within a shopping mall or barometric pressure that might indicate the exact floor within a mall at which the customer finds itself). Based on such data, advertisers can target consumers who are in the immediate vicinity of their stores. Such data is as valuable as high-level intent data described above. In response to our interim report, Oracle Moat highlighted the significant value to advertisers of location data.

[41] For advertisers who want to eg raise brand awareness the value of very detailed data such as consumer purchase behaviour or whether consumers are in-market is lower than for advertisers who are aiming to increase user's purchases of their products.

128. In summary, we have found that the value of different types of data may vary according to the setting such as the types of digital advertising and the sector in which such advertising is delivered. Nonetheless, at a high level, types of data that are closer to the bottom of the marketing funnel, such as purchase intent data, are valued more than data, such as demographic data, that are more removed.

Value of verification and attribution services

129. As discussed in the section on advertising verification, attribution and measurement of ad effectiveness above and in Appendix O, data is essential to provide these services to advertisers. In response to our interim report, some advertisers have expressed concerns about the lack of sufficient data that the large platforms release to advertisers and that this impedes to reach their own conclusions on the effectiveness of their campaigns (see Appendix B).

130. A study conducted by LinkedIn provides further evidence. This study shows that the use of LinkedIn conversion tracking was associated with a substantially faster increase in advertiser spend vis-à-vis non-users of conversion tracking, because the tracking enables advertisers to optimize campaigns and better understand the value being driven by their spend. [✂] Other case studies conducted by Microsoft show the benefits provided when advertisers implement their tracking technology to take advantage of the features that it enables, such as remarketing or enhanced cost-per-click (CPC) bidding. For example, Microsoft conducted one such study with Air France and found that Air France reduced its CPC by 26% and increased sales by 43%.

131. Although limited, this study, alongside views expressed by stakeholders during the course of our study, provides a useful indication of the overall value of verification and attribution services.

Value of incremental data

132. An important feature of data that might affect its value and, as a result of platforms' differential access to data and certain types of data, platforms' competitive advantage is scale. The higher the incremental value of additional data, the greater is the competitive advantage that large platforms are likely to enjoy. This would also hinder the ability of smaller platforms to successfully enter and grow into digital advertising.

133. In 2016 Google changed its privacy policy allowing itself to combine DoubleClick data with users' names and personal identifiable information that Google had previously collected from Gmail consumers and its other Authenticated Users. Google said that this change enabled it to improve ad personalisation and measurement as well as provide greater transparency and control to its users. In an internal document Google says that this change 'allow us to treat them [users] as one consistent identity whenever and wherever we see them' and continues saying 'we believe this is a pro-user, pro-privacy evolution, and it's one we need to make in order to remain competitive in the display ads business'. It further sets out in its 'commercialisation strategy' that in phase one of this change Google will provide cross-device measurement and targeting [✂]'.

134. This indicates that the value of incremental data is positive as by increasing the information available about one consumer platforms can target consumers more accurately.

135. This is supported by some of the academic literature, which suggests that the combination of data on the same consumer increases the value of such set of data with respect to the sum of values of the individual pieces of data.[42] In other words, there are economies of scale and scope in data whereby larger datasets lead to greater precision and smaller prediction errors. However, the academic literature also finds that there are diminishing returns to scale in data such that adding extra information does not improve the predictive power after a certain threshold.[43]

Conclusions on value of data

136. The academic literature as well as the evidence we have collected suggest that data is valuable to advertisers, in that it allows them to better target consumers, track attribution and improve the efficiency of their advertisement; and to publishers, as they can earn greater revenue than by providing contextual advertising, when competing with other publishers who can sell personalised advertising.

137. Data is not homogeneous and, as such, its value depends on the types of data and the setting in which advertising is delivered. Nonetheless, it is still the case that more detailed data able to provide direct signals on consumers' purchase intents is considered more valuable and useful to provide personalised advertising. Other factors that have an impact on the value of

[42] See Matz, S.C., Menges, J.I., Stillwell, D.J. and Schwartz, H.A., 2019. Predicting individual-level income from Facebook profiles. PloS one, 14(3), p.e0214369.
[43] See Tucker, 2019. Submission on data in UK advertising markets.

data are quantity and variety of data. The evidence indicates that there are economies of scale and scope in data. Although greater insights can be extracted by increasing data size, these are diminishing, but it is unclear at what threshold this occurs.

Conclusions

138. Data gives platforms a competitive advantage in the provision of digital advertising. Platforms provide targeting capabilities which allow advertisers to retarget their current customers and also to target potential new customers. For these purposes, detailed data on consumers' demographics, interests, preferences and behaviours is most valuable in terms of profiling consumers, predicting consumers' potential response to advertising and tailoring advertising messages.

139. Platforms also provide verification and attribution services to advertisers. For this purpose, platforms' ability to collect data, beyond their own consumer-facing services, from third-party sites and apps, and to combine it with analytics data to present a unified view of campaign performance to advertisers, is very important to demonstrate their effectiveness in digital advertising.

140. Google and Facebook have a competitive advantage in terms of being able to carry out attribution accurately for campaigns that advertisers run, at least in part, on their own 'walled garden' platforms. At the same time, Google and Facebook's tracking solutions are widely distributed across many other websites. Rival platforms such as Microsoft and Amazon have access to some detailed high-quality data about consumers and to other types of data, but this does not extend widely to the rest of the internet. This more limited access to data in terms of quantity and/or quality of analytics data on Google and Facebook's properties constitutes a barrier to entry and expansion (although the extent to which may vary by sector) for smaller rivals in the provision of personalised advertising. In particular, lack of access inhibits independent providers of attribution services, and it could make it more difficult for advertisers to compare the relative performance of ads on Google and Facebook against ads on other websites and apps. Therefore, restrictions on third-party access to granular analytics data on Google and Facebook's properties give Google and Facebook a competitive advantage.

141. This finding has several implications for the role of data in digital advertising that we discuss and assess throughout the report. In Chapter 5, we discuss the extent to which data, coupled with other barriers to entry and expansion, impedes effective competition between smaller platforms and Google and

Facebook. Chapter 6 discusses how platform's data advantages may lead to weaker competition and poor returns to consumers. In Chapter 10 we set out the next steps the CMA will be taking in relation, among other things, to data availability and data protection. In Appendix K we review the choices available to consumers to control their data and in Appendix X we evaluate potential interventions to allow consumers a choice over whether to receive personalised advertising. Lastly, in Appendix Z we present remedy options aimed at reducing or eliminating the competitive advantage that data confers to large platforms.

Annex

142. In the summer of 2019, Google ran a randomised controlled trial (RCT) on display ads served by Google's adtech services on non-Google sites across the internet. The primary purpose of the RCT was to assess the short-run impact of disabling cookie information (ie as if the user's browser had disabled third-party cookies) on publisher revenues from display ads. As part of the Online Platforms and Digital Advertising Market Study, Google made the data from the experiment available to the CMA.

143. This annex presents the results of our analysis of the Google experiment. We first describe the experiment and the data collection process. We then detail some potential selection issues that might arise in this setting. Next, we aim to replicate Google's own treatment effect estimate. We then expand its scope by addressing potential sample selection issues using econometric approaches and explore the heterogeneity of treatment effects using a machine learning approach. Finally, we reflect on what the results of this experiment can (and cannot) tell us about the value of data in adtech.

The experiment

144. Google ran a randomised controlled trial (RCT) on display ads served by Google's adtech services on non-Google sites across the internet, covering a small proportion of global traffic, from May to August 2019. The primary purpose of the RCT was to assess the short-run impact of disabling cookie information (i.e. as if the user's browser had disabled third-party cookies) on publisher revenues from display ads. The results of that experiment were described in a short paper published by Google.[44]

145. Following our request, Google preserved and submitted the data associated with the experiment for UK-based web-users that had not already been deleted in Google's ordinary course of business prior to our request, and extended the duration of the experiment to September 2019. The data submitted us covers the impressions of a random sample of UK-based web-users in the period from 21 June to 23 September 2019 that were served display ads via inventory that was either a) sold by publishers using Google's supply-side platform (SSP) solutions (either Google AdSense or Google AdX), or b) bought by advertisers using Google's demand side platform (DSP) systems (either Google Ads or Google Dv360[45])), or c) both.

[44] *Effect of disabling third-party cookies on publisher revenue*, available here; and *Next steps to ensure transparency, choice and control in digital advertising*, available here.
[45] There is also a negligible amount of traffic from Google Consumer Surveys.

146. Google randomly sampled UK users, identified by a cookie ID (user ID), and randomly assigned user IDs (and all queries and impressions associated with each user ID) either to an intervention (treatment) group or to a control group. The '*control*' group was representative of 'business as usual' transactions in the stack. For traffic in the '*treatment*' group, Google Systems were not able to access the cookie information, while keeping all other conditions unaffected.

147. Google stated that the blocking of cookie information was achieved in different ways, depending on the role played by Google in the transaction. This situation is summarised in Table F.3.

- For impressions served by Google's SSP, the cookie ID was simply removed from any bid request passed over to all DSPs (both Google and non-Google DSPs).

- For impressions served by other (non-Google) SSPs, Google could not prevent the cookie ID from being included in the bid request; in such cases, Google blocked its own DSPs from accessing any information associated with the cookie ID that might have been present in its repositories. This 'blinding' implies that the Google DSPs may have been at a disadvantage compared to other DSPs, which were able to exploit any information associated with the cookie ID at their disposal in setting their bids.

- Impressions served where Google was neither the SSP nor the DSP of the winning bid (ie outside the Google adtech stack) are not recorded – as Google cannot observe them.

Table F.3: Cookie information in Google RCT

	Google role	*Google has cookie info*	*Third-party DSPs have cookie info*
A	DSP + SSP	No	No
B	SSP only	No	No
C	DSP only	No	Yes
D	Neither	Not recorded	Not recorded

Source: CMA, Google.

148. Importantly, the experiment excludes a fraction of the users who were navigating while logged-in to their Google Account. During the experiment, Google had been rolling out the use of Google Account data for display impressions sold via Google SSP solutions. The users for whom this functionality was used were excluded from the sample submitted to us. They account for [20-30]% of the total traffic seen by Google SSP solutions as part

of the experiment.[46] These users might differ from the users that can be observed in this study in terms of characteristics, browsing habits, and the revenue they generate for publishers. In this sense, the sample is not representative of traffic through the Google stack from UK users.[47]

149. Another aspect to be considered concerns ad fraud and the use of cookies as quality signals. A large proportion of traffic, particularly from Safari and Firefox with tracking prevention enabled, is cookie-less regardless of the experiment. DSP algorithms are already used to bidding on such cookie-less queries. However, for non-Safari or Firefox queries, the lack of cookie ID in a bid request could increase DSPs' probabilistic assessment that the request originates from non-human (ie 'bot') or otherwise invalid traffic. The intervention might thus lead to lower bidding by DSPs not only due to their inability to access user profiles linked with a cookie ID, but also due to increased likelihood of traffic being deemed of poor quality or potentially fraudulent. The data does not allow to put a size on this concern, but neither Google nor the CMA expect this to play a large role in the analysis.

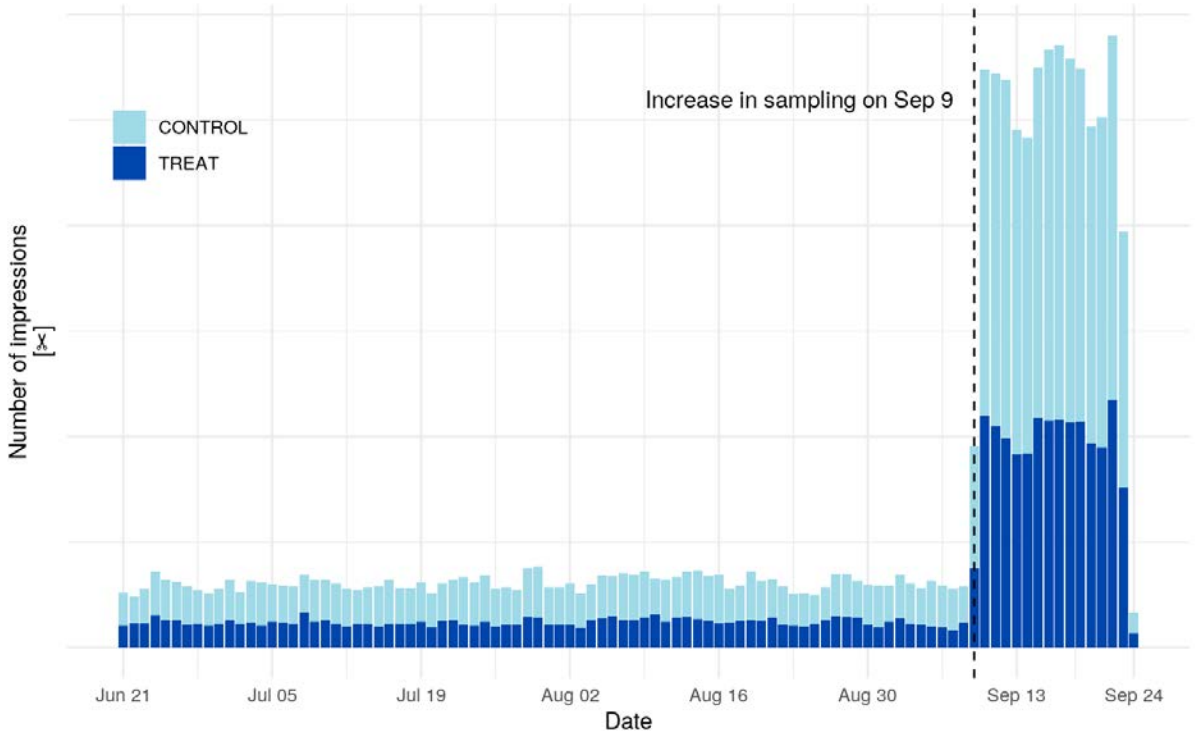
Sampling and randomisation

150. From the universe of UK user IDs that Google observed in its display advertising systems, user IDs were randomly selected to participate in the experiment. Users were blind to their selection and their allocation to either the treatment or control group. They were allocated to each of the control (with cookies/business as usual) and treatment (no cookies) arms, with equal probability.[48] Once a user ID was included in the study, data on all of the impressions served as part of their browsing activity during the study period was collected.

151. A timeline of the number of impressions included in the experiment is displayed in Figure F.2. Note that the sampling rate was increased on 9 September 2019 to achieve increased accuracy.

[46] The remaining [70-80]% of Google SSP traffic was either from users who were not logged-in to their Google Account ([60-70]%), or users who were logged-in but for whom this information was not used for the purpose of display advertising ([5-10]%).
[47] The primary reason why Google excluded these users from the experiment is because Google's adtech systems are able to access equivalent information for users logged into their Google Account as the information associated with the user's cookie ID stored in Google's repositories. Therefore, for these users, the treatment of blocking cookie information would have been ineffective (or had reduced effectiveness), and their inclusion in the study may lead to an underestimate of the effect of removing third-party cookies on publisher revenues.
[48] The overall probability for a user ID of being included in the study was [0.001-0.01]% and increased to [0.01-0.1]% on 9 September 2019 until the end of the experiment on 23 September 2019. User IDs were subsequently randomised with equal probability to each of the treatment and control arms, resulting in an overall probability of [0.001-0.1]% ([0.01-0.1]% after 9 September 2019) of being included in each arm.

Figure F.2: Number of impressions by date



Source: CMA computations on Google data. The sample includes all UK impressions in the study, regardless of user cookie availability.

152. Sampling was performed at the user level. However, not all users can be consistently identified throughout their traffic. Users who allow cookies during navigation can be identified by their cookie ID. This ID is persistent – at least until cookies are deleted by the user using their browser or device settings – and can thus be used to match impressions served to the same user in time across different domains. Once a user identified by a cookie ID is randomised to be in either control or treatment, all the impressions that were served to them can be allocated to the same group.

153. On the other hand, users navigating without cookies cannot be identified across their browsing; from a sampling perspective, these are treated as 'new' users every time they appear in Google's traffic. This means that the same user might appear in both control and treatment.

154. There are two main reasons why Google did not have access to a cookie ID for some users in this experiment:

- In some cases ([10-20]% of total impressions in the control group), the cookie information is simply not available. This is mostly due to device- or browser-level privacy settings – eg users browsing with Safari where Intelligent Tracking Prevention is enabled.

- In other cases ([10-20]% of total impressions in the control group) users had explicitly opted out of cookie-based advertising through Google's own Ads Preferences.[49]

Covariates and sample restrictions

155. Throughout the analysis, we condition on a set of user, publisher, and impression characteristics denoted by X. These are assumed to be determined independently of the treatment and its effects. We also use these variables as a basis for some sample restrictions that exclude small subsets of the original sample with less frequently occurring characteristics. These characteristics and restrictions are detailed in Table F.4.

Table F.4: Description of covariates and sample restrictions

Level	*Variable*	*Description*	*Sample restrictions*
User	Operating System	iOS, Android, Windows 10, older versions of Windows, MacOS	Exclude other operating systems (~1.8% of total impressions)
User	Platform	Mobile, desktop, tablet	Exclude other platforms (TV, console, ~0.2% of total impressions)
User	Browser	Chrome, Safari, Edge, Samsung Browsers, Internet Explorer, Firefox, other browser, and missing browser information	
User	Cookie age	Five intervals of approximately equal size: up to one week (0-7 days), one to 8 weeks (8-56 days), 8 to 20 weeks (57-140 days), 20 to 35 weeks (141-245 days), and over 35 weeks (245 days)	
Publisher	UK domain	*True* if publisher's effective top level domain (eTLD) is .uk	Exclude impressions for which the publisher domain is not available (<0.001% of total impressions)
Publisher	Language	*True* if the publisher's page language is English	Exclude impressions for which page language is not available (<0.001% of total impressions)
Publisher	Type	Takes values *YouTube*, *News*, or *Other*. The *News* category is manually derived from the top 250 most common publishers in the control group, and includes local and international news outlets.	
Impression	Time of day	Four intervals based on UK local time: 0-6, 6-12, 12-18, 18-24	
Impression	Weekend	*True* if on a Saturday or Sunday	

Source: CMA, Google

156. As this was a randomised experiment, the inclusion of these variables does not affect the unbiasedness of the estimator but does increase the degree of

[49] See *Opt out of seeing personalised ads*, available here. This choice has recently been made persistent by Google, instead of being reset upon clearing cookies in the browser.

statistical precision. Furthermore, these covariates serve as the basis for two additional tasks: the imputation procedure we adopt to deal with sample selection (detailed below), and the analysis of treatment effect heterogeneity.

157. We also restrict the sample in two additional ways:

- The analysis is only focused on *users navigating with cookies*. It excludes users with no cookie information, and users who have opted out of cookie-based advertising in their Google Account settings. These users are not affected by the intervention.

- The sample excludes impressions with a publisher payout (the payment that the publisher received) above its 99.9th percentile value (approximately [0.1-0.2] USD) to reduce the impact of outliers on our estimates.[50]

158. The characteristics of the sample in terms of covariates X are show in Table F.5. The table is limited to the control group, as the composition of the treatment sample might be influenced by the treatment itself (see next section).

[50] The maximum value for publisher payout in the data is around [0.1-0.2] USD, which is an order of magnitude larger than the 99.9th percentile.

Table F.5: Summary statistics

Variable	*Group*	*Percentage*	
		Impression-level (%)	*Query-level (%)*
User OS	Windows 10	[30-40]	[30-40]
	Android	[30-40]	[30-40]
	iOS	[10-20]	[20-30]
	Windows < 10	[10-20]	[5-10]
	MacOS	[0-5]	[0-5]
User platform	Desktop	[40-50]	[40-50]
	Mobile	[40-50]	[40-50]
	Tablet	[5-10]	[5-10]
User browser	Chrome	[50-60]	[50-60]
	Safari	[10-20]	[10-20]
	Edge	[5-10]	[5-10]
	Missing	[5-10]	[5-10]
	Other	[5-10]	[0-5]
	Samsung Browser	[5-10]	[0-5]
	IE	[0-5]	[0-5]
	Firefox	[0-5]	[0-5]
User cookie age	1 to 8 weeks (8-56 days)	[20-30]	[20-30]
	over 35 weeks (> 245 days)	[20-30]	[10-20]
	20 to 35 weeks (141-245 days)	[20-30]	[20-30]
	8 to 20 weeks (57-140 days)	[10-20]	[10-20]
	1 week or less (0-7 days)	[10-20]	[20-30]
Publisher UK domain	Non-UK domain	[60-70]	[60-70]
	UK domain	[30-40]	[30-40]
Publisher language	English	[90-100]	[90-100]
	Non-English	[5-10]	[5-10]
Publisher type	All other publishers	[70-80]	[60-70]
	News providers	[10-20]	[20-30]
	YouTube	[5-10]	[10-20]
Impression time of day	6-12	[30-40]	[30-40]
	12-18	[30-40]	[30-40]
	18-24	[20-30]	[20-30]
	0-6	[5-10]	[5-10]
Impression day of week	Weekday	[70-80]	[70-80]
	Weekend	[20-30]	[20-30]

Source: CMA, Google.
Notes: Sample excludes users navigating without cookies, and is restricted to the control group. Other sample restrictions are on revenue outliers, publisher domain/language unavailable, user platform/OS (see *Covariates and Sample Restrictions* above for details).

Sample selection issues

159. The objective of Google's experiment was to assess the short-term effect of blocking access to third-party cookie information on publisher revenue from programmatic ad sales. In this subsection, we focus on a set of selection issues that might be induced by features of the sample and the experiment, how they may cause biased estimates of this short-run effect, and what direction this bias plausibly takes.

160. A user's browsing activity can be thought of as a series of '*queries'* – ie occasions when the content of a publisher's page is loaded by the user's browser. This query *might* result in one or more impressions (ads) being loaded and shown to the user. It is important to note some aspects of this process for the purposes of this analysis:

- For any query, only the impressions that have won the supply-side auction are recorded; no information is available for losing bids.
- For any query, only the impressions that are served via the Google adtech stack are recorded; no information is available for impressions that were served to the user by other SSPs and DSPs without Google's involvement.
- Not all queries (pages visited by the user) generate impressions; some queries might result in no impressions being served and would thus not be present in the data.

161. The features of the experimental setup and the sampling process, combined with the fact that only winning impressions are recorded, imply a number of selection issues. For the purposes of this analysis, the term *'selection'* is used to signify that the intervention might not only affect outcomes in the treatment group (such as publisher payout) but might also affect *which impressions are selected* into the sample we observe. This would mean that users in the treatment group are *selected*, and not comparable to the control group. Therefore, a simple comparison between impressions in the treatment and control groups would lead to biased results, because the composition of each group is different to begin with.

162. Table F.6 outlines the main selection issues we have identified. They are described in detail in the following sections, together with our proposed solutions where applicable.

Table F.6: Selection issues and proposed solutions

	Issue	*Bias direction*	*Proposed solution*
1	Treatment blinds Google DSP selectively	*Over*estimation (larger negative effect)	Exclude DSP-only impressions
2a	Treatment reduces number of impressions per query	*Under*estimation (smaller negative effect)	Aggregate impressions at the query level
2b	Treatment increases queries without impressions	*Under*estimation (smaller negative effect)	Impute lost queries using control group distribution
3	Treatment increases impressions served by non-Google SSP	*Over*estimation (larger negative effect)	Outside the scope of this study

Source: CMA.

Table F.7: Impressions and queries

	Impressions per user		*Queries per user*		*Impressions per query*	
Google role in stack	*Control mean (SD*	*Treatment mean (SD)*	*Control mean (SD*	*Treatment mean (SD)*	*Control mean (SD*	*Treatment mean (SD)*
DSP + SSP	[40-50] (244.9)	[30-40] (179.5)	[20-30] (167.8)	[20-30] (109.1)	[1.8-2.1] (1.36)	[1.8-2.1] (1.38)
DSP only	[30-40] (243.5)	[5-10] (37.9)	[30-40] (215.4)	[5-10] (28.0)	[1.2-1.5] (0.48)	[0.9-1.2] (0.58)
SSP only	[20-30] (155.8)	[5-10] (25.8)	[20-30] (119.1)	[5-10] (21.8)	[1.2-1.5] (0.56)	[1.2-1.5] (0.57)

Source: CMA, Google.
Sample excludes users navigating without cookies. Standard deviations (SD) in parentheses. Other sample restrictions are on revenue outliers, publisher domain / language unavailable, user platform/OS (see *Covariates and Sample Restrictions* above for details).

Issue 1: Treatment blinds Google DSP selectively

163. As indicated above, impressions that are served by SSPs other than Google's still contain a cookie ID in the bid request. In such situations, the treatment implies blocking the Google DSP from accessing information associated with the cookie ID – which could put Google at a disadvantage with respect to other DSPs that can access information associated with that cookie ID in determining whether and how much to bid.

164. The direction of this effect on the probability of Google's DSP of winning RTB auctions is *ex ante* unclear: once 'blinded', Google DSPs might be induced to bid either too little or too much for a given impression. In the data fewer impressions per user are recorded in the treatment group compared to the control group, as seen in Table F.7. Given that only winning impressions are recorded, this appears to indicate that Google DSPs win auctions at a lower rate in the treatment group than in the control group.

165. This issue can be expected to result in estimates of the treatment effect that appear larger in magnitude than what would happen if every other DSP was also prevented from using cookie information, hence overestimating the negative effect of blocking access to cookies.

166. We mitigate this issue by focusing exclusively on the subsample of impressions in both the treatment and the control group where Google plays the role of SSP. Google might still be competing in these auctions with its DSPs, but other DSPs are equally deprived of the cookie ID in the bid request.[51]

ISSUE 2a: Treatment reduces number of impressions per query

167. For each query, users in the treatment group are shown fewer impressions than users in the control group when Google is the SSP (Table F.7). A reason for this may be that, as the SSP does not include cookie IDs in bid requests, bidders might find the opportunity to show an ad to these users less valuable and thus lower their bids or not bid at all.[52]. As a consequence, fewer ad slots would get filled for the same page visit for a treated user compared to an equivalent control user. Any treatment effect estimate that compares the mean publisher revenue for impressions between the treatment and control groups would therefore not fully capture the actual publisher revenue loss; this is because the number of impressions in each group is different. Our expectation is that this selection dynamic would lead to an underestimate of the short-run loss of revenue to publishers, as we are not observing impressions that would have been served absent the treatment.

168. Note that some ad slots, rather than not being served altogether, might be filled by impressions supplied by other non-Google SSPs. We would similarly not observe these impressions. Issue 3 below outlines this problem in more detail.

169. We address issue 2a by aggregating publisher revenue at the query level, summing revenues for all impressions served as part of the same query. This ensures that treatment-control differences in revenues are normalised by number of impressions.[53]

[51] This is in accordance with Google's own suggestion upon submission of this data. Notice however that excluding DSP-only impressions is likely to generate additional selection in the sample, as these impressions might be significant sources of revenue for some publishers.
[52] This could be the case if, for instance, the absence of cookie ID with associated information is perceived by advertisers as increasing the likelihood that the traffic is invalid or a bot.
[53] As there is no query identifier in the dataset supplied by Google, we generate query IDs for impressions that are served to the same user, on the same publisher domain, within 500 milliseconds of each other. Google have confirmed that the choice of this pragmatic threshold does not alter the broad findings of the analysis.

ISSUE 2b: Treatment increases queries without impressions

170. The query-level approach considered above should mitigate selection issues driven by the different number of impressions being served for otherwise comparable queries in the treatment and control groups. However, there may be instances where the treatment causes *no ads* to be served to users, meaning that no query is recorded. As an example, consider the scenario where a user visiting a website would have, if assigned to the control group, been shown two ads for this query, which would be recorded in the dataset, but if assigned to the treatment group, would be shown no ads, and thus their visit to the website is completely unobserved and not recorded in the dataset.

171. We suspect that this might be happening in the dataset because there are fewer queries per user in the treatment group compared to the control group (see Table F.7). Our expectation is that this selection dynamic would lead to an underestimate of the short-run loss of revenue to publishers, as we are not observing treatment group queries that generate no revenue.

172. Again, rather than generating no revenue, some queries might actually comprise impressions that are served by non-Google SSPs. See Issue 3 below for more details.

173. To mitigate this issue, we try to impute the unobserved queries in the treatment group, using the distribution of query characteristics in the control group. The intuition is that, in a balanced RCT with proper random assignment, we expect query characteristics to be similar in the treatment and control groups. Imputing queries in the treatment group will move the distribution of queries in treatment closer towards the distribution of queries in the control group.

174. To do this, we sample additional queries from the control group and add them to the treatment group, proportional to the extent to which the treatment group is underrepresented according to their characteristics. We can match treatment and control queries only with respect to characteristics that are not themselves affected by treatment assignment. These correspond to the user, publisher, and impression variables in Table F.4, denoted by X. Imputed queries, as they stand in for queries that did not carry any impressions, provide zero revenues for the publisher.

175. In practice, the imputation procedure is implemented as follows:

- Start from the query-level dataset, where the publisher payout for each query is summed across all impressions in that query.

- Use the set of characteristics X to define all 'cells' with unique combinations of publisher and user characteristics occurring in the data. An example cell might be: queries from Chrome users on an Android mobile device with a cookie aged 0-7 days, visiting a UK news publisher page in English between 18:00-24:00 on a weekday.

- Count the number of queries in each X-cell, separately for treatment and control. We drop impressions in X-cells that have zero queries in either the treatment or control group (approximately 3.5% of queries).

- Compute the difference in number of observations in each X-cell between control and treatment. If the difference is negative (more observations in treatment than in control), cap it at zero.

- Normalise the (capped) difference to a probability summing to one. This probability represents the extent to which the treatment group is under-represented in each X-cell, in terms of number of queries.

- Use the probabilities defined in the previous step to sample additional observations to be added to the treatment group, approximating those queries that have 'disappeared' due to the intervention, until the number of queries is the same in treatment and control. Assign a publisher payout value of zero to the newly imputed queries, since they were not actually shown to users in the treatment group.

176. This procedure produces a query-level imputed dataset that has an equal number of queries for users in both groups, and is closer to being balanced in terms of observable characteristics X.[54]

Issue 3: Treatment increases impressions served by non-Google SSP

177. As pointed out above, users in the treatment group are associated with fewer queries than in the control group, and their queries are associated with fewer impressions. In our approach to issues 2a and 2b we assume that these differences are entirely due to queries being likely to generate fewer overall impressions where cookies are blocked, or not to generate any impressions at all. To compensate for the queries that are lost, we aggregate at the query level and impute zero-revenue queries to the treatment group.

[54] The composition of query characteristics before and after the imputation step is presented in Table F.11. The table shows that imputing queries to the treatment group significantly narrows the difference in the distribution of characteristics X between treated and control queries.

178. However, another explanation for the lower number of impressions served in the treatment group is that the intervention increases the proportion of impressions that are served completely outside of the Google stack. This could be the case for publishers who offer their inventory to multiple SSPs, for example via header bidding or open bidding. By not passing the cookie ID downstream, the Google SSP would be putting itself at a disadvantage compared to other SSPs. This might cause non-Google SSPs to successfully sell impressions at a higher rate in the treatment group. As a consequence, some of the impressions and queries that we cannot observe in the dataset, for which we impute as if they were generating no publisher revenue, might actually have been filled by other sources of supply and have generated revenue.

179. Our expectation is that this selection dynamic would lead to an overestimation of the short-run negative effect of blocking cookies on publisher revenue. The available data does not allow us to address this issue in a convincing manner. We thus consider the magnitude of our estimates as upper bounds to the detrimental effect of the removal of cookie IDs for publisher revenue.

Results

Treatment effects

180. This section outlines the estimation of average treatment effects (ATEs). Following the selection issues identified in the previous section, we examine how the estimate for the ATE changes as we implement different econometric approaches to tackle each selection issue.[55]

181. Our main outcome of interest is publisher payout, a measure of publisher revenue for the impressions included in the dataset.[56] For each specification, we present both the ATE in terms of monetary value and as a percentage of the mean payout in the control group – to make it comparable with the original Google working paper.

182. We estimate average treatment effects with a simple linear regression approach, using the following specification:

$$\text{Payout}_{ij} = \beta_0 + \beta_1 \text{Treat}_i + u_{ij} \,,$$

[55] All the calculations in this annex are produced using **R** 3.6.2 (R Core Team, 2019). Data wrangling is performed using **R** packages dplyr 0.8.5 (Wickham et al., 2020) and data.table 1.12.8 (Dowle and Srinivasan, 2019), while plotting is performed using ggplot2 (Wickham, 2016).
[56] We measure payout in US Dollar cents, to simplify the presentation of very small absolute payout numbers.

where:

- Payout_{ij} is the publisher payout for impression (or query) j initiated by user i (in US dollar cents);
- Treat_i is a binary indicator taking the value of one if user i was randomly allocated to the treatment group, and therefore for whom the cookies were turned off;
- u_{ij} is the error term.

Under the assumption of randomised treatment assignment, the estimate of the coefficient β_1 corresponds to the ATE. In this simple model, it is equivalent to a difference in means between the payout in the treated and control groups.

183. In Table F.4 we identified some user, publisher, and impression characteristics that are independent of treatment assignment – denoted by X. We additionally estimate ATEs adjusting for this set of covariates:

$$\text{Payout}_{ij} = \beta_0 + \beta_1 \text{Treat}_i + \gamma X_{ij} + u_{ij}.$$

Under random assignment, this is not needed to interpret β_1 as an average treatment effect, but it can increase the precision of our β_1 estimate.

Table F.8: Average treatment effects at the impression level

	Outcome: Publisher revenue at the impression level (USD cents)					
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Control group mean (USD cents)	[0.1-0.2]	[0.1-0.2]	[0.1-0.2]	[0.1-0.2]	[0.1-0.2]	[0.1-0.2]
Effect (USD cents)	- [0.04-0.06] (0.003)	- [0.04-0.06] (0.004)	- [0.06-0.08] (0.004)	- [0.06-0.08] (0.004)	- [0.06-0.08] (0.004)	- [0.04-0.06] (0.003)
Effect (% of control mean)	-42.3 (1.0)	-36.4 (1.1)	-47.6 (1.0)	-58.1 (0.4)	-55.0 (0.4)	-50.9 (0.4)
Observations	[2,000,000 - 3,000,000]	[1,000,000 - 2,000,000]	[2,000,000 - 3,000,000]	[2,000,000 - 3,000,000]	[1,000,000 - 2,000,000]	[1,000,000 - 2,000,000]
Level	Impression	Impression	Impression	Impression	Impression	Impression
Top 500 publishers only		Yes				
Excludes users with no cookies			Yes	Yes	Yes	Yes
Other sample restrictions				Yes	Yes	Yes
Excludes DSP-only impressions					Yes	Yes
Controls for covariates X						Linear

Source: CMA computations on Google data. Standard errors in parentheses, clustered at the user level. Other sample restrictions are on revenue outliers, publisher domain/language unavailable, user platform/OS (see *Covariates and Sample Restrictions* above for details).

184. Table F.8 shows estimates of the effect of blocking access to cookie information on publisher revenue at the impression level. Column (1) is a simple comparison between the mean revenue in the treatment and control group; here, the average revenue loss per impression is [✂], around 42% of the mean in the control group.

185. Column (2) restricts the sample to impressions from the 500 publishers with the largest total revenue in our dataset. This is intended to be the closest replication of the headline result presented by Google in its original short paper.[57]

186. When users with no available cookies are excluded from the sample in Column (3), the effect becomes larger in both absolute value and percentage. This is to be expected, as users navigating with no cookies are not affected by their blocking.[58]

187. In Column (4), we implement some additional sample restrictions as detailed in the *Covariates and Sample Restrictions* section above. Among these restrictions, we drop the top 0.01% of the sample in terms of publisher payout, which we interpret as outliers. This has a significant impact on the control group mean, and consequently on the effect as a percentage of that mean.

188. The results discussed so far do not attempt to mitigate any of the selection issues outlined above in Table F.6. In Column (5), we exclude impressions where Google only played the role of DSP (and not SSP), to exclude cases where the Google DSP would have been at a disadvantage (Issue 1). As expected, the impression-level effect of blocking cookies for this subsample is significantly smaller ([✂]) and becomes slightly smaller when controlling for the covariates defined in X in Column (6) – although not in a statistically significant way as can be indirectly evinced by the overlap in standard errors Table F.8. Still, these estimates do not account for the fact that queries in the treatment group might generate less impressions (Issues 2a and 2b above).

[57] Note that the original paper defined a 'Top 500' publisher globally, as 'ordered by revenue earned by publisher when served programmatic ads through Google Ad Manager'. We do not have access to global publisher rankings on Google Ad Manager, so our ranking of publishers is within impressions in the dataset. This means that the publishers encompassed in the two definitions of 'Top 500' will differ.
[58] In fact, the effect of the intervention on the subsample of users with no cookie information is small and statistically insignificant.

Table F.9: Average treatment effects at the query level

	Outcome: Publisher revenue at the query level (USD cents)			
	(1)	*(2)*	*(3)*	*(4)*
Control group mean (USD cents)	[0.1-0.2]	[0.1-0.2]	[0.1-0.2]	[0.2-0.3]
Effect (USD cents)	- [0.09-0.12] (0.001)	- [0.12-0.15] (0.001)	- [0.12-0.15] (0.001)	- [0.12-0.15] (0.001)
Effect (% of control mean)	-52.3 (0.5)	-69.9 (0.5)	-72.4	-71.7 (0.5)
Observations	[1,000,000 – 1,500,000]	[1,000,000 – 1,500,000]	[1,000,000 – 1,500,000]	[1,000,000 – 1,500,000]
Level	Query	Query	Query	Query
Excludes users with no cookies	Yes	Yes	Yes	Yes
Other sample restrictions	Yes	Yes	Yes	Yes
Excludes DSP-only impressions	Yes	Yes	Yes	Yes
Controls for covariates X	Linear	Linear	Causal Forests	Linear
Imputed queries		Yes	Yes	Yes
Excludes YouTube impressions				Yes

Source: CMA computations on Google data. Heteroskedasticity-robust standard errors in parentheses. Standard error is omitted for percentage effect for causal forests, as it cannot be straightforwardly derived using the delta method. Other sample restrictions are on revenue outliers, publisher domain/language unavailable, user platform/OS (see *Covariates and Sample Restrictions* above for details). The causal forests model is estimated using the **R** package `grf` (Tibshirani et al., 2020).

189. In Table F.9, we present treatment effect results on data that has been aggregated at the query level. As mentioned above, query aggregation deals with the possibility of queries by treated users generating less impressions than control users, by aggregating revenue for all impressions recorded for the same page visit (Issue 2a). In Column (1), we show that blocking cookies implies an average loss of revenue of [0.1-0.2] USD cents for each query. As expected, this is a larger loss in percentage terms than the impression-level effect in Column (6) of Table F.8, as it accounts for the lower number of impressions shown to treated users.

190. Column (2) of Table F.9 shows results for the same query-level data, with the addition of imputed queries to compensate for the loss of queries caused by the blocking of cookies in the treatment group (Issue 2b). The average effect of blocking cookies on publisher revenue is estimated to be [0.12-0.15] USD cents, 70% of the mean revenue in the control group.

191. The results in Column (2) are the farthest we have been able to advance in tackling selection issues with the information available in the data. As previously noted, the estimate of a 70% average revenue loss at the query

level is likely to be an upper bound to the loss of revenue actually experienced by publishers in the short run, since:

- it does not account for the possible replacement of Google impressions by non-Google supply sources (Issue 3); and
- it is limited to users navigating with cookies, excluding users with no cookies where the effect is plausibly null.

Treatment effect heterogeneity

192. Above, we estimate an average treatment effect. This average estimate may mask significant underlying heterogeneity. That is, the effect of blocking access to cookie information might not be the same across all types of query, but might differ based on user or publisher characteristics. To explore this issue, we adopt two different approaches: a simple regression model with interactions, and a more flexible causal forest approach.

Regression-based heterogeneity

193. As a simple but interpretable approach to estimate treatment effect heterogeneity, we augment the linear regression model from the previous section with interactions between treatment and covariates:

$$\text{Payout}_{ij} = \beta_0 + \beta_1 \text{Treat}_i + \gamma X_{ij} + \delta X_{ij} \cdot \text{Treat}_i + u_{ij}.$$

The additional term $\delta X_{ij} \cdot \text{Treat}_i$ denotes all pairwise interactions between the treatment indicator and the categories in covariates X.

194. Estimates for model parameters are presented in Table F.10. The size and statistical significance of interaction terms informs about the presence of heterogeneity in the average treatment effects, which are masked in the non-interacted specification of the model. There is significant heterogeneity across all dimensions of X, apart from user platform – which is likely being absorbed by the information in the operating system and browser dimensions.

195. In this way, the predicted marginal treatment effect for a given query j by user i can be computed using estimates for β_1 and δ. Figure F.3 shows the average values for these marginal effects across the values of each covariate in X.

Table F.10: Regression-based heterogeneity

		Outcome: Publisher revenue at the query level (USD cents)
		Coefficient (SE)
	Constant	[0.09-0.12] (0.156)
	Treat	- [0.03-0.06] (0.211)
User OS (reference: Android)	iOS	[0.03-0.06]*** (0.003)
	MacOS	[0.12-0.15] (0.156)
	Windows 10	[0.06-0.09] (0.156)
	Windows < 10	[0.06-0.09] (0.156)
	Treat * iOS	-[0.06-0.09]*** (0.004)
	Treat * MacOS	-[0.15-0.18] (0.211)
	Treat * Windows 10	-[0.09-0.12] (0.211)
	Treat * Windows < 10	-[0.06-0.09] (0.211)
User platform (reference: Desktop)	Mobile	[0.06-0.09] (0.156)
	Tablet	[0.03-0.06] (0.156)
	Treat * Mobile	-[0.06-0.09] (0.211)
	Treat * Tablet	-[0.06-0.09] (0.211)
User browser (reference: Chrome)	Edge	-[0.00-0.03]*** (0.002)
	Firefox	-[0.09-0.12]*** (0.003)
	IE	-[0.03-0.06]*** (0.002)
	Safari	-[0.06-0.09]*** (0.003)
	Samsung Browser	-[0.00-0.03]*** (0.002)
	Other	[0.00-0.03] (0.003)
	Missing	-[0.12-0.15]*** (0.003)
	Treat * Edge	[0.03-0.06]*** (0.003)
	Treat * Firefox	[0.09-0.12]*** (0.004)
	Treat * IE	[0.03-0.06]*** (0.003)
	Treat * Safari	[0.12-0.15]*** (0.004)
	Treat * Samsung Browser	-[0.00-0.03]* (0.003)
	Treat * Other	[0.03-0.06]*** (0.004)
	Treat * Missing	[0.12-0.15]*** (0.005)
User cookie age (reference: 1 to 7 days)	1 to 8 weeks (8-56 days)	[0.03-0.06]*** (0.001)
	8 to 20 weeks (57-140 days)	[0.03-0.06]*** (0.001)
	20 to 35 weeks (141-245 days)	[0.03-0.06]*** (0.001)
	over 35 weeks (> 245 days)	[0.06-0.09]*** (0.001)
	Treat * 1 to 8 weeks (8-56 days)	-[0.03-0.06]*** (0.002)
	Treat * 8 to 20 weeks (57-140 days)	-[0.03-0.06]*** (0.002)
	Treat * 20 to 35 weeks (141-245 days)	-[0.03-0.06]*** (0.002)
	Treat * over 35 weeks (> 245 days)	-[0.06-0.09] *** (0.002)
	Publisher page language Non-English	-[0.09-0.12]*** (0.002)
	Treat * Publisher page language Non-English	[0.06-0.09]*** (0.003)
	Publisher UK domain	[0.00-0.03]*** (0.001)
	Treat * Publisher UK domain	-[0.00-0.03]*** (0.002)
Publisher type (reference: All other providers)	News providers	-[0.00-0.03]*** (0.001)
	YouTube	-[0.15-0.18]*** (0.002)
	Treat * News providers	[0.00-0.03]*** (0.002)
	Treat * YouTube	[0.15-0.18]*** (0.002)
Time of day (reference: 0-6)	6-12	[0.00-0.03]** (0.002)
	12-18	-[0.00-0.03]** (0.002)
	18-24	-[0.00-0.03]*** (0.002)
	Treat * 6-12	-[0.00-0.03]** (0.002)
	Treat * 12-18	[0.00-0.03] (0.002)
	Treat * 18-24	[0.00-0.03]*** (0.002)
	Weekend	[0.00-0.03]*** (0.001)
	Treat * Weekend	-[0.00-0.03]*** (0.001)
	Observations	[1,000,000-2,000,000]

Source: CMA computations on Google data. The table displays estimated coefficients for a linear regression of publisher payout on treatment status, covariates, and their first-level interactions. Heteroskedasticity-robust standard errors in parentheses. Significance: * = *p<0.1;* ** = *p<0.05;* *** = p<0.01

196. While some dimensions don't exhibit much variation around the average effect, there is marked heterogeneity along some relevant dimensions. In particular, blocking cookies for users browsing on an iOS device, or using Safari or Firefox as browsers, has a relatively smaller negative effect on publisher revenue. This is consistent with efforts from Apple and Firefox to disable tracking functionality on their browser, which might cause queries from such users to appear less valuable to advertisers.[59] Larger average effects are observed for Android users and browsers like Chrome and Samsung Browser.

197. The age of the cookie can be broadly interpreted as a measure of potential data quality associated with the cookie. Although data quality is affected by many factors, a cookie that has been used for a long time can be linked with more browsing data and can allow for building a more detailed user profile. As expected, there is an approximately monotonic relationship between the effect of blocking cookies and the cookie age. The effect on queries generated by users with older cookies is larger than for users with younger cookies, indicating that there might be more value in serving ads to users that have been observed for longer. This is indirect evidence of a relationship between user data quality and publisher revenue.

198. On the publisher side, the effect of blocking cookies is larger for English language pages, and pages with a UK domain. It is not possible to determine whether this is due to the characteristics of the users visiting this type of pages or the pages themselves. Keeping in mind that the sample only includes users in the UK, it might be that UK publishers find it more valuable to serve personalised adverts to such users, given they might be closer to the bottom of the conversion funnel by virtue of their location. This could explain why a larger effect is observed for UK publishers.

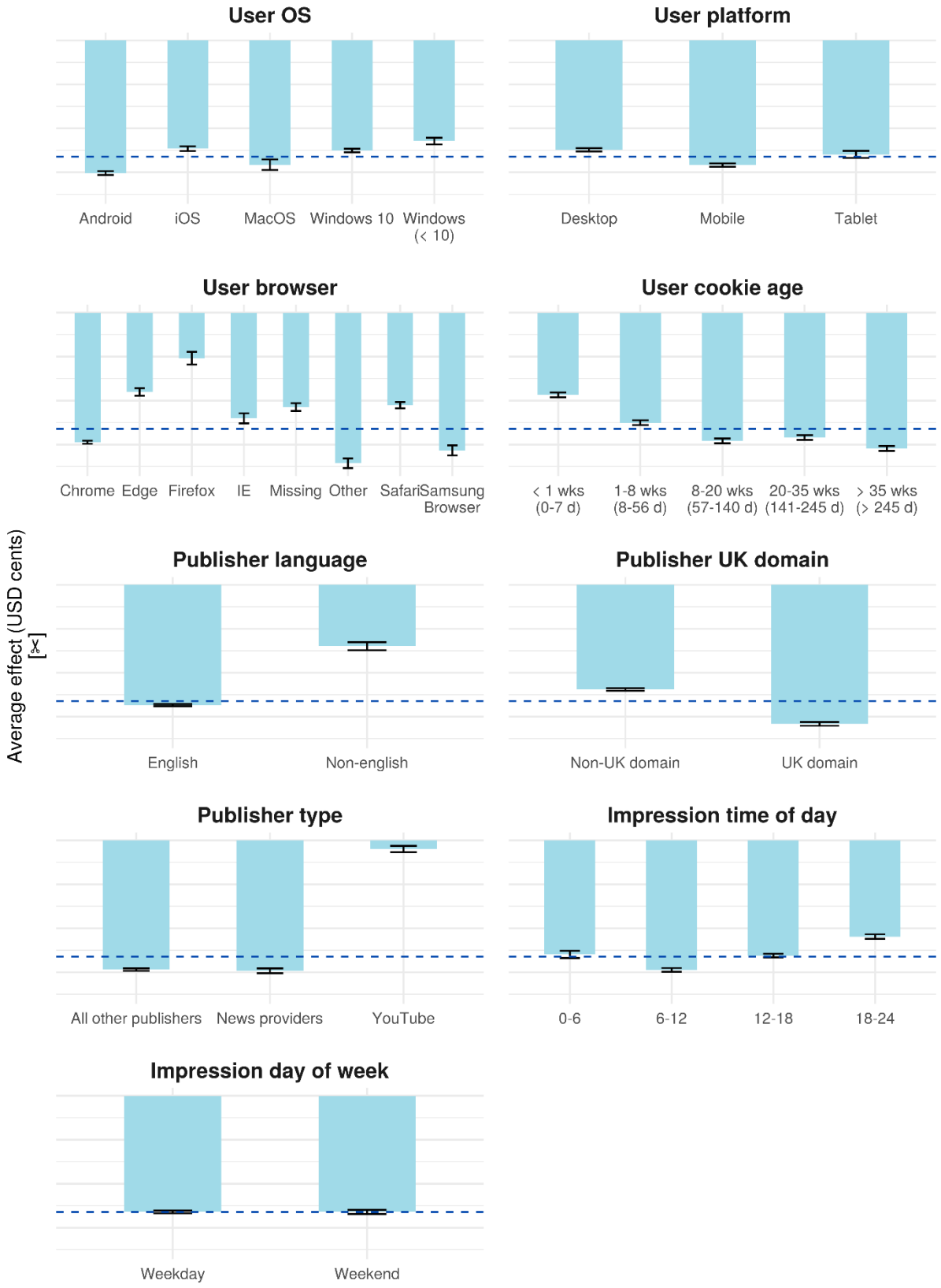
199. The negative effect of blocking cookies on publisher revenue is similar for large news publishers (see Table F.4 for the definition) and other types of publishers. The effect is almost null for impressions served on YouTube. This is because YouTube does not use third-party cookies to begin with, but rather Google's own first-party cookies. We present an additional estimate of the average effect excluding YouTube in column (4) of Table F.9 as a robustness check. Despite a slightly higher mean revenue in the control group and larger

[59] At the time of the Google experiment, the Intelligent Tracking Prevention tool in WebKit (the engine behind Safari) was in its version 2.2, which provided near-total third-party cookie blocking and stringent rules on cookie persistence by default. In fact, even for Safari users with available cookie information, the cookie age is significantly shorter than for other browsers. As seen in Figure A1, at the moment of inclusion in the study almost all Safari users have a cookie that is less than a day old. Similarly, Firefox announced default blocking of third-party cookies, but this occurred only in the version of the Enhanced Tracking Protection tool that was rolled out on 3rd September 2019 – towards the end of the study period.

absolute effect, the effect as a percentage is qualitatively similar to the estimates in columns (2) and (3). This shows that the effects we estimate are robust to the exclusion of YouTube as a publisher.

200. Finally, effects are broadly homogeneous according to whether the query occurred on a weekend or a weekday. Slightly larger negative effects are observed for queries during the morning than at night.

Figure F.3: Heterogeneous effects, regression



Source: CMA computations on Google data. The figure shows the estimated average effect of blocking cookies on publisher revenue at the query level, broken down by the covariates in Table F.4. The black whiskers correspond to 95% confidence intervals with heteroskedasticity-robust standard errors. All estimates are obtained from the regression model in Table F.10. The dashed blue line represents the average treatment effect obtained from the same model.

Causal forests

201. One of the key limitations with the linear regression approach is that it does not account from any higher order interactions between the variables contained in X and the treatment indicator. In theory, these higher order interactions could be added to the model above, but this would become exponentially complicated and vulnerable to overfitting.

202. As a robustness check to the regression model, we estimate heterogeneous effects in a fully flexible and data-driven way using causal forests. The main advantage of this method is that the appropriate level of interaction between X variables does not have to be pre-specified and can instead be learned by training a random forests model.[60]

203. We have elected to use an *honest causal forests* approach (Athey & Imbens, 2016; Athey, Tibshirani & Wager, 2018; Wager & Athey, 2018). We use the same exogenous (independent to treatment assignment) covariates we identified in Table F.4 above, denoted by X.[61]

204. The estimate of the average treatment effect obtained by the causal forests approach is reported in column (3) of Table F.9. Reassuringly, this estimate is very close to the regression estimate in column (2). This acts as a sense check on the data-driven approach. If the average treatment effect from this process were to differ significantly from that estimated using a simple linear model this would cast doubt on the reliability of the heterogenous treatment effects that we have estimated.

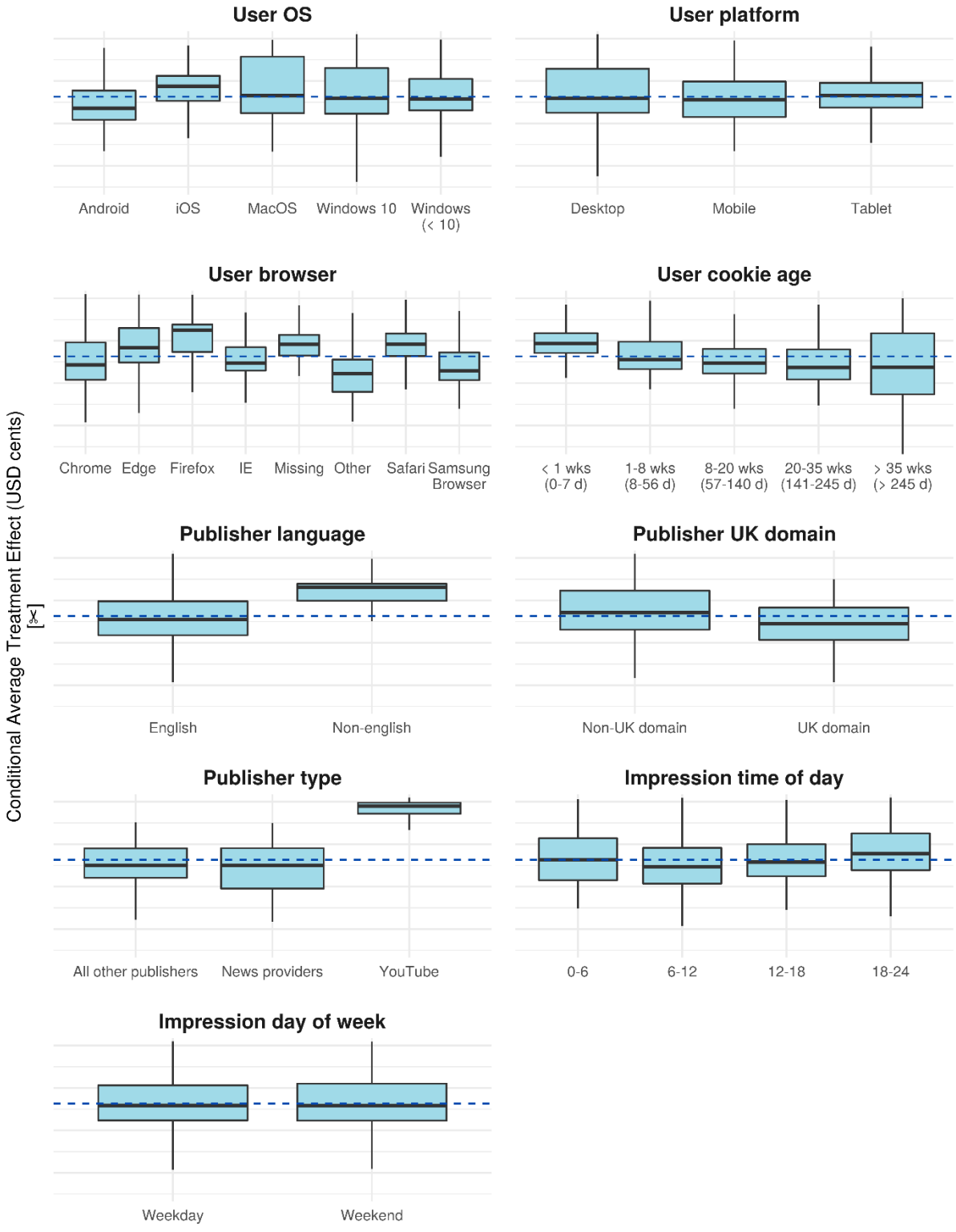
205. Figure F.4 displays the heterogeneity results graphically. Each box in the plot represents the distribution of predicted conditional average treatment effects (CATEs) for users in the subgroup labelled on the horizontal axis. The dashed line indicates the average treatment effect from Table F.9. Note that the

[60] The object of interest in heterogeneous treatment effect models is the individual or conditional average treatment effect. Conceptually this can be defined using the Rubin causal model, often referred to as the potential outcomes framework. The parameter of interest (τ_j, the 'individual treatment effect') is specified as $Y_{T=1,j} - Y_{T=0,j} = \tau_j$. The individual treatment effect τ_j in our setting is equal to the difference between the outcome for query j when exposed to treatment and the outcome for query j when not exposed to treatment. Randomisation ensures that the differences in the outcome between those queries exposed to treatment and those who were not is attributable to the treatment.

[61] The causal forest explicitly searches for the subgroups, defined by combinations of X, where the treatment effects differ most. If left unchecked the causal forest approach would arrive at query-level predictions based on every conceivable combination of variables and variable values. To avoid this overfitting, the data is split into two, one subsample is used for splitting and another is used for prediction. The causal forest algorithm is first applied to the splitting data to build a causal tree, the tree is then used to classify the prediction dataset and it is the difference in the treatment effect for observations within the leaves that is used to estimate the treatment effect. We use a random subset of 50% of the query-level imputed dataset to train the model, and the other half to predict. For the training, we grow 2000 trees, and use a fraction of 0.25 (1/4) of the training data to grow each tree. Given the sample size at our disposal, we set the minimum number of observations in each tree leaf to 100.

effects are expressed in USD cents, and thus a lower box corresponds to a larger negative effect. The conclusions that can be derived from these estimates are substantially equivalent to those derived from the regression estimates in Figure 2. This shows that modelling heterogeneity in a fully flexible way does not change the qualitative interpretation of the results.

Figure F.4: Heterogeneous effects, causal forests



Source: CMA computations on Google data. The figure shows boxplots of predicted conditional average treatment effects (CATE) of blocking cookies on publisher revenue at the query level, broken down by the covariates in Table F.4. In each boxplot, the thick middle line corresponds to the median, the hinges of the box to the interquartile range (between 25th and 75th percentiles), and the whiskers to the farthest observations within 1.5 times the interquartile range. Outliers are excluded from the plot. All estimates are obtained from the causal forest model in this section. The dashed blue line represents the average treatment effect obtained from the same model. The causal forests model and the corresponding predicted revenue values are estimated using the **R** package `grf` (Tibshirani et al., 2020).

Conclusions

Summary of results

206. In this annex we have re-examined the results of an experiment run by Google where access to third-party cookie IDs was disabled within the Google adtech stack for a random subset of users. Using impression-level data provided by Google for a sample of UK users, we have investigated the effect of disabling access to cookie IDs on publisher revenue.

207. Considering the particular characteristics of the experiment and of the data collection process, we identify a number of selection issues that would be expected to arise when simply comparing average publisher revenue for impressions in the treatment and control group. For most issues, we are able implement econometric solutions that allow us to better estimate the effect of blocking cookie access from the perspective of publishers.

208. Depending on the underlying assumptions about sample selection, our estimates of the average short-run effect of blocking cookies on publisher revenue range between 40% and 70% of the mean revenue in the control group, which approximates business as usual during the study period. Our best estimate is that blocking cookies decreases publisher revenue by .00132 USD per query in the short term. This corresponds to approximately 70% of the mean in the control group.

209. There is significant heterogeneity across user and publisher characteristics. Revenue from iOS users, and users browsing on Safari and Firefox, are less impacted by the blocking of cookies than Chrome users. The effect is larger for UK publishers and for pages in English, and is null for YouTube impressions (as they do not rely on third-party cookies).

What does the experiment tell us about the value of data?

210. The results of the experiment make it clear that, for individual publishers competing against other publishers that offer personalisable inventory using cookies, blocking access to cookie identifiers reduces publisher revenue from users navigating with cookies by a significant amount. In this sense, having access to cookies has value for individual publishers.

211. However, the nature of the data and the experiment imply some limitations on the generalisability of our estimates. These limitations can be grouped into two categories:

- '*short-run*, ie how accurately the estimates approximate the short-term effect of the blocking of cookies as experienced by publishers at the margin, as induced by the experiment;

- '*long-run market-wide'*, ie how well these short-term responses approximate what would happen at scale, if the cookie-less ecosystem implied by the intervention was commonplace and the actors in this ecosystem had time to adjust their strategies.

212. Let's start from short-run caveats. Our best estimate for the size of this loss is 70% of average revenue for each query. As touched upon above, this estimate should be regarded as an upper bound for the following reasons:

- The estimate is derived from a sample that excludes users with no available cookies (accounting for 28% of total impressions) for whom the effect of the intervention is null. Publishers actually face a mix of users with and without cookies enabled, and thus their revenue loss will be smaller.[62]

- Our estimates also do not account for the possibility that publishers might be able to replace impressions from Google supply paths with impressions from other SSPs, in the cases when Google blocks cookie IDs by excluding them from bid requests. This could partially compensate for the negative effect of Google blocking access to cookies. As explained in the previous section, this issue – which we have dubbed Issue 3 in our taxonomy – would result in an overestimation of the effect.

- Omitting cookie IDs from bid requests might generate adverse selection issues, where advertisers interpret the lack of cookie information as a signal of poor quality – especially from browsers that do not have tracking prevention enabled by default. In this perspective, any effect on publisher revenue for this traffic will not only approximate the value of the data associated with the cookie, but will also include the value of merely observing a cookie for the purposes of detecting fraudulent traffic, leading to an overestimation of the negative effect.

213. In addition, the data we received from Google excludes a significant fraction of users for whom the information in their Google Account was used to serve personalised ads. This implies that the results will not be representative of the whole population of UK traffic in the Google adtech stack. We cannot know a

[62] Our decision to focus on users with cookies stems from the ability to re-identify them across multiple impressions and queries. This allows us to tackle selection issues due to different numbers of impressions and queries by treatment status. However, we can't aggregate to the query level for queries by users for whom we don't have any identifier.

priori what the effect of blocking cookies for these users would have been. On one hand, the functionality of cookies might be superseded by Google's first party data, leading to a lower impact of the intervention on publisher revenue. On the other hand, signed-in users might be different from the rest of the sample in their monetisability from a publisher's perspective.

214. All the above consideration about the impact of the intervention on publisher revenue are true in a 'short-run' perspective, i.e. for a relatively small subset of impressions in an ecosystem where third-party cookies are the commonplace means of identifying users. In this ecosystem, there is no incentive for advertisers to bid on impressions for unidentifiable users, when there are billions more users that can be identified through their cookies.

215. A question this experiment cannot answer is what the impact would be in a 'long-run market-wide' perspective, where third-party cookies are unavailable throughout the ecosystem. In such a world, the impact on publisher revenue is likely to be mitigated by dynamic responses from actors in the ecosystem, for example:

- Heavier use of contextual targeting rather than personalised targeting;
- Increased reliance on first-party data for targeting, as well as integration of third-party trackers in first-party contexts;
- Other methods of cross-site tracking, including browser and/or device fingerprinting.

216. By its nature the experiment conducted by Google cannot circumvent these limitations – not due to poor study design or implementation, but due to the fact that the sample is necessarily limited to what Google can observe in its own traffic, and in the current adtech ecosystem. We expect that the estimates we provide for the short-run effect are significantly larger than the hypothetical effect on aggregate publisher revenues of an ecosystem-wide prohibition of the use of third-party cookies and other tracking technologies with similar effect to personalise advertising.

References

Athey, S. & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*.

Athey, S., Tibshirani, J., & Wager, S. (2018). Generalized random forests. *Annals of Statistics*.

Dowle, M., & Srinivasan, A. (2019). data.table: Extension of `data.frame`. R package version 1.12.8. https://CRAN.R-project.org/package=data.table

Leeper, T.J. (2018). margins: Marginal Effects for Model Objects. R package version 0.3.23.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Tibshirani, J., Athey, S., & Wager, S. (2020). grf: Generalized Random Forests. R package version 1.1.0. https://CRAN.R-project.org/package=grf

Wager, S. & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.

H. Wickham (2020). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

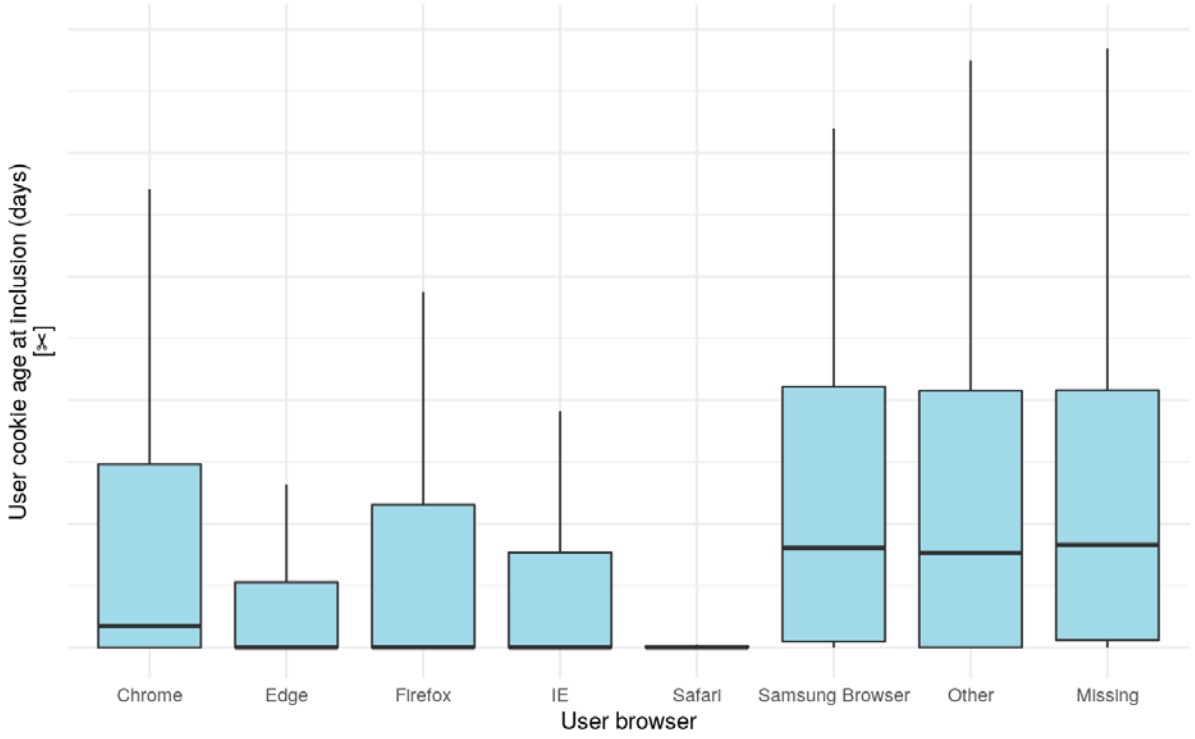
Additional tables and figures

Table F.11: Balance in query-level data

Variable	*Group*	*Percentage*		
		Control group (%)	*Treatment group (%)*	*Treatment group, including imputed queries (%)*
User OS	Windows 10	[30-40]	[30-40]	[30-40]
	Android	[30-40]	[20-30]	[20-30]
	iOS	[20-30]	[20-30]	[20-30]
	Windows < 10	[5-10]	[10-20]	[10-20]
	MacOS	[0-5]	[0-5]	[0-5]
User platform	Desktop	[40-50]	[40-50]	[50-60]
	Mobile	[40-50]	[40-50]	[40-50]
	Tablet	[5-10]	[5-10]	[5-10]
User browser	Chrome	[50-60]	[50-60]	[50-60]
	Safari	[10-20]	[10-20]	[10-20]
	Edge	[5-10]	[5-10]	[5-10]
	Missing	[5-10]	[5-10]	[5-10]
	Other	[0-5]	[0-5]	[0-5]
	Samsung Browser	[0-5]	[0-5]	[0-5]
	IE	[0-5]	[0-5]	[0-5]
	Firefox	[0-5]	[0-5]	[0-5]
User cookie age	1 to 8 weeks (8-56 days)	[20-30]	[10-20]	[10-20]
	over 35 weeks (> 245 days)	[10-20]	[10-20]	[10-20]
	20 to 35 weeks (141-245 days)	[20-30]	[10-20]	[20-30]
	8 to 20 weeks (57-140 days)	[10-20]	[20-30]	[20-30]
	1 week or less (0-7 days)	[20-30]	[20-30]	[20-30]
Publisher language	English	[90-100]	[90-100]	[90-100]
	Non-English	[5-10]	[5-10]	[5-10]
Publisher UK domain	Non-UK domain	[60-70]	[70-80]	[60-70]
	UK domain	[30-40]	[20-30]	[30-40]
Publisher type	All other publishers	[60-70]	[70-80]	[60-70]
	News providers	[20-30]	[10-20]	[10-20]
	YouTube	[10-20]	[10-20]	[10-20]
Impression time of day	6-12	[30-40]	[30-40]	[30-40]
	12-18	[30-40]	[30-40]	[30-40]
	18-24	[20-30]	[20-30]	[20-30]
	0-6	[5-10]	[5-10]	[5-10]
Impression day of week	Weekday	[70-80]	[70-80]	[70-80]
	Weekend	[20-30]	[20-30]	[20-30]

Source: CMA computations on Google data. Sample excludes users navigating without cookies. Other sample restrictions are on revenue outliers, publisher domain/language unavailable, user platform/OS (see *Covariates and Sample Restrictions* above for details).

Figure F.5: Cookie age at inclusion, control group



Source: CMA computations on Google data. The figure shows boxplots of each user's cookie age in days, broken down by the browser. In each boxplot, the thick middle line corresponds to the median, the hinges of the box to the interquartile range (between 25th and 75th percentiles), and the whiskers to the farthest observations within 1.5 times the interquartile range. Outliers are excluded from the plot. The sample is limited to the control group, and excludes users navigating without cookies. Other sample restrictions are on revenue outliers, publisher domain/language unavailable, user platform/OS (see *Covariates and Sample Restrictions* above for details).