Appendix E: The role of data

Introduction

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

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.

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

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

4. Our emerging thinking on the role of data is summarised below:

   a) Data gives platforms a competitive advantage in the provision of both consumer-facing and digital advertising services. In the provision of search services to consumers, having access to a greater volume of users and click-and-query data enables search engines to deliver more relevant results. This is particularly important for uncommon or new queries. For this reason, the greater scale of English-language queries seen by Google is likely to support its ability to deliver more relevant search results compared to its competitors, especially in relation to uncommon and fresh queries.

   b) In digital advertising, platforms provide targeting capabilities which allow advertisers to retarget their current and potential customers as well as reach wider audiences. For these purposes, detailed data on users' demographic characteristics, interests, preferences and behaviours is most valuable to predict consumers' potential response to advertising. Platforms also provide measurement, 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 is very important to demonstrate their effectiveness in digital advertising.

c) Google and Facebook have a competitive advantage because they collect a large amount and variety of data types from their widely used consumer-facing services and their broad coverage of third-party sites and apps. Rival platforms such as Microsoft and Amazon have access to some detailed high-quality data about users and other types of data, but we understand this is largely limited to their own services or does not extend widely to the rest of the internet. In order to compete, these platforms as well as intermediaries in the open display market can supplement their own data with data from other market participants, such as data management platforms. However, this requires rival platforms and intermediaries to extensively share data between one another and match different identifiers in order to compile rich user profiles. This process is less precise than the matching that takes place uniquely within closed ecosystems and we have heard that it also gives rise to privacy concerns caused by data flows.

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 however is not homogeneous, but diverse in content and nature, and its usefulness and value are unlikely to be equally important when data is used for different purposes. In order to understand whether and the extent to which 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 and previous reports to identify four broad categories of data useful for our purposes, namely:

- user data;
- contextual data;
- campaign 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. At this stage we consider this is the most useful categorization to assess the role that data and specific types of data play in platforms' market power in the provision of consumer-facing services and the supply 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 her 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 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

[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 (GDPR), 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?

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

and use (as described in concentration and quality of data section below) and the role of data in personalised advertising.

Contextual data

15. Contextual data refers to data on the context in which an advertisement 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.

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

Campaign data

18. Campaign 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 or campaign 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, 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 (measurement and attribution).

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). Modern 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 the provision of consumer-facing services in more detail below.

Summary of relevant types of data

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

Table E.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.
Campaign data		Number of consumers that click on an advertisement impression.
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 the extent to which 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. At this stage we have distinguished between two broad sources that platforms use to collect data: (i) their own consumer-facing services, and (ii) third-party providers.

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-party providers. 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

29. Platforms collect a wide range of data from the services 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) consumers' activities such as search history, clicks, content created and shared; and (iii) location through users' device information. The amount and types of data may vary based on the context in which consumers access platform's services, 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 for several reasons:

    a) Certain consumers may access some services, or platforms offer them, only through one type of device (eg apps on mobile devices). If this is the case and platforms cannot track consumers on several devices, they would capture data from fewer consumers (eg desktop-only users).

    b) Being able to follow consumers across devices also allows platforms to observe a wider range of behavioural data by capturing a larger part of

[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.

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 and measurement services.

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

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 in the UK, including Google Search, and gathers data through them. This data includes:

a) 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 28-days 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).

b) 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

[6] Non-mobile devices (such as laptops) 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.

advertising identifiers, such as the Android Advertising ID (AdID) or iOS Identifier for Advertisers (IDFA).

c) 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.

d) 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. The figure below shows an example of the detailed information collected.

Figure E.1: 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]

[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.

a) 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.

b) 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.

c) 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 you use), device signals (Bluetooth signals, etc), network and connections (ISP, language, time zone, mobile phone number) and cookie data (cookie IDs and settings).

Third-party providers

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 at this stage we understand that the main are the following:

a) Data is actively shared by third-party providers;

b) Data is collected directly from third-party sites or apps through technology;

c) Through sign-in functionality on third-party sites or apps; and

d) Through advertising services on publishers' sites or apps.

Data actively shared by third-party providers

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

    a) 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 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 similar audience section below).

    b) 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.

    c) 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 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 or other content such as videos from YouTube. 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, 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.

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

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

44. 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.[10] 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.

45. 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

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

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 date/time the cookie was installed, etc.); 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).

Through advertising services to publishers

46. Platforms can collect data through the advertising services they provide to other websites and apps. 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 campaign data and additional user data when providing measurement, verification, and attribution services (more detail in role of data in digital advertising section below).

47. 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:

a) 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;

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

c) 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 AdID for Android). In the case of web browsers, any cookie IDs that Google has previously set on the user's device; and

d) Event data such as impression, click or conversion data.

48. 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

49. This section draws on the description of the types and sources of data above to assess the relative competitive positions of platforms in relation to the amount of data they collect and use.

50. Google is the platform with the largest dataset collected 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.

51. 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,[11] from which it collects very granular user data.

52. The evidence indicates that the reach of Google and Facebook tools on third-party sites and apps is very extensive. Evidence submitted to the CMA estimates that Google tags (ie Google Analytics, Google Ads and Foodlight tags) cover about 88% of UK websites, whereas Facebook's tags cover about 50% of UK websites. This compares with Microsoft, whose tags cover less than 1% of UK websites. Other submissions also indicate that Google and Facebook account for the vast majority of third-party data collectors across the web.[12] Oracle 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

[11] Comscore MMX MP, Total Digital Audience, Desktop aged 6+, Mobile aged 13+, June 2019, UK. See further Appendix C.

[12] For example, Digital Content Next response to the statement of scope.

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.

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

a) Through their extensive reach on third-party sites and apps, Google and Facebook collect a large amount of campaign data that, as described in the advertising verification, measurement and attribution section below, they use to provide measurement and attribution services. Because they track consumers extensively and have the advantage of scale, Google and Facebook are able to better show their effectiveness compared to other platforms and attract larger advertisers' budgets. Moreover, they collect both demographic and behavioural data from users who access third-party services.

b) 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 2018, 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 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.

c) 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.

54. Overall, Google and Facebook collect many types of high-quality data from across the web and other sources at scale, combine all this data together and use it to compile accurate user profiles, on which basis they provide precise targeting capabilities to advertisers.[13] Compared with Google and Facebook, we consider that other platforms' data and targeting capabilities are relatively limited to user data from their own services, and are extremely limited in their

[13] 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. In the second half of the study, we will investigate this further.

ability to collect data about consumers on third-parties' websites and apps and combine it with their own first-party data.

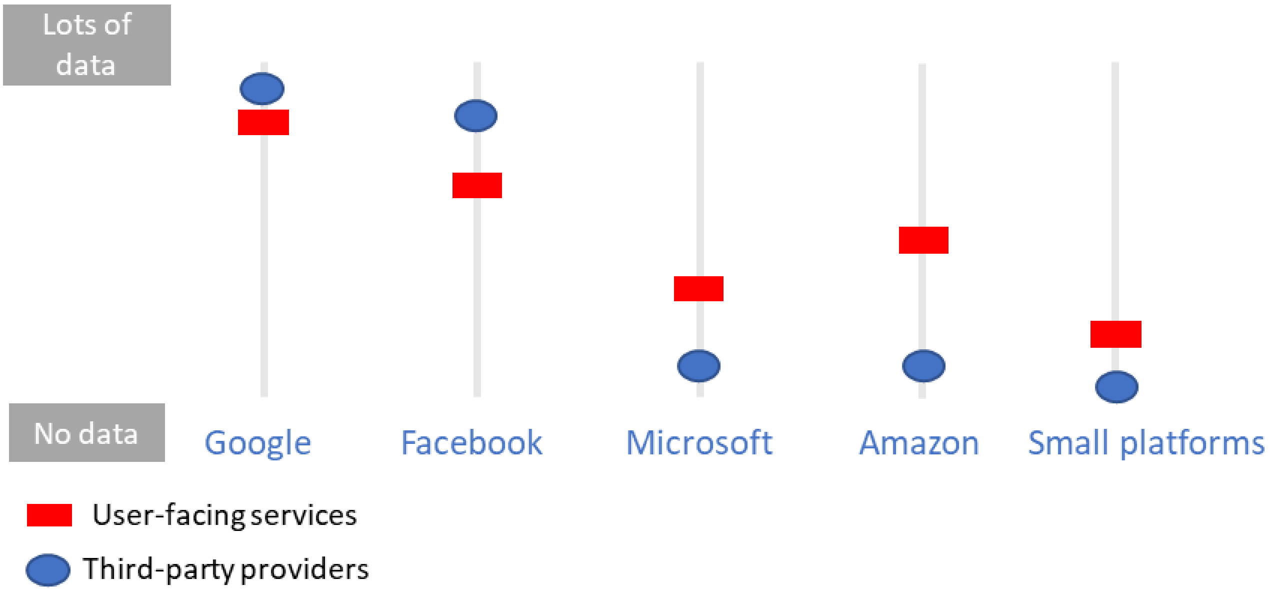
55. Amazon collects a large high-quality dataset from consumers of its owned and operated services (such as its online shopping, Prime Video, Kindle, Amazon Music, etc.). However, at this stage we are of the view that this data is more limited in breadth compared to Google and Facebook, as it relates to consumers interactions in a retail environment. Moreover, Amazon collects and uses data from third-party providers to a more limited extent than Google and Facebook. For example, data collected from Amazon pixels placed on an advertiser's website can only be used by that specific advertiser for its Amazon advertising campaigns. It also does not include information received about consumers' activities from third-party publisher websites in its interest-based ads profiles used for targeting ads.

56. 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 gathers also search and contextual data through its search engine, the amount of this data it can collect is significantly limited given that in 2018 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).

57. At this stage we understand that although most recent entrant platforms such as Twitter, TikTok, Pinterest and Snap possess high-accuracy data about their consumers, this is limited by the reach of their consumer-facing services and their inability to collect extensive off-platform data.

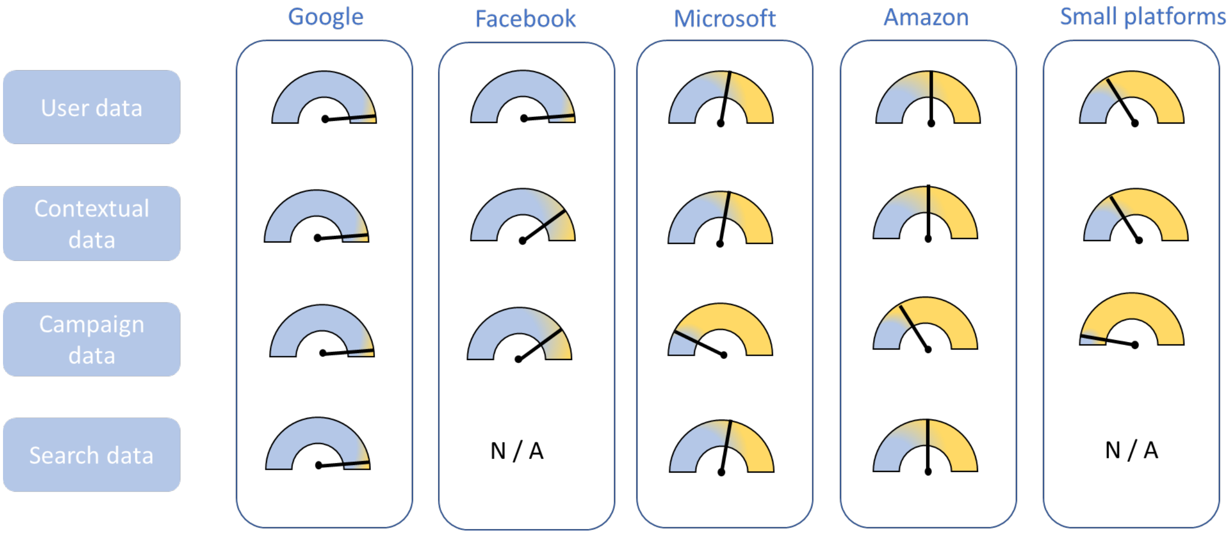
58. The figures below illustrate our current understanding on the volume and types of data that certain large platforms and a group of smaller platforms possess.

Figure E.2: 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.

Figure E.3: Illustration of the scale of data collection by certain platforms, split by types of data



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

59. In the next two sections we discuss the implications of data concentration for competition in the provision of consumer-facing services and advertising services.

Role of data in search

60. Data is often an important input to many products and services provided by online platforms both in search and social media. In this subsection we discuss the data used in search and, referring also to evidence discussed in Chapter 3, assess the extent to which this data is a barrier to entry and expansion in the provision of consumer-facing services. In doing so, we draw on academic literature, submissions from parties and other research. The role of data in social media is discussed in Chapter 3.

61. Search engines help people to find useful information and what they are looking for from the vast amounts of available information on the internet. Modern search engines have moved beyond organising webpages, and now also provide search results for other goods and services, including news, maps, videos, images, shopping, etc. They also provide information relevant to queries often in the form of quick references and answers (eg unit conversions, info boxes, etc.), so consumers may find what they are looking for without needing to navigate to any of the webpages listed in the results.

62. To provide a high-quality search engine that is attractive to consumers, search engine providers use a range of data:

a) Web crawling and indexing – some search engines collect data on websites and webpages, primarily using automated software called 'web crawlers' but also using information that is submitted by webmasters (eg sitemaps and crawl requests) that want their site to be found. These search engines then organise this information into an index from which the relevant and useful results can be returned to queries.

b) click-and-query data – search engines constantly refine their algorithms and processes for selecting relevant results that are responsive to a query (serving) and presenting them in a format which is most useful (ranking, eg with the most relevant and useful result at the top of the page). Having a large amount of data about what consumers search for, which results they select, and whether they spend time on the web page or return quickly ('bounce') back to the search engine's results page helps the search engine to:

- work out the intent of the consumer behind the query, particularly if the query contains spelling mistakes or words with several meanings (homonyms); and

- obtain feedback and learn which results consumers think are relevant and useful for their query.

c) Contextual data – in the context of search engines contextual data, described above, such as the user's location and their search history (particularly their immediate prior searches), can be used to personalise results to the consumer.

d) Additional data feeds – data from third-party publishers and other sources for certain queries, such as those relating to sport scores, exchange rates and weather forecasts.

63. We discuss how this data is used by search engines below.

Web-crawling and indexing

64. In order to return relevant results in response to a range of consumer queries, search engine algorithms must be able to draw on indices that cover a very wide range of relevant webpages, and that provide an up-to-date picture of the content of those webpages.

65. To crawl the web for new or updated webpages, search engines follow links from other known webpages and use data on known webpages' URL addresses and the links on webpages. Search engines also make use of crawl requests and sitemaps submitted by webmasters (ie people who are responsible for maintaining websites). Webmasters often want their websites to be found on search engines, and it is likely that webmasters will prioritise submitting crawl requests and notifying updates to leading search engines, such as Google, and may neglect to do so for smaller search engines. This may be a source of competitive advantage for large incumbent search engines relative to smaller search engines.

66. Search engines record and organise data and metadata collected from crawlers on the content of webpages to form an index, from which relevant search results are drawn. This data can include the title of a webpage, the words it contains and their location within the webpage, as well as metadata on the author of the page and the time the page was last updated.

67. The key challenges associated with crawling and indexing are the costs of this activity, as set out in Chapter 3, and the fact that some webmasters block or prevent web crawlers from crawling their webpages.

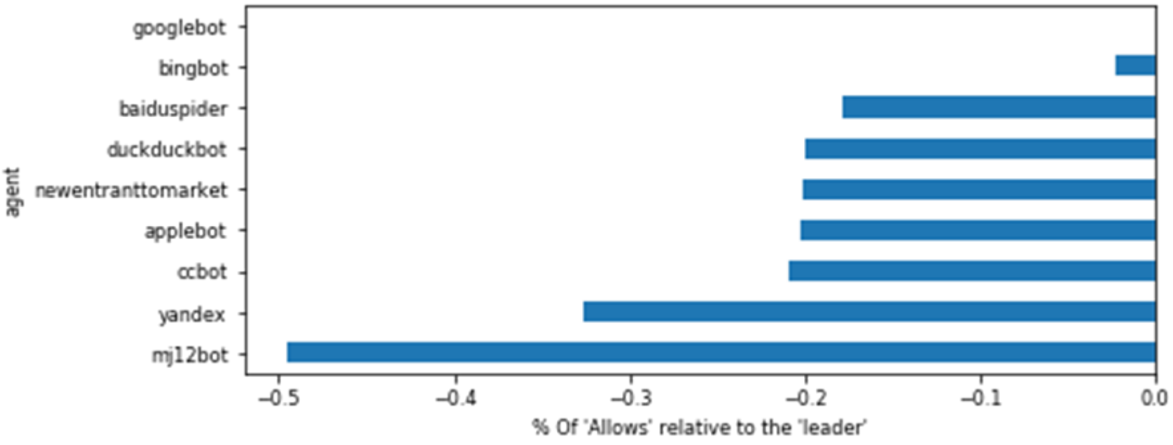
68. Webmasters can permit crawling by some or all bots through a 'robots.txt' file. Reasons for not permitting certain bots to crawl include prevention of fraud

and avoiding the latency problems and increased bandwidth costs that can result from too many automated bots crawling a website. DuckDuckGo submitted that 'small as well as large websites regularly include this blocking code'. When this occurs, search engine providers that operate web-crawlers can contact webmasters to seek a change of policy. However, we heard that this entailed significant effort and cost.

69. To test the significance of webmasters blocking the crawlers of smaller or new entrant search engines, we analysed a sample of 57 million domains from Common Crawl and found that 60% of sites in the sample hosted a robots.txt.[14, 15] This means that almost two thirds of sites in the sample are implicitly blocking or limiting crawling.

70. Figure E.4 below shows the relative access to sites for different bots as the percentage less than the 'leader', the bot with 'access' to the most sites (ie Googlebot).

Figure E.4: Relative access to sites for different bots



Source: CMA

71. The figure shows that Google's bot has the greatest access, followed by Bing's, with DuckDuckGo and new entrants more frequently denied access to

[14] The data for this sample were taken from Common Crawl's July 2019 archive of robots.txt. Common Crawl describe themselves as a non-for-profit dedicated to providing a copy of the internet to researchers, companies and individuals at no cost for research and analysis. Their monthly samples are intended to be representative of the internet.

[15] We set a rule to each robots.txt such that a bot is (i) 'allowed' if it had complete access to every part of a website (ii) 'denied' if it had no access to any part of a website and (iii) 'denied partially' if it had access to some parts but not others. A bot is considered as having 'access' to a domain if it is either (i) allowed or (iii) partially denied access to the domain. The following bots were compared: (i) Googlebot: Crawler for Google's searchable index, (ii) Bingbot: Crawler for Microsoft Bing search engine (iii) DuckDuckBot: DuckDuckGo's web crawler (iv) Yandex: Crawler for Yandex, a Russian search engine; (v) Baiduspider: Crawler for Baidu, a Chinese search engine; (vi) Applebot: Crawler for Apple; (vii) MJ12Bot: Crawler for a UK based specialist search engine; (viii) CCBot: Crawler for Common Crawl, the source of this dataset; and (ix) NewEntrantToMarket: a fictitious crawler, used to assess how a new crawler would behave at the time the robots.txt were collected

sites. The effect is relatively small, with entrants having access to approximately 0.2% fewer sites than Google and Bing. This is ten times greater than the difference between Google and Bing; where Bingbot has access to 0.02% fewer sites than Googlebot.

72. This suggests that the bot for a new entrant honouring[16] all robots.txt instructions is disadvantaged compared to incumbent bots for large search engines (eg Googlebot, Bingbot). A new entrant to the search market honouring robots.txt files may therefore be less able to return relevant content for searches. Similarly, the discovery of new domains relevant to search may be limited as links to those sites which allow crawling may be hosted on blocked pages. This would result in a further reduction in the coverage of their index compared to a competitor with greater access. As discussed in Chapter 3, we would welcome further evidence and views on whether crawler-blocking is a significant barrier to the development of at-scale web-indices.

Click-and-query data

73. As explained above, when a consumer enters a query, the search engine uses algorithms to select the most relevant result from its index. Search engines collect and store aggregated click-and-query datasets containing information about what consumers searched for and how they interacted with the results that they were served.

74. Click-and-query data helps search engines to understand how well their product is performing and to identify and test potential improvements, such as changes to ranking and spelling correction algorithms. It is primarily used to enhance the quality of the search algorithm, as results that consumers have clicked on previously in response to the same or similar queries are a useful signal of the relevance of a webpage. Useful interaction data may include, for example, dwell time on the search results page, whether the consumer scrolled down the page, whether the mouse hovered over a particular element, what links the consumer clicked and whether they hit the back button after following a link.

75. Other uses of click-and-query data include the development of natural language processing models to better understand the natural meaning of a query and suggest corrections to common spelling mistakes, as well as the

[16] The Robots Exclusion Protocol has not been made an official internet standard which has led to differences in how developers have interpreted it over the years. Further robots.txt files are explicitly ignored by some crawlers. For example, the Internet Archive have stated that they ignore robots.txt files as they do not serve their archival purposes and instead respond to removal requests.

construction of logical relationships among queries to suggest related searches.

76. In addition to the evidence discussed in Chapter 3, we have also reviewed some of the key papers that examine the performance of search engines depending on the click-and-query data available and found:

a) In relation to the relative importance of serving and ranking over other data, the literature is split. Some authors attach more weight to crawling and indexing and others to serving and ranking.[17]

b) The extent to which the quantity of user data, and click-and-query data in particular, affects the quality of the serving and ranking's outcome depends on the extent of personalisation. We can distinguish two cases:

- Changes in ranking quality not related to result personalisation. While there is relatively limited empirical evidence that investigates the issue, the literature that exists appears unanimous in claiming that additional click-and-query data at query level improves the quality of ranking.[18] The size of this effect, however, varies greatly depending on the scale of the data held for individual queries. Indeed, the empirical evidence finds rapidly diminishing returns to scale, hence making this effect more relevant for rare queries (which make up a substantial share of the queries submitted to search engines).[19]

- Changes in ranking quality related to result personalisation. The empirical evidence on this point is split, with some authors finding a positive effect and others finding no effect of larger quantities of click-and-query data at individual-user level on the quality of ranking.[20]

Other data feeds

77. There is other data used by search engines which may hinder rivals' ability to enter and expand in the provision of search services to consumers. This includes user data about webpages and other non-user data.

[17] See Lerner, AV (2014); 35. Varian, HR (2014), 'Beyond big data', Business Economics, pp27-31; Schaefer, M, Sapi, G and Lorincz, S (2018), 'The Effect of Big Data on Recommendation Quality. The Example of Internet Search', DICE.

[18] See Schaefer, M, Sapi, G and Lorincz, S (2018); He, D, Kannan, A, Liu, TY, McAfee, RP, Qin, T and Rao, JM (2017), 'Scale Effects in Web Search', International Conference on Web and Internet Economics pp294-310.

[19] See He, D, Kannan, A, Liu, TY, McAfee, RP, Qin, T and Rao, JM (2017).

[20] See Schaefer, M, Sapi, G and Lorincz, S (2018); Chiou, L and Tucker, C (2017), 'Search engines and data retention: Implications for privacy and antitrust', National Bureau of Economic Research.

User data about webpages

78. Search engines use location data to personalise results for queries that consumers generally want to refer to their specific location, such as ‘restaurants’ and ‘weather’. As discussed in Chapter 3, Google may have a particular advantage in obtaining location data due to its strong position in mobile search, derived from Android devices which send data directly to Google, and due to its other apps like Google Maps/Waze which are popular also on non-Android devices.

79. In addition, Google said that its search algorithms may use various other types of data to tailor search results and maintain a positive consumer experience. These include user search and web history and user data from other Google products and services. However, it also said that the most significant signal for providing search services is the text of the query itself. Consumer-specific personalisation is only a mildly important signal for search queries. Many queries are not affected at all by personalisation signals, even if historic data are available.

Non-user data about webpages

80. Search engines make use of data about webpages to determine whether a webpage is likely to be high-quality. In general, and for any specific query, relevant webpages that are of higher quality are more likely to be returned in results pages and in higher positions.

81. For example, Google’s PageRank algorithm counts the number and quality of links to a page to estimate how important a webpage is, as more important webpages are likely to receive more links from other pages. Google also employs human reviewers to make judgements about the quality of a webpage and its authors.[21] Google also takes account of various technical quality measures, such as the page’s load time and error rates.

Summary of role of data in search

82. In order to compete effectively, search engines must be able to access consumers and deliver relevant results to a wide range of queries. The key inputs to achieving relevant results include click-and-query data and a web-index.

83. Currently, only Google and Bing maintain an at-scale English-language web-index. Other search engines use search results from either Google or

[21] Google’s search quality rater guidelines.

Microsoft, which they access through syndication agreements. Crawling and indexing the web represents a significant cost for search engines and crawler-blocking is an issue that smaller search engines may encounter.

84. The evidence that we have reviewed indicates that having access to a greater volume of consumers and click-and-query data enables search engines to deliver more relevant results. The strength of this effect varies depending on the type of query entered and the frequency with which that query is seen. Both Google and Microsoft said that a substantial proportion of queries that they see are uncommon or new. This indicates that, for a substantial proportion of search queries, there is a material benefit from receiving more observations of those queries and situations.

85. Overall, our initial assessment is that the greater scale of English-language queries seen by Google is likely to support its ability to deliver more relevant search results compared to its competitors, especially in relation to uncommon queries and fresh queries (ie queries where the user intent changes over time). Given the importance of search relevance to consumers, lack of comparable scale in click-and-query data is likely to limit the ability of other search engines to compete with Google.

Role of data in digital advertising

86. Data has a key role in digital advertising as it is an essential input used to provide targeting digital advertising to consumers and measurement, verification 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.

87. 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 advertisement and affect publishers' and intermediaries' revenues. The value of data may be different for different types or sources of data and, if there is differential data access, may have competition implications.

88. Platforms compete with one another and with other publishers in the open display market and use data to provide targeting advertising and measurement services. Access to data or certain types of data and the way such data flows within platforms and between intermediaries in the open display market may therefore be an important driver of competition and determinant of large platforms' market power. Data flows can also raise

privacy concerns if consumers are unaware of what data is shared with which market participant and are not in control.

89. The following subsections discuss the role of data in targeted digital advertising and in verification, measurement and attribution services, the value of data and data flows. In doing so, we draw on the academic literature and parties' submissions.

Targeting digital advertising

90. 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).

91. 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

92. 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.

93. 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.

94. Contextual advertising typically uses relatively limited user data such as search terms, device, location and language, in order to show ads in the right size, format and language. However, advertisers can target audiences (eg demographics, affinity, in-market, similar audiences, etc), content and devices

(computer, mobile, etc) in a way that increasingly blurs the line between contextual and personalised advertising.[22]

Personalised advertising

95. 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 from a variety of sources into profiles about consumers, which often include information about demographics, interests, home and work addresses, online and offline behaviours.

96. 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.

97. Platforms 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.

98. 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

[22] As explained in the Google support page here.

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').[23] Each is described in more detail below.[24]

Demographic audience

99. 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) 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;

b) a consumer's stated or inferred age to enable age-based targeting; and

c) a consumer's stated or inferred gender, education and language to enable targeting on this basis.

100. 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.[25,26] These details allow advertisers to refine their bidding strategies and improve efficiencies.

101. 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.

[23] Terms such as 'custom audience' and 'lookalike audience' are not used consistently across the industry. In this annex we define custom audience as retargeting.

[24] 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.

[25] About audience targeting - Google Ads Help.

[26] This audience can be used only for advertising on certain properties, namely Gmail, Discovery, Search, Shopping and Video.

Interest-based audience

102. 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.

103. 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.[27] 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.[28]

104. In order to create interest-based segments, Amazon uses information such as search 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.[29] We understand that Microsoft uses primarily data entered by consumers in a logged-in environment and infer consumer's age, gender and interests to build targeting segments.

105. 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 as set out by Google's internal document in paragraph 107 below, because consumers might not be currently in the market for a specific product or service.

In-market audience

106. 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.

[27] 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), signals 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 (e.g. Shopify).

[28]About audience targeting - Google Ads Help.

[29] Amazon only obtains and processes personal data in accordance with its privacy notice.

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

[✂]

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

Retargeting

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

a) 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.[30] 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.

b) 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.

c) 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.

[30] 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. However, customer data does not necessarily have to be hashed.

d) 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).

110. 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

111. 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 to construct a model to predict and identify which other consumers are similar to the seed group.[31]

112. 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.

113. 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

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

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

114. 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.

115. 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, measurement and attribution

116. The second main purpose of data in digital advertising is to provide verification, measurement and attribution services. Platforms and publishers use data to provide advertisers with information on the success and reach of their advertising across inventories. Advertisers can then decide how to adjust their spend on different advertising media within a campaign and to make high-level marketing budget allocation decisions for future advertising campaigns on the basis of this information.

117. Measurement and verification services allow advertisers, publishers and platforms to confirm whether and the extent to which ads were shown to the right number and kinds of people.

118. 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.

119. 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.

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

121. 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.

122. 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. Search advertising is often used more for increasing conversions and less for raising awareness or sentiment whereas display advertising can be used for both. As a result, attribution analysis is more commonly linked to search advertising.

123. To measure conversions, advertisers need to be able to track consumer actions online (and to some extent offline). As described above, Google and Facebook tags are widely available on advertiser websites and apps, much more so than other platforms. In addition, Google's mobile data allows it to track consumer actions offline to some extent (eg to identify store visits) and the integration of Google Analytics and Google Ads also gives advertisers a clearer view of which adverts are translating into specific conversions and enables them to adjust which ads are served and which advertising channels to bid on accordingly.

124. A number of responses from advertisers and media agencies have pointed to the difficulty in accurately measuring attribution across different platforms. In order to be able to track the 'user journey', to see what adverts they are exposed to and to assess the impact of those adverts, an advertiser needs to be able to follow a consumer across the internet. Tracking a user's activities within a walled garden is possible and can be done with accuracy. However, it is not possible to do the same across different platforms and there is a tendency to regard the major platforms as 'silos'. Either certain amount of manual intervention is required by advertisers and media agencies to 'de-duplicate' the information they receive from the different platforms or they can use alternative approaches such as using data from advertisers' ad servers.

125. The ability to show effectiveness of advertising is an important driver of advertisers' decisions on how to allocate their advertising spend across publishers and platforms. In Microsoft's experience, when advertisers use their tags they are more likely to continue to advertise with Microsoft. A study LinkedIn have conducted 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 are able 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%.

126. Nonetheless, Microsoft's trackers cover less than 1% of UK websites and Microsoft has continued to account for a very small proportion of the UK search advertising revenue (more details in Appendix C). On the contrary, 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 result, they are better able to

demonstrate the effectiveness of using their platforms relative to others. This finding is supported by advertisers' submissions. For example, John Lewis said that Google is able to offer a better suite of measurement services, by being able to track consumers both online and offline, than its competitors in search advertising as they can connect data across their ecosystems of Search, Android phones and the Chrome browser.

The value of data

127. The value of data used in digital advertising may be different for parties involved in the supply chain: publishers, advertisers and intermediaries. Data has value if advertisers can use it to better target consumers, increasing the returns on their investments and the efficiency of their advertisement; this in turn impacts publishers and intermediaries that are paid by advertisers. Data also has value as platforms use it to improve their consumer-facing services.

128. There are two main reasons why it is helpful to measure how valuable or useful different types of data are:

a) Firstly, in order to understand the extent of competitive advantage that access to data or 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, how data flows and whether it is easily shared between market participants, as well as data attributes such as freshness and velocity.[32]

b) Secondly, in order to make informed decisions about potential remedies that may change the availability of data to market participants.

129. In the sections above, we have described how data is used to improve consumer-facing services and provide targeting capabilities to advertisers. This section discusses the evidence on the value of data and, in particular: (i) the value of personal advertising relative to non-targeted advertising and (ii) the incremental value of additional data. In doing this assessment, this section draws on the academic literature and on parties' quantitative and qualitative submissions.

[32] In its statement of scope response, Google submitted that the role of data in digital advertising is indeed a fundamental question. It submitted that 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.

Value of personal advertising

130. Despite measurement and attribution challenges, the academic literature seems to concur that targeted advertising is effective and useful to advertisers.[33] This is also supported by Google internal documents, one of which, used to pitch YouTube advertising services to advertisers, says, [✂].

131. 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:

a) 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.

b) 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 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

[33] See Marotta, V, Vibhanshu, A and Acquisiti A (2019), 'Online Tracking and Publishers' Revenues: An Empirical Analysis'.

advertisers.[34] For example, the aforementioned Google document shows 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.[35]

c) Thirdly, 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 campaign intended to drive increased footfall to the 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 the position along the 'marketing funnel'. Although this is also affected by the campaign's objective that advertisers want to meet, data may be more valuable the closer it is to the bottom of the funnel:[36]

- Data indicating consumers' purchase behaviour is very desirable. 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.

[34] 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.

[35] These are global statistics that do not refer solely to the UK.

[36] 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.

- 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 geo-location data may vary significantly. Broad-based geo-location data, such as a postal code, is helpful to narrow down the gap of desired consumers. However, geo-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 customers 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.

132. At an aggregate level, recent empirical evidence consistently finds that publisher revenues increase as a result of targeted advertising as opposed to contextual ads; however, the magnitude of this impact is unclear.[37] 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. Google has recently run its own experiment to test this result and we describe it below.

133. Google ran a Randomized Controlled Trial (RCT) that involved disabling all third-party cookie information for a small fraction of randomly selected consumers (the treatment group) so that these consumers would see only non-personalised ads.[38,39] Google compared the revenue that publishers earned from this group to that generated from another group of randomly selected consumers who continued to see personalised ads (the control group). The experiment was run from May 2019 for 96 days and covered

[37] See Marotta et al (2019).

[38] The experiment is described here and the methodology is described here.

[39] 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, [✂] the matching of cookies was prevented. 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.

[✂]% of global traffic for each of the control and treatment group.[40,41] This experiment covered traffic from all browsers and that went through both Google and other intermediaries.[42] It also affected both third-party inventory as well as Google's inventory, namely Play Browse, Gmail and a small portion of YouTube that serves non-video ad display.

134. The results indicate that UK publishers earned [50%-65%] less revenue when advertisers could not target their advertising.

135. We have discussed with Google the methodology and, at this stage, have the following comments on the results of this experiment and the implications for an assessment on the value of data:

a) There may be an adverse selection issue where advertisers seeing a bid request with no cookie information might believe that there is a higher probability of fraud (eg the site is being viewed by a 'bot'). In this case the experiment would overestimate the value of data as it would actually be assessing the value of cookies.

b) The results of the RCT seem to suggest that by removing cookies advertisers decrease their willingness to pay and in the medium to long term their advertising budget. However, advertisers may well redistribute their spending from personalised ads (ie targeted personal advertising) to non-personalised advertising. If this is the case, the results are likely to over-estimate the negative impact on publisher's revenue in the medium or long run.

c) For the treatment group, third-party cookies were disabled and Google logged-in users were excluded. This means that the interpretation of the experiment is focused on the removal of third-party cookies, but does not allow us to assess whether additional information on consumers, such as Google's internal data, may affect publisher revenues.

136. Although the experiment is likely to have some limitations and may well over-estimate the impact on publisher's revenue for the reasons stated above, our preliminary view is that it gives a helpful indication of the overall value of personal targeting advertising. In the second phase of the market study, we

[40] Although the traffic affected is small in percentage terms, in absolute terms the amount of traffic was significant.

[41] This experiment was run for a 96-day interval from May to August 2019 and affected all programmatic demand from Exchange Bidders, Authorised Buyers, Google Display Ads and DV360.

[42] It covered traffic going through Google Ad Manager's serving system (Google SSP) as well as traffic handled by non-Google SSP solutions, but only to the extent that such traffic passed through Google's DSP ad serving system.

plan to assess in more details the RCT experiment, including interrogating the data, in order to evaluate the merit of our observations above.

137. In summary, the academic literature as well as the evidence we have collected to date suggest that data is valuable to advertisers, in that it allows them to better target consumers and improve the efficiency of their advertisement, and to publishers, as they can earn greater revenue than otherwise. We have taken this into account in our assessment of the proposed interventions.

Value of incremental data

138. An important feature of data that might affect its value and, as a result of platform's 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.

139. 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. This appears to suggest that the value of incremental data is positive as by increasing the information available about one consumer platforms can target consumers more accurately. 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.[43]

140. In the second half of the study we plan to explore further the value of incremental data and whether it decreases with the volume of data.

Flow and matching of data

141. As discussed above, market participants collect data from a wide variety of sources and use it to provide targeting digital advertising and verification, measurement and attribution services to advertisers. In order to do so, they match user identifiers and share these and the associated user information with each other. As a result, there are significant transfers of data that take

[43] 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.

place within the large platforms and between intermediaries in the open display market. We discuss below these differences and their implications.

142. Google, Facebook and other platforms such as Twitter are referred to as 'walled gardens', in the sense that they require advertisers to use their ad management tools in order to purchase inventory on their owned and operated properties. Walled gardens are 'closed ecosystems' in which the platform provides a complete end-to-end technical solution for advertisers and publishers, and advertisers and publishers are restricted in their ability to choose other technical solutions.[44] Google and Facebook do work with a number, but not all, third-party measurement and verification providers and have also restricted the user-level data that they have access to. For instance, Google has restricted access to log-level user data, which prevents the advertiser from conducting reach, frequency and conversion analysis quickly and easily. Some analysis is still possible via Google's Ads Data Hub but it is no longer in real time and can take a number of days after campaign activity to be produced. This has implications, discussed in Chapter 5 and Appendix H, for advertisers' ability to address potential conflicts of interest that the platform faces, and for competition between providers of verification and measurement services.

143. In principle, there are relatively few transfers of personal data between different entities for a transaction conducted entirely within a vertically integrated platform. For example, in a situation where an advertiser buys advertising on YouTube, which is owned by Google, using Google's demand side platform and Google's exchange, Google does not transfer personal data to the advertiser. Instead, Google analyses and uses its data about individuals on behalf of advertisers to serve targeted advertising. Google may communicate aggregate campaign data to advertisers, for the purposes of reporting, verification and measurement. If the advertiser is making use of Google's retargeting or lookalike targeting services, it may transfer its first-party data about its customers to Google.

144. Although there are few transfers of data between different legal entities, large platforms like Google are nevertheless pooling data about consumers from their different services and across third-party webpages and apps to offer personal targeting of advertising. Two of the principles of processing personal data set out in Article 5 of the GDPR are that personal data should be collected for specified, explicit and legitimate purposes ('purpose limitation') and adequate, relevant and limited to what is necessary for these purposes

[44] Open and closed systems can both have advantages for competition and efficiency. In *The Economics of Open and Closed Systems*, the CMA and Autorité de la Concurrence outlined the advantages for competition and efficiency from both open and closed systems.

('data minimisation'). Articles 13 and 14 of the GDPR further provide that information about the purposes for processing should be provided to the data subject, including when a data controller decides to 're-use' the data for a different purpose.

145. At this stage, we have not assessed whether Google and Facebook have successfully met their obligations under GDPR for purpose limitation and data minimisation. However, it seems that Google and Facebook are using data collected to provide consumer-facing services and while providing these services, to also sell, target and measure the effectiveness of advertising. A discussion is in Chapter 4 of the update paper.

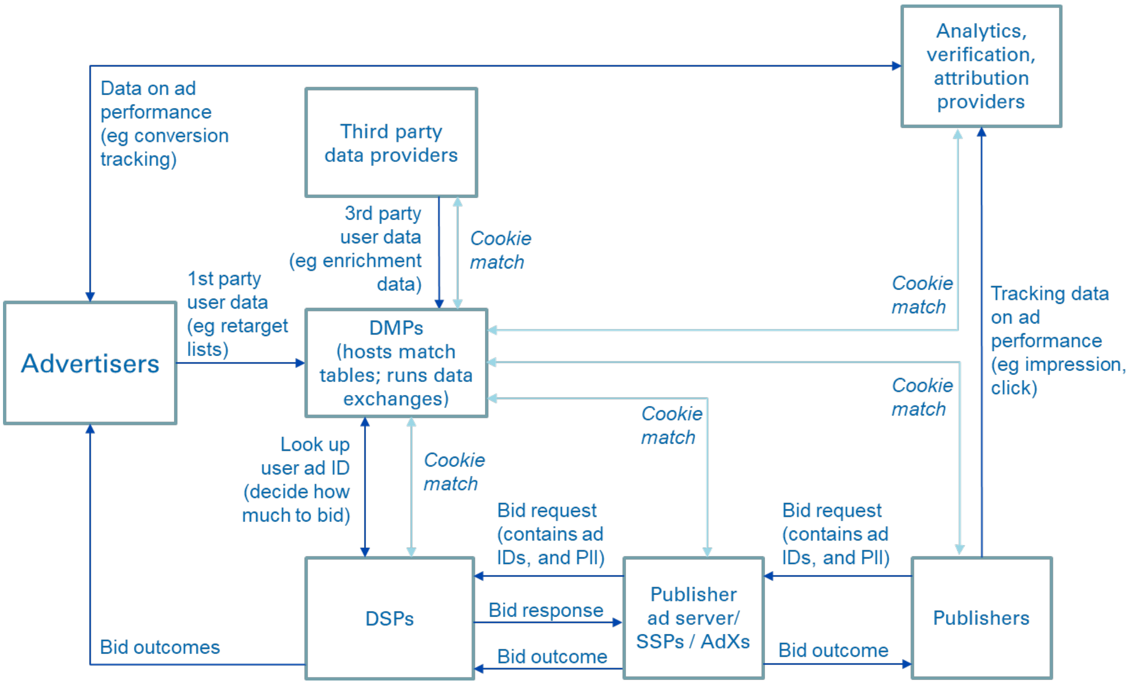
146. The CMA has also not assessed whether publishers in general are compliant with these obligations under GDPR to inform consumers that data is collected for the purpose of providing advertising.

147. In contrast to digital advertising transactions that make use of a vertically integrated 'walled garden' closed system, there can be very significant transfers of personal data in the digital advertising ecosystem in open display advertising (real-time bidding).

148. As explained in the section on personal advertising, market participants use various techniques and technologies to identify consumers, assign identifiers for consumers (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.

149. Figure 5 below shows that for open display advertising, bid requests are sent to potentially hundreds of intermediaries and advertisers, and bid requests contain personal data (and sensitive personal data), including a user ID, the webpage or app that the person is using, highly specific information about the consumer's exact combination of device and settings (which can be used to 'fingerprint' consumers), location data (including IP address or GPS coordinates), and, in some cases, audience segments that the consumer is associated with (a more detailed description of open display advertising is in Appendix H).

Figure E.5: Example flows of data in open real-time bidding (browser impression)


Analytics, verification, attribution providers
Data on ad performance (eg conversion tracking)
Third party data providers
3rd party user data (eg enrichment data)
Cookie match
1st party user data (eg retarget lists)
Advertisers
DMPs (hosts match tables; runs data exchanges)
Cookie match
Tracking data on ad performance (eg impression, click)
Cookie match
Cookie match
Look up user ad ID (decide how much to bid)
Cookie match
Bid request (contains ad IDs, and PII)
Bid request (contains ad IDs, and PII)
DSPs
Bid response
Publisher ad server/ SSPs / AdXs
Publishers
Bid outcomes
Bid outcome
Bid outcome

Source: CMA

150. The ICO has expressed concern that, in its view, consent is the only valid basis for processing the data in bid requests, and that it is not possible for consumers to provide valid consent to their personal data (including sensitive personal data) to be shared with an unknowable (from the perspective of the consumer) and large number of parties, with unknowable controls and security measures.

151. Some information is transferred in real-time, within the bid requests. Other information is transferred at lower frequencies (eg daily) using batch files. Information about people is bought and sold on data marketplaces, and advertisers and publishers can use data management platforms to hold and organise this information. Currently we do not have much information on the volume or kinds of information that is bought and sold about people in the UK, and this is an area that we intend to explore further in the second half of the study with data management platforms (DMPs), data brokers, marketplaces and exchanges.

152. The different access and quality of data between large platforms like Google and Facebook relative to other market participants affects competition in digital advertising. In general, it is difficult and costly for advertisers to assemble information on consumers, compared to Google/Facebook, from their own first-party data and other (non-Google/Facebook) third-party data

providers. Google and Facebook have high reach, as many people use them, and they have very valuable data on consumers that can be used to target and measure the effectiveness of advertising. This data is high-quality (eg often close to real-time, likely to be true, very informative about consumer interests and intent). As explained, Google and Facebook do not sell this data on open data exchanges, so the only way for advertisers to get (indirect) access to it and use it for targeting is to use their ad management tools.

153. Targeting and measurement is more effective if advertisers have a more complete picture of individuals. Google and Facebook are more likely to be able to provide this because typical match rates between datasets held by multiple parties are significantly below 100%, and vary greatly with different third-party sources of data. If an advertiser has to get and match several data from different sources, it will probably only get a patchy and incomplete picture. By contrast, Google and Facebook represent a single unified source of high-quality data, covering many people. By using them, advertisers only need to do a single match with their own first-party data, which is likely to have a good match rate because of Google and Facebook's high reach.

154. In addition, people use multiple browsers and multiple devices. This fragments the picture that advertisers have of those people, reducing the effectiveness of targeting and preventing certain kinds of measurement (eg whether people go to a physical store after they see an ad online). Because people typically log into Google and Facebook, they are well-placed to identify all the devices belonging to one individual, since if two different devices log into the same account, it is very likely that the same person owns both devices.

Conclusions

155. Data gives platforms a competitive advantage in the provision of both consumer-facing and digital advertising services. In the provision of search services to consumers, having access to a greater volume of users and click-and-query data enables search engines to deliver more relevant results. This is particularly important for uncommon or new queries. For this reason, the greater scale of English-language queries seen by Google is likely to support its ability to deliver more relevant search results compared to its competitors, especially in relation to uncommon and fresh queries.

156. In digital advertising, platforms provide targeting capabilities which allow advertisers to retarget their current and potential customers as well as reach wider audiences. For these purposes, detailed data on consumers'

demographics, interests, preferences and behaviours is most valuable to predict consumers' potential response to advertising.

157. Platforms also provide measurement, 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 is very important to demonstrate their effectiveness in digital advertising.

Google and Facebook have a competitive advantage because they collect a large amount and variety of high-quality data from their widely used consumer-facing services and their broad coverage of third-party sites and apps. Rival platforms such as Microsoft and Amazon have access to some detailed high-quality data about consumers and other types of data, but this is largely limited to their own services or does not extend widely to the rest of the internet. In order to compete, these platforms as well as intermediaries in the open display market can supplement their own data with data from other market participants, such as data management platforms. However, this requires rival platforms and intermediaries to extensively share data between one another and match different identifiers in order to compile rich user profiles. This process is less precise than the matching that takes place uniquely within closed ecosystems and we have heard that it also gives rise to privacy concerns caused by data flows.