Data Collection in the Age of Surveillance Capitalism

Douglas C. Schmidt[1]

1. Surveillance Capitalism and Google Data Collection

The proliferation of personal computing devices—coupled with advances in network bandwidth, geolocation, "big data" analytics, and targeted advertising—has led to the rise of "surveillance capitalism," where large high-tech companies monetize personal data they collect by monitoring the online behavior and physical movement of their users around the world. The term "surveillance capitalism" was popularized by Shoshana Zuboff in the mid-2010's (i). Surveillance capitalism constitutes practices conducted by large high-tech companies to monitor, predict, and control the behavior of consumers who use their products and platforms, either directly or indirectly.

Awareness of how personal data is collected by large high-tech companies is intensifying, with many people beginning to realize who is invested in knowing their online behaviors. For example, Facebook's enabling of data collection by Cambridge Analytica and the effect on the 2016 elections changed the conversation in the United States of America. Many people were rightly surprised by that vast collection and non-transparent use of their personal data, as well as the ways in which this information can be used to monitor, predict, and control many aspects of their lives.

Earlier means of monitoring, predicting, and controlling people relied on fear and shame to maintain order in society and ensure conformity with the interests of those in power. A classic example is the totalitarian regime represented by "Big Brother" in Orwell's dystopian novel 1984. In contrast, surveillance capitalism is based on a more subtle—and ultimately more pernicious—form of monitoring, predicting, and controlling that Zuboff calls "Big Other," which are digital networks trained to monitor and shape users actions remotely based on machine learning technologies that are regulated by neither national nor international laws (ii).

At the forefront of today's surveillance capitalism practices is Google, which is the world's largest digital advertising company (iii). Among its many popular products and platforms, it provides Chrome, which is the most popular web browser (iv), Android, which is the most popular mobile platform (v), and Google Search, which is the most popular search engine (vi) worldwide. Google's video platform (YouTube), email service (Gmail), and map application (Google Maps) each have over 1 billion monthly active users (vii). The tremendous popularity and reach of Google's products enable it to collect detailed information about user online behaviors and physical movements, which it applies to target users with paid advertising. Google's revenues increase significantly as it refines its targeting technology and data collection methods.

Google collects data on its users by various means. The most obvious are "active" means, where the user directly and consciously conveys information to Google. For example, users may login to any of Google's popular applications, such as YouTube, Gmail, Search etc. Less obvious ways that Google collects data are its "passive" means, whereby its products are instrumented to gather information while they are running, possibly without user knowledge. For example, Google's passive data gathering methods arise from platforms (e.g., Android and Chrome), applications

[1] This paper is an extension of my earlier Vanderbilt Technical Report entitled "Google Data Collection" (xii).

(e.g., Search, YouTube, and Maps), publisher tools (e.g., Google Analytics, and AdSense), and advertiser tools (e.g., AdMob, and AdWords), all of which are discussed later in this paper.

Although prior studies listed in Appendix A have explored ways in which Google tracks users via active means, the extent and magnitude of Google's passive data collection has largely been overlooked. To help fill this gap, this study provides a detailed view of Google's data collection methods by drawing on (1) data intercepted as it is sent to Google servers while Google or 3rd-party products are used and (2) Google's My Activity (viii) and Takeout (ix) tools, which describe information collected during the use of Google's user-facing products.

The remainder of this paper is organized as follows: Section 2 presents an example "day in the life" scenario, where a real user with a new Google account and Android phone goes through her daily routine to determine how much Google learns about her personal interests during a single day of typical Internet usage; Section 3 shows how Google's Android and Chrome products frequently send user location and device-related information to Google without direct user knowledge; Section 4 shows how purportedly anonymous advertising identifiers can get connected with a user's Google identity; Section 5 analyzes the results of an experiment that show how Google collects considerable information through its advertiser and publisher products, even if a user does not interact with any key Google applications; and Section 6 presents concluding remarks and summarizes lessons learned from research conducted for this paper.

2. Evaluating Google Data Collection in Practice

To demonstrate the multitude of touchpoints between Google's data collection methods and an individual, as well as to showcase the extent of information collected during these interactions, an experiment was conducted where a user carried an Android The LG X Power Android mobile device during a day's activities. The device used for this experiment was

- wiped clean to the default factory settings and given new SIM card to ensure that no data was stored on the phone and that phone numbers could not be linked with any past usage and
- wiped by conducting a factory data reset, which deleted all login data for Google services and other accounts, system and app data and settings, all downloaded apps, digital rights management licenses, music, images, documents and backups, and other usage data from the internal storage of the device.

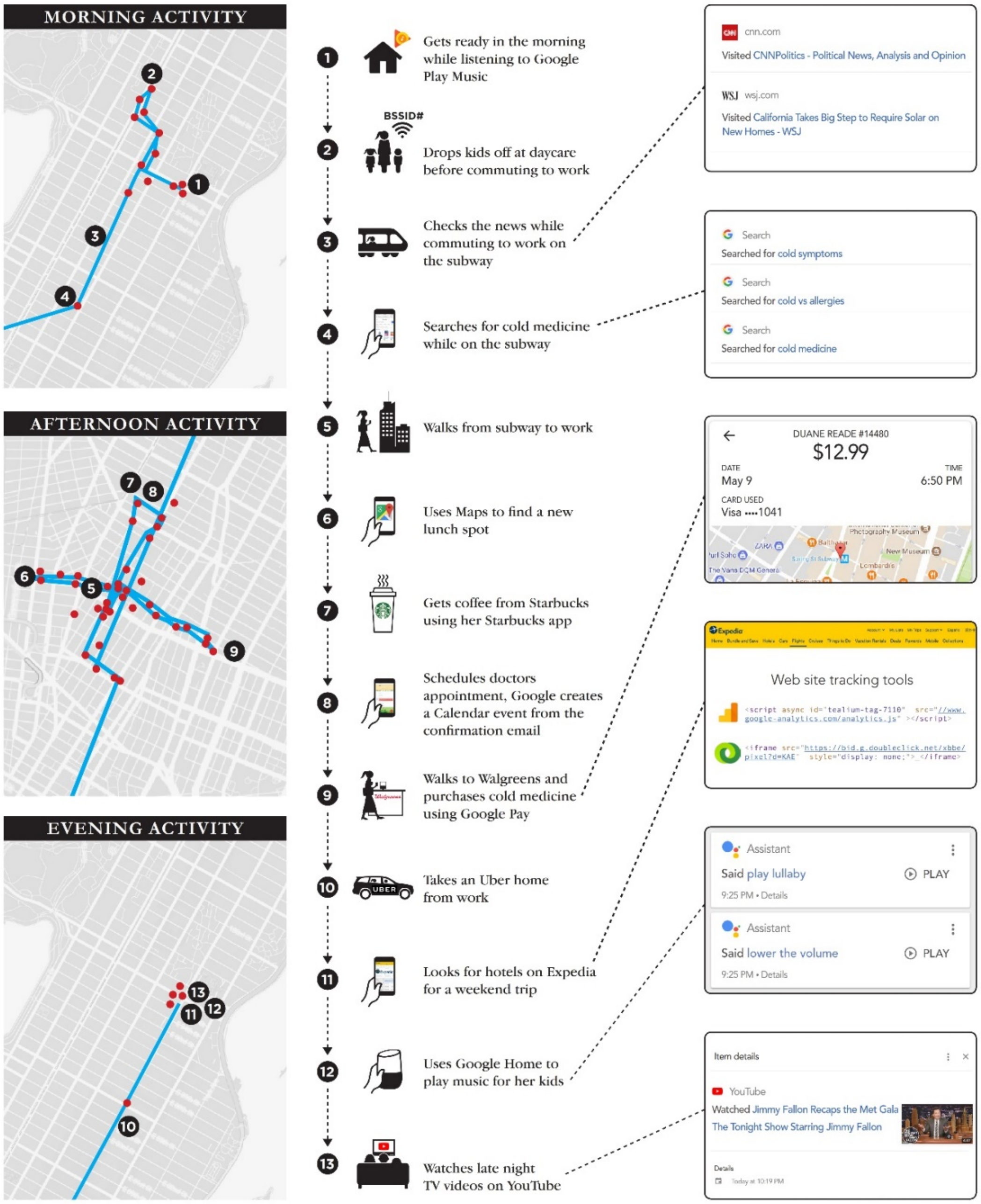
A new Google account was created (username "Jane"), so that Google had no prior knowledge of the user and had no advertising interests associated with the account. This user then went about a normal day using the mobile phone associated with this new Google account.

The data collected by Google was checked using two tools: My Activity (x) and Takeout (xi). The My Activity tool shows data collected by Google from any Search-related activities, use of Google applications (e.g., YouTube video plays, Maps search, Google Assistant), visits to 3rd-party web pages (while logged in to Chrome) and clicks on advertisements. The Google Takeout tool provides more comprehensive information about all historical user data collected via Google's applications (e.g. it contains all past email messages on Gmail, search queries, location collection, and YouTube videos watched). The collected data and used it to depict key information collection events were synthesized in the form of a "day in the life" of the user "Jane," as shown in Figure 1.

In the activity shown in Figure 1, as well as throughout the rest of this paper, the collected data is categorized in the following two broad subgroups:

- ***Active data collection***, where information is directly exchanged between the user and a Google product. An example of active data collection occurred when Jane submitted a keyword in the Search tool bar and that search query was collected by Google.
- ***Passive data collection***, where information is exchanged in the background without any obvious notification to the user. An example of passive data collection occurred when Jane's location was sent to Google after she entered a search query.

Analysis of key touch points during a normal day in Jane's life indicated that the "passive" data collection events outnumbered the "active" events by approximately two-to-one (xii).



1 Gets ready in the morning while listening to Google Play Music

2 Drops kids off at daycare before commuting to work

3 Checks the news while commuting to work on the subway

4 Searches for cold medicine while on the subway

5 Walks from subway to work

6 Uses Maps to find a new lunch spot

7 Gets coffee from Starbucks using her Starbucks app

8 Schedules doctors appointment, Google creates a Calendar event from the confirmation email

9 Walks to Walgreens and purchases cold medicine using Google Pay

10 Takes an Uber home from work

11 Looks for hotels on Expedia for a weekend trip

12 Uses Google Home to play music for her kids

13 Watches late night TV videos on YouTube

cnn.com
Visited CNNPolitics - Political News, Analysis and Opinion

wsj.com
Visited California Takes Big Step to Require Solar on New Homes - WSJ

Search
Searched for cold symptoms

Search
Searched for cold vs allergies

Search
Searched for cold medicine

DUANE READE #14480
$12.99
DATE
May 9
TIME
6:50 PM
CARD USED
Visa ••••1041

Web site tracking tools

```
<script async id="tealium-tag-7110"  src="//www.google-analytics.com/analytics.js" ></script>
<iframe src="https://bid.g.doubleclick.net/xbbe/pixel?d=KAE"  style="display: none;">_</iframe>
```

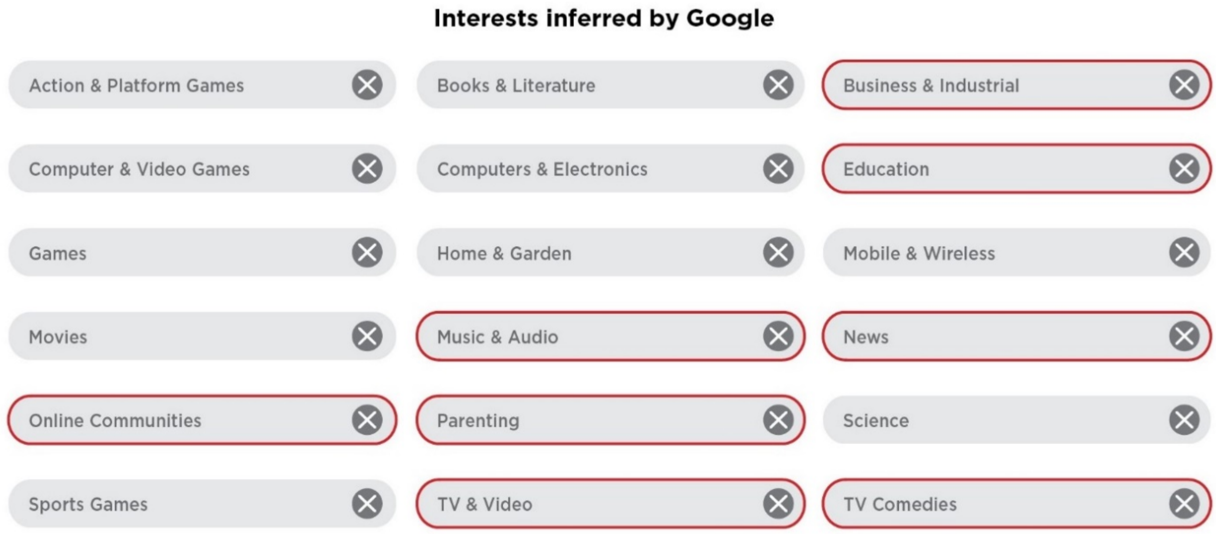
Assistant
Said play lullaby ⊙ PLAY
9:25 PM • Details

Assistant
Said lower the volume ⊙ PLAY
9:25 PM • Details

Item details
YouTube
Watched Jimmy Fallon Recaps the Met Gala
The Tonight Show Starring Jimmy Fallon
Details
Today at 10:19 PM

Figure 1: A Day in the Life of Jane, Highlighting Touchpoints Where Google Collects Data

Google analyzes the collected data to assess user interests and then applies it to target users with appropriate ads. For example, Google provides a list of interests it inferred from a user's activities, available via the "topics you like" section in the Google's Ad Personalization webpage (xiii). Figure 2 shows the list that Google associated with Jane's account after a day's worth of activity. In total, Google attributed 18 interests to Jane, eight of which (shown by colored borders) closely matched Jane's usage activities.[2]



Interests inferred by Google

Action & Platform Games	Books & Literature	Business & Industrial
Computer & Video Games	Computers & Electronics	Education
Games	Home & Garden	Mobile & Wireless
Movies	Music & Audio	News
Online Communities	Parenting	Science
Sports Games	TV & Video	TV Comedies

Figure 2: Google's Assessment of Jane's Interests at the End of One Day

3. Google Data Collection via the Android and Chrome Platforms

Android and Chrome are Google's primary platforms that collect significant user data due to their extensive reach and frequency of usage. By January 2018, Android captured 53% of the total US mobile OS market, where as Apple iOS held 45% (xiv). As of May 2017, there were more than 2 billion monthly active Android devices worldwide (xv).

Google's Chrome browser held more than 60% share of all internet browser usage in the world with over 1 Billion monthly active users as reported in the 2017 Q4 10K filing (xvi). Both platforms facilitate the use of Google and 3rd-party content (e.g. 3rd-party websites and 3rd-party apps) and hence provide Google access to a wide range of personal, web activity, and location information.

3.1. Personal Information and Activity Data Collection

To download and use apps from Google Play Store on an Android device, a user must have (or create) a Google Account, which becomes a key gateway through which Google collects personal information, including username, email, and phone number. If a user registers for services such as Google Pay (xvii), Android also collects the user's credit card information, zip code, and birth date. All this information becomes part of a user's personal information associated with their Google Account.

[2] It's unclear as to why other interests that have no connection with Jane's activities during the day show up in this list, though perhaps Google uses historical analysis of similar interests from other users to create associated recommendations.

While Chrome does not mandate sharing additional personal information gathered from users, it does have the capability to capture such information. For example, Chrome collects a range of personal information via its form "autofill" feature, and such form fields typically include user name, address, phone number, login name, and passwords (xviii). Chrome stores form fill information on a user's local drive, however, if the user logs in to Chrome using Google Account and enables its "Sync" feature, this information gets sent to and stored on Google servers. Chrome could also learn about the language(s) a person speaks during their interactions with its translate feature, which is enabled by default (xix).

In addition to personal data, Chrome and Android send Google information about a user's web browsing and mobile app activities, respectively. Any webpage visit is automatically tracked and collected under user credentials by Google if the user is login to Chrome. Chrome also collects information about a user's browsing history, passwords, website-specific permissions, cookies, download history, and add-on data (xx). Android sends periodic updates to Google servers, including device type, cell service carrier name, crash reports, and information about apps installed on the phone (xxi). It also notifies Google whenever any app is accessed on the phone (e.g., Google knows when Android users access their Uber apps).

3.2. User Location Data Collection

Android and Chrome platforms meticulously collect user location and movement information using a variety of sources. For example, a "coarse location" assessment can be done by using GPS coordinates on an Android phone or through a network's IP address on a desktop/laptop device. The user location accuracy can be improved further ("fine location") through the use of nearby cell tower IDs or via scanning the device-specific base-station identifiers or basic service set identifiers, assigned to the radio chipset used in nearby Wi-Fi access points.[3] Android phones can also use information from the Bluetooth beacons registered with Google's Proximity Beacon API (xxii). These beacons indicate a user's geolocation coordinates and can also pinpoint exact floor levels in buildings (xxiii).

It's hard for an Android mobile user to "opt out" of location tracking. For example, even if a user turns off an Android device's Wi-Fi, the device's location is still tracked via its Wi-Fi signal. To prevent such tracking, Wi-Fi scanning must be explicitly disabled in a separate user action. The ubiquity of Wi-Fi hubs has made location tracking quite frequent. For example, during a short 15-minute walk around a residential neighborhood, an Android device sent nine location requests to Google. The request collectively contained ~100 unique identifiers for public and private Wi-Fi access points. Google also ascertains with a high degree of confidence whether a user is standing still, walking, running, bicycling, or riding on a train or a car by tracking an Android mobile user's location coordinates at frequent time intervals in combination with the data from onboard sensors (such as an accelerometer) on mobile phones.

3.3. Assessing Passive Data Collection by Google via the Android and Chrome Platforms

Active data that Android or Chrome platforms collect and send to Google as a result of users' activities on these platforms can be assessed through Google's MyActivity and Takeout tools. Of potentially greater interest, however, is the passive data that these platforms collect, which goes beyond location data and which remains largely unrecognized by the users. To assess the type and frequency of occurrence of such collection in greater detail an experiment was conducted that

[3] To understand how location data is sent to Google servers in more depth, we analyzed the data traffic from a mobile phone from a user in motion.

monitored traffic data sent to Google from mobile phones (both Android and iPhone). For comparison's sake, this experiment also included the analysis of data sent to Apple via an iPhone device.

For simplicity, the phones were kept stationary, with no user interaction. On the Android phone a single Chrome browser session remained active in the background, whereas on the iPhone the Safari browser was used. This configuration provided an opportunity for systematic analysis of the background collection that Google performs purely through Android and Chrome, as well as collection that occurs in the absence of those (i.e., from iPhone device), without any additional collection requests generated by other products and applications (e.g., YouTube, Gmail, and app usage).

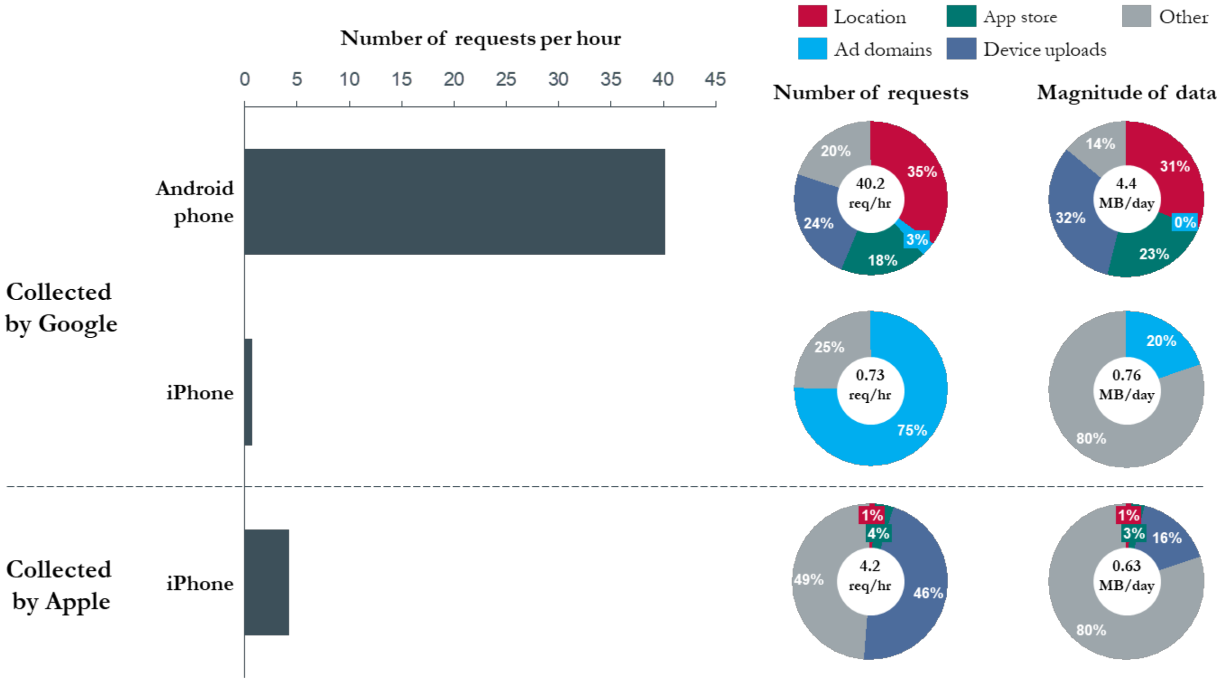


Figure 3: Traffic Data Sent from Idle Android and iPhone Mobile Devices

Figure 3 shows a summary of the results obtained from this experiment. The x-axis indicates the number of times the phones communicated with Google (or Apple) servers, whereas the y-axis indicates the phone type (Android or iPhone) and server domain type (Google or Apple) with which data packets were exchanged by the phones. The colored legend describes the broad categorization of the type of data requests identified by the domain address of the server.

During a 24-hour time period the Android device communicated ~900 data samples to a variety of Google server endpoints. Of these, ~35% (or approximately 14/hour) were location-related. Google ad domains received only ~3% of the traffic since the mobile browser was not actively used during the collection period. The remaining ~62% of communications with the Google server domains were roughly divided between requests to Google's Play App store, Android's uploads of device-related data (such as crash reports and device authorization), and other data which were predominantly in the category of Google services background calls and refreshes.

Figure 3 shows that the iPhone device communicated with Google domains at more than an order of magnitude (~50x) lower frequency than the Android device, and that Google did not collect any user location during the 24-hour experiment timeframe via iPhone. This result highlights the fact that the Android and Chrome platforms play an important role in Google's data collection.

Moreover, the iPhone device's communication with Apple's servers were 10x less frequent than the Android device's communications with Google. Location data made up a very small fraction (~1%) of the net data sent to Apple servers from the iPhone, with Apple receiving location-related communications once every day on an average.

Android phones communicated 4.4 MB of data per day (~130MB per month) with Google servers, which is 6x more than what Google servers communicated through the iPhone device. This experiment was conducted using a stationary phone with no user interaction. As a user becomes mobile and starts interacting with their phone, the frequency of communications with Google's servers increases considerably. Section 5 of this paper analyzes results from such an experiment.

4. Google Data Collection via Publisher and Advertiser Technologies

A major source for Google's user activity data collection stems from its publisher- and advertiser-focused tools, such as Google Analytics, DoubleClick, AdSense, AdWords, and AdMob. These tools have tremendous reach, e.g., over 1 million mobile apps use AdMob (xxiv), over 1 million advertisers use AdWords (xxv), over 15 million websites use AdSense (xxvi), and over 30 million websites use Google Analytics (xxvii).[4]

There are two main groups of users of Google's publisher- and advertiser-focused tools:

- ***Website and app publishers***, which are organizations that own websites and create mobile apps. These entities use Google's tools to (1) make money by allowing the display of ads to visitors on their websites or apps, and (2) better track and understand who is visiting their websites and using their apps. Google's tools place cookies and run scripts in the browsers of website visitors that help determine a user's identity, track their interest in content, and follow their online behavior. Google's mobile app libraries track use of apps on mobile phones.
- ***Advertisers***, which are organizations that pay to have banner, video, or other ads delivered to users as they browse the Internet or use apps. These entities apply Google's tools to target specific profiles of people for advertisements to increase the return on their marketing investments (better targeted ads generally yield higher click-through rates and conversions). Such tools also enable advertisers to analyze their audiences and measure the efficacy of their digital advertising by tracking which ads were clicked with what frequency and by providing insight into the profiles of people who clicked on ads.

Together, these tools collect information about user activities on websites and in apps, such as content visited, and ads clicked. They work in the background—largely unnoticeable by users. Figure 4 shows some of these key tools, with arrows indicating data collected from users and ads served to users.

The information collected by such tools includes a non-personal identifier that Google can use to send targeted advertisements without identifying the unique individual's personal information. These identifiers can be device- or session-specific, as well as permanent or semi-permanent. To

[4] Google has rebranded AdWords as "Google Ads" and DoubleClick as "Google Ad Manager", though there were no changes instituted in the core product functionalities including information collection by these products. Therefore, for the purpose of this paper the names are kept unchanged to avoid confusion that may occur with related domain names (such as doubleclick.net).

provide users greater anonymity during information collection for ad targeting, Google has recently shifted towards using semi-permanent device unique identifiers, such as GAIDs, which are alphanumeric strings for Android and iOS devices that enable accurate targeted mobile ads (xxviii).

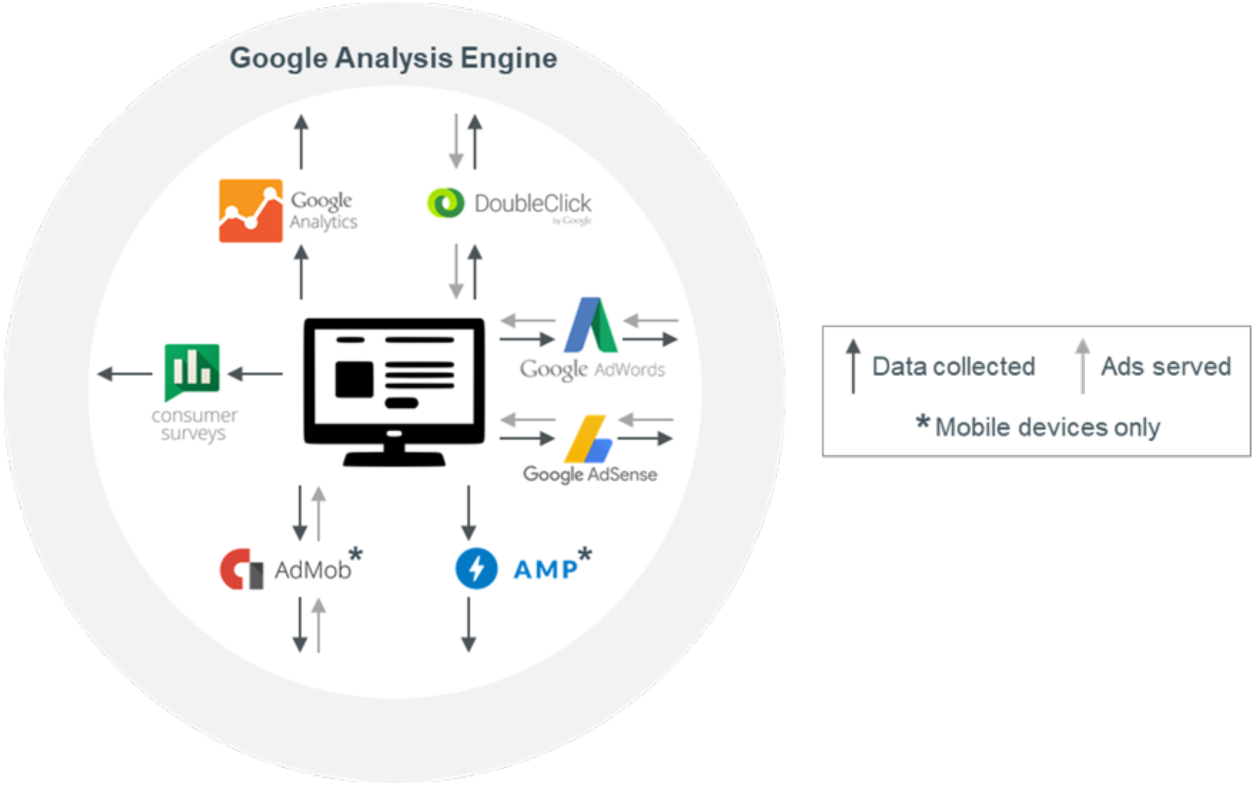


Figure 4: Google Products Aimed at Publishers and Advertisers

4.1. Google Analytics and DoubleClick

Google Analytics and DoubleClick are Google's leading products in user behavior tracking and webpage traffic analyses on desktop and mobile devices. Google Analytics is used by ~75% of the top 100,000 most visited websites (xxix). DoubleClick cookies are associated with more than 1.6 million websites (xxx).

Google Analytics uses short pieces of tracking code (called "page tags") embedded in a website's HTML code (xxxi). After a webpage loads per a user's request, the Google Analytics code calls an "analytics.js" program residing on Google's servers to transfer a "default" snapshot of user data at that moment, which includes visited webpage address, page title, browser information, current location (derived from IP address), and user language settings. Google Analytics scripts use cookies to track user behavior.

The first time when a Google Analytics script is run it generates and stores a browser-specific cookie on the user's computer. This cookie has a unique client identifier or Client ID (xxxii). Google uses the unique identifier to link previously stored cookies that capture a user's activity on a particular domain as long as the cookie does not expire, or the user does not clear the cookies cached on their browser (xxxiii).

While a Google Analytics cookie is specific to the domain of the website that user visits (called a "1st-party cookie"), a DoubleClick cookie is typically associated with a common 3rd-party domain (such as doubleclick.net). Google uses such cookies to track user interaction across multiple 3rd-party websites (xxxiv). When a user interacts with an advertisement on a website, DoubleClick's conversion tracking tools (e.g., Floodlight) places cookies on a user's computer and generates a unique client ID (xxxv). Henceforth, the stored cookie information gets accessed by the DoubleClick server if the user visits the advertised website, thereby recording the visit as a valid conversion.

4.2. AdSense, AdWords, and AdMob

AdSense and AdWords are Google tools that serve ads on websites and in Google Search results, respectively. AdSense collects information about whether an ad was displayed on the publisher's webpage and how the user interacted with the ad, such as clicking an ad or tracking the curser movement over an ad (xxxvi). AdWords enables advertisers to serve search ads on Google Search, display ads on publisher pages, and overlay ads on YouTube videos. To track user click-through and conversion rates, AdWords ads place a cookie on users' browsers to identify the user if they later visit the advertiser's website or complete a purchase (xxxvii).

AdSense and AdWords also collect data on mobile devices. Their ability to get user information on mobile devices is limited, however, since mobile apps do not share cookie data between them, an isolation technique known as "sandboxing."[5] It is therefore hard for advertisers to track user behavior across mobile apps.

To address this issue, Google and other companies use mobile "ad libraries" (such as AdMob) that are integrated into the apps by their developers for serving ads in mobile apps. These libraries compile and run with the apps and send data to Google that are specific to the app to which they belong, including GPS locations, device make, and device model when apps have the appropriate permissions. As observed through the data traffic analyses and confirmed through Google's own developer webpages (xxxviii), such libraries can also send Google user-personal data (such as age and gender) whenever app developers explicitly pass these values to the library.

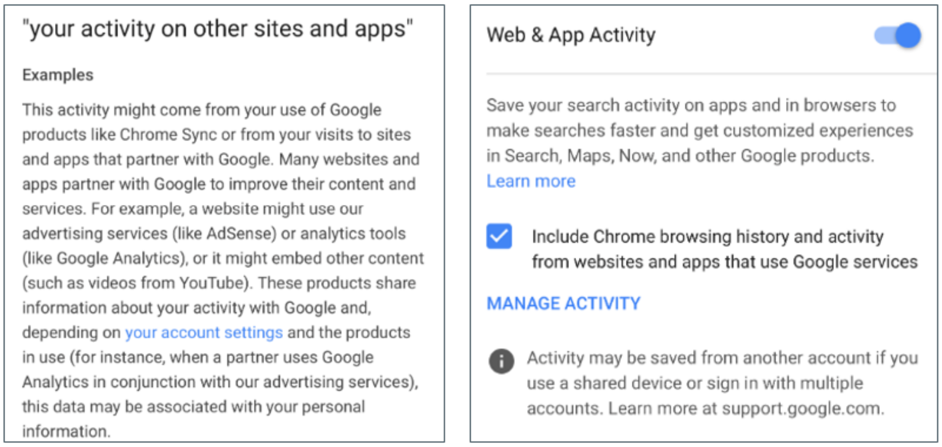
More than 15 million websites have AdSense installed to display sponsored ads (xxxix). Likewise, more than 2 million websites and apps that make up the Google Display Network (GDN) and reach over 90% of Internet users display AdWords ads.

4.3 Association of Passively Collected Data with Personal Information

As discussed above, Google collects data through publisher and advertiser products and associate such data with a variety of semi-permanent, anonymous identifiers. Google itself, however, can associate these IDs with a user's personal information, which is insinuated by statements made in Google's privacy policy, excerpts of which are shown in Figure 5 (xli)(xlii).

The left text box in Figure 5 clearly states that Google may associate data from advertising services and analytics tools with a user's personal information, depending upon the user's account settings. This arrangement is enabled by default, as shown in the right text box. Moreover, an analysis of data traffic exchanged with Google servers (summarized below) identified two key examples (one on Android and the other in Chrome) that indicate how Google can correlate anonymously collected data with users' personal information.

[5] This approach is similar to desktops, where cookies are not shared between browsers.



"your activity on other sites and apps"

Examples

This activity might come from your use of Google products like Chrome Sync or from your visits to sites and apps that partner with Google. Many websites and apps partner with Google to improve their content and services. For example, a website might use our advertising services (like AdSense) or analytics tools (like Google Analytics), or it might embed other content (such as videos from YouTube). These products share information about your activity with Google and, depending on your account settings and the products in use (for instance, when a partner uses Google Analytics in conjunction with our advertising services), this data may be associated with your personal information.

Web & App Activity

Save your search activity on apps and in browsers to make searches faster and get customized experiences in Search, Maps, Now, and other Google products. Learn more

☑ Include Chrome browsing history and activity from websites and apps that use Google services

Activity may be saved from another account if you use a shared device or sign in with multiple accounts. Learn more at support.google.com.

Figure 5: Google's Privacy Page for 3rd-party Website Collection and Association with Personal Info

4.3.1. Mobile Advertising Identifier May Get De-anonymized via Data Sent to Google by Android

Analyses of data traffic communicated between an Android phone and Google server domains suggest how anonymous identifiers (a "GAID" in this case) can get associated with a user's Google Account. Data exchanges with Google servers from an Android phone show how Google can connect anonymized information collected on an Android mobile device via DoubleClick, Analytics or AdMob tools with the user's personal identity. During a 24-hour experiment, data collected from a stationary and dormant Android phone two instances of checkin communications with Google servers were observed.

4.3.2. DoubleClick Cookie ID Gets Connected with User's Personal Info on Google Account

Section 4.3.1 explained how Google can de-anonymize user identity via the passive, anonymized data it collects from an Android mobile device. De-anonymization can also occur on a desktop/laptop device, where anonymized data is collected via cookie-based identifiers (e.g., Cookie ID), which are typically generated by Google's ad and publisher products (e.g., DoubleClick) and stored on a user's local mass storage. The experiment presented below assessed whether Google can connect such identifiers (and hence information associated with them) with a user's personal information. This experiment involved the following ordered steps:

1. Opened a new (no saved cookies, e.g. Private or Incognito) browser session (Chrome or other),

2. Visited a 3rd-party website that used Google's DoubleClick ad network,

3. Visited the website of a widely used Google service (Gmail in this case),

4. Signed into Gmail.

After completing step 1 and 2, as part of the page load process, the DoubleClick server received a request when the user first visited the 3rd-party website. This request was part of a series of requests comprising the DoubleClick initialization process started by the publisher website, which resulted in the Chrome browser setting a cookie for the DoubleClick domain. This cookie stayed on user's computer until it expired or until the user manually cleared coolies via the browser settings.

Thereafter, in step 3, when the user visited Gmail, they are prompted to log in with their Google credentials. Google manages identity using a "single sign on (SSO)" architecture, whereby credentials are supplied to an account service (signified by *accounts.google.com*) in exchange for an "authentication token," which can then be presented to other Google services to identify the users. In step 4, when a user accesses their Gmail account, they are effectively signing into their Google Account, which then provides Gmail with an authorization token to verify the user's identity. The advantage of the extra authentication step is that the user's browser can later use the same authentication token to confirm user identity on other Google services (due to this process a sign-on in any particular Google application enables an automatic sign-on all others in the same browser session).

In step 3 of this sign-on process, a request is sent to the DoubleClick domain containing the authentication token provided by Google and the tracking cookie set when the user visited the 3rd-party website in step 2. Google uses these data to connect the user's Google credentials with a DoubleClick cookie ID. If users do not clear browser cookies regularly, therefore, their browsing information on 3rd-party webpages that use DoubleClick services can be associated with their Google Account personal information.

5. Amount of Data Collected During a Minimal Use of Google Products

This section examines the details surrounding Google's data collection through its publisher and advertiser services. To understand such data collection, an experiment is designed which entailed a user going through her daily life using a mobile phone (akin to "day in the life" described in Section 2), while deliberately *avoiding* the use of any direct Google products (i.e., avoiding Search, Gmail, YouTube, Maps, etc.), except for the Chrome browser.

To make the experiments realistic, various consumer usage studies (xliii)(xliv) were used to form a daily usage profile of a typical mobile user. Thereafter, any direct interactions with Google's products were omitted from the profile. The experiment was replicated on both Android and iOS devices and the HTTPS data sent to Google and Apple servers were monitored and analyzed using a similar method explained in previous sections. The results are summarized in Figure 6.

During the 24-hour time period (which includes the nighttime stationary/dormant timeframe), most calls from the Android phone were made to Google's location and publisher/advertisement service domains (e.g. DoubleClick, Analytics). Google collected user location in ~450 instances, which is ~1.4x times the amount shown by the experiment presented in Section 3, which involved a stationary phone.

Google servers communicated significantly lower number of times with an iPhone device compared to Android (45% less). The number of calls to Google's advertising domains, however, were similar from both devices, which was expected since the usage of 3rd-party webpages and apps was similar on both devices. One notable difference was that the location data sent to Google from an iOS device is practically non-existent. In the absence of Android and Chrome platforms—or the use of any other Google product—Google becomes significantly limited in its ability to track the user location.

The total number of calls to Apple servers from an iOS device was much lower, just 19% the number of calls to Google servers from an Android device. Moreover, there are no ad-related calls to Apple servers, which may stem from the fact that Apple's business model is not as dependent on advertising as Google's. Although Apple does obtain some user location data from iOS devices, the volume of data collected is much (16x) lower than what Google collects from Android.

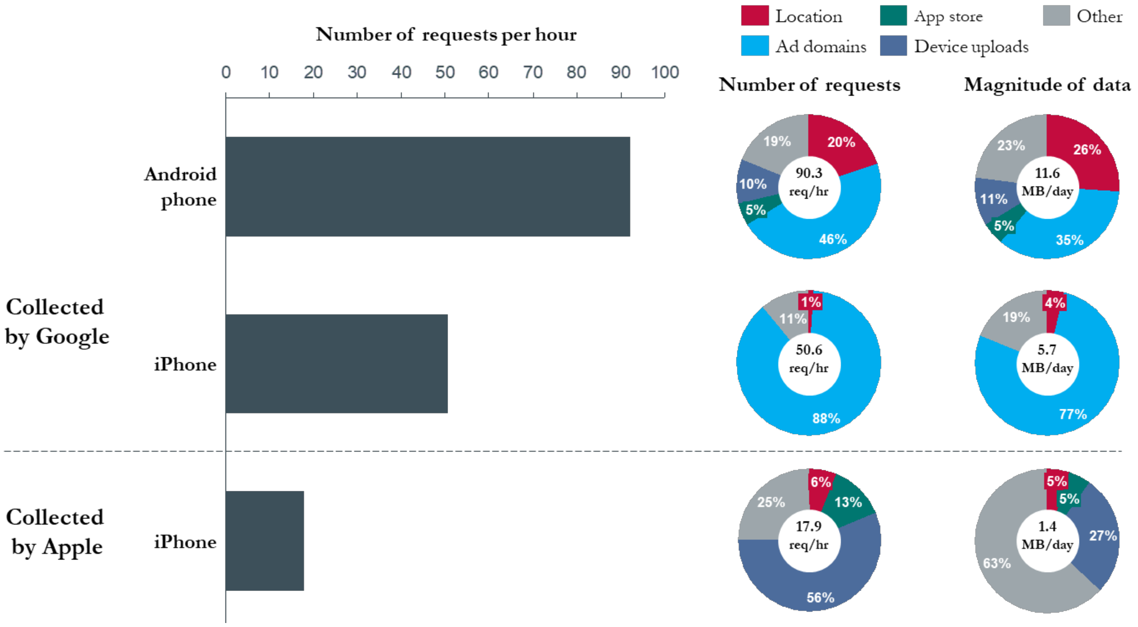


Figure 6: Information Requests from Mobile Devices During a Day of Typical Use

Android phones communicated 11.6 MB of data per day (~350 MB per month) with Google servers. On the other hand, the iPhone device communicated just half that amount. The amount of data particularly associated with Google's ad domains remained quite similar across both devices. The iPhone device communicated an order of magnitude less data to Apple servers than what the Android device exchanged with Google servers.

Even in the absence of user interaction with Google's most popular applications, a user of an Android phone and the Chrome browser still sends substantial amounts of data to Google. Much of this data are associated with location and calls to ad server domains. Although an iPhone user is insulated from Google's location collection in this narrow experiment, Google still captures a similar amount of ad-related data.

6. Concluding Remarks

A large percentage of the world's population directly consume Google products that lead their markets globally and surpass 1 billion monthly active users. These products collect user data via various means that may typically users may not understand. Much of Google's data collection occurs "passively," i.e., while a user is not directly engaged with any of its products. While such information is typically collected without identifying a unique user, Google distinctively possesses the ability to utilize data collected from other sources to de-anonymize such collection.

The following summarize the lessons learned from the research reported in this paper:

a. Google learns a great deal about a user's personal interests during even a single day of typical internet usage. As shown by the "day in the life" scenario in Section 2, Google collected data about Jane at numerous activity touchpoints, such as user location, routes taken, items purchased, and music listened to. Surprisingly, Google collected or inferred over two-thirds of the information through passive means. At the end of the day, Google identified user interests with remarkable accuracy.

b. Android is an essential enabler of data collection for Google, with over 2 billion monthly active users worldwide (xv). Android helps Google collect personal user information, activity on the mobile phone, and location coordinates. Section 3 showed how frequently Android sends Google user location and device-related information passively, i.e., without direct user knowledge.

c. The Chrome browser helps Google collect user data from both mobile and desktop devices, with over 2 billion active installs worldwide (xlv). Section 3 described how the Chrome browser collects personal information (e.g., when a user completes online forms) and sends it to Google as part of the data synchronization process. It also tracks webpage visits and sends user location coordinates to Google.

d. Android and Chrome send data to Google even without *any* user interaction. Experiments in Section 3 showed that a dormant, stationary Android phone running Chrome in the background communicated location information to Google 340 times during a 24-hour period. In fact, location information constituted 35% of all data samples sent to Google. In contrast, Google cannot collect any appreciable data on an iOS Apple device with Safari (where neither Android nor Chrome were used) in the absence of user interaction with the device.

e. Google can use advertising identifiers (which are purportedly "user anonymous" and collect activity data on apps and 3rd-party webpage visits) to connect with a user's Google identity. This connection happens via passing of device-level identification information to Google by Android devices, as discussed in Section 4. Likewise, the DoubleClick cookie ID is another purportedly "user anonymous" identifier that Google can connect to a user's Google Account.

f. After a user starts interacting with an Android phone, passive communication to Google server domains increases significantly, even when the user runs no prominent Google. This increase is driven by data activity from Google's publisher and advertiser products. The results presented in Section 5 suggest that even if a user does not interact with any key Google applications, Google still collects considerable information through its advertiser and publisher products.

g. While using an iOS device, if a user decides to forgo the use of *any* Google product (i.e., no Android, no Chrome, no Google applications), and visits only non-Google webpages, the number of times data is communicated to Google servers still remains surprisingly high. This communication is driven purely by advertiser/publisher services. An experiment in Section 5 shows how the total magnitude of data communicated to Google servers from an iOS device is approximately half of that from an Android device.

Google's data collection practices are particularly problematic in this age of surveillance capitalism. As shown in this paper and related studies, if you use Android mobile devices Google tracks your movements, whether you like it or not (xlviii). If you use popular products like Chrome, Gmail, and Search, Google also knows who you are and what you do, whether you like it or not. Likewise, if you use the Google Photos app, then Google knows who your friends and family are, whether you like it or not. In the words of former Google CEO Eric Schmidt "We know where you are. We know where you've been. We can more or less know what you're thinking about" (xlvi). Eric Schmidt has also pointed out that "Your digital identity will live forever... because there's no delete button" (xlvii), which is yet another reason why Google's vast trove of collected data on their users is particularly unsettling.

Appendix A. Related Work on Google's Data Collection Practices

Table 1 summarizes other studies relating to Google's data collection practices.

Table 1: Summary of other Google data collection studies

Title	Relevant Findings
AP Exclusive: Google tracks your movements, like it or not (xlviii)	Google is tracking users' location even when location services are disabled
Australian regulator investigates Google data harvesting from Android phones (xlix)	Google "harvest" about 1GB of data from Android devices per month
How to Keep Google From Owning Your Online Life (l)	It is hard for the average consumer to avoid Google products
Google tracking phones even when they are disconnected? (li)	Google tracks phones even when phones are "disconnected" (no SIM cards, airplane mode, Wi-Fi off)
Google collects Android users' locations even when location services are disabled (lii)	Google collects Android location when location services are turned off
Google is permanently nerfing all Home Minis because mine spied on everything I said 24/7 (liii)	The Google Home mini was saving recording when the device was not activated with "OK Google" (Google claims to have resolved the issue)
Online Tracking: A 1-million-site Measurement and Analysis (liv)	Google can track users ~80% of websites using its cookies
Why Do Android Smartphones Guzzle the Most Data? (lv)	Android devices consume more data (2.2GB/month) than other smartphones
Data leakage from Android smartphones (lvi)	Android passes anonymous IDs along with devices IDs such as Mac address and IMIE

References

i. Zuboff, Shoshana. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization," *Journal of Information Technology* (2015) 30, 75–89.

ii. Zuboff, Shoshana. "The Surveillance Threat Is Not What Orwell Imagined," https://time.com/5602363/george-orwell-1984-anniversary-surveillance-capitalism/

iii. "Google and Facebook tighten grip on US digital ad market," *eMarketer*, Sept. 21, 2017, available at https://www.emarketer.com/Article/Google-Facebook-Tighten-Grip-on-US-Digital-Ad-Market/1016494

iv. "Market share or leading internet browsers in the United States and worldwide as of February 2018," *Statista,* February 2018, available at https://www.statista.com/statistics/276738/worldwide-and-us-market-share-of-leading-internet-browsers/

v. "Global OS market share in sales to end users from 1st quarter 2009 to 2nd quarter 2017," *Statista*, August 2017, available at https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/

vi. “Worldwide desktop market share of leading search engines from January 2010 to October 2017,” *Statista,* Feb. 2018, available at https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

vii. Google 10K filings with the SEC, 2017, available at https://abc.xyz/investor/pdf/20171231_alphabet_10K.pdf

viii. “My Activity,” *Google*, available at https://myactivity.google.com/myactivity

ix. “Download your data,” *Google*, available at https://takeout.google.com/settings/takeout?pli=1

x. “My Activity,” Google, available at https://myactivity.google.com/myactivity

xi. “Download your data,” Google, available at https://takeout.google.com/settings/takeout?pli=1

xii. Schmidt, Douglas C. “Google Data Collection,” Vanderbilt University Technical Report #ISIS-20-201, August 15, 2018 https://www.dre.vanderbilt.edu/~schmidt/PDF/google-data-collection.pdf.

xiii. “Ads personalization,” *Google*, last accessed on August 15, 2018, available at https://adssettings.google.com/authenticated

xiv. “Subscriber share held by smartphone operating systems in the United States from 2012 to 2018,” *Statista,* May 2018, available at https://www.statista.com/statistics/266572/market-share-held-by-smartphone-platforms-in-the-united-states/

xv. Burke, David. “Android: celebrating a big milestone together with you,” *Google,* May 17, 2017, available at https://www.blog.google/products/android/2bn-milestone/

xvi. Google 10K filings with the SEC

xvii. “Google Chrome privacy whitepaper,” *Google,* March 6, 2018, available at https://www.google.com/chrome/privacy/whitepaper.html#payments

xviii. “Google Chrome privacy whitepaper,” *Google,* March 6, 2018, available at https://www.google.com/chrome/privacy/whitepaper.html#autofill

xix. “Google Chrome privacy whitepaper,” *Google,* March 6, 2018, available at https://www.google.com/chrome/privacy/whitepaper.html#translate

xx. “Google Chrome Privacy Notice,” *Googl*e, March 6, 2018, available at https://www.google.com/intl/en/chrome/browser/privacy

xxi. https://policies.google.com/privacy?hl=en&gl=us#infocollect

xxii. “Google beacon platform, proximity beacon API,” *Google,* last accessed on August 15 2018, available at https://developers.google.com/beacons/proximity/guides

xxiii. “Google beacon platform, proximity beacon API,” *Google,* last accessed on August 15 2018, available at https://developers.google.com/beacons/proximity/guides

xxiv. “AdMob by Google,” *Google*, last accessed on August 15 2018, available at https://www.google.com/admob/

xxv. “Hear from our happy customers,” *Google,* last accessed on August 15 2018, available at https://adwords.google.com/home/resources/success-stories/

xxvi. “Websites using Google Adsense,” *BuiltWith,* last accessed on August 15 2018, available at https://trends.builtwith.com/websitelist/Google-Adsense

xxvii. Google Analytics usage statistics,” *BuiltWith,* April 2018, available at https://trends.builtwith.com/analytics/Google-Analytics

xxviii. “Best practices for unique identifiers,” *Google,* last accessed on August 15 2018, available at https://developer.android.com/training/articles/user-data-ids

xxix. Google Analytics usage statistics," *BuiltWith,* April 2018, available at https://trends.builtwith.com/analytics/Google-Analytics

xxx. "DoubleClick market share," *Datanyze,* last accessed on August 15 2018, available at https://www.datanyze.com/market-share/ad-exchanges/doubleclick-market-share

xxxi. GA or other tags can also be implemented through Google Tag Manage (GTM) without changing the functionality of the page tag

xxxii. "Cookies and user identification," *Google*, last accessed on August 15 2018, available at https://developers.google.com/analytics/devguides/collection/analyticsjs/cookies-user-id

xxxiii. "Cookies and user identification," *Google*, last accessed on August 15 2018, available at https://developers.google.com/analytics/devguides/collection/analyticsjs/cookies-user-id

xxxiv. "DoubleClick search help," *Google,* last accessed on August 15 2018, available at https://support.google.com/ds/answer/7298761?hl=en

xxxv. "DoubleClick search help," *Google,* last accessed on August 15 2018, available at https://support.google.com/ds/answer/2903014?hl=en&ref_topic=6054260

xxxvi. "AdSense help, privacy and security," *Google,* last accessed on August 15 2018, available at https://support.google.com/adsense/answer/9897?hl=en

xxxvii. "Evaluating ad performance on the Search Network," *Google,* last accessed on August 15 2018, available at https://support.google.com/adwords/answer/2404037?hl=en; "About conversion tracking," *Google,* last accessed on August 15 2018, available at https://support.google.com/adwords/answer/1722022?hl=en

xxxviii. "Google APIs for Android," *Google,* last accessed on August 15 2018, available at https://developers.google.com/android/reference/com/google/android/gms/ads/doubleclick/PublisherAdRequest.Builder

xxxix. "Websites using Google Adsense," *BuiltWith,* last accessed on August 15 2018, available at https://trends.builtwith.com/websitelist/Google-Adsense

xl. "Google Ads Help," *Google,* last accessed on August 15 2018, available at https://support.google.com/google-ads/answer/2404191?hl=en

xli. "Google privacy and terms," *Google,* last accessed on August 15 2018, available at https://policies.google.com/privacy/example/your-activity-on-other-sites-and-apps

xlii. "Google https://myaccount.google.com/activitycontrols

xliii. "Nielsen provides topline U.S. web data for March 2010," *Nielsen,* April 2010, available at http://www.nielsen.com/us/en/insights/news/2010/nielsen-provides-topline-u-s-web-data-for-march-2010.html

xliv. 55% of web traffic comes from mobile devices. Eric Enge, "Mobile vs desktop usage: mobile grows but desktop still a big player in 2017," *Stone Temple,* April 2017, available at https://www.stonetemple.com/mobile-vs-desktop-usage-mobile-grows-but-desktop-still-a-big-player-in-2017/

xlv. Lardinois, Frederic. "Google says there are now 2 billion active Chrome installs," *TechCrunch,* Nov. 10, 2016, available at https://techcrunch.com/2016/11/10/google-says-there-are-now-2-billion-active-chrome-installs/

xlvi. Orlowski, Andrew. "Google's Schmidt: We know what you're thinking", https://www.theregister.co.uk/2010/10/04/google_ericisms/

xlvii. Levine, Yasha. "Googles For Profit Surveillance Problem," pando.com/2013/12/16/googles-for-profit-surveillance-problem.

xlviii. Nakashima, Ryan. "AP Exclusive: Google tracks your movements, like it or not," *AP*, August 13, 2018, available at https://apnews.com/828aefab64d4411bac257a07c1af0ecb xlix. Davis, Anne. “Australian regulator investigates Google data harvesting from Android phones,” *The Guardian,* May 13, 2018, available at https://www.theguardian.com/technology/2018/may/14/australian-regulator-investigates-google-data-harvesting-from-android-phones

l. Pierce, David. “How to Keep Google From Owning Your Online Life,” *The Wall Street Journal,* May 8, 2018, available at https://www.wsj.com/articles/how-to-keep-google-from-owning-your-online-life-1525795372

li. Larson, Brett. "Google tracking phones even when they are disconnected?,” *Fox News,* Feb 11, 2018, available at http://video.foxnews.com/v/5731183327001/?#sp=show-clips

lii. Collins, Keith. “Google collects Android users’ locations even when location services are disabled,” *Quartz,* November 17, 2017, available at https://qz.com/1131515/google-collects-android-users-locations-even-when-location-services-are-disabled/

liii. Russakovskii, Artem. “Google is permanently nerfing all Home Minis because mine spied on everything I said 24/7,” *Android Police,* October 10, 2017, available at https://www.androidpolice.com/2017/10/10/google-nerfing-home-minis-mine-spied-everything-said-247/

liv. Englehardt, Steven. and Narayana, Arvind. “Online Tracking: A 1-million-site Measurement and Analysis,” *ACM CCS,* 2016, available at http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf

lv. Chen, Brian. “Why Do Android Smartphones Guzzle the Most Data?,” *The New York Times,* December 31, 2013, available at https://bits.blogs.nytimes.com/2013/12/31/why-do-android-smartphones-guzzle-the-most-data/

lvi. Øverlier, Lasse. “Data leakage from Android smartphones,” *Norwegian Defense Research Establishment,* June 6, 2012, available at https://www.ffi.no/no/Rapporter/12-00275.pdf