

# Third Party Tracking in the Mobile Ecosystem

Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert\*, Nigel Shadbolt

Department of Computer Science, University of Oxford

\*Reuters Institute for the Study of Journalism, University of Oxford

Oxford

reuben.binns|ulrik.lyngs|max.van.kleek|jun.zhao|nigel.shadbolt@cs.ox.ac.uk

timothy.libert@politics.ox.ac.uk

## ABSTRACT

Third party tracking allows companies to identify users and track their behaviour across multiple digital services. This paper presents an empirical study of the prevalence of third-party trackers on 959,000 apps from the US and UK Google Play stores. We find that most apps contain third party tracking, and the distribution of trackers is long-tailed with several highly dominant trackers accounting for a large portion of the coverage. The extent of tracking also differs between categories of apps; in particular, news apps and apps targeted at children appear to be amongst the worst in terms of the number of third party trackers associated with them. Third party tracking is also revealed to be a highly trans-national phenomenon, with many trackers operating in jurisdictions outside the EU. Based on these findings, we draw out some significant legal compliance challenges facing the tracking industry.

## CCS CONCEPTS

• **Security and privacy** → **Economics of security and privacy**; *Software reverse engineering*; • **Applied computing** → **Law**; • **Networks** → *Mobile and wireless security*;

## KEYWORDS

privacy, tracking, behavioural advertising, mobile, android, static analysis, data protection

### ACM Reference Format:

Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert, Nigel Shadbolt. 2018. Third Party Tracking in the Mobile Ecosystem. In *WebSci '18: 10th ACM Conference on Web Science, May 27–30, 2018, Amsterdam, Netherlands*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3201064.3201089>

## 1 INTRODUCTION

Billions of people use smartphones every day, generating vast amounts of data about themselves. Much of the functionality afforded by these devices comes in the form of applications which derive revenue from monetising user data and displaying behaviourally targeted advertising. Firms with the ability to collect such data have

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*WebSci '18, May 27–30, 2018, Amsterdam, Netherlands*

© 2018 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.

ACM ISBN 978-1-4503-5563-6/18/05...\$15.00

<https://doi.org/10.1145/3201064.3201089>

become a significant part of the digital economy [3], with the online advertising industry earning \$59.6 billion per year in the U.S. alone [20].

This business model is primarily enabled through ‘third-party’ trackers [27], which track users via ‘first-party’ mobile applications, whose developers embed their technology into application source code. Such networks link activity across multiple apps to a single user, and also link to their activities on other devices or mediums like the web. This enables construction of detailed profiles about individuals, which could include inferences about shopping habits, socio-economic class or likely political opinions. These profiles can then be used for a variety of purposes, from targeted advertising to credit scoring and targeted political campaign messages.

This paper aims to provide a high-level empirical overview of the extent of third party tracking on the mobile ecosystem. In particular, we aim to answer the following:

- (1) How are third party trackers distributed across apps on the Google Play Store?<sup>1</sup>
- (2) Which companies ultimately own these tracking technologies, and in which jurisdictions are they based?
- (3) Do different trackers prevail amongst different genres of apps?

Our motivation is to shed light on the status quo, in order that future efforts to address and mitigate third party tracking can be more informed and targeted.

## 2 BACKGROUND

We begin by introducing previous work on tracker detection methods, and on large-scale field studies of tracking on the web and mobile. Then, to motivate some of the present analysis, we provide an overview of existing approaches to addressing mobile tracking, including end-user controls, OS provider rules, and legal regulation. The shortcomings of the first two approaches have driven a renewed focus on the latter; by surveying the existing state of mobile tracking, we aim to provide insights into the extent to which current tracking activities may be affected by certain key data protection regulations.

### 2.1 Detecting third party tracking at scale in the wild

The third party tracking ecosystem has been studied on both the web and mobile using a variety of methods. Large scale web tracking studies detect third-party trackers by inspecting network traffic associated with a website. Some approaches use crowd-sourcing

<sup>1</sup>We did not study the Apple iOS App Store because there are no equivalently scalable iOS app collection and analysis methods

(e.g. [36, 39]) while others use automated web crawlers (e.g. [15, 23, 32, 39]). In all cases, a small number of dominant trackers are observed.

Several studies of third-party tracking have also been conducted on mobile platforms [10, 36], using both dynamic and static detection methods. Dynamic methods, as in web-based tracking studies, involve inspecting network traffic from the browser / device and identifying any third party destinations that relate to tracking. One common approach has been OS-level instrumentation, such as those of TaintDroid [14], and AppTrace [29]. An alternative to low-level OS instrumentation is to analyse all communications traffic transmitted by an app whilst it is in use [31]. Other methods involve unpacking an application’s source code (on Android systems, this comes as an Android Application Package (APK)) and detecting use of third-party tracking libraries [5, 8, 13, 24].

Other aspects of tracking have been studied, including the variety of techniques that are used, from cookies [6, 15, 16] to fingerprinting [2]. A more recent field study by Yu et al. provided a finer-grained view into tracker behaviour, by classifying data being transmitted to trackers as either ‘safe’ or ‘unsafe’ [39]. Another factor is the permissions requested by an app, which constrain the kinds of data a third party can obtain; longitudinal research has found that Android apps request additional privacy-risking permissions on average every three months [34].

The crossover between the mobile and web tracking ecosystem has also attracted attention in recent research. Various comparisons have shown that web and mobile tracking are different, both in terms of the companies that operate on each environment [36], and the specific kinds of personal information that are shared by web and mobile versions of the same service [22]. In previous work comparing 5,000 apps and 5,000 websites, it was found that while certain companies dominate both environments, the overlap between top trackers is only partial, even for web and mobile versions of the same service [9].

## 2.2 Existing approaches to addressing risks of tracking

There are three main approaches for addressing the risks of tracking: end-user privacy controls, industry self-regulation, and traditional legal regulation.

*2.2.1 End-user privacy controls.* Tracking exists on both the web and on mobile apps, but web browsers have traditionally enabled end-users to control tracking via default browser settings or through third party plugins. By contrast, no major smartphone platform OS currently gives end-users the ability to block or otherwise control third party tracking by apps (although tracker blocking is available on mobile web browsers). The privacy settings are primarily focused on app-by-app permissions, or permissions regarding certain data types (e.g. location, contacts, etc.). While various changes have been introduced like run-time permissions, and advertising identifier controls [28], these do not address the distinction between first party apps and third party trackers. More recently, awareness-raising tools have been proposed which do reveal the presence of third-parties. They make use of techniques including reverse-engineering of app source code and network traffic analysis [5, 8, 13, 14, 18, 29, 40], allowing identification of personal data

flows from apps to first and third parties. These tools have been used to map data flows and display them to end-users [7, 11, 33, 37]. Such focus on third-party data collection, rather than app-level permissions, may be a more meaningful way to enact privacy choices. However, until such controls are enabled by the OS providers, third party tracking via apps remains largely invisible to end-users. This is in contrast to the web, where millions of users make use of tracker protection tools such as uBlock Origin or Ghostery.

*2.2.2 Self-regulation by platforms.* In response to the development and proliferation of trackers, and the lack of wide-scale deployment of effective end-user tracker controls, various efforts have been made by mobile OS platform developers to address the risks. Mobile application developers are required to follow the rules of the app market providers in order for their apps to be listed [4]. Since few consumers use multiple app stores on a single smartphone, these platforms are in a stronger position to impose industry self-regulation than browser vendors, because they have the ability to effectively kick an application off the platform entirely.

Industry-led self-regulatory initiatives have thus far attempted to strike a balance between protecting users from malicious behaviour and creating a relatively permissive environment. With respect to smartphone operating systems, Apple and Google have the power to exert varying degrees of control over the behaviour of apps appearing in their default app stores. Thus far, both of their respective developer agreements permit third-party tracking, although certain user-protective practices are required, such as collecting a replaceable advertising identifier (IDFA / AAID) rather than the permanent device identifier.

More stringent action against third party tracking may also have been held back by vested interests of the OS providers. Both Google and Apple have historically had a stake in the digital advertising industry. Google own several tracker companies such as DoubleClick and others. Apple used to take a cut of advertising revenue from ad network trackers in iPhone apps, through the iADs program, but this scheme ended in 2016.

*2.2.3 Legal regulation.* These self-regulatory efforts, such as they are, sit alongside a variety of specific legal regulations with varying levels of enforcement in different countries around the world. Perhaps the most stringent and far-sighted of these is the data protection legal regime in Europe. With updated rules incoming this year in the form of the European Union’s General Data Protection Regulation, new enforcement powers including the issuing of larger fines and scope for indefinitely suspending processing may substantially curtail the activities of third party trackers.

For instance, the specific identities and purposes of third party trackers will have to be made transparent to the data subject (i.e. the user of the app); and special safeguards must be applied in the case of children. While profiling of children is not outright prohibited by the GDPR, the Article 29 Working Party (the EU body responsible for providing guidance on data protection), advise that organisations should ‘refrain from profiling them for marketing purposes’.

Regarding transfer of data across borders, while existing requirements are not fundamentally different under the GDPR, transnational data transfer is likely to receive additional scrutiny in light of

the introduction of stronger enforcement powers. Under the existing regime, personal data is permitted to flow from one jurisdiction to another, subject to compliance with certain conditions. The least onerous condition is if the recipient organisation is based in a country whose existing data protection regime has been assessed by the European Commission and deemed ‘adequate’. Otherwise, special arrangements such as standard contractual clauses and binding agreements between organisations in both jurisdictions may be necessary in order to make cross-jurisdictional data flows legitimate. Similar data flow agreements exist between other countries. In some cases these are reciprocal (such as between the EU and Andorra), while others are not (e.g., the Russian privacy regulator allows personal data to flow from Russia to EU countries<sup>2</sup>, but the reverse is not true).

Such cross-border rules and data ‘trade blocs’ have consequences for the legal basis for third party tracking when tracking companies, app developers, app stores and end-users are located in different jurisdictions. While the transfer of data from people residing in the EU to countries whose data protection regime is deemed inadequate could be legitimate in principle, more onerous conditions would need to be met. As such, any efforts to assess the legality of current practices must consider the extent to which tracking occurs across borders.

### 3 DATA COLLECTION & METHODOLOGY

#### 3.1 Play Store Indexing and App Discovery

The first step was to identify available apps. We programmatically identified popular search terms in the Play Store by autocompleting all character strings of up to a length of five, and then issued each search term to get a list of apps, ranked by popularity [17]. The identified apps were then downloaded using the `gplaycli` [25], a command line tool for interacting with the Play Store.

**3.1.1 Static analysis method.** An Android Package Kit (APK) is an Android file format that contains all resources needed by an app to run on a device. Upon download, each APK was unpacked and decoded using APKTool [35] to obtain the app’s assets, in particular its icon, bytecode (in the DEX format) and metadata (in XML format). Finally, permission requests were parsed from the XML and hosts were found in the bytecode using a simple regex<sup>3</sup>.

**3.1.2 Mapping hostnames to known tracker companies.** While this static analysis process effectively identified references to hosts in the APKs, it did not provide a means of mapping them to companies, let alone selecting only those companies who are in fact engaged in tracking. A large number of the hostnames found in the static code analysis refer to a wide range of benign external resources which are not necessarily engaged in tracking. In order to isolate only those engaged in tracking, we combined two lists of trackers derived from previous research. One list is compiled by the Web X-Ray project [23]. It maps third party web tracking domains to companies that own them, as well as parent-subsidiary

<sup>2</sup><https://www.huntonprivacyblog.com/2017/08/16/russian-privacy-regulator-adds-countries-list-nations-sufficient-privacy-protections/>

<sup>3</sup>We note that this method has the inherent problem that we cannot confirm if bytecode relating to or referencing such hosts is ever called. More sophisticated static analysis methods might better distinguish but this is left for future work. The regex used to identify hosts in the bytecode is available on [osf.io/4nu9e](https://osf.io/4nu9e)

relationships. The second list is compiled from previous research by the authors of the present paper [9, 38], which also maps domains to companies, and companies to their owners, but incorporates mobile app-centric trackers which are missing from web-oriented tracker lists. An example of domain-company ownership in the resulting aggregated list is shown in Figure 1, and parent-subsidiary relationship in Figure 2.

Host names in the tracker lists were shortened to 2-level domains using the python library `tlldextract`<sup>4</sup> (e.g. for ‘subdomain.example.com’, the domain name ‘example’ and top-level domain suffix ‘.com’ were kept and any subdomains were omitted). Tracker hosts were then matched to hosts identified in app bytecode with a regular expression which excluded matches that was followed by a dot or an alphabetic character (matching ‘google.com’ to ‘google.com/somepath’ but not ‘google.com.domain’ or ‘google.coming’).

#### 3.2 Data analysis

Most of the data analysis was conducted in R, using RStudio<sup>5</sup>.

## 4 RESULTS

#### 4.1 Numbers of tracker hosts in apps

The distribution of number of tracker hosts per app was highly right-skewed (see Figure 3). Gini inequality coefficient was 0.44. Across all analyzed apps (n = 959,426), the median number of tracker hosts included in the bytecode of an app was 10. 90.4% of apps included at least one, and 17.9% more than twenty.

#### 4.2 Numbers of distinct tracker companies behind hosts

The distribution of number of distinct tracker companies (at the lowest subsidiary level) behind the hosts in an app was similarly right-skewed (see Figure 4). The median number of companies was 5, 90.4% of apps included hosts associated with at least one company, and 17.4% with more than ten companies.

There were 13 apps for which our analysis identified 30 or more different tracking companies referred to via hosts in the bytecode. In some cases, these high numbers can be explained by the particular function of the app; for instance, some of these apps integrate multiple different services into one app (e.g. ‘Social Networks All in One’); in such cases, any tracking domains associated with those integrated services will be identified by our method. For others, mostly gaming apps, the high numbers of trackers serve no obvious function other than the usual kinds of behaviourally targeted advertising and analytics.

Rather than simply counting number of companies, we can query the proportion of apps containing hosts associated with specific companies. As illustrated in Figure 2, however, many companies have been acquired by larger parent or holding companies, such as Alphabet. The result of grouping by ‘root parent’ the percentages of apps which include hosts associated with specific companies is shown in Table 1.

<sup>4</sup><https://github.com/john-kurkowski/tldextract>

<sup>5</sup>Analysis scripts plus data are available via the Open Science Framework at [osf.io/4nu9e](https://osf.io/4nu9e). For access to the full data set, contact the authors.

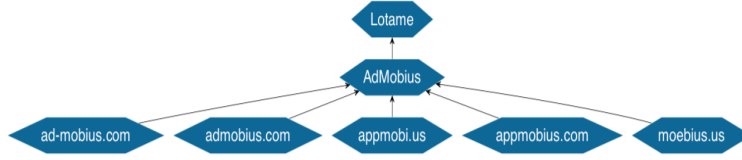


Figure 1: Example of domain-company ownership. The domain Admobi.us is owned by the company AdMobius, which is owned by the parent company Lotame.

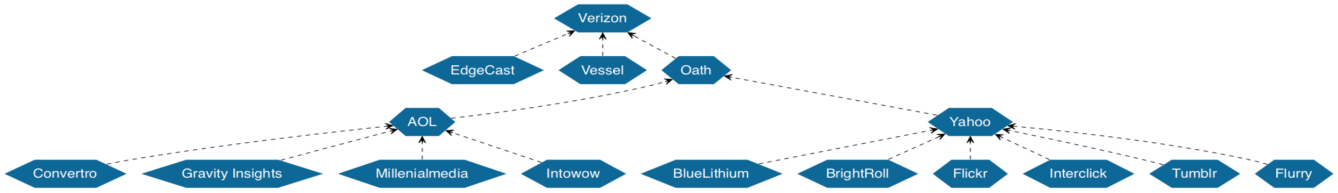
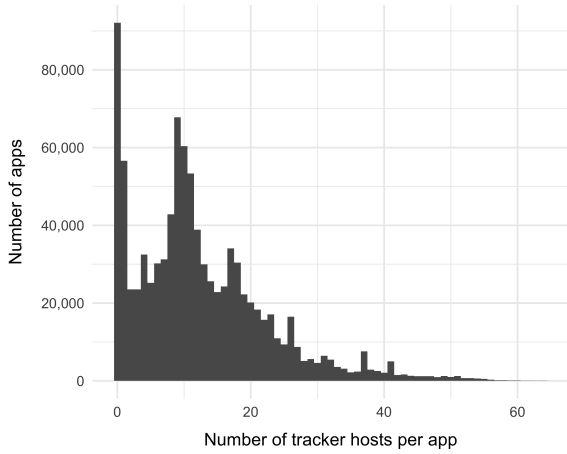
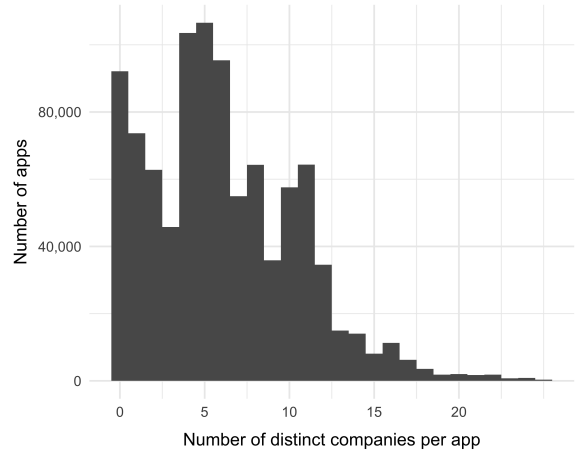


Figure 2: Example of parent-subsidiary company ownership (domains omitted). Flurry is owned by Yahoo, which is owned by Oath, which is owned by Verizon (the ‘root parent’).



Median	Q1	Q3	>20 hosts	No hosts
10	5	18	17.9%	9.6%

Figure 3: Histogram and descriptive statistics for number of tracker hosts per app (free apps on the Google Play store).



Median	Q1	Q3	>10 companies	No companies
5	3	9	17.4%	9.6%

Figure 4: Number of distinct tracker companies behind hosts in apps (free apps on the Google Play store).

### 4.3 Company prevalence by genre

The Google Play store metadata divides apps into 49 different genres (no less than 17 of these are subcategories of games, e.g. ‘Casino Games’ and ‘Adventure Games’). To provide a high-level analysis, we grouped these genres into 8 more succinct ‘super genres’

(by e.g. clustering all game genres, plus the genres ‘Comics’, ‘Entertainment’, ‘Sports’ and ‘Video Players’ into a single ‘Games & Entertainment’ category<sup>6</sup>). In addition, given concern of in particular tracking of children[1], we created a super genre consisting of

<sup>6</sup>See [osf.io/4nu9e](https://osf.io/4nu9e) for details of this grouping.

Root parent	% apps	Subsidiary	% apps	Country		
Alphabet	88.44	Google	87.57	US		
		Google APIs	67.51	US		
		DoubleClick	60.85	US		
		Google	39.42	US		
		Analytics				
		Google Tag	33.88	US		
		Manager				
		Adsense	30.12	US		
		Firebase	19.20	US		
		Admob	14.67	US		
		YouTube	9.51	US		
		Blogger	0.46	US		
		Facebook	42.55	Facebook	42.54	US
				Liverail	1.03	US
Lifestreet	<0.01			US		
Twitter	33.88	Twitter	30.94	US		
		Crashlytics	5.10	US		
		Mopub	2.51	US		
Verizon	26.27	Yahoo	20.82	US		
		Flurry	6.28	US		
		Flickr	1.37	US		
		Tumblr	1.22	US		
		Millennialmedia	0.71	US		
		Verizon	0.11	US		
		AOL	0.06	US		
		Intowow	<0.01	US		
		One By AOL	<0.01	US		
		Brightroll	<0.01	US		
		Gravity	<0.01	US		
Microsoft	22.75	Microsoft	22.11	US		
		Bing	0.12	US		
		LinkedIn	20.62	US		
		Amazon	11.57	US		
Amazon	17.91	Amazon Web				
		Services				
		Amazon	7.72	US		
		Amazon	1.73	US		
		Marketing				
		Services				
		Alexa	<0.01	US		
Unitytechnologies	5.78	Unitytechnologies	5.78	US		
Chartboost	5.45	Chartboost	5.45	US		
Applovin	3.95	Applovin	3.95	US		
Cloudflare	3.85	Cloudflare	3.85	US		
Opera	3.20	Adcolony	3.12	US		
		Admarvel	0.09	US		

**Table 1: The most prevalent root parent tracking companies and their subsidiaries (full list available on [osf.io/4nu9e](https://osf.io/4nu9e)).**

Genre	K	$\sum K$
Productivity & Tools	0.14	5.5
Games & Entertainment	0.13	5.41
Health & Lifestyle	0.1	5.5
Communication & Social	0.09	5.29
Art & Photography	0.09	5.12
Family	0.04	4.33
News	0.03	4.5
Education	0.03	5.42
Music	0.02	5.24

**Table 2: K distances between tracker rankings for each genre compared to all apps (K), and sum of pairwise distances between each genre and every other genre ( $\sum K$ ).**

apps included in one of the Google Play store’s ‘family’ categories.<sup>7</sup> For each super genre, we reran the company analysis, which revealed some important differences between the nature of tracking by genre.

First, there are differences in the number of distinct tracking companies associated with apps from different genres. Figure 5 shows the number of apps in each super genre, and descriptive statistics of number of distinct tracker companies associated with apps within each. *News* and *Family* apps have the highest median number of tracker companies associated with them, and over 20% of apps in the *News*, *Family*, and *Games & Entertainment* super genres are linked to more than ten tracker companies. Meanwhile, the lowest median number of trackers are found within *Productivity & Tools*, *Education*, *Communication & Social*, and *Health & Lifestyle* apps, and over 10% of *Productivity & Tools*, *Education* and *Communication & Social* apps have no trackers at all.

Second, there are differences in which particular trackers are associated with apps from each super genre. By comparing rankings for each, we can see the extent to which different trackers dominate each super genre. In addition to comparing the difference in rankings for any given tracker, we use an overall distance metric, the Kendall tau distance, in order to measure the extent to which rankings differ between super genres [21].

The Kendall Tau distance may be defined as:

$$K(\tau_1, \tau_2) = \sum_{\{i,j\} \in P} \bar{K}_{i,j}(\tau_1, \tau_2)$$

where:

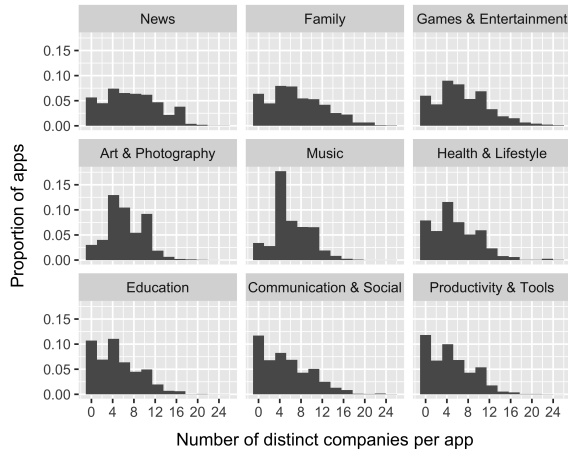
- (1) "P" is the set of unordered pairs of distinct elements in  $\tau_1$  and  $\tau_2$
- (2)  $\bar{K}_{i,j}(\tau_1, \tau_2) = 0$  if "i" and "j" are in the same order in  $\tau_1$  and  $\tau_2$
- (3)  $\bar{K}_{i,j}(\tau_1, \tau_2) = 1$  if "i" and "j" are in the opposite order in  $\tau_1$  and  $\tau_2$ .

In this context, "P" is the set of unordered pairs of trackers (e.g. ‘DoubleClick’ and ‘AdChina’), in one genre ranking  $\tau_1$  (e.g. ‘Games’)

<sup>7</sup>All apps on the Google Play store have an ordinary genre classification, but some apps are in classified into one of the Play store’s family genres.

<i>Super genre</i>	<i># apps</i>	<i>Med.</i>	<i>Q1</i>	<i>Q3</i>	<i>&gt;10</i>	<i>None</i>
News	26281	7	4	11	29.9%	6.5%
Family	8930	7	4	11	28.3%	7.2%
Games & Entertainment	291952	6	4	10	24.5%	7.3%
Art & Photography	27593	6	4	10	16.8%	3.6%
Music	65099	6	4	8	13.5%	4.1%
Health & Lifestyle	163837	5	3	8	15.4%	9.0%
Communication & Social	39637	5	2	8	16.2%	13.4%
Education	79730	5	2	8	13.3%	11.9%
Productivity & Tools	265297	5	2	8	11.9%	13.5%

(a)

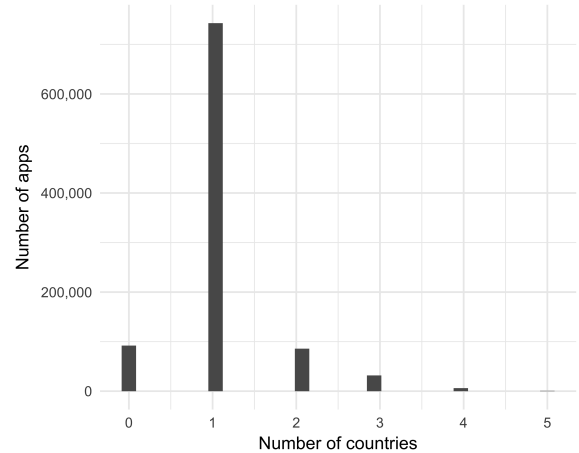


**Figure 5: Descriptive statistics (a) and histograms (b) of number of distinct tracker companies behind hosts referenced in apps, grouped by super genre.**

and another genre ranking  $\tau_2$  (e.g. ‘News’).  $K$  is based on the number of discordant pairs between  $\tau_1$  and  $\tau_2$ , where a higher  $K$  indicates greater distance.

We find that the Productivity & Tools and Games & Entertainment categories exhibit the biggest differences in ranking of trackers compared to the overall ranking of trackers across the whole Play Store, while the ranking of trackers in the Music category is the closest to the overall ranking (see Table 2).

In addition to calculating the distance between the rankings of each genre and the rankings for the entire Play Store, we also calculated the distances between each distinct pair of genres and summed them to get an idea of the overall distance of a single genre from every other genre. When considering the distance in tracker rankings from the tracker rankings of all other categories, Productivity & Tools and Health & Lifestyle appear to be the biggest outliers; the top 20 trackers in the former include companies not present in the top 20 for all apps, like Mapbox (rank #64 across all apps) as well as Chinese companies Alibaba and Baidu.



**Figure 6: Number of distinct countries in which tracker companies behind hosts in an app (free apps on the Google Play store) are based.**

<i>Country</i>	<i># apps present</i>	<i>% apps</i>
U.S.	865369	90.2%
China	48451	5.1%
Norway	30674	3.2%
Russia	24889	2.6%
Germany	24773	2.6%
Singapore	19323	2.0%
UK	14451	1.5%
Austria	4754	0.5%
South Korea	3366	0.4%
Japan	1801	0.2%

**Table 3: Apps including at least one tracker associated with a subsidiary or root parent within a given country.**

#### 4.4 Country differences

We also analysed the prevalence of countries in which the tracker companies are based (including both subsidiary and root parent level; see Table 3). Just over 90% of all apps contained at least one tracker owned by a company based in the United States. China, Norway, Russia, Germany, Singapore, and the United Kingdom were the next most common destinations. The median number of unique countries associated with the companies referred to in an app was 1 (see Figure 6).

We also calculated the country prevalence figures on a genre-by-genre basis. While the US remained the most prevalent in every case, (between 86-96%), the prevalence rankings for other countries differed by super genre. For instance, UK-based trackers were the second-most prevalent in ‘Art & Photography’, despite being only 7th overall.

## 5 DISCUSSION

We begin by discussing the limitations of our data collection methods. Next we consider some differences between tracking on websites and on mobile apps, and finally we draw out implications for the regulatory approaches outlined in section 2.2.3.

### 5.1 Limitations of data collection methods

There are several limitations to our tracker detection methods. First, it is incomplete; our knowledge base of tracker domain to company mappings is limited to those trackers which have been discovered in the course of previous research (namely [9, 23, 38]). While these lists were compiled in a systematic way, focusing on the most prevalent tracking domains, including the entire long tail of less prevalent domains might change the results reported. The inclusion and exclusion criteria for what constitutes a ‘tracker’ are also open to debate; the list compiled in prior works, and relied on here, defines a third-party tracker as ‘an entity that collects data about users from first-party websites and / or apps, in order to link such data together to build a profile about the user’, but the definition and its application are debateable.<sup>8</sup> Another issue is that without dynamic network traffic analysis of all apps, including successful man-in-the-middle proxying and ability to interpret the data payloads, we cannot confirm precisely what data is sent to each tracker. Finally, different trackers serve different purposes; some facilitate targeted advertising, while others are used for analytics. Without further fine-grained distinctions between such purposes, the figures presented here do not represent the full nuance and variety of third party tracking and its impacts.

### 5.2 Web vs. Mobile

Previous large-scale studies of tracking have largely focused on the web. The distribution model of the web allows measurement of tracking to scale in a way that the model for smartphone app distribution does not; web services are delivered in a standardised way through a browser which can easily be automated. As a result, large-scale web tracking studies typically include millions of sites. By contrast, the largest smartphone app tracking study to our knowledge at the time of writing is derived from network traffic detected by the Lumen app, which includes the data flows of 14,599 apps installed on Lumen user’s devices [30]. While such crowd-sourced methods have many advantages in terms of the granularity of the data flows and ecological validity, at best they scale to tens of thousands of apps. By contrast, our method is scalable to hundreds of thousands of apps (indeed, our dataset of apps is close to a million).

### 5.3 Implications for tracker regulation

While the distribution of trackers across apps is of general interest from a privacy and data protection regulation perspective, we focus here on several particular regulatory implications arising from our findings.

*5.3.1 Cross-jurisdictional data flow.* As explained in Section 2.2.3, the rules regarding transfers of data outside the EU under the

GDPR are similar to the previous regime (under the Data Protection Directive), but with some new details as well as larger associated fines. In so far as these developments result in more investigation and enforcement by authorities, the impact will be different for companies depending on their jurisdiction. There will be no impact on those based in the EU, such as Germany (the fifth-most prevalent country in which trackers are based), who benefit from rules permitting the free flow of data within the Union. Some third countries such as Canada also benefit from being on the EU Commission’s list of legal regimes that are deemed ‘adequate’ and therefore data transfers to trackers in those jurisdictions are legitimate without further measures in place.

However, amongst the top-10 most prevalent countries there are several which lie outside the E.U. and are not deemed adequate, such as China, Russia, Singapore, South Korea and Japan. In order for transfers to these countries to be legitimate, additional safeguards must be in place as explained in Section 2.2.3. We cannot determine whether such arrangements have been put in place by the identified companies based in non-approved jurisdictions, but our figures give an indication of the volume of companies to whom these more onerous rules apply. While the percentages of apps which include trackers from such jurisdictions are small compared to the US—China (5.1%), Russia (2.6%), Singapore (2%) versus US (90%)—they are still significant, numbering in the tens of thousands.

*5.3.2 Profiling.* The GDPR uses the term ‘profiling’ to describe any fully or partly automated processing of personal data with the objective of evaluating personal aspects of a natural person (Article 4(4)). Many of the tracking companies included in our knowledge base engage in data processing activity that would likely constitute ‘profiling’ under this definition. For instance, the purpose of many of the most common trackers is behaviourally targeted advertising, whereby individuals are evaluated along demographic and behavioural dimensions to determine their propensity to respond to certain marketing messages. Profiling is prohibited if it has ‘legal or significant’ effects on the data subject. While the definition of ‘significant effects’ is not entirely clear, the Article 29 Working Party has advised that even profiling for marketing purposes could potentially give rise to significant effects, including if it is: intrusive; targets vulnerable, minority groups, or those in financial difficulty; involves differential pricing; or deprives certain groups of opportunities.<sup>9</sup> Trackers which enable such activities without consent of the data subject could therefore be in breach of Article 22 (unless such profiling is necessary for entering or performing a contract, or it is authorised by another member state law). Many of the most prevalent trackers observed in our study have the capacity to be used in such ways, and evidence of such practices is beginning to emerge. For instance, DoubleClick (present on 60% of apps analysed) has been shown to target adverts for higher-paid jobs to men at a higher rate than to women [12]; while web-based price discrimination has also been documented by numerous studies in recent years [19, 26].

*5.3.3 Rights and obligations regarding children.* Like the old Directive, the GDPR defines certain additional rights and obligations

<sup>8</sup>The principles behind the criteria used here are discussed in the aforementioned prior works

<sup>9</sup>Article 29 Working Party: Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 [http://ec.europa.eu/newsroom/article29/item-detail.cfm?item\\_id=612053](http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053)

regarding processing the personal data of children (defined as anyone under the age of 16, and for certain additional protections, 13). If a tracker is relying on consent as a legitimating ground for processing, then such consent would not be valid from a child under 13; instead a parent or guardian would need to consent. Furthermore, as discussed above, Recital 38 states that special protections should be in place if children's data are being processed for marketing and user profiling. This description would likely cover many of the trackers which are embedded in apps from the Family and Games & Entertainment genre categories, which are clearly targeted at children. Problematically, apps from these two genres are especially exposed to third party tracking, with the average app including hosts associated with 7 distinct tracker companies for Family apps, and 6 for Games & Entertainment apps (only News apps are more exposed). Given the relatively higher level of protection set in the law regarding profiling children for marketing, it seems that tracking is most rampant in the very context in which regulators are most concerned to constrain it.

## 6 CONCLUSION

We believe that by undertaking analysis of the distribution of tracking technology on close to 1 million smartphone apps, we gain insight into the breadth and scale of this highly important phenomenon. Unlike previous studies whose coverage of apps numbers in the tens of thousands, and may be skewed towards the app choices of the users from whom data is gathered, our study is a systematic analysis of apps on the Play Store.

Our genre-by-genre analysis suggests that there are differences in the behaviour and distribution of trackers depending on the functionality or purpose the app provides. News and Games apps appear amongst the worst in terms of the number of tracker companies associated with them. Tracking is also a substantially trans-national phenomenon; around 100,000 apps we analysed send data to trackers located in more than one jurisdiction.

These findings suggests that there are challenges ahead both for regulators aiming to enforce the law, and for companies who intend to comply with it. Full audits of mobile app stores such as this could help regulators identify areas to focus on. Previous privacy enforcement 'sweeps'<sup>10</sup> have focused on the most popular apps, and their terms of service and privacy policies. But the analysis here suggests that apps may not necessarily be the most efficient point of analysis; rather, identifying and investigating the most prevalent trackers might be a better target. Some of the practices likely to be involved - such as allowing profiling of children without attempting to obtain parental consent - may be downright unlawful. It remains to be seen how and if regulators will attempt to detect and prevent behavioural targeting that has 'significant effects' on data subjects.

The governance of these activities is complex, involving many stakeholders, including: users, smartphone operating system developers, equipment manufacturers, alternative app market operators, app developers, and tracking companies (who also operate multi-sided markets with advertisers and therefore have the ability to impose constraints on what ads can be served). Effective regulation

will require collaboration between regulators and these myriad other actors.

## ACKNOWLEDGMENTS

All authors are supported under *SOCIAM: The Theory and Practice of Social Machines*, funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/2 and comprises the University of Oxford, the University of Southampton, and the University of Edinburgh. Reuben Binns and Max Van Kleek are also supported by *ReTiPS: Respectful Things in Private Spaces*, a project funded through the PETRAS IoT Hub Strategic Fund, which, in turn, was funded by the EPSRC under grant number N02334X/1. Timothy Libert is also supported by the Google Digital News Project at the Reuters Institute for the Study of Journalism. Jun Zhao is also supported by KOALA (<http://SOCIAM.org/project/koala>): Kids Online Anonymity & Lifelong Autonomy, funded by EPSRC Impact Acceleration Account Award, under the grant number of EP/R511742/1.

## REFERENCES

- [1] 2010. EU kids online. *Zeitschrift für Psychologie - Journal of Psychology* 217, 4 (2010), 236–239. <https://doi.org/10.1027/0044-3409.217.4.233>
- [2] Gunes Acar, Marc Juarez, Nick Nikiforakis, Claudia Diaz, Seda Gürses, Frank Piessens, and Bart Preneel. 2013. FPDetective: dusting the web for fingerprinters. In *Proc. of ACM SIGSAC conference on Computer & communications security*. ACM, 1129–1140.
- [3] Alessandro Acquisti, Curtis R Taylor, and Liad Wagman. 2016. The economics of privacy. *Journal of Economic Literature* 52, 2 (2016).
- [4] Jonathan Anderson, Joseph Bonneau, and Frank Stajano. 2010. Inglorious Installers: Security in the Application Marketplace. In *WEIS*. Citeseer.
- [5] Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Ochteau, and Patrick McDaniel. 2014. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. *ACM SIGPLAN Notices* 49, 6 (2014), 259–269.
- [6] Arslan Aziz and Rahul Telang. 2015. *What is a Cookie Worth?* Technical Report. Technical Report.
- [7] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. 2013. Little brothers watching you: Raising awareness of data leaks on smartphones. In *Proceedings of the Symposium on Usable Privacy and Security*. ACM, 12.
- [8] Leonid Batyuk, Markus Herpich, Seyit Ahmet Camtepe, Karsten Raddatz, Aubrey-Derrick Schmidt, and Sahin Albayrak. 2011. Using static analysis for automatic assessment and mitigation of unwanted and malicious activities within Android applications. In *Malicious and Unwanted Software (MALWARE), 2011 6th International Conference on*. IEEE, 66–72.
- [9] Reuben Binns, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2018. Measuring third party tracker power across web and mobile. *arXiv preprint arXiv:1802.02507* (2018).
- [10] Theodore Book and Dan S Wallach. 2015. An empirical study of mobile ad targeting. *arXiv preprint arXiv:1502.06577* (2015).
- [11] Saksham Chitkara, Nishad Gothoskar, Suhas Harish, Jason I Hong, and Yuvraj Agarwal. 2017. Does this App Really Need My Location?: Context-Aware Privacy Management for Smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 42.
- [12] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.
- [13] Manuel Egele, Christopher Kruegel, Engin Kirda, and Giovanni Vigna. 2011. PiOS: Detecting Privacy Leaks in iOS Applications. In *NDSS*. 177–183.
- [14] William Enck, Peter Gilbert, Seungyeop Han, Vasant Tendulkar, Byung-Gon Chun, Landon P Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N Sheth. 2014. TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems (TOCS)* 32, 2 (2014), 5.
- [15] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of ACM Computer and Communications Security 2016*.
- [16] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W Felten. 2015. Cookies that give you away: The surveillance implications of web tracking. In *Proc. of the 24th International Conference on World Wide Web*. ACM, 289–299.

<sup>10</sup>See <https://www.privacyenforcement.net/node/906>



- [17] Google. 2017. Search using Autocomplete. (2017). <https://support.google.com/websearch/answer/106230?co=GENIE.Platform%3DAndroid&hl=en-GB>
- [18] Michael I Gordon, Deokhwan Kim, Jeff H Perkins, Limei Gilham, Nguyen Nguyen, and Martin C Rinard. 2015. Information Flow Analysis of Android Applications in DroidSafe. In *NDSS*.
- [19] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 conference on internet measurement conference*. ACM, 305–318.
- [20] IAB. 2016. IAB Internet Advertising Revenue Report 2015. (2016).
- [21] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1/2 (1938), 81–93.
- [22] Christophe Leung, Jingjing Ren, David Choffnes, and Christo Wilson. 2016. Should You Use the App for That? Comparing the Privacy Implications of App- and Web-based Online Services. In *Proc. of the 16th ACM Internet Measurement Conference*. To appear.
- [23] Timothy Libert. 2015. Exposing the Invisible Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites. *International Journal of Communication* 9 (2015), 18.
- [24] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I. Hong. 2014. Modeling Users' Mobile App Privacy Preferences: Restoring Usability in a Sea of Permission Settings. In *Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, Menlo Park, CA, 199–212. <https://www.usenix.org/conference/soups2014/proceedings/presentation/lin>
- [25] Matlink. 2017. Google Play Downloader via Command Line. Website. (2017). <https://github.com/matlink/gplaycli>
- [26] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2012. Detecting price and search discrimination on the internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks*. acm, 79–84.
- [27] Rodrigo Montes, Wilfried Sand-Zantman, and Tommaso M Valletti. 2015. The value of personal information in markets with endogenous privacy. (2015).
- [28] Mohammad Nauman, Sohail Khan, and Xinwen Zhang. 2010. Apex: extending android permission model and enforcement with user-defined runtime constraints. In *Proceedings of the 5th ACM symposium on information, computer and communications security*. ACM, 328–332.
- [29] Lingzhi Qiu, Zixiong Zhang, Ziyi Shen, and Guozi Sun. 2015. AppTrace: Dynamic trace on Android devices. In *2015 IEEE International Conference on Communications*. IEEE, 7145–7150.
- [30] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. 2018. Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem. (2018).
- [31] Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, and David Choffnes. 2016. Demo: ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic. In *Proceedings of the International Conference on Mobile Systems, Applications, and Services Companion (MobiSys '16 Companion)*, 117–117.
- [32] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and defending against third-party tracking on the web. In *Proc. of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 12–12.
- [33] Gaurav Srivastava, Saksham Chitkara, Kevin Ku, Swarup Kumar Sahoo, Matt Fredrikson, Jason Hong, and Yuvraj Agarwal. 2017. PrivacyProxy: Leveraging Crowdsourcing and In Situ Traffic Analysis to Detect and Mitigate Information Leakage. *arXiv preprint arXiv:1708.06384* (2017).
- [34] V. F. Taylor and I. Martinovic. 2017. To Update or Not to Update: Insights From a Two-Year Study of Android App Evolution. In *ACM Asia Conference on Computer and Communications Security (ASIACCS'17)*. <https://doi.org/10>
- [35] Connor Tumbleson. 2017. Apktool - A tool for reverse engineering 3rd party closed binary Android apps. (2017). <https://ibotpeaches.github.io/Apktool/>
- [36] Narseo Vallina-Rodriguez, Srikanth Sundaresan, Abbas Razaghpanah, Rishab Nithyanand, Mark Allman, Christian Kreibich, and Phillipa Gill. 2016. Tracking the Trackers: Towards Understanding the Mobile Advertising and Tracking Ecosystem. *arXiv preprint arXiv:1609.07190* (2016).
- [37] Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J Weitzner, and Nigel Shadbolt. 2017. Better the devil you know: Exposing the data sharing practices of smartphone apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5208–5220.
- [38] Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J. Weitzner, and Nigel Shadbolt. 2017. Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5208–5220. <https://doi.org/10.1145/3025453.3025556>
- [39] Zhonghao Yu, Sam Macbeth, Konark Modi, and Josep M Pujol. 2016. Tracking the Trackers. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 121–132.
- [40] Jinyan Zang, Krysta Dummit, James Graves, Paul Lisker, and Latanya Sweeney. 2015. Who knows what about me? A survey of behind the scenes personal data sharing to third parties by mobile apps. *Proceeding of Technology Science* (2015).