Understanding Worldwide Private Information Collection on Android

Yun Shen
NortonLifeLock Research Group
yun.shen@nortonlifelock.com

Pierre-Antoine Vervier
NortonLifeLock Research Group
pierre-antoine.vervier@nortonlifelock.com

Gianluca Stringhini
Boston University
gian@bu.edu

***Abstract*—Mobile phones enable the collection of a wealth of private information, from unique identifiers (e.g., email addresses), to a user's location, to their text messages. This information can be harvested by apps and sent to third parties, which can use it for a variety of purposes. In this paper we perform the largest study of private information collection (PIC) on Android to date. Leveraging an anonymized dataset collected from the customers of a popular mobile security product, we analyze the flows of sensitive information generated by 2.1M unique apps installed by 17.3M users over a period of 21 months between 2018 and 2019. We find that 87.2% of all devices send private information to at least five different domains, and that actors active in different regions (e.g., Asia compared to Europe) are interested in collecting different types of information. The United States (62% of the total) and China (7% of total flows) are the countries that collect most private information. Our findings raise issues regarding data regulation, and would encourage policymakers to further regulate how private information is used by and shared among the companies and how accountability can be truly guaranteed.**

I. Introduction

Data has become the commodity that sustains much of the Web. In recent years, the research community has raised awareness on the threats linked to sensitive user data collection by third parties. For example, specialized companies collect information from Web users to uniquely identify them across websites, potentially to provide them with more tailored advertisements [4], [32], [35], [46], [67]. In some cases, rogue browser extensions collect information that is supposed to remain private, such as a user's browsing history [15], [82]. As mobile devices become more central in the computing experience of users, the threats linked to private information collection increase. Mobile devices can provide a wealth of sensitive information [36], [53] that goes beyond identifiers to uniquely fingerprint users [58], including location information [22], [44], [78], call logs, text messages, and even information on which applications are installed a device [91]. This information can be used by third parties to deliver targeted advertisements [50] as well as for nefarious reasons, from stalking a victim by monitoring her location [14] to defeating two factor authentication by stealing text messages [34].

Network and Distributed Systems Security (NDSS) Symposium 2021
21-25 February 2021, Virtual
ISBN 1-891562-66-5
https://dx.doi.org/10.14722/ndss.2021.24076
www.ndss-symposium.org

There exist insightful research efforts [12], [16], [17], [27], [32], [33], [41], [58], [61], [65], [68], [77] to understand the impact and the threats posed by information collection on mobile devices. It remains however very challenging to obtain a comprehensive view of the information collected by mobile apps, given the wealth of potential information collected, the software diversity of mobile platforms, and the geographic diversity of mobile users and of the actors that they interact with. To shed light on the problem, previous research resorted to running apps in a sandbox environment [16] or analyzing network traffic [12], [61], [65] to monitor the information that they leaked. While this approach can be useful to identify trackers, it has two limitations: first, by running apps from a single vantage point it is challenging to replicate the geographic diversity of real users; second, apps could detect sandboxed environments and act in a different way than they would on real devices (for example by not leaking any sensitive information), and this could bias the results [48], [62], [79]. Alternatively, previous work collected network data from an ISP, looking for information leaks [32], [33], [77]. While this approach solves the sandbox detection problem, it still has a geographic bias, since different users around the world might be using different apps and might be subject to different types of sensitive data collection. As a third approach, researchers recruited participants to install an app on their mobile phones; the app would then monitor the device for information leaks [58]. This approach solves the issues mentioned above and offers insightful findings, but it remains a challenging task to attract a population of users that is large and diverse enough to represent worldwide trends at scale. Additionally, previous research [58] either mainly looked at information collections that can be used to identify a device (e.g., IMEI numbers or SIM card information) or considered limited types of sensitive information that can be collected by third parties [36], [52], [61], such as birthday, username/passwords, contacts, media files, etc.

In this paper, we provide the most comprehensive view of private data collection by Android apps to date. To achieve this, we tap into the analysis infrastructure of a popular mobile security product. The company behind this product runs Android apps in its backend infrastructure and identifies dangerous information flows by performing static and dynamic analysis. It then builds signatures of method calls that are indicative of privacy invasive activity and pushes them to the mobile devices that installed the security product, which use them to identify privacy invasive and malicious apps that have been installed. This infrastructure allows us to monitor the information collected by apps for a population of 17.3M

TABLE I: Summary of datasets used.

Dataset	Data	Count
Mobile app activity log (01/2018 - 09/2019)	Total records	6B
	Days	634
	Countries and regions	201
	Devices	17.3M
	Distinct app names	2.13M
	Distinct app SHA2s	6.5M
	Distinct PIC FQDNs	76,451
	Distinct PIC domains	40,851
Mobile app reputation log	Low reputation SHA2s	3.4M
VT	Total reports	6.5M
	PHA SHA2s (detections $\geq$ 6)	3.5M
	Benign SHA2s (no detection)	2.3M
	Not found SHA2s	401K
Domain to owner org. (01/2018 - 09/2019)	Domains	10,736
	Organizations	9,593
Blacklists (01/2018 - 09/2019)	Domains/IPs	7,670
Geolocation (01/2018 - 09/2019)	Domains/IPs	40,851

devices daily for 21 months between 2018 and 2019. This is three orders of magnitude more devices than what previous work analyzed [58]. Compared to previous work, we go beyond tracking, contact, and credential information, and trace 22 categories of private information, 13 of which were not considered by previous work [12], [52], [58], [61] (see Section II). This allows us to paint an unprecedented picture of the state of sensitive information collection on Android in the wild, identifying the big players in this space (both legitimate companies and malicious actors), together with geographic trends.

Among others, this paper makes the following findings:

- Private information collection is widespread on Android, with 87.2% of all devices in our dataset sending information to at least five distinct domains. While most PIC domains collect identifiers to track a user or a device (e.g., device information or email addresses), an alarmingly high number of domains collect other types of private information such as a device's location or a user's contacts.
- Looking at the destinations where private information is sent to, we find that most information flows terminate in the United States. We do however find that China, trailing the US at the second place, collects 7% of all data flows. This is three times higher than what was reported in previous work [58]. We also find that there was no significant difference in the number of information flows leaving the European Union after the implementation of GDPR. These findings highlight the challenges involved in implementing data protection regulations.
- We find that potentially harmful applications (PHAs) [28] are more aggressive in collecting private information than benign apps, especially when it comes to information related to the apps installed and running on a device. We also find that a small number of devices (4k) had apps installed that steal the user's text messages, potentially enabling the circumvention of two factor authentication.

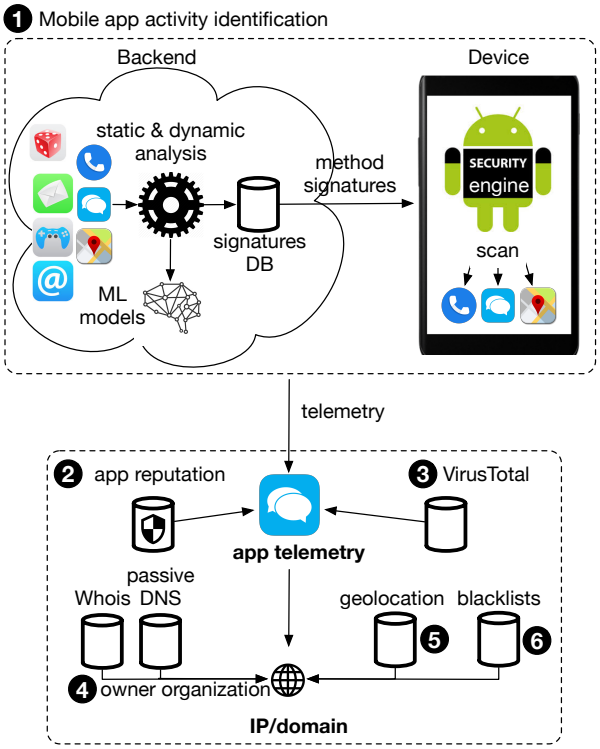


Fig. 1: Workflow of our measurement study.

Our findings highlight a number of challenges faced by the research community when studying private information collection on Android. We show that looking at device penetration is critical to observe the distribution of information collection actors in the wild, and looking at application penetration only can provide a biased view. We also highlight how looking at users located in different regions is important to get a comprehensive view, since actors operating in different countries are interested different types of information.

II. Datasets

This section details the approach that we follow for data collection (summarized in Figure 1) and summarizes the datasets used in this study (see Table I).

Workflow. The overall workflow of our measurement study is as follows (see Figure 1). We use *mobile app activity data (❶)* to identify the private information collection activities from 2.13M apps (6.5M SHA2s) installed on 17.3M devices across 200 countries and regions. We then augment this data by using *Mobile app reputation data (❷)* and *VirusTotal (VT) reports (❸)* to identify the potentially harmful apps (PHAs). Finally we use *domain and IP Whois and passive DNS (❹)* to extract domain ownership information (e.g., parent company, business category, etc), *IP and domain geolocation (❺)* to identify the country where apps send data, and *IP and domain blacklists (❻)* to identify domains associated with malicious activity. We provide details of each step in the rest of this section.

Telemetry data collection. In this paper, we use mobile telemetry data collected by the security company's mobile security product, which has been installed on millions of

TABLE II: The 22 types of private information monitored by the security app.

Group	Category	Description	Previously studied or novel
Tracking	Phone number	Phone number	[36], [58]
	Device info	IMEI, OS/kernel version, phone producer, phone model	[36], [58], [61]
	SIM card info	Information about SIM serial number, IMSI, voicemail number	[58]
	Location info	GPS or cell tower coordinates	[36], [61]
	Operator info	Information about the network operator	✓
	Setting info	Information about the device configurations	✓
Activity and social profiling	Account info	Details about the configured accounts can be exported (including user names of entries under Settings/Accounts)	[36], [61] (partially)
	Email info	Details about the email address such as Gmail address can be exported	[36]
	Contact info	Contact list can be exported	[61]
	Social network account	Details about the social network accounts such as Facebook account can be exported	✓
	Voice mail account	Details about the voice mail accounts can be exported	✓
	Call log	Call log can be exported	✓
	SMS info	App can send the content or sender/recipient details from SMS/MMS messages	✓
	Calendar info	Calendar can be exported	✓
Usage preference	Installed app info	Details about apps installed on the phone are/can be exported (full or partial list of installed package names, or app titles)	✓
	Running app info	Details about apps running at a certain time are/can be exported	✓
	Browser history info	Browser history can be exported	✓
	Browser bookmark info	Browser bookmarks can be exported	✓
Audio/Video	Audio info	Recorded audio clips can be exported (e.g., recorded by the app, or picked from saved)	[52]
	Photo info	Photo can be exported	✓
	Video info	Video can be exported	[52]
	Camera info	App can take pictures or picks them from gallery and exports them	✓

mobile devices. This company has a dedicated infrastructure to collect apks (one app may have multiple apks) from popular Android markets and various intelligence sources. These apks are then analyzed by a sophisticated infrastructure with both static and dynamic analysis pipelines. For instance, the static analysis pipeline can identify if an apk directly invokes any suspicious and sensitive API (including reflection [1], dynamic code loading [55], native code [38], etc.), requests permissions not related to its advertised description [56], as well as perform fine-grained permission analysis [64], [23], flow and context sensitive taint analysis [7], etc. Third-party libraries/SDKs used by the apps are also analyzed using the same procedure stated above. Following the static analysis, the backend can build an initial report on control-flow, data-flow, and permissions related to an apk. In addition to static analysis, the security company also performs dynamic analysis by running an apk in a sandbox environment with various Android OS versions. Through network and system instrumentation, the dynamic analysis pipeline runs an apk in different conditions (e.g., UI-automation [29], input generation [83], apk fuzzing [86], etc.) with varied execution time to capture its activities under different contexts. For example, the dynamic analysis pipeline reports if advertisements appear outside of an app in unexpected places (e.g., notification bar, shortcut, etc.) or exhibit unusual behaviors (e.g., change the user's home page). State-of-the-art commercial products are also employed by the security company to deal with challenges such as emulator/motion evasion, obfuscated code/libraries, etc. to assist the aforementioned analysis pipelines. At the same time, several machine learning models are built using the features generated from the pipelines to enable the backend to detect sophisticated PHAs. This way, the mobile security product can fingerprint activities with high accuracy and minimize false positives which may lead to an undesirable high customer churn rate. Note that the infrastructure continuously inspects apks. An apk that has been analyzed before may also be subject to regular reinspection. By combining the results of the static and dynamic analysis, the security company can rigorously fingerprint traces of app activity, including the types of private information that the app is collecting. These traces are later used to develop signatures for the apps collecting private information, in the form of sequences of method signatures.

These signatures are then deployed in the security product on the mobile devices to identify installed apps that have been linked to private information collection. If the user permits telemetry data collection, meta-information related to the app detections are sent to the telemetry data collection infrastructure and used to improve the app security features and its privacy leakage detection capability. The collected data is safeguarded by the global privacy policy of this security company. Devices are identified by a unique anonymized identifier, but it is not possible to link such an identifier back to the device. The mobile security app only collects detection metadata, and it cannot inspect network traffic data, hence the company does not collect any actual communication/user data, or other types of PII. We provide a detailed discussion about ethics and data privacy at the end of this section.

Mobile app activity data ❶. Following the aforementioned data collection stage, we extract the following meta information associated with detected app activities from the telemetry data: anonymized device identifier, device country code, timestamp, app SHA2, app package name, category of private information accessed, and domain to which such information was sent. Note that we only collect activities detected with high confidence by the security engine using the aforementioned signatures. Table II lists the 22 categories of private information monitored by the security app and used in this study. We organize these types into *four* functional categories: tracking, activity & social profiling, usage preference, and audio/video/photo data. To perform a comprehensive study, we collected 634 days (i.e., 21 months) of data between January 5, 2018 and September 30, 2019. On average, we collect 10M raw events from 17.3M devices daily. In total, our dataset covers 2.13M unique package names with 6.5M unique app hashes across 200 countries and regions. Note that the distribution of

devices used this study is not heavily screwed towards any specific region.

Mobile app reputation data ❷. We also use the mobile app reputation data from this security company to identify potentially harmful applications (PHAs) [28], which have been identified as malware or unwanted applications in general by the company's analysis infrastructure. This data contains meta information associated with PHAs that are detected based on pre-defined signatures or whose behaviors violate the pre-defined rules. We collect meta information of apps with a negative reputation score, which indicates PHAs. In total, we collect 3.4M PHA SHA2s.

VirusTotal ❸. We augment our PHA dataset by querying all 6.5M SHA2s collected from mobile app activity data on VirusTotal, to minimize false positives and false negatives potentially incurred by the mobile malware detection data from the security company. We consider an app as PHA if VirusTotal returns a minimum of six detections and a file as benign if none of the AV companies flag it. Combining this with the reputation data described in step ❷, we identify 3.5M PHA SHA2s. Besides, we also retain additional information associated with these apps including signer info, detections, etc from the VirusTotal reports.

Domain/IP to owner organization ❹. As part of our analysis, we aim to map the domain names that collect personal information to the organizations that own them. This mapping enables us to group multiple, sometimes seemingly unrelated domain names under the umbrella of their parent organization. For instance, `google.com` and `youtube.com` would be both mapped to `Alphabet`. This task has been discussed in the previous literature such as [42]. To better perform this mapping we combine information related to domain names from different data sources. Each data source typically enables us to link multiple domains to a single organization or identify a canonical name for the organization.

First, we take advantage of domain Whois to identify domains that have been registered by an organization or its subsidiaries. We then extract the IP footprint of organizations from the Internet Routing Registries (IRRs) by identifying all public IP addresses owned by these organizations. Further, we use Rapid7's passive DNS [57] to identify the IP addresses to which domains resolve as well as uncover additional domains controlled by the organizations. To limit the impact of IP address churn and the dynamics of IP address and domain registration we continuously update the domain to organization mapping throughout the 21 months so that it always reflects the most up-to-date view of the data controllers involved in private information collection.

We then build a relationship graph by taking advantage of previously extracted connections between `domains`, `IP addresses` and `organization names`. We later perform a graph-based label propagation to map domains to their most likely owning organizations. To avoid wrong mappings, we filter out relationships involving domain registrars, Whois privacy protection services, ISPs and cloud providers as these services may hide the real owner of a domain and therefore pollute our results.

Overall our dataset contains 76,471 PIC FQDNs, which correspond to 40,851 (second-level) domains. We can map 10,736 domains to a total of 9,593 organizations. To assess the accuracy of these results we inspect the mapping results to remove corner cases and found that 107 domains (0.13% of the total) were mapped to wrong organizations. Such false positives were mainly due to domains using the same domain registrars and technical support email addresses,e.g., some public relations (PR) companies possibly register and maintain domains for their different customers. Given the small number of false positives, we consider these results accurate enough to conduct further analysis in this paper.

Domain and IP geolocation ❺. One critical aspect of measuring private information collection in mobile apps is the ability to locate, i.e., identifying the country where this information is sent to. To achieve this goal, we systematically extract the geolocation of all domains and IP addresses with which apps communicate. Note that these domains and IP addresses are identified on the backend when analyzing the apps. They do not come from actual communication/user data. To obtain the most accurate and fine-grained location of the PIC parties we first attempt to geolocate each domain from its country-related top-level domain, e.g., `.be` for `Belgium`. If not possible, we revert to geolocating the domain based on its IP-level hosting infrastructure. To this end, we first try to locate the individual IP address with Maxmind [45] and, if unsuccessful, consider the coarser-grained location of its encompassing network using the IRR (IP whois). The impetus behind our domain geolocation approach is that the country code of a top-level domain should provide more accurate geolocation than its resolving IP address location, as observed by recent work [63]. Note that to limit the impact of CDNs (e.g., IP anycast) on the geolocation of the domains they serve, we identify these domains by relying on CNAME DNS records for the PIC domains and a list of CDN domains [1]. Cases of generic ccTLDs such as `.io`, `.me`, `.tv` or `.co` are handled with care and only represent 2.7% of all domains in our dataset. We have also noticed that some domains (about 6%) appear to have been incorrectly extracted by the dynamic analysis infrastructure leading to invalid domains that are therefore discarded.

Domain and IP blacklists ❻. To determine whether some apps in our dataset communicate with domains or IP addresses that have been previously associated with malicious activities we query large domain and IP blacklists. These feeds cover a wide range of activity including sending spam emails, hosting malware, phishing, and fraud websites. They are pulled from different sources, such as Spamhaus' SBL and DBL [66], Team Cymru's Botnet Analysis and Reporting Service [73], Abuse.ch's malware sources blacklist [3], and IPS and AV alerts collected by the security company. We take daily snapshots of these feeds and can assess the reputation of each domain and IP address throughout the period covered by our main dataset (Jan 2018-Sep 2019).

Ethics and data privacy. At the very first screen when using the mobile security product for the first time, users are shown a dialog about the purpose of the telemetry collection in the license agreement, and how the global privacy policy of the security company safeguards the data. The license agreement specifies that the telemetry is "processed for the purposes

[1] https://github.com/appurify/OpenSource-Software-Bundle/blob/master/webpagetest-wpt/agent/wpthook/cdn.h

of delivering the product by alerting the User to potentially malicious applications as well as improving the app security feature" and is "kept in an encrypted pseudonymized form." Detecting private data leakage from apps is one of the detection capabilities offered by the security product. The stated purpose for data collection therefore meets the use of the data in this paper, since the measurements performed in this paper are being used by the security company to refine and improve data protection on the customer devices. The telemetry used in this paper does not contain any PII. The anonymized device identifier that characterizes each device is only used to compute device-based prevalence rates in this study and discarded after processing.

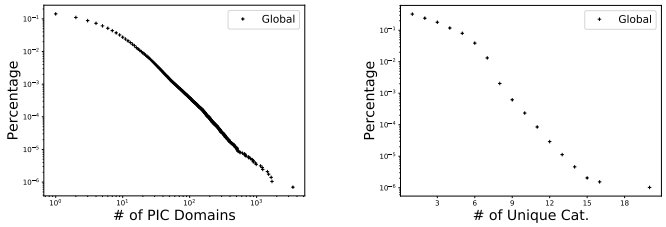
Roadmap. The roadmap of our measurement study is laid out as follows. In Section III, we discuss the landscape of private information collection (PIC) in Android ecosystems. More specifically, we start with the measurement of the pervasiveness of PIC apps installed on our user base, then focus on the app presence rate and the device penetration rate to uncover the most pervasive PIC domains globally as well as important regional players. Once we identify the global/regional players, we then study the destination of those private information flows, aiming to understand the countries where these flows terminate in Section IV, and understanding the characteristics of the data controllers who ultimately obtain and process the private information in Section V. Finally, we investigate whether the type of information collected by PHAs is different from the one collected by regular apps in Section VI.

III. Landscape of Private Information Collection (PIC) in Mobile Ecosystems

In this section, we study the landscape of private information collection (PIC) in mobile apps. First, we look at the pervasiveness of PIC apps installed on the user base of the security vendor. We then focus on the app presence rate (i.e., the number of PIC domains in apps) to identify the global/regional top players. We later focus on the device penetration rate (i.e., the number of devices that a PIC domain collects information from) to uncover the most pervasive PIC domains globally as well as important regional players, understand what types of private information are collected by these PIC domains, and if we can observe behavioral differences in different regions regarding private information collection.

A. Pervasiveness of PIC in Mobile Apps

In this section, we demonstrate the pervasiveness of private information collection in mobile apps. The results are shown in Figure 2. Apps send private information collected to 2 unique PIC domains on average. As we can observe in Figure 2a, over 175K apps (approximately 8.2% of total apps) send collected data to at least 5 unique PIC domains. These apps are installed on 15.1M devices in our dataset (87.2% of all devices). At the same time, as we can observe in Figure 2b, over 156K apps collect at least 5 unique categories of private information (see Table II). This covers 13M devices (74.9% of all devices). The overlapping 57.6k apps between the aforementioned two categories of apps cover 12.8M devices (73.8% of all devices). In other words, 73.8% of all devices in our dataset have at least one app collecting at least 5 unique categories of private information and sending them to at least 5 unique PIC domains. Our findings show that private information collection in mobile apps is universal and diversified at the same time.



(a) (b)

Fig. 2: **(log-scale)** Complimentary cumulative distribution (CCDF) of mobile apps in terms of unique PIC domains (a) and unique categories of private information collected (b).

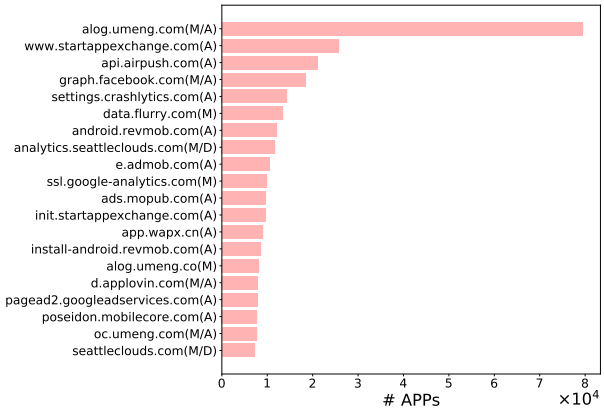


Fig. 3: Global top 20 PIC domains ranked by app presence. Domain's primary function- **M**: Metrics/Analytics, **A**: Advertising, and **D**: Development.

B. PIC Domains: App Presence Study

PIC organizations generally benefit from collecting data about more users. To reach this goal, one of the strategies adopted by these organizations is increasing their presence in mobile apps to reach out to more users. For example, an ad library will entice developers into including it in their apps. Figure 3 shows the 20 PIC domains with the largest app presence globally (i.e., the domains that were contacted by the largest number of apps). Based on the information we collect from Crunchbase and the company websites, we attribute these PIC domains to three functions - Metrics/Analytics (M), Advertising (A), and Development (D). As we can see in Figure 3, the majority of these PIC domains (15 out of 20) offer advertising services. For example, in addition to the PIC domains owned by Google and Facebook, several known PIC domains operated by online advertisement companies (e.g., `api.airpush.com`, `android.revmob.com`, `e.admob.com`, `ads.mopub.com`) have considerable global app presence, being contacted by 10K apps or more. Additionally, 8 out of the top 20 PIC domains offer metrics/analytics services. One noticeable finding from our study is `alog.umeng .com` (part of `Alibaba Group`).

Rank	Top 20 Domains			Top 100 Domains			1K Domains			10K Domains		
	Cat.	# Domains	# apps	Cat.	# Domains	# apps	Cat.	# Domains	# apps	Cat.	# Domains	# apps
1	device info	20	192,488	device info	99	255,794	device info	993	327623	device info	9866	353662
2	settings info	20	16,096	location info	95	62098	sim card info	891	184712	sim card info	7448	224159
3	email address	19	4,638	settings info	92	21987	location info	816	106313	location info	5415	126247
4	location info	18	36,083	email address	87	6546	phone number	646	37972	settings info	2833	33491
5	social network account	17	196	phone number	85	15450	settings info	594	299501	phone number	2700	51989
6	phone number	16	7,572	sim card info	85	100204	email address	475	9250	email address	1839	14088
7	sim card info	16	41,249	social network account	68	2623	social network account	280	4042	account info	778	6299
8	account info	15	2,876	account info	61	3255	account info	247	4537	social network account	741	6051
9	contact info	14	171	call log	47	275	call log	178	364	installed app info	489	21191
10	call log	14	162	contact info	41	258	installed app info	164	18074	call log	366	596
11	sms info	13	117	sms info	39	234	contact info	139	475	contact info	350	1384
12	installed app info	10	9,795	installed app info	33	11616	sms info	129	337	sms info	340	917

TABLE III: Top 12 Private Information collected by top 20, 100, 1K and 10K PIC domains (ranked by global app presence). We also show the number of apps collecting such information and communicating with these domains.

This domain has the largest global app presence and is contacted by 79,402 apps (3% of total apps). This domain was not reported by previous measurement studies [58], and might be because our dataset contains three orders of magnitude more devices, distributed across the globe (recall that over 7M users are located in Asia). Note that a high app penetration rate does not necessarily lead to a high device penetration rate, while the latter is directly proportional to the real amount of data collection. We will discuss this aspect in Section III-C.

We then investigate the diversification of private information collection by these PIC domains from a global perspective. Previous literature focused on unique hardware- and user-identifiers (UIDs) to study Advertising and Tracking Services (ATS) [58]. It remains an open question if these PIC domains only collect UIDs given their wide presence in mobile apps. In this study, we move beyond UIDs and leverage 22 categories of private information monitored by the security company to show a holistic picture of private information collection in the mobile ecosystem. We summarize our findings in Table III. As it can be seen, the top 20 domains collect a wide spectrum of private information (e.g., 14 out of 20 collect call log information, 13 out of 20 collect SMS information, etc). We can also observe that the top 10,000 PIC domains converge to collecting three types of private information - device (9,866 PIC domains), sim card (7,448 PIC domains), and location information (5,415 PIC domains) - which enable them to uniquely identify and track the end users for potential targeted advertising purpose [74], [37]. In contrast, the top 100 PIC domains focus on collecting more types of private information (i.e., on average, the top 100 PIC domains collect over 8 types of private information) and build a holistic profile of users (e.g., 61 out of top 100 PIC domains collect social network account information from the end users, in contrast to only 741 out of top 10,000 PIC domains collecting such information).

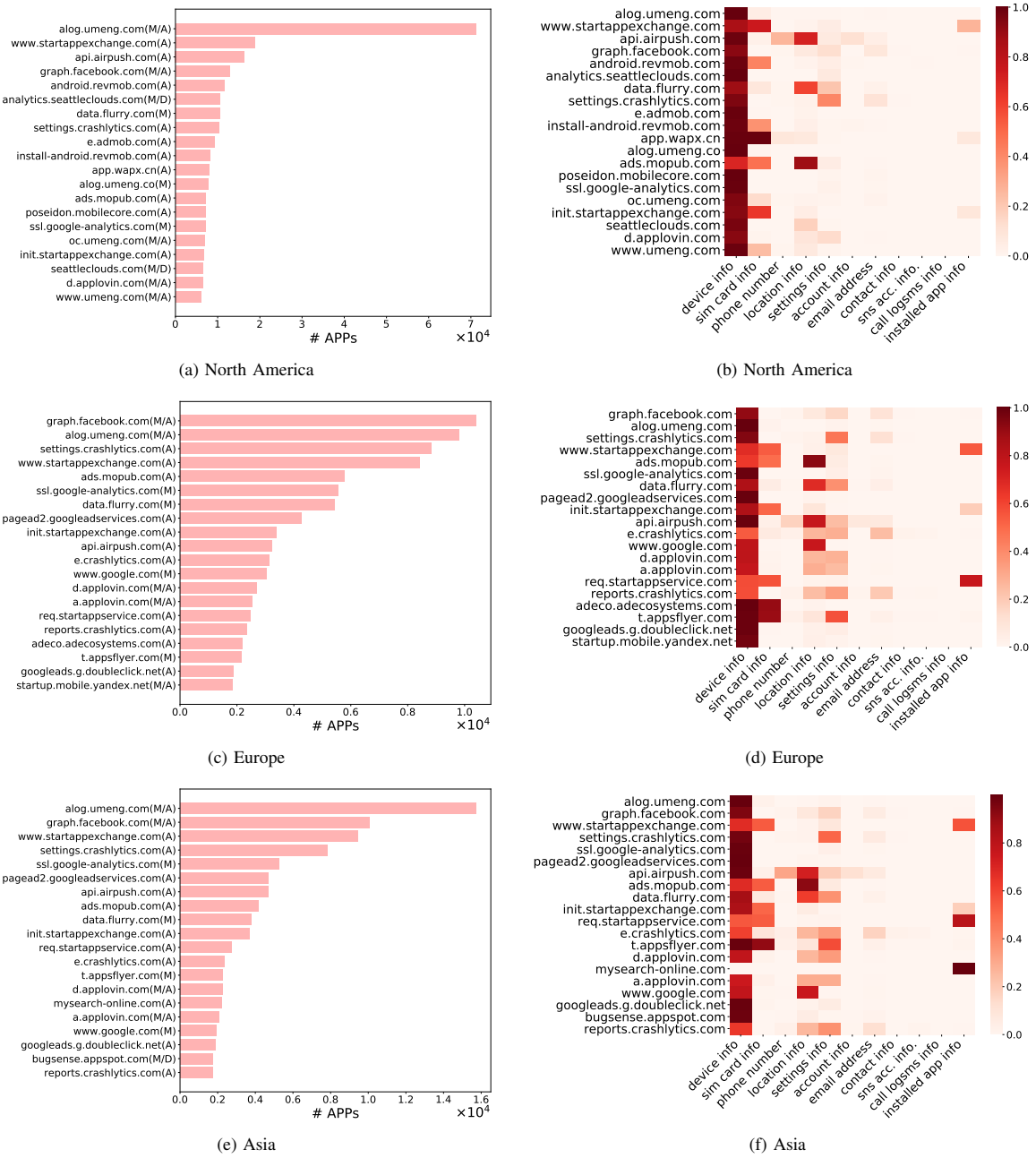
Geographic differences in PIC domains. Figure 4a, 4c and 4e show the top 20 PIC domains with the largest regional app presence in North America, Europe and Asia respectively. In addition to the top global PIC domains, we uncover that certain PIC domains have a high regional app presence and were not previously reported. For example, `poseidon.mobilecore.com` (7,046 apps, 91% of its global presence) and `seattleclouds.com` (89% of its global presence) have high app presence in North America, Russia-based `startup.mobile.yandex.net` (1,832 apps, 72.5% of its global presence) and `mysearch-online.com` (2,194 apps, 70% of its global presence), respectively, have high app presence in Europe and Asia. Regarding this regional presence phenomenon, we can only speculate that it is due to the business models adopted by these companies by focusing on serving regional markets.

At the regional level, we find that the top 20 PICs contacted by apps installed on devices in different geographical regions collect different categories of private information. Note that we consider a PIC domain *notably* collects a certain kind of private information if 20% of apps with its presence collect such information. In North America, we observe that the top 20 PIC domains (Figure 4b) mainly collect device information and sim card information, and only 3 PIC domains (`api.airpush.com`, `data.flurry.com` and `ads.mopub.com`) collect location information. In contrast, top PIC domains in Europe (Figure 4d) and Asia (Figure 4f) collect more diversified categories of private information. For example, 8 out of 20 top PIC domains collect location and settings information in both Europe and Asia, 4 out of top 20 PIC domains in Asia prevalently collect installed app information, with `mysearch-online.com` exclusively gathering such data.

C. PIC Domains: Device Penetration Study

In this section, we investigate the top PIC domains from the mobile device penetration rate perspective. We show that looking at device penetration provides different results than looking at app presence only. In fact, some of the actors who manage to get their libraries installed in many apps do not manage to have a large number of users running them.

Top PIC domains by device penetration rates. In reality, a high app presence does not necessarily lead to a high device penetration rate (i.e., the number of users sending information to PIC domains), whereas the latter is directly proportional to the real amount of information collection. In the rest of the section, we focus on the PIC domains that have high device penetration rates to uncover their private information collection dynamics in the real world. Figure 5a shows the 20 PIC domains with the largest device penetration rate globally. As we can see in Figure 5a, the top 3 PIC domains (`settings.crashlytics.com`, `graph.facebook.com`, and `ssl.google-analytics.com`) cover 8.03M, 7.8M, and 4.5M devices respectively, which are proportional to their app presence (see Figure 3). `alog.umeng.com`'s high app presence strategy also pays off covering roughly 3M devices.



(a) North America (b) North America

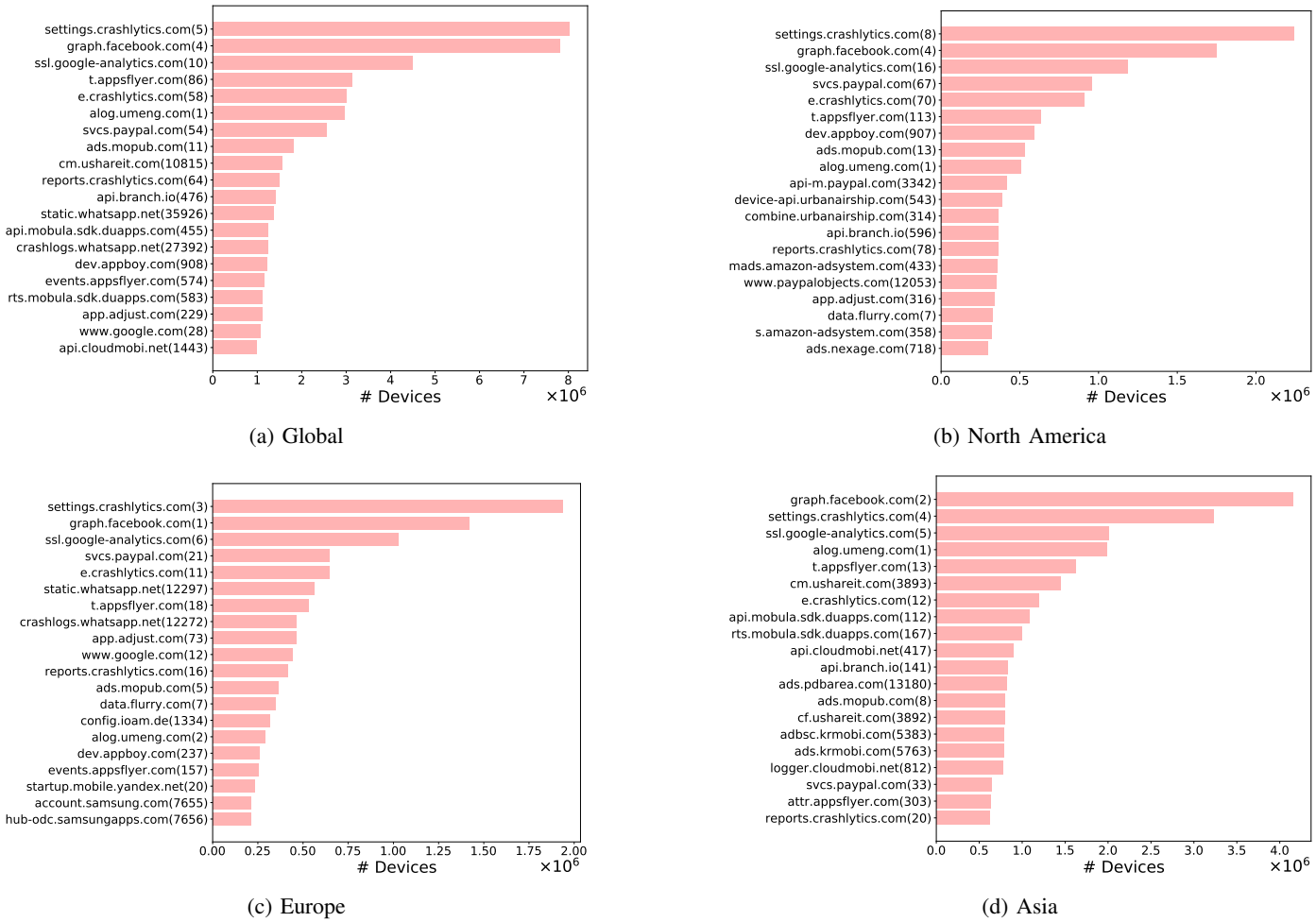
(c) Europe (d) Europe

(e) Asia (f) Asia

Fig. 4: **(left column)** Regional top 20 PIC domains ranked by app presence. Domain's primary function - **M**: Metrics/Analytics, **A**: Advertising, and **D**: Development. **(right column)** Heatmap illustration of top 12 categories of private information collected by these PIC domains. Each row is normalized to [0, 1] by a PIC domain's total app presence. The darker the red implies that the more apps that a PIC domain collects information from.

However, quite a few PIC domains with high app presence failed to gain high device penetration rates. For example, `api.airpush.com` only covers 68K devices despite of its high app presence (21K apps, Figure 3). Besides, PIC domains controlled by `revmob.com`, `seattleclouds.com` and `mobilecore.com` also did not manage to have high prevalence in the devices.

Geographic differences in PICs. We also discover that different regions present different dominant PIC domains. For example, `*.urbanairship.com` and `mads.amazon-adsystem.com` have high device penetration rate in North America. `config.ioam.de` (99.4% of its global presence) solely operates in Europe. `cm.ushareit.com` (92.4% of its global presence), `api.mobula.sdk.duapps.com` (88% of its global presence), `ads.pdbarea.com` (825K, 95.4% of its global presence) and `adbsc.krmobi.com` (95.4% of its global presence) are almost exclusively contacted by devices located in Asia. We further investigate the types of private information collected by PIC domains with high device penetration rates,



(a) Global

(b) North America

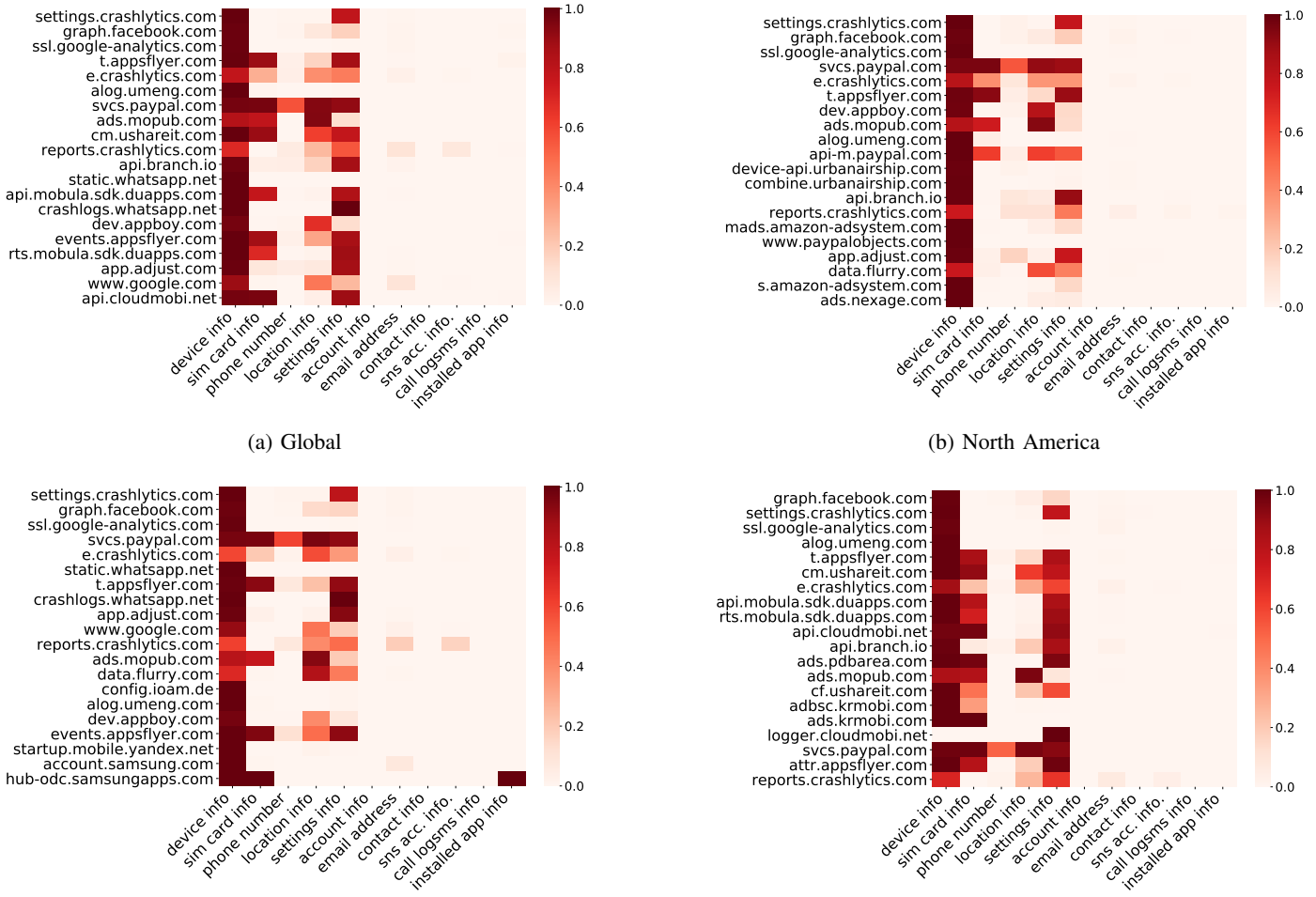
(c) Europe

(d) Asia

Fig. 5: Top 20 PIC domains ranked by device penetration rate. The number next to a PIC domain represents its ranking by app presence.

to check if different players active in different regions are interested in different types of private information. Our findings are summarized in Figure 6. Each row is normalized by a PIC domain's total device penetration rate. The heatmaps illustrate the main types of information collected by the PIC domains. First, it is interesting to see in Figure 6 that the top 20 global and regional PIC domains with high device penetration rate focus on collecting *four* types of private information from the end users - *device*, *sim card*, *location* and *settings* information. For example, all of the top 20 PIC domains in Figure 6 collect device information. The only exception is `logger.cloudmobi.net`, a prominent PIC active in Asia (see Figure 6d), which predominantly collects device setting information. Approximately 50% of the top PIC domains collect sim card, location, and setting information at both global and regional levels. Our findings also show that certain PIC domains consistently collect multiple types of private information from devices, potentially enabling them to track the end users more systematically. For example, in Europe `events.appsflyer.com` (1.16M global device penetration rate) collects device information from all devices that connected to it, and sim card information (and setting information) from 95% of them (see Figure 6c). Similarly, `ads.mopub.com` with 1.8M global device penetration rate (see Figure 6a, 6b and 6c) exhibits a similar behavior, i.e., collects location, device, and sim card information from over 80% of the devices that connected to it. In Section III-B, we show that these two behavior patterns are different from the ones observed when looking at the top PIC domains ranked by app presence, where the intention is to collect more diversified private information (see Table III). Our findings can be treated as the profiles of PIC domains, and help the community understand their behavior in fine granularity (e.g., understanding the correlation between domain naming conventions and the types of private information collected).

Summary of findings. We found that looking at app presence can provide misleading results. In fact, some of the actors who managed to get their libraries installed in many apps failed to have many users running them. Further information can be found in Section III-B. We also found that certain PIC domains consistently collect multiple types of private information from the devices and are capable of tracking the end users more systematically. We observed different regional players targeting users in different continents, and collecting different types



(a) Global

(b) North America

(c) Europe

(d) Asia

Fig. 6: Heatmap illustration of top 12 types of private information collected by both global and regional top 20 PIC domains. Each row is normalized to [0, 1] by a PIC domain's total device penetration rate. The darker the red implies that the more devices that a PIC domain collects information from.

of private information. Following these observations, we will further discuss the data controllers behind these PIC domains and the implications of data protection in Section V.

IV. Private Information Destinations

In the previous section, we focused on end user devices, looking at the top PIC domains that collected private information from them. In this section, we focus on the destination of those private information flows, aiming to understand the countries where these flows terminate.

Geolocation of PIC Domains. We leverage the technique detailed in Section II to uncover the geolocation of the PIC domains and summarize our findings in Figure 7. Our analysis reveals that United State and China are the largest two countries hosting the PIC domains. The United States hosts 44% of PIC domains, which is in line with the previous literature [58] and China hosts 26.1% of PIC domains. This figure is three times higher than previously reported [58]. PIC domains hosted in the US and China collect private information from 14M devices (80.9% of global devices) and 4.6M devices (26.5% of global devices) respectively. Other countries host significantly fewer PIC domains compared to the United States and China (e.g., South Korea, ranked 3rd in the list, hosts merely 2.6% of the PIC domains). Note that the geolocation of 6.4% of PIC domains could not be identified because our approach cannot trace their historical domain records.

Global private information flow. As we saw in Section III, a PIC domain can collect multiple types of private information from the end user. We further investigate the global private information flow from the mobile devices to the PIC domains. The result is shown in Figure 8a. PIC domains hosted in the United States collect 62% (of which 42.3% coming from out of the country) of global private information flows. PIC domains hosted in China collect 7% of private information flows from 4.59M devices globally. This figure is almost four times more than previously reported [58]. At the same time, PIC domains hosted in Singapore collect 6.53% of global private information flows (mainly from India). The rest of the

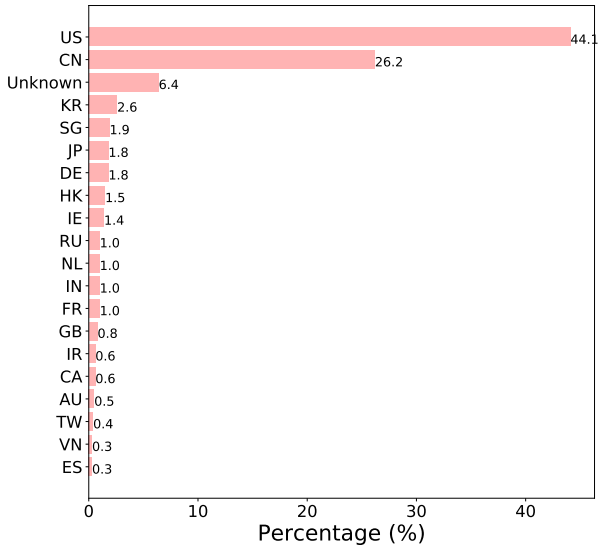


Fig. 7: Global top 20 countries ranked by the number of PIC domains hosted.

countries shown in Figure 8a notably collect much less private information comparing to these three countries.

European private information flow and the effect of GDPR [80]. The European Union's (EU) General Data Protection Regulation (GDPR) entered into effect on May 25th, 2018. It imposes obligations onto organizations in any country so long as they target or collect data related to people in EU countries (EU28). If data is being transferred to a third-party and/or outside the EU28, GDPR requires that data subjects must be clearly informed about the extent of data collection, the legal basis for the processing of personal data, how long data is retained. In light of this legislation, we measure the private information flows originated from EU countries before (January 5th, 2018 - May 24th, 2018) and after (May 26th, 2018 - September 30th, 2019) the GDPR effective date, and check if GDPR has a real-world impact to private information collection. Our findings are shown in Figure 8b and 8c. As we can see, private information confinement within the EU is low. PIC domains hosted in the United States dominate the private information collection in the EU, collecting 68% and 66% of European private information flows respectively before and after the GDPR. This figure is 30% lower than previously reported 89.2% [58]. At the same time, Germany and Ireland are the only two European countries that host a reasonable portion of PIC domains and good control of private information can be applied, while the other European countries hosting a very small fraction of PIC domains and US remains the largest hosting country. Notably, we uncover that approximately 4.4% and 1.7% of private information flows are collected by PIC domains hosted in Russia and China respectively [87], [88].

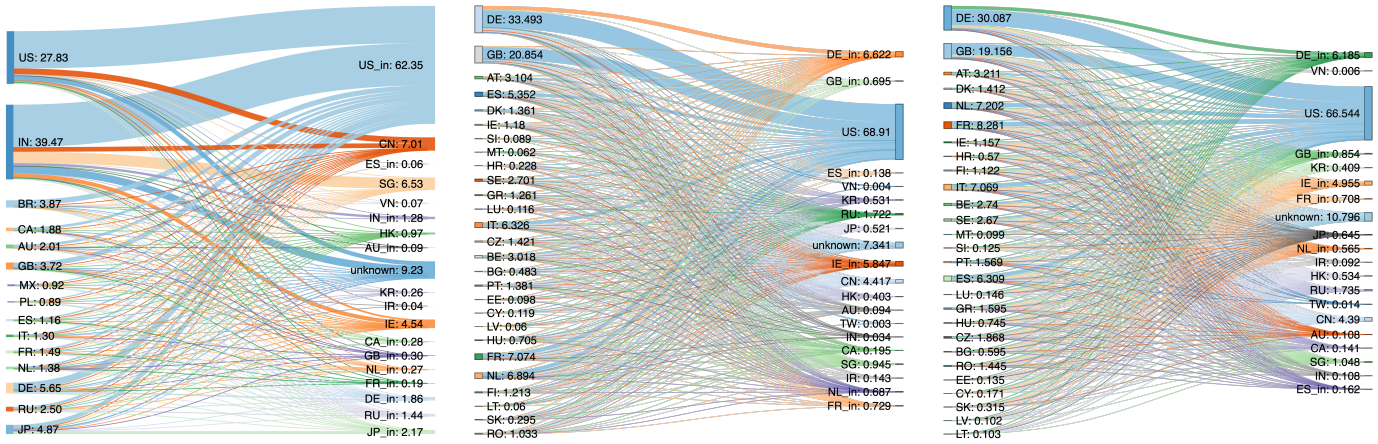
It is also interesting to see that private information collection in Europe is not affected by GDPR in general. As we can see in Figure 8b and Figure 8c, the fractions of private information collected by these PIC domains (and consequently the countries hosting them) remains stable regardless of the implementation of GDPR. Our results show that GDPR has not stopped companies from collecting private information from end users as long as their services are GDPR-compliant, partially because that the GDPR treats first-party data uses more leniently [30]. However, it remains an unanswered question, especially to the consumers, how to trace their private information after sharing with the GDPR-compliant companies, and how accountability can be truly guaranteed [81], [47], [49]. For instance, which company should be held accountable if a device identifier was abused (e.g., targeted advertising) while the majority of apps in mobile devices collect device identification information as shown in Section III-C? We aim at studying this question in Section V.

V. Data Processors and Controllers

In the previous sections, we provided an overview of the landscape of private data collection from mobile devices (Section III) and of the countries where private information is sent to (Section IV). In this section, we aim at understanding the characteristics of the data processors and controllers who ultimately obtain and process the private information and the implications of their privacy policies to the end users.

Overview of top data processors and controllers. We select the top 10k PIC domains covering all the devices in this study, and use the technique detailed in Section II to uncover the ownership of the PIC domains. The top 25 data processors and controllers (ranked by the fraction of devices they collect private information from) are shown in Figure 9. In total, these 25 data processors and controllers collect private information from 13.9M devices (80.2% of all devices used in this study). Facebook and Alphabet are the two dominant data controllers, collecting private information from 9.3M and 9.1M devices respectively. AppsFlyer is the third largest data processor/controller collecting information from 3.4M devices. It is worth noting that there are six Chinese companies among the global top 25 data processors and controllers: Alibaba (3.1M), Baidu (1.6M), CloudMobi (1.0M), MobVista (880K), Tencent (650K), and Intsig(Shanghai) (646K). In total, these six companies are collecting private information from 4.55M devices (i.e., 26% of total devices).

Operation of top data processors and controllers. We analyze the domain distribution of these top 25 data processors and controllers to understand more details on their infrastructure and on their operational strategies. Our findings are summarized in Figure 10. In total, 16 out of the top 25 data processors and controllers have no more than 21 PIC domains. For example, AppsFlyer, the third largest data processor/controller, has only 11 PIC domains in our dataset. It is evident that the majority of the data controllers prefer to control data flow via several API gateways. At the same time, Baidu (425 PIC domains), Tencent (531 PIC domains), and Adobe (374) prefer to use many loosely coupled services to collect data since their operational strategies rely on the Cloud infrastructure. For example, DU Ad Platform (`*.duapp.com`, part of Baidu) almost exclusively runs in the AWS infrastructure, QQ platform (`*.qq.com`, part of Tencent) operates in Tencent owned Cloud infrastructure, and 2o7 (`*.2o7.net`) is part of Adobe Marketing Cloud. Note that previous literature [58] found that "*292 parent organizations that own nearly 2,000 ATS and ATS-C domains.*" Our findings, however, indicate that these data controllers may own more PIC domains that previously thought.



(a) (b) (c)

Fig. 8: Sankey diagrams illustrating 1) Global private information flows between the top 15 countries (ranked by the number of devices) and top 20 PIC domain locations (a) and 2) Private information flows between EU28 and top 20 PIC domain locations before (b) and after (c) GDPR. Note that the left side of the diagrams represents the origin of information flows and the right side represents where the information flows terminate. We add a postfix '_in' to the country code at the right hand side in case of private information flows originating and terminating at the same country.

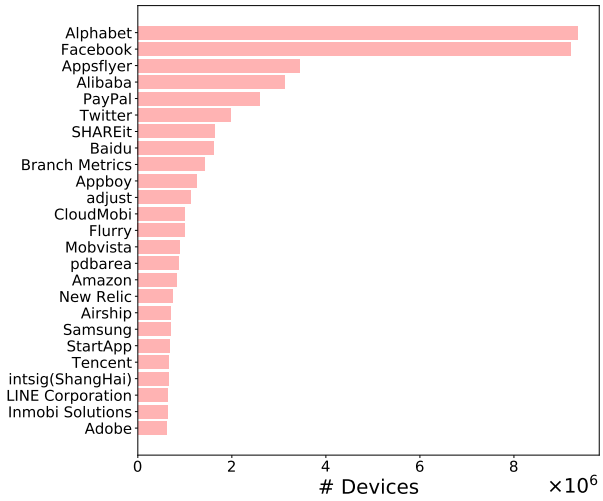


Fig. 9: Global top 25 data controllers ranked by the fraction of devices they collect private information from. These 25 data controllers collect private information from a total of 13.9M devices covering 80.2% of all devices used in this study.

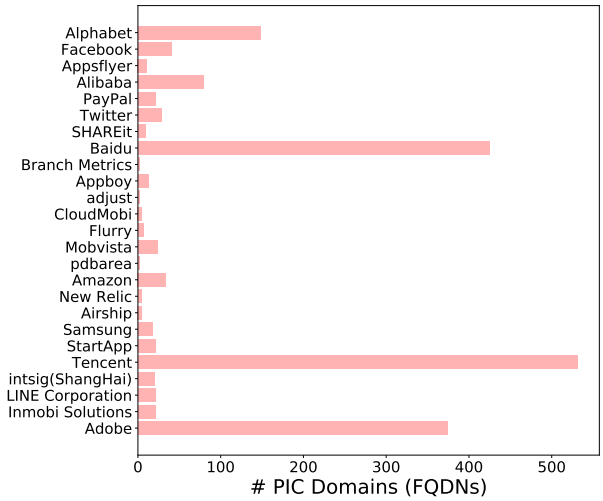


Fig. 10: Domain distribution: top 25 data controllers.

Cross-border transfer, non-EU data processor and controllers, and implications to data protection. Based on our factual findings, we use Chinese companies as a case study to quantitatively and objectively understand the implications of users' private information collection when involving cross-border data transfer [81], [47], [49], [75] and how it becomes more difficult to trace how this data flows. As mentioned before, the top six Chinese companies are collecting private information from 4.55M devices. Superficially, such coverage seems in line with our findings in Section IV where we found that 7% of private information flows from 4.59M devices globally flow to China. We further investigate the geolocation of the PIC domains controlled by these companies and see if these domains are hosted in China using the technique detailed in Section II.

For example, Baidu has 210 PIC domains hosted outside China, mainly because the `*.duapp.com` PIC domains (owned by its subsidized DU Ad Platform) are hosted in AWS (USA). Besides, `*.mobvista.com` is hosted in Amazon Web Services (AWS) and `*.cloudmobi.net`, has a mixture of hosting environments in the US and Singapore. We report more details about the country distribution of Chinese data controllers in Figure 11. The figure confirms that many PIC domains owned by these companies are not hosted in China. Such operational strategy employed by these data controllers, however, leads to undesirable implications for data protection.

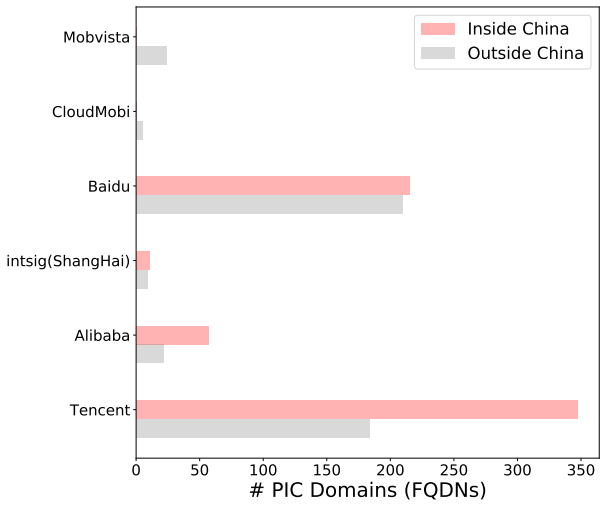


Fig. 11: Domain distribution: top 6 Chinese data controllers.

For example, the DU Ad Platform (partnering with Facebook, Alphabet, appnext, etc.) states in its privacy policy[2] that personal information could be "*shared with any organization part of Baidu Group*" and "*may be transferred to countries which provide an adequate level of protection*." In this case, even though private information flows terminate in the AWS Cloud, such data could still be transferred to third countries. Moreover, Mobvista (partnering with Baidu, TikTok, etc) explicitly claims in its privacy policy[3] that private information would be "*transferred to recipients in countries located outside the EEA (including in Singapore where the Site is hosted) which do not provide a similar or adequate level of protection to that provided by countries in the EEA*."We acknowledge that without knowing more about the actual underlying contractual relationships it is difficult to draw conclusions on how data is further processed by those entities. Nevertheless, it remains an open yet important question on how to protect and audit the usage of such data flows terminated at the PIC domains owned by these companies with data transfers to third countries explicitly stated in the privacy policy. We hope that our findings will motivate lawmakers to consider how to address such issues in the future legislations, and more importantly, encourage the commercial partners of these companies to design rigorous policies to protect user private information when sharing data cross-border.

Summary of findings. We found that the top 25 data processors and controllers can collect private information from an overwhelming 80.2% of all devices. 6 top Chinese data controllers provided privacy policies and hosted part of their infrastructure in countries with rigorous data protection laws. However, they also allow data transfer to third countries and may incur technical and legal complications on how to further protect private information [87], [88], [81], [47], [49].

[2]http://ad.duapps.com/gdpr/index.html#title-2
[3]https://www.mobvista.com/en/privacy/

Rank	Cat.	# PHAs	# Dev.
1	Device Info	295K	1.45M
2	Sim Card Info	167K	993K
3	Location Info	127K	670K
4	**Operator Info**	**116K**	**393K**
5	Installed App Info	91K	486K
6	Phone Number	75K	364K
7	**Running App Info**	**63K**	**280K**
8	Account Info	17K	73K
9	Settings Info	10K	376K
10	Email Address	4K	107K

TABLE IV: Top 10 private information collected by PHAs on a global scale.

Fig. 12: Heatmap illustration of regional private information collection by PHAs.

VI. Characterization of PHA Private Information Collection

Potentially harmful applications (PHAs)[28] are apps that could put users, user data, or devices at risk (e.g., trojan, spyware, etc.). Some of them aren't strictly malware but are harmful to the software ecosystem (e.g., impersonating other apps). These PHAs have been substantially discussed and studied in the previous literature [40], [90], [24], [21], [14]. In this section, we focus on understanding what private information is collected by PHAs. In particular, we aim to understand whether the type of information collected by PHAs is different from the one collected by regular apps.

Private information collection by PHAs. We consider a SHA2 as potentially harmful if it is flagged by at least 6 AV companies in VirusTotal (see Section II). Together with mobile app reputation data, we identify 3.5M SHA2s associated with 1.2M unique PHA app names that were installed on 3.8M devices. Following the analytical process used in Section III, we uncover the top 10 types of private information collected by PHAs and summarize our findings in Table IV. We can see that PHAs mainly collect tracking information, e.g., device info, sim card, location, etc. Besides, 116K PHAs (covering 393K devices) collect operator information and 63K PHAs (covering 280K devices) also collect running app information on a global scale. This is more aggressive comparing to the private information collection behavior comparing to 43K/42K benign apps respectively collecting such information. As we can see in Figure 12, the majority of these aggressive PHAs are installed on devices in North America. Note that such aggressive private information collection behavior enables adversaries to better profile end users and may lead to some intrusive monetization actions. For example, we uncover that 590K devices with PHAs presence are affected by notification bar ads (i.e., ads are displayed as app notifications) and 317k devices suffer

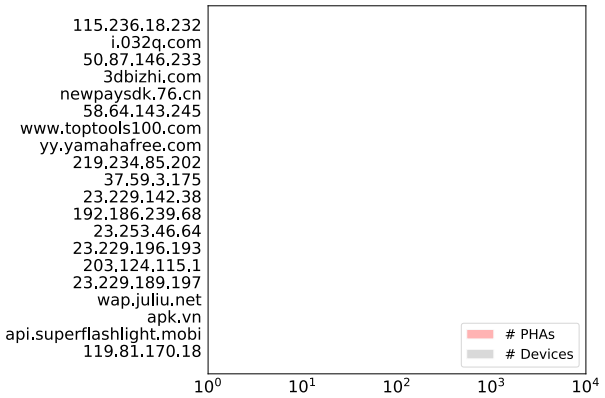


Fig. 13: **(log-scale)** Top 20 malicious IPs/Domains ranked by the number of PHAs.

from short-cut ads (i.e., targeted ads are placed on the home screen). Yet, only 230K devices with PHA installations exhibit in-context ads behavior (i.e., normal behavior as ads are displayed inside an app). However, due to the limitation of our system, we are not able to measure the content correlation between private information collected by PHAs and the subject of advertisement displayed as shortcuts on the devices. We also identify 1,549 PHAs (4,930 SHA2s) that read/sent SMS from 4,461 devices. Even though such SMS leakage is minor in terms of device prevalence ratio, in light of the recent discussion of limitation of SMS-based 2FA authentication[4], our findings show that the possibility of such breaches still exists in the wild.

Communications with malicious domains. We compile a blacklist of IPs/domains that have been involved with malicious activities from various sources (see Section II), and aim at understanding if PHAs send private information collected from these devices to malicious domains. Figure 13 shows the 20 malicious domains with the largest app presence and the fraction of devices connecting to them. As we can see in Figure 13, `115.236.18.232` has the largest app presence and was contacted by 550 PHAs collecting data from 686 devices. `www.toptools100.com` and `yy.yamahafree.com` have higher device penetration rates, respectively showing communications with 3,789 and 6,455 devices respectively. In general, we find that only a small portion of PHAs communicate with known malicious hosts and domains, and such domains have limited device coverage. This is different from PC malware while a considerable fraction of malware connect with malicious domains and are part of botnets [6], [54], [31].

Summary of findings. We found that PHAs are more aggressive comparing to generic private information collection behavior, leading to intrusive monetization actions. However, communications with malicious domains are less pervasive comparing to desktop applications.

[4]https://krebsonsecurity.com/2018/08/reddit-breach-highlights-limits-of-sms-based-authentication/

VII. Discussion and Limitations

Implications for the research community. Our study shows that looking at device penetration provides different results than looking at apps only. Designing measurement studies focused on executing apps could lead to conclusions that are biased and do not reflect real malicious activity in the wild. In fact, some of the actors who manage to get their libraries installed in many apps do not manage to have many users running them. In light of this, we hope that our study can inspire security researchers to design measurement studies that are representative of the real world as possible.

Implications for policymakers. We observe that private information confinement within the EU is low. GDPR has not stopped companies from collecting private information from the end users as long as their services are GDPR-compliant. In light of these findings, we hope that our study would encourage policymakers to further regulate how private information is used by and shared among the companies and how accountability can be truly guaranteed (e.g., , which company should be held accountable if a device identifier was abused by targeted advertising while the majority of apps on that mobile device collect device identification information).

Study limitations. Our study relies on the static and dynamic analysis, and layered security engines at the backend to identify and fingerprint that certain API calls lead to specific private information leakage. This prevents us to capture private information collection activities that happen at runtime but are not captured by the company's analytical infrastructure, which the on-device security engine relies on. Therefore, our work covers the lower bound of global private information collection activities. Nevertheless, despite such limitations, our study provides the most comprehensive view of private data collection by Android apps to date and actionable insights.

While this paper is based on measurements collected from a user base that is three orders of magnitude larger than previous work, our dataset is biased towards the end users of a single mobile security product, and therefore still presents some biases. For example, the distribution of devices used this study is not heavily screwed towards any specific region. However, the device distribution in Asia is skewed towards India and Japan and does not have as many devices in China which is one of the top countries/markets in terms of mobile users. In terms of the representativeness of the analyzed apps, it is challenging to ascertain the coverage of our study since it is infeasible to determine the total number of all Android apps, given such a fragmented ecosystem and many alternative markets. Still, by analyzing 2.1M apps this study is covering one of the largest sets of apps to date and is in line with the largest datasets collected by the academic community [5].

Our analysis of *data controllers* presented in Section V rely on the identification of the organizations behind PIC domains. As detailed in Section II the mapping of domains to their owner organization relies on multiple data sources providing connections between the domains, the networks that host them as well as the organizations supposedly maintaining these resources. Such connections are cross-checked in the different data sources to compensate for inaccuracies in each of the sources. We also take a conservative approach and automatically discard all connections that are not seen in all

data sources. This naturally hurts the number of domains to which we can map an organization. However, we favor the accuracy of the domain to owner organization mapping over its coverage. It is also important to note that some apps communicate with raw IP addresses instead of relying on domains. Moreover, we have seen that more than 97% of these IP addresses refer to CDNs or hosting or cloud providers which hinder the identification of their owner organization.

Finally, while in this paper we studied how private information is collected from devices, and where this information flows to, our study does not allow us to understand how this information is acted upon by data controllers (i.e., whether and how it is used to track users). This remains an open question for the research community.

VIII. Related Work

In this section, we selectively review previous studies on PHA characterization, Android permission system, private data leakages and prevention, and third party advertising and tracking services. We refer the readers to [72], [23], [20], [51], [84], [69] for in-depth studies and surveys on securing Android devices in general.

PHA characterization. Previous studies mainly focused on analyzing PHAs and systematically characterize them from various aspects such as evasion mechanism [21], installation methods [90], malicious payloads [90], repackaging mechanism [89], [69], [40], behaviors [39], [85], monetization [24], etc. These efforts shed light on how Android PHAs operate in the wild [90], main incentives of mobile malware [24], [40], weaknesses of some of the popular mitigation solutions [21], etc. However, they did not discuss potential threats posed by information collection on mobile devices as these efforts center on app analysis and offer a less comprehensive view of the real device prevalence.

Android permission system. Android permission system has been extensively covered in the previous literature [11], [51], [23], [8], [10]. These studies on Android permissions have mainly leveraged static analysis techniques to understand the role of a given permission [8], [23], [26], potential privacy violation incurred by overprivileged apps [23], [64], permission circumvention [59], description-to-permission fidelity [56], and improve mapping of Android permissions to framework/SDK API methods [9], [2]. Some recent research efforts also utilize dynamic analysis systems to distinguish and trace the permissions requested by apps at the runtime and those requested by the app's core functionality [18] and generate a more precise call graph enabling the system to extract the permission specification and improve the mapping [43]. Our study complements these studies by showing the scales and the prevalence of private information collection in the real world devices.

Third-party advertising and tracking services (ATSes). There are two main approaches on studying third-party advertising and tracking services (ATSes). One approach leverages static tools to decompile apps and identified the embedded trackers from API calls and quantify various aspects of trackers [65], [12]. These methods offer a view of tracker behavior and prevalence from app perspective. Another approach is leveraging network traffic either captured on device or by ISP providers to provide insights into the mobile advertising and tracking ecosystem from an information flow perspective [33], [77], [58], [32].

PII leakage detection and protection. The root cause of PII leakage is because the end users are presented with a set of required permissions by the apps but not how they handle the data after permissions are granted. Previous studies showed that mobile apps leak more privacy information than their web counterpart [53], [36]. To this end, research efforts mainly focused on monitoring private information flows [19], [71], detecting potential privacy leaks by apps [27], [60], [59], sensitive data leakage via third party libraries [68], [17], [41], privacy implication caused by targeted advertising in apps [13], private data leakage via network traffic analysis [61], [16], impact of GDPR notices [76], and privacy implications incurred by pre-installed apps [25]. Our work complements the previous work and shows a holistic picture of the state of sensitive information collection on Android in the wild, identifying the big players in this space (both legitimate companies and malicious actors), together with geographic trends.

IX. Conclusion

In this paper, we presented the most comprehensive measurement study on private information collection on Android to date. We showed that PIC is widespread on Android, and that various types of information are collected, with actors operating in different geographic areas interested in different types of information. While most information flows terminate in the US, 7% of the flows that we observe are directed to China. We also find that data regulation laws like GDPR have not been effective in limiting the amount of personal information that flows to third countries **outside EU**.

Acknowledgments

We wish to thank the anonymous reviewers for their feedback and our shepherd Adwait Nadkarni for his help in improving this paper.

References

[1] Y. Aafer, W. Du, and H. Yin. Droidapiminer: Mining api-level features for robust malware detection in android. In *SecureComm*, 2013.

[2] Y. Aafer, G. Tao, J. Huang, X. Zhang, and N. Li. Precise android api protection mapping derivation and reasoning. In *ACM CCS*, 2018.

[3] Abuse.ch. Fighting Malware and Botnets. https://www.team-cymru.com/bars-feed.html, 2019.

[4] G. Acar, M. Juarez, N. Nikiforakis, C. Diaz, S. Gürses, F. Piessens, and B. Preneel. Fpdetective: dusting the web for fingerprinters. In *ACM CCS*, 2013.

[5] K. Allix, T. F. Bissyandé, J. Klein, and Y. Le Traon. Androzoo: Collecting millions of android apps for the research community. In *MSR*, 2016.

[6] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, and D. Dagon. Detecting malware domains at the upper dns hierarchy. In *USENIX Security*, 2011.

[7] S. Arzt, S. Rasthofer, C. Fritz, E. Bodden, A. Bartel, J. Klein, Y. Le Traon, D. Octeau, and P. McDaniel. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. *Acm Sigplan Notices*, 2014.

[8] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie. Pscout: analyzing the android permission specification. In *ACM CCS*, 2012.

[9] M. Backes, S. Bugiel, E. Derr, P. McDaniel, D. Octeau, and S. Weisgerber. On demystifying the android application framework: Re-visiting android permission specification analysis. In *USENIX Security*, 2016.

[10] M. Backes, S. Bugiel, O. Schranz, P. von Styp-Rekowsky, and S. Weisgerber. Artist: The android runtime instrumentation and security toolkit. In *EuroS&P*, 2017.

[11] D. Barrera, H. G. Kayacik, P. C. Van Oorschot, and A. Somayaji. A methodology for empirical analysis of permission-based security models and its application to android. In *ACM CCS*, 2010.

[12] R. Binns, U. Lyngs, M. Van Kleek, J. Zhao, T. Libert, and N. Shadbolt. Third party tracking in the mobile ecosystem. In *ACM WebSci*, 2018.

[13] T. Book and D. S. Wallach. An empirical study of mobile ad targeting. *arXiv preprint arXiv:1502.06577*, 2015.

[14] R. Chatterjee, P. Doerfler, H. Orgad, S. Havron, J. Palmer, D. Freed, K. Levy, N. Dell, D. McCoy, and T. Ristenpart. The spyware used in intimate partner violence. In *IEEE S&P*, 2018.

[15] Q. Chen and A. Kapravelos. Mystique: Uncovering information leakage from browser extensions. In *ACM CCS*, 2018.

[16] A. Continella, Y. Fratantonio, M. Lindorfer, A. Puccetti, A. Zand, C. Kruegel, and G. Vigna. Obfuscation-resilient privacy leak detection for mobile apps through differential analysis. In *NDSS*, 2017.

[17] S. Demetriou, W. Merrill, W. Yang, A. Zhang, and C. A. Gunter. Free for all! assessing user data exposure to advertising libraries on android. In *NDSS*, 2016.

[18] M. Diamantaris, E. P. Papadopoulos, E. P. Markatos, S. Ioannidis, and J. Polakis. Reaper: Real-time app analysis for augmenting the android permission system. In *CODASPY*, 2019.

[19] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B.-G. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems (TOCS)*, 2014.

[20] Z. Fang, W. Han, and Y. Li. Permission based android security: Issues and countermeasures. *computers & security*, 43, 2014.

[21] P. Faruki, A. Bharmal, V. Laxmi, V. Ganmoor, M. S. Gaur, M. Conti, and M. Rajarajan. Android security: a survey of issues, malware penetration, and defenses. *IEEE communications surveys & tutorials*, 17(2), 2014.

[22] K. Fawaz and K. G. Shin. Location privacy protection for smartphone users. In *ACM CCS*, 2014.

[23] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner. Android permissions demystified. In *ACM CCS*, 2011.

[24] A. P. Felt, M. Finifter, E. Chin, S. Hanna, and D. Wagner. A survey of mobile malware in the wild. In *SPSM*, 2011.

[25] J. Gamba, M. Rashed, A. Razaghpanah, J. Tapiador, and N. Vallina-Rodriguez. An analysis of pre-installed android software. In *IEEE S&P*, 2020.

[26] X. Gao, D. Liu, H. Wang, and K. Sun. Pmdroid: Permission supervision for android advertising. In *SRDS*, 2015.

[27] C. Gibler, J. Crussell, J. Erickson, and H. Chen. Androidleaks: automatically detecting potential privacy leaks in android applications on a large scale. In *TRUST*, 2012.

[28] Google. Android Security & Privacy 2018 Year In Review. 2019.

[29] S. Hao, B. Liu, S. Nath, W. G. Halfond, and R. Govindan. Puma: programmable ui-automation for large-scale dynamic analysis of mobile apps. In *MobiSys*, 2014.

[30] C. J. Hoofnagle, B. van der Sloot, and F. Z. Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1), 2019.

[31] C. C. Ife, Y. Shen, S. J. Murdoch, and G. Stringhini. Waves of malice: A longitudinal measurement of the malicious file delivery ecosystem on the web. In *AsiaCCS*, 2019.

[32] C. Iordanou, G. Smaragdakis, I. Poese, and N. Laoutaris. Tracing cross border web tracking. In *IMC*, 2018.

[33] C. Joe-Wong, S. Ha, and M. Chiang. Sponsoring mobile data: An economic analysis of the impact on users and content providers. In *INFOCOM*, 2015.

[34] Kaspersky. How banking Trojans bypass two-factor authentication. https://www.kaspersky.com/blog/banking-trojans-bypass-2fa/11545/, 2019.

[35] A. Lerner, A. K. Simpson, T. Kohno, and F. Roesner. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *USENIX Security*, 2016.

[36] C. Leung, J. Ren, D. Choffnes, and C. Wilson. Should you use the app for that? comparing the privacy implications of app-and web-based online services. In *IMC*, 2016.

[37] K. Li and T. C. Du. Building a targeted mobile advertising system for location-based services. *Decision Support Systems*, 54(1), 2012.

[38] L. Li, T. F. Bissyandé, M. Papadakis, S. Rasthofer, A. Bartel, D. Octeau, J. Klein, and L. Traon. Static analysis of android apps: A systematic literature review. *Information and Software Technology*, 2017.

[39] M. Lindorfer, M. Neugschwandtner, L. Weichselbaum, Y. Fratantonio, V. Van Der Veen, and C. Platzer. Andrubis–1,000,000 apps later: A view on current android malware behaviors. In *BADGERS*, 2014.

[40] M. Lindorfer, S. Volanis, A. Sisto, M. Neugschwandtner, E. Athanasopoulos, F. Maggi, C. Platzer, S. Zanero, and S. Ioannidis. Andradar: fast discovery of android applications in alternative markets. In *DIMVA*, 2014.

[41] X. Liu, J. Liu, S. Zhu, W. Wang, and X. Zhang. Privacy risk analysis and mitigation of analytics libraries in the android ecosystem. *IEEE Transactions on Mobile Computing*, 2019.

[42] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey, and M. Liu. Cloudy with a chance of breach: Forecasting cyber security incidents. In *USENIX Security*, 2015.

[43] L. Luo. Heap memory snapshot assisted program analysis for android permission specification. In *SANER*, 2020.

[44] S. Ma, Z. Tang, Q. Xiao, J. Liu, T. T. Duong, X. Lin, and H. Zhu. Detecting gps information leakage in android applications. In *IEEE GLOBECOM*, 2013.

[45] Maxmind. Maxmind IP Geolocation. https://www.maxmind.com/, 2019.

[46] J. R. Mayer and J. C. Mitchell. Third-party web tracking: Policy and technology. In *IEEE S&P*, 2012.

[47] T. Minssen, C. Seitz, M. Aboy, and M. C. Compagnucci. The eu-us privacy shield regime for cross-border transfers of personal data under the gdpr. *European Pharmaceutical Law Review*, 4(1), 2020.

[48] N. Miramirkhani, M. P. Appini, N. Nikiforakis, and M. Polychronakis. Spotless sandboxes: Evading malware analysis systems using wear-and-tear artifacts. In *IEEE S&P*, 2017.

[49] T. Mulder and M. Tudorica. Privacy policies, cross-border health data and the gdpr. *Information & Communications Technology Law*, 28(3), 2019.

[50] S. Nath. Madscope: Characterizing mobile in-app targeted ads. In *MobiSys*, 2015.

[51] M. Nauman, S. Khan, and X. Zhang. Apex: extending android permission model and enforcement with user-defined runtime constraints. In *ASIACCS*, 2010.

[52] E. Pan, J. Ren, M. Lindorfer, C. Wilson, and D. Choffnes. Panoptispy: Characterizing audio and video exfiltration from android applications. *PETS*, 2018(4), 2018.

[53] E. P. Papadopoulos, M. Diamantaris, P. Papadopoulos, T. Petsas, S. Ioannidis, and E. P. Markatos. The long-standing privacy debate: Mobile websites vs mobile apps. In *WWW*, 2017.

[54] D. Plohmann, K. Yakdan, M. Klatt, J. Bader, and E. Gerhards-Padilla. A comprehensive measurement study of domain generating malware. In *USENIX Security*, 2016.

[55] S. Poeplau, Y. Fratantonio, A. Bianchi, C. Kruegel, and G. Vigna. Execute this! analyzing unsafe and malicious dynamic code loading in android applications. In *NDSS*, 2014.

[56] Z. Qu, V. Rastogi, X. Zhang, Y. Chen, T. Zhu, and Z. Chen. Autocog: Measuring the description-to-permission fidelity in android applications. In *ACM CCS*, 2014.

[57] Rapid7 Labs. Forward DNS (FDNS). https://opendata.rapid7.com/sonar.fdns_v2/, 2019.

[58] A. Razaghpanah, R. Nithyanand, N. Vallina-Rodriguez, S. Sundaresan, M. Allman, C. Kreibich, and P. Gill. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *NDSS*, 2018.

[59] J. Reardon, Á. Feal, P. Wijesekera, A. E. B. On, N. Vallina-Rodriguez, and S. Egelman. 50 ways to leak your data: An exploration of apps' circumvention of the android permissions system. In *USENIX Security*, 2019.

[60] J. Ren, M. Lindorfer, D. J. Dubois, A. Rao, D. Choffnes, and N. Vallina-Rodriguez. Bug fixes, improvements,... and privacy leaks: A longitudinal study of pii leaks across android app versions. In *NDSS*, 2018.

[61] J. Ren, A. Rao, M. Lindorfer, A. Legout, and D. Choffnes. Recon: Revealing and controlling pii leaks in mobile network traffic. In *MobiSys*, 2016.

[62] C. Rossow, C. J. Dietrich, C. Grier, C. Kreibich, V. Paxson, N. Pohlmann, H. Bos, and M. Van Steen. Prudent practices for designing malware experiments: Status quo and outlook. In *IEEE S&P*, 2012.

[63] I. Sanchez-Rola, M. Dell'Amico, D. Balzarotti, P.-A. Vervier, and L. Bilge. Journey to the center of the cookie ecosystem: Unraveling actors' roles and relationships. In *IEEE S&P*, 2021.

[64] B. P. Sarma, N. Li, C. Gates, R. Potharaju, C. Nita-Rotaru, and I. Molloy. Android permissions: a perspective combining risks and benefits. In *SACMAT*, 2012.

[65] S. Seneviratne, H. Kolamunna, and A. Seneviratne. A measurement study of tracking in paid mobile applications. In *WiSec*, 2015.

[66] Spamhaus. The Spamhaus Project. https://www.spamhaus.org/, 2019.

[67] O. Starov and N. Nikiforakis. Extended tracking powers: Measuring the privacy diffusion enabled by browser extensions. In *WWW*, 2017.

[68] R. Stevens, C. Gibler, J. Crussell, J. Erickson, and H. Chen. Investigating user privacy in android ad libraries. In *MoST*, 2012.

[69] G. Suarez-Tangil and G. Stringhini. Eight years of rider measurement in the android malware ecosystem. *IEEE Transactions on Dependable and Secure Computing*, 2020.

[70] G. Suarez-Tangil and G. Stringhini. Eight years of rider measurement in the android malware ecosystem: evolution and lessons learned. *IEEE Transactions on Dependable and Secure Computing*, 2020.

[71] M. Sun, T. Wei, and J. C. Lui. Taintart: A practical multi-level information-flow tracking system for android runtime. In *ACM CCS*, 2016.

[72] D. J. Tan, T.-W. Chua, V. L. Thing, et al. Securing android: a survey, taxonomy, and challenges. *ACM Computing Surveys (CSUR)*, 47(4), 2015.

[73] Team Cymru. BARS. https://www.team-cymru.com/bars-feed.html, 2019.

[74] R. Unni and R. Harmon. Perceived effectiveness of push vs. pull mobile location based advertising. *Journal of Interactive advertising*, 7(2), 2007.

[75] U.S. Department of Commerce. *The Privacy Shield Framework*, Accessed July 23, 2020.

[76] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz. (un) informed consent: Studying gdpr consent notices in the field. In *ACM CCS*, 2019.

[77] N. Vallina-Rodriguez, J. Shah, A. Finamore, Y. Grunenberger, K. Papagiannaki, H. Haddadi, and J. Crowcroft. Breaking for commercials: characterizing mobile advertising. In *IMC*, 2012.

[78] E. Vanrykel, G. Acar, M. Herrmann, and C. Diaz. Leaky birds: Exploiting mobile application traffic for surveillance. In *International Conference on Financial Cryptography and Data Security*, 2016.

[79] T. Vidas and N. Christin. Evading android runtime analysis via sandbox detection. In *ASIACCS*, 2014.

[80] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[81] W. G. Voss and K. A. Houser. Personal data and the gdpr: Providing a competitive advantage for us companies. *American Business Law Journal*, 56(2), 2019.

[82] M. Weissbacher, E. Mariconti, G. Suarez-Tangil, G. Stringhini, W. Robertson, and E. Kirda. Ex-ray: Detection of history-leaking browser extensions. In *ACSAC*, 2017.

[83] M. Y. Wong and D. Lie. Intellidroid: A targeted input generator for the dynamic analysis of android malware. In *NDSS*, 2016.

[84] M. Xu, C. Song, Y. Ji, M.-W. Shih, K. Lu, C. Zheng, R. Duan, Y. Jang, B. Lee, C. Qian, et al. Toward engineering a secure android ecosystem: A survey of existing techniques. *ACM Computing Surveys (CSUR)*, 49(2), 2016.

[85] C. Yang, Z. Xu, G. Gu, V. Yegneswaran, and P. Porras. Droidminer: Automated mining and characterization of fine-grained malicious behaviors in android applications. In *ESORICS*, 2014.

[86] H. Ye, S. Cheng, L. Zhang, and F. Jiang. Droidfuzzer: Fuzzing the android apps with intent-filter tag. In *MoMM*, 2013.

[87] B. Zhao and W. Chen. Data protection as a fundamental right: The european general data protection regulation and its exterritorial application in china. *US-China Law Review*, 16(3), 2019.

[88] B. Zhao and G. Mifsud Bonnici. Protecting eu citizens' personal data in china: a reality or a fantasy? *International Journal of Law and Information Technology*, 24(2), 2016.

[89] W. Zhou, Y. Zhou, X. Jiang, and P. Ning. Detecting repackaged smartphone applications in third-party android marketplaces. In *CODASPY*, 2012.

[90] Y. Zhou and X. Jiang. Dissecting android malware: Characterization and evolution. In *IEEE S&P*, 2012.

[91] Y. Zhou, X. Zhang, X. Jiang, and V. W. Freeh. Taming information-stealing smartphone applications (on android). In *TRUST*, 2011.