Tracking the Trackers: Towards Understanding the Mobile Advertising and Tracking Ecosystem

Narseo Vallina-Rodriguez[1,2], Srikanth Sundaresan[3], Abbas Razaghpanah[4]
Rishab Nithyanand[4], Mark Allman[1], Christian Kreibich[1,5], Phillipa Gill[6]
[1]ICSI, [2]IMDEA Networks, [3]Samsara, [4]Stony Brook University, [5]Lastline, [6]UMass

ABSTRACT

Third-party services form an integral part of the mobile ecosystem: they allow app developers to add features such as performance analytics and social network integration, and to monetize their apps by enabling user tracking and targeted ad delivery. At present users, researchers, and regulators all have at best limited understanding of this third-party ecosystem. In this paper we seek to shrink this gap. Using data from users of our ICSI Haystack app we gain a rich view of the mobile ecosystem: we identify and characterize domains associated with mobile advertising and user tracking, thereby taking an important step towards greater transparency. We furthermore outline our steps towards a public catalog and census of analytics services, their behavior, their personal data collection processes, and their use across mobile apps.

1. INTRODUCTION

Mobile apps provide services to billions of users worldwide. These apps often rely on third-party providers for services that enhance user experience, such as crash and bug reporting and social network integration, but also for monetizing their app with user tracking and ad integration.

Third-party services typically collect information about the user to provide their service. They typically rely on granted app permissions to collect this information, some of which may be privacy-sensitive. While mobile platforms typically enable users to grant or disable permissions for each app, this model has several shortcomings. First, users usually remain unaware that by granting permissions to an app their information might be harvested by third-party services. Second, users are not informed of which apps share the same third-party services, rendering them unaware of the potentially rich data (spanning a super-set of permissions across apps) that the third-party services aggregate.

This lack of transparency means that the third-party service ecosystem remains fundamentally mysterious to users, researchers, and regulators—to the extent that we are not even fully aware of the identities of the major service providers. Current techniques to explore this ecosystem require arduous effort and produce only limited understanding. For instance, some techniques require manual supervision such as static analysis of app source code followed by manual assessment of embedded libraries. Other approaches such as network-based trace collection and analysis yield (i) less than desirable coverage due to on-the-network encryption and (ii) at-best a fuzzy understanding of the relationship between traffic flows and the apps that generate them due to the absence of access to device context.

In this work, we aim to transform our understanding of the third-party service ecosystem by studying, at scale, how user-installed apps communicate with it. We leverage the data provided by the ICSI Haystack, an on-device app that provides us with rich and deep insight into user traffic and device operation stemming from real user stimuli, to identify and characterize third-parties associated with advertising and tracking services (ATS) at the traffic-level.

The identification and characterization of third-party tracking services is a fundamental step towards building mechanisms to improve the transparency of mobile tracking and to develop methods to protect users from abusive practices. Our results additionally point to places where targeted analysis using more traditional techniques, *e.g.*, static and dynamic analysis in a dedicated testbed, will strengthen our understanding of the ecosystem.

2. RELATED WORK

The research community has used diverse techniques to identify advertising and tracking libraries on Android apps. A large corpus of research characterized the presence of ad networks across mobile apps by analyzing network traces [35, 24, 29, 31]. These methods rely on data available on the payload (*e.g.*, `User-Agent` field) to associate flows to apps. However, due to the increasing use of encryption on mobile apps, these methods may fail to accurately associate network flows to apps.

Static and dynamic analysis of apps have also had limited success in identifying the prevalence of advertising and tracking services. The work by Chen *et al.* [20] used dynamic analysis of Android apps to uncover pervasive leakages of sensitive data and to measure the penetration of libraries for advertising and analytics across apps. Other studies instead leveraged static analysis of app source code to identify 190 embedded tracking libraries [32].

Techniques relying on static and dynamic analysis fall short in terms of scalability and app coverage [30]—*i.e.*, they rely on Google Play crawlers to obtain the executable and cannot access pre-installed services. In fact, they may generate false positives as the presence of a library in an app's source-code does not necessarily imply that it actually gets invoked at runtime.

3. THE ICSI Haystack APP

ICSI Haystack is an Android app, available free via Google Play [25], that helps mobile users understand how their mobile apps handle their private information [30], including the sensitive data their mobile apps leak and with whom they share it. Haystack leverages Android's VPN permission to capture and analyze network traffic locally on the device, and in user space: it implements a simplified network stack via standard user-level sockets to act as a local middleware that transparently transmits packets between the app and the network interface.

Haystack offers a unique vantage point to understand the mobile ecosystem at scale with real user stimuli. By operating locally on the device, Haystack can correlate disparate and rich contextual information, such as app identifiers and process IDs, with flows; *e.g.*, it can match DNS queries to outgoing flows and accurately identify the process owning a given socket.

Haystack analyzes app traffic payload and searches for personal information that it retrieves from the device subject to Android's permissions. Moreover, with user consent, Haystack also performs TLS interception by implementing a local TLS proxy that injects forged certificates on the flows during TLS session establishment [33]. Examining user traffic—especially encrypted flows—raises ethical issues that we consider carefully. We provide further details about Haystack's design, goals and performance, in addition to a discussion of the privacy precautions and ethical standards Haystack employs, in our technical report [30]. Given we do not export payload or user identifiers to our database for analysis, our IRB views our efforts as a non-human subjects research; we analyze the behavior of software, not people.

4. CLASSIFYING THIRD-PARTY SERVICES

This section presents our method for identifying and classifying third-party advertising and tracking services (ATS). We leverage the data provided by 690 Haystack users, summarized in Table 1. It includes 1798 K flows generated by 1,732 apps. We exclude mobile browsers from our analysis to avoid polluting our dataset with web trackers.

4.1 Identifying third-party services

We identify third-party services by analyzing how mobile apps interact with online services. We create a graph with two types of nodes: domains (identified by their DNS FQDN) and apps (identified by their Google Play ID) [1]. We create an edge between a pair of nodes if we observe a flow between them. We simplify domains to their second level using the Mozilla public suffix library [15]. Figure 1 shows an example interaction between six apps and two domains. Using the above graph, we label the second-level domains with a degree greater or equal to two as potential third-party services.

Users	Flows	Apps	Domains	Second-level domains
690	1798 K	1,732	12,206	4,678

Table 1: Summary and scale of our user study.

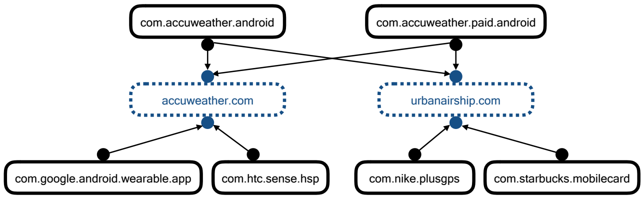

Figure 1: Communication between 6 mobile apps (in black) and two online services (in blue): accuweather.com and urbanairship.com.

This approach may result in false positives caused by first-party domains shared between apps developed by the same company or app developer. We illustrate an example of a false positive in Figure 1. Here, two Accuweather apps—one free, one paid—communicate with `accuweather.com` which could be labeled (incorrectly) as a third-party service when reached by those apps but not when used by other weather apps. We avoid such errors by matching tokens found in the app package name (*e.g.*, `com.accuweather.android`) with the domain names (*e.g.*, `accuweather.com`). If a domain and an app have matching tokens[2] we mark the domain as a first-party service. By applying these heuristics, we identify 446 domains as third-party services.

4.2 Identifying ATS domains

The 446 third-party domains that we identify range from CDNs and news sites to advertising and tracking services. In our work, we are specifically interested in the latter category; those that provide advertising and tracking services to app developers.

Accurately classifying the services provided by each domain proves challenging, as demonstrated by the fact that even popular commercial domain classification systems do not completely classify all identified third-party domains. Consider OpenDNS [16], which features a domain classification service maintained by a user community. It does not

[1] We differentiate between free and paid versions of the same app.

[2] We do not consider frequent tokens such as "com", "android", "free" or "paid".

Category	Keyword Sample
Ad-network	“ads”, “interstitial”, “advertising”, “ppi”
Analytics	“analytics”, “intelligence”, “bug report”
User Engagement	“push notification”, “crm”, “a/b test”

Table 2: Sample of reference keywords used by the service classifier.

contain records for 213 of the 446 third-party domains. Even when such systems provide a classification, it often remains vague and uninformative. For example, McAffee’s URL categorization service [27] classifies Crashlytics [7]—a popular crash-reporting analytics service—simply as “software”. While manually curated ATS-specific lists such as the ones provided by AdBlockPlus [21] and hpHost [28] provide better accuracy than general-purpose services, our results show that they are primarily web-centric and often miss mobile-specific ATS domains.

In order to overcome the incompleteness and inaccuracy of current domain classification systems, we propose a new classifier that extends the insight provided by commercial domain categorization systems with data gathered from crawling the domains to be classified. The classifier identifies three types of ATS (ad networks, analytics/tracking services and services to promote user engagement) by comparing the keywords present on their landing page with a reference set partially listed in Table 2. We pre-populated the reference set by crawling the websites of well-known ad networks and analytics services (*e.g.*, Google’s AdMob, Google Analytics, comScore and Yahoo’s Flurry).

In particular, the classifier follows two steps: First, it uses the McAffee and OpenDNS URL categorization services to identify and remove well-known non-ATS domains such as news sites, email services, and CDNs which are also absent from the manually curated AdBlockPlus and hpHost ATS lists. Second, we use our web crawler to analyze the content of the web pages of the remaining domains and the description provided by top search results from “<domain>+about” queries on the DuckDuckGo search engine. Our crawler also checks domains categorized in ambiguous categories such as Software, Internet, and Business services by McAffee and OpenDNS. Finally, the crawler analyzes and compares the keywords present on the landing pages of the domains (when available) with our reference set. This allows us to infer the services a domain offers.

4.3 Results

Our classifier identifies 280 second-level domains associated with ATS activity. Table 3 breaks down the 280 services identified per subcategory. Many of the ATS domains (80 %) cannot be uniquely categorized into a single category. This is the case of services like Flurry [10] and Localytics [14] that offer both analytics and ad services.

Of the 280, only 61 and 205 were reported as ATS services by the manually curated AdBlockPlus and hpHost ATS lists, respectively. All of the 61 domains listed by AdBlockPlus are also included in hpHost list, therefore our classification method reports 75 previously unreported ATS-related third-party services.

Category ($N = 446$)	#	%	Example
Non-ATS Domains	166	37	
ATS Domains	280	63	
Ad Network	177	40	mathtag.com
Analytics	153	34	crashlytics.com
User Engagement	77	17	pushwoosh.com
ATS (ABP)	61	14	baidu.com
ATS (hpHosts)	205	46	ubermedia.com

Table 3: Service classification for all domains identified as third party services. A third-party service can fall in multiple ATS categories.

In order to verify the correctness of our classifier we manually inspect the 75 new domains classified as ATS. We find that 58 domains were correctly classified as ATS, while 17 are false positives. Our results show that third-party domains such as `measurementapi.com` (the Google Play tracker) and Facebook’s Graph API [23] were correctly labeled by our classifier and absent from hpHosts ATS list. We speculate that this is a result of the web-specific focus of these manually curated lists and the multi-purpose nature of modern trackers such as Facebook’s Graph API which state-of-the-art ad-blockers cannot block at the domain level. On the other hand, the 17 false positives reported by our method include Google API subdomains (`fonts.googleapis.com`), A-GPS services (`izatcloud.net`) [34], the AVG anti-virus service (`avg-hrd.appspot.com`), and domains associated with IoT vendors (*e.g.*, `netatmo.net`) due to the presence of relevant keywords in their landing pages. In our ongoing research efforts we are exploring new methods to improve the accuracy of our classifier.

5. ATS PREVALENCE IN MOBILE APPS

Figure 2 shows the distribution of the number of ATS services prevalent in each app. We find that 60% of the apps monitored by Haystack connect to at least one ATS domain and 20% of the apps use at least 5 ATS services. The analysis reveals that users of news and social media apps are exposed to the largest number of ATS services (Facebook: 106, Twitter: 65) due to web trackers embedded in content shared via these platforms. More alarmingly, we find that popular games typically connect to a large number of ATS services. Given the popularity of general-audience apps—according to their ESRB rating [8]—among children, it remains unclear if they violate the FTC’s Children’s Online Privacy Protection Act (COPPA) [6] which requires app developers to obtain parental consent before collecting children’s sensitive information and sharing it with third-party services.

Figure 3 shows the top 25 third-party ATS domains by the percentage of apps actively reaching them. We find

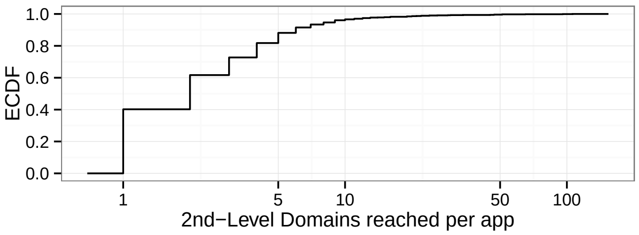

Figure 2: Empirical CDF of the number of ATS domains per mobile app.

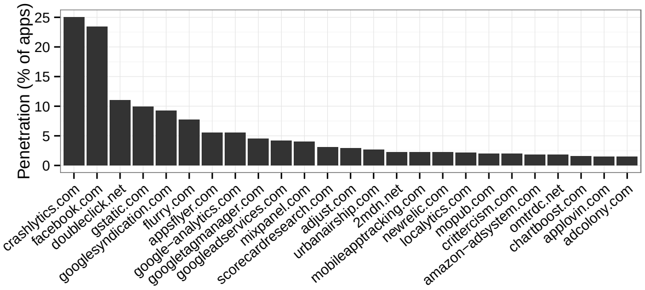

Figure 3: ATS domains ranked by the percentage of apps actively using them.

ATS Domain	ABP	hpHosts	#Apps	#Sites
crashlytics.com	False	False	434	0
facebook.com	False	True	406	623
doubleclick.net	True	True	190	621
gstatic.com	False	True	172	509
googlesyndication.com	False	True	160	441
flurry.com	True	True	133	0
appsflyer.com	False	True	95	9
google-analytics.com	True	True	95	664
googletagmanager.com	True	True	78	200
googleadservices.com	True	True	72	470

Table 4: Top 10 ATS domains (sorted by app penetration) with their presence on manually curated ATS lists and penetration in the Alexa Top 1000 Websites.

that over 20% of the apps monitored by Haystack connect to the Crashlytics [7] analytics service and the Facebook Graph API [23]. Crashlytics's SDK offers app developers a wide range of services beyond crash reporting, including app testing and real-time analytics. The Facebook Graph API—which also includes `graph.instagram.com`—is a unified and comprehensive service that offers app developers the opportunity to integrate their app on Facebook's social network and analytic services as well as cross-platform ad delivery. An interesting feature of Facebook's Graph API is its resilience to blocking by conventional mobile ad-blocking techniques by an in-path observer: its traffic runs entirely over TLS and re-uses non-tracking domains of other Facebook apps. The browser context is a different scenario. It thus becomes necessary to identify the process generating the flow to identify its use as a third-party or perform TLS interception to inspect its purpose on the URL. If not done carefully, non-tracking Facebook services can also be disrupted.

Over 10% of the apps we analyze utilize DoubleClick ad service. Other popular mobile ad-networks are provided by Amazon, AOL (Millenial Media) and comScore. Some ATSs specialize in assisting mobile game monetization, as in the case of Appsflyer [3] and Applifier by Unity3d [2]. App promotion services like Chartboost [5], Liftoff [13] and TapJoy [17] are a new type of ad-network specializing in promoting other apps via advertising. They implement a PPI (pay-per-install) model that allows app developers to monetize their apps by advertising other apps participating in the network. Promotion services aim to increase app audiences and, therefore, the number of installs on Google Play. The number of apps using promotion services is still small compared to traditional ad-networks and analytic services.

The "User Engagement" category groups services offering a broad range of features to app developers: push notifications mechanisms [19], in-app messages and surveys to increase user loyalty and obtain user feedback. UrbanAirship [18], jPush [12] and Apptentive [4] are among the most popular ones. Finally, Gigya [11] allows app developers to collect and manage customer identities while collecting social, behavioral, interest and transactional data from them.

5.1 Cross-platform tracking

Cross-platform ATS services have the ability to collect richer behavioral and contextual information about users. This poses a higher privacy risk than single platform trackers. In order to understand how common cross-platform tracking is, we also measure ATS presence on a non-mobile platform: the Web.

In particular, we measure how many mobile ATS domains are also present in the Alexa Top 1000 websites. Table 4 shows the ten most popular mobile ATS services and the number of Alexa Top 1000 websites which use their services. Across all the ATS domains identified in our analysis, we find that 68.5% are cross-platform and operate on at least one website in the Alexa Top 1000. We find that two of the most popular mobile ATS services—Crashlytics and Flurry— have no presence in the Web. However, Facebook, DoubleClick, and Google Analytics are present on over 60% of all the Alexa Top 1000 websites. Additionally, Table 4 shows that manually curated Web-specific ATS lists fail to identify mobile-only ATS domains such as Crashlytics and more unpopular services like Adjust and Urbanairship.

5.2 Traffic Overhead of ATS services

Having the ability to identify and label ATS domains allows us to estimate the data volume—which also translates to battery costs [35]—of mobile tracking. Figure 4 shows the distribution of the percentage of app traffic flowing to ATS third-parties. We limit our analysis to the 200 most data-hungry apps.

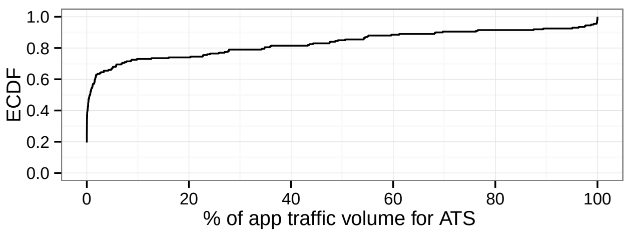

Figure 4: Empirical CDF of the percentage of traffic going to ATS domains per mobile app.

On average, 17% of app traffic is associated with ATS services. If we inspect in detail the distribution, we can see that 70% of the analyzed apps dedicate at least 10% of their traffic to tracking and advertising activities, while more than 7% of mobile apps have at least 90% of their traffic associated with ATS activities. If it were not for ATS-related activities, many mobile apps would operate mostly offline. However, the results may vary depending on how users interact with their apps and the nature of the service they provide as for data-hungry apps like audio/video streaming ones.

6. FUTURE WORK

Dynamic analysis of mobile apps: Thanks to its ability to capture traffic under real user and network stimuli, Haystack has effectively revealed interesting interactions between mobile apps and third-party services at the network level. However, as it focuses primarily on network traffic, Haystack does not allow us to analyze in depth how apps and libraries access sensitive resources during runtime [22]. Moreover, Haystack fails to inspect the payload of network flows when apps employ techniques against TLS interception [33, 30]. To overcome these limitations, we plan to deploy a purpose-built testbed to automatize the acquisition of comprehensive traces both at the network- and system-level.

Privacy leaks: Most apps and third-party ATS services upload sensitive user information using HTTPS. However, we have found instances of highly sensitive information (ranging from unique identifiers like the IMEI to WiFi SSID) being uploaded by popular apps to 11 ATS services in the clear. Apps actively uploading user metadata without encryption expose mobile users to in-path profiling and surveillance. Additionally, we have also identified app developers tracking users without their consent by gathering unique identifiers like the device serial number and MAC address which are not protected by any Android permission[3]. We will investigate and report techniques app developers employ to profile and track mobile users and instances of severe privacy leaks.

ATS detection accuracy: Our current method (Section 4.1) ignores unpopular domains accessed solely by a single app which translates into false negatives. In our future efforts, we will investigate domains present in the long-tail to identify instances of third-party tracking activity. We also plan to explore more advanced text-mining techniques to compile a more comprehensive set of keywords hoping to improve the scalability and accuracy of our detection method.

Cross-platform tracking: A significant number of ATS services offer cross-platform support (Section 5). This feature gives ATS services the ability to gather richer behavioral and personal data from users, we plan to investigate how those services aggregate, link and leverage personal information from different platforms to build accurate user profiles and for advertising purposes.

Contextualizing privacy leaks and tracking activity to regulatory jurisdictions: App developers and ATS domains must comply with a diverse set of rules enforced by regulatory jurisdictions. However, whether apps correctly comply with them in the wild is unclear. For example, the European General Data Protection Regulation controls how personal data is exported outside the EU [9]. Yet it remains unclear which organizations are behind each ATS domain and where they reside geographically. Another interesting case is the FTC Children's Online Privacy Protection Act (COPPA) which aims to protect the privacy of minors when using commercial websites and mobile apps [6]. According to COPPA rules mobile apps can only collect childrens' personal information such as unique identifiers (*e.g.*, IMEI), telephone number or geo-location with parental consent. We are working to develop methods that would allow us to contextualize the results of our app's behavioral analysis to each regulatory jurisdiction.

7. CONCLUSIONS

In this paper we presented our ongoing research efforts to illuminate the mobile ecosystem. Our first step in this endeavor was identifying the organizations responsible for user tracking and how mobile apps interact with them by leveraging the data provided by the ICSI Haystack tool. To that extent, we implemented a classifier which has allowed us to identify 58 domains that remained unreported by well-known tracking and advertising domain lists like AdBlock's Easylist and hpHost's ATS list. The results of our analysis are incorporated to tools and services to promote mobile transparency and develop techniques to protect mobile user's privacy like the ICSI Haystack Panopticon [26] and the ICSI Haystack Android app itself.

Acknowledgments:

This project is partially funded by the Data Transparency Lab Grants (2016) and the NSF grant CNS-1564329. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding bodies.

[3] The app developer only needs to invoke and parse the information provided by the undocumented `getprop` command [1].

8. REFERENCES

[1] ADB Shell. Getprop. http://adbshell.com/commands/adb-shell-getprop.

[2] applifier (Unity3D). http://applifier.com/.

[3] Appsflyer. https://www.appsflyer.com/.

[4] Apptentive. http://www.apptentive.com/.

[5] Chartboost. http://www.chartboost.com/.

[6] Children's Online Privacy Protection Rule. https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.

[7] Crashlytics. http://www.crashlytics.com.

[8] Entertainment Software Rating Board. http://www.esrb.org/ratings/ratings_guide.aspx.

[9] EU General Data Protection Regulation. http://data.consilium.europa.eu/doc/document/ST-9565-2015-INIT/en/pdf.

[10] Flurry. Yahoo Mobile Developer Suite. http://www.flurry.com/.

[11] Gigya. The leader in customer identity management. http://www.gigya.com/.

[12] jPush. https://www.jiguang.cn/.

[13] Liftoff. http://liftoff.io/.

[14] Localytics. http://www.localytics.com/.

[15] Mozilla public suffix library. https://pypi.python.org/pypi/publicsuffix/.

[16] OpenDNS Domain Tagging. https://domain.opendns.com.

[17] TapJoy PPI. https://home.tapjoy.com/developers/acquire/.

[18] Urban Airship. https://www.urbanairship.com.

[19] A. Aucinas, N. Vallina-Rodriguez, Y. Grunenberger, V. Erramilli, K. Papagiannaki, J. Crowcroft, and D. Wetherall. Staying online while mobile: The hidden costs. In *ACM CoNEXT*, 2013.

[20] T. Chen, I. Ullah, M. A. Kaafar, and R. Boreli. Information leakage through mobile analytics services. In *ACM HotMobile*, 2014.

[21] EasyList, AdBlockPlus. https://easylist.to/.

[22] W. Enck, P. Gilbert, B. Chun, L. Cox, J. Jung, P. McDaniel, and A. Sheth. TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones. In *USENIX OSDI*, 2010.

[23] Facebook for Developers. The Graph API. https://developers.facebook.com/docs/graph-api.

[24] P. Gill, V. Erramilli, A. Chaintreau, B. Krishnamurthy, K. Papagiannaki, and P. Rodriguez. Follow the money: Understanding economics of online aggregation and advertising. In *ACM IMC*, 2013.

[25] Google Play. Icsi haystack. https://play.google.com/store/apps/details?id=edu.berkeley.icsi.haystack&hl=en.

[26] ICSI Haystack Panopticon. https://www.haystack.mobi/panopticon/.

[27] Intel Security/McAfee. Customer URL Ticketing System. http://www.trustedsource.org/.

[28] MalwareBytes. hpHosts. http://hosts-file.net/.

[29] A. Rao, J. Sherry, A. Legout, A. Krishnamurthy, W. Dabbous, and D. Choffnes. Meddle: Middleboxes for Increased Transparency and Control of Mobile Traffic. In *ACM CoNEXT Student Workshop*, 2012.

[30] A. Razaghpanah, N. Vallina-Rodriguez, S. Sundaresan, C. Kreibich, P. Gill, M. Allman, and V. Paxson. Haystack: In Situ Mobile Traffic Analysis in User Space. *ArXiv e-prints*, 2015.

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

[32] S. Seneviratne, A. Seneviratne, P. Mohapatra, and A. Mahanti. Your installed apps reveal your gender and more! *ACM SIGMOBILE Mobile Computing and Communications Review*, 2015.

[33] N. Vallina-Rodriguez, J. Amann, C. Kreibich, N. Weaver, and V. Paxson. A Tangled Mass: The Android Root Certificate Stores. In *ACM CoNEXT*, 2014.

[34] N. Vallina-Rodriguez, J. Crowcroft, A. Finamore, and K. Grunenberger, Y.and Papagiannaki. When Assistance Becomes Dependence: Characterizing the Costs and Inefficiencies of A-GPS. *ACM SIGMOBILE Mobile Computing and Communications Review*, 2013.

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