

Using Passive Measurements to Demystify Online Trackers

Original

Using Passive Measurements to Demystify Online Trackers / Metwalley, Hassan; Traverso, Stefano; Mellia, Marco. - In: COMPUTER. - ISSN 0018-9162. - STAMPA. - 49:3(2016), pp. 50-55. [10.1109/MC.2016.74]

Availability:

This version is available at: 11583/2637929 since: 2016-03-20T16:01:04Z

Publisher:

IEEE - INST ELECTRICAL ELECTRONICS ENGINEERS INC

Published

DOI:10.1109/MC.2016.74

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

Online Trackers Demystified from Passive Measurements

Hassan Metwalley, Stefano Traverso, Marco Mellia
Politecnico di Torino, Italy – {firstname.lastname}@polito.it

Abstract—While on the Internet, individuals encounter invisible services that collect personal information, also known as third-party online trackers. Linked to advertisement, social sharing, and analytic services in general, hundreds of companies de facto track and build profiles of people. In this work, we present a measurement study to understand the extensiveness of this practice. We use passive measurements, naturally factoring users in the picture. In our measurements, we count more than 800 active trackers, of which 100 are regularly contacted by more than 40% of active users. The pervasiveness of trackers across the web is high, with websites that host hundreds of them, attracted by the chance to monetize visits. Conversely, privacy enhancing plugins are actually installed by few users (12.5%), and they mostly fail to protect people privacy. The resulting picture calls for a debate around privacy in the Internet, and for possible initiatives to regulate and control these practices.

I. INTRODUCTION

Internet is the revolution that changed our life, allowing us to stay informed, buy goods, enjoy shows, play games, keep in touch with friends, and freely express our opinions. Smartphones and tablets let people access to information from anywhere at anytime. Companies have consistently increased their investments in the Internet, where they leverage the web to attract customers, stay in contact with them and offer products.

Unsurprisingly, the web advertisement market has been growing consistently, overcoming revenues from TV broadcast since 2005 [1]. This is easily justified by the fact that online advertisement – ads for short – provides companies the capability to design campaigns tailored for very specific groups of users, and based on the knowledge about their interests and taste. However, the collection of such knowledge has rapidly built a new business. The modern web is populated by hundreds of online services – usually third-party –, which track users during everyday online activity, and use the information they collect to create per-user profiles. Later, these profiles are made available to advertisers to build ads campaigns, or to offer tailored suggestions to, e.g., recommend goods to buy, or a movie to watch.

To shadow users during browsing activity, online tracking services (trackers for short) identify a user leveraging different techniques, e.g., by using the IP address, storing a cookie, injecting javascript code in webpages, or using fingerprinting techniques that uniquely identify a user's browser [2], [3], [4]. Following then users across different websites and along time, profiles are easily built. The tracker business model varies greatly. Some directly manage ads. Some provides information to website owners and designers. Some act as data brokers

by selling personal profiles to third parties. The full list of companies that build their business around this information includes thousands of companies, the majority of which are mostly unknown, and whose business is unclear.

On the one hand, the mechanisms associated to online users' profiling can be beneficial for both companies and consumers. On the other hand, they raise many privacy concerns. Ultimately, the consciousness of people about their privacy being violated in the Internet is growing day by day. Regulators are also becoming more active. In Europe, for instance, the ePrivacy directive mandates prior consent to inform users that the website uses third-party elements [5].

Our goal in this work is to quantify the pervasiveness and extensiveness of online tracking. We leverage passive measurements, which have the major advantage to naturally factor the users into the picture. We address questions as how invasive are trackers? How many trackers does an Internet user face during her activity? How different is the picture from past years? Are browser plugins effective in protecting users?

The research community studied the pervasiveness of trackers, typically leveraging active measurement campaigns [6], [7], [8]. Our study complements the body of work available in the literature that addresses the problem of understanding how pervasive tracking services are. The most remarkable examples are [9], which analyzes third-party tracking services based on browser extensions, and [8], which builds on the crawling of the top websites in Alexa rank for different countries, and measure the per-country pervasiveness of third-party trackers. Our previous work, [10], provides preliminary results on how extensive tracking is. In this work, we revisit our results, providing an updated and more comprehensive picture of the phenomenon. We rely on a longitudinal dataset composed by (anonymized) traces we collected by passively observing users that access the web from home. We compare data collected in 2013, 2014 and 2015, from which we quantify the amount of traffic, penetration, and pervasiveness of services that either are tracking systems, or ads providers.

Results confirm that online tracking is ubiquitous. We count more than 800 trackers that are active in the country where the traces have been collected. About 100 of them are regularly contacted by more than 40% of users, with the most pervasive ones that observe 98.5% of online people. We observe websites that embed more than 100 third-party services, attracted by the chance to monetize visits. Interestingly, very unpopular services also participate in this rush, thus increasing the probability for a user to be identified in the mass. Our measurements highlight another phenomenon: the increasing

adoption of encryption (via HTTPS) as the standard mean to exchange data. This exacerbates the tension on the need to protect users' privacy. In fact, this mines the possibility to develop in-network solutions to limit trackers. For instance, this results critical for companies that would like to control which information their employees exchange on the Internet.

We also resume the list of available counter measures users may adopt to oppose tracking services. We focus on privacy-enhancer plugins, and, thanks to the perspective offered by our passive approach, we can quantify their popularity and effectiveness. Surprisingly, we find that those are used only by 12.5% of users, and, more critical, their efficacy is limited, mostly due to users accessing the Internet with multiple devices, with only few of these eventually protected by those plugins.

We hope the picture we draw can contribute to increase the sensibility of people, researchers and regulators towards privacy in the Internet, and to stimulate a debate around these topics.

II. METHODOLOGY

In this work, we focus on three datasets that we built following the procedure explained in our previous work [10]. Each dataset refers to three working days (Tuesday to Thursday) of July 2013 (09-11), 2014 (08-10) and 2015 (07-09), respectively. Each dataset aggregates the traffic of 9,500 active users, who contact in total 53,000 distinct service hostnames per day. For this study, we analyze almost 260 million TCP flow summaries. Despite its finitude and locality, this dataset allows us to obtain results comparable with those presented in other studies [9], [11], which build on datasets older and smaller than ours and collected in different scenarios.

A. Identifying Active Users and Number of Connected Devices

The client IP address the probe sees refers to the access gateway (ADSL/FTTH routers) customers are given by the ISP. As such, the IP address is an identifier of the household, in which several actual devices and applications may connect to the Internet via the NAT provided by the access gateway, using WiFi or Ethernet LANs. In some cases, traffic may be generated by connected devices without any actual users. For instance, the access gateway acts also as VoIP gateway, thus we expect some households to appear as "active" (the IP address generates traffic) even if no user is present (we observe VoIP data only). To filter these outliers, we only focus on HTTP and HTTPS traffic, and consider as active those households for which some significant web browsing activity is detected: at least one HTTPS flow, and at least 300 HTTP or HTTPS total flows are present in the three days. This filters out those sources of traffic we are not interested in (e.g., smart TVs, VoIP gateways, or pure P2P clients), or users that generate very little web traffic.

To detect the presence of multiple devices that are hidden behind the home gateway NAT, we leverage the User-Agents Python module. By parsing the user-agent field in HTTP requests, it identifies device, operating system, version, and browser, e.g., Google Chrome on OS X 10.10, on Samsung

Galaxy S5. The number of unique user-agents per household is an estimate of the number of multiple devices or applications that were in use in a household.

We leverage this information to get an estimate of the number of different devices connecting the Internet through the same household: We compute the distribution of unique user-agents seen for a given active household, considering only those associated to actual browsers, for PC and mobile terminals. We find that between 2 to 10 different devices are present in about 70% of households, and only in 2% of households the number of devices is above 20. Manually checking them, we observe that smartphones and tablets are very popular, with sometimes multiple browsers being used on the same device. In few cases we see more than 20 user-agents. A manual check shows the presence of suspicious behaviour with lots of HTTP requests toward few advertisement servers, with a rotating set of legitimate browser user-agents. We suspect this to be related to some device being infected by a malware involved in click fraud activity, i.e., a malicious software artificially generating clicks on ads servers by forging user-agents. We removed these cases before performing the measurements presented in the following.

B. Identifying Online Tracking Services

We build a list of third-party tracking services (trackers) by merging together lists we obtain from different sources. We extract tracking services from the Ghostery plugin, and augment it with a list we obtained from the developers of Abine. The latter includes hostnames referring to trackers specifically tailored to track mobile clients. Finally, we include also some trackers we identify using the procedure presented in [12]. To simplify the matching, and to make it more general, we extract the second-level domain name. For instance from *cnt2.acmetrackscopyou.com* and *srv1.acmetrackscopyou.com* we consider *acmetrackscopyou* only. This improves the accuracy of the matching since some trackers use multiple names, some of which were not present in the Ghostery lists, but which match the second-level domain name.

The final list consists of 2450 distinct trackers, and includes only services that we classify as services collecting users' information. These include tracking services that profile users explicitly (e.g., *scorecardresearch*), or that track users when on a website (e.g., *google-analytics*), or ads servers (e.g., *adnxs*). We do not consider social network buttons and plugins.

In the remainder of the paper, we rely on this list to pinpoint traffic that clients exchange with tracker servers. When analyzing the logs, we match the server hostname against the list to identify traffic to trackers.

III. RESULTS

A. Penetration of trackers

We start our analysis by measuring the penetration of each tracker that appears in our list. We consider the July 2015 trace. We compute this metric as the percentage of users that contact at least once a given tracker with respect to active users. Results are impressive: the top trackers— *doubleclick*, *googleanalytics*, and *googlesyndication* — are contacted by

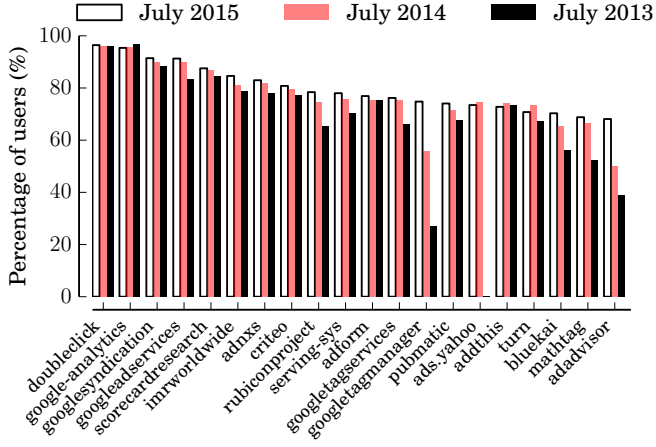


Fig. 1: Penetration of the 20 most pervasive trackers compared for different years.

98.8%, 98.7% and 97.4% of active users. Fig. 1 details the 20 most pervasive trackers. Notice how names of tracking services are mostly unknown even to expert internauts.

Overall, from the entire list of 2450 trackers, 800 are contacted by at least one user. Clearly, the set of contacted trackers would change based on the country the users are, but observing that several hundreds of these systems can be encountered during web browsing is in any case impressive.

In Fig. 1 we compare the share of users that contacted top trackers over 2013, 2014 and 2015. It shows marginal changes over the years, reflecting the fact that top trackers have saturated the market. Going down in the list, we see that most trackers show an increase in the penetration over years, with only few exceptions. For instance, *googletagmanager* has tripled its coverage in the three years. Some new players shows up as well, e.g., *ads.yahoo*. Notably, no top services went out of business (or disappeared).

At last, we measure to which extent the top trackers rely on encrypted channels, i.e., HTTPS, to collect information about the users. To this end, we measure how many TCP flows the users exchange with the trackers, and how many of these flows are HTTPS. We observe some trackers do use encryption to collect users' information. Some of them have double or even tripled the usage of HTTPS in 2015 with respect to 2014. Almost all the top 20 tracker has increased the usage of HTTPS over the last three years, on average, by more than 400%. For instance, *googleadvertisives*, *doubleclick* and *serving-sys*, to name a few, now encrypt 55%, 52% and 46% of their flows, respectively. In the 2013 dataset, they were found using encrypted flows in 18%, 10% and 12% of the cases. This is also mandated by the general increase of HTTPS-enabled websites that enforce HTTPS for all third-party content too.

We complement above observations by reporting the number of trackers contacted by users that are active in the 2015 dataset. Fig. 2 shows the results. Observe that few users contact more than 300 trackers over the three days of the trace. Those are very active users who spend lot of time on the web. More interestingly, almost 40% of the overall active population contacts at least 100 different trackers. Only 73

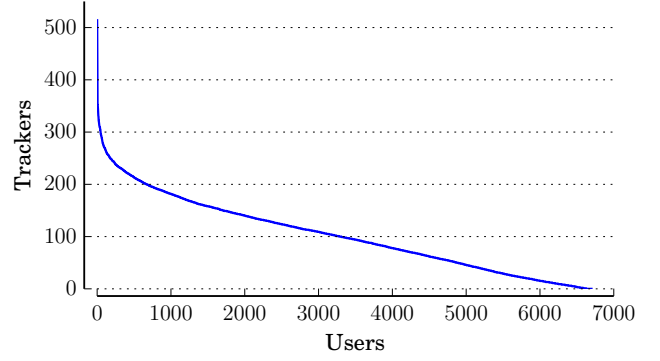


Fig. 2: Number of trackers contacted at least once from each user in 2015 trace.

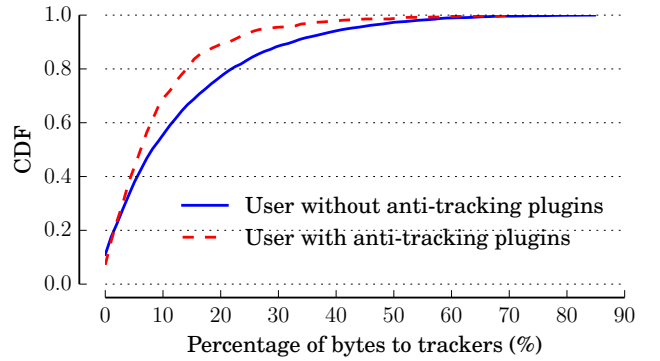


Fig. 3: Percentage of bytes sent to trackers for users who install anti-tracking plugins, and those who do not. Trace 2015.

active users (over 6699) never contact any tracker, i.e., a mere 1% of population is capable of escaping all trackers.

B. Popularity and efficacy of anti-tracking plugins

Several solutions promise to protect users from tracking services. We expose them in Sec IV. Here, we first evaluate how effective anti-tracking plugins such as, e.g., Adblock Plus, AdBlock, Blur, Ghostery and Web of Trust, might be at reducing the amount of information sent to trackers. We consider again the 2015 trace and split the population of users in two sets. The first set includes users who are seen running an anti-tracking plugin with at least one device. The second set includes users with no anti-tracking plugin (we discriminate users running an anti-tracking plugin by checking if they contact the correspondent service update server – see [10] for more details). 12.5% of the overall households fall in the first set.

Next, for each user in the two sets, we compute the fraction of data exchanged with tracking servers with respect to all data exchanged with all web services. Fig. 3 shows the CDF with respect to users, for the two sets. First, notice how for some users a significant fraction of the traffic they generate is exchanged with trackers. For instance, 50% of user without anti-tracking plugins exchange more than 10% of their data

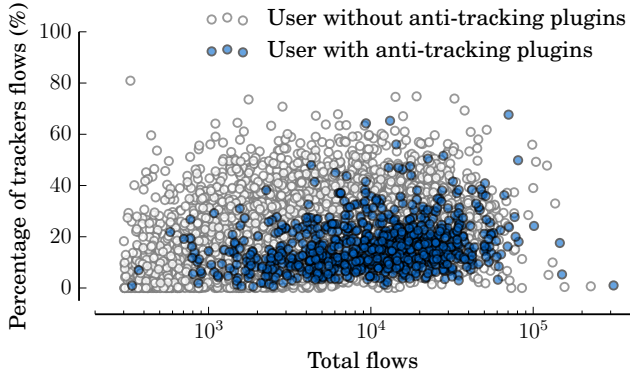


Fig. 4: Scatter plot reporting for each user (dot), the total number of generated flows, and the fraction of flows to trackers. White (red) dots refer to users with (without) anti-tracking plugins. 2015 trace.

with trackers. Second, the plot confirms the intuition that users installing an anti-tracking plugins do exchange less data with trackers. However, the difference between the two curves is smaller than expected. We further investigate this, and notice that most of anti-tracking plugins are available for PC browsers only (see Sec IV). This leaves mobile devices unprotected against trackers. Hence, even if a user can limit trackers when browsing the web from her PC, she has practically little means to block trackers when using her mobile device, a situation which is very popular.

We complement above finding by conducting a second experiment. We compute the percentage of flows exchanged with trackers over the total number of generated flows. Again, we perform the computation discriminating households with or without anti-tracking plugins. We plot the result as a scatter plot in Fig. 4. Each dot represents a user for whom the x axis reports the total number of flows she generates during the three days of measurement, while the y axis reports the percentage of flows to trackers. White (blue) dots refer to households with no (at least one) anti-tracking plugin installed on any device. Several observations hold. First, not surprisingly, we observe that the more the user is active, the more likely she contacts some trackers. Second, the users installing a plugin are also the most active ones. Third, the fraction of flows exchanged with trackers is large also for those users with anti-tracking plugins, as previously depicted in Fig. 3. More worryingly, some users, despite installing anti-tracking plugins, exchange more than 50% of flows with trackers. By manually inspecting, we observe that these are users using tablets and smartphones and browsing a large number of news portals, which typically embeds a large number of trackers too.

C. Pervasiveness of Trackers

Finally, we investigate the pervasiveness of trackers among different web services. We consider the trace of July 2015, and we use the HTTP summaries produced by the probe. From each URL in the trace where the *hostname* is a given (third-party) tracker, we check the *Referer* field to look at the (first-party) service embedding it. For simplicity, we consider only

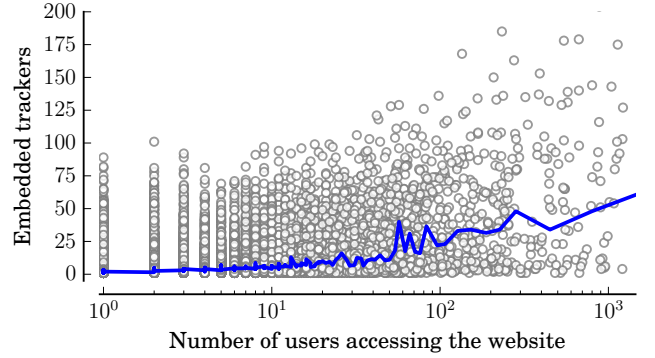


Fig. 5: Scatter plot of the number of users contacting a service, and the number of trackers embedded by the same service. Trace 2015.

the second level domain as the name of the service. We count more than 44500 services hosting trackers. That is one third of services hosts at least one tracker. For each of them, we count how many users contact them, i.e., the service popularity, and how many trackers they embed. In the scatter plot in Fig. 5, each dot represents a first-party service; the x-axis (in log scale) reports the number of distinct users accessing it; the y-axis reports the number of embedded trackers. The scenario is rather heterogeneous, with many services embedding several tens of trackers. We observe both unpopular services hosting many trackers—e.g., the services contacted by one or two users only, but hosting more than 80 trackers—and popular services hosting a few trackers—e.g., the rightmost bottom corner of the plot. The reason why we observe so many trackers being embedded in a single website is explained by a popular mechanism to serve ads in websites: the website owner offers a space for advertisement to a mediator company. In turn, each time the webpage is accessed, the mediator company runs an auction in background, selling the space to possible advertisers in real-time. The winner then embeds the ads into the page, thus resulting a third-party tracker. The more pages are visited, the more spaces are offered, and the larger are the number of auction competitors, which appear as distinct trackers that the website embeds [13].

In general, the number of trackers per service tends to increase with the popularity of the service. To better quantify this, the blue curve reports the average number of trackers for every subset of 100 services grouped by popularity. As it can be seen, the more popular the site is, the higher the average number of embedded trackers.

IV. EXISTING SOLUTIONS

The catalogue of countermeasures proposed in recent years against trackers is rich. We classify them based on several angles, and list them in Tab. I. We put particular emphasis on three aspects: i) their compatibility with mobile terminals, ii) their capability of monitoring the traffic generated by the users' device, and iii) their capability of handling the content carried in encrypted HTTP channels. This latter is particularly significant when the tracker uses domain names which look

| | Browser Plugins | | | | | On-device-based tools | | | In-network-based tools | | |
|------------------------|-----------------|----------------|---------------|----------------|----------------------|-----------------------|---------|----------|------------------------|-----------|---------|
| | Ghostery | Privacy Badger | Blur | Adblock Plus | iOS Content Blocking | AdGuard | PrivDog | AdFender | Privoxy | SafeSquid | OpenDNS |
| Block tracking | Yes | Yes | Yes | Opt-In | Possibly | Yes | Yes | Yes | No HTTPS | No HTTPS | No |
| Block ads | Yes | No | Yes | Acceptable ads | Possibly | Yes | Yes | Yes | No HTTPS | No HTTPS | Partly |
| Customizable | Yes | No | Premium | Yes | Possibly | Yes | No | Yes | Yes | No | Yes |
| Support Mobiles | Their browser | No | Their browser | Proxy | Only Safari | Proxy | No | No | No HTTPS | No HTTPS | Yes |
| Open-Source | No | Yes | No | Yes | No | No | No | No | Yes | No | No |

TABLE I: Comparison between existing solutions.

legit, but the actual content it delivers contains some piece of tracking code. This is the case, for instance, of Facebook and its social sharing buttons.

A. On-device solutions

A practice to hinder communications with trackers is installing some piece of software on the user's device. The most popular subgroup of solutions in this family are browser plugins, software to install inside the web browser that checks and filters the traffic generated uniquely inside the browser. The most notable examples of anti-tracking browser plugins are Ghostery and EFF's Privacy Badger. Other popular ones, e.g., Adblock Plus, do not explicitly target the problem of protecting personal information, but partially achieve this goal by blocking online advertisement services. However, this family of solutions has some notable limitations: First, they control the transactions established by the browser only, and have no visibility on the traffic generated by other applications on the device. Second, they are often not available for mobile devices, and when available, they are implemented as standalone browsers (e.g., Ghostery) or as content blocking policies that however apply to browsers only (e.g., iOS/Safari).

A second group of on-device solutions is composed by tools which work as local proxies, with AdGuard, PrivDog and AdFender being notable examples. This approach allows the tool to intercept all the HTTP transactions generated by (properly configured) applications, and not only those generated by the browser. However, they often lack visibility on encrypted transactions, and they are not available for mobile devices.

In general, only a few of on-device tools are open-source and offer the possibility to customize their functionalities.

B. In-network solutions

Apparatuses like firewalls, proxies (Privoxy, SafeSquid) and DNS resolvers (OpenDNS) are typically installed in the network demilitarized zones (DMZs) to process the traffic of all devices connected by the network. Those can be easily instrumented to block connections to trackers as well. However, since they either work at TCP/IP (firewalls) or DNS

level, they have no visibility on traffic when encryption is enabled. To overcome the problem one might rely on man-in-the-middle solutions, but these are very intrusive. For DNS based solutions the anti-tracking capabilities build on lists of domains to block thus with very poor granularity. Finally, being installed in the network and centralized, these solutions lack of scalability and the rules they run can hardly be customizable by the users.

In summary, we lack a comprehensive solution capable of limiting the connections headed to trackers. However, there exist proposals in the literature like [14], whose aim goes beyond the anti-tracking task, but which might be easily employed for this specific end.

V. CONCLUSION

We presented in this paper a passive characterization of online tracking in the wild. We leveraged a large dataset of traffic summaries we collected from an ISP to passively quantify the pervasiveness and the intrusiveness of online tracking practice in our online lives.

Our results show that trackers' intrusiveness is astonishing: the top 100 trackers collect information from 40% of the users on a regular basis, with some of these being able of tracking 98% of the Internauts and embedded into more than 70% of websites, including the most popular ones.

We also observed that trackers are increasingly embracing HTTPS to collect data. While this is possibly driven by the increase of HTTPS usage, it complicates the task of controlling and, possibly, limiting the information trackers can collect.

Our results show that the consciousness of the users about their activity being monitored by trackers is limited. Indeed, a small fraction of users rely on privacy-enhancer browser plugins as Ghostery. Moreover, we showed that the efficacy of these tools at blocking transactions to trackers is limited as these are not available for all devices users use to browse the web. Existing countermeasures are indeed lacking a comprehensive approach.

We believe that the information contained in this paper can contribute to increase the consciousness of people about the fragility of their privacy in modern web. Our findings may be of stimulus for regulators, researchers and practitioners to

design solutions to let the users take control on the information they exchange with the Internet.

REFERENCES

- [1] IAB internet advertising revenue report, 2013 full year results, http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2013.pdf.
- [2] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, "The Web Never Forgets: Persistent Tracking Mechanisms in the Wild," in *ACM SIGSAC*, 2014.
- [3] B. Krishnamurthy, K. Naryshkin, and C. E. Wills, "Privacy leakage vs. Protection measures: the growing disconnect," in *W2SP*, 2011.
- [4] T.-F. Yen, Y. Xie, F. Yu, R. P. Yu, and M. Abadi, "Host Fingerprinting and Tracking on the Web: Privacy and Security Implications," in *NDSS*, 2012.
- [5] ePrivacy, http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm.
- [6] B. Krishnamurthy and C. Wills, "Privacy Diffusion on the Web: A Longitudinal Perspective," in *WWW*, 2009.
- [7] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and Analyzing Online Display Ads," in *WWW*, 2014.
- [8] M. Falahrestegar, H. Haddadi, S. Uhlig, and R. Mortier, "The Rise of Panopticons: Examining Region-Specific Third-Party Web Tracking," in *TMA*, 2014.
- [9] C. Castelluccia, S. Grumbach, and L. Olejnik, "Data Harvesting 2.0: from the Visible to the Invisible Web," in *WEIS*, 2013.
- [10] H. Metwalley, S. Traverso, M. Mellia, S. Miskovic, and M. Baldi, "The Online Tracking Horde: A View from Passive Measurements," in *TMA*, 2014.
- [11] E. Pujol, O. Hohlfeld, and A. Feldmann, "Annoyed users: Ads and ad-block usage in the wild," in *ACM IMC*, 2015.
- [12] H. Metwalley, S. Traverso, and M. Mellia, "Unsupervised Detection of Web Trackers," in *IEEE Globecom*, 2015.
- [13] L. Olejnik, M.-D. Tran, and C. Castelluccia, "Selling off Privacy at Auction," in *ISOC NDSS*, 2014.
- [14] H. Metwalley, S. Traverso, M. Mellia, S. Miskovic, and M. Baldi, "CrowdSurf: Empowering Transparency in the Web," *ACM SIGCOMM Comput. Commun. Rev. "October 2015 Issue"*, vol. 45, no. 4, 2015.

Hassan Metwalley is a Ph.D. student of TNG group of Politecnico di Torino. In 2015 he has visited the NEC Laboratories (Heidelberg, Germany), to study web tracking techniques and targeted online advertising. His research interests include privacy-preserving systems and network measurements.

Stefano Traverso (M'12), Ph.D. His research interests include privacy-preserving systems, network measurements and content delivery networks. During his Ph.D. and Post-doc he has been visiting Telefonica I+D research center (Barcelona, Spain), NEC Laboratories (Heidelberg, Germany) and Alcatel-lucent Bell Labs (Paris, France). He is currently a Post-doc Fellow of the TNG group of Politecnico di Torino.

Marco Mellia (SM'08), Ph.D., research interest are in the design of energy efficient networks (green networks), in the area of traffic monitoring and analysis, and in cyber monitoring in general. He is the coordinator of the mPlane Integrated Project that focuses on building an Intelligent Measurement Plane for Future Network and Application Management. Marco Mellia has co-authored over 200 papers published in international journals and presented in leading international conferences.