AdPExT: Designing a Tool to Assess Information Gleaned from Browsers by Online Advertising Platforms

Joseph Waldu Woensdregt
Data and Systems Team
All Response Media
London, UK
Waldu@allresponsemedia.com

Haider M. al-Khateeb
Wolverhampton Cyber Research Institute (WCRI)
University of Wolverhampton
Wolverhampton, UK
H.Al-Khateeb@wlv.ac.uk

Gregory Epiphaniou
Wolverhampton Cyber Research Institute (WCRI)
University of Wolverhampton
Wolverhampton, UK
G.Epiphaniou@wlv.ac.uk

Hamid Jahankhani
Computer and Information Sciences
Northumbria University and QAHE
London, UK
Hamid.Jahankhani@northumbria.ac.uk

***Abstract*— The world of online advertising is directly dependent on data collection of the online browsing habits of individuals to enable effective advertisement targeting and retargeting. However, these data collection practices can cause leakage of private data belonging to website visitors (end-users) without their knowledge. The growing privacy concern of end-users is amplified by a lack of trust and understanding of what and how advertisement trackers are collecting and using their data. This paper presents an investigation to restore the trust or validate the concerns. We aim to facilitate the assessment of the actual end-user related data being collected by advertising platforms (APs) by means of a critical discussion but also the development of a new tool, AdPExT (Advertising Parameter Extraction Tool), which can be used to extract third-party parameter key-value pairs at an individual key-value level. Furthermore, we conduct a survey covering mostly United Kingdom-based frequent internet users to gather the perceived sensitivity sentiment for various representative tracking parameters. End-users have a definite concern with regards to advertisement tracking of sensitive data by global dominating platforms such as Facebook and Google.**

Keywords*— *AdPExT, online advertising, tool, privacy, trust, advertising platforms, GDPR, Sensitivity Perception

I. Introduction

A recent Gartner study found that marketing budgets in the United Kingdom and the United States will rise to 12% of company revenue in 2017 and this is largely due to website-based advertising even though there is a known increase in ad-blockers [1]. These budget increases are in part due to improved targeting by advertising platforms who build and sell visitor cookies in grouped bundles called "audiences" to enable effective retargeting and this practice is evolving exponentially with advanced data merging, cookie syncing and data analytics. This eco-system is dependent on web tracking data by third-parties and end-users that visit websites where advertisements are shown are concerned about their data being collected and whether they can trust the practices of the data collectors. This research investigates these practices in detail.

There are three main entities involved in the online advertisement process (apart from the end-user): the advertiser, publisher and advertising platform [2]. The *Advertiser*, also known as retailer, is the company that has a service or product to sell or promote (e.g. EasyJet) and they generally buy advertising space to generate interest and brand knowledge from website visitors (end-users) that could lead to a direct visit from a click on the advertisement or a future visit or purchase. The *Publisher* is the owner of the website where the advertisement is shown (e.g. the Daily Mail) and they are paid for showing the advertisement. The *Advertising Platform* (AP) connects publishers with advertisers (e.g. Google and AppNexus) and Estrada-Jiménez et al. (suggests that APs represent the heart of online advertising through merging user interests with the relevant advertisers [2], [3]. It is important to note that one publisher can have multiple APs tracking their site at the same time (e.g. the Daily Mail normally has more than sixty APs active on their website). The modern AP consists of various sub-entities (including ad networks and ad exchanges) that facilitate advanced targeting, data merging and advertising optimisations. Fig. 1 presents a diagram influenced by Estrada-Jiménez et al. (2017) study of the online advertising environment and how these entities interact [2].

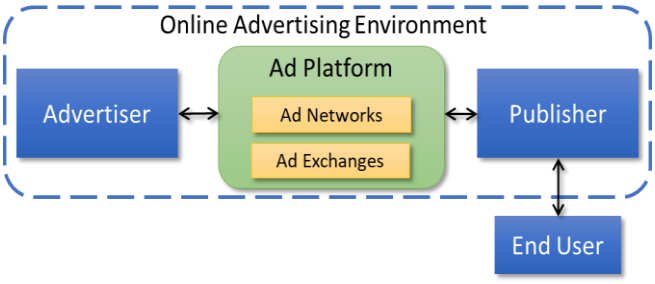

Fig. 1. Main Online Advertising Environment Entities

To track and target website visitors, APs drop a file (cookie) onto a visitor's computer and then build and update a user profile linked to the cookie that is used to target the visitor with relevant adverts. The user cookie and data

J. W. Woensdregt, H. M. al-Khateeb, G. Epiphaniou and H. Jahankhani, "AdPExT: Designing a Tool to Assess Information Gleaned from Browsers by Online Advertising Platforms", In *12th Annual International Conference on Global Security, Safety and Sustainability (ICGS3)*, London, UK: IEEE, 16 – 18 Jan 2019. DOI: 10.1109/ICGS3.2019.8688328. [online]. Available: https://doi.org/10.1109/ICGS3.2019.8688328.

transmission to and from the AP are managed through a JavaScript-based tracking code (pixel) or a transparent 1x1 image (image pixel) which posts the data through URL parameters during the HTTP request-response protocol transmission. This is demonstrated in Fig. 2, influenced by Puglisi, Rebollo-Monedero and Forne (2016) study) [4]. This tracking is referred to as *Third-party tracking* since it is activated by URL domains that belong to the APs (e.g. google.com) and not the publisher's local domain (e.g. dailymail.co.uk). *Parameters* can be noted within the third-party URL strings as everything that follows the question mark ("?") with each parameter key and value being separated by an equal sign ("=") and different parameter key-value pairs separated with an ampersand ("&"). For example, in the string presented in Fig. 3 the parameter key "param_key1" has the value "param_val1" while "param_key2" has the value "param_val2".

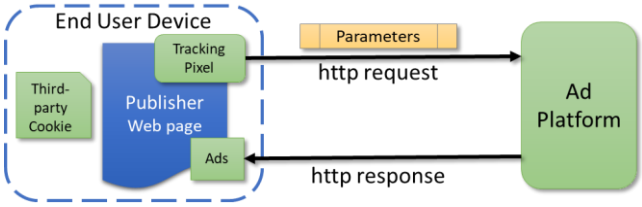

Fig. 2. Tracking parameter transfer via HTTP request and response process

```
https://sampleAPsubdomain.com/page.html?param_key1=param_val1&param_key2=param_val2
```

Fig. 3. Sample URL string with parameter keys and values

User profile based *Targeted Advertising* is effective and beneficial to the advertiser since it can increase revenue up to 2.68 times and it benefits the visitor with personalised adverts while enabling a high amount of free content all over the internet, but it can also pose a significant privacy and security risk to end-users [2], [3], [5], [6], [7], [8], [9]. Some of the main concerns over advertisement tracking is privacy disclosure during retargeting and data merging of third-party tracked data from various sites and sources (including ISPs) which can be used to build personally intrusive profiles of individuals [2], [3], [4], [5], [8], [10]. Increasingly IP addresses, browsers, e-mails and mobile devices contain or identify user location information (sometimes with GPS accuracy) which is collected by APs to amend and improve advert relevancy even though end-users are not explicitly aware of this [2], [3], [6], [8].

Browsers are making it easier for users to delete their cookies and some like Safari block third-party cookies (cookies from sites not directly visited) by default. To circumvent this occurrence trackers can use *Device Fingerprinting* (also known as *Stateless Tracking*) whereby the device and browser information (collected through the tracking parameters) is used to build a user profile across multiple sites and browsers when cookies have been deleted or blocked and in some cases the original cookies are recreated (respawned) on the user's computer after deletion [2], [6], [9], [11], [12]. These recreated cookies are sometimes referred to as "zombie cookies" [13].

Various *Ad-blocking* systems and techniques are available, and they can be effective for more advanced users, but they can also cause site breakage which can then cause users to disable them [9], [13]. The IAB reported that 45% of people state they are less likely to use ad-blockers if it does not affect their browsing directly, but ad-blocking still grew by 30% globally in 2016 with an estimated quarter of all internet users reported to be using these blockers by end 2016 [14], [15]. The online advertising industry's dependency on tracking data collection and processing to enable effective re-targeting means blocking poses a direct threat to the advertising ecosystem and blocking tools will continuously be opposed or bypassed by APs which creates a false sense of security to ad-blocking users [12], [15].

The increase in data mining techniques and processing capacity exposes the threat of unintentional privacy leakage. Publishers that transfer user data including address details and user identifiers between their own website pages via URL parameters can leak these metrics to third-parties that collect the URL strings, albeit unintentionally [13], [16]. This information can contribute to fingerprinting techniques and provide details to third-parties that can reveal sensitive information without the knowledge of end-users or even the direct knowledge of the publisher [6].

Trust is an important element when these users consider the release of their private information to third-parties and in addition to brand trust, the website privacy policies are expected to provide a clear picture with regards to the use of private and sensitive data, although this is not always the case with some privacy policies being vague, unnecessarily long or just full of legal wording that is difficult to comprehend. Previous surveys have found that end-users are not always aware that their browsing activity and sensitive information including their location is being collected by third-parties and when informed they are opposed to the practice and surprised that they were not better informed by the website they are visiting [6], [13], [17]. There is an immediate and increasing need for website visitors to have more clarity about what is being done with their tracked data and there is a lack of clear existing research to investigate the actual parameter collection practices and the variance from privacy policies [5].

The aim of this paper is to design a tool that can be used to assess the information gleaned from browsers when websites provide advertising space served by online advertising platforms (APs). This information can be used to ascertain privacy concerns and identify inaccurate privacy statements for a subset of APs. Therefore, this study creates and uses a new application to extract the various website visitor parameters (cookie ID, IP address, browser, etc.) that are collected by online APs when an online advert is shown. Our aim was addressed by achieving two main objectives related to the tools development and user perception of privacy issues related to the parameters collected by APs.

The first objective can be described as to design and implement a method to read website tracking parameters from online websites. During this stage of the work, design attention will be given to related tracking tools and how their techniques can be integrated and evolved when applicable. The expected output for objective one is a process flow diagram of the new method including research sources and proof of concept code. The second objective covers users' perceptions. This is more concerned with perceiving and

understanding which parameters collected by APs are considered more private by end-users compared to what they are more willing to share.

The research scope has been defined to control the quality and duration of the design and validation process. The design was created within a VirtualBox virtual machine within the United Kingdom with a focus on the use of the Mozilla Firefox browser for any testing purposes (Proof-of-Concept). Online tracking within mobile and tablet browsers and apps is outside the scope of our testing but we believe that our results can be generalised and extended to these other environments. Furthermore, we will survey end-users in response to objective two with a focus on the United Kingdom.

In the remaining part of this paper, we review the literature and related work in Section II, demonstrate the design of the new tool namely AdPExT supported by a link to its source code within Section III. The design and results of the sensitivity perception survey will be presented in Section IV. Finally, conclusions and future work are shared in Section V.

II. Literature Review

The privacy concern related to online tracking is well documented, but still contemporary, unresolved and expected to increase since the introduction of the EU GDPR legislation enforcement in May 2018 [20]. Specific relevant papers are noted in this section following a literature review related to third-party online tracking privacy data.

A. User Identification

To sustain the thriving advertising industry, advertising platforms (APs) track unique users across multiple sites and devices which allow them to build and update a user profile enabling them to effectively target the user with relevant advertising content. Various literature examines how cross-site user tracking is achieved online.

Puglisi, Rebollo-Monedero and Forne (2016) analyse the methods employed by APs to build user profiles across different sites and devices and investigate the accuracy of these profiles while considering the privacy risk and how adverts are amended depending on user profiles [4]. It is found that profiles are built quickly across only a few web page visits which then allow adverts to be adjusted accordingly and they suggest data obfuscation and user profile visualisation tools are required to preserve privacy and inform users [4]. The research result is somewhat vague on the accuracy of profiles built during testing but provides some proof of how swiftly such profiles can affect what users are exposed to.

Beck (2015) highlights the industry practice of online tracking across both the visible digital identity which users are aware of and the invisible digital identity which APs can build by combining data from various sites and user actions [5]. The research concludes that more education is required with regards to online tracking and privacy and that future research is needed to expose the actual parameters collected by APs [5]. This suggestion is directly in line with the aim of this research paper.

With a similar focus Acar et al. (2013) demonstrate FPDetective as their framework to detect and analyse web-based fingerprinting (tracking via device signatures) and analyse one million websites to find that fingerprinting is quite pervasive and counter fingerprinting tools like the Tor Browser and Firegloves from Firefox are ineffective [11]. They raise concern that web-users are in general unaware of (or do not understand) web-based fingerprinting and although some third-party privacy policies might mention fingerprinting, the websites implementing those third-party scripts rarely do [11]. Fingerprinting is not as trustworthy (for unique user identification) as user cookies since multiple computers can have the same fingerprint, and although FPDetective is a good tool to detect it, without an effective method to avoid fingerprinting the tool has limited progressive value.

Jain, Javed and Paxson (2016) developed a methodology to detect unofficial and usable user identifiers via network traffic analysis (mainly HTTPS request elements) and examines the traffic at the border of an enterprise network for fifteen days [6]. Their method identifies repeated strings in the network traffic (not contained in the cookie header information) which are then manually approved by an analyst. The research concludes that a fair amount of first-party as well as third-party identifiers are transmitted through various techniques including HTTP headers, URL parameters and non-HTTP messages [6]. These findings emphasise the lack of transparency from third-party trackers (knowingly as well as unknowingly) and further justifies the need for more detailed third-party tracker parameter investigations in relation with what is declared by trackers.

From a data aggregation perspective Rao, Schaub and Sadeh (2015) highlighted a concern related to data from various sources (including offline sources) being combined to build behavioural profiles of individuals [8]. The accessible cookie-based profiles of BlueKai, Google and Yahoo are found to sometimes include full names and addresses, albeit with poor accuracy and the data aggregation processes provide detailed profiles of individuals in contrast to claims of anonymity by the companies [8]. These research findings are in line with the concerns raised within the GDPR with regards to extensive privacy violations occurring when data from multiple sources are combined and this concern will increase with the improvements of behavioural profiles built through machine learning [21], [22].

B. Detecting New and Unsavoury Tracking Practices

With the ever-expanding internet and exponential increases in advertisers and advertisement practices, it is difficult to stay abreast of the latest tracking practices based purely on what APs reveal. There is a need to easily and automatically capture the tracking cookies and data to allow for in-depth data investigations and the following literature provides such solutions.

Purra and Carlsson (2017) released a measurement platform that captures web request and response headers and then evaluates third-party tracking by differentiating tracking types (including advertising), site popularity and reach, and the use of HTTP and HTTPS protocols by publishers [7]. The extensive site crawl across more than 130 million sites found that HTTPS usage by publishers correlates with an increase in trackers and that the more well-known APs are by far more widely used with Google being the most prolific tracker detected [7]. It is interesting to note that the implementation of HTTPS by publishers does not mean less

third-party tracking is in place, although at least the encrypted traffic is less susceptible to privacy leakage through sniffing.

Mayer and Mitchell (2012) presents the FourthParty third-party web-tracking data extract tool and discusses the techniques and risks related to third-party tracking [13]. Their tool works as a Firefox browser extension and stores results within an SQLite database [13]. This tool is one of the first efficient third-party tracking data extraction tools and the research discussion provides good detail of the various third-party tracking and blocking techniques in use, although it does not address any of the concerns with regards to actual data collection versus user perception and privacy policy statements.

By building on the FourthParty system Englehardt and Narayanan (2016) created the OpenWPM Python package and then scrapes the top one million websites (as per Alexa.com in January 2016) to analyse various tracking methods including cookie and device fingerprint-based tracking, browser-based privacy tools and their effect as well as cookie syncing techniques (third-parties sharing cookies) [16]. The research confirms the effectiveness of their OpenWPM platform, detects a large amount of cookie syncing (90% of the top fifty third-parties) and confirms the effectiveness of third-party cookie blocking by Firefox and Ghostery [16], [29]. This is important research that provides a platform for future research with the efficient OpenWPM package and delivers a considerable overview of various third-party tracking techniques and practices that are not widely known by the average internet user.

C. Ad-blockers

The increase in advertisement tracking awareness and privacy concerns by end-users are driving a large uptake of ad-blocking tools like Ghostery, Disconnect, AdBlock and Adblock Plus, but often these browser-based tools are outdated (i.e. new APs are not on their block lists), are detected by websites which then request (or force) them to be disabled or break the website structure such that the user cannot effectively use the website. The following recent research attempts to address this concern through new privacy-enhancing blocking methods.

Wu et al. (2015) proposed a new machine learning based system to detect and block third-party tracker scripts automatically based on their actions and signatures and validated a high level of accuracy although future improvements are suggested [12]. The concern with this approach is that without the ability to optimise targeting much of the internet content will need to be changed to paid access models and the site-breakage concern is not addressed.

A somewhat more sustainable solution is proposed by Yu et al. (2016) through a novel approach to online tracking privacy improvement where users identify data they believe can uniquely identify them [9]. The researcher tests their new system that removes this data from third-party data requests in real-time across 200,000 users in Germany and find that this approach achieves better protection than the well-known ad-blocker Disconnect, causes less website and advertisement tracking breakage (than other ad-blockers) and concludes that they could block user identifiers across 78% of their test sites [9]. The removal of sensitive data parameters will improve website visitor privacy protection, but APs that are affected by this method could start to encrypt these data parameters before transmission to circumvent the data removal.

D. Trust in Advertising

The younger generation are growing up with easy access to everything everywhere, but they also better understand the online world and its various hazards. With a better understanding of the information that browsers could collect about them (including location information), these increasingly heavy web users browse with less trust. The effect of trust on advertising is considered by these recent literature papers.

Cottrill and 'Vonu' Thakuriah (2013) reviews public and private sector privacy policies to ascertain how well end-users can comprehend them and how privacy is addressed in general as well as whether website visitor location privacy elements are explicitly addressed [17]. The research finds that there is a definitive lack of focus on privacy policies with regards to location-based data collection and sharing and that different industry sectors address privacy inconsistently. They suggest some privacy policy guidelines to address location-based data concerns [17]. Although this privacy policy analysis concentrates on location information, it highlights the lack of clarity within some privacy policies that could provide a false sense of security to end-users and undermine trust.

More directly Bleier and Eisenbeiss (2015) studied how effective retargeting is in relation to the user's trust of advertisers by considering the accuracy (depth) and the number of interest items correctly covered within an advert (breadth) of banner adverts [10]. They find that advertisers with higher consumer trust can get effective results with increased depth and less breadth without raising concerns about privacy from end-users [10]. Trust is very important in the internet world where there are seemingly always other options for customers to switch to and this research emphasises the importance for websites to be transparent and honest within their privacy policies to ensure trust is built and maintained.

E. Limiting Online Privacy Leakage

Privacy leakage is one of the main concerns with the advances in online tracking, data merging and data analytics. This allows sensitive or private information collected by third-parties to be more easily found, shared and used without the end-user's explicit knowledge. The following research papers investigated this concern and some proposed changes to the ad delivery model.

Estrada-Jiménez et al. (2017) examines the advertising tracking infrastructure and privacy technologies and performs a survey to compare privacy mechanisms [2]. They conclude that the existing ad delivery model needs significant modification before visitor privacy can be effectively improved [2]. The finding seems almost obvious, but with this research proof, the next step should be to find ad delivery methodology that can protect end-users while also ensuring advertising revenue generation.

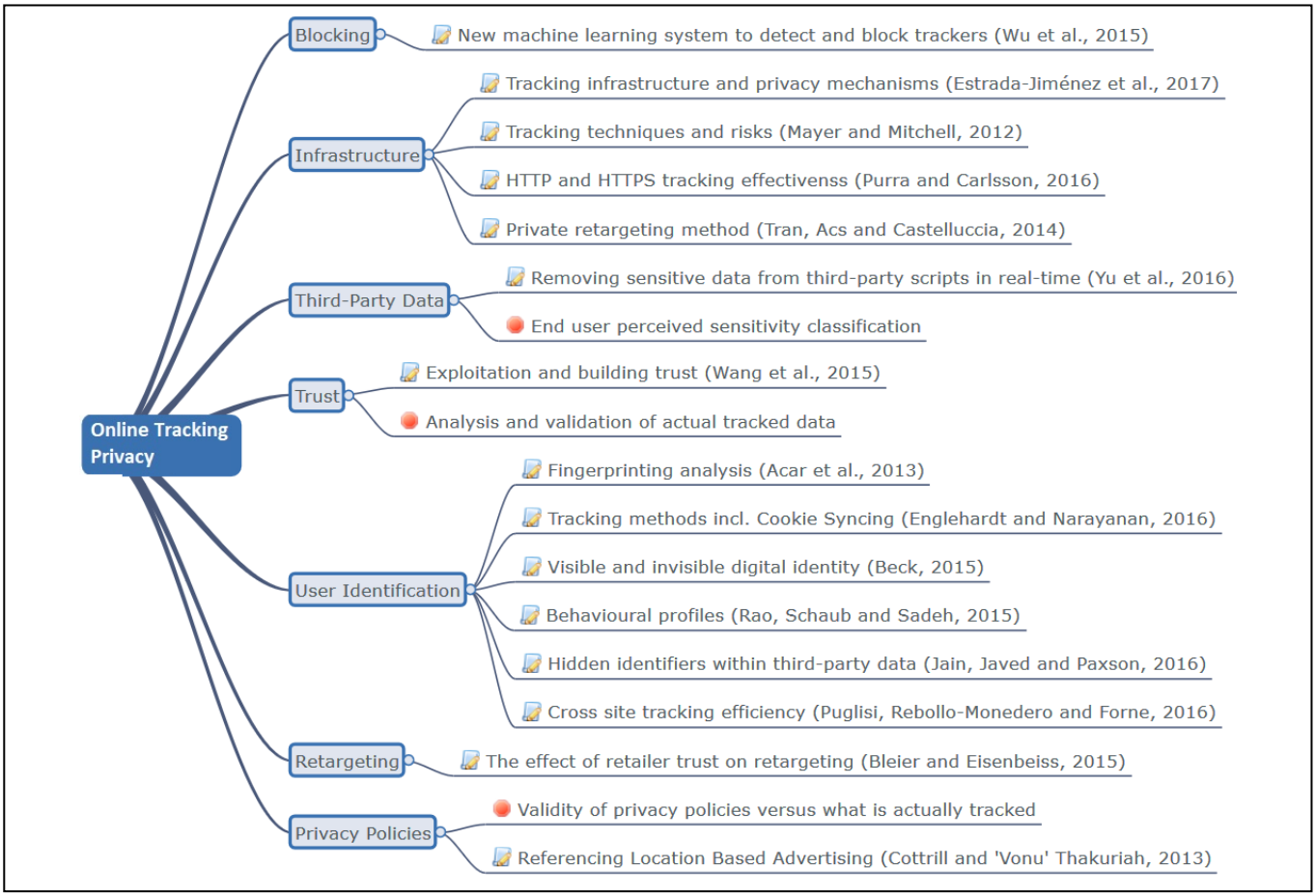

Fig. 4. Online Tracking Privacy Literature Review BOK

One such option is provided by Wang et al. (2015) who stresses the concern around aggressive ad-brokers that can abuse private information during tracking and targeting and suggests a new framework that uses compensation to incentivise website visitors for the use of their private information and allows end-users to control the level of sensitive and private information that is shared [3]. This model has real potential to bridge AP requirements and visitor protection.

A seemingly similar but more thorough methodology is suggested by Tran, Acs and Castelluccia (2017) who offers an advertisement retargeting system that circumvents the risk of unwanted privacy leakage to third-parties through utilising user-localised profiles that are homomorphically encrypted and shared with APs during the Real-Time Bidding (RTB) process [23]. Homomorphic encryption negates the need to decrypt encrypted data before calculations are performed on it and doing this encryption on the user's computer avoids speed concerns during RTB and ensures privacy confidentiality [23], [24]. This is a surprisingly innovative and feasible approach that would mitigate many malicious attack vectors exposed through third-party JavaScript tracking and resolves advertising industry struggles to balance user privacy and effective targeting. It will still require a big effort to roll out, but the wide-ranging benefits should make this realistic.

F. Gaps in the Literature

The extensive tracking privacy-related literature review in the earlier part of this literature review section is presented visually within a Body of Knowledge (BOK) diagram in Fig. 4. We highlight the gaps in the literature with red icons indicating that there is a lack of clarity with regards to sensitivity perception of end-users towards the actual parameters being collected by third-parties for the use of advertising purposes. We also describe the actual data being collected.

III. Designing AdPExT

To collect and categorise third-party parameter data belonging to advertising platforms (APs), a method was required that could read these parameters from websites at an individual parameter level and in a timely manner. No existing tool could be found through literature and internet searches that completely satisfied this need and a new tool called the Advertising Parameter Extract Tool (AdPExT) was designed as presented in this research paper to enable this study and improve efficiency for future researchers. The tool can be downloaded from a GitHub repository (https://github.com/waldu/AdPExT).

During the design research phase, it was found that although various Python packages exist to help to scrape data from websites (Django, Selenium, urllib2, BeautifulSoup4, html5lib, PyQt5, requests, lxlm, mechanicalsoup), the majority do not collect the third-party JavaScript or image pixel parameters transferred as part of the HTTP request to APs [25], [26]. This is seemingly specifically challenging due to it not being part of the page source. A website's HTML elements (i.e. the page source) can be scraped with a small amount of code, but integrating the third-party HTTP request elements required more advanced methods and it was found that the OpenWPM project by Englehardt provided the most efficient way to achieve this [19]. Furthermore, OpenWPM best simulates a true user experience (albeit without user login processes) with full-featured browser usage rather than a stripped-down browser version as is generally employed during web scraping automation. For example, some websites have been found to not serve ads to the well-known Python web scraping library PhantomJS [16].

```
https://stats.g.doubleclick.net/r/collect?v=1&aip=1&t=dc&_r=3&
tid=UA-1645798-21&cid=927237433.1498998715&jid=420245
892&gid=519006367.1498998715&gjid=1657828343&v=j56&z
=661177318
```

Fig. 5. Third-party URL with parameters as collected by OpenWPM (Englehardt, 2016)

The OpenWPM Python package populates the HTTP request and response information (captured during the page load) into the "http_requests" and "http_responses" tables in an SQLite database and the third-party tracking scripts along with the parameter data are inserted into the URL columns in these tables as one string. Fig. 5 shows an example of the DoubleClick (owned by Google) tracking image URL string with the parameter information after the question mark ("?"). For this research into online advertisement tracking these parameters needed to be broken up into individual key and value pairings to allow for the detailed categorisation and analysis and thus the OpenWPM project needed to be extended. During testing, it was found that the third-party tracking data URLs that were added into the "http_requests" and "http_responses" tables are the same in both tables and thus the parameter cleansing process design only focused on one of the tables ("http_responses"). The new coding was done in Python since OpenWPM is a Python library and the researcher had familiarity with the programming language. All new files and SQLite tables created with AdPExT were prefixed with "msc_" for easy identification within the research paper and database.

A. Pre-design Validation

Pre-design validation was run before the extensive design process to ensure the planned data extraction and parameter cleansing method was reliable. A manual data retrieval was completed for www.smallestwebsitetotheworld.com and www.reddit.com of two relevant third-party AP parameter strings (from different AP sub-domains) in real-time through the Mozilla Firefox browser's network monitor [18]. To extract the parameter data via the OpenWPM library new proof of concept (POC) code was created and run for the two test URLs.

The first test was specifically done with www.smallestwebsitetotheworld.com to reduce the data noise. It can be noted in Table I that for this website the Google Analytics tracking URL and parameters collected via the POC code is very similar to the URL and parameters of the browser network monitor in Fig. 6. Not all the values within this comparison are the same and this is expected since the data collection session is automatically closed during the OpenWPM code execution and the parameters from the network monitor information in Fig. 6 is from a separate manual session a few minutes later albeit within the same computer and browser. This close similarity provides proof that the OpenWPM library collects the correct third-party parameters in sufficient detail as required for the research aim.

Fig. 6. www.smallestwebsitetotheworld.com manual data validation for Google Analytics image pixel

TABLE I. WWW.SMALLESTWEBSITETOTHEWORLD.COM DATA COLLECTED FOR METHOD VALIDATION

Website URL where data collected from	**Third-party tracking URL request with parameters**
http://www.smallestwebsitetotheworld.com/	http://www.google-analytics.com/r/collect?v=1&_v=j56&a=214507101&t=pageview&_s=1&dl=http%3A%2F%2Fwww.smallestwebsitetotheworld.com%2F&ul=en-us&de=UTF-8&dt=SMALLEST%20WEBSITE%20TO%20THE%20WORLD&sd=24-bit&sr=1920x975&vp=1366x697&je=0&fl=11.2%20r202&_u=IEBAAEABI~&jid=802918663&gjid=1496508818&cid=1362621894.1500218735&tid=UA-56270256-2&_gid=1256393134.1500218735&_r=1&z=1851836601

The further validation across a much bigger and tracker-heavy website www.reddit.com succeeded as well with the "quantserve" pixel reflected in both the data and the manual screen print via the browser network monitor tool.

B. Design

The process required to collect the third-party AP parameter information from websites is shown in Fig. 7 and it covers the initialisation and setup processes, data collection via the OpenWPM package and the parameter cleansing process which provide the individual key-value pairs [19]. Three new Python files ("msc_Collectdata.py", "msc_UseOpenWPM" and "msc_PramCleansing.py") were created to facilitate this OpenWPM extension and their main functions are described in this section along with clarification around the additional database expansion required to hold the new parameter detail as well as the eventual categorisation and AP mappings.

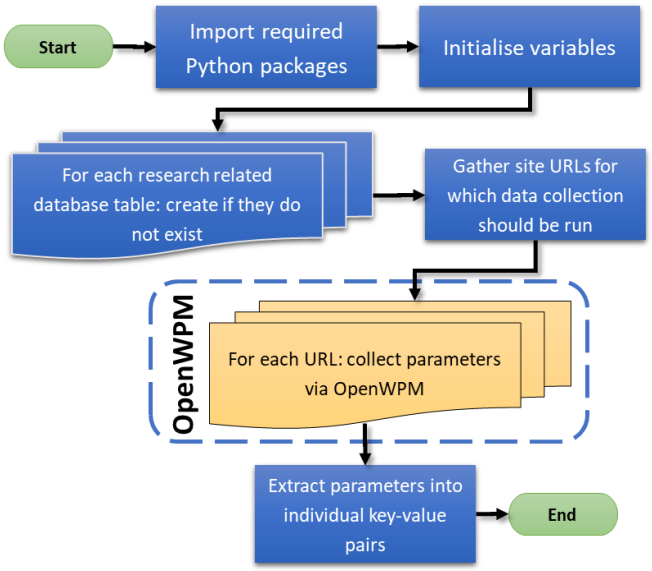

Fig. 7. AdPExT Third-party Parameter Data Collection Process Flow

a) Launching ADPExT: msc_Collectdata.py

A new Python "launch" file was created as "msc_Collectdata.py" which imports the OpenWPM Python library as well as the relevant required functions from the new "msc_ParamCleansing" library. The "msc_" prefix is also defined as a variable within this code file to allow for easy amendment during testing or by future researchers. The function "get_site_urls_to_extract" contains the list of web URLs that will be crawled and when this launch file is activated it sends this list to the "extract_via_openwpm" function in the "msc_UseOpenWPM" library and once the data is collected the parameter cleansing process is activated via the "extract_parameters" function from the "msc_ParamCleansing" library.

b) Activating OpenWPM Data Collection: msc_UseOpenWPM.py

The Python file "msc_UseOpenWPM" is based on the "demo.py" demonstration file provided by Englehardt although it was rewritten to fit into this extended code process [19]. After importing the OpenWPM "TaskManager" function from the "automation" library, a function called "extract_via_openwpm" activates the OpenWPM data collection with one active browser that processes each provided website URL consecutively.

c) Database Operations and Parameter Cleansing: msc_PramCleansing.py

The "msc_PramCleansing.py" Python file manages the database connections (function "open_db_conn"), additional table creations (function "setup_db_tables"), extraction of individual parameter key-value pairs (function "extract_parameters") as well as writing the key-value pairs into the database (function "add_param_into_db"). Using Python rather than SQLite to extract the key-value pairs within the "extract_parameters" function is very efficient with only one database hit performed to retrieve all the collected third-party URLs that contain parameter data (i.e. containing "?" in the URL string) before using a Python list variable and Python "split" function to separate out the parameter keys and values as pairs that are then written into the database table "msc_param_values".

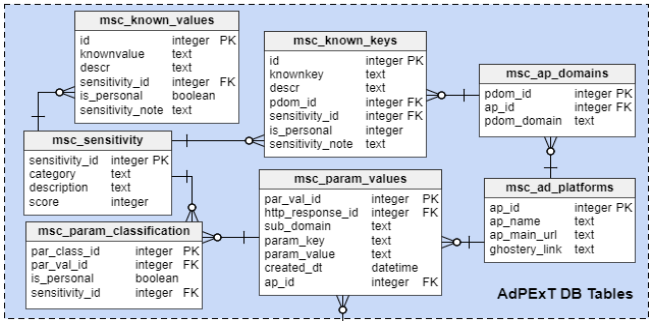

Fig. 8. Tables created for this research

d) Database Expansion

OpenWPM creates the SQLite database as per the configuration in a JSON file located in the "OpenWPM/automation" folder which by default stores the database file as "crawl-data.sqlite" within a desktop folder "openwpm". In addition to the ten OpenWPM tables that are automatically created, for AdPExT a further seven tables (shown in Fig. 8) were created to facilitate the parameter cleansing and categorisation. The main table is "msc_param_values" which contains the separated parameter key-values and is linked to the APs in the "msc_ad_platform" table and the classification in the "msc_param_classification" table. The remaining tables are used during the classification process to identify various keys and values of interest for the research.

e) Design Challeneges

The design process had a few struggles that is noteworthy for future researchers utilising AdPExT. The SQLite database in use sometimes gets locked for an unknown reason. The problem seems to be related to the loops writing to the database, but although the newly written code implements explicit database connection close commands (i.e. "conn.close()") the issue persists intermittently. Running the connection close command manually sometimes released the lock, but most often the best solution was to save everything and restart the VirtualBox computer. Additionally, the Sqliteman software within the virtual machine would sometimes freeze and require to be closed and re-opened which would lose all unsaved SQLite code within the platform. Frequent backups are proposed and versions of Sqliteman after V1.2.2 might resolve this issue.

Additionally, the current code method does not log into websites or share any user name, address and e-mail data to detect whether that info is tracked.

IV. Sensitivity Perception Survey

As part of this study, we investigated the perception of Internet users towards the sensitivity of the several tracking parameters gathered by APs.

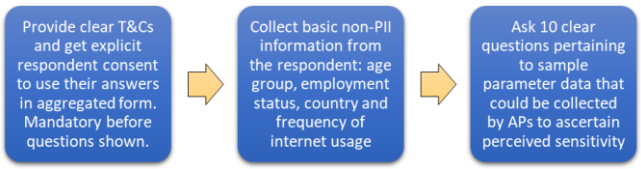

Fig. 9. Survey Questions Flow

A. *Survey Design*

Google Forms provides an easy, clear and reportable survey engine with an online form that could be shared across numerous contact streams and social media platforms [27]. The survey question flow (Fig. 9) was structured to ensure explicit consent is provided by respondents before any survey questions are shown and answered based on guidance from SurveyMonkey [28]. The first page of the survey stipulates the purpose of the survey and provides clear terms and conditions that the respondent needs to agree with. On agreement, four non-personal questions are put to the respondent to determine their age group, employment status, country and frequency of use of the internet. These metrics assist to measure the appropriateness of the audience with regards to the research aim and objectives. Finally, the last section of the survey provides ten clear sample parameter types that APs might collect and request the respondent to rate how sensitive they consider these parameters to be according to a categorical scale of sensitivity from one (public knowledge) to five (high risk) as shown in Table II.

TABLE II. PERCEIVED DATA SENSITIVITY CATEGORIES

Perceived Sensitivity	Category Description	Sensitivity Score
Public Knowledge	I do not care about this data	1
Insensitive	Not concerned about this	2
Sensitive	I am somewhat concerned about this	3
Private	I am concerned about this and will want to know this is being collected	4
High Risk	I would not freely allow this information to be collected by 3rd parties and want explicit notification	5

The survey was distributed mainly within the United Kingdom through online methods including direct e-mail to the researcher's contacts, postings on Facebook and LinkedIn and additional distribution amongst professional contacts. The survey data collection process took three months from June 2017 and a total of 164 respondents agreed to the terms and conditions and completed the survey questions.

B. *Survey Results*

From the classifying questions within the survey, it can be noted that the respondents are mostly between the ages of 26 and 55 (76%) while 86% are full time employed (Table III) and 74% are based within the United Kingdom. Additionally, only 2 out of the 144 respondents reportedly do not use the internet on a daily basis (both in the age group "56 years old or older"). The good proportional presentation by the age groups between 26 and 55 lines up well with the results from a recent large study by the Office for National Statistics which found 97% of adults aged between 35-54 frequently used the internet in 2017 [30]. But it would have been better to also get a higher proportion of participants aged between 19 and 25 since this younger age group is similarly active on the internet [30]. Overall these metrics verify the relevance of the survey audience since employed, UK based adults that frequently use the internet are well placed to measure the sensitivity perception of their information that might be used by APs across the top twenty UK websites.

TABLE III. EMPLOYMENT STATUS OF VARIOUS AGE GROUPS

In which age group do you fall?	Are you currently Employed?			
	No, I am a student	*No, not currently employed*	*Yes, full time*	*Yes, part time*
18 years old or younger	1			
19 - 25 years old	3		23	2
26 - 35 years old	1	1	50	3
36 – 55 years old		4	48	3
56 years old or older		2	3	
Total	**5**	**7**	**124**	**8**

The summarised results of the perceived sensitivity of each of the sample parameters are visualised in Fig. 10 with the Facebook ID, Google ID, e-mail and IP address notably as the most sensitive.

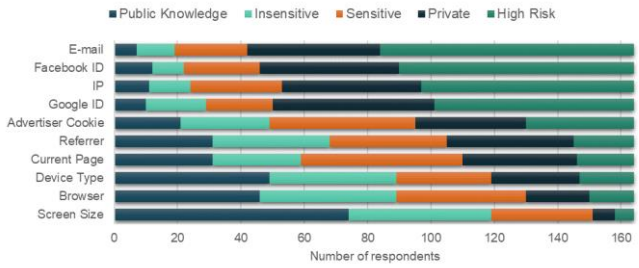

Fig. 10. Survey Parameter Perceived Sensitivity Results

V. CONCLUSIONS AND FUTURE WORK

Empowering end-users to monitor, understand and validate the data collection practices of advertising platforms (APs) was a key motivation for this study. The sensitivity perception survey highlighted how e-mail and IP addresses have been perceived as high-risk data items and that Google and Facebook IDs also generate concern due to the implied relation to the additional private information linked to those IDs. This justifies the need for a tool that can enable future researchers to investigate the AP parameter values efficiently and correctly and the design of the new AdPExT (Advertising Parameter Extraction Tool) platform presented in this paper sufficiently satisfies that requirement.

As part of our future work, the tool will be tested through a validation process that will compare results generated by the tool with manual data extraction results via the freely available Firefox Developer Tools network monitor [18]. An

extension of this study would also enquire to investigate the suitability of publishers' privacy policies as well as the actual practices of smaller APs. The AdPExT tool can also be enhanced with additional automation and reporting capabilities. Other related topics that can be researched include the impact of using a mixture of normal and private browsing (e.g. Incognito mode in Google Chrome) [31] on the quality of information gleaned by APs.

REFERENCES

[1] C. Pemberton, "Gartner CMO Spend Survey 2016-2017 Shows Marketing Budgets Continue to Climb - Smarter With Gartner". 2017. [online] Gartner. Available at: http://www.gartner.com/smarterwithgartner/gartner-cmo-spend-survey-2016-2017-shows-marketing-budgets-continue-to-climb/ [Accessed 5 Aug 2018].

[2] J. Estrada-Jiménez, J. Parra-Arnau, A. Rodríguez-Hoyos, and J. Forné, "Online advertising: Analysis of privacy threats and protection approaches," Computer Communications, vol. 100, pp. 32–51, Mar. 2017. doi: 10.1016/j.comcom.2016.12.016

[3] W. Wang, L. Yang, Y. Chen, and Q. Zhang, "A privacy-aware framework for targeted advertising," Computer Networks, vol. 79, pp. 17–29, Mar. 2015. doi: 10.1016/j.comnet.2014.12.017

[4] S. Puglisi, D. Rebollo-Monedero, and J. Forne, "On Web user tracking: How third-party http requests track users' browsing patterns for personalised advertising," in 2016 Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net), 2016. doi: 10.1109/MedHocNet.2016.7528432

[5] E. N. Beck, "The Invisible Digital Identity: Assemblages in Digital Networks," Computers and Composition, vol. 35, pp. 125–140, Mar. 2015. doi: 10.1016/j.compcom.2015.01.005

[6] S. Jain, M. Javed, and V. Paxson, "Towards Mining Latent Client Identifiers from Network Traffic," Proceedings on Privacy Enhancing Technologies, vol. 2016, no. 2, pp. 100–114, Apr. 2016. doi: 10.1515/popets-2016-0007

[7] J. Purra and N. Carlsson, "Third-Party Tracking on the Web: A Swedish Perspective," in 2016 IEEE 41st Conference on Local Computer Networks (LCN), 2016. doi: 10.1109/LCN.2016.14

[8] A. Rao, F. Schaub and N. Sadeh "What do they know about me? Contents and Concerns of Online Behavioral Profiles". arXiv preprint arXiv:1506.01675. 2015 [online] Available at: https://arxiv.org/abs/1506.01675 [Accessed 3 Aug 2018].

[9] Z. Yu, S. Macbeth, K. Modi, and J. M. Pujol, "Tracking the Trackers," in Proceedings of the 25th International Conference on World Wide Web - WWW '16, 2016. doi: 10.1145/2872427.2883028

[10] A. Bleier and M. Eisenbeiss, "The Importance of Trust for Personalized Online Advertising," Journal of Retailing, vol. 91, no. 3, pp. 390–409, Sep. 2015. doi: 10.1016/j.jretai.2015.04.001

[11] G. Acar et al., "FPDetective," in Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13, 2013. doi: 10.1145/2508859.2516674

[12] Q. Wu, Q. Liu, Y. Zhang, and G. Wen, "TrackerDetector: A system to detect third-party trackers through machine learning," Computer Networks, vol. 91, pp. 164–173, Nov. 2015. doi: 10.1016/j.comnet.2015.08.012

[13] J. R. Mayer and J. C. Mitchell, "Third-Party Web Tracking: Policy and Technology," in 2012 IEEE Symposium on Security and Privacy, 2012. doi: 10.1109/SP.2012.47

[14] Internet Advertising Bureau. IAB UK reveals latest ad blocking behaviour | IAB UK. 2016. [online] Available at: https://www.iabuk.net/about/press/archive/iab-uk-reveals-latest-ad-blocking-behaviour [Accessed 16 Aug 2018].

[15] B. Shiller, J. Waldfogel, and J. Ryan "Will Ad Blocking Break the Internet?". National Bureau of Economic Research Working Paper Series. 2017. [online] Available at: http://www.nber.org/papers/w23058 [Accessed 22 Aug 2018].

[16] S. Englehardt and A. Narayanan, "Online Tracking," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16, 2016. doi: 10.1145/2976749.2978313

[17] C. D. Cottrill and P. 'Vonu' Thakuriah, "Privacy in context: an evaluation of policy-based approaches to location privacy protection," International Journal of Law and Information Technology, vol. 22, no. 2, pp. 178–207, Nov. 2013. doi: 10.1093/ijlit/eat014

[18] Mozilla Developer Network. Network Monitor. 2017 [online] Available at: https://developer.mozilla.org/en-US/docs/Tools/Network_Monitor#Network_request_details [Accessed 9 Aug 2018].

[19] S. Englehardt, "Setting up OpenWPM". 2016. [online] GitHub. Available at: https://github.com/citp/OpenWPM/wiki/Setting-Up-OpenWPM [Accessed 27 Aug 2018].

[20] EU GDPR Portal, "Timeline of the EU GDPR". [online] Available at: http://www.eugdpr.org/gdpr-timeline.html [Accessed 6 Aug 2018].

[21] H. Bitar and B. Jakobsson, "GDPR: Securing Personal Data in Compliance with new EU-Regulations". 2017. [online] Available at: http://ltu.diva-portal.org/smash/record.jsf?pid=diva2%3A1113478&dswid=1029 [Accessed 22 Aug 2018].

[22] B. van Loenen, S. Kulk, and H. Ploeger, "Data protection legislation: A very hungry caterpillar," Government Information Quarterly, vol. 33, no. 2, pp. 338–345, Apr. 2016. doi: 10.1016/j.giq.2016.04.002

[23] M. Tran, G. Acs, and C. Castelluccia, "Retargeting Without Tracking. CoRR". 2014. [online] Available at: http://arxiv.org/abs/1404.4533 [Accessed 10 Aug 2018].

[24] M. Mazza, "Homomorphic encryption: a new potential for cryptography?". 2016. [online] The Kroll Ontrack UK Blog. Available at: https://www.krollontrack.co.uk/blog/the-world-of-data/homomorphic-encryption-a-new-potential-for-cryptography/ [Accessed 3 Oct. 2017].

[25] Python For Beginners. "How to use urllib2 in Python". 2013 [online] Available at: http://www.pythonforbeginners.com/python-on-the-web/how-to-use-urllib2-in-python/ [Accessed 22 Aug 2018].

[26] L. Richardson, "Beautiful Soup: We called him Tortoise because he taught us". 2017. [online] Crummy.com. Available at: https://www.crummy.com/software/BeautifulSoup/ [Accessed 19 Aug 2018].

[27] Google.co.uk, "Google Forms - create and analyze surveys, for free". [online] Available at: https://www.google.co.uk/forms/about/ [Accessed 20 Aug 2018].

[28] SurveyMonkey, "Adding a Consent Statement or Privacy Policy". [online] Available at: https://help.surveymonkey.com/articles/en_US/kb/How-do-I-create-a-consent-form-or-disqualify-respondents-from-a-survey [Accessed 20 Aug 2018].

[29] Ghostery. "Ghostery Makes the Web Cleaner, Faster and Safer!". [online] Available at: https://www.ghostery.com/ [Accessed 16 Aug 2018].

[30] Office for National Statistics, "Internet users in the UK - Office for National Statistics". 2017. [online] Available at: https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2017 [Accessed 24 Aug 2018].

[31] C. Flowers, A. Mansour, and H. M. Al-Khateeb, "Web browser artefacts in private and portable modes: a forensic investigation," International Journal of Electronic Security and Digital Forensics, vol. 8, no. 2, p. 99, 2016. Doi: 10.1504/IJESDF.2016.075583