What's in the Community Cookie Jar?

Aaron Cahn*, Scott Alfeld†, Paul Barford*†, S. Muthukrishnan‡
*comScore, †University of Wisconsin-Madison, ‡Rutgers University
*acahn@comscore.com, †{salfeld, pb}@cs.wisc.edu, ‡muthu@cs.rutgers.edu

***Abstract*—Third party tracking of user behavior via web cookies represents a privacy threat. In this paper we assess this threat through an analysis of anonymized, crowd-sourced cookie data provided by Cookiepedia.co.uk. We find that nearly 45% of the cookies in the corpus are from Facebook and of the remaining cookies 25% come from 10 distinct domains. Over 65% are Maximal Permission cookies (*i.e.,* 3rd party, non-secure, persistent, root-level). Cookiepedia's anonymization of user data presents challenges with respect to modeling site traffic. We further elucidate the privacy issue by conducting targeted crawling campaigns to supplement the Cookiepedia data. We find that the amount of traffic obscured by Cookiepedia's anonymizing procedure varies dramatically from site to site – sometimes obscuring as much as 80% of traffic. We use the crawls to infer the inverse function of the anonymizing procedure, allowing us to enhance the crowd-sourced dataset while maintaining user anonymity.**

I. Introduction

Since the inception of cookies as a mechanism to maintain state in web browsing sessions, their use has grown widely. However, cookies are a double edged sword, repeatedly attracting the attention of the popular media due to exploitable vulnerabilities and user privacy concerns; particularly with the emergence of 3rd party cookies, which are used primarily to track user behavior. These issues have led to proposals for new mechanisms that selectively block or allow the setting of cookies on the browser [7], [12].

The goal of our work is to improve the general understanding of cookie setting behavior and it's implications for privacy. We approach this problem from an empirical perspective using crowd-sourced data provided by Cookiepedia[1]. While the Cookiepedia data is useful for understanding a variety of cookie characteristics, the way in which the data is obfuscated for user privacy limit its utility for inferring user browser behavior. In this paper, we first analyze a month-long snapshot of the Cookiepedia data, examining cookie characteristics and domain footprints [2]. In the second part of this paper, we enrich the crowd sourced cookie data set by conducting a set of targeted crawls to the top sites indicated by Cookiepedia. Our intention is not to reverse engineer the privatization provided by Cookiepedia, rather to gain further insights into privacy-related issues using Cookiepedia as a guide.

Through crawling, we find that the view of each individual site, in terms of the cookies it sets, is affected by the Cookiepedia anonymization procedure in a variety of ways. In particular, we observe that from 29% to **up to 85%** of total traffic to a site is obscured. In addition, the perceived proportion of Maximal Permission [2] (*i.e.,* non-secure, persistent cookies with a root-level path) cookies is almost universally increased by anonymizing the data (by up to 33 percentage points for an individual site).

In summary, this paper makes the following contributions. First, we provide a first look at characteristics of cookies that are commonly placed on user's browsers in their daily web surfing activities. We find that the vast majority are Maximal Permission. Second, guided by our crowd sourced data, we perform targeted crawls on the most commonly visited sites to examine the details of cookie setting behavior. Our crawls reveal that the Cookiepedia anonymization procedure has various, dramatic effects on different sites, and provides a supplemental data set to provide a more accurate view of the cookie-scape.

II. Cookiepedia Data Set

The Cookiepedia project "aims to build a comprehensive knowledge base about website cookies and similar technologies." Cookiepedia collects data in a crowd-sourced fashion through freely available Firefox- and Chrome-based plug-ins called the Cookie Collector [3]. The plug-ins collect all attributes of cookies as users surf to sites with the exception of the *value* attribute, which is omitted for privacy reasons. When the plug-in transmits cookies back to Cookiepedia, a small amount of additional metadata is computed during storage into their database.

Cookiepedia provided us with a one month snapshot of their cookie database. The corpus contains aggregate data from January 1st, 2014 to January 31st, 2014. The data was anonymized in a three-fold fashion. First, as mentioned above, the value attribute of all cookies was removed to preserve the privacy of individual users. Second, the path of the web page a user was visiting when a particular cookie was collected is removed. For example, if a cookie is collected on google.com/account, Cookiepedia will log this as coming from google.com. Third, and most limiting to our analysis, Cookiepedia computes a set operation (that is, removes duplicates and forms an unordered collection) on a tuple of cookie attributes. The tuple consists of the name, host, urlDomain, path, isSecure, isHttpOnly, and isSession attributes. We were provided a single example of each unique (defined by this tuple) cookie.

We acknowledge Cookiepedia is a UK-based entity, and therefore there is the potential for user browsing behavior to skew toward UK or European-based web sites. However, due to data obfuscation, it is difficult to directly assess geographic bias. Our methods for enhancing the dataset are general, and may be freely applied to other anonymized (*e.g.*, user web browsing) data.

[1] www.cookiepedia.co.uk

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA 978-1-5090-2846-7/16/$31.00 ©2016 IEEE

Total 3rd Party Cookies		Host Footprint	
adnxs.com	27,878	google.com	7,885
rfihub.com	24,446	adnxs.com	6,684
rubiconproject.com	21,910	doubleclick.net	6,205
facebook.com	17,070	facebook.com	4,442
invitemedia.com	16,616	youtube.com	3,567
google.com	14,424	quantserve.com	2,763
pubmatic.com	13,745	twitter.com	2,582
rlcdn.com	12,987	turn.com	2,547
bluekai.com	12,347	addthis.com	2,521
youtube.com	12,275	criteo.com	2,404

TABLE I: The top 10 3rd party hosts setting the most unique 3rd party cookies (columns 1 & 2) and the top 10 3rd party hosts with the largest footprints (columns 3 & 4) over the Cookiepedia corpus.

III. Cookie Characteristics

For brevity, we provide a summary of the raw statistics (and percentages) for the Cookiepedia corpus in Table II. The corpus contains a total of 1,364,041 cookies. Note, the numbers in Table II are conservative because, as mentioned above, we do not have the aggregate counts for non-unique cookie tuples. A total of 49,096 unique web sites were visited by users of Cookiepedia's cookie collection plug-in to obtain this data set. The cookies collected during visits to these 49,096 unique web sites came from a total of 30,998 distinct hosts.

Table II provides interesting insights into the way in which cookies are being used and attributes are being set. Most notably is the use of what we call *Maximal Permission* cookies. A cookie is Maximal Permission if it is persistent, non-secure, and has its path set to the root level [2]. This definition encompasses both 1st and 3rd party cookies, but we focus here on 3rd party cookies.

The Cookiepedia corpus contains 500,568 Maximal Permission 3rd party cookies. A full **80.3%** of 3rd party cookies are Maximal Permission. These cookies give 3rd party entities the ability track users both across pages within a site and across separate sites which set their cookie(s). Additionally, Maximal Permission cookies are persistent and often stay active on a client's browser for months or years. This means 3rd party entities can correlate user information across multiple browsing sessions. This gives the information collected by Maximal Permission 3rd party cookies both breadth and depth. We note that often users have no direct interaction with a 3rd party entity when visiting a webpage, and thus are likely unaware of the magnitude of 3rd party tracking and dissemination of personal information taking place (see [2] for a study of user information leakage via cookies).

Any given domain may place 3rd party cookies elsewhere on the Web. The number of 3rd party cookies a domain places varies dramatically, as well as the number of sites on which the domain places a cookie. A domain's *footprint* [2] is defined as the set of sites on which the domain places at least one cookie. Table I shows the 10 domains which set the most 3rd party cookies, as well as the 10 with the largest footprints. We note that while adnxs.com (the domain used by the AppNexus ad sever) placed the most 3rd party cookies, google.com has an $\sim 18\%$ larger footprint with nearly half as many cookies. It is also important to point out that ad servers routinely convey cookie-based information to advertisers to enhance targeting.

IV. Enriching Crowd Sourced Data

The observations of the prior section provide a baseline for understanding cookie attributes from a crowd sourced perspective. However, the anonymization procedure Cookiepedia performs on the data set limits the scope and applicability of an investigation of the data, as is common for publicly available user data. Specifically, the multi-step anonymization procedure skews certain aspects of the data set in a way that can lead to inaccurate conclusions about cookies and cookie placement mechanisms. Therefore, we expand the perspective of the crowd sourced data through a series of targeted crawling campaigns.

We utilize the crawling infrastructure of Cahn *et al.* [2]. In essence, we are crawling sites to infer the inverse operation of the anonymization procedure, and then applying it to Cookiepedia's aggregated data. This allows us to supplement the (anonymized) Cookiepedia dataset, yielding more accurate inference of actual user traffic while still maintaining anonymity of users. Obtaining the data from the automated crawls, rather than additional human users, allows us to more broadly understand the crowd sourced data set in a manner that preserves user anonymity.

A. Experimental Setup

We crawled a subset of sites that were responsible for setting a large number of cookies so that we could discern the methods used from site to site which lead to the observed behavior. Understanding a site's cookie naming pattern and subsequent cookie setting policy allows us to investigate the variance from site to site in the number of truly unique cookies being set per site. To this end, we chose the top ten sites which set the most 1st party cookies as the set of target sites for the crawling campaign.

As a pre-processing step, for each of the ten target sites, we utilize the crawling infrastructure to spider down within the site to a depth of twenty. During the spidering process the path the Crawler took (*i.e.,* the pages within the site the Crawler visited) was logged. This created a set of twenty web pages per target site which constitutes the basis for the supplemental crawling campaign. The crawling campaign over these two hundred web pages was run five times. We were careful to clear the Firefox profile between each site (*i.e.,* after visiting the set of twenty web pages for a particular site) thereby wiping out any accumulation of cookies. This was done to isolate the data obtained from a particular site so it could not influence the data from a subsequent crawl.

B. Results

For the 10 sites crawled, we collect a set $\mathscr{C}_{\text{ori}}$ cookies. We then perform Cookiepedia's anonymization procedure on $\mathscr{C}_{\text{ori}}$ to obtain $\mathscr{C}_{\text{anon}}$. With these two data sets, we are able to infer the effects of the anonymization procedure – and apply the inverse operation to the original Cookiepedia data.

TABLE II: Raw Cookiepedia Data

	Total	Session	Persistent	isSecure	isHTTPOnly	Path Depth = 1	Path Depth > 1
Total Cookies	1,364,041	742,460	621,581	17,459	639,880	1,329,116	34,925
1st Party Cookies	740,650 (54.3%)	641,843 (86.4%)	98,807 (15.9%)	2,999 (17.2%)	587,640 (91.8%)	731,260 (55.0%)	9,390 (26.9%)
3rd Party Cookies	623,391 (45.7%)	100,617 (13.6%)	522,774 (84.1%)	14,460 (82.8%)	52,240 (8.2%)	597,856 (45.0%)	25,535 (73.1%)

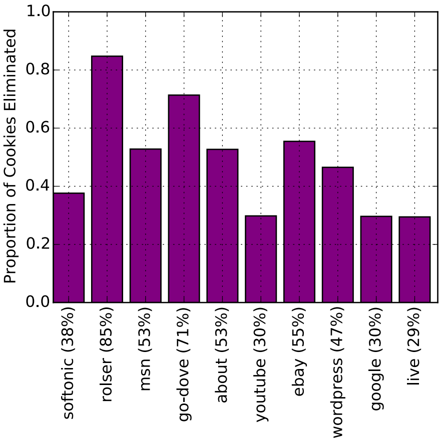

Fig. 1: **Proportion of Traffic Obscured.** Proportion of an individual site's traffic (measured by number of cookies) which is hidden by the aggregation step of Cookiepedia's anonymization procedure.

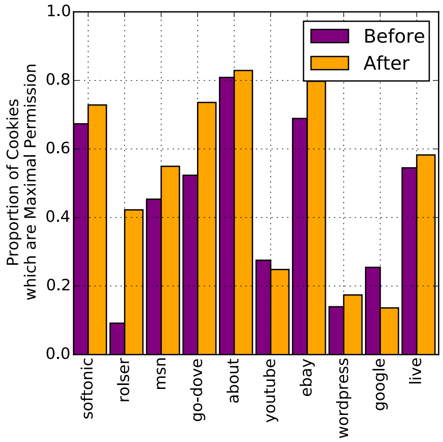

Fig. 2: **Prevalence of Maximal Permission Cookies.** Values shown in purple are from the raw crawl, before any additional processing. Values shown in yellow are from after the anonymization procedure.

For each site S, we calculate the proportion of cookies hidden by the aggregation step: $1-(|\mathscr{C}_{\text{anon}}|/|\mathscr{C}_{\text{ori}}|)$. As seen in Figure 1, the amount of traffic obscured varies dramatically between sites. At a minimum, the anonymization procedure obscured (hid) roughly 29% of traffic (live.com). For rolser.com, a staggering **85% of traffic was obscured**. That is, for every 100 cookies originally collected, only 15 are present (due to aggregation) in the resulting data set. This demonstrates how the overall view of user traffic is altered – when the data suggests that site A receives more traffic than site B based on the number of cookies collected from users, it may only be due to the skew induced by the anonymizing step (obscuring more of B's traffic than A's).

We now turn to examining sites in isolation. While the view of total user traffic is skewed as described above, the distribution of any specific site's cookie makeup is also altered by the anonymization procedure. We focus first on *Maximal Permission* cookies, as described in Sec III, and then move to a more general view.

1) Maximal Permission Cookies: For each site crawled, we compute the proportion of 3rd party Maximal Permission cookies in both $\mathscr{C}_{\text{ori}}$ and $\mathscr{C}_{\text{anon}}$ (see Figure 2). We note two striking features. First, for nearly every site, the relative number of Maximal Permission cookies increases with the anonymization procedure. The exceptions are google.com and youtube.com, for which the proportion of Maximal Permission cookies decreases instead. Note, this is caused by the cookie naming conventions used by each site – if every cookie is given a unique name, for example, then no cookies are removed by the anonymization procedure. Second, the amount of change we see after the anonymization procedure varies dramatically with each site. The minimal increase we observe is roughly 2 percentage points (about.com), while rolser.com shows an increase of over 33 percentage points. In total, 84.7% of the 3rd party cookies in the original (unanonymized) crawl data are Maximal Permission, and 85.9% in the anonymized data (compared to 69.2% in the Cookiepedia collection when only considering these 10 sites.) We posit that the increase (from 84.7 to 85.9) is small due to the prevalence of Google cookies.

2) General Features: Moving to a broader perspective, we now examine the effects of the anonymization procedure on aspects of sites beyond total cookies and Maximal Permission cookies. We describe each site s as a fixed length vector:

$$[p^s_{3^{\text{rd}}\text{ Party}}, p^s_{\text{Not-Secure}}, p^s_{\text{Long-Term}}, p^s_{\text{Root-Path}}]$$

where p^s_f is the proportion of all cookies from s for which f is True. For example, $p^s_{\text{Not-Secure}}$ is the proportion of s's cookies which have isSecure set to False. Note that other, equally valid representations exist because we are considering proportions (for example, $p^s_{3^{\text{rd}}\text{ Party}} + p^s_{1^{\text{st}}\text{ Party}} = 1$). We select these particular features (rather than their negations) to facilitate easier interpretation of results – in general, p^s_f being high indicates a prevalence of Maximal Permission cookies. For example, a site with 100% 3rd party Maximal Permission cookies would be represented as: [1.0, 1.0, 1.0, 1.0].

For each site crawled, we examine its location in this 4-dimensional space before and after the anonymization procedure. For illustrative purposes, we show only 5 of the 10 sites crawled (see Figure 3), but the findings presented are consistent with results from the remaining sites. As Figure 3 illustrates, the anonymization procedure distorts the view of individual sites, both in magnitude and direction.

For example, consider the first (upper left) subfigure of Figure 3: 3rd Party vs Not-Secure. In this 2-dimensional space, each of the five sites are affected differently by the anonymization procedure. In particular, anonymizing the original data leads to Google having fewer 3rd party cookies and more secure cookies, while Ebay and msn show more 3rd party cookies than before anonymizing.

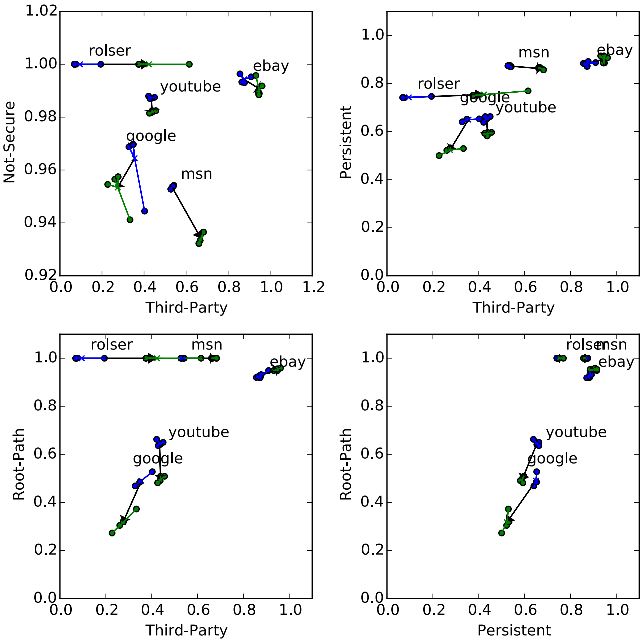

Fig. 3: **Effects of Anonymizing.** Each site is shown as five dots, one for each time it was crawled, connected by lines. The blue dots/lines show the values obtained from $\mathscr{C}_{\text{ori}}$, and the green dots show after anonymization has been performed. Black arrows show the explicit transformation. Note that the upper left y-axis has been adjusted for illustrative purposes.

V. Related Work

Web cookies have been mentioned in prior studies of web vulnerabilities (*e.g.,* [4], [5]). Tappenden *et al.* [10] use crawling to collect cookies from publisher websites. However, their work is focused on testing cookie setting mechanisms within a web development framework. Cahn *et al.* [2] perform a large scale crawl and model user information leakage through cookies – we utilize their crawler infrastructure for this work. Web user privacy and the potential for information leakage via web cookies has been considered in a number of prior studies (*e.g.,* [9], [11]). Mayer and Mitchell surveys privacy implications related to 3rd party tracking and discusses measurement methods that can be used to gather cookies [7]. Li *et al.* [6] perform an empirical study of cookie-based 3rd party tracking on top Alexa sites. Within the research community, there is a balance between providing useful real-world datasets and maintaining privacy. One framework for addressing this balance is provided in [1]. Finally, Roesner *et al.* examine the prevalence and details of 3rd party tracking using a small set of web crawls similar to ours, however the cookie characteristics are not the focus of their study [8].

VI. Summary and Conclusions

In this paper, we assess the distribution and privacy implications of cookie setting behavior in the internet. We do this by analyzing a collection of crowd-sourced cookies from the Cookiepedia project. We find that over 65% of the cookies are Maximal Permission. We used crawling to enrich the base data set without compromising user privacy. The results from our crawls highlight the fact that the anonymization methods used by Cookiepedia can lead to incorrect conclusions about global user traffic. Specifically, we find that a large percentage of traffic is obscured by the aggregation step of Cookiepedia's anonymization process based on the cookie setting mechanisms that are used by web sites. We also find that Maximal Permission cookies are likely to be underrepresented in Cookiepedia data. Finally, we develop a simple analysis and visualization to show how individual sites are affected by anonymization. This analysis highlights the spectrum of impact on top sites in the Cookiepedia data. In future work, we plan to continue to investigate web privacy through crowd-sourced and crawl based methods.

Acknowledgements

The authors would like to thank Cookiepedia for supplying the data used in this study. This material is based upon work supported by the DHS grant BAA 11-01 and AFRL grant FA8750-12-2-0328. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of the DHS or AFRL.

References

[1] M. Allman and V. Paxson. Issues and etiquette concerning use of shared measurement data. In *Proceedings of the ACM SIGCOMM conference on Internet measurement*, October 2007.

[2] Aaron Cahn, Scott Alfeld, Paul Barford, and S Muthukrishnan. An empirical study of web cookies. In *Proceedings of the World Wide Web Conference*, 2015.

[3] The Cookie Collective. About the Cookie Collector, 2015.

[4] I. Dacosta, S. Chakradeo, M. Ahamad, and P. Traynor. One-time Cookies: Preventing Session Hijacking Attacks with Stateless Authentication Tokens. *ACM Trans. Internet Technol.*, 12(1), July 2012.

[5] K. Fu, E. Sit, K. Smith, and N. Feamster. Dos and Don'ts of Client Authentication on the Web. In *Proceedings of the USENIX Security Symposium*. USENIX Association, August 2001.

[6] T. Li, H. Hang, M. Faloutsos, and P. Efstathopoulos. TrackAdvisor: Taking Back Browsing Privacy from Third-Party Trackers. In *Proceedings of the Passive and Active Measurement Conference*. Springer, March 2015.

[7] J. R. Mayer and J. C. Mitchell. Third-Party Web Tracking: Policy and Technology. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE Computer Society, March 2012.

[8] F. Roesner, T. Kohno, and D. Wetherall. Detecting and Defending Against Third-party Tracking on the Web. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*. USENIX Association, April 2012.

[9] A. Soltani, S. Canty, Q. Mayo, L. Thomas, and C. J. Hoofnagle. Flash Cookies and Privacy. aug 2009.

[10] A. F. Tappenden and J. Miller. Cookies: A Deployment Study and the Testing Implications. *ACM Trans. Web*, 3(3), July 2009.

[11] R. Tirtea. *Bittersweet cookies some security and privacy considerations* . feb 2011.

[12] C. Yue, M. Xie, and H. Wang. An Automatic HTTP Cookie Management System. *Comput. Netw.*, 54(13), September 2010.