DATA COLLECTION THROUGH COOKIES: COMPETITION AT THE COST OF AN INDIVIDUAL'S PRIVACY

by

Nandini Tripathy

I. Introduction

Cookies are files that may be sent to the user's hard drive by a website a user visits.[1] They allow an entity to uniquely identify users each time they visit the website, and allow the website to remember information about the user each time.[2] They help entities in tracking the browsing activities and preferences of users, sometimes even without users' knowledge.[3] Cookies can make a website more useful to users. Cookies are mainly of three types- Session cookies, which expire when a user's browsing session ends, Third party cookies, which are set in websites visited by a user not by the website itself but by other third parties, and Persistent cookies, which are active for a longer period than session cookies, which is set by the entity. The duration of such cookies, if used by entities, differs from entity to entity and is in certain countries restricted by law.[4] In the EU, persistent cookies cannot last longer than 12 months, as per the e-Privacy Directive.[5]

The algorithms an entity creates define how their cookies function and what data it will collect about a person. How their algorithms work is known only to the entity using the cookies. Machine-

[1] Steven Tyler Englehardt, 'Automated Discovery of Privacy Violations on the Web' (Doctor of Philosophy Dissertation, Princeton University 2018) <https://senglehardt.com/papers/princeton_phd_dissertation_englehardt.pdf> accessed 2 November 2021.

[2] Jonathan P. Cody, 'Protecting Privacy Over the Internet: Has the Time Come to Abandon Self-Regulation?' [1999] (48) (4) Catholic University Law Review 1183 <https://scholarship.law.edu/cgi/viewcontent.cgi?article=1430&context=lawreview > accessed 10 November 2021.

[3] Joshua B. Sessler, 'Computer Cookie Control: Transaction Generated Information and Privacy Regulation on the Internet' [1997] (5) (2) Journal of Law and Policy 627 <https://brooklynworks.brooklaw.edu/cgi/viewcontent.cgi?article=1433&context=jlp > accessed 10 November 2021.

[4] 'Different types of internet cookies' (*RocketLawyer)* < https://www.rocketlawyer.com/gb/en/quick-guides/different-types-of-internet-cookies> accessed 9 November 2021.

[5] Richie Koch, 'Cookies, the GDPR and the ePrivacy Directive' (*GDPR.EU)*< https://gdpr.eu/cookies/> accessed 9 November 2021.

learning or self-learning algorithms also play a role in how the data collected by cookies is used by the entity. We know lesser about the former aspect than the latter.

In both stages, the use of algorithms raises several important concerns regarding user privacy, which this paper discusses and proposes steps to mitigate. A privacy-friendly alternative to third party cookies, being developed by Google, will also be critically analysed. The paper concludes with the observation that entities tend to compete and profit at the cost of an individual's privacy.

II. Privacy Issues arising out of the Use of Cookies by Entities:

A. **Privacy issues arising from opaque aspects of the algorithm used by the entity in designing cookies and their cookie policies:** These problems arise when the cookies are programmed by entities and when data is being collected through them. We do not know much about them. They are:

i. **Vague Cookie Policies of Entities:** The purposes for cookie use are often very broadly expressed, and therefore not properly known to consumers. This is the root of many privacy concerns. Users do not know exactly what data is collected about them, and why.

ii. **Algorithmic opacity in terms of other tracking technologies used by entities:** On a slightly tangential note, there are data collection technologies other than cookies as well, which are used by entities for collection of data pertaining to individuals. One such example is widgets. They may function in a manner similar to cookies, and therefore contribute to raising similar, if not identical user privacy concerns. What these alternative technologies are and how they function is unknown, as they are not explained in the cookie policy of entities. They may complement cookies in user tracking and profiling. This is a form of algorithmic opacity which invades upon the privacy of individuals, in our view, and needs to be addressed along with cookies.

iii. **Algorithmic Opacity as to what data is collected in a cookie:** As consumers, if we were to open the file location where cookies are stored on our devices, we will not be able to read the data that has been stored in them, because it is encrypted. Therefore, we essentially

do not know what information is being collected about us. Further, we also do not know how the algorithm collects data- what its sources are.

iv. **C-Name Cloaking:** This is a phenomenon where a third- party disguises itself as a first-party by using the first- party's domain name as its website name. Here, the website URL shows the first party's name, but it is linked to a third party's domain name. "It misleads web browsers into believing that a request for a subdomain of the visited website originates from this particular website, while this subdomain uses a CNAME to open to a tracking-related third-party domain." It is essentially a practice involving deceptive programming, which is also algorithmically opaque, where a third- party is disguised or "cloaked" as the first party or main website which an individual intends to visit.[6]

v. **Persistent Cookies:** The duration of a persistent cookie is not standardised or known to internet users, as it largely depends on how it is coded, except in the EU, where a time limit is set. Persistent cookies have no set expiry date. We do not know when it stops collecting data, or even when it is used by entities, and the purposes for which such a cookie is used by entities. This can also be seen as a form of algorithmic opacity.

B. **Privacy issues arising from use of cookies because of how the data collected by entities through cookies is used, such as latent profiling of individuals:**. These privacy issues arise after the data is collected, and when it is used by entities. Explained below are the issues falling under this category

i. **User Tracking and Profiling by Third Parties and Websites:** Given the nature of user data including of a personal nature that may be stored and used by websites due to use of cookies, there are serious privacy concerns raised. The use by websites and their chosen third parties of tracking cookies can enable entities to compile long-term records of individuals' browsing activities. This is a huge privacy concern, considering that users are, more often than not, unaware of who is collecting their data, when it is being collected or is it ever not being collected, and what data exactly is being collected about them. This tracking is usually therefore done in a latent manner. The ePrivacy Directive and General Data Protection Regulation, read together, require that all websites targeting European Union Member States obtain, among other

[6] Ha Dao and Johan Mazel and Kensuke Fukuda, 'CNAME Cloaking-based Tracking on the Web: Characterization, Detection, and Protection' [2021] <http://www.fukuda-lab.org/publications/2021/DMF_tnsm2021.pdf> accessed 2 November 2021.

requirements, "informed" consent from users before storing non-essential cookies on their device.[7] Further, there is no standard understanding as to what constitutes "essential" and nonessential purposes for which data can be collected through cookies. What constitutes an essential purpose presently depends on the entity using the cookies. Moreover, the concern that data may be collected by entities through cookies for a purpose which is in fact not essential, under the guise of an essential purpose.

ii. **Over-processing:** Over processing can happen in the following ways:

a. Data collected through cookies may have a defined general purpose for collection. A user may consent at the time of collecting but he or she may not become aware when or how his or her data is actually being processed. It may so happen that a person wouldn't want his data to be processed at such a later time, but had no opportunity to stop this usage because he or she essentially did not know when it was processed by the entity. There is a significant mismatch between consent and processing.[8]

b. **Limitless storage of user information:** Data may be collected on a website server and never truly get deleted. Data protection laws mandate that data be stored for only as long as is necessary to fulfil the necessary purposes. Further, data may be indiscriminately collected irrespective of whether it is actually required to fulfil the purpose for which it was collected. Another tenet of data protection is violated in this way, that only the data which is relevant for the purpose should be collected.

c. **Place of Storage on the Device may be unknown to the user:** Cookie respawning is the process by which trackers store identifiers in multiple locations in the browser and use this information to reactivate cookies if they expire or are deleted.[9] This may be seen as another aspect of algorithmic opacity, deceptive programming and latent profiling of users. Cookies may be stored in several locations on the user's device, which may be unknown to a user, apart from where they are actually supposed and known to be stored.

[7] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1 <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN> accessed 9 November 2021.

[8] Use of Cookies- https://www.loyalsolutions.eu/use-of-cookies (last accessed on 8th November 2021)

[9] Englehardt, note 1.

d. **Deletion:** Whether deletion of cookies amounts to deletion of actual data stored by entities is also a question worth raising. Cookies are a vehicle for collection and transfer of data to the entity. Once the cookies transfer the data, entities have the data even if the cookies are deleted. This essentially amounts to creation of a copy of the data belonging to users with the entity, which may permanently remain with them.

iii. **End of a browsing session:** It may not be well known that a browsing session ends when the browser, and not any particular tab or website accessed through a browser, is closed. Individuals may unintentionally be subjecting themselves to additional tracking and profiling, even when they are not on a particular website. By not clarifying this aspect for its users, an entity is making its data collection operations more latent and independent of the consent of the user concerned.

III. Solutions to privacy concerns resulting from use of Cookies:

1. Algorithms should be mandated to be more transparent and readable to the end-user, to allow them to know at least, what data is collected in a cookie.

2. Entities should explain to users how their algorithm which processes the information after it has been collected works. This would help in reducing fears that data is indiscriminately collected and used, even if it is irrelevant for the purpose,

3. Users should get notifications not only when cookies are actually being used by the entity in order to collect data about visitors to their website, but also when it is transferred from the browser of the user to the entity's server, and further also when it is actually processed by the entity. This will help in reducing uncertainty as to how the algorithm works, and lso preserve a sound consent framework, by allowing a user to rescind his or her consent given earlier for processing of their data through cookies.

4. The cookie should be coded in such a way that deletion of cookies from a user's device also leads to deletion of the data therein stored on the entity's server.

5. Entities should explain when a browsing session ends so that a consumer is aware of the fact that he or she is subject to tracking even when he or she is not browsing on a particular website, and on another one.

6. The duration of persistent cookies should be regulated upon globally and their meaning explained by regulation. Else, they should be done away with, by regulation. Tracking and profiling of users must not be perpetual.

7. Technologies which work in a manner similar to cookies, and are used by entities complementarity along with cookies, such as widgets, should be described by entities. Individuals should be informed of their being used. Else, it will amount to the same privacy concerns as cookies pose, being posed by these alternative technologies. They should be regulated as well, as is done in the EU, under the E-Privacy Directive.

IV. Alternatives to Cookies: Federated Learning of Cohorts (FloC), being developed by Google

Federated Learning of Cohorts (FLoC) is a machine- learning algorithm based facilitator of targeted advertising, which is said to protect user privacy.[10] It is being developed by Google as an alternative to third-party cookies, which the company is in the process of phasing out. This section of the paper is about how FLoC will function, and a critical analysis of whether it will indeed be an effective privacy protecting-mechanism for data collection.

A. How FLoC works:

1. A FLoC service needs to be used by browsers.
2. This service creates a model of numerous cohorts, which are essentially clusters or segments of identical or similar browsing activity of several individuals. Each of these segments is assigned a number, which is shared with advertisers and websites. It is said that these segments are not based on the knowledge of any browsing history, and are created randomly. A cohort is comprised of several browsers. A cohort model resembles a honeycomb.
3. The browser then gets this cohort model from the FLoC service. As the user browses the internet, the browser uses a clustering machine-learning algorithm to find the cohort which matches its browsing behaviour. Browsing history is not shared with the FLoC service or with any third party. No user data is stored by the FLoC service.[11]

[10] Sam Dutton, 'What is FLoC?' (*web.dev*, 30 March 2021) <https://web.dev/floc/> accessed 2 November 2021.
[11] Ibid

4. Data of the users grouped in a cohort is so similar that one person's browsing data cannot be distinguished from another.
5. Websites and users can choose whether to opt-in or out of this system.
6. Advertisers can include code on their own websites in order to gather and provide cohort data to their ad tech platforms, which are companies that facilitate advertising for the advertiser. This information can be used by ad tech platforms to accurately recommend ads which are most relevant to the cohort data. This helps in targeted advertising.[12]
7. The clustering algorithm used to create the FLoC cohort model analyses whether a cohort may be correlated with sensitive categories, without learning why a category is sensitive. Cohorts that may disclose sensitive information categories will be blocked. When finding its relevant cohort, a browser will only be choosing between cohorts that do not disclose sensitive information.[13]

B. **Critical Analysis:**

1. This mechanism only replaces third party cookies. Cookies are not being completely done away with, so the risks from their use to user privacy still remain. The risk of C-Name cloaking still exists[14]
2. It is claimed that the cohorts and the model encompassing them are created randomly by the FloC service, with no knowledge of users' browsing history. Considering that cohorts are essentially categories of browsing activity of users, it is clear that they are based on browsing activity. Further, the browser selects the relevant cohort which it may become part of. The cohorts may not be created randomly.
3. It is said that no browsing data is shared by the FLoC service with the browser vendor or any third parties. Browsing history of the user concerned need not be shared with the browser or browser vendor by the FLoC service as browsing history is stored by the browser, as a user browses the internet, unless the user changes these default settings. Also, the fact that the browser selects the cohort which it falls under in the model created by the FLoC service likely entails sharing of browsing history by the browser with the FLoC service. This feature may therefore not be reassuring to users, as it appears that the browser facilitates user profiling as

[12] Ibid
[13] Ibid
[14] Ha Dao and Johan Mazel and Kensuke Fukuda, 'CNAME Cloaking-based Tracking on the Web: Characterization, Detection, and Protection', note 6.

much as the FLOC service. What is meant by the fact that entities can code their websites to allow for such sharing to take place is not made clear.

4. Sizes of cohorts may be small, no matter how many browsers they include within them, because they essentially divide internet users into groups. If an unauthorised tracker starts with a cohort, he will have to differentiate one user's information from very few others within the cohort. This may prove dangerous considering the risk of browser fingerprinting and user identification.[15] Also, there are always differentiators between the activities of several users, even if they have similar browsing activity, which may aid in their individual identification by hackers.
5. If a machine- learning clustering algorithm is used to classify sensitive and non-sensitive information in the process of making a FLoC cohort model and finding a cohort which best applies to the browsing activity, it may be that the algorithm itself knows certain sensitive information about users. Else, it would not be able to correctly classify this information and block that which is sensitive. The process of machine-learning necessitates storage of data. This may be a sifting mechanism, to aid in deciding which data to block from further processing and which to send through the FLoC system. The algorithm may store sensitive and non-sensitive information, possibly temporarily, to function better. The first-party website which uses the FLoC system and learning algorithm also may store it indirectly.
6. The claim that the algorithm classifies sensitive and non-sensitive information and blocks sensitive information without learning why it is sensitive, appears illogical. Whether data is sensitive or not depends on the data itself. The algorithm only needs to, problematically, know the data to categorise it. Fundamentally, there need be no other reason why particular data is considered sensitive, relative to other data. In case information is erroneously classified by the algorithm, how it is to be addressed is not clarified.

   Which parties qualify as third parties has not been explained.

At this stage, the FLoC framework can potentially create new privacy problems and exacerbate existing ones, rather than effectively mitigating them. Many aspects of its functioning are unclear. However, this is a rapidly- developing framework, and needs to be analysed in greater depth as it

[15] Bennett Cyphers, 'Google's FLoC Is a Terrible Idea' (*Electronic Frontier Foundation*, 3 March 2021) <https://www.eff.org/deeplinks/2021/03/googles-floc-terrible-idea> accessed 2 November 2021.

develops, to cure any potential privacy lapses in the framework which may defeat the objectives with which it was made. This framework has the potential to be more privacy-friendly. However, at this stage, if entities were to use this framework, they can, in our view, compete and profit at the cost of a user's privacy.

IV. **Conclusion**

Cookies are both advantageous and disadvantageous in terms of impact. Whether their impact is positive or negative depends on how ethically they are designed and used by entities. We have come to infer that in today's global data-driven economy, technologies used by entities to collect data regarding their individuals pose inherent and multifarious privacy risks to individuals. Owing to rapid technological development, these risks may only be mitigated, and not fully eliminated, regardless of how sophisticated, or user-centric technology evolves to become in due course of time. Entities may always compete with user data, and profit therefrom at the cost of a user's privacy- the gravity of which it may never be possible to accurately gauge, but the risks to privacy can and must be mitigated.

By:

1. Ankita Thakur (ID No.: 21011808)
2. Nandini Tripathy (ID No.: 21011862)
3. Sayee S. Tandale (ID No.: 21011804)