Catch me if you can: Achieving complete internet anonymity using open source technologies

Roshan Bangar[1], Varun Narkar[1], Atharva Phand[1], and Jay Lohokare[2]

[1] *Computer Engineering, College of Engineering Pune, Maharashtra, India*

[2] *Computer Science, Stony Brook University, New York, USA*
**Email: phandah18.comp@coep.ac.in*

Abstract

Search engines aggressively capture user data for targeted advertisements at the cost of user privacy. We present a framework for complete internet anonymity that blocks search engines from tracking user data based on cookies, IP address and search queries . The framework consists of a browser plugin to enable anonymous search by masking browser cookies, a VPN service to enable masking IP address and finally Natural Language Processing models for creating multiple queries to mask user data. Based completely on open source technologies, the framework serves as the first of its kind open source internet anonymity tool.

1 Introduction

Every day on the internet, users keep getting recommendations for various things, come across more relevant content, advertisements related to what they like and so on. This is known as tracking or online tracking, or user tracking. When users are surfing the web, they want their privacy to be intact. Nevertheless, websites keep tracking people when they browse the web. They collect data such as name, geographical location, what are people doing on the web, and many other things. This information is used in many ways by companies and trackers all around the world and can cause harm to internet users if used in the wrong way. Hence we see technology has its pitfalls as well. With every click on the internet, information such as IP Address, the URL of the previously loaded web page can be captured. However, tracking is useful as well. For example, websites often use first-party cookies to remember user's language, layout preferences, or the contents of a shopping cart.

In this paper, we present 'Catch me if you can', a framework for achieving complete anonymity while using search engines, so they can't profile users via cookies, IP addresses and search queries.

We have structured the paper as follows. Section 1 describes web tracking in some depth. Section 2 focuses on related works on internet anonymity and gives a background on the same. Sections 3 and 4 explain the framework in detail. Finally, we conclude with a view on the user journey using the anonymity tool.

Web tracking means tracking a person on the internet by various means when he/she is browsing a certain website. Techniques of web tracking are as follows.

1. Cookies: These are small blocks of data created by a web server. When someone visits a website, it sends a cookie back to the device which is then stored in a file inside the browser. The purpose of cookies is to help websites keep track of user visits and activities. It helps in improving the user's browsing experience. Other cookies like session cookies are used while navigating a website, they are stored temporarily in memory. Persistent cookies remain on the device till their expiration, they are used for tracking and authentication. Examples are login information, preferences such as sports versus politics, customized advertising, recommendations based on past views.

2. IP address: Every device has an IP address that needs to be exchanged with the websites one visits. This is to ensure reliable data transfer between them. Due to this handshake, IP address trackers can easily hack into a user's geographical location, monitoring the user's viewing pattern, etc.

2 Related Works

We now discuss various references on internet anonymity.

2.1 *Literature Review*

Jinbao Wang, et al. [1] introduces differential privacy combined with k-anonymity in their paper. They focused on query masking in location-based services (LBS). A k-anonymizer produces k-1 random but relevant queries along with the original search query. The system lies in the user's mobile device. The aim of this research was to make the LBS provider unable to distinguish the actual query interest and the k-1 dummies through probabilistic inference, and differential privacy is adopted to achieve this goal. In paper [2], Yabo Xu, et al. introduced a technique to achieve and maintain the user's anonymity: a user pool in between the web service and ¡d, q¿ (personal information, query). First, the user is anonymized on d through the user pool, then generalizes it to some d' and

then send ¡d', q¿ to the web service. The user pool is supposed to track the online users who issued queries during a specific time interval and anonymize their personal information d in an online fashion. But this paper lacks in covering the query component which contains user information. In paper [3], authors Oguri, Hidenobu, and Noboru Sonehara propose that the approach to k-anonymity suffers from data loss in the case of databases and computational time complexity increases as well. In paper [4] Eckert, Claudia, and Alexander Pircher summarize the main results of the "Anonymous-project". They explain the problems and limitations of current anonymizing services and present the new services. El Emam, Khaled, and Fida Kamal Dankar in their paper [5] through a simulation, evaluated the re-identification risk of k-anonymization and three different improvements on three large data sets. In [6] Regner, Tobias, and Gerhard Riener discuss the impact of anonymity on the online sales and revenue. Why online anonymity is precious to the consumers and methods to acquire online anonymity successfully are discussed in this paper. [7] Wang, Jinbao, Zhipeng Cai, and Jiguo Yu discuss the client based personalized k-anonymity algorithm and the performance of (CPKA). In paper [8], authors have studied behavior-based user tracking in real life. They have designed a technique using Naive Bayes classifier which is around 88 percent accurate as mentioned.

2.2 *Background*

The World Wide Web is in use 24/7 for personal and professional work. A huge amount of user data is leaked every second, stored on servers of the companies. To tackle such threats, we introduce "internet anonymity" - through which users enjoy a safe, secure search to an untrusted web service with their anonymity preserved. A user's identity can involve geographical location, name (age, gender and other basic details), dislikes, interests, and so on. It also involves tracking which sites the user visits, how often, and related data. To keep user privacy intact to some extent, we combine the following methods in our internet anonymity tool.

1. Query masking: In today's data driven world, to ensure data privacy, protecting confidential as well as personal information of an individual is very important. When the user enters a query on the search engine, directly or indirectly some of the user's information is exposed to the outside world. Lot of information is useless is the basic fact behind the working of query masking. When something needs to be conveyed to the outside world, the world is fed with a lot of useless information. Natural Language Processing (NLP) techniques are used to generate similar and relevant queries. The original query is used as the input and the n - 1 similar queries are generated. The original query is buried in this set of similar queries. In this way, query masking is implemented to make query anonymous.

2. IP masking: IP address helps get a device's approximate location and possibly gives away information about the user's online activity. IP masking is a method of hiding real IP addresses. It protects users from identity theft and conceals their location, which can otherwise be easily traced. The most basic way to hide the IP address is to use the Virtual Private Network (VPN). Another method is to use a web-based proxy server wherein we send data to this server, which sends it to the internet. The proxy server acts as a middleman between the user and the internet, hiding the real IP address. Open-source tools like Tor can also be used. It uses the mechanism of routing all the traffic through the tor network, also known as the tor circuit. Finally, it spits the traffic through the final exit node, making it difficult to identify the actual IP of the user.

3. Cookie masking: When a user visits shopping websites, shopping cart information is stored in cookies. Cookies also hold site preferences and session tracking information. When a third party tracks cookies, this valuable and important information is exposed to them. Cookie masking can be implemented to make users anonymous. Cookie masking can be implemented with the help of techniques like using Tracking Prevention Lists (TLPs), disabling third-party cookies, enabling private browsing mode or using Do Not Track (DNT) headers.

3 System Components

System components of internet anonymity tool are browser plugin, flask [9] server, and proxy server. Cookie masking, IP masking, and Query masking is implemented on each search query given by users for making users anonymous, which is the important aim of our anonymity tool.

• Flask server: An important backend component of the anonymity tool is flask server, which implements the vital task of query masking. When a user types a search query, it is first given to the backend flask server. Flask server uses NLP techniques to implement query masking, i.e. similar queries are generated. These similar queries along with the original query are then sent to the proxy server with the help of a secure SSH channel. We selected Python Flask to create our APIs due to it's ease of deployment and development.

• Browser plugin: IP masking is another crucial aim of the anonymity tool which helps in not knowing the IP address of the user. An intermediate proxy server is selected by browser plugin to hide the user's IP address. Tracking Prevention Lists (TPLs) are used to disable third party cookies. Adding a do not track (DNT) header allows users to request websites not track them. In this way, cookie masking is implemented to ensure user anonymity.

• Proxy server: After receiving response from search engine, intermediate proxy server sends it back to the user through the same secure SSH channel mentioned above. It acts as an important bridge between user and search engine. It does not let search engines know the actual IP address of the user, which is the important goal of our anonymity tool. In this way, it plays a vital role in query anonymity.

4 System Operation

The proposed tool presents a search interface to users in the form of a browser plugin. The users can enter their search queries in this plugin, and the framework will serve appropriate results to the users and also prevent the search engines

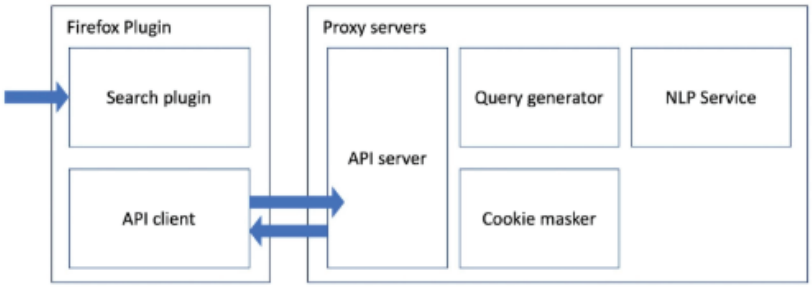

Figure 1: Web interface for control center operators

from capturing user data. Once the query is entered in the browser plugin, the plugin makes an API call passing the query as API parameters. The API call goes to the framework's backend where the query generator generates a set of queries similar to the input query. The API server then makes a call to the search engine using all the queries. Once the responses are fetched, they are sent back to the user. The API server ensures that the search engine doesn't see any user cookies, and uses SSH tunneling to use multiple IP addresses while making the search calls. This ensures that the search engine doesn't have any data points to track the users. Thus, the framework enables users to use search engines without compromising on data privacy.

This framework is completely non-intrusive to users. The tool can be deployed on any browser that supports plugins (Chrome, Firefox, Edge, etc) and it has no effect on user experience using the browser as it runs in the background.

5 Pilot

To study the tool we proposed, we ran a pilot where we asked 6 candidates to use our tool for internet search. We asked the candidates to search 10 queries on Google search using our tool. We then asked them to provide us the list of queries they ran and then the Google activity data for the time they ran the queries. To validate that the search engine was not able to profile the users based on the queries they ran, we checked if the queries were present in their activity data. None of the queries were present in the activity recorded by google, thereby proving that our tool ensures anonymity.

6 Conclusion

In this paper, we introduced a tool for achieving internet anonymity by leveraging data masking and Natural language processing. We present the architecture of the framework, while demonstrating how it tackles the data capture techniques used by search engines. However, the use of VPN or torr browser may sometimes slow down the internet speed due to long circuit formation. Also, blocking third-party cookies sometimes make websites crash or not work as expected. Using k-anonymity makes it slightly more costly in terms of time because k queries will be searched instead of only one. In spite of the limitations stated above, the proposed anonymity tool is successful in preventing search engines from profiling users for targeted advertisements.

References

[1] Jinbao W., Zhipeng C., Yingshu L., et al: 'Protecting query privacy with differentially private k-anonymity in location-based services'. Personal and Ubiquitous Computing, March 2018, 22, (3), pp. 453–469.

[2] Yabo X., Ke W., Guoliang Y., Ada W.C. F.: 'Online anonymity for personalised web services'. Proceedings of the 18th ACM conference on Information and knowledge management, 2009.

[3] Hidenobu O., Nobori S.: 'A k-anonymity method based on search engine query statistics for disaster impact statements'. 2014 Ninth International Conference on Availability, Reliability and Security. IEEE, 2014.

[4] Claudia E., Alexander P.: 'Internet anonymity: Problems and solutions'. IFIP International Information Security Conference. Springer, Boston, MA, 2001.

[5] Khalel E. E., Fida K. D.: 'Protecting Privacy Using k-Anonymity'. Journal of the American Medical Informatics Association, 2008, 15, (5), pp. 627–637.

[6] Tobias R., Gerhard R.: 'Privacy is precious: On the attempt to lift anonymity on the internet to increase revenue'. Journal of Economics and Management Strategy, 2017, 26, (2), pp. 318–336.

[7] Jinbao W., Zhipeng C., Jiguo Y.: 'Achieving Personalized k-Anonymity-Based Content Privacy for Autonomous Vehicles in CPS'. IEEE Transactions on Industrial Informatics, 2019, 16, (6), pp. 4242–4251

[8] Christian B., Dominik H., Hannes F., et al: 'Tracking users on the internet with behavioural patterns: Evaluation of its practical feasibility.' IFIP International Information Security Conference. Springer, Berlin, Heidelberg, 2012, pp. 235–248.

[9] 'Flask Documentation', https://flask.palletsprojects.com/en/2.0.x