Project Proposal

(Due: Sep 25)

5

Group Members		
Name	SBU ID	% Contribution
Jay Lohokare	111492930	33.3
Revati Damle	111461639	33.3
Nittin Aggarwal	111401512	33.3

Who Am I ?

1 Problem:

Targeted advertisements are on the rise as social media companies are aggressively trying to increase their revenues. Even when there are multiple ways to prevent search engines/social media websites from knowing who you are, its inevitable that these websites get to create a data profile for an user. Search queries are one of the most rich source of data that can be used to profile a user. IP address, cookies are also used to extract demographic and browsing history information.
Privacy and non-traceability are an important power in todays world. With numerous websites and applications aiming to capture data, its getting difficult to remain untraceable. In this project, we aim to create a tool that not only implements basic features of cookie/IP untraceability but also to anonymize data present in search queries. We also provide a thorough analysis of the outcomes of such anonymization by studying the resultant advertisements (If any).

2 Context :

Search engines are using various methods to capture data from users, to generate data profiles of users. Browsing history, location history, shopping cart, SMS, emails, call logs are call captured by these internet giants for targeted advertisements. For most of the users, capturing this data is objectionable, as no one likes their privacy being invaded. Not only is confidential personal information a part of the data collected by such websites, but due to rise of hacks and attacks on servers of such internet websites, this data collected is at risk of being exposed to public.

The data profiles created by internet companies from search queries often tells details about a users work, personal and social life. For example, searching for a costly restaurant indicates high social status. Search engines can target advertisements of similar hotels to the user. More importantly, the search engine now knows the social status and buying power of the user. These search queries reveal multiple such small details, together which can be used to construct a complete profile for the user. Hence, masking the search queries is often important. The so-called incognito mode of browsers is hardly of any use. The internet companies already have mapped IP addresses with users, and can easily know what user is trying to connecting without having access to the cookies.

The concept of **Plausible deniability** is a very old one - Its ability of people to deny certain events due to lack of evidence confirming their participation. We base our study on this concept - to build a tool that ensures privacy by ensuring plausible deniability for users. This plausible deniability is achieved by IP masking, Query masking and cookie masking.

3 Approach :

When connecting to internet, internet website/servers can see the public IP transmitted online and keep a record of these IPs in log files, to map a user profile to IPs. Removing these IPs will make it difficult for internet websites to trace a user, ensuring privacy. Though the IP address can not be completely removed, it can be masked by using various methods:-

1. **VPN**

In this method, the user machine becomes part of a virtual network that can belong to an geography, thereby masking the actual users IP.

2. **Proxy servers**

In this method, users connect to a intermediate server (Between users and internet). This intermediate server reroutes the user requests to the internet, making it appear that the requests are coming from the intermediate server, thereby masking the user.

There are multiple browser extensions for both - VPN and proxy. We plan to implement both these methods in the tool we develop, so as to make it one tool to do it all.

Hiding Cookies of users from the internet is well documented process. We wont provide details for how we achieve it in this proposal.

The process of masking the user queries is what we want to emphasize on in this paper. There has been some work done in past to achieve such masking. However with dawn of deep learning, the extent to which we can achieve this masking has highly increased. We use the concept of Plausibly deniable Query set, which is a set of queries similar to the user queries that can be used to mask the users identity, but get similar results. We construct a new query from the input query using a deep learning model (**LSTM**). We have covered various scenarios like Sequentially edited queries (Where user queries are linked to past queries), Location based queries (Where queries need to be mapped with user location), Time correction queries (Where current query is after previous query gave incorrect results). We explore various approaches to constructing the query set based on clustering and using similarity scores.

Unlike previous studies, we leverage the growing computing abilities for browsers and the increased internet bandwidth to achieve a better query set. We also include a notion of background random queries (A set of random queries run in background to confuse the advertisement targeting algorithms), and analyze efficiency of such an approach. By achieving a multiple countries based VPN, we also evaluate use of such rotating IP approach for queries.

Finally, we provide analysis of the performance of this tool we developed (Qualitative and Quantitative)

4 Evaluation:

The evaluation of the tool that we develop has different criteria, but all aim to achieve the same goal that is to test the relevance and efficacy of the tool brought about its presence or absence. The tool should not disturb the search results much that were directed from search strings. Secondly, the search strings if logged by an adversary should not reveal the true identity or traits that can help in profiling users.

For these metrics, we need to test the following,
1. Relevance of search results in presence and absence of our tool.
This can be seen by comparing the search results before and after usage of tool for same set of queries.
2. The diversity of search queries which can be pattern based / timestamp based. This can be seen by using clustering approach to see the number of clusters formed by the transformed queries. If the number of clusters are large then the diversity is high in the search queries.

We also need to test whether randomization of queries which is an age old technique used to betray the data miners, is required or not. So, the metrics can be verified without and with randomization.

Both of the above metrics need to be verified and tested without using VPN and with VPN.
Another metric that has to be considered is the size of the model that needs to be bound to the browsers. The query latency also needs to be almost similar to those without the tool so that the tool does not hamper user experience.

5 Scope :

The main idea is to test how efficiently we can eliminate the inherent traits that the queries can divulge related to the end user. The tool developed should be able to transform query and use masked query to generate close to previous results in addition to achieving plausible deniability. The queries generated by the tool should be diverse enough so that randomization methods are not required.

We will try to keep our project timeline aligned with the guidelines provided by the professor. We intend to develop our initial model by project update 1. We intend to integrate the VPN functionality by project update 2. Post the project update 2 we intend to follow our evaluation plan to post results and project and document them by our Demo Date.