Literature Review

(Due: Oct 9)

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 Related Project Work:

Query-based web search is an integral part of many peoples daily activities. But this search process is used to identify the individuals by the internet giants. Government requests for such logs increases the concern. To address this problem, [8] proposes a client-centered approach of plausibly deniable search. 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. As the paper suggests, just anonymizing users is less effective, as the query itself contained identifying information. The paper address this concern by generating a set of cover queries for each real query. This eliminates timing based attacks to discover the real query. While not completely hiding the real query, if done properly it supports a notion of plausible deniability. While it is possible that an individual may have issued the actual query, it is equally plausible that they issued one of the generated cover queries. This places a burden of proof on the user of the query log when trying to imply something about an individuals interests. 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, as introduced in [8], 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). 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.

Google announced in 2007 that it will anonymize server logs that it collects every 18-24 months to protect the privacy of it users. Prior to this, google retained server log data in its original form indefinitely, which made it possible for anyone with access to those logs such as government agencies possibly gaining them through legal processes to potentially track queries back to users. Amongst the information that was stored was IP address and cookie that can help the onlooker to reach the user even if his connection changes. But this project does not anonymize the query on the fly allowing Google itself to draw conclusions on profiling the end user. That year AOL also released some anonymous data where a woman was identified based on cookie data from the captured logs. Google as of 2007 did not find a way to anonymize cookies. The Search history stored at ones computer and ISP might again reveal the identity of the user and become target for Ads. So, a solution starting from ones browser is essential. Although there are arguments against anonymization but they also forebear the need for privacy of users. The paper commands that there can be potential loss of data and data ownership [1].

[2] provides two solutions to the problem of web search anonymization. One is using a client side software to inject noise queries and second is using a network of relay servers to hide the source. Our solution is inspired to take into account both the techniques by means of using a VPN server to make the queries anonymous and also by generating canonically similar queries as original ones without giving away any identity information that can be derived even after removing the IP information and cookie information. These 2 constraints are not covered by [2]. The objective of machine learning is to extract useful information from data, while privacy is preserved by concealing information. Thus it seems hard to reconcile these competing interests. [3] follows from AOL revelations and provide mechanisms by which we can not deduce the users from the search query logs too. This paper suggests the use of differential privacy to anonymize a query log. Differential privacy is the state-of-the-art approach which provides a strong privacy notion. Differential privacy provides guarantees that every individual user in the datasets would not be identified. Unlike k-anonymity, differential privacy does not make assumptions about the amount and scope of an adversarys background knowledge. The first mechanism is that the queries can be encrypted using a hash that can be decrypted if one has sufficient number of instances of that query from multiple users. The second mechanism is that the users can be divided into multiple sessions to disassociate the users from themselves. Both these methods again neglect the fact that there might be other levels where the search anonymity might be compromised.[4] provides an interesting way by which we can test whether personalized search can be done even after anonymization of user profiles is done. Usage of VPN to promote anonymity in addition to query obfuscating might not work as proposed by various organizations as Nortel as it might limit the user to certain sites as many sites these days can detect a user being operating behind a vpn and they prevent or block this access and consider it as default on IP restrictions.[5] tells us how the web search engines build user profiles to provide personalized search results to the end user based on persons past search keywords which breaches users privacy.[6] and [7] both discuss about ways in which the search engines as well as government agencies or adversaries who can lay their hands on the search query logs can trace back to users and target them appropriately. All these efforts are directed in order to identify the information that the users are revealing in hands of adversaries by means of their search queries which tend to reveal more than they should. These efforts propel for solutions on the client side to hide away the information such as IP address, gender revealing keywords, age oriented keywords etc.

2 References:

[1] Search Query Privacy: The Problem of Anonymization
[2] Web Search Query Privacy: Evaluating Query Obfuscation and Anonymizing Networks
[3] User 4XXXXX9: Anonymizing Query Logs
[4] Anonymizing User Profiles for Personalized Web Search
[5] Web Search Personalization by User Profiling
[6] User Profiling Based on Keyword Clusters for Improved Recommendations
[7] Modified Approach of User Profiling from Search Engine Logs
[8] Providing Privacy through Plausibly Deniable Search
[9] Anonymizing Query Logs by Differential Privacy