Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.8 Issue 1 Page No 31-40

Privacy Protection in Personalized Web Search

[1] Krishan Kumar, [2] Mukesh Kumar Gupta, [3] Vivek Jaglan
[1] Dept. of Computer Science & Engineering
[2] Dept. of Electrical Engineering
[3] Dept. of Computer Science & Engineering
DPG Institute of Technology & Management
krishanchhillar@gmail.com, mkgupta72@gmail.com, jaglanvivek@gmail.com

Abstract— Web Search Engines are the tools which provide information to the user based on queries entered. Search engines prepare a user profile from search history and information provided. Rich user profile can provide personalized results to the user but at the stake of privacy. Unsolicited advertisement, disclosure of sensitive information and identity are enormous challenges. Rich profiles can attract malicious interest and may lead to reveal personally identifiable information. In this paper we are going to study various techniques which can improve privacy of the user. We will study various privacy enhancing techniques and classification of tools which are used to enhance user privacy. We have proposed a system which can maintain anonymity of individual and reveal partial information to get benefits of personalization as well. The proposed system is using well-defined protocols and implemented on a proxy level with privacy protection.

Keywords—Personalized web search, Private Information Retrieval, Privacy, Anonymity

I. INTRODUCTION (*HEADING 1*)

Most of the modern Web Search Engines like Yahoo, Google collect, analyse and store user information. They performed these tasks to track, exploit personal information of user and search behaviour to provide better customized search results. The main purpose of storing this information is to provide relevant and useful content to the user to improve the effectiveness of the search results. So, personalization is a convenient way to access user information, but they abused this personal information in different ways like digital discrimination and targeted advertisement.

Nowadays, user data is considered as most important asset for any organization because firms are mining this data for value creation. By exploiting this data, organizations can understand the preferences and valuations of their users and customers, which facilitates the firms' relevant advertising, efficient targeting, and personalized services. Quality and quantity of the customer data also play very important role in value creation for the organization. Therefore, organization is investing a lot of money to collect, store and analyze consumer data.

So firms are collecting different personal and vast information about the users by hook or crook. This huge personal information collected from different sources is not only facing privacy threat from within the organization which collected this information because of lack of protection arrangements but also facing threats from the potential data breaches as well. The number of exposed data records and data breaches are increasing daily. In a survey, 79 percent of US people reported they don’t feel confident about the firms that firms will keep that their personal information as mentioned in the terms and conditions. Firms are admitting mistakes and responsibilities if they misuse the personal information.

To fight the increasing threat of privacy invasion, users are looking for more solutions to protect their privacy [1], [2]. Privacy protection can be divided into two categories-regulations and self-regulations. Several regulations have been proposed and implemented to some extent to mitigate data breaches and privacy concerns in recent times, but results are not clear. So, it is up to end users only to adopt self-regulated approaches to protect their own privacy. We end users are proactively fighting against the privacy invasion threat and not showing trust in the privacy laws and self-regulations of organizations. As per a survey report, over 55 percent people in America are using privacy tools to safeguard their personal information rather than believing in laws and regulations for the privacy concerns [3]. 41 percent people in Europe give false personal data and information while signing up for the online services and products. So privacy enhancing tools (PETs) are IT products which are used by end user to protect personal information by removing or minimizing personal data to protect the privacy. The industry has witness growing number of PET users on a rapid rate with time. The daily users of Tor [4], [5] has crossed 2 million mark. However, the use of PETs is not beneficial for the firms as they cannot collect quality data from the users and hurt their profits.

Correspondence to: Krishan Kumar, Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur
E-mail addresses: krishanchhillar@gmail.com

Now a days, Web Search Engines (WSE) are very common to cater day to day life information needs. Most of the information retrieval techniques are implemented to search new information but Stuff I've Seen (SIS) [6] is a system for personal information retrieval and re-use. A lot of knowledge work is required to find and re-use previously accessed information. The system work in two steps, in first step unified index is created for the information which the user has accessed. It may be from email, web page, document, appointment etc. In second step contextual cues can be used to search the information in search interface. Web Search Engines work on a very short query and find relevant information using anchor text, user history, link and popularity cues. This process involves integrating and re-using previously accessed information also. It has been observed that 58-81 percent of web pages accessed were revisited during the web surfing[7]–[9].

First benefit of SIS is all the information is indexed in one source regardless of where from information is originated. Second benefit is user can revisit pages from the information saved in SIS and can easily avoid WSE which will improve privacy of user. There are some more personal information organizers similar to SIS are also available in the market like- Haystack, MyLifeBits, Enfish Personal, PC Data Finder, 80-20 Retriever and Scopeware etc. which indirectly improve privacy of user.

Hide-n-Seek (HS) is an intent aware privacy protection plugin for personalized web search [10]. The system works on link-ability aspect for the privacy protection of the user. Link-ability means multiple queries can be linked with a user. It can reveal detailed information about the user behavior and background and can be collected by the internet service provider. Although true identity of the user cannot be revealed but privacy is invaded/breached. So, HS is a plugin used with web browser to protect link-ability aspect of the privacy of users. Its countermeasure privacy breach by search log-based personalization techniques adopted by search engines. It submits several cover queries in addition to the true query to hide actual search intentions of the user. The plugin discards the search results generated for the fake queries and re-rank the results against the true query only. TrackMeNot[11], [12] and GooPIR[13] are also developed on same notion of generating fake queries to hide the general intent of the user which is known as obfuscation, noise addition or perturbation.

UCAIR (User Centered Adaptive Information Retrieval)[14][15] is a personalized search toolbar. UCAIR works on the concept of client-side personalization and embedded in web browser as a plug-in. Unlike most of the web search engines, it does not store any information on server-side and works on the implicit information collected from the user to prepare results. It stores user interaction history log, performs implicit user modeling based on past queries and click thorough results, can modify queries based on implicit user modeling and re-rank the search results. Tor [4], [5] is a circuit based low latency based communication solution based on anonymity.

Congestion control, perfect forward secrecy, directory servers, configurable exit policies, integrity checking, design which support location-based services are the advantages of second-generation Tor over its previous version. It provides tradeoff between anonymity, usability and efficiency with very little coordination and synchronization between nodes. It does not require any kernel modification and works with real world internet. With second generation Tor, many TCP streams can share one circuit. Beside this, no mixing, padding, or traffic shaping is required.

Personalized Web Search (PWS) is good way to get better web search results which need collection and aggregation of information about the user to be more effective which pose severe privacy infringement threats for users. It is found that if personalization is performed on the client machine, better results can be achieved than the existing server side personalized web search in terms of privacy protection. [3] "One size fits all" is a big drawback of existing search engines as they are not tailored as per the privacy need of individual users. Search engine record individual search query logs, location information, click through history, user cookies, browsing history, IP address and conduct user profiling. The collection of personally identifiable information (PII) by the search engine about user is considered as tracking. This information is collected from the user while interacting with web search engine to provide personalization and user profiling but at the same time web search engine use this information for targeted advertisement to get monetary benefits and improving search quality. But once the information is revealed individual's privacy can be compromised because it is no longer under their own control, how and by whom it is used. Although some web search engine organization online publish the privacy policies about their practices for the sake of public knowledge. But these policies are full of legal and technical terms which are difficult to understand for the users. most of the users are worried about monitory of their activities. Some users try to avoid this monitoring by using tools to anonymize their queries and by rejecting cookies. users register themselves on web search engines and enhance services by personalizing and customizing as per their needs. beside this user require sides to garner data to their end.

To fill the bridge off conflicting requirements we developed a system which manage anonymity while sharing information with the web search engines. This web-based framework takes care of privacy concerns while using web search engines and balance personalized web services and privacy concerns. In this system masks are used as anonymity barriers between user's private data and web search engine. it also controls the information flow between web search engine and user. Mask act as a filter which prevents exposure of user's information and allow service personalization. It does not allow third parties to create user profiles based on click-through and privacy issue at data collection are well addressed.

II. RELATED WORK

Anonymity of Twitter users is analyzed [16] to check user anonymity and correlation with the sensitive content. It was observed that the people were supporting, fighting, sharing and discussing on the topics like sexual orientations, marital and relationship issues, health related issues, personal experience and feelings, social anxieties, depression, suicidal tendencies and disclosing their own. Anonymity can provide them an opportunity to solicit support.

Correspondence to: Krishan Kumar, Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur
E-mail addresses: krishanchhillar@gmaild.com

In the recent times, users are providing their personal information to get excellent web services. This personal information generally contain name of the user, contact number, address , social, IDs credit card numbers etc. privacy of the user can be compromised, because the web services can provide it to some third parties which may not be obliged to keep the privacy protected. Companies generally implement anonymization or de-identification techniques on the user’s data to keep the privacy intact. anonymized data cannot be associated with individual users in any way[17]. anonymization is a way of converting open personal information of the user into aggregated data. Few techniques of anonymization are suppression, encryption, generalization, and perturbation. These techniques are combined to make the data anonymous. Data anonymization models includes k-anonymity[18]–[20], l-diversity[21], t-closeness[22], b-likeness etc. But all the techniques are implemented after storing information about the user. We are working on a system which does not allow web service providers to store user personal information.

Now Nowadays privacy can be breached in many ways like hackers can steal data from email, computers, user groups, and online service providers can also steal habits and user activity. Service providers also garner personal information about the user to personalize websites which also create big concerns in the minds of users. it may vary user to user how much privacy they want to give up when they're making their personal information publicly available. Beside this it is totally different how much information they reveal while interacting with web search engine. One more question how much information users are interested to obtain better services. It may vary person to person how much privacy they want to protect, or we can say privacy can be determined individually. We cannot decide privacy needs which may fit to all users.

III. LEVELS OF PRIVACY PROTECTION IN PERSONALIZED WEB SEARCH

Everyone has different requirement of privacy protection so the level of privacy protection can be decided as per individual need and there is trade-off between personalization and privacy protection. In the given below table, four levels of privacy protection in PWS as discussed.

A. Level I: Pseudo Identity

The user identity ID(U) is not used directly rather, a pseudo-identity IDp(U) is created which contains less personal information about the user and used.

TEXT (N, i)can be aggregated according to IDp(U) at server side.

They safeguard identification and classification of user. We can map pseudo profile with user information like queries and click through. We can exploit user profile to Personalized Web Search. AOL replaced IP addresses of users with a pseudo Id in August, 2006 User log release.[23] New York Times Reporter identified A lady in Lilburn, Georgia with this log.

B. Level II: Group Identity

a) Some users can create a group and identity for the entire group is treated as single user identity ID(U).

b) The description of information needs TEXT(N; i) for the users of group is aggregated to ID(U).

In this technique, a proxy for a group of users is created and users of whole group communicate with web search engine through proxy. So the identity and descriptions of user information need mixed with group users and it is made very difficult to identify individual user.

C. Level III: No Identity

a) The user identity ID(U) is completely hidden from the search engine.

b) Information needs TEXT(N; i) for the user are also not be aggregated on the search engine side, not even at the group level.

Here, user profile can be kept on local machine and personalization of search results at local user personal computer by re-ranking the results. Anonymous networks like Torpark are used to communicate with web search engine.

D. Level IV: No Personal Information

a) User identity ID(U) and information need TEXT(N) of the user are not provided to the web search engine.

To achieve level IV of privacy protection, cryptography techniques can be used. For example, the user does not send any query directly to the web search engine, but it sends the query to the trusted third party and the third party performs the search operation on behalf of client and sends back search results to client. Government agencies also can force search engine companies not to store any sort of data which can avail level IV privacy protection to the user.

IV. SOFTWARE ARCHITECTURE FOR PERSONALIZED SEARCH

On the basis of location of personally identifiable information about the user and way of exploitation for personalization, they divided the architectures in three categories as given below table.

A. Server-Side Personalization

Personal information P(U) of the user is stored and updated on server side using user specified interests (explicitly) and queries, click through history (implicitly) etc. Information collected implicitly is a richer way of data collection although both ways require an account to store information.

In current scenario, most of the personalized search systems like Google, Yahoo use server side personalization which have some advantages like resources of search engine(e.g. common search patterns, document index) can be used in personalization algorithms with no requirement of changes in client side software. Although Search Engines store and holds personally identifiable information of the user with his/her consent used for personalization and claim first level of privacy. But still many users are doubtful about the potential privacy threats by the search engine, so the adoption of this architecture is hindered by some users.

If the search engine replaces identity of the user ID(U) by a pseudo identity IDp(U), then it is possible to achieve level I

Correspondence to: Krishan Kumar, Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur

E-mail addresses: krishanchhillar@gmaild.com

privacy. Then search logs can be shared with corporate partners, public or researchers with pseudo identity and it is possible to achieve level I privacy.

In current scenario, level II of privacy cannot be achieved even if user communicate through proxy and use group profile technique. Because search engines use the user login ID and locale machine address (MAC address and IMEI numbers) to aggregate the user information not only the IP address. Level III and level IV privacy protection cannot be attained with this architecture.

B. Client-Side Personalization

Personal information of user P(U) is stored on client machine. Client side personalization agent makes changes in the query at the time of submission and re-ranks the search results as per the requirements of the user after receiving results from the search engine. In this architecture, user search behaviour, contextual activities, page viewed, browser bookmarks and emails also can be considered to personalize the user profile. Sensitive information, computation and storage of user profiles are distributed among various client machines and no more overhead of server but there is a drawback also that is algorithms available at server side for personalization cannot be used.

C. Client Server Collaborative Personalization

Here, all the information about the User profile is kept on the client machine only and it is not shared with the server. Client only submit contextual information extracted from the user profile with the query to the web search engine. Server performs personalization on the contextual information received from the user and provides the results to the user. Advantage of this architecture is to utilize the resources of server and drawback is only contextual information is not sufficient for good search results as compared to the profile. This architecture cannot provide a better level of privacy as the server can keep storing the contextual information received from the client and provides almost same privacy as obtained from server side personalization.

V. ANONYMITY

Meaning of anonymity is personas identity unknown or namelessness which originated from Greek word ‘anonymia’. In other words, we can say a person is non-identifiable, untrack-able, or unreachable. Other similar terms like identity, pseudonymity and privacy also come up with time.

First of all, we need to determine what type of anonymity service we are concerned with in personalised web search. As discussed by [24] Anonymity can be segregated into parts data anonymity and connection anonymity. data anonymity is de-identification of data, which means removing identity linkage or filtering any personal identifiable information from the data. So de-identification is an issue related to privacy preserving data mining and carried out on data sets [25]. Whereas connection anonymity deals with the issue of stealing identities during the interaction. So in privacy protection in personalised web search, we are concerned with connection anonymity.

Further, we can define three types of anonymity. First one is environmental anonymity which can be defined by external factors like number of users, diversity of the users and their prior knowledge. Second one is procedural anonymity which is defined by underlying protocol, intrinsic qualities and design of the system. It can be discussed to improve the privacy of the system. Third one is content based anonymity, which deals with mitigating contextual information in the data transfer.

Different levels of anonymity can be defined on the basis of properties, which can be defined as follows. Identifiability is the case of a possibility of revealing personally identifiable information or identity of user during the communication with the system when actual data exchange takes place. Traceability is the case for obtaining PII about the user by observing the communication context. Unlikability means that two or more items of interest are no more and no less related to each other than to a priori knowledge. Identifiability and traceability are almost identical terms, the way information is collected makes them different. Recognisability is a common term which means collection of PII, irrespective of how information is observed. Every user wants to hide identity, but at the same time, fraudsters should be held accountable for their actions, which creates a strong conflict of interest over unrecognizability. So we can say unlikability and recognizability are not required to provide anonymity.

Personally Identifiable Information (PII) can be of two types- direct PII and indirect PII. Direct PII does not need any external assistance to trace individual user whereas indirect PII needs third party involvement to access PII records. Whereas unresolvable PII cannot be revealed with a third party as well. This is highest level of anonymity. Pseudo-anonymity or Pseudonym is a single identifier which associate with an individual user. Public pseudonyms is a class of direct PII, whereas non-public pseudonyms can be considered as indirect or unresolvable PII. Group anonymity precludes unconditional recognizability and link-ability, which involves trusted and dedicated group manager who are responsible for removing or adding members and reveal identities in case of disputes.

VI. PRIVACY CONCERNS IN WEB SEARCH

Let us suppose a User (U) submitted a query (q) to a Search Engine (S) which returned Results (R={R1, ...,Rn}) back to the user. Then User selects Ri ε R and then, Search Engine provided the content of Ri to the user. In this whole process between user and search engine, user reveals potential personal information which can be inferred as:

1. User Identity: This can be IP address of user machine or personal user ID if user has registered account with search engine.
2. Queries: Queries submitted by the user.
3. Viewed Search Results: Web pages viewed by the user. (Click through, Time spent, url footprints etc)

All the submitted queries and viewed search results are about to pose serious privacy concerns for the user. Abbreviations are mentioned in Table 1.

Correspondence to: Krishan Kumar, Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur
E-mail addresses: krishanchhillar@gmaild.com

TABLE 1 *Abbreviations used*

Abbreviation	Description
U	User
S	Search Engine
q	Query submitted to search Engine
R={R1, ...,Rn}	Search results from the search engine for query q
Ri ε R	Chosen to view Result Ri by the user
ID(U)	Some ID revealed about user like user ID or IP address
TEXT(N)	Text description of information need N of user e.g. viewed results and/or related queries
P(U)	Personal Information of User U.

So sensitive personal information a user can reveal while conducting k search activities can be expressed as follow:

P(U) = { ID(U,i) , TEXT (N, i)} where i = 1, ..., k

So P(U) is what a search engine need to personalize web search for user U. The challenge is to protect user privacy in PWS to exploit P(U) to improve services for user (U) but to protect P(U) to keep identity of user safe from the outer world.

VII. PRIVACY PROTECTION STEPS

We have an idea of six steps to implement personalised privacy protection. Users can improve and enhance their privacy by following these six steps. It will be totally up to the user how much information it want to expose through their actions and situation. In other words, we can say users can select information disclosure as per time place or some other entities involved. Each step is independent of other one and existence of each step doesn't affect others like previous or next. However, if more than one step is not available, they're still available in the same fashion and order.

A. *Step 1: Awareness*

Users can provide information on the web search engine voluntarily or involuntarily. When user full form and submit information about itself, we call it voluntarily. But when web search engine collect information about user and its activities without user consent, we call it involuntarily. The information collected from the user without its consent during the interaction for the web services is considered as a privacy risk. Many users are not aware that they can block cookies to enhance privacy but they have to sacrifice personalization for this task. Many of them don't know that web server is storing their each click, website searched and creating their profile to provide personalised results. The best and easiest way to enhance the privacy is to aware the user about these privacy risks. Privacy Critics [26] is a privacy protection tool which helps user while interacting with the Internet and issues warnings and suggestions. It is considered as first step privacy protection and spreads awareness among users about privacy risks and help them to understand their exposure with the web. However it is not protecting privacy of the user by itself. It just inform and suggest the user.

B. *Step 2: Control*

Step 2 help users to define a mechanism which can fight against privacy invasion which support user behaviour analysis like- History file access, Web Bugs, Third Party Cookies. The major platform to interact with the interact is Web Browser which help user with the help to reject and filter-out undesired data collection techniques. Hackers use malicious codes to collect information about history files and can easily reveal to third party about the webpages traversed which is considered as a privacy breach. So web browsers should delete this information automatically or provide facility to the user to delete these files. Cookies, hidden form fields, session Ids, URL rewriting, Web Brower Ids are different techniques to track user and collect information about it. Web conglomerates generally collect user information on one platform and use it to create profile and can use it on other web platform. Emails, News sites, web search engines, Shopping sites, Social sites can be used to track user and create profile web search engine is not the only platform. Most of the web browser provide facility to reject cookies, even it is cumbersome task for the user and only conscious user opt it and it may stop some desired services. Filters can also be used to block cookies advertisements and web bugs but there is disadvantage with filters that is they block all the cookies and which lose access to personalised services also. it is true that filters can only lower down chances of privacy invasion but cannot guarantee hundred percent privacy . Still geographical location, interaction time and IP address can be used to track the user.

C. *Step 3: Privacy Improving Tools*

In step 3, privacy protection tools are used to enhance user privacy. In this step location of privacy protection mechanism also matters which make it different from step 2 as well. Privacy specialist say that a user should not trust privacy policies laid down by website but should control privacy with own tools and techniques.

Pseudonyms are virtual names, individually or as part of a collective, used to maintain anonymity while interacting with internet. Generally, it’s not easy to associate a real user with the pseudonym but a groups of messages can be associated if they carry PII about the user. The Anonymizer act as a proxy and submit queries on behalf of user. In this case, Web services cannot trace user’s IP Address but there are some drawbacks also like web server or search engine cannot get information about the real user and so cannot provide personalization and customization.

Lucent Personalized Web Assistant (LPWA) [27] is a pseudonym based tool which allow user to interact with web site for identification based services without revealing actual identity. But there is drawback with this system also, if real identity of user is revealed, all past action are exposed.

Unlike LPWA and Anonymizer which involve third parties to interact with internet, Crowds[28] and Onion[29] are using groups to hide identity of user. Onion have a static path defined on other hand Crowds have dynamic path. Just

Correspondence to: Krishan Kumar, Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur
E-mail addresses: krishanchhillar@gmaild.com

like other tools these also do not provide personalization facility.

D. Step 4: Privacy Policies

This step is about the privacy policies laid down by the websites, how they collect and personal information of the user. It is mandatory for the websites to publicly disclose how the information collected from the users is handled and also describe their privacy preferences. Privacy preference Project (P3P) of World Wide Web Consortium enabled website need to provide information in a format which can be read by machine. P3P enabled web browsers can easily read this information and can perform a comparative analysis with the privacy preferences of the user. Generally, web browser continue requisition of the web pages. If the policy match with security configuration of the user else agent, notify the user about the disparity. P3P does not monitor sites for their minimum standard, user have to trust them and sites can change their privacy policy also.

E. Step 5: Privacy and Trust Certification

On the basis of Huaiqing Wang and Colleagues taxonomy website are graded. This grading takes into consideration access, collection, monitoring, analysis, transfer, solicitation and storage of the user data. Now a days, consumers and businesses are very sensitive about privacy policies and approach privacy certification very cautiously. Bankrupt companies transfer user private information and assets to other companies and these purchasing companies are generally not obliged to keep the user information safe and private.

F. Step 6: Privacy Protection Laws

In many countries, laws to regulate privacy are discussed and proposed. A mechanism to take action against companies and individuals for breaking rule is created. But until these laws are enforced universally, companies hardly respecting and protecting user privacy. One of the major reason is user is not even aware that their privacy can be violated and behaviour on web is very tough to control. Another issue is to create international laws which is dependent on diverse culture and political will. Still we can address some common concerns of privacy invasion like-

1) Use data should not be collected and analysed without user consent

2) User data should be used the way it has been consented.

3) Use of data should not be disclosed or sent to others without permission and knowledge.

Even after implementing international laws, some countries may not follow those and we can say that only laws cannot protect user data and user should trust some mechanism which van be trusted to protect the data.

VIII. Classification of Privacy Enhancing Tools:

Before we categories privacy enhancing tools in different categories, it is important to understand the difference between privacy and security as the terms are inter-related and used interchangeably also. Security is protection of personal information in terms of integrity, authentication and confidentiality, whereas privacy protection is to decide what information about himself should be communicated to others. Privacy Enhancing Tools are divided in six categories on the basis of technique used in these tools.

A. Communication Anonymizers

Communication anonymizers protect the users' IP address or other network information through anonymous communication networks e.g., Tor, mixes and mix networks, onion network, garlic routing etc. it help the user to browse the network by maintaining anonymity of the user. These applications can be web search engines (e.g., duckduckgo, lxquick), web browsers (e.g., Epic, Tor) network layer application (e.g., I_2P). These applications protect privacy by replacing or hiding user’s real online information with a non-traceable information. These are most adopted end user PETs in all categories because they provide high level of privacy protection. These applications are slower than normal browsers or search engines which don’t provide anonymization facility even blocked by some websites as well. These are sometimes used for illegal activities and are considered as part of dark web as well.

B. Privacy settings

Some web browsers, smart phones and social media services are providing few controls to the users to decide who can access their personal information and up to what extend it can be accessed. For example, in Facebook allow users even to lock their profile so that visitors cannot access their profile. Even users on Facebook can limit what kind of information is visible to whom and who has access to their personal information. Web browsers are providing new feature to access information in private mode which does not store information like browsing history, passwords and cookies. But it cannot avoid the internet service provider to store user’s personal information. Privacy setting is not providing any privacy at all, rather it is considered as user may end up sharing more sensitive information where they have illusion that they are more secure which may lead to more privacy threats[30], [31].

C. Transparency Enhancing Technologies

How information is collected, what kind of information is collected and how information is processed if very important for the user to know and a clear visibility is provided by the Transparency Enhancing Technologies. These tools compare the users’ personal privacy preferences with the policies of the website. Personalized Privacy Assistant is learning users privacy requirements and take decision on user’s behalf[32]. Data Track allow users to know what personal information has been collected by whom and for what purpose[33].

D. Trackers and Evidence Erasers

Trackers and evidence erasers help individuals in removing electronic traces related to online activities. Tools like CCleaners cleans the cookies and browser search history. Privacy Eraser remove information permanently so that information cannot be retrieved once deleted. DeleteMe is also similar tool which help users to delete personal information from data brokers and search sites. Recent web-browsers like Safari also have inbuilt trackers which protects the privacy of the users. These techniques and tools help individuals to protect privacy by deleting information which can track them from their own devices and evidence from websites which stops data collection. Capabilities of these tools are very limited as service providers can user new trackers to garner personal information of the users.

E. Filters and Blockers

Filters and blockers generally work on unsolicited and unwanted emails, messages, web-content from reaching individuals like cookies blockers, ad blockers. These tools do not protect privacy of individual rather eliminate post-hoc negative effects of loss of privacy. But cookie blocker protects privacy of individual be blocking third party cookies which indirectly stops data collection from third party firms.

F. Personal Data Store

Personal Data Store (PDS) are used to store, manage and maintain individual's digital information. As compared to traditional information storage systems, PDS is considered as new and final solution to protect privacy of the users. PDS provides full control of personal information because these are structured in decentralized way[34]. As individual's information is stored locally or trusted third party, illegal use and collection of personal information is abolished. Blockchain Technology can improve it more, but it is costly, require high computation and more carbon footprint are some disadvantages. Mydex, CitizenMe and Hub-of-All-things are examples of good PDS.

G. Common Privacy Enhancing Technologies

We can divide the privacy enhancing technologies into three broad categories. These are Cryptographic algorithms, data masking and AI & ML algorithms.

1) Cryptographic algorithms

- Homomorphic Encryption: It is an encryption technique that allows computational operation on encrypted data. This technique allows encrypted data to be transferred, analyzed and returned to the sender, which can decrypt the data and see the results of original data. With the help of this technique, companies can share sensitive information with the third parties. Encrypted data can be stored in cloud and there are many more applications of homomorphic encryption. Partial homomorphic encryption, somewhat homomorphic encryption, and fully homomorphic encryption are different types of homomorphic encryption.

- Secure multi-party computation: Secure multi-party computation is also considered as a subfield of homomorphic encryption which has one difference, user can compute values from multiple encrypted data sources. Secure multi-party computation is used for large volumes of data and machine learning models can be used to encrypt data.

- Zero knowledge proofs: This technique uses a set of cryptographic algorithms. Using this technique, information can be validated without revealing data that proves it.

- Differential privacy: This technique protects the individual from sharing any sensitive information. Cryptographic algorithm adds a "statistical noise" to dataset which maintain the privacy of individual user.

2) Data masking techniques

Data masking techniques can be implemented to protect privacy of the user by hiding sensitive information from the web search engine. A number of data masking techniques can be used in tools to protect individual.

- Obfuscation: This technique is also known as data perturbation or noise addition because some dummy or fake data is added with the sensitive information of the user to distract or mislead the web search engine.

- Data minimization: Using this technique, minimum amount of personal data is collected from the user end to protect privacy by the search engine. Quality of service is compromised here, to protect the privacy.

- Pseudonymization: Personally, identifier information is replaced with fictitious data like characters or other data. Pseudonymization is used a lot and anonymity is compromised if the user reveals his or her identity.

- Communication anonymizers: Anonymizers replace the digital online identity of the user, like IP address, MAC address by onetime untraceable identity.

- Shared bogus online accounts: One person creates an account for web search engine by providing fake name, address, phone number and other preferences. Then, the user can share user-IDs and password on internet which can be used by everyone comfortably. Here, user need to be sure that no personal information should be provided while creating the account. In this case, privacy can be maintained, but there are no chances of personalization.

3) Using AI & ML algorithms

- Federated learning: It is a machine learning technique which trains algorithm which work on servers keeping local data samples. In case of decentralized servers, user can minimize data by reducing data on a centralized server or in cloud storage.

- Synthetic data generation: Synthetic data is artificially created data by using different algorithms, like ML algorithms. Privacy-enhancing techniques transform data by generating synthetic data, where third parties can access it and data have same statistical characteristics.

IX. MULTIPLE PROXY SERVER

Step 5 and 6 cannot be implemented in any mechanism as these steps belong to different countries. But we are proposing a system which implements first three steps of privacy protection. The proposed system has proxy servers, consent based distributed privacy architecture. Multiple proxies use Network Address Translation (NAT) to improve security and deal with IP address shortage. A proxy is a temporary identification used to interact with the internet. The proxy is selected on the basis of user's area of interest in or a particular site. When a user interact with a website, website store information about the proxy instead of visitor. Users can interact with a site through different proxies depending on interest at a given time.

As shown in Figure 1, multiple proxy server has two main components- proxy servers and Privacy and security agents (PSA). Proxy servers work between user and websites whereas PSA work in coordination with web browser. PSA

Correspondence to: Krishan Kumar, Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur
E-mail addresses: krishanchhillar@gmaild.com

warn users about privacy invasion, allow users to configure proxies and cipher user request to avoid eavesdropping. It filters and blocks general privacy violations like web bugs and cookies. PSA can allow to directly interact with internet by giving up anonymity. It provide privacy protection which belong to first two steps.

Multiple proxy servers can be deployed on locations like-intranet or ISP proxy. User request goes to the group selector which decide to which group the request should be forwarded based on semantic context of user request. Every group represent a different area of interest. Groups then forward the request to any of the proxy servers randomly. Proxy servers forward the request to the web search engine and individual identity is hidden but interest are shown which help in personalization and protect privacy of the user.

Role of group selector is very important. A user can have diverse interest in a session and it is very difficult to predict its behaviour, so group selector choose a group on the basis of each request which makes it very difficult for the adversary to track the user. Multiple proxy server does not store any information about the user which makes almost impossible for the web search engine to detect sequence of requests from different proxies belong to individual or group of people with common area of interest.

TrackMeNot (TMN) is a tool for query obfuscation and Tor anonymizing network for concealing the source of queries to protect the user privacy. Beside this, solutions were analyzed against the adversarial search engine. It was found that search engine with even short-term history of user's search queries, can break the privacy guarantee of Tor and TMN with the help of machine learning techniques.

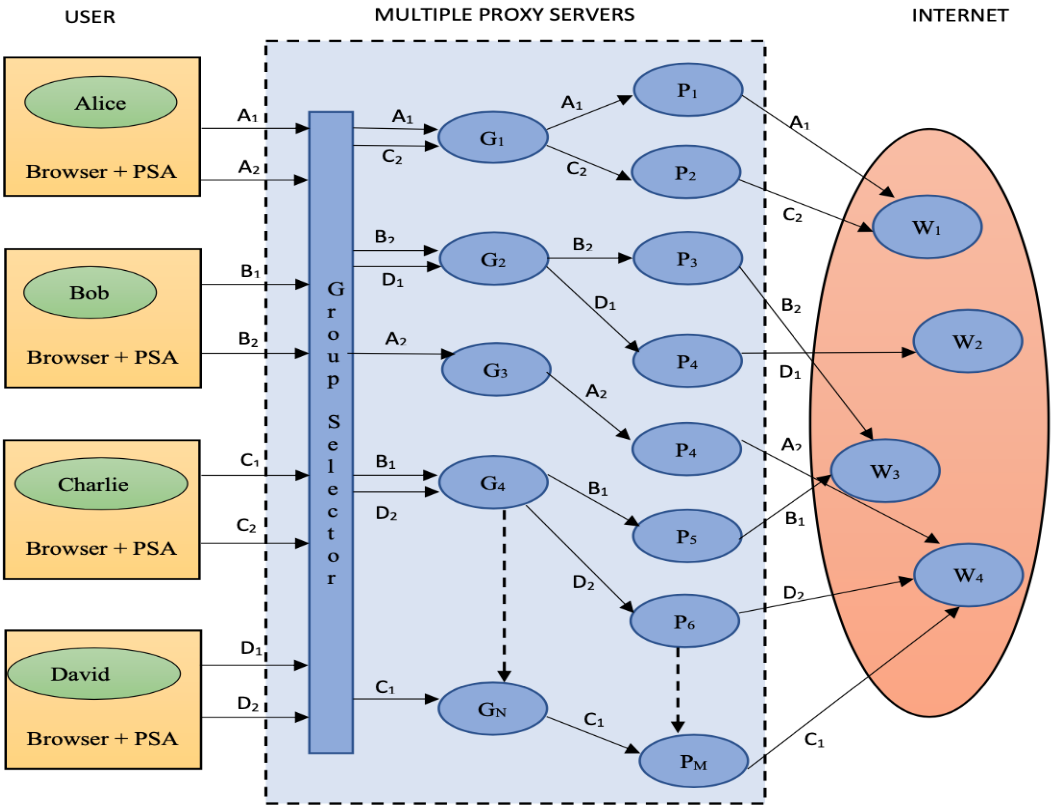

Figure 1:Multiple Proxy Server Architecture

It has been found that quantified levels of privacy provided by query obfuscation and tor anonymizing networks. 60 users were selected from AOL search logs and it was found that user queries were identified with a positive rate of 48.88 percent when TMN was used whereas false positive rate was 0.02%. 60 users from AOL search logs were used over anonymizing networks like Tor and it was found that 25.96% queries were identified with average true positive rate when N =100. Where N is size of user set performing on Tor. When N =1000, the average true positive rate decreased to 18.95%. It is true that the average true positive rate is not much high but in some cases for N=1000, the average true positive rate was 80%-98%. Authors identified reasons behind query classification and drop-in true positive rate when number of users increase in anonymization services. Therefore, results confirms that anonymizing networks and query obfuscation tools are not much effective in protecting privacy of the user. The attacks were carried out on minimal information (query content) for identification of user queries and off-the-self classification techniques, and it was still reasonably successful. But in actual conditions, the attack will be much stronger as other information like query timestamp will also be available with search engine. Search histories of the users can help the search engine to build better classifiers. Geographics locality information with queries and contextual information can further improve the results.

X. Conclusion

There is no doubt that the privacy of individual user is on stake when he is looking for the information on internet. Most of the web search engines are profiling user and storing sensitive information about the user which is a big concern for every user. There are different techniques to protect user information but privacy enhancing tools are considered as one of the fine way as it does not require anything from the client side and no modifications are needed on web server as well. These easily available tools are not able to protect the privacy 100 percent but can safeguard the user up to a great extent. To avail the free services provided by various websites and web search engines, we generally, create accounts and share personal information. These websites are selling or using that data for targeted advertisement and many more money-making businesses by putting the privacy of individual on stake.

Multiple proxy servers can provide a number of benefits including spreading awareness among the user about the privacy risks involved while using internet. The system can control the amount of private information to be shared with the internet. Anonymity is the basic feature behind the system to preserve user privacy. Unlike other privacy tools, Multiple proxy servers can disclose partial information about a group of users which can personalize services without profiling single user. It stores only one request and multiple request can use different groups and proxies which does not allow adversary to relate queries with each other. Users can select exposure level also which make the system more transparent. Complex data mining techniques and clustering algorithms are avoided and simple data structures like lists , trees used to enhance efficiency. HTTP and TCP protocols are used to communicate, and no special protocols are created.

XI. Reference:

[1] A. Acquisti et al., “The Economics of Privacy †,” J. Econ. Lit., vol. 54, no. 2, pp. 442–492, 2016, doi: 10.1257/jel.54.2.442.

[2] F. Bélanger and R. E. Crossler, “THEORY AND REVIEW PRIVACY IN THE DIGITAL AGE: A REVIEW OF INFORMATION PRIVACY RESEARCH IN INFORMATION SYSTEMS 1.” Accessed: Mar. 28, 2021. [Online]. Available: http://www.misq.org.

[3] X. Shen, B. Tan, and C. Zhai, “Privacy protection in personalized search,” ACM SIGIR Forum, 2007, doi: 10.1145/1273221.1273222.

[4] R. Dingledine, N. Mathewson, and P. Syverson, “Tor: The Second-Generation Onion Router,” Sov. At. Energy, vol. 46, no. 4, pp. 337–337, 1979, doi: 10.1007/bf01118387.

[5] R. Dingledine, “Tor and Circumvention: Lessons Learned,” 2011, pp. 485–486.

[6] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Bobbins, “Stuff I’ve Seen: A System for Personal Information Retrieval and Re-Use,” SIGIR Forum (ACM Spec. Interes. Gr. Inf. Retrieval), vol. 49, no. SPEC. ISS., pp. 72–79, 2003.

[7] L. D. Catledge and J. E. Pitkow, “Characterizing browsing strategies in the World-Wide web,” Comput. Networks ISDN Syst., vol. 27, no. 6, pp. 1065–1073, 1995, doi: 10.1016/0169-7552(95)00043-7.

[8] B. McKenzie and A. Cockburn, “An empirical analysis of web page revisitation,” in Proceedings of the Hawaii International Conference on System Sciences, 2001, p. 128, doi: 10.1109/HICSS.2001.926533.

[9] L. Tauscher and S. Greenberg, “How people revisit web pages: Empirical findings and implications for the design of history systems,” Int. J. Hum. Comput. Stud., vol. 47, no. 1, pp. 97–137, 1997, doi: 10.1006/ijhc.1997.0125.

[10] P. Yu, W. U. Ahmad, and H. Wang, “Hide-n-Seek,” in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Jun. 2018, pp. 1333–1336, doi: 10.1145/3209978.3210180.

[11] V. Toubiana, L. Subramanian, and H. Nissenbaum, “TrackMeNot: Enhancing the privacy of Web Search,” Sep. 2011, [Online]. Available: http://arxiv.org/abs/1109.4677.

[12] D. C. Howe and H. Nissenbaum, “Trackmenot: Resisting Surveillance in web search,” in Lessons from the Identity Trail : Anonymity, Privacy and Identity in a Networked Society, 2018, pp. 418–434.

[13] J. Domingo-Ferrer, A. Solanas, and J. Castellà-Roca, “H(κ)-private information retrieval from privacy-uncooperative queryable databases,” Online Inf. Rev., vol. 33, no. 4, pp. 720–744, 2009, doi: 10.1108/14684520910985693.

[14] X. Shen, B. Tan, and C. Zhai, “UCAIR A Personalized Search Toolbar,” in SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2005, p. 681, doi: 10.1145/1076034.1076193.

[15] X. Shen, B. Tan, and C. X. Zhai, “Implicit user modeling for personalized search,” in International Conference on Information and Knowledge Management, Proceedings, 2005, pp. 824–831, doi: 10.1145/1099554.1099747.

[16] S. T. Peddinti, K. W. Ross, and J. Cappos, “User Anonymity on Twitter,” IEEE Secur. Priv., vol. 15, no. 3, pp. 84–87, 2017, doi: 10.1109/MSP.2017.74.

[17] K. Liu, C. Kuo, W. Liao, and P. Wang, “Optimized data de-identification using multidimensional k-anonymity,” in 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 2018, no. 2, pp. 1610–1614, doi: 10.1109/TrustCom/BigDataSE.2018.00235.

[18] N. Man, X. Li, and K. Wang, “A Privacy Protection Model Based On K-Anonymity,” in International Conference Advanced Engineering and Technology research, 2018, vol. 153, no. Aetr 2017, pp. 15–19, doi: 10.2991/aetr-17.2018.4.

[19] J. Domingo-Ferrer and V. Torra, “A Critique of k-Anonymity and Some of Its Enhancements,” in 2008 Third International Conference on Availability, Reliability and Security, Mar. 2008, no. May 2014, pp. 990–993, doi: 10.1109/ARES.2008.97.

[20] K. El Emam, F. K. Dankar, K. El Emam, and F. K. Dankar, “Protecting Privacy Using k-Anonymity,” J. Am. Med. Informatics Assoc., vol. 15, no. 5, pp. 627–637, 2008, doi: 10.1197/jamia.M2716.

[21] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-Diversity : Privacy Beyond k -Anonymity,” in Proc. of 22nd International Conference on Data Engineering, 2006, vol. 1.

[22] J. Domingo-Ferrer and J. Soria-Comas, “From t-closeness to differential privacy and vice versa in data anonymization,” Knowledge-Based Syst., vol. 74, pp. 151–158, Jan. 2015, doi: 10.1016/j.knosys.2014.11.011.

[23] M. Barbaro and T. Zeller, “A Face Is Exposed for AOL Searcher No. 4417749,” New York Times, no. 4417749, pp. 1–3, 2006, [Online]. Available: papers3://publication/uuid/33AEE899-4F9D-4C05-AFC7-70B2FF16069D.

[24] C. Díaz, S. Seys, J. Claessens, and B. Preneel, “Towards measuring anonymity,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2003, vol. 2482, pp. 54–68, doi: 10.1007/3-540-36467-6_5.

Correspondence to: Krishan Kumar, Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur
E-mail addresses: krishanchhillar@gmaild.com

[25] C. C. Aggarwal and P. S. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," 2008, pp. 11–52.

[26] M. S. Aekerman and L. Cranor, "Privacy critics: UI components to safeguard user's privacy," in Conference on Human Factors in Computing Systems - Proceedings, 1999, pp. 258–259, doi: 10.1145/632716.632875.

[27] D. Kristol, E. Gabber, and P. Gibbons, "Design and implementation of the Lucent Personalized Web Assistant (LPWA)," Submitt. …, 1998, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.53.5909&rep=rep1&type=pdf.

[28] M. K. Reiter and A. D. Rubin, "Crowds: Anonymity for Web Transactions," ACM Trans. Inf. Syst. Secur., vol. 1, no. 1, pp. 66–92, Nov. 1998, doi: 10.1145/290163.290168.

[29] M. G. Reed, P. F. Syverson, and D. M. Goldschlag, "Anonymous connections and onion routing," IEEE J. Sel. Areas Commun., vol. 16, no. 4, pp. 482–493, 1998, doi: 10.1109/49.668972.

[30] L. Brandimarte, A. Acquisti, and G. Loewenstein, "Misplaced Confidences: Privacy and the Control Paradox," Soc. Psychol. Personal. Sci., vol. 4, no. 3, pp. 340–347, 2013, doi: 10.1177/1948550612455931.

[31] D. J. Solove, "Privacy Self-Management and the Consent Dilemma," 2013. Accessed: Mar. 31, 2021. [Online]. Available: https://scholarship.law.gwu.edu/faculty_publications.

[32] B. Liu et al., Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions. 2016.

[33] J. Angulo, S. Fischer-Hübner, T. Pulls, and E. Wästlund, "Usable Transparency with the Data Track-A tool for visualizing data disclosures," doi: 10.1145/2702613.2732701.

[34] T. Kirkham, S. Winfield, S. Ravet, and S. Kellomaki, "A personal Data Store for an Internet of Subjects," Online (Wilton, Connect., vol. 34, no. 3, pp. 26–28, 2011, doi: 10.1109/mp.2006.1692290.