IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 24, 2021, accepted April 27, 2021, date of publication May 10, 2021, date of current version June 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3078431

Multi-Group ObScure Logging (MG-OSLo) A Privacy-Preserving Protocol for Private Web Search

MOHIB ULLAH[1,2], RAFIULLAH KHAN[1,2], MUHAMMAD INAM UL HAQ[3], ATIF KHAN[4], WAEL ALOSAIMI[5], MUHAMMAD IRFAN UDDIN[6], AND ABDULLAH ALHARBI[5]

[1]Institute of Computer Sciences and Information Technology, The University of Agriculture, Peshawar, Peshawar 25120, Pakistan
[2]Department of Computer Science, Capital University of Science and Technology Islamabad 44000, Pakistan
[3]Department of Computer Science and Bioinformatics, Khushal Khan Khattak University, Karak 27200, Pakistan
[4]Department of Computer Science, Islamia College Peshawar, Peshawar 25120, Pakistan
[5]Department of Information Technology, College of Computers and Information Technology, Taif University, Taif 21944, Saudi Arabia
[6]Institute of Computing, Kohat University of Science and Technology, Kohat 26000, Pakistan

Corresponding author: Rafiullah Khan (rafiyz@gmail.com)

This work was supported by the Taif University Researchers, Taif University, Taif, Saudi Arabia, under Project TURSP-2020/231.

ABSTRACT The Web Search Engine (WSE) is a software system used to retrieve data from the web successfully. WSE uses the user's search queries to build the user's profile and provide personalized results. Users' search queries hold identifiable information that could compromise the privacy of the respective user. This work proposes a multi-group distributed privacy-preserving protocol (MG-OSLo) and tries to investigate the state-of-the-art distributed privacy-preserving protocols for computing web search privacy. The MG-OSLo comprises multiple groups in which each group has a fixed number of users. The MG-OSLo measures the impact of the multi-group on the user's privacy. The primary objective of this work is to assess local privacy and profile privacy. It aims at evaluating the impact of group size and group count on a user's privacy. Two grouping approaches are used to group the users in MG-OSLo, i.e. a non-overlapping group design and overlapping group design. The local privacy results reveal that the probability of linking a query to the user depends on the group size and group count. The higher the group size or group count, the lower the likelihood of relating the query to the user. The profile privacy computes the profile obfuscation level using a privacy metric Profile Exposure Level (PEL). Different experiments have been performed to compute the profile privacy of the subset of an AOL query log for two situations: i) self-query submissions allowed and ii) self-query submissions not allowed. The privacy achieved by MG-OSLo is compared with the modern privacy-preserving protocol UUP(e), OSLo, and Co-utile protocols. The results show that the MG-OSLo provided better results as compared to OSLo, UUP, and Co-utile. Similarly, the multi-group has a positive impact on local privacy and user profile privacy.

INDEX TERMS Web search privacy, profile obfuscation, anonymity, profile exposure level.

I. INTRODUCTION

Being an enormous warehouse of data, the World Wide Web (WWW) is a storehouse to various documents, including text, images, video, audio, etc. Today, information about anything and everything is uploaded on the Web. We depend on the Web Search Engines (WSEs) to search for specific information on the Web. The WSE has become an integral part of life these days. Information from the Internet is obtained by people from all walks of life all over the world [1], [2].

The current Internet live statistics show that Google alone answers around 72 thousand queries in one second [3]. The WSEs record all the submitted queries in a query log during the information retrieval process. WSEs also make the profile of a user based on his or her web search queries. A typical query log contains a user-submitted query, the machine IP address, operating system details, browser type, the query's temporal information, some critical preferences, and cookies possibly used to recognize users' browsers matchlessly [1], [4]. WSEs claim that the query log is evaluated through specific algorithms for an extended period to infer users' interests for providing personalized search results [4]–[6].

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

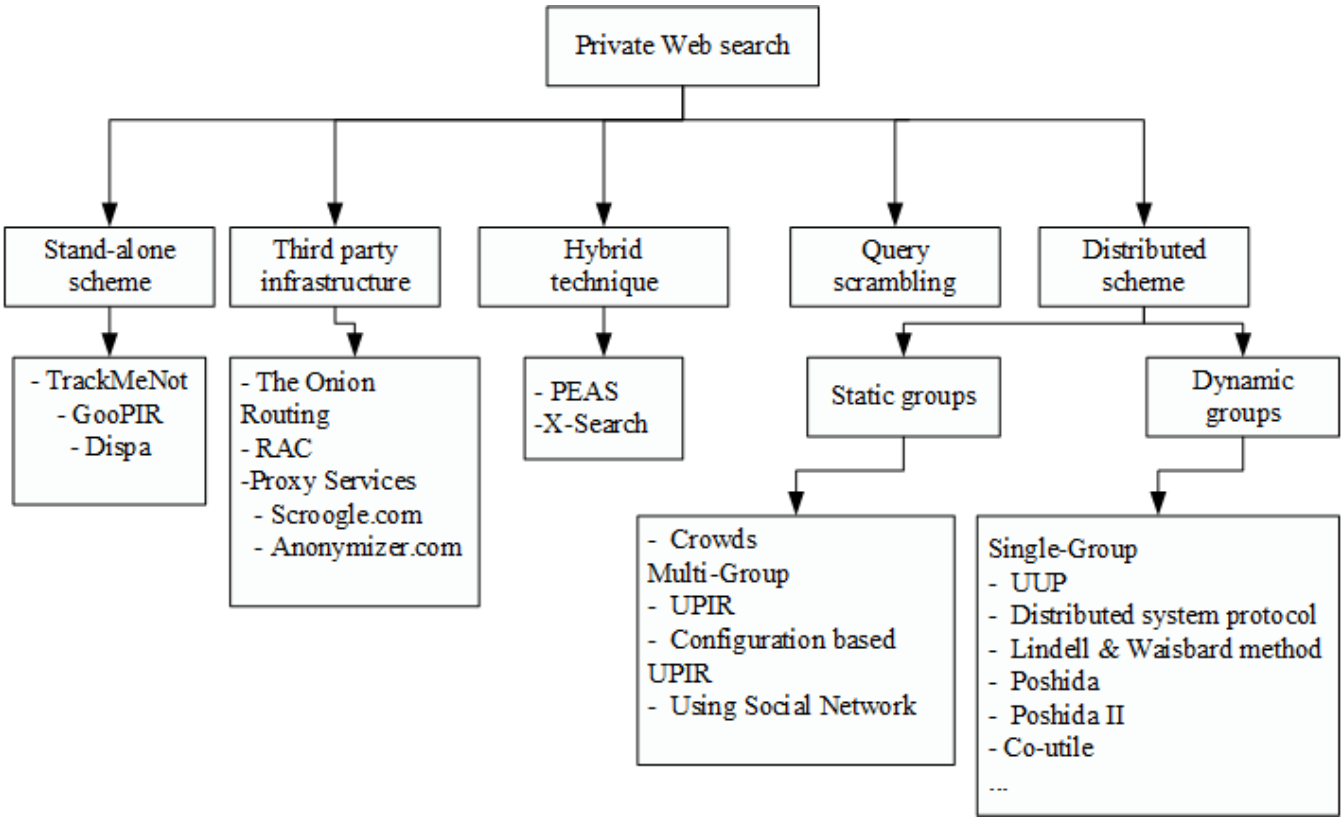

FIGURE 1. Taxonomy of private web search.

The user's profile is an asset for WSEs because it exposes sensitive data about the user. A user's queries often contain important information like Unique User ID, name, user's employer's details, location, etc. Moreover, a user's query may reveal health information, gender orientation, religion, politics, faith, believes, etc., which may be deemed sensitive for a possessor [7]. The disclosure of a query log poses severe risks in terms of privacy. Preserving the web search privacy of a user is the genuine concern of today's online life.

Researchers have presented various types of techniques to preserve the privacy of a user. Figure 1 shows the taxonomy of existing approaches that provide Web search privacy. The standalone scheme, Third-party infrastructure, Hybrid techniques, Query scrambling, and distributed schemes are popular solutions used to achieve the private web search. The standalone schemes focus on sending fictitious (machine-generated) queries with the original query to mask the user's original intent. TrackMeNot, GooPIR, and DisPA are the famous standalone scheme to obfuscate the user's profile [2], [8]. However, the machine-generated queries are distinguishable from actual queries. Users who want to hide their identity from WSE often utilize third-party infrastructure (TOR) and proxy services (scroogle.com and anonymizer.com, and others). TOR provides network-layer anonymity, and the WSE can identify a user from the application layer or cookies. Likewise, the users utilize the hybrid techniques (PEAS and X-Search) that send fictitious queries through third-party to obfuscate their profile and hide their identity for achieving web search privacy. Query scrambling splits the query into multiple terms; each term is sent to WSE separately to conceal the user's actual query [9], [10]. Nevertheless, the query results retrieved through query scrambling are still far from what the users desire. Distributed protocol is another approach that functions by collaborating multiple users where each user forwards other users' queries to hide his or her identity and obfuscate his or her profile [6], [11]–[20]. The existing solutions achieve Web search privacy through indistinguishability and unlinkability. The indistinguishability is the process of obfuscating the user profile maintained by the WSE. Simultaneously, unlinkability hides an individual's identity, i.e., disassociating a query from a user. The distributed protocols achieve both the advantages of indistinguishability and unlinkability, while other privacy-preserving techniques may provide one or the other. However, in modern distributed protocols, either the query or results remain visible to group users or suffer a functionality issue. Sometimes, A user has to wait for a long time to get the query answer. This paper proposes multi-group obscure logging (MG-OSLo) being a distributed privacy-preserving protocol to tackle the aforementioned limitations. We aim at achieving confidentiality (i.e., hiding the query and results), indistinguishability (obfuscating the user's profile maintained by the

WSE with the queries of other group members), unlinkability (the group users shall not be able to link a query with the originator), and functionality (a user gets an answer for every query sent) for searching data on the WSE. MG-OSLo consists of entities like a user, Query Forwarding Node (QFN), Central Server (CS), and WSE. MG-OSLo creates multiple groups for achieving Web search privacy by encryption, query shuffling, and forwarding queries of other group members.

To evaluate a user's privacy, Ullah *et al.* [15] have defined the local privacy and profile privacy for scaling the Web search privacy. The local privacy computes the unlinkability, whereas the profile privacy measures the indistinguishability. The local privacy is considered preserved if no entity of the distributed protocol can see the query and link it with the user. Similarly, a user's profile privacy is deemed preserved if WSE cannot build a reliable profile of a user and hence he or she attains indistinguishability.

The rest of the paper is organized as: section II explains the related work. MG-OSLo model is described in section III. The privacy evaluation is detailed in section IV, followed by results and discussion in section V whereas the conclusion and future work are described in the last section.

II. RELATED WORK

As mentioned earlier, the distributed protocol requires multiple users' collaboration to forward queries of group members to the WSE. There have been many distributed protocols proposed from time to time. The distributed protocols are divided into static and dynamic group categories based on the users' grouping nature. In the static group category, a group once created would remain with the same users every time the protocol executes [6]. A dynamic group is the second category that constitutes a new group, possibly consisting of different users each time the protocol executes.

Crowds presented by Reiter and Rubin [21] employed static group to preserve the web search privacy of a user. The authors of [12], [19] had utilized the social networks (e.g., Facebook, MSN messenger) to establish a path to WSE for achieving Web search privacy. However, this technique offered no confidentiality and remained vulnerable to the predecessor's attack. In the same static group category, Domingo-Ferrer *et al.*, introduced user-private information retrieval (UPIR) by adopting memory location and organizing users into multiple groups to retrieve data from the database. UPIR allowed a user to hide their identity among the group members [20]. The concept of UPIR was later extended to various grouping techniques like balanced incomplete block design (BIBD) and pairwise balanced design (PBD) [16], [17]. In these protocols, a user was supposed to write a query in the memory location; another member linked with the exact memory location was required to proxy a query to the database. However, the authors did not mention any technique of asking a peer user to proxy a query anonymously. Further, they did not evaluate the profile privacy of a user by executing UPIR. In 2018, Domingo-Ferrer *et al.*, introduced a self-enforcing protocol called Co-utile protocol to promote social welfare [22], [23]. A user of the Co-utile protocol was initially required to check if a sending query would obfuscate or expose their profile maintained by the WSE. In case the query obscured the profile, the user was supposed to forward the query to WSE on their own.h Otherwise, the same user asked the peer user to forward the query to WSE on his or her behalf. The peer user followed the same steps. He or she checked if the query obfuscates his or her profile, then he or she forwarded it to WSE, otherwise drops the query.

In the dynamic group category, researchers introduced protocols consisting of entities like a user, a central server (CS), and a WSE [6], [14], [24]. These protocols initially created a group after receiving a connection request from the users. Each user in the group forwards a query of another member to the WSE. In this way, the user's identity remains concealed, and the profile gets obfuscated. Useless User Profile (UUP) [24] is the first dynamic group distributed protocol. It makes a group for 'n' users, where users shuffle their queries before forwarding them to the WSE. In UUP, when users dispatch all the queries, the group concludes. Similarly, the whole process of group creation and query shuffling must get repeated for a user to send a new query. Lindell and Waisbard concluded by investigating UUP being insecure in the presence of a single adversary and suggested using double encryption to preserve Web search privacy [25]. However, double encryption puts an extra delay on the system resulting in twofold costly as UUP. Romero-Tris *et al.* improved UUP's privacy in the presence of an untrusted partner (UUP(e)) [6]. They used El-Gamal group key encryption for confidentiality and optimized bens network for query shuffling for achieving anonymity. However, the results were broadcasted in an unencrypted form and disclosed what queries were being searched inside the group. Furthermore, the extended UUP(e) [6] was secure in the presence of an untrusted partner; however, it was also susceptible to data mining attacks [10], [26]. In 2019, Ullah *et al.* introduced ObScure logging (OSLo) by employing a single dynamic group to achieve local privacy and profile privacy in private web search [15].

A. LIMITATION IN THE EXISTING DISTRIBUTED PROTOCOLS

We have identified several limitations in the existing privacy-preserving distributed protocols. Functionality is a significant issue in the Co-utile protocols [22], [23]. A peer user may turn down the request for forwarding a query because the peer user only forwards the query if it obfuscates his or her profile. Even though a higher number of users in a single group may cause network overhead, OSLo tackles functionality issues. In the extended UUP(e), the query result is broadcasted in clear text form, which lets the group users know what is searched inside the group [6]. Likewise, scalability is another issue in UUP(e), and OSLo [6], [15]. These protocols are simulated only for the group size of three users, four users, and five users.

Moreover, it is not clearly stated in the multi-group protocols whether a user may ask a peer user associated with the exact memory location to proxy a query on his or her behalf [16], [17], [20]. Once a user writes a query in the memory location, they have to wait for the peer user to read the query and proxy it to the WSE, which ultimately causes a significant delay. To the best of our knowledge, the existing multi-group distributed protocols evaluate a user's privacy only from one dimension i.e. relative to the group users. The profile privacy (the privacy of users relative to the WSE) in multi-group distributed protocol has never been evaluated to compute the magnitude of profile obfuscation.

B. CONTRIBUTION

This work proposes a multi-group distributed privacy-preserving protocol called MG-OSLo (Multi-Group ObScure Logging) that tackles the aforementioned limitations. Furthermore, it also preserves and evaluates the local privacy and profile privacy in a private web search. The list of the main contribution of this work is as follows:

1) Our primary objective is that the user's query and its results must remain concealed from the group members. The unlinkability between the user and the query must be assured, and WSE should not be allowed to build an accurate user profile.
2) Functionality is the secondary objective of MG-OSLo, and it strives to make possible that a user gets an answer for all his or her queries.
3) Additionally, MG-OSLo also aims to evaluate the impact of group count and group size on the user's local privacy and profile privacy.
4) To compute the local privacy, users are grouped using a non-overlapping group design and overlapping group design and calculating the probability of linking the query with the originator by a curious entity..
5) Likewise, a user's profile privacy is calculated for two situations: first, in which a self-query submission is allowed, and second, in which a self-query submission is not allowed.

A privacy metric termed as Profile Exposure Level (PEL) is used to measure the magnitude of profile obfuscation.

III. MULTI GROUP ObScure LOGGING (MG-OSLo) MODEL

The MG-OSLo consists of an 'M' number of groups, and each group accommodates a 'K' users to secretly perform a web search. MG-OSLo does not use the memory location; instead, the Central Server (CS) groups the users. This section describes the entities required for the execution of MG-OSLo and explains the working of the MG-OSLo.

A. ENTITIES

The following are the entities required for the executing of MG-OSLo.

1) USER

An individual who intends to search a query over the WSE covertly.

2) CENTRAL SERVER (CS)

A machine dedicated to supervises the working of the MG-OSLo. The CS is responsible for group creation, selection of QFN, and query shuffling. In this work, we have considered the CS as an honest but curious machine that performs its duties genuinely and accordingly.

3) QUERY FORWARDING NODE (QFN)

The CS selects a user as a QFN that forwards group users' queries to the WSE. Every group have one QFN; CS will choose all users as QFN in a round-robin fashion.

4) WEB SEARCH ENGINE (WSE)

A software system, that is used to search information on Internet based on queries.

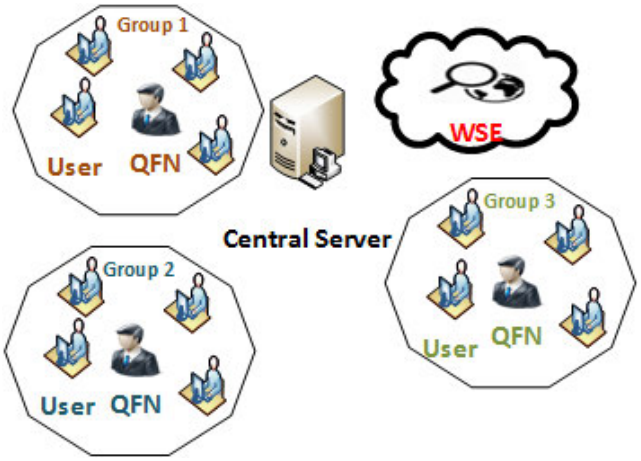

FIGURE 2. MG-OSLo entities and group design.

B. MG-OSLo EXECUTION PROCESS

Figure 2 shows MG-OSLo entities and group design. The privacy requirement of MG-OSLo is to achieve the following objectives:

1) The query of the user and its results must remain concealed from the group users.
2) The unlinkability between the user and query must be assured, i.e., neither a group entity nor WSE can link a query with the user.
3) WSE should not be able to build an accurate profile of the user.

The following are steps required in the execution of MG-OSLo:

- Connection setup, group creation, and QFN selection
- Query sending process
- Query shuffling
- Query forwarding to WSE and result processing

1) CONNECTION SETUP, GROUP CREATION, AND QFN SELECTION

As mentioned earlier, there is an 'M' number of groups, and each group has 'K' users in MG-OSLo. The Central server (CS) continuously listens to the connection request from the users. A user must send a connection request to the CS to search a query covertly through MG-OSLo. The CS places the user in a group having a vacant slot when it receives a connection request from a user, and it also records the user's credential (IP and port number). It creates a new group if there is no empty slot in a group. It selects one user as a QFN for each group in the next step once the group count and group size are complete. It asks the selected QFNs for the encryption key and the group ID (*G_ID*). Each QFN provides the public key and *G_ID*. In this work, we have used the RSA encryption scheme for query encryption. The user uses the public key to encrypt a query for achieving confidentiality, whereas QFN uses the *G_ID* to confirm that a query is encrypted with their public key. When the process of group creation and QFN selection gets complete, the CS broadcasts the information about all groups, users in each group, and the details about the QFNs. A QFN is supposed to forward queries to the WSE on behalf of other users. Every user of MG-OSLo plays the role of QFN in a round-robin fashion.

2) QUERY SENDING PROCESS

After the CS broadcasted the group's information, each user holds details of users in each group and information about QFNs. Figure 3 shows the query sending and result retrieval process. A user is required to perform the following steps to send a query to WSE. In the first step, the user generates a query 'q,' an encryption key *K_Ui*, and a query ID (*q_ID*). The *K_Ui* is a 128bits AES share key used to encrypt the result (r) for the query 'q'. At the same time, a user matches the *q_ID* to recognize that the result(r) is for his or her query(q). The user concatenates the 'q,' *K_Ui*, and *q_ID* and makes a packet called a query message (QMsg). In the next step, the user selects a QFN from the list and encrypts the QMsg with his or her public key, making an encrypted query packet (eQ). The user then concatenates *G_ID* and eQ, producing an encrypted query message *eQ_Msg*. QFN uses the *G_ID* to confirm that query (q) is encrypted with their public key, and he/she is supposed to forward the query (q) to the WSE. After completing the query encryption process, query shuffling is performed as the next step of the query sending process in MG-OSLo. The *eQ_Msg* is shuffled in two stages to disassociate the query and the user.

3) QUERY SHUFFLING

The process of query shuffling is performed in two steps, i.e., intra-group shuffling and inter-group shuffling. The query is shuffled among the group users so that no user inside the group or the group's QFN can link a query with the originator in the intra-group shuffling. The user tosses a coin to decide the destination for *eQ_Msg*. If the tossing produces head, the *eQ_Msg* is forwarded to QFN, and intra-group shuffling ends. Otherwise, *eQ_Msg* is forwarded to a random user of the group. The shuffling continues until *eQ_Msg* reaches the QFN. After receiving the *eQ_Msg*, QFN checks the *G_ID* to find if the query(q) is encrypted with his or her public key. If the value of *G_ID* matches, the *eQ_Msg* is decrypted. Otherwise, the second step (inter-group) of shuffling starts. The inter-group shuffling hides the user's group identity and disassociates the query and group. The QFN flips a coin in the inter-group shuffling; if the outcome of coin tossing is head, the *eQ_Msg* is forwarded to CS, which forwards the *eQ_Msg* to all QFNs. In the case of a tail outcome, the QFN sends the *eQ_Msg* to another random QFN. Any QFN who receives an *eQ_Msg* from CS checks the *G_ID*. If the *G_ID* matches, the QFN decrypts *eQ_Msg*; otherwise, the *eQ_Msg* is dropped. Once the *eQ_Msg* reaches the QFN, the process of shuffling concludes.

4) QUERY FORWARDING TO WSE AND RESULT PROCESSING

Once the *eQ_Msg* arrives at the destined QFN, the process of decryption starts. It first decrypts the *eQ_Msg* using the private key to acquire the query message *Q_Msg*. The QMsg is dis-concatenated to acquire the query 'q', query ID (*q_ID*), and encryption key (*eK_Ui*). The query (q) is forwarded to the WSE, which processes the 'q' and returns the query results (*Q_Result*). The QFN encrypts the result (*Q_Result*) with the user's encryption key (*eK_Ui*), making an *En_Res*. The *En_Res* are concatenated with *q_ID*, thus creating an encrypted answer message *eAns_Msg*. It forwards *eAns_Msg* to CS, which sends the results to all QFNs. Afterward, each QFN broadcasts the *eAns_Msg* in their respective group. The user who has the decryption key decrypts the *eAns_Msg*. The user uses the *q_ID* to confirm that the results are for the query 'q' what he or she has sent. Once the *q_ID* matches, the user decrypts the packet with the decryption key and retrieves the query's result.

IV. PRIVACY EVALUATION OF MG-OSLo

The query sending and result retrieval process of MG-OSLo depicts that users can achieve local privacy and profile privacy in web searching. The user attains local privacy by encryption and shuffling. The encryption conceals a query and the query results. On the other hand, query shuffling breaks the link between the user and query by making it unlinkable with the originating user. The user achieves profile privacy by polluting the user's profile with the queries of group users. Local privacy and profile privacy are the two dimensions used to evaluate the user's web search privacy. The local privacy measures the unlinkability, whereas the profile privacy quantifies the indistinguishability. The local privacy calculates the probabilistic advantage being a curious entity that links a query with the originating user, whereas the profile privacy computes the level of profile obfuscation. The section below gives a comprehensive evaluation of local privacy and profile privacy.

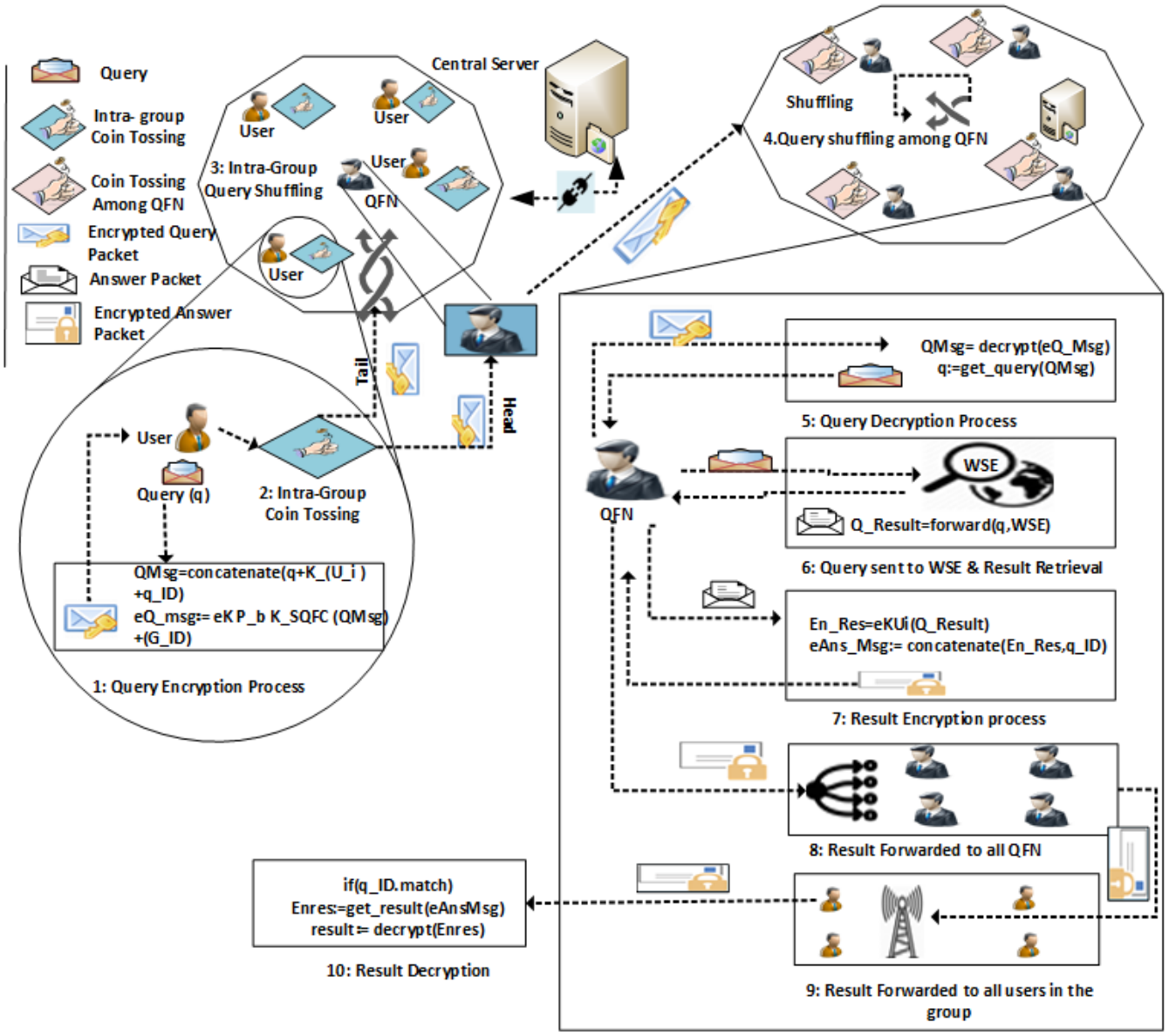

FIGURE 3. Query sending and result retrieval process of MG-OSLo

A. LOCAL PRIVACY

The local privacy of a user depends on the way the CS groups the users. The users can be grouped in different ways, whereby a user's privacy gets affected by the users' grouping. In this work, we have considered two approaches to the grouping design. First, Non-overlapping group design: Each user appears in one group; all groups are distinct. Second, overlapping group design: A user appears in multiple groups simultaneously.

B. NON-OVERLAPPING GROUP DESIGN

In this design, each user appears in a single group. Each group has its QFN, and a user can send a query through any QFN. We have adopted a (v, b, r, k) approach for non-overlapping group design.

'v' total number of users, i.e., $U = \{U_1, U_2, \ldots U_v\}$

'b' total number of groups, i.e., $M = \{S_1, S_2, ..S_b\}$

'r' degree of a user, i.e., how many groups a user is associated with. In a non-overlapping design, r is one because each user appears in a single group.

'k' is the number of users in a single group.

We have considered three random variables, S,M, and P. Where S represents a query source, M denotes a group, and P indicates proxy (QFN or CS), an entity that forwards the query to QFN. The subsection below describes the local privacy of a user relative to the entities of MG-OSLo. When a non-overlapping design is considered for grouping users, and

a curious entity of MG-OSLo wants to link a query with the user, the probability of linking query is given as:

1) PRIVACY RELATIVE TO QFN

What advantages a QFN has in linking a query with the user? When it receives a suspicious query on an esoteric topic and interested in finding the query source. According to the working of MG-OSLo, the query is shuffled well before reaching the QFN. The probability of connecting query with the user.

$$Pr(S=U_i, M=S_l \mid P=QFN) = Pr(M=S_l \mid P=QFN) \cdot Pr(S=U_i \mid M=S_l, P=QFN) \quad (1)$$

$$Pr(M=S_l \mid P=QFN) = \frac{1}{(b-1)} \quad (2)$$

where, b is the total number of block (groups). QFN knows that query does not belong to their group so he excludes their group.

$$Pr(S=U_i \mid M=S_l, P=QFN) = \frac{1}{K} \quad (3)$$

where K is the total number of users in each group.

Equating 2 and 3

$$Pr(S=U_i, M=S_l \mid P=QFN) = \frac{1}{(K(b-1))} \quad (4)$$

If QFN wants to link query with the user, the probability depends on a number of users and groups, shown in (4). Considering a situation when there are four groups and each group has four users, then according to (4), the probability of linking the query to the users is $\frac{1}{12}$. QFN does not get any advantage in linking query with the user except excluding his or her group.

2) PRIVACY RELATIVE TO THE CORE SERVER

The CS receives the encrypted query at the end of the shuffling process. A user achieves confidentiality due to the encryption of both query and result. As a consequence, the CS will not be able to discover the content of a query. Considering the similar situation when there are four groups and each group has four users, If QFN or any other user does not collaborate with CS, the probability of linking the query to the user is $\frac{1}{v}$ or $\frac{1}{16}$ because all users are equally probable, where 'v' is the total number of users in all groups ($v = b.K$). The CS does not get any advantage in linking the query with the user.

However, when QFN & CS form a coalition to find the query source, the probability of linking the query to the user is given by:

$$Pr(S=U_i, M=S_l \mid P=QFN, S \notin M_c) = Pr(M=b_i \mid P=QFN, S \notin M_c) \cdot Pr(S=U_i \mid M=S_l, P=QFN, S \notin M_c) \quad (5)$$

M_c is a group whose QFN has made a coalition with CS.

$$Pr(M=S_l \mid P=QFN, S \notin M_c) = \frac{1}{(b-1)} \quad (6)$$

$$Pr(S=U_i \mid M=b_i, P=QFN, S \notin M_c) = \frac{1}{K} \quad (7)$$

Equating 6 and 7

$$Pr(S=U_i, M=S_l \mid P=QFN, S \notin M_c) = \frac{1}{(K(b-1))} \quad (8)$$

The CS can exclude only those groups whose QFN has made a coalition with the CS shown in (8); all other groups and users are equally probable. However, if 'C' number of QFNs builds a coalition with CS, all of those groups would be excluded from the anonymity list, then the probability of link query would be $\frac{1}{(b-C)K}$.

3) PRIVACY RELATIVE TO GROUP USERS

As discussed earlier, the query and the results are encrypted under the public key of the QFN. The query originating user achieves both unlinkability and confidentiality relative to users in the same group or other groups. A malicious user would not be able to see either the query or answer to the query. However, suppose a situation when all QFNs are compromised (i.e., the one who has sent the query to WSE & the QFNs of other groups), along with the CS, build a coalition to find the query originator. Equation (9) gives the probability of linking the query to the user when the entities mentioned earlier form a coalition to find the query originator. For example, in the previous-considered situation, when there are four groups, each group has four users, based on (9), the probability of a linking query is $\frac{1}{(4)}$.

$$Pr(S=U_i \mid M=S_l, P=QFN) = \frac{1}{K} \quad (9)$$

When C users collaborate with the QFN and CS to find the user, the probability of linking the query to the user is presented (10). In such case, if C = 2 (in case of two compromised users) based on (10) the probability is $\frac{1}{(2)}$.

$$Pr(S=U_i \mid M=S_l, P=QFN) = \frac{1}{(|K|-C)} \quad (10)$$

C. OVERLAPPING GROUP DESIGN

In the overlapping group design, a user appears in multiple groups. One user (QFN) is supposed to forward the queries to WSE. The description of the overlapping group design is detailed below. If an overlapping design is used, i.e., a user appears in multiple groups, one user (QFN) is supposed to forward the queries to WSE. The design is described as

v total number of users i.e., $U = \{U_1, U_2, \ldots U_v\}$.

b total number of groups i.e., $M = \{S_1, S_2, \ldots S_b\}$.

r degree of user, i.e., the user association with number of groups.

k number of users in a single group.

λ pair of user appearance in a group if $\lambda = 1$ means, a pair of users appears in a single block, $\lambda = 2$ means a pair appears in two block.

Definition: Balance Incomplete Block Design (BIBD): A (v, b, r, k, λ) design in which every pair of points occurs in exactly λ blocks [27], [28]. In the MG-OSLo, users are grouped using the BIBD approach to evaluate the impact of overlapping design on local privacy. Let's consider the first case where (v, b, r, k, λ) design is used, S = source, P = proxy (a user who forwards the query to QFN), M = group, Group Query Forwarding Node (QFN) user who forwards the query to the WSE.

1) QUERY SOURCE AND QFN BELONG TO DIFFERENT GROUPS

A user appears in multiple groups in the overlapping group design and the QFN is supposed to forward the queries of all peer users to the WSE. Considering the situation when QFN and query source are not from the same group, the probability of linking a query by the QFN with the source is computed as, (11)–(15), shown at the bottom of the page. Equating 14 and 15

$$Pr(S = U_i, M = S_l) = \frac{r}{b} \cdot \frac{r}{b \cdot K} \tag{16}$$

$$Pr(P = U_j, QFN \notin S_l) = Pr(P = U_j) \cdot Pr(QFN \notin S_l) \tag{17}$$

$$Pr(P = U_j) = \sum_{j=1, U_i \in S_l}^{b} Pr(U_j \mid M = S_l) \cdot Pr(M = S_l)$$

$$= \sum_{j=1, U_i \in S_l}^{b} \frac{1}{K} \cdot \frac{1}{b} = \frac{r}{(b \cdot K)} \tag{18}$$

$$Pr(P = U_j) = \frac{r}{b \cdot K} \tag{19}$$

$$Pr(QFN \notin S_l) = \frac{(b - r)}{b} \tag{20}$$

Equating 19 and 20

$$Pr(P = U_j, QFN \notin S_l) = \frac{r}{b \cdot K} \cdot \frac{(b - r)}{b} \tag{21}$$

$$Pr(P = U_j, QFN \notin S_l \mid S = U_i, M = S_l) = Pr(P = U_j \mid S = U_i, M = S_l) \cdot Pr(QFN \notin S_l \mid S = U_i, M = S_l) \tag{22}$$

$$Pr(QFN \notin S_l \mid S = U_i, M = S_l) = \frac{(b - r)}{b} \tag{23}$$

$$Pr(P = U_j \mid S = U_i, M = S_l) = \frac{1}{K} \tag{24}$$

Equating (23) and (24)

$$Pr(P = U_i, QFN \notin S_l \mid S = U_i, M = S_l) = \frac{1}{K} \cdot \frac{(b - r)}{b} \tag{25}$$

Equation 16, 21, and 25

$$= \frac{r}{(b - 1) \cdot K} \tag{26}$$

Equation (26) shows the probability of linking query with the source when QFN is supposed to forward all user queries from multiple groups to the WSE. Let suppose we have BIBD with configuration values as: $v = 4$, $b = 4$, $r = 3$, $k = 3$, and $\lambda = 2$. Based on the (26) the probability of linking query with user is $\frac{1}{3}$.

Source and QFN Belong to the Same Group: When overlapping group design is used, QFN and source occur in the same group, two case may occur, i. $QFN = U_j$ ii. $QFN \neq U_j$. The probability of such cases is given by.

Case 1: When QFN and the query source user occur in the same group such that $QFN = U_j$, (27)–(29), as shown at the bottom of the next page. As U_i and QFN are pair in λ groups so

$$Pr(P = U_j \mid M = S_l, S = U_i, QFN \in S_l) = \frac{1}{K - 1} \tag{30}$$

$$Pr(S = U_i \mid QFN \in S_l) = \frac{Pr(S = U_i \cap QFN \in S_l)}{Pr(QFN \in S_l)} \tag{31}$$

$$Pr(S = U_i, M = S_l \mid P = U_j, QFN \notin S_l) = \frac{Pr(P = U_j, QFN \notin S_l \mid S = U_i, M = S_l) \cdot Pr(S = U_i, M = S_l)}{Pr(P = U_j, QFN \notin S_l)} \tag{11}$$

$$Pr(S = U_i, M = S_l) = Pr(S = U_i) \cdot Pr(M = S_l \mid S = U_i) \tag{12}$$

$$Pr(S = U_i) = \sum_{i=1}^{b} Pr(M = S_l) \cdot Pr(U_i \mid M = S_l) = \sum_{i=1}^{b} \frac{1}{b} \cdot \frac{1}{K} \tag{13}$$

$$Pr(S = U_i) = \frac{r}{(b.K)} \tag{14}$$

$$Pr(M = S_l \mid S = U_i) = \frac{r}{b} \tag{15}$$

$$= \frac{Pr(S = U_i \cap QFN \in S_l)}{\sum_{i=1}^{b} Pr(S_l) \cdot Pr(QFN \mid S_l)} \tag{32}$$

$$Pr(S = U_i \cap QFN \in S_l) = \frac{\lambda}{k \cdot K} \tag{33}$$

$$\sum_{i=1}^{b} Pr(S_l) \cdot Pr(QFN \mid S_l) = \sum_{1=1}^{b} \frac{1}{b} \cdot \frac{1}{k} = \frac{r}{b \cdot k} \tag{34}$$

Equating 33 and 34

$$Pr(S = U_i \mid QFN \in S_l) = \frac{\lambda}{r} \tag{35}$$

$$Pr(M = S_l, P = U_j \mid QFN \in S_l)$$
$$= Pr(M = S_l \mid P = U_j, QFN \in S_l) \cdot Pr(P = U_j \mid QFN \in S_l) \tag{36}$$

$$Pr(M = S_l \mid P = U_j, QFN \in S_L) = \frac{1}{r}$$
$$as\ QFN = U_j \tag{37}$$

$$Pr(P = U_j \mid QFN \in S_l) = 1 \quad as\ U_j = QFN \tag{38}$$

Equating 29, 30, 35 and 37 we get

$$Pr(S = U_i \mid M = S_l, P = U_j, QFN \in S_l) = \frac{1}{K - 1} \tag{39}$$

When QFN, User U_i and belongs to the same group the probability of linking query with the source is given in (39). In the above-considered BIBD configuration, based on (39) the probability of link query with user is $\frac{1}{2}$.

Case 2: When overlapping group design is used, such that $(QFN \in S_l)$ and $U_j \neq QFN)$ then the probability is given by (40)–(43), shown at the bottom of the page. Equating 42 and 43

$$Pr(S = U_i, M = S_l) = \frac{r}{b} \cdot \frac{r}{b \cdot k} \tag{44}$$

$$Pr(P = U_j,\ QFN \in S_l) = Pr(P = U_j) \cdot Pr(QFN \in S_l) \tag{45}$$

$$Pr(S = U_j) = \sum_{j=1}^{b} Pr(M = S_l) \cdot Pr(U_j \mid M = S_l)$$
$$= \sum_{j=1}^{b} \frac{1}{b} \cdot \frac{1}{k}$$
$$Pr(S = U_j) = \frac{r}{b \cdot k} \tag{46}$$

$$Pr(QFN \in S_l) = \frac{\lambda}{b} \tag{47}$$

Equating 46 and 47, (48)–(55), as shown at the bottom of the next page, equating 53, 54 and 55

$$Pr(QFN \in S_l \mid S = U_i, M = S_l) = \frac{1}{r} \tag{56}$$

Equating 44, 48, 50, and 56

$$Pr(S = U_i, M = S_l \mid P = U_j, QFN \in S_l)$$
$$= \frac{1}{\lambda \cdot (K - 1)} \quad if\ i = j\ and\ j \neq QFN \tag{57}$$

$$Pr(S = U_i \mid M = S_l, P = U_j, QFN \in S_l)$$
$$= \frac{(Pr(S = U_i \mid GQFN \in S_l) Pr(M = S_l, P = U_j \mid S = U_i, QFN \in S_l))}{Pr(M = S_l, P = U_j \mid QFN \in S_l)} \tag{27}$$

$$Pr(M = S_l, P = U_j \mid S = U_i, QFN \in S_l)$$
$$= Pr(M = S_l \mid S = U_i, QFN \in S_l)$$
$$\cdot Pr(P = U_j \mid M = S_l, S = U_i, QFN \in S_l) \tag{28}$$

$$Pr(M = S_l \mid S = U_i, QFN \in S_l) = \frac{1}{\lambda} \tag{29}$$

$$Pr(S = U_i, M = S_l \mid P = U_j, QFN \in S_l)$$
$$= \frac{Pr(P = U_j, M = S_l \mid S = U_i, M = S_l) \cdot Pr(S = U_i, M = S_l)}{Pr(P = U_j, QFN \in S_l)} \tag{40}$$

$$Pr(S = U_i, M = S_l) = Pr(S = U_i) \cdot Pr(M = S_l \mid S = U_i) \tag{41}$$

$$Pr(S = U_i) = \sum_{i=1}^{b} Pr(M = S_l) \cdot Pr(U_i \mid M = S_l)$$
$$= \sum_{i=1}^{b} \frac{1}{b} \cdot \frac{1}{k}$$
$$Pr(S = U_i) = \frac{r}{b \cdot k} \tag{42}$$

$$Pr(M = S_l \mid S = U_i) = \frac{r}{b} \tag{43}$$

Equating 44, 48, 52, and 56

$$= \frac{1}{\lambda \cdot (K-2)} \quad if\ i \neq j\ and\ j \neq QFN \tag{58}$$

When *QFN* and source belongs to the same group the probability of linking query is given by (57) and (58). Additionally, based on the previously-considered BIBD configuration values, the probability of linking query with the users is $\frac{1}{4}$ and $\frac{1}{2}$ respectively.

MG-OSLo provides better local privacy as compared to the memory-based multi-group distributed protocols in the following ways. The query and results stay hidden from the group users, in contrast to the existing protocols [16]–[18]. Unlike the UPIR proposed by Stoke *et al.* [18], the user gets an answer for each query without waiting for the memory location to get free. The user is not required to contact another user anonymously to proxy a query on his or her behalf. MG-OSLo is dynamic compared to other memory-based multi-group protocols because different users may group each time the protocol executes.

D. PROFILE PRIVACY

The profile privacy validates the level of profile obfuscation by simulating a privacy-preserving protocol. In protocol (MG-OSLo), a user's profile is obfuscated by sending other users' queries to the WSE. A privacy metric termed as Profile Exposure Level (PEL) is used to measure the magnitude of profile obfuscation. PEL measures the difference between the user's original profile and the obfuscated profile. The users' original profile is built from queries sent directly to the WSE without executing the privacy-preserving protocol. In contrast, the obfuscated profile is constructed from queries after implementing the privacy-preserving protocol. An experiment consisting of three steps is performed to measure the level of profile obfuscation: i. simulating the MG-OSLo with the dataset mentioned in the section below. ii. Building the user profile, iii, measuring the profile privacy with PEL relative to the WSE. To compute the impact on profile privacy and to have a fair comparison with the state-of-the-art distributed privacy-preserving protocol, the MG-OSLo is simulated for two situations; a) self-query submission is allowed, b) self-query submission is not allowed. We have experimented by changing the group size (number of users in a group) and group count (number of groups). The MG-OSLo is simulated with three users, four users, & five users in each group with the group count of three, four, five, and six groups.

TABLE 1. Attribute description.

AnonID	Anonymous ID, Distinctively ascertain users in the query log
Query	The query contents submitted to the AOL search engine.
Query Time	Temporal information including date and time of the query.
ItemRank	The rank assigned to each clicked URL
ClickURL	Address of the clicked URL

1) DATASET

America Online (AOL) released a query log of more than 650 thousand users for research [10], [26], [29]. The query log consists of around twenty million queries generated by the users in three months from March 2006 through May 2006 [30], [31]. The users were unaware of their queries being released and freely accessible [2]. Before releasing the queries, AOL had pseudo-anonymized the query log so that attackers could not link queries back to the originator. AOL achieved the anonymization of the query log by removing all identifiers and personal information such as the name, email address, IP address, etc. The AOL query log dataset consisted of five attributes: i.e., AnonID, Query, Query Time, ItemRank, and ClickURL [10]. AOL query log is considered the primary experimental data source in query log privacy [32]. Researchers in the field of Web search privacy have extensively used the AOL query log. Tab. 1 shows the attributes of the AOL query log along with its description. Piddinti and Saxena worked on the AOL query

$$Pr(P = U_j, QFN \in S_l) = \frac{\lambda}{b} \cdot \frac{r}{b \cdot k} \tag{48}$$

$$Pr(P = U_j, QFN \in S_l \in S = U_i, M = S_l) = Pr(P = U_j \mid S = U_i, M = S_l) \cdot Pr(QFN \in S_l \mid S = U_i, M = S_l) \tag{49}$$

$$Pr(P = U_j \mid S = U_i, M = S_l) = \frac{1}{K-1} \quad if\ i = j\ and\ j \neq QFN \tag{50}$$

$$= \frac{1}{K-2} \quad if\ i \neq j\ and\ j \neq QFN \tag{51}$$

$$Pr(QFN \in S_l \mid S = U_i, M = S_l = \frac{Pr(M = S_l \mid S = U_i, QFN \in S_l) \cdot Pr(QFN \in S_l \mid S = U_i)}{Pr(M = S_l \mid S = U_i)} \tag{52}$$

$$Pr(M = S_l \mid S = U_i, QFN \in S_l) = \frac{1}{\lambda} \tag{53}$$

$$Pr(QFN \in S_l \mid S = U_i) = \frac{\lambda}{b} \tag{54}$$

$$Pr(M = S_l \mid S = U_i) = \frac{r}{b} \tag{55}$$

TABLE 2. Dataset: range of queries sent by a user.

Queries range	Number of User
20-30	210
31-40	180
41-50	245
51-100	140
101-200	105
201-300	50
300-400	30
400-600	20
600-1515	20

log and analyzed its different aspects; the statistics show that 98.72% have performed less than 100 searches over three months [8], [15]. About 70% of the users have sent less than 30 queries to the AOL log. In this work, to evaluate the profile privacy, experiments are performed with three-month queries of users selected from the AOL query log. To measure the profile privacy, a subset of one thousand random users from highly active users to the least active users extracted from the AOL query log. The statistical selection of the dataset for experimentation is shown in Tab 2. The chosen users have sent a minimum of 20 queries up to a maximum of 1514 queries.

2) USER PROFILE BUILDING

The profile of a user is built from the queries a user sends to the WSE. The WSEs use this profile to retrieve the personalized results, also considered a source of revenue for WSEs. The authors of [6], [15] have proposed steps to build the user profile. These steps involve the morpho-syntactic analysis and semantic analysis of the queries. Details about each step are described below.

a: MORPHO-SYNTACTIC ANALYSIS

The primary step of profile building is to identify the main topic of the query. In the morpho-syntactic analysis of the query content, Natural Language Process (NLP) techniques based on maximum entropy are used to analyze the user query syntactically. The NLP techniques like sentence detection, syntactic-parsing, tokenization, stop words removal, stemming, and part of speech tagging are followed to acquire the user's query's main category. Cohen and Dolbey described the NLP techniques to extract the query's main topic [33]. After getting the query's main subject, the next step is the semantic analysis of the query.

b: SEMANTIC ANALYSIS

The keywords acquired in the previous step are sent to DMOZ[1] to discover the query topic's hierarchy. DMOZ is an open-content directory of World Wide Web links, the community that maintains DMOZ is also known as the Open Directory Project (ODP). ODP is the largest human editable web directory preserved by a community of volunteers [6], [34]. Figure 4 shows the ODP hierarchy categorization; there are sixteen different categories at the top level (first degree)

[1] Dmoz.org

TABLE 3. Profile of a user X at different degrees.

Query	Degree 1	Degree 2	Degree 3	Degree 4
Snooker	Sports	Cue	Sports	Snooker
Rugby	Sports	Football	Rugby	Union
Java	Computers	Programming	Languages	Java
XML	Computers	Data	Formats	Markup
Honda	Recreation	Motorcycles	Models and make	Honda
Jeep	Recreation	Autos	Models and make	Jeep
Herpes	Health	Conditions and Disease	Infectious	Diseases
Boeing	Recreation	Aviation	Aircraft	Fixed

of hierarchy. There are around 1 million different categories in the ODP hierarchy. When a user sends their queries to ODP, it categorizes the user's query into a hierarchy of categories [34] e.g., a first degree, the query is categorized into one out of 16 top-level categories. At the following degree (second level of hierarchy), the query is classified into subcategories and so on so forth. Consider a user query ''mac.com'', the ODP categories this query as ''Computers: Software: Operating Systems: MacOS: Internet.'' The query ''mac.com'' at first degree is categorized as ''computers,'' at a second degree as ''Software,'' at the third degree as ''Operating Systems'', at the fourth degree at ''MacOS,'' and ''Internet'' at the fifth degree. Thus, the user whose query is ''mac.com'' will have computers, software, operating systems, MacOS in their profile. Table 3 shows an example of a few queries categorized by ODP into a hierarchy of categories. The first degree represents a more general category of a user query. Many, unlike queries, may fall under the same category at degree 1 of the ODP hierarchy. For example, Snooker and Rugby's user queries are categories as sports at degree 1 of hierarchy, although snooker is a cue sport whereas Rugby is a field game.

The syntactical analysis and semantic analysis are the two analyses applied to the queries of a user. The corresponding profile of the user is built to the first four degrees of the ODP hierarchy. Consider a user 'X' with eight queries: snooker, Rugby, Java, XML, Honda, Jeep, herpes, and Boeing. When those queries are sent to the ODP, it categorizes X's queries into a hierarchy of categories, as shown in Tab. 3. The 'X' profile at the first degree contains ''Sports, sports, computer, computer, Recreation, Recreation, Health, Recreation, and Health.'' The 'X' profile at the second degree contains the categories: ''Cue, Football, Programming Language, Data formats, Motorcycle, Autos, Condition and disease, aviation, Condition and disease.'' Similarly, the profile of 'X' at the third-degree contains categories: ''Sports, Rugby, Java, Markup, makes and model, Makes and model, aircraft, and immune.''

In this work, two types of user-profiles are built. The first profile is called the original profile built from the user's original queries without executing any privacy-preserving protocol. An obfuscated profile is the second type of profile built

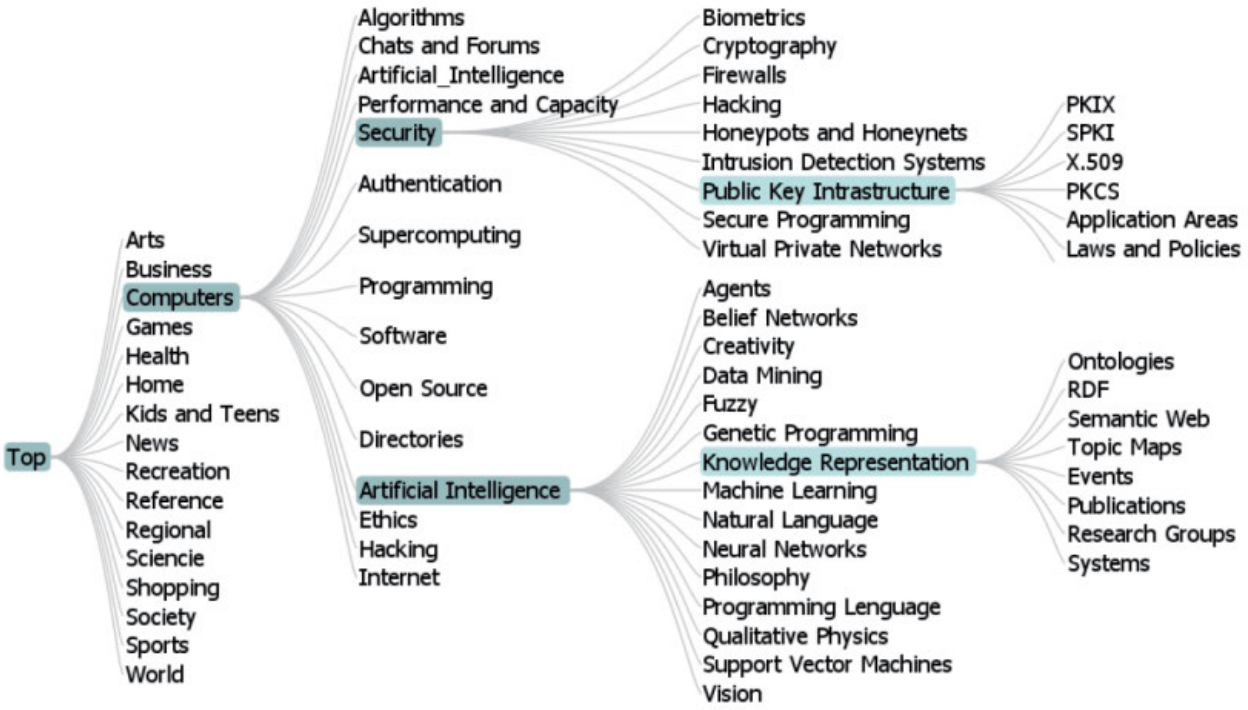

FIGURE 4. ODP hierarchy of categories [11].

from the queries that a user sends to the WSE after executing the privacy-preserving protocol. Romero-Tris *et al.* demonstrated that ODP categories provide a consistent, sufficient specificity level to evaluate a profile at the second degree [6]. However, we exhibit the results for the first, second, third, and fourth degrees of ODP hierarchy for exhaustiveness.

3) PROFILE EXPOSURE LEVEL (PEL)

The authors of [6], [11]–[15] have used PEL to measure the profile privacy achieved by a user relative to the WSE. PEL uses mutual information and entropy to measure the level of user profile exposure.

$$PEL = \left(\frac{I(M,N)}{H(M)} * 100 \right) \tag{59}$$

Viejo and Castella-Roca have defined M, and N as random variables having a sample space ΩM and ΩN [12]. M represents a set of categories of queries that a user generates, N represents a set of categories of queries that a user sends to the WSE. N frequently contains the queries of other users' categories. $H(M)$ is the entropy of M,

$$H(M) = -\sum pr(m_i).(log_2)pr(m_i) \tag{60}$$

where, $H(M)$ is the entropy of M, $I(M,N)$ is the mutual information

$$I(M,N) = H(m) - H(M \mid N) \tag{61}$$

$$I(M,N) = \sum_{m,n} pr(m \mid n).pr(n)log_2 \left(\frac{pr(m \mid n}{pr(n)} \right) \tag{62}$$

TABLE 4. Simulation details under controlled environment: simulation tools.

Computer	CPU	Intel(R) Core(TM) i3-231M
	RAM	8 Gbytes
	OS.	Microsoft Windows 8.1 pro 64-bit
	Java version	Java(TM) SE Runtime Environment (1.8.0_65)
Network	4Mb internet connection	Java Multi-threading

$H(M|N)$ is the conditional entropy. $pr(m)$ and $pr(n)$ are the probabilities of each element of M and N proportional to its cardinal. In this work, PEL is used to measure the percentage of information exposed when the user forwards other users' queries.

V. RESULTS AND DISCUSSION

This section gives a detailed description of experiments performed to compute the profile privacy by simulating MG-OSLo. A java-based simulator is developed to execute MG-OSLo using multi-thread socket programming to create multiple groups of MG-OSLo. The CryptoUtil library and keypair generator methods are exercised to create RSA public-private key pairs for query and result encryption. The experiments are performed over Intel(R) Core(TM) i3-231M CPU with 8192MB RAM over Windows 8.1 Pro 64bits. Table 4 shows the details of the simulation equipment used to execute the MG-OSLo. As mentioned above, the MG-OSLo is simulated for two situations, i.e., i) self-query submission is allowed, ii) self-query submission is not allowed. In the first situation, a user forwards one of his/her queries when

forwarding other group users' queries. Whereas, in the latter case, the user only delivers the queries of group users but not his/her query to the WSE. The user's profile privacy attained by executing MG-OSLo is compared with the state-of-art OSLo [15], and Co-utile protocol [22] for the first situation. Similarly, for the second situation, the profile privacy is compared with UUP(e) [6] and OSLo [15].

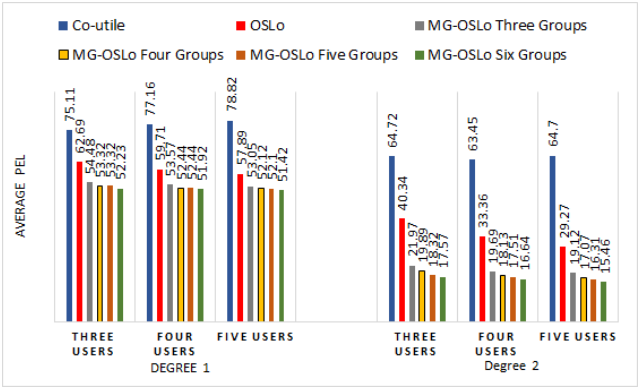

FIGURE 5. Average PEL of MG-OSLo VS OSLo VS Co-utile at Degree 1 and Degree 2 of ODP hierarchy.

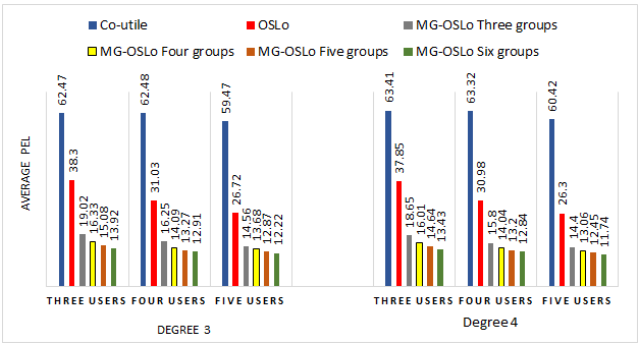

FIGURE 6. Average PEL of MG-OSLo VS OSLo VS Co-utile at Degree 3 and Degree 4 of ODP hierarchy.

A. SELF-QUERY SUBMISSION ALLOWED

Figures 5 and 6 show the profile privacy comparison of MG-OSLo, Co-utile, and OSLo for a situation where self-query submission is allowed. The result indicates that MG-OSLo achieves an average PEL of 54.48% for the group count of three, each having three users at degree 1 of the ODP hierarchy. When the group count increased to four, the value dropped to 53.32%. The results depict that increasing the group count declines the average PEL value to 52.23%. Likewise, when the group size is increased to four users, the results illustrate that MG-OSLo attains an average PEL of 53.57% for the group count of three. Further extending the group counts decreased the average PEL to 52.44% and 51.92% for the group size of four users. Similarly, at higher degrees (degree 2, degree 3, and degree 4) of the ODP hierarchy, the average PEL inversely affects when the group count and group size increase. Furthermore, the results show that the average PEL of a user simulating OSLo and Co-utile for the group size of three users are 62.69% and 75.11% at degree 1 of the ODP hierarchy. The MG-OSLo succeeds 13.09% and 27.45% better profile privacy than OSLo and Co-utile protocol for the group count of three. When the group count increased to four with the same three user groups, the MG-OSLo attained 14.94% and 29% better profile privacy than OSLo and Co-utile protocol. Similarly, once the group size is raised to four users, the average PEL value for OSLo dropped to 59.71% and 77.16% for Co-utile. In comparison, the average PEL values of MS-OSLo for the group count of three, four, five, and six decreased to 53.57%, 52.44%, 52.34%, and 51.92%. The average PEL declined when the number of users increases in the group both in OSLo and MG-OSLo. Moreover, for the five users' group size, a user achieves an average PEL of 53.05% with a three-group count, 52.11% with a group count of four, and so on by executing MG-OSLo, approximately 8.34% better than OSLo. Likewise, at degree 2, degree 3, and degree 4 of the ODP hierarchy, the MG-SOLo has a lower average PEL value than OSLo. Based on the results, MG-OSLo displayed better profile privacy than the OSLo and Co-utile for all group counts and group sizes for a situation where self-query submission is allowed.

The reason for the MG-OSLo to achieve better profile privacy compared to Co-utile is because, in the Co-utile protocol, the user (forwarding agent) only forwards the initiator's query if it muddies the profile of the forwarding user; otherwise, the query is denied. In such a case, the initiator has to forward the query himself, resulting in no obfuscation of the initiator's profile. However, a QFN in MG-OSLo has to forward all other users' queries, resulting in a more prominent obfuscation of the user's profile. Functionality (to retrieve an answer to the query) is another prime issue in the Co-utile protocol; the responder may deny the initiator's request causing a notable delay in the query answering. In contrast, a user gets an answer for every query sent in MG-OSLo. Further, A user achieves better profile privacy with MG-OSLo than OSLo and Co-utile because the profile of a user in MG-OSLo is obfuscated with a higher number of users. The chances of getting a group with users with diverse interests are much higher in MG-OSLo than OSLo.

B. SELF-QUERY SUBMISSION NOT ALLOWED

Figures 7 and 8 show the comparison based on profile privacy of a user simulating UUP(e), OSLo, and MG-OSLo for a situation when self-query submission is not allowed at degree 1 to degree 4 of the ODP hierarchy. The results for three users' group size show that MG-OSLo attains 47.40% average PEL for the group count of three at degree 1 of the ODP hierarchy. When the group count increased to four, MG-OSLo exhibited an average PEL of 48.34%. Similarly, when the group count increased to six, the MG-OSLo presented an average PEL of 50.16%. At degree 2 of the ODP hierarchy, MG-OSLo obtained a 12.58% average PEL for the group count of three. The value went slightly up to 12.60% for

TABLE 5. Security analysis of distributed protocols.

Name	Encryption type	Query Confidentiality	Answer Confidentiality	Unlinkability	Indistinguishability	Functionality (availability)	scalability
UPIR	Symmetric	Query visible to the group members	No,	Yes, but did not explained how to contact the group members	Never computed	Yes, But a user had to wait for the group member to the query from memory to proxy the query	No
UUP (e)	Asymmetric	Yes	No	Vulnerable to data mining attack	Yes	Yes	No
Co-utile	No	No	No	No	Yes	Critical issue	No
OSLo	Asymmetric and Symmetric	Yes	Yes	Yes	Yes	Yes	No
MG-OSLo	Asymmetric and Symmetric	Yes	Yes	Yes	Yes Better than previous techniques	Yes	Yes

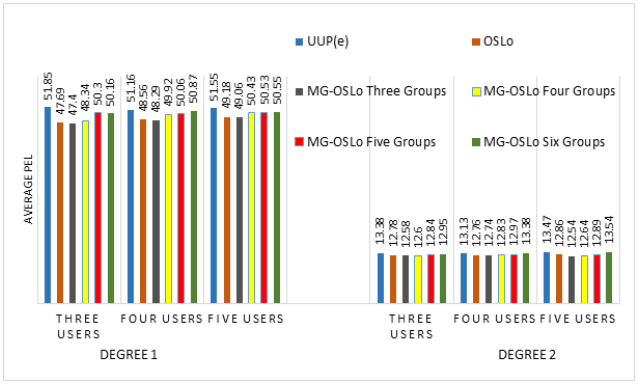

FIGURE 7. Average PEL of MG-OSLo VS OSLo VS UUP(e) at Degree 1 and Degree 2 of ODP hierarchy.

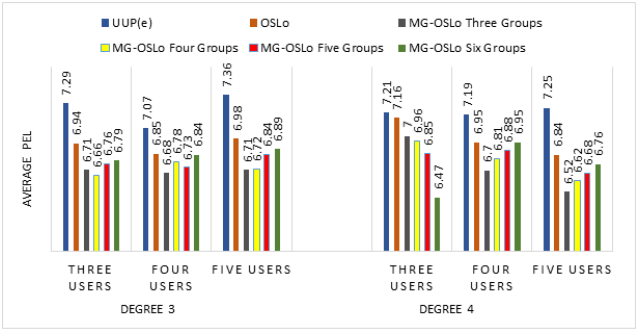

FIGURE 8. Average PEL of MG-OSLo VS OSLo VS UUP(e) at Degree 3 and Degree 4 of ODP hierarchy.

the group count of four and reached 12.95% for six groups, where each group has three users. Likewise, at degree 3 of the ODP hierarchy (representing a more specific category of the user's query), MG-OSLo attained a 6.71% average PEL for the group count of three, each having three users. The value of average PEL raised to 6.79% for the group count of six. Average PEL results at degree 4 for the three users' group show that MG-OSLo attained 7.0% for the group count of three; this value dropped to 6.47% for six groups. Besides, when four users are grouped, MG-OSLo provided a 6.70% average PEL for the group count of three and went up to 6.95% for six groups. The similarity pattern is shown for the group of five users at degree 4 of the ODP hierarchy.

When three users are grouped, a user achieves an average PEL of 51.85% when executing UUP(e) and 47.69% with OSLo as shown in Figure 7. The MG-OSLo provides 8.58% and 1.17% better profile privacy than UUP(e) and OSLo at degree 1 of the ODP hierarchy. When the group size is raised to four, UUP(e) and OSLo provided 51.16% and 48.56%, whereas the MG-OSLo result shows an average PEL of 48.29%. The MG-OSLo with three group count provides 5.59% and 0.55% better profile privacy than the UUP and OSLo. It is heeded that by increasing group count in MG-OSLo, the average PEL unexpectedly increased. The results infer that when self-query submission is not allowed, a group count of three, each containing three users, achieves the minimum average PEL because the user's profile is obfuscated to its maximum level. The results show that further increasing the group count or group size has no significant impact on average PEL. Instead, it slightly increases in contrast to the situation when the self-query submission was allowed. Therefore, based on the result, it is recommended that the group count of three and each group having three users provide the best results in terms of average PEL for a situation when self-query submission is not allowed. However, a user can forward more queries with a group count of four and five with a little compromised average PEL.

C. SECURITY ANALYSIS

The proposed MG-OSLo achieves confidentiality, indistinguishability, unlinkability, and availability (functionality) aspects of security. The previously distributed protocols either achieved a query's confidentiality or did not attain any confidentiality relative to the group members. Table 5 shows the security analysis of modern distributed protocols.

UPIR and Co-utile do not provide confidentiality of a query and its results. Though UUP only provides a query's confidentiality, the results remain visible to the group members. However, MG-OSLo succeeds in establishing the confidentiality of both the query and its results. The user's query is encrypted with the QFN's public key using the RSA encryption algorithm and its results with AES shared key. Furthermore, Co-utile provides no unlinkability because each user knows the exact query of the peer user, compromising a user's local privacy. Whereas MG-OSLo accomplishes the unlinkability through two levels of shuffling, breaking a link between a user and a query. Additionally, UPIR does not explain how the peer members contact each other to proxy a query to WSE after a query is written in the memory location by a user. Conversely, MG-OSLo requires no such work; instead, a query is forwarded to WSE when it reaches the desired QFN.

Moreover, based on the results mentioned in section V, MG-OSLo succeeds well in achieving indistinguishability (profile obfuscation) in a better way than the modern distributed protocols (OSLo, Co-utile, and UUP(e)). Compared with Co-utile and UPIR, MG-OSLo also accomplishes functionality (availability) by receiving answers for each query sent by a user. Also, MG-OSLo is scalable as compared to the other analogue protocols. MG-OSLo can handle a higher number of users by putting them in multiple groups.

VI. CONCLUSION AND FUTURE WORK

WSE builds a user's profile based on the queries it receives. WSE exploits the user profile for various purposes. The user's profile often contains sensitive information, and the disclosure of such information threatens the user's privacy. Among many privacy-preserving methods, a distributed protocol has the advantage that provides both local privacy and profile privacy. The existing multi-group distributed privacy-preserving protocols evaluate only privacy relative to the group users (local privacy), but they do not consider the profile privacy relative to the WSE. The query contents and results of the query are visible to the group users. This work proposes a Multi-Group ObScure Logging (MG-OSLo) to minimize the limitations in the existing techniques. The results prove that group count and group size directly affect a user's local privacy and profile privacy when the self-query submission is allowed. However, three users' group count provides the best possible result when the self-query submission is not allowed.

In the future, profile-based groups need to be created in which a user can be grouped with those users with diverse interests. The effect of a random grouping on result quality also needs to be investigated. The delay caused by query group creation, query shuffling, encryption, and result processing must be examined in the future.

ACKNOWLEDGMENT

This research was supported by Taif University Researchers Supporting Project number (TURSP-2020/231), Taif University, Taif, Saudi Arabia.

REFERENCES

[1] A. J. Biega, J. Schmidt, and R. S. Roy, "Towards query logs for privacy studies: On deriving search queries from questions," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2020, pp. 110–117.
[2] D. Pàmies-Estrems, J. Castellà-Roca, and J. Garcia-Alfaro, "A real-time query log protection method for Web search engines," *IEEE Access*, vol. 8, pp. 87393–87413, 2020.
[3] *Internet Live Stats—Internet Usage & Social Media Statistics*. Accessed: Nov. 16, 2020. [Online]. Available: https://www.internetlivestats.com/
[4] A. Cooper, "A survey of query log privacy-enhancing techniques from a policy perspective," *ACM Trans. Web*, vol. 2, no. 4, pp. 1–27, Oct. 2008.
[5] A. Raza, K. Han, and S. O. Hwang, "A framework for privacy preserving, distributed search engine using topology of DLT and onion routing," *IEEE Access*, vol. 8, pp. 43001–43012, 2020.
[6] C. Romero-Tris, J. Castella-Roca, and A. Viejo, "Distributed system for private Web search with untrusted partners," *Comput. Netw.*, vol. 67, pp. 26–42, Jul. 2014.
[7] F. Saint-Jean, A. Johnson, D. Boneh, and J. Feigenbaum, "Private Web search," in *Proc. 2007 ACM Workshop Privacy Electron. Soc.*, 2007, pp. 84–90.
[8] S. T. Peddinti and N. Saxena, "On the privacy of Web search based on query obfuscation: A case study of trackmenot," in *Proc. Int. Symp. Privacy Enhancing Technol. Symp.* Berlin, Germany: Springer, 2010, pp. 19–37.
[9] A. Arampatzis, G. Drosatos, and P. S. Efraimidis, "Versatile query scrambling for private Web search," *Inf. Retr. J.*, vol. 18, no. 4, pp. 331–358, Aug. 2015.
[10] R. Khan, A. Ahmad, A. O. Alsayed, M. Binsawad, M. A. Islam, and M. Ullah, "QuPiD attack: Machine learning-based privacy quantification mechanism for PIR protocols in health-related Web search," *Sci. Program.*, vol. 2020, pp. 1–11, Jul. 2020.
[11] A. Viejo, J. Castella-Roca, O. Bernadó, and J. M. Mateo-Sanz, "Single-party private Web search," in *Proc. 10th Annu. Int. Conf. Privacy, Secur. Trust*, 2012, pp. 1–8.
[12] A. Viejo and J. Castellà-Roca, "Using social networks to distort users' profiles generated by Web search engines," *Comput. Netw.*, vol. 54, no. 9, pp. 1343–1357, Jun. 2010.
[13] M. Ullah, R. Khan, and M. A. Islam, "Poshida II, a multi group distributed peer to peer protocol for private Web search," in *Proc. Int. Conf. Frontiers Inf. Technol. (FIT)*, Dec. 2016, pp. 75–80.
[14] M. Ullah, R. Khan, and M. A. Islam, "Poshida, a protocol for private information retrieval," in *Proc. 6th Int. Conf. Innov. Comput. Technol. (INTECH)*, Aug. 2016, pp. 464–470.
[15] M. Ullah, M. A. Islam, R. Khan, M. Aleem, and M. A. Iqbal, "ObSecure logging (OSLo): A framework to protect and evaluate the Web search privacy in health care domain," *J. Med. Imag. Health Informat.*, vol. 9, no. 6, pp. 1181–1190, Aug. 2019.
[16] C. M. Swanson and D. R. Stinson, "Extended results on privacy against coalitions of users in user-private information retrieval protocols," *Cryptogr. Commun.*, vol. 7, no. 4, pp. 415–437, Dec. 2015.
[17] C. M. Swanson and D. R. Stinson, "Extended combinatorial constructions for peer-to-peer user-private information retrieval," 2011, *arXiv:1112.2762*. [Online]. Available: http://arxiv.org/abs/1112.2762
[18] K. Stokes and M. Bras-Amorós, "Optimal configurations for peer-to-peer user-private information retrieval," *Comput. Math. with Appl.*, vol. 59, no. 4, pp. 1568–1577, Feb. 2010.
[19] A. Erola, J. Castellà-Roca, A. Viejo, and J. M. Mateo-Sanz, "Exploiting social networks to provide privacy in personalized Web search," *J. Syst. Softw.*, vol. 84, no. 10, pp. 1734–1745, Oct. 2011.
[20] J. Domingo-Ferrer, M. Bras-Amorós, Q. Wu, and J. Manjón, "User-private information retrieval based on a peer-to-peer community," *Data Knowl. Eng.*, vol. 68, no. 11, pp. 1237–1252, Nov. 2009.
[21] 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.
[22] J. Domingo-Ferrer, S. Martínez, D. Sánchez, and J. Soria-Comas, "Co-utility: Self-enforcing protocols for the mutual benefit of participants," *Eng. Appl. Artif. Intell.*, vol. 59, pp. 148–158, Mar. 2017.
[23] J. Domingo-Ferrer, S. Martínez, D. Sánchez, and J. Soria-Comas, "Co-utile P2P anonymous keyword search," in *Co-utility*. Cham, Switzerland: Springer, 2018, pp. 51–70.
[24] J. Castellà-Roca, A. Viejo, and J. Herrera-Joancomartí, "Preserving user's privacy in Web search engines," *Comput. Commun.*, vol. 32, nos. 13–14, pp. 1541–1551, Aug. 2009.
[25] Y. Lindell and E. Waisbard, "Private Web search with malicious adversaries," in *Proc. Int. Symp. Privacy Enhancing Technol. Symp.* Berlin, Germany: Springer, 2010, pp. 220–235.

[26] R. Khan, M. A. Islam, M. Ullah, M. Aleem, and M. A. Iqbal, "Privacy exposure measure: A privacy-preserving technique for health-related Web search," *J. Med. Imag. Health Informat.*, vol. 9, no. 6, pp. 1196–1204, Aug. 2019.
[27] B. N. Mandal, V. K. Gupta, and R. Parsad, "Balanced treatment incomplete block designs through integer programming," *Commun. Statist.-Theory Methods*, vol. 46, no. 8, pp. 3728–3737, Apr. 2017.
[28] R. E. Bechhofer and A. C. Tamhane, "Incomplete block designs for comparing treatments wth a control: General theory," *Technometrics*, vol. 23, no. 1, pp. 45–57, Feb. 1981.
[29] K. Kenthapadi, I. Mironov, and A. G. Thakurta, "Privacy-preserving data mining in industry," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, 2019, pp. 840–841.
[30] H. Wang, W. Liu, and J. Wang, "Achieve Web search privacy by obfuscation," in *Proc. Int. Conf. Secur. Intell. Comput. Big-data Services*. Cham, Switzerland: Springer, 2019, pp. 315–328.
[31] C. Wei, Q. Gu, S. Ji, W. Chen, Z. Wang, and R. Beyah, "OB-WSPES: A uniform evaluation system for obfuscation-based Web search privacy," *IEEE Trans. Dependable Secure Comput.*, early access, Dec. 25, 2020, doi: 10.1109/TDSC.2019.2962440.
[32] C. Carpineto and G. Romano, "A review of ten year research on query log privacy," in *Proc. 7th Italian Inf. Retr. (IIR) Workshop*, 2016, pp. 11–22.
[33] K. B. Cohen and A. Dolbey, "Foundations of statistical natural language processing," *Language*, vol. 78, no. 3, p. 599, 2002.
[34] N. Senthilkumar and P. R. Ch, "Prediction of user interest fluctuation using fuzzy neural networks in Web search," *Int. J. Intell. Unmanned Syst.*, vol. 8, no. 4, pp. 307–319, Jun. 2020.

MOHIB ULLAH received the M.S. degree from Birmingham City University, U.K., and the Ph.D. degree from the Capital University of Science & Technology, Islamabad, Pakistan. He is currently working as a Senior Lecturer with the Institute of Computer Sciences and Information Technology (ICS/IT), The University of Agriculture, Peshawar, Pakistan. He has published 15 research articles in well-reputed journals and international conferences. His research interests include the security and privacy issues associated with computer networks, WSN, and the IoT.

RAFIULLAH KHAN received the B.S. degree in computer science from Islamia College Peshawar, Peshawar, Pakistan, in 2007, the M.S. degree in internetworking and digital communication from the Institute of Management Sciences (IMS), Peshawar, in 2010, and the Ph.D. degree in computer science from the Capital University of Science & Technology, Islamabad, Pakistan, in 2020. He has been working as a Senior Lecturer with the Institute of Computer Sciences and Information Technology, The University of Agriculture, Peshawar, Pakistan, since 2011. His research interests include data mining, machine learning, web user privacy, sentiment analysis, and computer networks.

MUHAMMAD INAM UL HAQ received the M.S. degree in computer science from the University of Peshawar, Pakistan, and the Ph.D. degree in computer science from France. He is currently working as an Assistant Professor with the Department of Computer Science & Bioninformatics, Khushal Khan Khattak University, Karak. His research interests include image processing, computer vision, computer graphics, security, and privacy.

ATIF KHAN received the M.Sc. degree in computer science from the University of Peshawar, Pakistan, in 2004, and the Ph.D. degree in computer science (text mining) from Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia, in 2016. Since 2016, he has been an Assistant Professor with Islamia College Peshawar, Pakistan. His current research interests include data mining, text mining, sentiment analysis and opinion mining, recommender systems, and machine learning. He is currently a technical committee member in many international conferences. He serves as an Associate Editor for *ACM Transactions on Asian and Low-Resource Language Information Processing*. He also serves as a reviewer for many international conferences and journals.

WAEL ALOSAIMI was born in Saudi Arabia, in 1979. He received the B.Sc. degree in computer engineering from King Abdulaziz University, in 2002, and the M.Sc. degree in computer systems security and the Ph.D. degree in cloud security from the University of South Wales, in November 2016. From 2002 to 2004, he worked with Saline Water Conversion Corporation (SWCC) as an Instrument and Control Engineer. Then, he has served as a Trainer for Technical and Vocational Training Corporation (TVTC) until 2008. Next, he joined Taif University as a Teaching Assistant. It provides him with a scholarship to pursue his studies in the U.K. Since 2017, he has been an Assistant Professor with the Computer Engineering Department, Taif University. He has many publications in peer-reviewed conferences and journals. His current research interests include cloud computing, cloud security, information security, network security, e-health security, the Internet of Things security, and data science.

MUHAMMAD IRFAN UDDIN received the B.Sc. degree in computer science, the M.Sc. degree in computer science, the M.S. degree in grid computing, and the Ph.D. degree in computer science. He worked as a Postdoctoral Research Fellow. He worked as a faculty member with different institutes. He is currently working with the Institute of Computing, Kohat University of Science and Technology, Kohat, Pakistan. He has been actively involved with academia and research. He has published several articles in reputed journals and conference proceedings. He also serves as a reviewer for different journals. His research interests include machine learning, data science, deep learning, convolutional neural networks, reinforcement learning, computer vision, and parallel programming.

ABDULLAH ALHARBI received the Ph.D. degree from the University of Technology Sydney, Australia. He is currently an Assistant Professor with the Information Technology Department, Taif University. His research interests include human–computer interaction, information systems, cybersecurity, and data science.