A Dummy Query-Based User Privacy Protection Scheme in Named Data Networking

Jie Duan(✉), Wenyu Tang, Chunxia Sun, Zihao Yan, Weidan Cheng, and Chaojiang Li

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
duanjie@cqupt.edu.cn

Abstract. Named Data Networking (NDN) can distribute content efficiently due to its characteristics of content naming and in-network caching, but these characteristics also raise privacy concerns. However, the existing NDN privacy protection schemes cannot protect user's privacy completely because they neglect the correlation among NDN contents, content names and content caches. To solve this problem, this paper proposes a dummy query-based privacy protection scheme. Firstly, a similarity-based privacy metric applicable to NDN is formulated to measure the dispersion between queries, and the constraints of attacker's background knowledge in NDN, communication overhead and NDN characteristics are established. Based on the above indicators, the two-step dummy query set selection algorithms are proposed to construct dummy query set. The algorithms select the dummies that satisfy the constraint on NDN distribution. From the selected dummies, the algorithms then filter out dummies that can be identified by taking account of decentralization, anonymity and communication overheads. Security analysis shows that our privacy protection scheme can effectively resist attacks against the user privacy in NDN. Furthermore, experimental results indicate that the proposal greatly improves user privacy compared with the existing schemes.

Keywords: Named Data Networking · User privacy · Privacy protection · Dummy queries

1 Introduction

Named Data Networking (NDN) emerges as a new trend that utilizes its own characteristics (e.g., in-network caching, content naming and content name-based routing) to efficiently delivery content and improve the robustness of network [1]. These characteristics contribute to the tremendous growth of NDN in

Supported by National Natural Science Foundation of China (No. 61701058), the Key Project of State Grid of Sichuan Electric Power Company (No. 52199922005) and Yibin City Introduced High level Talents Project 2022YG05.

F. Gao et al. (Eds.): Chinacom 2023, LNICST 590, pp. 162–181, 2024.
https://doi.org/10.1007/978-3-031-67162-3_11

the fields of intelligent healthcare and transportation systems, but they raise privacy attacks. Specifically, the privacy disclosure occur in three ways. The privacy of sensitive information is exposed through the semantics of content names, the plaintext of contents [2], and the cache contents (by timing attacks) [3]. The sensitive information of users such as identities, health statuses, and addresses can be exposed, which inevitably pose privacy threats to users.

Currently, existing works on NDN privacy preservation including encryption mechanisms [4–8] and obfuscation mechanisms [9–15]. Encryption mechanisms perform different encryption operations to maintain the confidentially of contents and content names. Study [4] analyzed the trade-offs among privacy, efficiency and scalability in encryption mechanisms. Works [5–7] encrypted the full content names by homomorphic cryptography. Works [8] utilized proxy re-encryption to encrypt both the content names and contents. However, these mechanisms need to design complex matching algorithms for forward route and maintain the in-network caching characteristic of NDN. Meanwhile, encryption mechanisms have inherent problems of high computational overhead and complicated key transmission. Obfuscation mechanisms disrupt attackers by increasing the ambiguity of data. Studies [9–11] achieved time obfuscation by designing different methods of adding delays for contents that have not been cached. Studies [12–14] realized spatial obfuscation by selectively caching contents in routers. However, the above obfuscation mechanisms increased the delay in acquiring content or incurred a large amount of computation overhead. Work [15] enabled content obfuscation by mixing the constituent blocks of the target content with the blocks of obfuscated data. Then users reorganized the target content based on the metadata sent by the content provider. But the mechanism was impractical due to the fact that an additional secure channel has to be created to transmit the metadata when the data is transferred.

Furthermore, none of the above studies have adequately considered the correlation among contents, content names and content caches from the user's perspective. Specifically, contents and content names have semantic association, thus they correspond to each other. Meanwhile, cache privacy is essentially ways for attackers to obtain sensitive information through content names and contents and then link the sensitive information to their neighboring users. If the above relevance is ignored, i.e., only ensuring the confidentiality of content names and contents during transmission, or only cutting off the linkages between the content caches and neighboring users, can cause user privacy being inferred from the other means. Therefore, it is crucial to protect privacy completely by ensuring the non-inferability of contents, content names, and content caches.

Nowadays, one of the main techniques in the field of privacy preserving research is the dummy-based method, which is widely used in Location Based Service (LBS) to protect location privacy and query privacy [16]. The main idea of this method is to construct k-1 dummies and send them to the LBS server along with the real location or query to confuse attackers, where k represents the degree of anonymity. In NDN, the dummy query-based method effectively protects the contents, content names and content caches simultaneously without combining any other mechanisms. Since this method can preserve NDN characteristics and route forwarding mode without any requirement for a third party

or key sharing, it effectively addresses the problem of increased delay in content acquisition. Furthermore, the dummy-based method is lightweight [17], because it does not need to increase the burden of key management and computation of routers. Therefore, the dummy-based mechanism is a potential solution to the NDN privacy protection issue.

However, due to the in-network caching characteristic, if the traditional dummy-based mechanism is directly applied to NDN, it will produce the problems of decreased cache hit rate and increased communication overhead caused by the change of request distribution and the increase of transmitted data volume. In a constrained communication environment, the system resources at the user's devices are limited. Communication overhead will reduce the quality of service for users [18]. Moreover, attackers can utilize the background knowledge such as content freshness to increase the probability of successfully inferring user privacy, resulting in a lower level of privacy protection. To overcome the above problems, we propose a dummy query-based privacy protection scheme for NDN. Our main contributions are summarized as follows:

(1) We analyze the advantages of the dummy-based method for solving the problems of existing schemes, as well as the challenges posed by applying the dummy-based method to NDN.
(2) The above challenges are formulated as a privacy-preserving optimization model. Specifically, a privacy metric based on semantic and name similarities to extend the dispersion of the dummy queries set is proposed. Meanwhile, three aspects of constrains was considered: establishing privacy constraints to resist the attackers' inference attacks, constructing bandwidth constraints by leveraging the NDN aggregation property to limit the communication overhead, and building distribution constraints to maintain the performance of the network.
(3) The problem of solving the privacy preserving optimization model is converted into the problem of selecting a dummy query set. The initial dummy query set generation algorithm is designed to select a candidate set that satisfies the distribution constraints, and then a dummy query filtering algorithm is designed to select the dummy query set that is well-dispersed, highly anonymous and satisfies the communication overhead from the candidate set.
(4) Theoretical analysis demonstrate that our algorithms are effective resist attacks against NDN privacy. The simulations based on real-world database is conduct, and the experimental results show that the proposed scheme outperforms in enhancing the degree of privacy protection of NDN compared to the existing schemes.

2 Scenario and Problem Analysis

2.1 Attack Model

In NDN, a passive attack is launched by an attacker who eavesdrops on the network to intercept the communication messages or performs traffic analysis to

detect privacy [3]. Since only passive attacks are aimed at NDN privacy, only passive attacks are considered. Suppose there are two types of attackers as follows: (i) attackers intercepting user-requested information in the communication channel, and (ii) attackers who are neighboring nodes accessing the same router as the legitimate user, and they can launch timing attacks to probe the user-requested information. The proposed privacy preserving scheme needs to resist the inference attacks of above two types of attackers and resist collusion attacks, which increase the probability of inferring sensitive information by sharing information among attackers.

2.2 Problem Description

The attack process against user privacy in NDN is shown in Fig. 1. Alice sends an interest packet with content name /University/Alice/video/student/v1/s2. Alice's privacy will be leaked through content names, contents and content caches as follows: (i) the way from content names is that attacker A intercepts and traces the interest packet and the data packet, then deduces Alice's location (i.e., university), the real identity (i.e., Alice), and the occupation (i.e., student) from the semantics of the content name, (ii) the way from contents is that attacker A obtains the content directly from the data packet, and (iii) the way from caches is that attacker B launches a timing attack. Specifically, at first, attacker B measures the round-trip delay RTT_S for obtaining content from the source server (by requesting low-popularity data that is not cached in the network). Then attacker B measures the round-trip delay RTT_A for obtaining content from the nearest router A (by requesting a data twice, and recording the second round-trip delay as RTT_A). Attacker B also requests the target data and records its round-trip delay as RTT_B. Finally, it is inferred whether Alice has requested the target data by comparing the relationship between the three delays, as shown in step 4 in Fig. 1. Therefore, there is an urgent to protect content names, contents and caches in NDN to increase the level of privacy protection, as well as to solve the problems of the existing schemes, i.e., increased computational load on routers and increased latency of data fetching.

Therefore, it is urgent need to protect contents, content names and content caches in NDN to protect privacy completely. Therefore, as shown in Fig. 1, compared to the existing mechanisms, the dummy query-based method has the following advantages: (i) users generates dummy queries locally and attackers cannot identify the real query, even if they obtain all queries in the communication channel and caches, (ii) no trusted third parties are required and no additional latency is added, so that this method can preserve NDN characteristics and route forwarding mode, and (iii) it is lightweight, this method eliminates the complex operations of encryption and decryption and effectively reduces the burden on routers. In summary, it is an effective potential solution to protect user privacy in NDN.

However, applying dummy query-based directly to NDN faces the following three challenges:

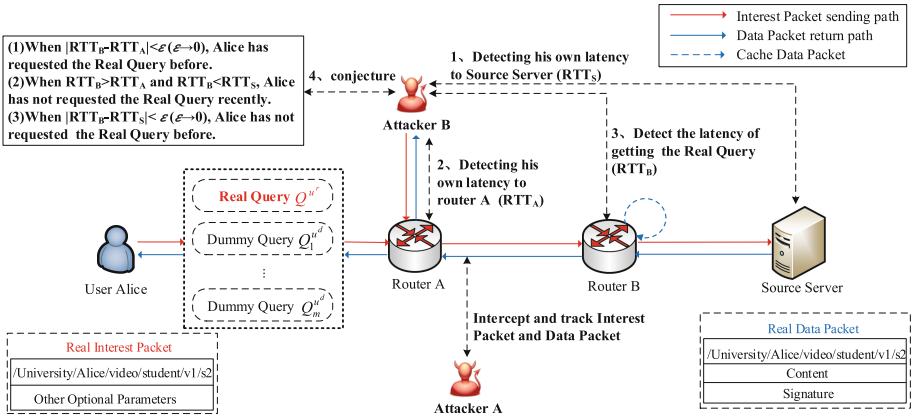


Fig. 1. The process of attacking user privacy in NDN

(1) How to construct high-quality dummy queries while ensuring NDN characteristics. This is an inherent challenge in applying the dummy query-based approach.
(2) How to address network performance degradation caused by the altered probability distribution of requests across the network. Since data request probability in the NDN satisfy a specific distribution, sending additional dummy queries will result in the distribution being altered.
(3) How to limit the communication overhead caused by additional dummy queries. Additional dummy queries increase communication overhead. Therefore, a dummy query-based scheme is proposed to effectively address the above three challenges.

3 Principles of Constructing Dummy Query Set

To address the challenges of privacy preservation in NDN, the principles of constructing the dummy query set DQS are proposed.

Since the content names of NDN identifies the only interest packet or data packet, it is considered first. Content names are categorized into two types: (i) flat names, which are hash strings with no hierarchy and no semantics, (ii) and hierarchical names, which are sequences of strings with multiple hierarchies and possible semantics. Broadly speaking, flat names can be regarded as hierarchical names with level 1 and no semantics, so both of them are seen as hierarchical names in this paper. Let $N = I_1/I_2/\cdots/I_i$ denotes a content name, where i represents the number of levels, $i \geq 1$, and each level contains one lexical element I_i, and user u send a real query Q^{u^r} with content name N^r. Let D denotes the content obtained from the data packet. Then let $DQS_u = \{N_1^{u^d}, N_2^{u^d}, \cdots, N_{k-1}^{u^d}\}$ be the dummy query set for u when he send Q^{u^r}, where $N_{k-1}^{u^d} = I_1^d/I_2^d/\ldots/I_j^d$. Then, the anonymous query set sent to the network by u is $FQS_u = \{N^{u^r}, N_1^{u^d}, N_2^{u^d}, \cdots, N_{k-1}^{u^d}\}$.

3.1 Privacy Metric

Since similarities exist among the content names, they are categorized into semantic and non-semantic contents name to measure the similarity.

Semantic Similarity. For semantic content names, when the content names of real query and the dummy query belong to the same type of semantics, the attacker can infer the semantics of the real query from them. For example, users query the traffic condition near the hospital and the location information of the clinic, the hospital and the clinic are related to health. So the attacker can infer user's personal health condition. Therefore, it is necessary to ensure that the semantic relationship between any two content names in FQS_u is as small as possible.

A WordNet semantic tree [19] which is formed by connecting the words according to semantic relationships is first constructed to portray the semantic similarity. Only the "is-a" semantic relation in WordNet semantic tree is required when measure semantic similarity [20], thus we calculate semantic similarity only based on the "is-a" structure in WordNet.

The method that is proposed by Wu and Palmer [21] is used to calculate the semantic similarity due to its effectiveness. Thus the semantic similarity between the I_i^x of N^x and the I_j^y of N^y is defined as follows:

$$Sim(I_i^x, I_j^y) = \frac{2 \times H}{len(I_i^x, lca(I_i^x, I_j^y)) + len(I_j^x, lca(I_i^x, I_j^y)) + 2 \times H} \tag{1}$$

where $lca(I_i^x, I_j^y)$ represents the nearest common parent node of I_i^x and I_j^y; len represent the distances between two nodes, respectively; H represents the distance from $lca(I_i^x, I_j^y)$ to the root node of the semantic tree.

The larger the $Sim(I_i^x, I_j^y)$, the higher the semantic similarity between I_i^x and I_j^y , $Sim(I_i^x, I_j^y) \in [0, 1]$. The semantic similarity between I_i^x and I_j^y is computed, so that the maximum resulting is taken to be the semantic similarity between N^x and N^y, as defined by the formula below:

$$Sim_s(N^x, N^y) = \max_{I_i^x \in I^x, I_j^y \in I^y} \left[Sim(I_i^x, I_j^y) \right] \tag{2}$$

Name Similarity. For non-semantic content names, if there are identical lexical elements in content names, it can exposes sensitive information, e.g., the same suffix indicates that the attributes of the two data are the same. To better distinguish any two content names intuitively, the degree of similarity between strings of content names, i.e., name similarity, needs to be considered. It is important to ensure that the name similarity between any two content names as small as possible.

The naming rule defines that the content name of NDN consists of three parts: the global routable field, the content field and the function field. The global routable field is used for routing forwarding; the content field and the

function field contain the type of version information, etc. In addition, Pending Interest Table (PIT) can aggregate interest packets, transmitting redundant data. However, PIT aggregation can be realized if the global routable field in multiple content names are the same. Therefore, only the content field and the function field are calculated to realize PIT aggregation as much as possible.

The generalized Jaccard-based similarity coefficient [22], generally denoted by EJ, is mainly used to compare the similarity and difference between a finite set of samples. The generalized Jaccard-based similarity coefficient is used to calculate name similarity, we have:

$$Sim_n(N^x, N^y) = EJ(N_o^x, N_o^y) = \frac{\overrightarrow{N_o^x} \times \overrightarrow{N_o^y}}{\left\|\overrightarrow{N_o^x}\right\|^2 + \left\|\overrightarrow{N_o^y}\right\|^2 - \overrightarrow{N_o^x} \times \overrightarrow{N_o^y}} \tag{3}$$

where N_o represent the content field and the function field of N; $\overrightarrow{N_o}$ represent N_o vectors. $Sim_n(N^x, N^y) \in [0, 1]$, the larger the $Sim_n(N^x, N^y)$, the higher the name similarity between N^x and N^y.

Similarity Metric. From formulas (1) and (3), both semantic similarity and name similarity need to be as small as possible. Then the similarity between any two content names can be represented by the following formula:

$$Sim(N^x, N^y) = \gamma \times Sim_s(N^x, N^y) + (1 - \gamma) \times Sim_n(N^x, N^y) \tag{4}$$

The parameter γ is used to control the weight of semantic similarity and name similarity, $\gamma \in [0, 1]$.

Since it is necessary to consider the similarity relationship between real query and dummy query, as well as dummies query and dummy query. The goal is to make the similarity between any two queries in the anonymous query set FQS as small as possible. Then the anonymous query set FQS needs to satisfy the following formulation:

$$FQS = \arg\min S_{mul} = \arg\min Sim(N^x, N^y) \tag{5}$$

where S_{mul} is the similarity product; N^x, N^y are the content names in the anonymous query set FQS.

3.2 Privacy Constraints

The privacy constraint mainly ensures that the attacker cannot distinguish the real query from the dummy query by the background knowledge, which contains user history query probability and content freshness.

User History Query Probability. User history query probability represents the history query set of a single user and his preferences. Attackers are more likely to believe that a user's real query is the query that the user has frequently

accessed before. Therefore, it is necessary to keep the user historical query probability similar with real query Q^{u^r} and dummy queries Q^{u^d}. Then we have:

$$\left|P_m^d - P^r\right| \leq Ph \tag{6}$$

where P_m^d and P^r represents the user history query probability of m-th dummy query $Q_m^{u^d}$ and Q^{u^r}; Ph represents the threshold of the maximum acceptable probability gap between P_m^d and P^r. Ph is a positive number very close to 0, which can be customized by the user.

Content Freshness. In NDN, the content delivery method adopts the "publish-request-response" model. The content provider first publishes the data, then the user sends an interest packet to request the data. In this process, the difference between the system time T_q^r for sending the interest packet and the time T_p^r for generating the data packet by the content provider is called the content freshness, which also represents the survival time of the data. The smaller the content freshness is, the better the timeliness of the content is represented. Then the content freshness T^r of Q^{u^r} is represented by the following formula:

$$T^r = T_q^r - T_p^r \tag{7}$$

Since a dummy query is not equivalent to a fake query, but a real query that exists in the network, thus it also has content freshness. Similarly, the content freshness T_m^d of $Q_m^{u^d}$ can be expressed as follows:

$$T_m^d = T_{q_m}^d - T_{p_m}^d \tag{8}$$

where $T_{q_m}^d$ represents the system time for sending the m-th dummy interest packet; and $T_{p_m}^d$ represents the time for the content provider to generate the data corresponding to the m-th dummy interest packet.

When constructing dummy query set, if the content freshness T^r of Q^{u^r} is much smaller than the content freshness T^d of $Q_m^{u^d}$, then the attacker is likely to exclude the dummy query based on its low timeliness to recognize the real query. Therefore, only when T_m^d and $T_{p_m}^d$ are close to each other, attackers cannot distinguish them easily. Then the T_m^d and $T_{p_m}^d$ satisfy the following equation:

$$\left|T^r - T_m^d\right| \leq Th \tag{9}$$

where Th represents the threshold of the maximum acceptable content freshness gap between T^r and T_m^d. And Th is a positive number very close to 0, which can be customized by the user.

Since Q^{u^r} and $Q_m^{u^d}$ are sent by user at the same time when the dummy query method is used, it should satisfy the equation $T_q^r = T_{q_m}^d$. Therefore, using the above equation and formulas (7), (8) and (9), we get:

$$\left|T_{p_m}^d - T_p^r\right| \leq Th \tag{10}$$

3.3 Bandwidth Constraint

A large number of interest packets and data packets take up a large bandwidth, resulting in increased communication overhead. Due to the size of the data packet is much larger than the size of the interest packet, the impact of the data size of interest packets on network traffic is negligible.

The amount of data transmitted can be reduced by collaborating with the neighboring users. Specifically, if a neighboring user u' has sent a request Q in period t, implying that the interest packet of Q has been recorded in the PIT, or the data packet of Q has been cached in their access router already. Due to the characteristics of in-network caching and the PIT aggregation of NDN, user u also sends request Q in period t, u get the data packet from the access router, effectively reducing the data traffic in the whole network. Assuming that u has neighboring users connected to the same router. Let the set of neighboring users of u is $U = \{u_1, u_2, \cdots, u_\omega\}$. In this case, the anonymous queries set that sent by u_ω in period t is FQS_{u_ω}. Then, the total anonymous query set that sent by U is $QS_U = FQS_{u_1} \bigcup \cdots \bigcup FQS_{u_\omega}$. Therefore, the set of repetition factors of u corresponding to U in period t is defined as $\lambda(u, U) = \{\lambda_1(u, U), \lambda_2(u, U), \cdots, \lambda_{k-1}(u, U)\}$, where each repetition factor $\lambda_i(u, U)$ is computed as follows:

$$\lambda_i(u,U) = \begin{cases} 1, & \text{if } N_i^{u^d} \notin QS_U \\ 0, & \text{if } N_i^{u^d} \in QS_U \end{cases} \quad (1 \le i \le k-1) \tag{11}$$

Then the bandwidth constraint is given as:

$$\sum\nolimits_{i=1}^{k-1} \lambda_i(u,U) \cdot b_i \le B \tag{12}$$

where B represents the threshold of link transmission bandwidth; b_i denotes the link bandwidth for transmitting the i-th data packet of DQS_u.

3.4 Distributional Constraint

Content popularity reflects the request probability of contents. Then content popularity is expressed as $P_\tau = \frac{C}{\tau^\sigma}$. Where τ and P_τ represents the ranking and frequency of the frequency of occurrence of the content corresponding to N_τ, respectively; C is the normalization factor; and σ is a parameter between 0.6 and 1. To construct DQS_u, it is necessary to ensure that the content popularity of Q^{u^r} and $Q_m^{u^d}$ are similar, thus confusing attackers. The similarity of the content popularity can be measured by the proximity of rank and we have:

$$\left|\tau_m^d - \tau_r\right| \le Ih \tag{13}$$

where τ_m^d and τ_r represent the rankings of $Q_m^{u^d}$ and Q^{u^r}, respectively; Ih is the threshold of the maximum acceptable ranking gap between $Q_m^{u^d}$ and Q^{u^r}.

The probability of content requests of all data in the NDN follow the zipf distribution. If dummy queries are generated in a randomized manner, the content popularity will no longer follow the original distribution. Attackers can distinguish dummy queries by comparing the original distribution with the current distribution through long-term observation, leading to the exposure of the real query. In addition, the change of distribution reduce the content caching hit rate and user service quality. Therefore, in order to keep the distribution unchanged, we have:

$$\frac{q}{\sum_{i=0}^{n} q_i} = \frac{q'}{\sum_{i=0}^{n} q_i{}'} \tag{14}$$

where $q/\sum_{i=0}^{n} q_i$ represents the request probability of the content before u constructs dummy queries; $q'/\sum_{i=0}^{n} q_i{}'$ represents the request probability of the content after the i-th query.

3.5 Problem Modeling

The goal is to construct k-1 high-quality dummy queries to form a dummy query set DQS for NDN privacy preservation. To achieve this goal, The privacy metric and constraints are formulated as an optimization problem as follows:

$$\begin{aligned} \min \prod_{x=1,x<y\le k}^{k} \Bigg\{ & \gamma \times \max_{I_i^x \in I^x, I_j^y \in I^y} \left[Sim(I_i^x, I_j^y) \right] \\ & + (1-\gamma) \times \frac{\overrightarrow{N_o^x} \times \overrightarrow{N_o^y}}{\left\| \overrightarrow{N_o^x} \right\|^2 + \left\| \overrightarrow{N_o^y} \right\|^2 - \overrightarrow{N_o^x} \times \overrightarrow{N_o^y}} \Bigg\} \\ & s.t., (6), (10), (12), (13), (14), \gamma \in [0,1] \end{aligned} \tag{15}$$

The optimization objective established is to minimize the privacy measure, i.e., the similarity product S_{mul}. Thus it can maximize the degree of regional dispersion. DQS should has the highest dispersion, the smallest probability of privacy leakage, and satisfies the overhead within a certain range.

4 Dummy Query Set Selection Algorithms

In order to obtain the optimal dummy query set, two algorithms are designed as follows: (i) dummy query distribution assurance algorithm. From constraints (13) and (14), the goal is to constitute initial dummy query set DQS_0 that do not change the original distribution and have similar content popularity of $Q_m^{u^d}$ and Q^{u^r}. (ii) dummy query filtering algorithm. The k-1 dummy queries with the largest degree of dispersion satisfying constraints (6), (10) and (12) are filtered out from DQS_0 to constitute the dummy query set DQS_u. Then obtain the final query set FQS_u that combine the real query and DQS_u. Additionally, the above two algorithms should not be switched in order, otherwise the gap between the content popularity of $Q_m^{u^d}$ and Q^{u^r} will increase, thus the probability that the attacker speculate the real query will increase.

4.1 Dummy Query Distribution Assurance Algorithm

The dummy query distribution assurance algorithm is shown in Algorithm 1. The interval threshold H affects the effectiveness of privacy preservation, $H \geq Ih$. Algorithm 1 preserve network characteristics. Lines 2–3 represent that when the number of required dummy queries is less than the number of queries in H, then all the queries in the interval are satisfied and all of them are added to DQS_0, otherwise, lines 4–13 of the algorithm are executed. In this case, the query that is closest to the real query in ranking is added to DQS_0 by executing lines 4–8 or 12.

Algorithm 1. Dummy Query Distribution Assurance Algorithm

Input: The maximum value of all rankings $\tau_{\max}$, number of dummy queries in the initial dummy query set $s_{\max}$, content request probability ranking table, the interval threshold H, the number s_{ih} of all queries in the interval threshold H

Output: Initial dummy query set DQS_0.

1: $DQS_0 \leftarrow \emptyset$, $s \leftarrow 0$, $b \leftarrow 0$, get the ranking τ of content names in real queries
2: **if** $s_{ih} \leq s_{\max}$ **then**
3: Add all queries in Ih to DQS_0
4: **else if** $\left|N_{\tau}^{u^d}\right| > s_{\max}$ **then**
5: **while** $1 \leq \tau \leq \tau_{\max}$ and $s \leq s_{\max}$ **do**
6: Randomly selecting a $N_{\tau_i}^{u^d} \in N_{\tau}^{u^d} \backslash N_{\tau}^{r}$, $s \leftarrow s + 1$
7: **end while**
8: **else**
9: **for** e **do**ach $N_{\tau_i}^{u^d}$ in $N_{\tau}^{u^d} \backslash N_{\tau}^{r}$
10: Add $N_{\tau_i}^{u^d}$ to the end of the DQS_0
11: **end for**
12: $b \leftarrow s_{\max} - \left|N_{\tau}^{u^d}\right|$, $b/2$ dummy queries are randomly selected before and after the ranking τ and added to the end of DQS_0
13: **end if**
14: **return** DQS_0

Theorem 1. *The DQS_0 obtained by Algorithm 1 conforms the data probability distribution when the number of dummy queries grows in equal proportions.*

Proof. In order to conform the distribution constant, it is necessary to show that constraint (14) holds. Since the following equation is satisfied:

$$q' = q + \sum\nolimits_{j \neq i}^{n} q_j \tag{16}$$

where, $\sum_{j\neq i}^{n} q_j$ represents the number of all other queries except the i-th true query. It is obtained by substituting the above equation into Eq. (14):

$$\begin{aligned} &\frac{q}{\sum_{i=0}^{n} q_i} = \frac{q + \sum_{j\neq i}^{n} q_j}{\sum_{i=0}^{n}\left(q + \sum_{j\neq i}^{n} q_j\right)} \\ &\Rightarrow q \cdot \sum_{i=0}^{n}\left(q + \sum_{j\neq i}^{n} q_j\right) = \sum_{i=0}^{n} q_i \cdot \left(q + \sum_{j\neq i}^{n} q_j\right) \\ &\Rightarrow \frac{q}{\sum_{i=0}^{n} q_i} = \frac{\sum_{j\neq i}^{n} q_j}{\sum_{i=0}^{n}\sum_{j\neq i}^{n} q_j},\ Q_i = \sum_{j\neq i}^{n} q_j \\ &\Rightarrow \frac{q}{\sum_{i=0}^{n} q_i} = \frac{Q_i}{\sum_{i=0}^{n} Q_i} \end{aligned} \tag{17}$$

The above equation is the identity, so that the original distribution is roughly similar to the distribution after the algorithm when the number of constituent dummy queries is increased in equal proportions.

4.2 Dummy Query Filtering Algorithm

The dummy query filtering algorithm is shown in Algorithm 2.
In lines 3–9, firstly, according to the constraints (6) and (10), $2k$ dummy queries are filtered from DQS_0 to form the candidate dummy query set DQS_1; Lines 10–21 are for filtering k-1 dummy queries from DQS_1 that minimize privacy metric and satisfy constraints (12), thus constituting the final dummy query set DQS_u.

After filtering the anonymous query set FQS_u, u sends the interest packets of FQS_u to network. Since different interest packets may be responded to by different content providers, the time and path of the returned data are different, which can also reduce the probability of path congestion when the packets are returned. In the end, u can filter out the data corresponding to his real query from all the returned data.

4.3 Algorithm Time Complexity Analysis

The proposed scheme contains two algorithms: the initial dummy query set generation algorithm and the dummy query screening algorithm. The first algorithm constructs $s_{\max}$ dummy queries to form the initial dummy query set DQS_0. Setting $s_{\max} = 4k$, then the time complexity is $O(4k)$. The second algorithm has three main iteration steps. The first one calculates the number of repetition factors by cyclic comparison, and the time complexity is $O(4k)$. The second one filters 2k dummy queries from DQS_0 to form DQS_1. Setting the number of dummy queries in DQS_1 as 2k, and the time complexity is $O(2k)$. The third one filters k-1 dummy queries from that can get the minimum S_{mul} with FQS_u. The time complexity is $O\left(k^2 \cdot (IJ+1)\right)$, where represent the maximum number of lexical elements contained in the real query or the dummy query, respectively. Since the output of the first algorithm is the input of the second algorithm, the total time complexity of the two algorithms is $O(k^2 \cdot I \cdot J)$.

Algorithm 2. Dummy Query Filtering Algorithm

Input: N^{u^r} , Wordnet semantic tree, B , content freshness table, content popularity table, k , DQS_0
Output: Final query set FQS_u

Initialize Ph , $DQS_1 \leftarrow \emptyset$, $FQS_u \leftarrow N^{u^r}$, $k \leftarrow 1$, $\psi \leftarrow 0$
Calculate $\lambda(u, U)$ and ψ of $\lambda_i(u, U) = 1$
while $|DQS_1| < 2k$ **do**
  for each N^{u^d} in DQS_0 **do**
    if $|P_m^r - P^d| \leq Ph$, $|T_{p_m}^d - T_p^r| \leq Th$ and $b_i < \frac{B}{\psi}$ **then**
      $DQS_1 = DQS_1 \cup N^{u^d}$
    end if
  end for
end while
while $|DQS_u| < k - 1$ and $b_{sum} < B$ **do**
  $\delta_{\min} \leftarrow \infty$
  for each N^x in $DQS_1 \backslash FQS_u$ **do**
    $\delta_{mul} \leftarrow$ current S_{mul} of FQS_u
    for each N^y in FQS_u **do**
      $\delta \leftarrow Sim(N^x, N^y)$, $\delta_{mul} \leftarrow \delta_{mul} * \delta$
      if $\delta_{mul} < \delta_{\min}$ **then**
        $\delta_{\min} \leftarrow \delta_{mul}$, $\xi \leftarrow x$, $DQS_u \leftarrow DQS_u \cup N_\xi^{u^d}$, $FQS_u \leftarrow FQS_u \cup N_\xi^{u^d}$, $b_{sum} = b_{sum} + \lambda_\xi(u, U) \cdot b_\xi$
      end if
    end for
  end for
end while
return FQS_u

4.4 Security Analysis

In this section the security defensibility of the algorithms against the attacks (inference attacks and collusion attacks) mentioned in attack model is verified.

Resisting Inference Attacks. Attacker A and B in attack model launch inference attacks by integrating background knowledge.

Theorem 2. *Let $negl(k)$ denote the negligible function of k. The attacker satisfies $Ph \leq negl_1(k)$ and $Th \leq negl_2(k)$ with known FQS. Our scheme can resist inference attacks.*

Proof. Ideally the probability that the attacker speculates the real query from FQS is $p_1 = 1/k$. Since our scheme satisfies formula (6). Hence, for any $1 \leq j \neq s \leq k$ in FQS, the probability p_2 that the attacker speculates the real query from the user history query probability satisfies $p_2 = |P_j - P_s| \leq Ph \leq negl_1(k)$. Similarly, from formula (10), $p_3 = |T_j - T_s| \leq Th \leq negl_2(k)$. So given the

knowledge of FQS, the probability p that the attacker speculates the real query satisfies $p = p_1 \cdot p_2 \cdot p_3 \leq negl_1(k) \cdot negl_2(k)/k \leq negl(k)$. Thus, for any two queries in FQS are indistinguishable, our scheme can resist inference attacks.

Resistance to Collusion Attacks. The collusion attack usually involves Attacker A, B and intermediate routers by sharing information to speculate the private information.

Theorem 3. *Our scheme can resist to collusion attacks.*

Proof. A user's real query will be confused with other k-1 dummy queries. The attacker cannot speculate the real query and link it to the user, even if the attacker locate the legitimate user by using the background knowledge (e.g., the network topology). Additionally, since all the information obtained by the attacker is the anonymous query set, even if information sharing is performed, there is no additional valid information can be launched for real query, i.e., the probability of attackers guessing the real query does not increase with information sharing. Thus the probability of attackers guessing the real query does not increase with information sharing, our scheme resists collusion attacks.

5 Experiments and Simulations

This section evaluates the performance of the proposed NDN privacy protection scheme.

5.1 Experimental Scenario

Experimental Data. In order to compare with existing schemes, the experiments will be conducted under NDN for LBS service queries. The experimental simulation dataset is derived from the New York Yellow Cab dataset of TLC [23] (Taxi and Limousine Commission). In the experiments, the route from the boarding point to the alighting point is taken as the query content of the vehicle, and the boarding time is taken as the query time. We use Python to implement the dummy query scheme on Windows 10 operating system, and conduct all experiments on a 12th Gen Intel(R) Core(TM) i9-12900H CPU 2.50 GHz 32 GB RAM. The specific parameters of the experiments are shown in Table 1.

Experimental Evaluation Indicators. The experimental evaluation mainly focuses on the performance of proposed algorithms in terms of the privacy protection effectiveness and cost overhead. Privacy protection effectiveness is demonstrated by the degree of decentralization and anonymity of the anonymous query set FQS_u. Cost overhead is measured by analyzing the effect of the number ψ of the repetition factor set $\lambda(u, U)$ with $\lambda_i(u, U) = 1$ and the link transmission

Table 1.

parameters	notation	value
size of the dummy query set	k	3,6,9,12,15,18
query packet size (M)	b	range in [1,300]
number of neighboring users	ω	5
time period size (s)	t	120
user history query probability gap threshold	Ph	0.05
content freshness gap threshold	Th	0.05
privacy metric weighting values	γ	0.5

bandwidth threshold B on the transmitted packet size. Then, the information entropy is defined as follows:

$$H = -\sum_{i=1}^{k} P_i \cdot \log_2 P_i = -\sum_{i=1}^{k} \frac{p_i}{\sum_{i=1}^{k} p_i} \cdot \log_2 \frac{p_i}{\sum_{i=1}^{k} p_i} \tag{18}$$

where H is the information entropy of the FQS_u; p_i is the user history query probability of the i-th query of the FQS_u; $i = 1, 2, \cdots, k$.

5.2 Experimental Result

Based on the above settings, the comparison algorithms include: (1) existing dummy-based privacy protection schemes (enhanced-DLS [24], RDG [25], and SPDDS [26]); (2) the random case (Random), which construct a dummy query set by randomly selecting the dummy query instead of taking into account the optimization problem (i.e., formulation (12)); and (3) the optimal case (Optimal), the degree of anonymity is k, and the query probability of each query in the anonymous query set is equal (i.e., they all equal to $1/k$).

Privacy Protection Effectiveness Analysis. Figure 2 demonstrates the negative logarithm of the similarity product S_{mul} with different k. At the same k, our scheme has a larger negative logarithm of the similarity product and outperforms other schemes in the degree of decentralization of FQS_u. This is because Random, enhanced-DLS and RDG do not consider semantic similarity and name similarity, resulting in higher queries similarity and lower decentralization in the set of anonymous query set. Although SPDDS guarantees semantic diversity, it ignores name similarity. Our scheme integrates the privacy metrics based on semantic similarity and name similarity, and ensures that the privacy metrics among all the queries in the FQS_u are as small as possible. Hence, our scheme forms an anonymous query set with a larger S_{mul} and a greater degree of decentralization to achieve better privacy protection.

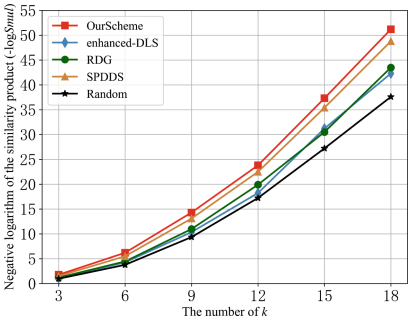
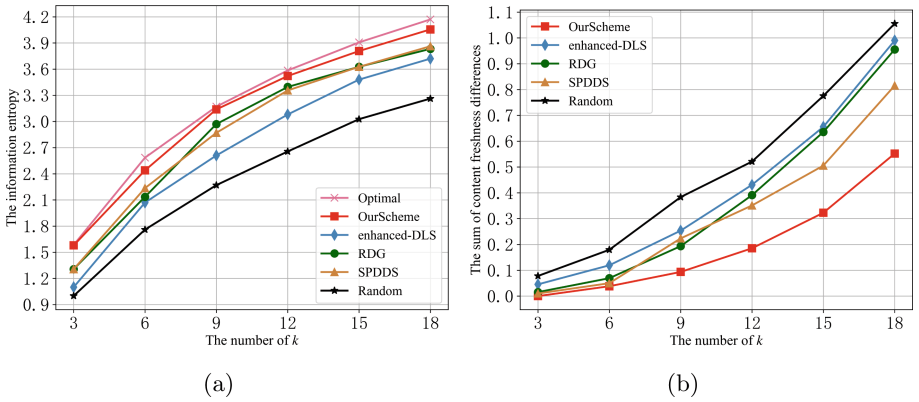


Fig. 2. Negative logarithm of the similarity product with different k

Figure 3 demonstrates the effect of k on the degree of anonymity. Specifically, the variation of information entropy with k in Fig. 3(a) and the sum of content freshness differences with k in Fig. 3(b). The degree of anonymity of our scheme outperforms the other schemes, mainly because (i) the remaining schemes consider the global content history query probability. Our scheme fully takes into account the history query probability of a certain user while considering the global content history query probability, which makes its information entropy greater than others, (ii) the remaining schemes do not consider the content freshness of the query, whereas our scheme ensures that the content freshness is similar among the dummy queries. So our scheme produces a better privacy protection effect with a better anonymity.



(a) (b)

Fig. 3. Degree of anonymity with different k. (a) The information entropy; (b) The sum of content freshness differences.

Cost Overhead Analysis. The experiments further evaluate the impact of parameters algorithms on cost overhead.

Figure 4 illustrates the variation of the total packet size of the anonymous query set FQS_u with different k. All the total packet sizes of FQS_u increase with the increase of the k. This is because the larger the k, the more dummy queries need to be sent and the total packet size of FQS_u increases. Figure 7 further shows that at the same k, our scheme has the smallest total packet size of FQS_u. And the gap of the total packet size of FQS_u gradually increases with the growth of k. The reason is that, the other schemes do not consider the impact of communication overhead, while our scheme sets the bandwidth constraint that the total packet size of FQS_u needs to be within a threshold range, effectively avoiding network congestion and excessive communication overhead.

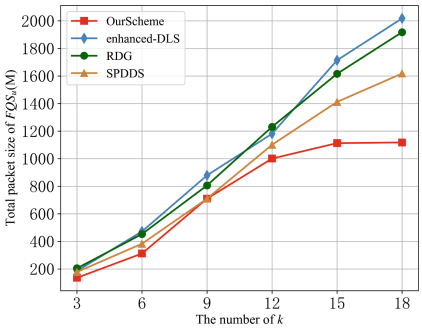


Fig. 4. Total packet size of FQS_u with different k

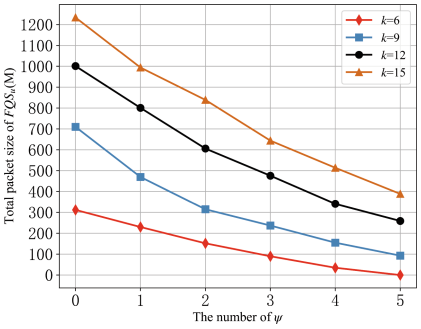
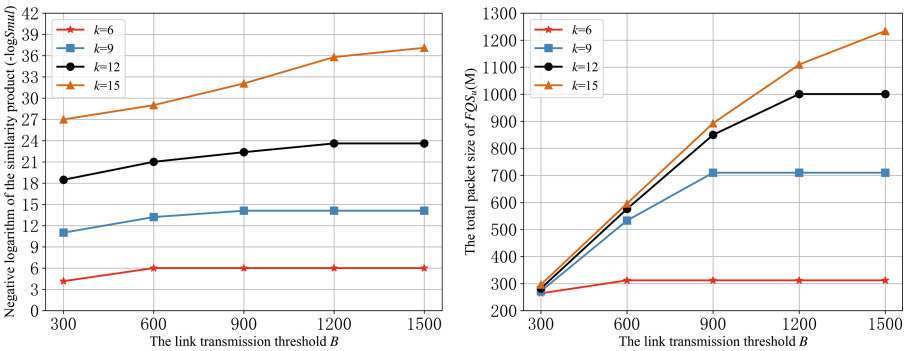


Fig. 5. Effect of ψ on the total packet size of FQS_u

Figure 5 demonstrates the effect of the number of repetition factors ψ on the total packet size of the anonymous query set FQS_u constituted by user u. It is observed that as the ψ increases, the number of dummy queries that need to be sent decreases. In addition, for the same ψ, the larger the k, the more the number of dummy queries required, thus the larger the total packet size of FQS_u. It represents that the total packet size of the anonymous query set of u keeps decreasing and the amount of data transmitted and the communication overhead will decrease. It shows that privacy and bandwidth overhead are a contradiction. When more dummy queries are constituted, the probability of attackers guessing the real query decreases. It means the better the privacy protection is, the bandwidth overhead increases. However, our scheme can take advantage of the characteristics of NDN multi-user collaboration to avoid the transmission of redundant data as much as possible by aggregating multiple identical data and utilizing the in-network caching characteristics.

Figure 6 depicts the effect of the link transmission bandwidth threshold B. Figure 6(a) reflects the variation of the negative logarithm of the similarity prod-

uct with different B. As shown in Fig. 6(a), the negative logarithm of the similarity product FQS_u gradually increases with the increase of B for different k. This is because the bandwidth limit becomes smaller as B increases, which means that the number of dummy queries with better privacy protection increases when constructing the dummy query set. Hence, the more decentralized the anonymous query set is, the better the privacy protection effect is, i.e., a certain amount of bandwidth overhead is sacrificed for a better user privacy protection effect. However, the negative logarithm of FQS_u no longer increases when B increases to a certain degree. This is because the optimal dummy query set has already been obtained and the threshold value will not have an impact on the construction of the dummy query set. Figure 6(b) reflects the variation of the total packet size of FQS_u with B. From Fig. 6(b), it can also be seen that as the B increases, the total packet size of FQS_u also shows the trend of increasing, which exhibits the constraining effect of the B. At the same time, when the B increases to a certain extent, the total packet size of FQS_u will no longer change, because the total packet size of FQS_u with the optimal degree of privacy protection already satisfies the threshold B.



(a) (b)

Fig. 6. The effect of the link transmission threshold B. (a) Negative logarithm of the similarity product; (b) The total packet size of $FQSu$.

Therefore, it is verified that our scheme is limited in increasing the communication overhead and can reduce the possibility of link congestion, which shows the feasibility of our scheme in terms of communication overhead.

6 Conclusion

This paper proposed a dummy query-based privacy protection scheme for NDN to avoid concurrently user privacy leakage from contents, content names and content caches. Specially, to increase the indistinguishability among dummy queries

and real query, the privacy, bandwidth and distributional constraints are taken into account. Based on these constraints, an optimization privacy model with the objective of minimizing the semantic and name similarity is formulated. To efficiently solve it and then construct the dummy query set, the initial dummy query set generation algorithm and dummy query filtering algorithm are proposed. Firstly the former algorithm select a set of dummy query that satisfies the distributional constrains, and then the later algorithm filters out the final dummy query set by privacy, bandwidth indicators. Theoretical analysis and experimental results show that the dummy query set generated by the scheme can effectively protect user privacy while keeping the network performance stable and the communication overhead within a certain range.

References

1. Conti, M., Gangwal, A., Hassan, M., et al.: The road ahead for networking: a survey on ICN-IP coexistence solutions. IEEE Commun. Surv. Tutorials **22**(3), 2104–2129 (2020)
2. Zhang, Z., Lung, C.H., Wei, X., et al.: In-network caching for ICN-based IoT (ICN-IoT): a comprehensive survey. IEEE Internet Things J. **10**(16), 14595–14620 (2023)
3. Tourani, R., Misra, S., Mick, T., et al.: Security, privacy, and access control in information-centric networking: a survey. IEEE Commun. Surv. Tutorials **20**(1), 566–600 (2017)
4. Zhang, Z., Won, S. Y., Zhang, L.: Investigating the design space for name confidentiality in named data networking. In: Proceedings of MILCOM 2021 IEEE Military Communications Conference (MILCOM), pp. 570–576. IEEE (2021)
5. Ko, K.T., Hlaing, H.H., Mambo, M., et al.: A PEKS-based NDN strategy for name privacy. Future Internet **12**(8), 130 (2020)
6. Guo, X., Chen, C., Zhang, M.J., et al.: Privacy-aware transmission scheme based on homomorphic proxy re-encryption for NDN. Int. J. Secure. Network. **13**(1), 58–70 (2018)
7. He, H., Chen, B.: An elliptic curve based name privacy protection mechanism for sensory data centric named data networking. In: Proceedings of 2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN), pp. 56–62. IEEE (2019)
8. Bernardini, C., Marchal, S., Asghar, M.R., et al.: PrivICN: privacy-preserving content retrieval in information-centric networking. Comput. Netw. **149**(1), 13–28 (2019)
9. Yang, J., Tang, J., Li, J.: Differential defense against distributed timing attack for privacy-preserving information centric network. In: Proceedings of 2022 IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–6. IEEE (2022)
10. Kumar, N., Srivastava, S.: A triggered delay-based approach against cache privacy attack in NDN. In: Proceedings of 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), pp. 22–27. IEEE (2018)
11. Acs, G., Conti, M., Gasti, P.: PrivICN: privacy-aware caching in information-centric networking. IEEE Trans. Dependable Secure Comput. **16**(2), 313–328 (2017)

12. Abani, N., Braun, T., Gerla, M.: Betweenness centrality and cache privacy in information-centric networks. In: Proceedings of the 5th ACM Conference on Information-Centric Networking, pp. 106–116. IEEE (2018)
13. Sivaraman, V., Sikdar, B.: A defense mechanism against timing attacks on user privacy in ICN. IEEE/ACM Trans. Netw. **29**(6), 2709–2722 (2021)
14. Jones, A., Simon, R.: A privacy-preserving collaborative caching approach in information-centric networking. In: Proceedings of 22nd International Symposium, pp. 133–150. IEEE (2020)
15. Arianfar, S., Koponen, T., Raghavan, B., et al.: On preserving privacy in content-oriented networks. In: Proceedings of the ACM SIGCOMM Workshop on Information Centric Networking (ICN), pp. 19–24. IEEE (2011)
16. Li, W., Li, C., Geng, Y.: APS: attribute-aware privacy-preserving scheme in location-based services. Inf. Sci. **527**(5), 460–476 (2020)
17. Zhao, P., Liu, W., Zhang, G., et al.: Preserving privacy in WiFi localization with plausible dummy locations. IEEE Trans. Veh. Technol. **69**(10), 11909–11925 (2020)
18. Jiang, H., Li, J., Zhao, P., et al.: Location privacy-preserving mechanisms in location-based services: a comprehensive survey. ACM Comput. Surv. **54**(1), 2373–2395 (2021)
19. Fellbaum, C., et al.: WordNet: An Electronic Lexical Database. MIT press, Cambridge (1998)
20. Liu, H.Z., Bao, H., Xu, D.: Concept vector for similarity measurement based on hierarchical domain structure. Comput. Inform. **30**(5), 881–900 (2011)
21. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: Proceedings of Association for Computational Linguistics, pp. 133–138. IEEE (1994)
22. Real, R., Vargas, J.M.: The probabilistic basis of Jaccard's index of similarity. Syst. Biol. **45**(3), 380–385 (1996)
23. TLC Trip Record Data, data publications (2022). https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data
24. Niu, B., Li, Q., Zhu, X., et al.: Achieving k-anonymity in privacy-aware location-based services. In: Proceedings of IEEE INFOCOM 2014-IEEE Conference on Computer Communications, pp. 754–762. IEEE (2014)
25. Shaham, S., Ding, M., Liu, B., et al.: Privacy preservation in location-based services: a novel metric and attack model. IEEE Trans. Mob. Comput. **20**(10), 3006–3019 (2020)
26. Yang, D., Ye, B., Chen, Y., et al.: A dummy location selection algorithm based on location semantics and physical distance. In: Proceedings of 16th International Conference of Information Security Practice and Experience, pp. 283–295. IEEE (2021)