IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 5, 2018, accepted November 24, 2018, date of publication December 5, 2018,
date of current version December 31, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2885073

Search Engine Based Proper Privacy Protection Scheme

GUIBING GUO, TIANZHI YANG, AND YUAN LIU
Software College, Northeastern University, Shenyang 110169, China

Corresponding author: Yuan Liu (liuyuan@swc.neu.edu.cn)

This work was supported in part by the National Natural Science Foundation for Young Scientists of China under Grant 61702090 and Grant 61702084, in part by the Natural Science Foundation of Liaoning Province of China under Grant 20170540319 and Grant 201602261, and in part by the Fundamental Research Funds for the Central Universities under Grant N162410002 and Grant N161704001.

ABSTRACT In the era of the "Internet of everything," people and devices are interconnected and keep exchanging private information, including confidential data about personal identity, locations, payments, transactions, and so on. The privacy leakage problem also becomes an essential security issue that challenges the management and authentication of these private information, which even restricts the communication efficiency in different "things." With the occurrence of serious information security incidents, such as the Facebook data breach event in March 2018, more and more attention has been taken into protecting private information from leakage or misuse. The existing private information protection mechanisms are mostly built on the basis of data encryption, access control, or user authentication, bearing two shortcomings: 1) strictly protecting data but sacrificing the data usage efficiency; and 2) rare consideration of the related application scenario information. In this paper, a privacy information protection scheme is proposed, where private data is protected differently based on the data attributes and application scenarios, ensuring that the private data is "properly protected". We first investigate the specific attributes of private data based on the search engine (Google) in different application scenarios, such as the data importance and dependence. Based on these attributes, private data are divided into four security domains in a given application scenario, which is further protected by applying different protection policies. By implementing the proposed scheme, it has been demonstrated that our protection scheme can balance well between the security and usage convenience of private information.

INDEX TERMS Privacy data, privacy protection scheme, proper protection.

I. INTRODUCTION

In today's society, private information has attracted widely attentions due to its ubiquitousness, which is inseparable with our daily life, work, study and entertainment. According to incomplete statistics, Facebook, a popular social network platform, has more than 2.23 billion active users by the second quarter of 2018 [1], which processes 2.5 billion messages, 500+ terabytes of data per day, and 2.7 billion hits of Likes, 300 million photos, and 105 terabytes per 30 minutes. It is conceivable that the personal privacy data involved will be large. Another scenario is the express delivery business. The average package number delivered daily by FedEx in 2018 is about 5.95 million [2], which merely covers a single express company.

In order to achieve personalized services, people always trade their privacy for convenience [3], even though more and more people sense that it is risky in imposing their personal information and gradually become unwilling to share such information [4], [5]. The basic concern is that their privacy data may not be properly protected. Serious privacy data breaches and misuse often happen in the past year. For example, thousands of FedEx customer records were exposed by unsecured servers in February 2018; 50 million Fackbook user accounts were sold to advertisers in March 2018, including user demographics, locations, interests and behaviors.

Many attempts have been investigated in order to protect private data from breaching and misuse. Many information protection schemes have been proposed in the field of encryption and user authentication. Encryption methods include private key encryption technology, public key encryption technology, digital certificates, etc; user authentication mechanisms utilize static passwords, smart cards, digital signatures, dynamic passwords, pass-codes, biometric authentication, etc. A common data protection scheme generally

encrypts and stores private data by a private key. The owner holds the key to control the data usage right, and other parties need to obtain the key authorized by the data owner to access the private data, or to be verified through a centralized identity authentication system to obtain the access permission to the private data. These protection schemes have two common drawbacks. (1) They give too much protection to the privacy data, such that the efficiency and availability of the system is decreased. For example, for the purchase of railway ticket online in China, in order to avoid the scalpers snapping up quantities of tickets through misusing user information and reselling them at inflated prices, the ticketing system (12306.com) is upgraded in 2015 by applying a verification system where the users are required to match prictures with a corresponding description. Such verification system prevents the malicious scalpers and normal users from access the system at the same time, bringing inconvenience and unsatisfactory experience for target users. (2) The existing schemes protect private data by not considering the specific application scenarios. For example, the drug use information in social network is credential information that the users are willing to protect such information from being vealed, however, such information should be accurately shared in medical service application scenario.

In this paper, in order to make up the above drawbacks, a ''Proper Privacy Protection'' scheme is proposed. In our scheme, the private data is protected differently based on the data attributes and application scenarios, ensuring the private data is ''properly protected''. We firstly investigate the specific attributes of private data based on the search engine (Google) in different application scenarios, such as the data importance and dependency. Based on these attributes, the private data is divided into four security domains in a given application scenario, which is further protected by applying different protection policies. Our contributions are summarized as follows. (1) Two quantified information attributes are proposed based on Google search engine, which can capture and track the information security dynamics; (2) The relationship of private data with application scenerios is dynamically captured, which further guides the classification of private information in order to be properly protected; (3) The proposed privacy protection scheme has been demonstrated to achieve well balance between the security and usage convenience through real data based case analysis.

II. RELATED WORK

A. PRIVATE INFORMATION CLASSIFICATION

The classification study of private information has a long history. A generalized classification system was proposed in 1976 [6], targeting at privacy protection purpose. In this classification system, information sensitivity levels, dissemination categories, integrity and security provisions are the main criteria in classifying private record-keeping systems. The private data regards as the general information independent of application scenarios or areas. With the private information ubiquitously applied in vast circumstances, the researchers in different domains have done further classification study. For example, in the field of mobile health care, the privacy and security apps are studied in [7]; in Internet of Things, a classification of private information was proposed in [8], where the confidentiality and universality attributes of private data are quantified based on Google search engine. Our work is inspired by the quantification method in [8] by quantifying several information attributes from query results in Google search engine. The main difference of our method with [8] is that we consider the attributes differently in different application scenarios.

Other private information classifications are also available by considering how the data is generated and communicated. In Tinghuai Ma's research [9], a classification of personal information is used to build a hierarchy of information sharing services, with each organization communicating with each other through different levels of security pipeline. In each level, an organization has the appropriate permissions to others. They classify information as private, protected, and public. At the same time, the privacy system data is divided into two parts: system definition data (SDD) and user-defined data (UDD).

B. ENCRYPTION BASED PRIVACY PROTECTION

Encryption theory has been widely used to improve the information security. In the study of Camenisch *et al.* [10], they designed a practical multiple anonymous certificate system (allowable), and proposed a general structure, which can realize hierarchical authorization by zero-knowledge proof. They provide a new approach that proves to be safe in the common public key model, requiring only one authenticator for each signer, with the involvement of anonymous identities and signature verification time.

Chase [11] proposed a multi-authority ABE scheme using the concepts of a trusted central authority (CA) and global identifiers (GID). In this construction, the use of a consistent GID allowed the authorities to combine their information to build a full profile with all of a user's attributes, which unnecessarily compromises the privacy of the user. Moreover, the CA in that construction has the power to decrypt every cipher-text, which seemed somehow contradictory to the original goal of distributing control over many potentially untrusted authorities. So, Chase and Chow [12] studied how to improve privacy and security in the multi-authority attribute based encryption(ABE). In a multi-authority ABE scheme, multiple attribute-authorities monitor different sets of attributes and issue corresponding decryption keys to users, and encryptors can ensure that a user would obtain keys for appropriate attributes from each authority before decrypting a message. The authors proposed a solution which removed the trusted central authority, and protected the users' privacy by preventing the authorities from pooling their information on particular users, thus making ABE more usable in practice.

GH Wolfond focused on credential authorization [13], and he proposed a patent for authenticating an identity which

name as "Multi-mode credential authorization". He divided the entire credential authorization process into two communication channels and two communication credentials. First of all, a computing device received a first credential over the first communications channel, and the first credential was provisionally associated with an identity, and determined the second communications channel provisionally associated with the first credential. Then, the computing device received a second credential over the second communications channel. He divided the process of receiving the second credential by the second communication channel into another two steps: channel opening step and credential receiving step. The channel opening step comprised the computing device initiating communication to a communications address uniquely associated with the first credential. In credential receiving step, the computing device opened the second communications channel at a predetermined time and received the second credential over the opened second communication channel. The predetermined time being associated with the first credential. Finally, the computing device authenticated the identity in accordance with a verification of the second credential.

III. THE PROPOSED MODEL

In our model, the private dataset is composed of predefined types of private information. The main components of our model are shown in Figure 1. More specifically, the private data is queried in a certain search engine (e.g. Google Chrome) to obtain the query data. Two attributes about the private data type are quantified: privacy concern degree (PCD) and application dependency degree (ADD). In a specific application scenario, the attributes values are then input in a clustering model to obtain four significant different security domains. The four security domains are then protected by four different privacy protection policies.

A. ATTRIBUTES OF PRIVATE DATA TYPE BASED ON A SEARCH ENGINE

A search engine can provide the private query results about private data. We construct two attributes based on the private data query results, which are privacy concern degree (PCD) and application dependency degree (ADD).

Suppose there are n types of private data (keywords), which are denoted by $X = [x_1, x_2, ..., x_n]$. Let $S = [s_1, s_2, ..., s_m]$ denote m application scenarios. Two attributes of private data are formally defined as follows.

Definition 1 (Privacy Concern Degree (PCD)): The private concern degree of a type of private data x_i is denoted by $p(x_i)$, which is calculated as

$$p(x_i) = \frac{\frac{q(x_i,k_2)}{q(x_i,k_1)} - \min\limits_{x_k \in X} \frac{q(x_k,k_2)}{q(x_k,k_1)}}{\max\limits_{x_k \in X} \frac{q(x_k,k_2)}{q(x_k,k_1)} - \min\limits_{x_k \in X} \frac{q(x_k,k_2)}{q(x_k,k_1)}} \tag{1}$$

where $k_1 =$ "privacy"; $k_2 =$ "privacy disclosure"; $q(x_i, k_1)$ is the web page numbers by inputing the query template as x_i and "privacy"; $q(x_i, k_2)$ is the web page numbers by inputing the query template as x_i and "privacy disclosure".

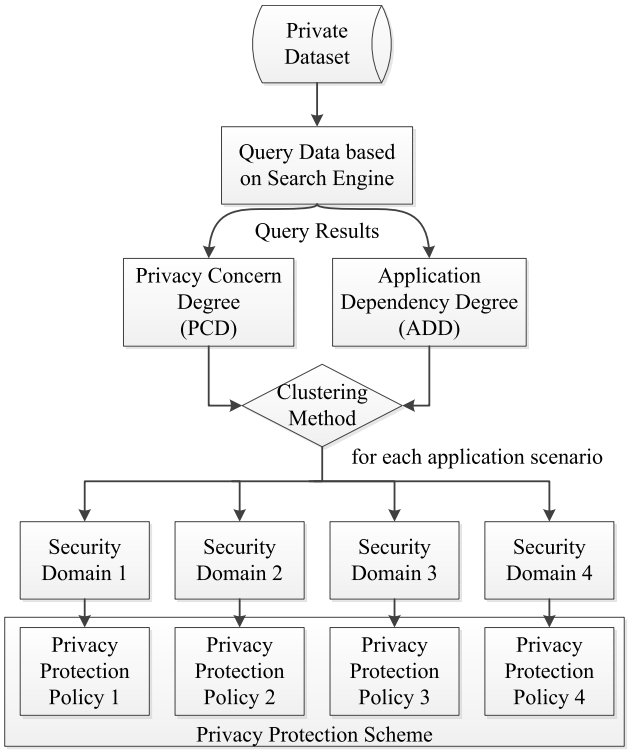

FIGURE 1. The Main Components of the Proposed Model.

It is noted that the PCD value of a private data type aims to capture the user concern extent over the leakage of the private data. When the value of $q(x_i, k_1)$ is large, users concern the privacy of the information very much by constructing large number of the related websites. Similarly, when the value $q(x_i, k_2)$ is large, users concern the privacy leakage of the information very much. The range of p_{x_i} is [0, 1]. The privacy concern degree of the private dateset is then denoted by $P(X) = [p(x_1), p(x_2), ..., p(x_n)]$.

The second attribute that we are to define is the application dependency degree.

Definition 2 (Application Dependency Degree): The application dependency degree of a type of private data x_i respecting to a specific scenario is denoted by $d(x_i, s_j)$, which is calculated as

$$d(x_i, s_j) = \frac{q(x_i, s_j) - \min\limits_{x_k \in X} q(x_k, s_j)}{\max\limits_{x_k \in X} q(x_k, s_j) - \min\limits_{x_k \in X} q(x_k, s_j)} \tag{2}$$

where $q(x_i, s_j)$ is the web page numbers by querying in a template of x_i and s_j.

The range of $d(x_i, s_j)$ is also [0, 1]. The ADD of the private dataset is then denoted by $D(X, S)$, which is a $m \times n$ matrix.

$$D(X,S) = \begin{pmatrix} d(x_1,s_1) & d(x_2,s_1) & \dots & d(x_n,s_1) \\ d(x_1,s_2) & d(x_2,s_2) & \dots & d(x_n,s_1) \\ \vdots & \vdots & \ddots & \vdots \\ d(x_1,s_m) & d(x_2,s_1) & \dots & d(x_n,s_1) \end{pmatrix} \tag{3}$$

B. SECURITY DOMAINS

With the quantified PCD and ADD attribute values of a type of private data, a private data type in a specific scenario is then represented as a point $(p(x_i), d(x_i, s_j))$. A clustering algorithm can be applied to classify the dataset into several clusters. We divide the security space into four domains, which are specified as in Table 1.

TABLE 1. The four security domains.

No	Security Domain	Description
1	Restricted	Both the PCD and ADD values are very low
2	Confidential	The PCD value is low but the ADD value is relatively high
3	Sensitive	The PCD value is high but the ADD value is relatively low
4	Highly Sensitive	Both the PCD and ADD values are very high

1) PRIVACY PROTECTION SCHEME

After the security domain is determined, different protection policies are determined according to the security requirements of the respective security domain. Then, the same private data could be protected in different ways in different scenario. The main motivation in applying different protection policies for private data in different scenario is to well balance the data security and usage convenience, so as to achieve ''proper protection''.

The privacy protection policies for the four security domains are designed as follows.

- **Restricted**: Data can be stored without encryption;
- **Confidential**: Data is stored in encrypted format, using dynamic authorization(DA), authorized for one time;
- **Sensitive**: Data is stored in encrypted format, using dynamic anonymous identity authorization(DAA);
- **Highly Sensitive**: Data is stored in encrypted format, using dynamic anonymous authorization and time limit, and user authorization is required in each single operation.

In other words, the protection policies are designed from two dimensions: data storage and user authentication. The private protection scheme is further shown in Table 2.

TABLE 2. Scheme of privacy domain.

Domain	Storage	Authorization
Restricted	Encryption(Opt)	One-time
Confidential	Encryption	DA
Sensitive	Encryption	DAA
Highly Sensitive	Encryption	DAA with time limit

IV. IMPLEMENTATION AND ANALYSIS

In order to evaluate the proposed model, it is implemented in a private dateset with 53 typical privacy data which is the same as that used in [8]. The application scenarios under our consideration include 5 types: sociality, medical care, occupation, entertainment, and finance. The search engine in our experiments is Google search because of its popularity.

We obtain the required query results through a Python crawler so as to calculate the PCD and ADD values in the proposed model. The total number of web pages crawled is 111,572 in October 2018. Next, we will present the experimental results and comparison results through analyzing the crawled web pages.

A. EXPERIMENTAL RESULTS

Table 3 is the PCD value $p(x_i)$ and the ADD value $d(x_i, s_j)$ of each private data type in each application scenario.

The clustering method in our experiment is k-means with $k = 4$. The classification results are shown from Table 4 to Table 8.

From the above results, we can observe that the same private dataset is classified into four security domains differently. We then analyze several cases to demonstrate the effectiveness of the proposed model in balancing security and convenience.

B. CASE ANALYSIS

We proceed to study four typical cases where validity and rationality are analyzed.

1) CASE 1 DRUG USE

In sociality and entertainment application scenarios, the drug use information is in the highly sensitive domain. It is reasonable, since people in a social network may refuse to construct a relationship with a person who has drug use habits, leading to the friend-making failure. However, in the medical setting, the drug use information is classified in the restricted domain. Since in medical treatments, the use of drugs is a necessary step that should be accessible to doctors. In the other two scenarios, it is unsurprising that the drug use information is in the sensitive domain.

2) CASE 2 PHONE NUMBER

In the finance application scenario, the phone number is highly sensitive information. It is understandable as the phone number is always closely binded with financial account, and it is highly risky if the phone number of users is leaked in this scenario. However, in the scenario of entertainment, the phone number is in the restricted security domain. The reason behind this observation may be that people may exchange their phone number frequently and have the willingness to contact with each other. In sociality, medical care, and occupation, the telephone number is either confidential or sensitive. It is also acceptable as people may feel disrupted if their phones numbers are abused for advertisements or other profit purpose.

3) CASE 3 MY PHOTO

The ''My photo'' information in occupation scenario is highly sensitive, which can be explained that people may seriously reluctance in sharing personal photo and such information

TABLE 3. *P*(*X*) and *D*(*X*, *S*) of the privacy data types in our dataset.

No.	Privacy Type (X)	$P(X)$	$D(X,S)$ Sociality	Medical Care	Occupation	Entertainment	Finance
1	Address book	0.4111	0.3871	0.0963	0.3750	0.3228	0.2351
2	Affiliation	0.2152	0.2028	0.3571	0.3942	0.4335	0.5268
3	Age	0.3277	0.4147	0.7298	0.3494	0.3196	0.5089
4	Bank account	0.3940	0.4470	0.1988	0.2147	0.4146	0.2232
5	Birth	0.1957	0.5622	0.6553	0.4776	0.2057	0.5744
6	Blood type	0.4131	0.3871	0.5217	0.2276	0.4367	0.3125
7	Booking hotel	0.0590	0.2719	0.0186	0.0353	0.3228	0.0774
8	Call records	0.2907	0.4931	0.0000	0.1795	0.2247	0.0744
9	Car	0.2224	0.8664	0.9565	0.4295	0.4051	0.9345
10	Chat	0.3527	0.5991	0.4596	0.2788	0.2278	0.4940
11	Children	0.2920	0.6636	0.5062	0.4808	0.7595	0.5298
12	Company address	0.3586	0.4562	0.0870	0.2051	0.4335	0.2143
13	Credit card	0.2318	0.0645	0.0497	0.0609	0.0728	0.0000
14	Credit card	0.3603	0.3180	0.6925	0.4583	0.4652	0.5714
15	Credit score	0.3730	0.6313	0.2671	0.2692	0.2405	0.5446
16	Criminal records	0.3373	0.4055	0.1863	0.1026	0.2247	0.1815
17	Disease	0.1999	0.3871	0.5963	0.3077	0.2753	0.4494
18	Driver's license	0.2056	0.2765	0.2329	0.3494	0.4842	0.2768
19	Drug use	0.700	0.5991	0.2609	0.2853	0.1551	0.2321
20	Email address	0.4232	0.2903	0.2360	0.4359	0.3608	0.2024
21	Family	0.2511	0.4194	0.8540	0.4327	1.0000	0.6577
22	Fingerprint	0.4689	0.3917	0.5342	0.5577	0.2880	0.5655
23	Height	0.3073	0.6129	0.5248	0.4744	0.4715	0.2827
24	Hobby	0.3971	0.7972	0.5248	0.4263	0.4557	0.5536
25	Home	0.2477	0.9355	1.0000	1.0000	0.5190	0.6280
26	House	0.4136	0.5945	0.6677	0.6122	0.3987	0.4464
27	Identification	0.2849	0.2074	0.4037	0.3429	0.5032	0.3810
28	Insurance	0.2112	0.6682	0.8323	0.6506	0.7468	1.0000
29	Investment	0.0000	0.5530	0.5155	0.4679	0.6076	0.7381
30	IP	0.4405	0.7097	0.2826	0.2532	0.7184	0.4643
31	Job	0.3219	0.5945	0.5901	0.4038	0.3449	0.3958
32	Location	1.0000	0.4240	0.4845	0.9968	0.5791	0.5000
33	Marriage	0.2651	0.4793	0.5093	0.3333	0.0759	0.5446
34	Mobile phone	0.4681	0.2857	0.3385	0.3237	0.4399	0.2768
35	Msn	0.3781	0.1705	0.2609	0.1346	0.0000	0.1101
36	My photo	0.1890	0.2120	0.0373	0.0000	0.2975	0.0536
37	Nation	0.2616	0.5760	0.7019	0.3429	0.2342	0.5417
38	Online records	0.2609	0.1198	0.0062	0.1026	0.1741	0.1548
39	Party	0.3617	0.5253	0.6087	0.6058	0.5253	0.5833
40	Passport	0.1105	0.4654	0.4193	0.4391	0.7215	0.4464
41	Password	0.0731	0.6267	0.8230	0.3109	0.5570	0.6429
42	Phone book	0.4267	0.3364	0.0217	0.1506	0.3829	0.1905
43	Phone number	0.5646	0.3733	0.1460	0.3237	0.3544	0.1607
44	Position	0.2400	0.6313	0.6708	0.4551	0.4715	0.8304
45	Property	0.4016	1.0000	0.4845	0.5577	0.6772	0.7827
46	Race	0.3102	0.3318	0.4099	0.3365	0.4272	0.4643
47	Religion	0.3298	0.3871	0.3913	0.4455	0.4209	0.3899
48	Salary	0.8275	0.0000	0.2578	0.3622	0.1614	0.2024
49	Shopping	0.1608	0.7143	0.3385	0.2083	0.7880	0.5327
50	Spouse	0.4036	0.0507	0.6118	0.6122	0.3671	0.4286
51	Stock	0.0293	0.4009	0.4503	0.3205	0.1424	0.3333
52	Travel	0.0345	0.7143	0.7950	0.4776	0.7911	0.7887
53	Weight	0.4494	0.4562	0.8602	0.3782	0.1994	0.4524

leakage makes people feel uncomfortable. In sociality, the photo information is always posted to update their status or increase users' attractiveness. In finance scenario, the photo information is often used in security verification, which should be also easily accessible. In entertainment scenario, personal photo is also very sensitive and people may represent them by pseudo images in such virtual circumstance.

4) CASE 4 AGE

In medical care scenario, the age information is in the highly sensitive domain, which may be explained by the fact that

TABLE 4. Private data security domains in sociality.

Restricted	Confidential	Sensitive	Highly Sensitive
Identification	Call records	Passport	**Drug use**
Affiliation	Family	Height	Location
Booking hotel	Disease	Birth	Salary
Online records	Credit card	Travel	
Stock	Email address	Insurance	
My photo	**Phone number**	Car	
Credit card	Race	Nation	
Spouse	Criminal records	Shopping	
Msn	Marriage	Position	
Driver's license	Mobile phone	Password	
	Fingerprint	Investment	
	Religion	Job	
	Weight	Chat	
	Bank account	Credit score	
	Address book	Children	
	Party	Hobby	
	Age	House	
	Company address	Home	
	Phone book	Property	
	Blood type	IP	

TABLE 5. Private data security domains in medical care.

Restricted	Confidential	Sensitive	Highly Sensitive
Drug use	Disease	Call records	Family
Location	Identification	Booking hotel	Credit card
Salary	Passport	Online records	Birth
	Affiliation	Email address	Travel
	Height	**Phone number**	Insurance
	Stock	Criminal records	Car
	Race	**My photo**	Nation
	Shopping	Credit card	Position
	Marriage	Bank account	Password
	Mobile phone	Address book	Weight
	Fingerprint	Credit score	**Age**
	Religion	Msn	House
	Spouse	Company address	Home
	Investment	Phone book	
	Job	Driver's license	
	Chat	IP	
	Party		
	Children		
	Hobby		
	Property		
	Blood type		

TABLE 6. Private data security domains in occupation.

Restricted	Confidential	Sensitive	Highly Sensitive
Location	Family	Credit card	Call records
Home	Disease	Email address	Booking hotel
	Identification	**Phone number**	Online records
	Passport	**Drug use**	Shopping
	Affiliation	Mobile phone	Criminal records
	Height	Fingerprint	**My photo**
	Birth	Weight	Credit card
	Stock	Spouse	Chat
	Travel	Address book	Bank account
	Insurance	Party	Credit score
	Race	Salary	Msn
	Car	Hobby	Company address
	Nation	House	Phone book
	Marriage	Property	Blood type
	Position		IP
	Religion		
	Password		
	Investment		
	Job		
	Age		
	Children		
	Driver's license		

TABLE 7. Private data security domains in entertainment.

Restricted Strong	Confidential	Sensitive	Highly Sensitive
Credit card	Family	Call records	**Drug use**
Identification	Passport	Disease	Location
Affiliation	Travel	Birth	Salary
Height	Insurance	Booking hotel	
Email address	Shopping	Online records	
Phone number	Password	Stock	
Race	Investment	Nation	
Car	Children	Criminal records	
Mobile phone	Property	Marriage	
Position	IP	**My photo**	
Fingerprint		Credit card	
Religion		Weight	
Spouse		Chat	
Job		Credit score	
Bank account		Msn	
Address book			
Party			
Age			
Hobby			
House			
Home			
Company address			
Phone book			
Blood type			
Driver's license			

the medical treatment may require the approximate age range of the patients, but patients biological information leakage frequently happen in the recent years. In entertainment scenario, the age information is mostly used to attract partners, which has low risk if the age information is leaked. In sociality, occupation, and finance, the age information is confidential due to the extent of the information people concern. By comparing the results of the existing privacy data classification, we can find that the two private data categories mentioned above for drug use and telephone number are presented in the PISC classification [8] as the same low security level, restricting such information to be managed in a more flexible way. In our model, different protection levels are given according to different application scenarios, and

TABLE 8. Private data security domains in finance.

Restricted	Confidential	Sensitive	Highly Sensitive
Call records	Disease	**Drug use**	Family
Height	Credit card	Location	Travel
Booking hotel	Identification	Salary	Insurance
Online records	Passport		Car
Email address	Affiliation		Position
Phone number	Birth		Password
Criminal records	Stock		Investment
Mobile phone	Race		Property
My photo	Nation		
Credit card	Shopping		
Bank account	Marriage		
Address book	Fingerprint		
Msn	Religion		
Company address	Weight		
Phone book	Spouse		
Blood type	Job		
Driver's license	Chat		
	Party		
	Credit score		
	Age		
	Children		
	Hobby		
	House		
	Home		
	IP		

for the privacy categories with less impact on the application scenario, we have little difference with the results obtained by other classification methods.

C. COMPARATIVE ANALYSIS

We also compare our private data classification results with an existing classification method in [8] denoted by PISC. There are two reasons of comparing our classification results with that in [8]. The first reason is that there are also four classes, making it comparable to our model. The second reason is that the PISC model was implemented in the same dataset. The classification results of the two models are shown in Table 9.

Regarding the number of private data in each security domain, it can be observed that our model has the similar performance with PISC in the sociality scenario, and significantly different performance in the other four scenarios. This observation can indicate (a) the proposed model is compatible with the PISC model in the sociality scenario; (b) the proposed model is adaptable with the application scenario, which indicates that the proposed model can self-adjusted to circumstance dynamics.

TABLE 9. Private data classification results in PISC and our PPP model.

Application Scenes	PISC			
	Low	Basic	Medium	High
Sociality Medical Care Occupation Entertainment Finance	12	21	17	3
Application Scenes	**Our Model (PPP)**			
	Constricted	Confidential	Sensitive	Highly Sensitive
Sociality	10	20	20	3
Medical Care	3	21	16	13
Occupation	2	22	14	15
Entertainment	25	10	15	3
Finance	17	25	3	8

V. DISCUSSION

In this section, we further evaluate the proposed model by discussing its advantages and disadvantages.

A. ADVANTAGES

The first advantage of our model is its good adaptability, which can been achieved in three aspects. In the first aspect, our model is adaptable with different application scenarios. Even for a new merging application scenario, the model can capture the proposed attribute values by querying the specific application scenario in a search engine. In the second aspect, our model is adaptable with different search engines. In our experiment, Google search engine is chosen due to its popularity, but our model can accept any other search engines. In the third aspect, our model is adaptable with unpredictable dynamics brought by security events. With the development of new technology, information security events happen frequently, which brings dynamics to people's privacy concern, and our model can well capture such dynamics by updating the attribute values.

The second advantage is its well balancing privacy protection and usage convenience in specific application scenario. It is always a challenging issue to protect privacy at a minimal cost of performance deduction. Our model insights a new attempt to protect privacy properly in the sense of fulfilling the security requirements in a personalized scenario.

B. DISADVANTAGES

The proposed model classified the private data into four security domains. However, several concerns are not studied. The first concern is whether the number of the security domains in our model (4 security domains) is the optimal choice. To achieve the answer, a large scale real database is needed to build and investigate how the private data distributes. The second concern is that whether each single private data belongs to a single domain. There should exist such private information that belongs to two or more domains. The overlapped classification methods could be investigated in our future work. Another concern is about the protection policies which should be studied in a complete and comprehensive way, which provides directions to improve the proposed model. Finally, the proposed two attributes may not be sufficient if the privacy data is more complicated. The integrity of private data should be a very relevant attribute which can be potentially investigated.

VI. CONCLUSIONS

In this paper, we have proposed a proper privacy protection scheme based on the collected query data in a search engine. Two attributes: privacy concern degree (PCD) and application dependency degree (ADD) have been defined and quantified. Based on the two attribute values, private data types are classfied into four security domains based on a clustering method. Finally, proper protection policies are designed for protecting the information privacy in each domain. The proposed model is implemented and analyzed in a real dataset, demonstrating that the proposed model can adaptively provide a proper privacy protection solution. To further evaluate the performance of the model, its privacy information classification results are compared with a comparable model in the literature, and the comparison results indicates our model outperforms the compared one in the sense of properly protecting private information in various application scenarios.

REFERENCES

[1] Statista(a). (2018). *Number of Monthly Active Facebook Users Worldwide as of 2nd Quarter 2018 (in Millions)*. Accessed: Sep. 2, 2018. [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[2] Statista(b). (2018). *Total Average Packages Delivered Daily by Fedex Express Between FY 2016 and FY 2018 (in Millions)*. Accessed: Sep. 2, 2018. [Online]. Available: https://www.statista.com/statistics/878354/fedex-express-total-average-daily-packages/

[3] Hutchinson SA. (2015). *Convenience vs Privacy: The Latest Study in the Data Tracking Debate*. Accessed: Sep. 1, 2018. [Online]. Available: https://www.socialmediatoday.com/technology-data/adhutchinson/2015-06-05/convenience-vs-privacy-latest-study-data-tracking-debate

[4] Natasha Singer. (2015). *Sharing Data, but not Happily*. Accessed: Sep. 1, 2018. [Online]. Available: https://www.nytimes.com/2015/06/05/technology/consumers-conflicted-over-data-mining-policies-report-finds.html

[5] Marvin the Robot. (2017). *Stranger Danger: The Connection Between Sharing Online and Losing the Data We Love*. Accessed: Sep. 1, 2018. [Online]. Available: https://www.kaspersky.com/blog/my-precious-data-report-three/16883/

[6] R. Turn, "Classification of personal information for privacy protection purposes," in *Proc. Nat. Comput. Conf.*, 1976, pp. 301–371.

[7] B. Martínez-Pérez, I. de la Torre-Díez, and M. López-Coronado, "Privacy and security in mobile health apps: A review and recommendations," *J. Med. Syst.*, vol. 39, p. 181, Jan. 2015.

[8] X. Lu, Z. Qu, Q. Li, and P. Hui, "Privacy information security classification for Internet of Things based on Internet data," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 8, pp. 932–941, 2015.

[9] T. Ma, S. Yan, J. Wang, and S. Lee, "Privacy preserving in ubiquitous computing: Classification & hierarchy," *Comput. Sci. Inf. Syst.*, vol. 8, no. 4, pp. 1185–1206, 2011.

[10] J. Camenisch, M. Drijvers, and M. Dubovitskaya, "Practical UC-secure delegatable credentials with attributes and their application to blockchain," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2017, pp. 683–699.

[11] M. Chase, "Multi-authority attribute based encryption," in *Proc. Conf. Theory Cryptogr.*, 2007, pp. 515–534.

[12] M. Chase and S. S. M. Chow, "Improving privacy and security in multi-authority attribute-based encryption," in *Proc. ACM Conf. Comput. Commun. Secur.*, 2009, pp. 121–130.

[13] G. H. Wolfond, J. Shapiro, and R. P. Mansz, "Multi-mode credential authorization," U.S. Patent 7 941 835, May 10, 2011.

GUIBING GUO received the Ph.D. degree in computer science from Nanyang Technological University, Singapore, in 2015. He is currently an Associate Professor with the Software College, Northeastern University, China. His research interests include recommender systems, deep learning, natural language processing, and data mining.

TIANZHI YANG received the bachelor's degree from the Information Science and Technology School, Shenyang Technology University. He is currently pursuing the master's degree with the Software College, Northeastern University, Shenyang, China. His research interests include privacy protection and blockchain-based information security.

YUAN LIU received the B.Sc. degree from the Honor School, Harbin Institute of Technology, China, in 2010, and the Ph.D. degree from the School of Computer Engineering, Nanyang Technological University (NTU), Singapore, in 2014. From 2014 to 2015, she was a Research Fellow at the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), NTU. She is currently an Associate Professor with the Software College, Northeastern University, Shenyang, China. Her research papers have been published in top international conferences in the area of artificial intelligence. Her research interests include trust-based incentive mechanism design, multi-agent systems, trust management, and blockchain technology-based reputation systems.