


Research Article

Individual Attribute and Cascade Influence Capability-Based Privacy Protection Method in Social Networks

Jing Zhang ¹, Si-Tong Shi,¹ Cai-Jie Weng,¹ and Li Xu²

¹School of Computer Science and Mathematics, Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou, 350108, China

²College of Computer and Cyber Security, Fujian Normal University, Fuzhou, 350108, China

Correspondence should be addressed to Jing Zhang; jing165455@126.com

Received 13 November 2021; Revised 17 December 2021; Accepted 9 January 2022; Published 1 February 2022

Academic Editor: Wenjuan Li

Copyright © 2022 Jing Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Users can obtain intelligent services by sharing information in social networks. Big data technologies can discover underlying benefits from this information. However, stringent security concern is raised at the same time. The public data can be utilized by adversaries, which will bring dire consequences. In this paper, the influence maximization problem is investigated in a privacy protection environment, which aims to find a subset of secure users that can make the spread of influence maximization and privacy disclosure minimization. At first, in order to estimate the risk level for each user, a Bayesian-based individual privacy risk evaluation model is proposed to rank the individual risk levels. Secondly, as the aim is to measure the influence capability for each user, a cascade influence capability evaluation model is designed to rank the friend influence capability levels. Finally, based on these two factors, a privacy protection method is designed for solving the influence maximization with attack constraint problem. In addition, the comparison experiments show that our method can achieve the goal of influence maximization and privacy disclosure minimization efficiently.

1. Introduction

Big data sharing through social media has meaningfully grown in the current era of social network [1, 2]. The social media has assumed great importance through WeChat, Facebook, social network sites, and Twitter. The issue of how information spreads through the social network has drawn more and more attention [3, 4]. The publicly available data can be utilized for market analysis, social research, and personalized service formulation [5]. Based on these analyses, more effective marketing strategies such as “viral marketing” can be found [6, 7]. The shared data may include a lot of individual information such as user’s occupation, family members, and religious affiliation. [8, 9]. These data are gathered and shared by many organizations, companies, institutions, and public websites. Unquestionably they bring valuable benefits for intelligent services [10]. However, they also pose a series of serious privacy risks [11–13]. Therefore, appropriate privacy protection should be undertaken for secure information spread [14–16].

Influence maximization problem in privacy protection environment is to find a subset of secure and reliable users that can make the spread of influence maximization and privacy disclosure minimization. It is a critical problem of finding the main factors of privacy risk and estimating the risk level of these factors [17]. There are many kinds of attributes for the shared big data, which can be classified into three categories: quasi-attributes, direct attributes, and sensitive attributes [18]. Quasi attributes are those that can be shared and do not belong to just one user, such as gender. Direct attributes have character of uniqueness such as e-mail and WeChat ID. Sensitive attributes contain individual private information such as personal health status. Sometimes other sensitive attributes are externally visible [19]. It is necessary to design an estimation to measure the risk level for each user. The user with a low individual privacy risk level is assumed to be safe. Because the attackers will pay no attention to the lower ones, the user with the lower risk level will be safe. While the user with high privacy risk level will be of great interest to attackers, all attributes of this user are leaked. Furthermore, with the

rapid development of communication technology, the world is getting smaller [4]. The influence of friends can pose privacy risks. It has been found that a great deal of privacy leakage comes from friends' indirect disclosure [7]. In a word, information attributes and friends' influence are two important factors for privacy risks evaluation.

This paper focuses on the design of the influence maximization method in privacy protection environment. Individual privacy risk and friends' influence are two important factors for privacy risks evaluation. It is necessary to design an estimation to measure the risk level of the user and friends' influence capability. Based on this, a privacy protection method can be designed for the social network. For these purposes, at first, an attribute risk level grading method is designed based on Bayesian Network. Secondly, the cascade influence model is employed for designing the friend influence capability model. Thirdly, a privacy protection method is designed for solving the influence maximization with attack constraint problem. Specifically, the contributions of this paper are summarized as follows:

- (1) Bayesian-based individual privacy risk evaluation model (IPREM) is proposed to evaluate the individual privacy risk levels. Since the actual multidimensional attribute data may not be completed, it is difficult to deal with the complex nonlinear relationship between the individual privacy risk and the multidimensional attribute evaluation index by using the regression analysis method. However, Bayesian Network has the function of reverse reasoning. Under the premise of some serious privacy risk, the trained Bayesian Network can be used to carry out reverse operation and analyze the objective factors causing risk.
- (2) Cascade influence capability evaluation model (CICEM) is designed to evaluate the friend influence capability levels based on the cascade influence model. According to the users' cascade influence capability, the benefits and threats for the friends' influence capability can be measured.
- (3) Based on the two evaluation models for two important factors, an IPREM and CICEM based Privacy Protection Method (ICPM) is designed for solving the influence maximization with attack constraint problem. It is the first attempt, to our knowledge, to consider the individual privacy risk and influence maximization on the privacy protection design.

The rest of the article is organized as follows: the related work is given in Section 2; the preliminaries are given in Section 3; the privacy protection method is designed in Section 4; the simulation analysis is discussed in Section 5; and finally, the conclusion is given in Section 6.

2. Related Work

There are many researches that focus on influence maximization issue, such as degree base heuristic algorithm [20] and greedy algorithm [4]. However, most of them do not

consider the privacy threat problem. Privacy threats toward social network have been extensively documented. To deal with these concerns, many privacy preserving techniques have been proposed in literature. The aspect of data protection, the behaviors of data collecting or publishing, and the privacy characterization and measurement method are three main methods.

As an aspect of data protection method, Li et al. considered data security by putting all data into a cloud [21]. A mobile-cloud framework is presented by eradicating the data over-collection. However, this kind of approach mainly involves restricting data sharing, which is not suitable for the social network. Some approaches are designed based on cryptography, such as a match-then-decrypt technique proposed by Zhang et al. [22]. The data can be decrypted only when the attribute private key can match the hidden access policy. Some approaches are designed by setting the access control permission, such as Li et al. proposed a lightweight approach to protecting privacy, which applies the information flow control in routers [23]. However, these approaches are controlled by the servers without considering the user's personalization, which are not suitable for the social network.

Some studies consider data security from data collection or publish behavior [19, 24]. For example, based on the theory of planned behavior and the privacy calculus model for social network, Li et al. proposed an integrated model to explain privacy disclosure behaviors [25]. In order to reduce disclosure risk and enhance data utility, Marmar et al. proposed an improved suppression method by targeting the highest risk records and keeping other records intact [18]. Three new theories extended parallel process model, self-control theory, and routine activity theory which are employed by Chen et al. to explore online privacy concerns [26]. Based on the anonymity technology, Javier et al. present a generalization of aggregation method, where the individual data are replaced by cluster mean for data publishing [19]. However, without considering the difference of the individual attributes, they use the same strategy for different data sets. In fact, different people can use different strategies. Therefore, it is very desirable to have a lightweight and scalable mechanism to protect privacy.

The above researches focus on data protect, which lack a proper privacy characterization and measurement. A quantification model with multi-variable privacy characterization is presented by Ref. [14], which can analyze the sensitivity of individual privacy characterization. Investigate how to optimize the tradeoff between latent-data privacy and customized data utility. He et al. proposed a data-sanitization strategy that does not greatly reduce the benefits brought about by social network data, while sensitive latent information can still be protected [8]. Based on defining user vulnerability, Gundecha et al. present a privacy setting model by keeping users away from the high threatening users [7]. In order to quantify the location privacy leak, Li et al. designed a model by matching the users shared locations with their real mobility traces [6]. Several link-prediction and attribute prediction algorithms are proposed in social attribute networks [27]. In order to predict sensitive

information, a data-sanitization strategy is proposed by harnessing link and attribute information simultaneously [28]. In order to resist inference attack, Cai et al. proposed a collective inference model with a mixture of nonsensitive attributes and social relationships [9]. However, these researches only consider the individual vulnerability, without considering the friendship influence. In fact, both individual vulnerability and friendship influence will affect the privacy risk. Therefore, it is very desirable to design some models to estimate the risk level of each attribute and the friend's influence capability. Based on these analyses, we focus on the design of the individual privacy risk evaluation model and friend influence capability evaluation model. Furthermore, based on these two models, we need to design a privacy protection method for solving the influence maximization with attack constraint problem.

3. Preliminaries

This section describes some necessary background of the privacy protection, such as the social network, the attribute set, the cascade model, and the Bayesian model.

Definition 1 (social network) (see [8]). Social network can be described as a graph $G(V, E, \mathcal{A})$, with node set V , edge set E , and attribute sets \mathcal{A} . An edge $(\mathbf{v}_i, \mathbf{v}_j) \in E$ exists if and only if nodes \mathbf{v}_i and \mathbf{v}_j can communicate with each other. $|V| = n$ and $|E| = m$ represent the total number of nodes and edges, respectively. For the uniformity, all users are referred to as nodes in this article.

Definition 2 (attribute set) (see [8]). The attribute set of node \mathbf{v}_i can be represented by an attribute vector $A \in \mathcal{A}$. $|A|$ represents the total number of attributes. Each attribute $x_i \in A (1 \leq i \leq |A|)$ takes value from the i -th dimension attribute.

Definition 3 (risk of individual index) (see [7]). Individual index is defined to estimate the risk of privacy, the risk may be incurred by allowing individual attributes to be visible. The risk of individual index \mathbf{R}_u for node \mathbf{u} can be defined as a function of individual attribute, which is shown as follows:

$$\mathbf{R}_u = \frac{\sum_{i=1}^{|A|} w_i \times a_i}{\sum_{i=1}^{|A|} w_i}, \quad (1)$$

where w_i is the sensitivity weight of an i -th attribute x_i , which will be defined in the IPREM model. $a_i = 1$ if i -th attribute is visible, otherwise the attribute is not visible. $\mathbf{R}_u \in [0, 1]$, where $\mathbf{R}_u = 1$ indicates that all attributes of node \mathbf{u} can be visible. On the other hand, $\mathbf{R}_u = 0$ indicates that the attribute of node \mathbf{u} is nonvisible.

Definition 4 (cascade model) (see [4]). The cascade model is an influence spreading model with probability. The activated node \mathbf{v}_i will attempt to activate its inactive neighbor \mathbf{v}_j under the probability p_{ij} . Furthermore, the active node has only one chance to activate each of its inactive neighbors. Such

attempts are mutually independent for different neighbors, namely, the activation of \mathbf{v}_i to \mathbf{v}_j will not be affected by the influences from other neighbors of \mathbf{v}_j .

Two kinds of cascade models, the Independent Cascade Model (ICM) with random p_{ij} and the Weight Cascade Model (WCM) with the weighted probability, will be utilized in this article.

Definition 5 (independent cascade model (ICM)) (see [4]). The ICM is an influence spreading model with probability. The activated node \mathbf{v}_i will attempt to activate its inactive neighbor \mathbf{v}_j under the random probability p_{ij} .

Definition 6 (weight cascade model (WCM)) (see [4]). The WCM is an influence spreading model with probability. The activated node \mathbf{v}_i will attempt to activate its inactive neighbor \mathbf{v}_j under the weighted probability p_{ij} .

Definition 7 (cascade index). Cascade index is defined to estimate the status of the cascade influence capability. It is defined as a function of transmission capability for cascade influences as follows:

$$\mathbf{C}_v = \sum_{G'} p(G') \times a_i, \quad (2)$$

where $p(G')$ is the probability of transmission subgraph and $a_i = 1$ if i -th level needs to be calculated, otherwise the level is not considered. $\mathbf{C}_v \in [0, 1]$, where $\mathbf{C}_v = 1$ indicates the highest level, where all nodes in the network can be influenced by node \mathbf{v} . $\mathbf{C}_v = 0$ indicates the lowest level, where none of the nodes can be influenced in the network.

Definition 8 (Bayesian model) (see [29]). If M_1, \dots, M_K are the models considered, and Δ is the quantity of interest, then its posterior distribution under given data D is shown as

$$pr(\Delta|D) = \sum_{k=1}^K pr(\Delta|M_k, D)pr(M_k|D), \quad (3)$$

The posterior probability for model M_K can be given by

$$pr(M_k|D) = \frac{pr(D|M_k)pr(M_k)}{\sum_{i=1}^K pr(D|M_i)pr(M_i)}. \quad (4)$$

Some important symbols and their definitions are presented in Table 1.

4. IPREM and CICEM Based Privacy Protection Method (ICPM)

In order to maximize the influence and minimize the privacy risk, the seed set with k -size needs to be selected. The node with high cascade influence capability but low individual privacy risk evaluation can satisfy this necessary criterion. The nodes with high cascade influence capability can improve the network influence. However, if it also has high private risk level, it will be of great interest to attackers. Then, it can pose a threat to his friends, and the threats are increased with the number of vulnerable nodes that are

TABLE 1: Symbols and definitions commonly encountered.

Symbols	Definitions
$G(V, E, \mathcal{A})$	Graph G for describing the network
\mathbf{v}_i	User or node in the network
$(\mathbf{v}_i, \mathbf{v}_j) \in E$	Edge in the network
$x_i \in A (1 \leq i \leq A)$ and $A \in \mathcal{A}$	Attribute vector
\mathbf{R}_u for Node \mathbf{u}	The risk of individual index
w_i	The sensitivity weight of an i -th attribute x_i
a_i	The visible value of the i -th attribute
I_u	Individual privacy risk
C_v	Cascade index
$pr(\cdot)$	The posterior distribution
σ	The vulnerable weight selected into the seed set
α, β	The weight factors
k	The threshold value
$p(G')$	The probability of the subgraph G
$f_{G'}(X, Y)$	The nodes set influenced by seed set X after i steps through cascade influence model, when the attacked set Y exists
T	The total number of the steps in the cascade influence model

influenced. So, the user with a low individual privacy risk level will paid less attention by the attackers, at the same time, the user with high cascade influence capability can improve the network influence. In other words, it is necessary to find the nodes with high cascade influence capacity and stay away from the vulnerable ones. This kind of problem can be defined as the Influence Maximization with Attack Constraint problem (IMAC).

IMAC (X, Y, k) : X is a selected seed set with high vulnerable weight σ , where $\sigma_i = \alpha C_i - \beta I_i$ and α, β represent the weight factors. The seed set size needs to satisfy $|X| \leq k$. Y is the set whose member is vulnerable to be attacked. The aim of the IMAC is to maximize the influence under the constraint environment of minimizing the privacy risk. So σ is positively correlated with the cascade influence capability C_i while negatively correlated with the individual privacy risk I_i . I_i and C_i can be calculated by (9) and (10), respectively.

For a graph G with n nodes and m edges: The probability of the subgraph G generated can be calculated as (5).

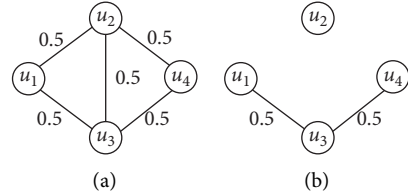
$$p(G') = \prod_{e \in G'} p_e \prod_{e' \in G/G'} (1 - p_{e'}), \quad (5)$$

where p_e can be defined by the concrete cascade influence model.

According to (5), the probability of the example subgraph as shown in Figure 1 is $p(G') = 0.015625$. Assume graph has l probability graphs $(G'_1, G'_2, \dots, G'_l)$, then the number of nodes that can be influenced by the nodes set S can be calculated by the arbitrary G' . The node conditional expected influence privacy is shown as (6).

$$f_G(X, Y) = \sum p(G') f_{G'}(X, Y) \quad (6)$$

$f_{G'}(X, Y) = \sum_{i=1}^T \sum_{v \in X} f_{G'}((v, Y), i)$ represents the nodes set influenced by seed set X after i steps through cascade influence model, when the attacked set Y exists. T is the total number of the steps in the cascade influence model.

FIGURE 1: Possible graph instance. (a) $G = (V, E)$. (b) $G' = (V', E')$.

Users in the social network can decide whether or not to reveal their individual attributes based on the risk levels. So, estimating the privacy risk is the basic precondition for the privacy protection. In this section, the evaluation models for quantifying privacy disclosure risks are discussed. Two factors are estimated, individual privacy risk and cascade influence capability. The Bayesian-based Individual Privacy Risk Evaluation model (IPREM) is designed for the hierarchy of individual attribute at first. Individual attributes include personal information such as name, age, gender, family members, e-mail, QQ ID, WeChat ID, occupation, and even religious affiliations. Furthermore, one node's vulnerability depends not only on the visibility of individual attributes but also on the exposure of the profile through his friends. Then, the Cascade Influence Capability Evaluation model (CICEM) is designed, which aims to rank the friend influence risk levels based on the cascade influence model. At last, an IPREM and CICEM based Privacy Protection Method (ICPM) is designed for solving the IMAC problem.

4.1. IPREM: Hierarchy of Individual Privacy Risk. The individual privacy risk is one of the most important factors for privacy protection. For example, if one of your friends who has most of your information has a high individual privacy risk, there is high probability that your information will be leaked indirectly. So how to evaluate each user's individual privacy risk level is the first important issue. The probability of individual privacy risk can be predicted based on the

Bayesian Network, under the condition that some risk is known. Since the actual multidimensional attribute data may not be completed, it is difficult to deal with the complex nonlinear relationship between the individual privacy risk and the multidimensional attribute evaluation index by using the regression analysis method. However, Bayesian Network has the function of reverse reasoning. Under the premise of some serious privacy risk, the trained Bayesian Network can be used to carry out reverse operation and analyze the objective factors causing risk. Bayesian Networks can be obtained by means of data analysis and expert experience.

The individual attribute exposure statuses are denoted as F_i^1 ($i = 1, 2, \dots, n$). Assume the prior probability is $pr(F_i^1)$. The risk level is denoted as \mathbf{R} and new additional information obtained from the investigation is $pr(\mathbf{R}|F_i^1)$ ($i = 1, 2, \dots, n$). According to Bayesian Network, the posterior probability $pr(F_i^1|\mathbf{R})$ can be calculated:

$$pr(F_i^1|\mathbf{R}) = \frac{pr(F_i^1)pr(\mathbf{R}|F_i^1)}{\sum_{i=1}^n pr(F_i^1)pr(\mathbf{R}|F_i^1)}. \quad (7)$$

The probability of each factor leading to its occurrence can be calculated by (7), when risk \mathbf{R} happens. In this paper, the levels of privacy risk is defined as the number of the attributions $|A|$, from level 1 to level $|A|$, with level $|A|$'s risk being the highest. The risk levels can be denoted as $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_{|A|}$.

The privacy risk level may be affected by the status of the factors. The level of exposed information determines the status of the node and thus the risk level of the node can be calculated. So, it can be considered that there is a causal relationship between various statuses and the levels of privacy risk. According to (7), there is

$$\begin{aligned} pr(F_j^1|\mathbf{R}_i) &= \frac{pr(\mathbf{R}_i|F_j^1)pr(F_j^1)}{pr(\mathbf{R}_i)} \\ &= \frac{pr(\mathbf{R}_i|F_j^1)pr(F_j^1)}{\sum_{j=1}^4 pr(\mathbf{R}_i|F_j^1)pr(F_j^1)}. \end{aligned} \quad (8)$$

Based on these theoretical foundations, the IPREM can be designed by 4 steps as follows:

Bayesian-Based Individual Privacy Risk Evaluation Model (IPREM)

Input: the network and the probability of each attribute being exposed.

Output: individual privacy risk I_i for each individual privacy.

Step 1. The first step is to design a disclosure risk measure. According to the analysis of network, the probability of exposure (e_i) for each attribute can be obtained, and sensitivity weight (w_i) for each value can be calculated by $w_i = 1 - e_i$.

An example of 4 sample attributes is shown in Table 2. According to the analysis of Facebook network, about

TABLE 2: Probabilities of exposure and sensitivity weights.

Attribute	Probability (e_i)	Weight (w_i)	Status
Gender	0.8177	0.1823	F_1^1
Education and work	0.2513	0.7487	F_2^1
Mobile number	0.0036	0.9964	F_3^1
Website	0.0626	0.9374	F_4^1

81.77% nodes reveal their gender, about 6.26% nodes reveal their individual websites. Then the sensitivity weight of the gender is 0.1823. Assume the four statuses of the individual privacy risk is depended by the attribute public situation:

- (1) The probability of exposure between 0.3 and 1 is defined as status F_1^1 , such as the gender, whose probability $p = 0.8177$. These attributes only trigger the lowest level of individual privacy risk
- (2) The probability of exposure between 0.15 and 0.30 is defined as status F_2^1
- (3) The probability of exposure between 0.04 and 0.15 is defined as status F_3^1
- (4) The probability of exposure between 0 and 0.04 is defined as status F_4^1

For example, the attribution phone number is set to be visible by only 00.36% of users, and then it has a sensitivity weight of 0.9964, which will trigger the highest level of leakage risk.

Step 2. The second step is to calculate the prior probability of each attribute based on (1).

The prior probability can be calculated as Table 3 based on (1). Since $|A| = 4$, the risk levels can be denoted as $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_4$. Take status F_2^1 as an example; the probability for the privacy risk at level 1, $\mathbf{R}_1 = 0.1958$ can be calculated. Then, the coefficient of individual privacy risk evaluation r_i can be calculated as the last column of Table 3.

Step 3. The modeling of Bayesian Network can be completed based on (8).

Assume $pr(F_j^1) = 0.25$, and based on (8), Table 4 shows the $pr(F_j^1|\mathbf{R}_i)$ ($i = 1, 2, 3, 4; j = 1, 2, 3, 4$). When the privacy risk is at \mathbf{R}_1 , the posterior probability $pr(F_1^1|\mathbf{R}_1) = 0.7369$, $pr(F_2^1|\mathbf{R}_1) = 0.1443$, $pr(F_3^1|\mathbf{R}_1) = 0.0719$ and $pr(F_4^1|\mathbf{R}_1) = 0.0469$ can be calculated. It is easy to find that the risk is higher when the individual privacy risk is in the higher status. For example, when risk \mathbf{R}_4 happens, the probabilities in status F_4^1 is 100%. However, if risk \mathbf{R}_1 happens, the probabilities in statuses F_1^1, F_2^1, F_3^1 , and F_4^1 are 73.69%, 14.43%, 7.19%, and 4.69%, respectively.

Step 4. The individual privacy risk can be calculated by

$$I_{\mathbf{u}} = \frac{\sum_{i=1}^{|A|} x_i \times r_i}{\sum_{i=1}^{|A|} x_i}, \quad (9)$$

where x_i represents the value of the i 's attribute value.

TABLE 3: Individual privacy risk evaluation.

	$pr(\mathbf{R}_i F_1^1)$	$pr(\mathbf{R}_i F_2^1)$	$pr(\mathbf{R}_i F_3^1)$	$pr(\mathbf{R}_i F_4^1)(r_i)$
\mathbf{R}_1	1.0000	0.1958	0.0976	0.0636
\mathbf{R}_2	0	0.8042	0.4007	0.2613
\mathbf{R}_3	0	0	0.5017	0.3272
\mathbf{R}_4	0	0	0	0.3478
Total	1	1	1	1

TABLE 4: Posterior probability.

	$pr(F_i^1 \mathbf{R}_1)$	$pr(F_i^1 \mathbf{R}_2)$	$pr(F_i^1 \mathbf{R}_3)$	$pr(F_i^1 \mathbf{R}_4)$
$i = 1$	0.7369	0.0000	0.0000	0.0000
$i = 2$	0.1443	0.5485	0.0000	0.0000
$i = 3$	0.0719	0.2733	0.6053	0.0000
$i = 4$	0.0469	0.1782	0.3947	1.0000
Total	1	1	1	1

For example, based on the data in Tables 3 and 5, since $|A| = 4$, for node \mathbf{u}_1 , $x_1 = x_2 = 1$, and $x_3 = x_4 = 0$, the individual privacy risk $I_1 = 0.3250$ can be calculated. In the same way, the individual privacy risk for \mathbf{u}_2 , \mathbf{u}_3 , and \mathbf{u}_4 can be calculated as $I_2 = 1$, $I_3 = 0.3908$, and $I_4 = 0.9364$, respectively. Node \mathbf{u}_2 has the highest individual privacy risk.

4.2. CICEM: Hierarchy of Influence Capability Based on Cascade Influence Model. Cascade influence capability is another important factor in the social network. On the one hand, users' cascade influence capability is a key factor for influence maximization. Users with high cascade influence capability are selected into the seed set and can make the spread of influence maximization. On the other hand, the friend influence risk is another factor for privacy protection. A friend with high cascade influence capability may have higher influence risk. For example, if one of your friends who has high cascade influence capability know most of your information, there is a high probability that your information may be leaked indirectly. So, how to evaluate each user's cascade influence capability is the second important issue.

The idea is similar to the IPREM, and based on the cascade influence model, the CICEM is designed as follows:

4.2.1. Cascade Influence Capability Evaluation Model (CICEM)

Input: one social network G

Output: cascade influence capability C_i

Step 1. Similar to IPREM, the statuses can be defined according to the cascade index C_v , which also can be dynamic regulated by the user or environment requirement. Here, assume the cascade influence has four kinds of statuses F_j^2 ($j = 1, 2, 3, 4$), the prior probability is $pr(F_j^2)$. According to (2), the probability of cascade influence $pr(\mathbf{L}_i|F_j^2)$ can be calculated. \mathbf{L}_i represents the level of the cascade influence capability, which can be set based on some established principles.

TABLE 5: One example.

	\mathbf{u}_1	\mathbf{u}_2	\mathbf{u}_3	\mathbf{u}_4
Gender (x_1)	1	1	1	0
Education and work (x_2)	1	1	0	1
Website (x_3)	0	1	1	1
Mobile number (x_4)	0	1	0	1

Take Figure 1 as an example. Four levels can be set based on the node degrees. The nodes are arranged in descending order of the degree. The top 25% nodes with the lowest degree are set as \mathbf{L}_1 , those arranged between 25% and 50% are set as \mathbf{L}_2 , those arranged between 50% and 75% are set as \mathbf{L}_3 , and the top 25% nodes with the highest degree are set as \mathbf{L}_4 . Figure 1(a) is the original graph. Figure 1(b) is the subgraph; nodes \mathbf{u}_3 and \mathbf{u}_4 are affected by node \mathbf{u}_1 through cascade influence model, that is to say that, in this subgraph, two nodes have been influenced. In this example, set node \mathbf{u}_j 's influence status to F_j^2 . According to (2), the total probability of influence is shown in Table 6. Then, the coefficients of cascade influence evaluation c_i can be calculated as shown in the last column of Table 6.

Step 2. The posterior probability $pr(F_j^2|\mathbf{L}_i)$ ($i = 1, 2, 3, 4$; $j = 1, 2, 3, 4$) can be calculated based on (9).

Take the same example, $pr(F_j^2|\mathbf{L}_i)$ ($i = 1, 2, 3, 4$; $j = 1, 2, 3, 4$) are shown in Table 7. It is easy to find that the cascade influence capability is higher when more nodes are influenced. For example, when cascade influence capability \mathbf{L}_4 happens, the influence set size more than 4 has a probability of 100%. If cascade influence capability \mathbf{L}_1 happens, only one node is influenced with a probability of 61.15%, and nodes 2, 3, and 4 have been influenced with probabilities of 22.93%, 10.19%, and 5.73%, respectively.

Step 3. Based on the cascade influence evaluation c_i , the cascade influence capability can be calculated as

$$C_i = \sum_{j=1}^n \left(\prod_{e \in \{\mathbf{u}_i \rightarrow \mathbf{u}_j\}} p_e \times c_{j-1} \right). \quad (10)$$

c_i is the cascade influence evaluation, which can be calculated, as shown in Table 6. The selection of coefficient is determined by the node. For example, when node \mathbf{u}_j is in level \mathbf{L}_3 , then $c_j = c_3 = 0.3125$ will be selected as the coefficient. p_e is the probability of the edges in the route between the node \mathbf{u}_i and \mathbf{u}_j . n is the total number of the nodes in the network. The cascade influence capability needs to be normalized as $C_i = C_i / \text{Max}\{C_i\}$.

4.3. ICPM: IPREM and CICEM Based Privacy Protection Method. Influence maximization problem in privacy protection environment is to find a subset of secure and reliable users that can make the spread of influence maximization and privacy disclosure minimization. For this purpose, at

TABLE 6: Cascade influence evaluation.

	$pr(\mathbf{L}_i F_1^2)$	$pr(\mathbf{L}_i F_2^2)$	$pr(\mathbf{L}_i F_3^2)$	$pr(\mathbf{L}_i F_4^2)(c_i)$
\mathbf{L}_1	1.0000	0.3750	0.1667	0.0938
\mathbf{L}_2	0.0000	0.6250	0.2778	0.1563
\mathbf{L}_3	0.0000	0.0000	0.5556	0.3125
\mathbf{L}_4	0.0000	0.0000	0.0000	0.4375
Total	1	1	1	1

TABLE 7: Posterior probability.

	$pr(F_i^2 \mathbf{L}_1)$	$pr(F_i^2 \mathbf{L}_2)$	$pr(F_i^2 \mathbf{L}_3)$	$pr(F_i^2 \mathbf{L}_4)$
$i = 1$	0.6115	0.0000	0.0000	0.0000
$i = 2$	0.2293	0.5902	0.0000	0.0000
$i = 3$	0.1019	0.2623	0.6400	0.0000
$i = 4$	0.0573	0.1475	0.3600	1.0000
Total	1	1	1	1

last, an IPREM and CICEM based privacy protection method (ICPM) is designed. The process of the ICPM is as follows:

- Step 1. Calculate the individual privacy risk I_i based on IPREM.
- Step 2. Calculate the cascade influence capability C_i based on CICEM.
- Step 3. Calculate the $\sigma_i = \alpha C_i - \beta I_i$ for each node \mathbf{u}_i .
- Step 4. Select some nodes into seed set X .
- Step 5. Calculate $f_G(X, Y)$.
- Step 6. For node \mathbf{u}_j , calculate $f_G(X \cup \{\mathbf{u}_j\}, Y)$, if $f_G(X \cup \{\mathbf{u}_j\}, Y) \geq f_G(X, Y)$, add \mathbf{u}_j into X .
- Step 7. Repeat Step 6 until $|X| = k$.

5. Performance Evaluations

In order to analyze the performance of the ICPM method, the Facebook network is selected for experimental analysis. Imitating the data source of literature [7], we captured some Facebook data containing user information. This network contains about 130,000 users and 1,000,000 edges. The profile information includes 26 attributes for users such as age, gender, mobile phone number, and address. Without invasion of privacy, each of the attribute information is defined as true or false. True means this attribute is visible, while false means nonvisible. Since there are 26 attributes in the simulation, the risk levels can be denoted as $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_{26}$. Figure 2 shows the percentages of people who enable the particular attribute to be visible. For example, it can be found that 0.36% users enable their mobile phone numbers to be visible. 81.77% users enable their gender to be visible.

In this section, three influence maximization methods, two cascade influence models, and two attack models are discussed for comparison. Three influence maximization methods are degree-based [20], random-based [30], and our ICPM method. For the degree-based method, k nodes with higher degree will be selected. For the random-based seed set

selected method, k nodes will be selected randomly. For our ICPM method, k nodes will be selected according to the method proposed in Section 4. The weight coefficients α, β can be set by the environments and requirements. In this simulation experiment, they are set as $\alpha = 0.6$ and $\beta = 0.4$. The nodes selected by these methods are taken as the initial active nodes. Furthermore, ICM and WCM are two cascade influence models we will use. For each influence maximization method, with or without edge weight modified models will be discussed. In addition, in order to test the security, two attack models will be modeled: (1) the attack model based on high individual privacy risk and (2) the attack model based on high degree. Two measurements, influence size and protection degree, are discussed. The simulation experiments are carried out in the MATLAB environment. The final influence effects and protection degrees are the average of 50 times simulation experiment.

5.1. Comparison Experiment Based on ICM. At first, the comparison experiment based on ICM will be discussed. Figure 3 shows the influence for three methods in Facebook network. For simplicity and lack of information, the activation probability between nodes is set as the same value of 0.05, which is also the probability value commonly used under this model [4]. Figure 3(a) is a comparison of the number of influenced nodes of different seed sizes by three methods in the Facebook network, where the x -coordinate is the seed set size, and the y -coordinate is the size of the set to be influenced to eventually. It can be found that the seed set selected by our ICPM method can spread much wider than other methods when no attack happens.

An important conclusion that can be drawn from Figures 3(b)–3(d) is that, when the high degree attack happens, the ICPM method is affected slightly, and its property of antiattack is the best. The degree-based method has the lowest influence set size. Assume that 200 nodes are selected as the seed set; for the degree method, the influence set sizes are decreased from 933 to 261, 163, and 51 when the attacked set sizes are 50%, 70%, and 90% of the seed set size, respectively. However, for the ICPM method, the influence set sizes are decreased from 1249 to 372, 355, and 331 when the attacked set sizes are 50%, 70%, and 90% of the seed set size, respectively. Take the attacked set size as 90% as an example, the influence set sizes fall to 94% and 73% by degree method and ICPM method, respectively.

However, It is not clear at a glance for the privacy protection level. Then, the protection degree is defined as the ratio of the influences set sizes under attack to that without attack. Figure 4 shows the protection degree under high individual privacy risk attack. Figures 4(a) and 4(b) show 10% and 20% nodes of the whole network are attacked, respectively. For example, when 10% nodes are attacked, the protection degree is about 0.4 by our ICPM method, while that is nearly 0 by the degree method. It also can be found that ICPM method is affected slightly under the individual privacy risk attack, which can protect the privacy more. The reason for this behavior is that the degree is not the only factor considered in our ICPM method. The nodes with high

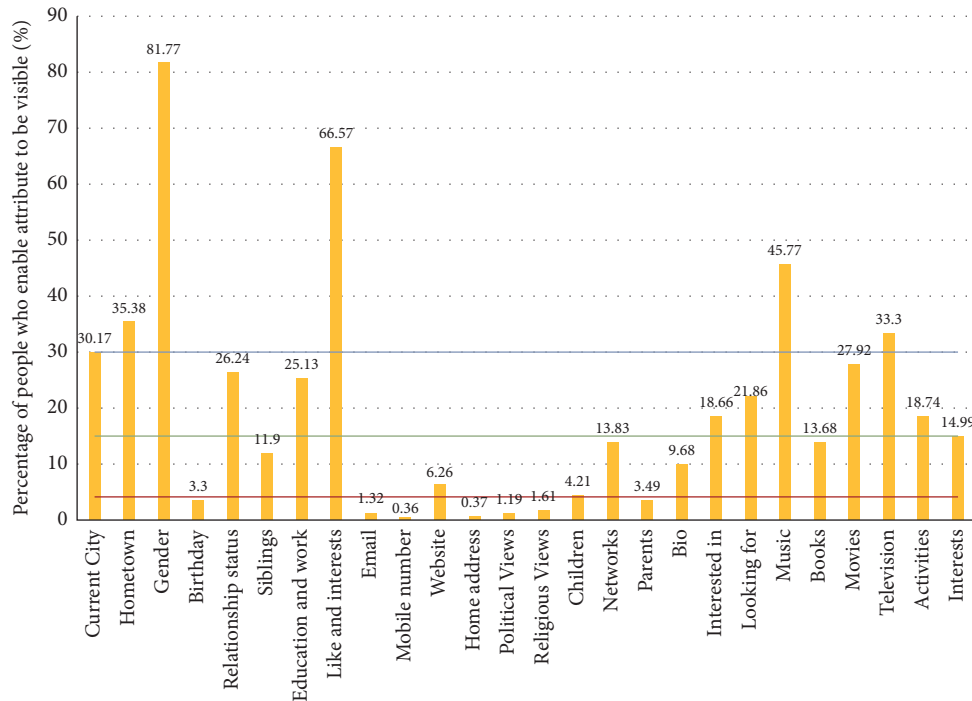


FIGURE 2: Attributes visibility distribution.

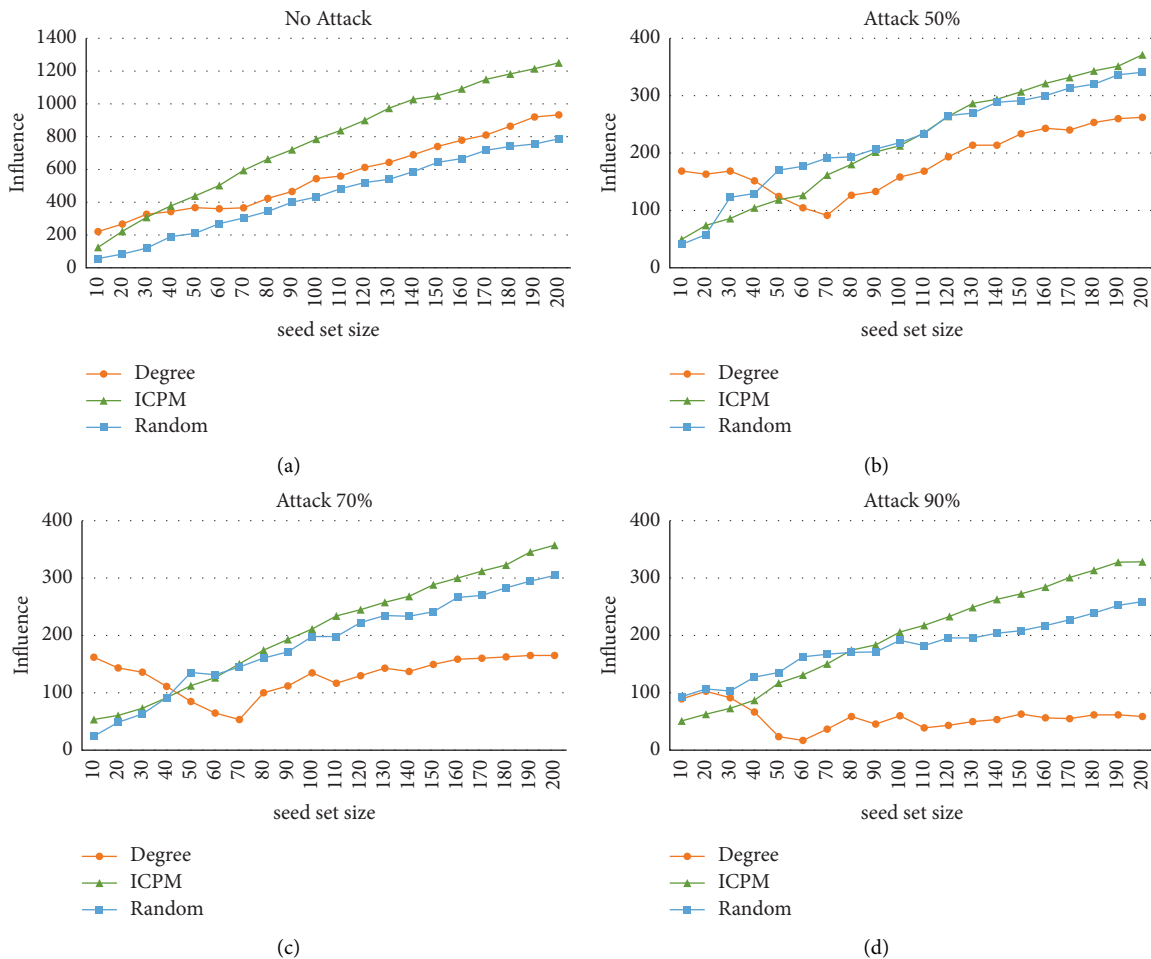


FIGURE 3: The influence based on ICM under high degree attack model in the Facebook network. (a) Not attacked, (b) the attacked set size is 50% of the seed set size, (c) the attacked set size is 70% of the seed set size, and (d) the attacked set size is 90% of the seed set size.

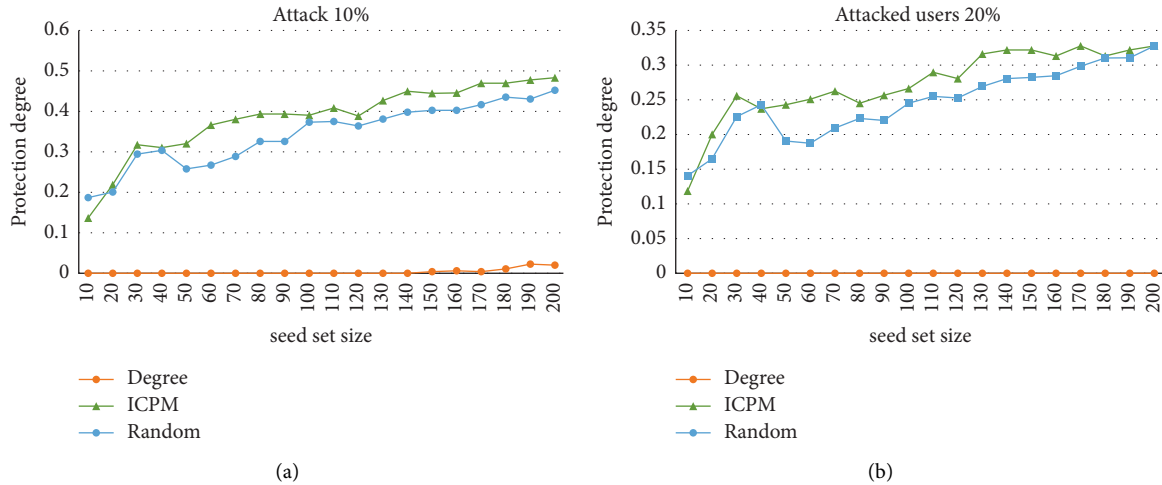


FIGURE 4: The protection degree based on ICM under high individual privacy risk attack in the Facebook network. (a) The attacked set size is 10% of the whole network and (b) the attacked set size is 20% of the whole network.

individual privacy risk have a lower probability to be selected as the seed node. That is to say, the ICPM method has the ability to find more security influential nodes than the other methods.

5.2. Comparison Experiment with Edge Weight Modification.

Second, based on the ICM, the influence with or without edge weight modification will be discussed. The edge weight modification means that the weight can be modified according to the actual environment and requirement. For example, three kinds of activation probabilities between nodes are set. For the nodes with top 33% highest individual privacy risk, the probabilities of the incidence edge are modified to $p_{ij}(1 - \lambda_1 I_i)$, and for the nodes with top 33% lowest individual privacy risk, the probability of the incidence edge are modified to $p_{ij}(1 + \lambda_2 I_i)$, where $\lambda_1 = 1$ and $\lambda_2 = 0.5$.

Figure 5 shows the influence under high degree attack model. Six kinds of influence maximization methods, random-based, degree-based, and ICPM method with or without edge weight modification, are discussed based on ICM. For example, assume that 200 nodes are selected as the seed set. Under high degree attack and without edge weight modification, the attacked set sizes are 70% and 90%, respectively, of the seed set size. For the degree method, the influence set sizes are 165 and 51, respectively. For the random method, the influence set sizes are 302 and 259, respectively. However, for the ICPM method, the influenced set sizes are 355 and 331, respectively. It can be found that the ICPM method is affected slightly under the degree attack.

Furthermore, the protection degrees under the six kinds of methods are discussed. Figure 6 shows the protection degree based on ICM under high individual privacy risk attack in the Facebook network. Figures 6(a) and 6(b) show 10% and 20% nodes of the whole network are attacked, respectively. For example, when 20% nodes are attacked, the protection degree is about 0.35 by our ICMP method, while that is 0 by the degree based method. The reason is that

according to equations (9)–(10) and the high individual privacy risk attack principle, the nodes with higher degree will have higher individual privacy risk. They are selected as the seed nodes by the degree-based method, that is to say, almost all the seed nodes will be attacked by the high individual privacy risk attack, and no information can be spread.

As mentioned above, some important conclusions can be drawn. At first, the attack effect for the ICPM method is less than other methods. Second, the influence set sizes are almost the same by the methods with or without edge weight modification. The reason for this behavior is due to the fact that the individual privacy risk and the cascade influence capability are two factors considered in the ICPM method. The edge weight modification is that the weight can be modified according to the actual environment and requirement.

5.3. Comparison Experiment Based on WCM.

At last, the comparison experiment based on WCM will be discussed. Different from the ICM, the activation probability between nodes in WCM is set as the inverse of the degree. The Facebook network is also utilized for the experimental analysis.

Figures 7(a)–7(d) show the influence set sizes when different number of nodes are attacked by high degree attacks in the Facebook network. It is easy to find that the ICPM method has higher influence set sizes than the other two kinds of methods. Assume 200 nodes are selected as the seeds. For the degree-based method, the influenced set sizes are decreased from 935 to 547, 365, and 136 when the attacked set sizes are 50%, 70%, and 90% of the seed set size, respectively. However, for the ICPM method, the influenced set sizes are decreased from 1250 to 1153, 1129, and 1096 when the attacked set sizes are 50%, 70%, and 90% of the seed set size, respectively. Furthermore, compared with Figure 3, it can be found that the influence set sizes based on WCM are higher than that based on ICM.

In addition, the protection degrees under six kinds of methods are discussed. Figure 8 shows the protection

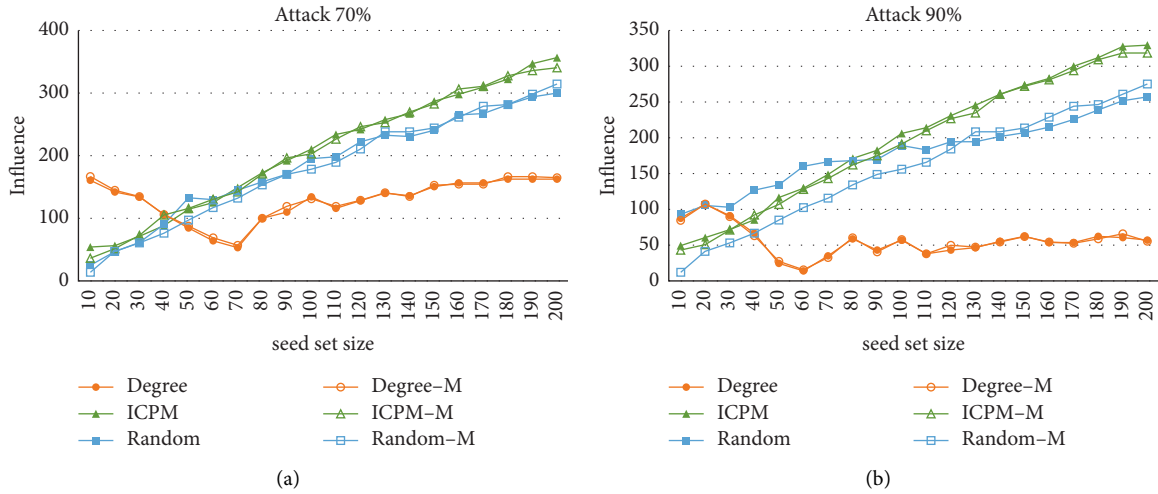


FIGURE 5: The influence based on ICM under high degree attack model in Facebook network.

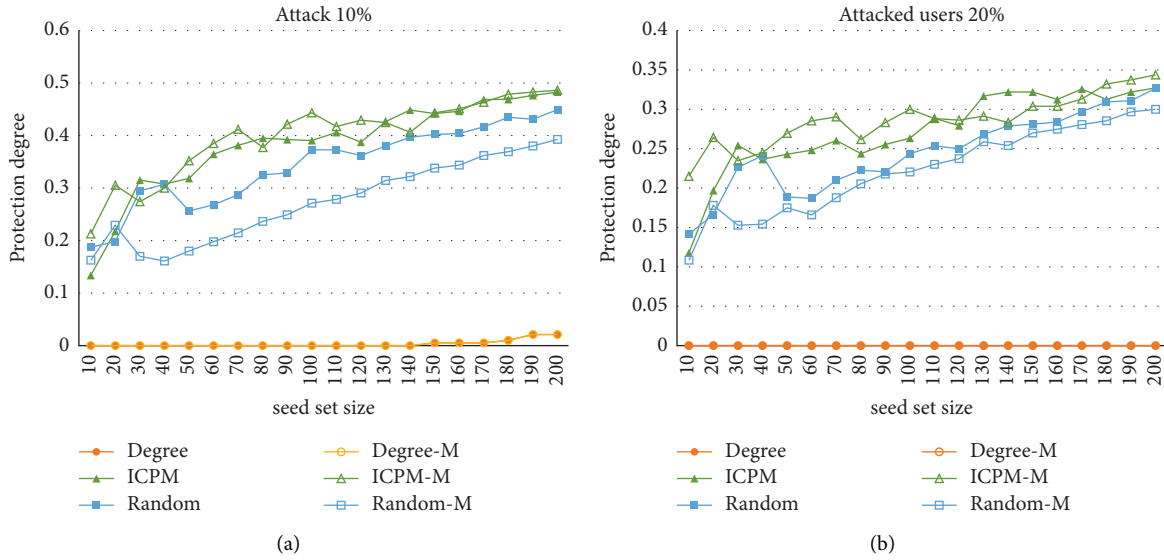


FIGURE 6: The protection degree based on ICM under high individual privacy risk attack in the Facebook network. (a) The attacked set size is 10% of the whole network. (b) The attacked set size is 20% of the whole network.

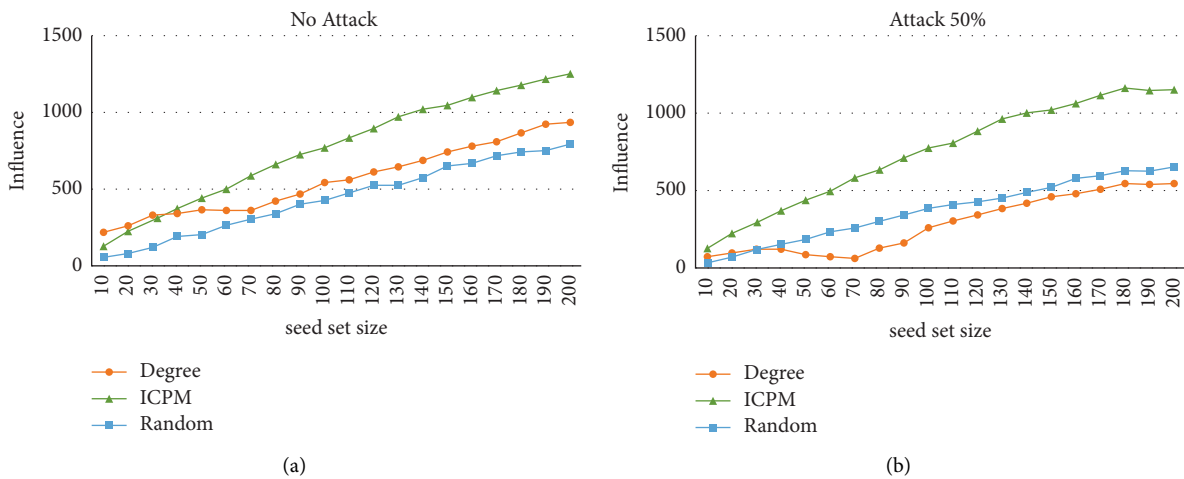


FIGURE 7: Continued.

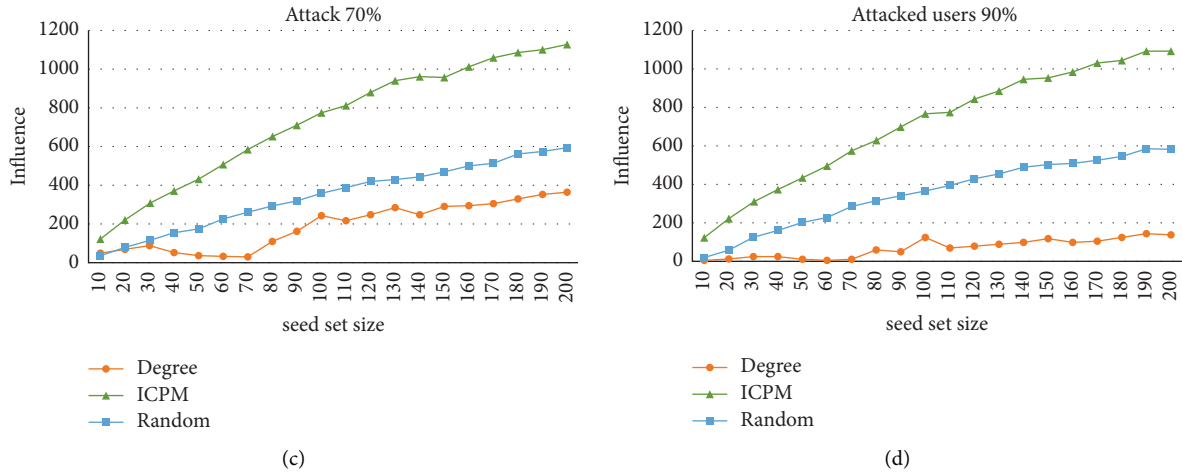


FIGURE 7: The influence based on WCM under high degree attack model in the Facebook network. (a) Not attacked, (b) the attacked set size is 50% of the seed set size, (c) the attacked set size is 70% of the seed set size, and (d) the attacked set size is 90% of the seed set size.

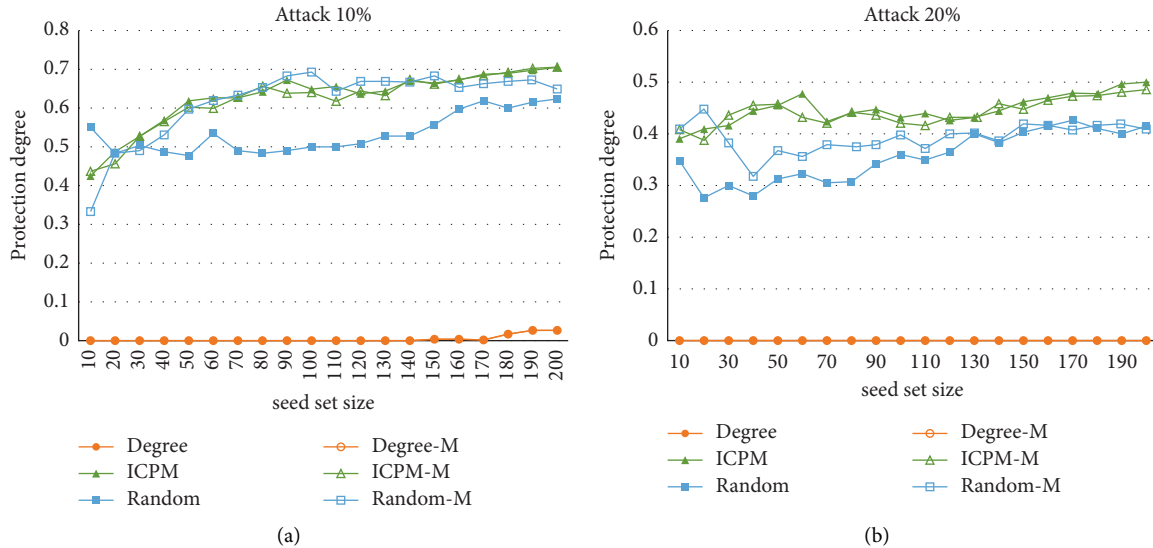


FIGURE 8: The protection degree based on WCM under individual privacy risk attack in the Facebook network. (a) The attacked set size is 10% of the whole network. (b) The attacked set size is 20% of the whole network.

degrees based on WCM under individual privacy risk attack in the Facebook network. Figures 8(a) and 8(b) show that 10% and 20% nodes of the whole network are attacked, respectively. It is easy to find that our ICPM method is affected slightly under the individual privacy risk attack, which can protect the privacy more. From these two figures, it can be found that the ICPM method has the highest protection degree, while the degree-based method is the worst method. From the analysis, it can be concluded that under different kinds of attacks, the ICPM method is affected slightly and its property of anti-attack is the best.

6. Conclusion

This paper focuses on the research of privacy protection model in social networks. One of our key methods beyond the existing literature is considering both the individual risk

and cascade influence capability: (1) Bayesian-based Individual Privacy Risk Evaluation Model (IPREM) is proposed to rank the individual risk levels; (2) by considering the influence capability, Cascade Influence Capability Evaluation Model (CICEM) is designed; and (3) an IPREM and CICEM based Privacy Protection Method (ICPM) is designed. It is the first attempt, to our knowledge, to consider jointly individual privacy risk and influence maximization on the privacy protection design. Finally, the performance and security are compared with different methods, and our method can obtain the highest influence set sizes and exhibit the best antiattack property when some attacks happened.

Our IPREM, CICEM models and ICPM method provide good starting points in the influence maximization privacy protection research in future social network. Further studies may concentrate on the temporal and spatial variation

environment, the case when the attacker has strong reasoning attack ability. Furthermore, the attributes are not discussed independent of analysis in this article. Next, the problem of what is the amount of private attribute leakage and privacy breach when the attacks happen will be discussed.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank the National Natural Science Foundation of China (nos. 61902069 and U1905211) and the Natural Science Foundation of Fujian Province of China (no. 2021J011068).

References

- [1] H. Kou, H. Liu, Y. Duan, W. Gong, and L. Qi, "Building trust distrust relationships on signed social service network through privacy-aware link prediction process," *Applied Soft Computing*, vol. 100, no. 5, Article ID 106942, 2021.
- [2] Z. Wang, Y. Li, D. Li et al., "Enabling fairness-aware and privacy-preserving for quality evaluation in vehicular crowdsensing: a decentralized approach," *Security and Communication Networks*, vol. 2021, Article ID 9678409, 11 pages, 2021.
- [3] J. Zhang, L. Xu, and P.-W. Tsai, "Community structure-based trilateral stackelberg game model for privacy protection," *Applied Mathematical Modelling*, vol. 86, pp. 20–35, 2020.
- [4] W. Liu, X. Chen, B. Jeon, L. Chen, and B. Chen, "Influence maximization on signed networks under independent cascade model," *Applied Intelligence*, vol. 49, no. 3, pp. 912–928, 2019.
- [5] Z. Hu, J. Yang, and J. Zhang, "Trajectory privacy protection method based on the time interval divided," *Computers & Security*, vol. 77, pp. 488–499, 2018.
- [6] H. Li, H. Zhu, S. Du, X. Liang, and X. Shen, "Privacy leakage of location sharing in mobile social networks: attacks and defense," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 646–660, 2018.
- [7] P. Gundecha, G. Barbier, J. Tang, and H. Liu, "User vulnerability and its reduction on a social networking site," *ACM Transactions on Knowledge Discovery from Data*, vol. 9, no. 2, pp. 1–25, 2014.
- [8] Z. He, Z. Cai, and J. Yu, "Latent-data privacy preserving with customized data utility for social network data," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 665–673, 2018.
- [9] Z. Cai, Z. He, G. Xin, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, pp. 1545–5971, 2016.
- [10] K. Oishi, Y. Sei, Y. Tahara, and A. Ohsuga, "Semantic diversity: privacy considering distance between values of sensitive attribute," *Computers & Security*, vol. 94, Article ID 101823, 2020.
- [11] P. Menard and G. J. Bott, "Analyzing IOT users' mobile device privacy concerns: extracting privacy permissions using a disclosure experiment," *Computers & Security*, vol. 95, Article ID 101856, 2020.
- [12] J. Zhou and C.-M. Pun, "Personal privacy protection via irrelevant faces tracking and pixelation in video live streaming," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1088–1103, 2021.
- [13] T. Ma, J. Jia, Y. Xue, Y. Tian, and A. Dhelaan, "Protection of location privacy for moving knn queries in social networks," *Applied Soft Computing*, pp. 1–14, 2018.
- [14] M. H. A. Ibrahim, K. Zhou, and J. Ren, "Privacy characterization and quantification in data publishing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 9, pp. 1756–1769, 2018.
- [15] B. Bostanipour and G. Theodorakopoulos, "Joint obfuscation of location and its semantic information for privacy protection," *Computers & Security*, vol. 107, no. 4, Article ID 102310, 2021.
- [16] Y.-C. Tsai, S.-L. Wang, H.-Y. Kao, and T.-P. Hong, "Edge types v.s privacy in k-anonymization of shortest paths," *Applied Soft Computing*, vol. 31, pp. 348–359, 2015.
- [17] H. Jianping, C. Lin, and G. Xinping, "Preserving data-privacy with added noises: optimal estimation and privacy analysis," *IEEE Transactions on Information Theory*, vol. 64, pp. 1–14, 2018.
- [18] M. Orooji and G. M. Knapp, "Improving suppression to reduce disclosure risk and enhance data utility," in *Proceedings of the SAVE Proceedings 2018 IISE Annual Conference*, Pennsylvania, PA, USA, May 2018.
- [19] J. Parra-Arnau, J. Domingo-Ferrer, and J. Soria-Comas, "Differentially private data publishing via cross-moment microaggregation," *Information Fusion*, vol. 53, pp. 269–288, 2020.
- [20] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 199–208, Paris, France, January 2009.
- [21] Y. Li, W. Dai, Z. Ming, and M. Qiu, "Privacy protection for preventing data over-collection in smart city," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1339–1350, 2016.
- [22] Y. Zhang, X. Chen, J. Li, D. S. Wong, H. Li, and I. You, "Ensuring Attribute Privacy protection and Fast Decryption for Outsourced Data Security in mobile Cloud Computing," *Information Sciences*, vol. 379, pp. 42–61, 2017.
- [23] Q. Li, R. Sandhu, X. Zhang, and M. Xu, "Mandatory content access control for privacy protection in information centric networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 5, pp. 494–506, 2017.
- [24] W. Dai, M. Qiu, L. Qiu, L. Chen, and A. Wu, "Who moved my data? Privacy protection in smartphones," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 20–25, 2017.
- [25] X. Li and X. Chen, "Factors affecting privacy disclosure on social network sites: an integrated model," *Electronic Commerce Research*, vol. 13, no. 2, pp. 151–168, 2013.
- [26] H. Chen, C. E. Beaudoin, and T. Hong, "Securing online privacy: an empirical test on internet scam victimization, online privacy concerns, and privacy protection behaviors," *Computers in Human Behavior*, vol. 70, pp. 291–302, 2017.
- [27] N. Z. Gong, "Joint link prediction and attribute inference using a social-attribute network," *ACM Trans. Intell. Syst. Technol*, vol. 5, no. 2, 2014.

- [28] R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Preventing private information inference attacks on social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1849–1862, 2013.
- [29] J. A. Hoeting, D. Madigan, and R. C. T. Volinsky, "Bayesian model averaging: a tutorial," *Statistical Science*, vol. 14, no. 4, pp. 382–401, 1999.
- [30] Z. Lei, M. Chunguang, Y. Songtao, and Z. Xiaodong, "Probability indistinguishable: a query and location correlation attack resistance scheme," *Wireless Personal Communications*, vol. 97, no. 4, pp. 6167–6187, 2017.