IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 31, 2020, accepted June 1, 2020, date of publication June 8, 2020, date of current version June 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3000780

Using User Behavior to Measure Privacy on Online Social Networks

XUEFENG LI[1,2], YANG XIN[1,2], CHENSU ZHAO[1,2], YIXIAN YANG[1,2], SHOUSHAN LUO[1,2], AND YULING CHEN[2]

[1]National Engineering Laboratory for Disaster Backup and Recovery, Information Security Center, School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guizhou 550025, China

Corresponding author: Xuefeng Li (lxf3710@bupt.edu.cn)

This work was supported in part by the Foundation of the National Key Research and Development Program of China under Grant 2017YFB0802300, in part by the Beijing University of Posts and Telecommunications Excellent Ph.D. Students Foundation under Grant CX2019319, in part by the Foundation of Guizhou Provincial Key Laboratory of Public Big Data under Grant 2017BDKFJJ015, and in part by the Major Scientific and Technological Special Project of Guizhou Province under Grant 20183001.

ABSTRACT Because social networks exemplify the phenomenon of homogeneity in complex networks, researchers generally believe that a user's privacy disclosure is closely related to that of the users around them, but we find that the related users studied in previous methods were not correct. That is, the analyzed user groups may have had nothing to do with the privacy disclosure of the target users. Since private information is time-sensitive, information held by users who are no longer in the same environment as the target user may no longer be true and have lost its value. For example, considering students and members of the working class, transfers to another school for further studies or job changes entail dramatic changes to most of their information. This lack of timeliness has an overarching impact on the effectiveness of social network analysis and privacy protection, but this problem has not been addressed by researchers. Therefore, we study and characterize this problem, add the user's behavior trace to solve this problem and measure the user's privacy status more accurately.

INDEX TERMS Social networks, privacy security, measurement, timeliness, network environment, behavior.

I. INTRODUCTION

Academic researchers focus on security issues in large complex networks, such as integration systems, the Internet of Things, cloud networks, and so on, because if these networks have security problems, it will cause great economic loss. In contrast, few researchers focus on individual security issues, for example, the leaking of large amounts of user information by Uber in 2017 [1] and Facebook's election incident and information disclosure incidents [2]. Although these incidents caused great repercussions, they did not cause great losses in a short time. Unlike incidents in industrial networks, the dangers of such incidents are persistent and far-reaching. Currently, OSNs (online social networks) have become an indispensable part of human life and the main way to obtain and share personal information. Additionally, operators can use users' personal information arbitrarily for many purposes, such as viral marketing, targeted advertising, instant advertising, network promotion, and censorship. As users disclose more personal information [3], [4], more malicious attacks may result in the real world and cyberspace [5]–[8], including but not limited to tracking, defamation, spam, phishing, identity theft, personal data cloning, and Sybil attacks [9].

The concept of privacy is subjective and complex, with no uniform definition [10]. It can be interpreted in different contexts and perspectives in disciplines such as law, health science, social sciences, and computer and information science. Privacy protection in different fields is not universal. The methods in mature industrial networks are not suitable for social networks. Generally, privacy can be defined as the right to be alone and to have freedom from interference or intrusion [11]. In addition, in social networks, facing such a complex and large network environment, homogeneity makes it impossible to consider privacy protection from only a personal perspective; the network environment in which the users are located also exposes private information [12], [13].

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

Moreover, the acceptance of privacy leakage by different users, such as public figures and ordinary users, varies widely. The existing methods of protecting users' privacy in social networks include anonymity, decentralization, encryption, information security regulations, fine-grained privacy settings, access control and enhancing users' privacy awareness and privacy behavior [14]. The first four methods rely on the operators of social networking sites in order to be implemented, which has proven to be unreliable [15]. At the same time, complex security settings compromise a user's experience. Studies have repeatedly identified users as the weakest link in security, but few studies used or focused on this concept and conducted research in this field [16]. In this paper, we utilize the last method, which is to cultivate the user's privacy awareness [17] and enhance privacy behavior [18] by privacy measurement to alleviate the problem of privacy leakage at the root.

However, we face a problem: this method needs to analyze too many users. According to the small world principle, any two users can reach each other within six steps. Just one user's link relationships can reach tens of millions of other users. With such a huge and complex network structure, it is difficult for traditional algorithms to deal with this situation effectively. At the same time, there are a large number of fake accounts, zombie accounts, marketing accounts, public figure accounts, official accounts, etc. [19]–[22]. These accounts do not affect the privacy of ordinary users. In other studies, this observation was not incorporated; they just analyzed all the friends around a user. This approach served as inspiration inspired: if the interference of these users can be reasonably excluded, we could just analyze the segment of users who affect the privacy disclosure of the target users, and the efficiency could be improved and the accuracy of privacy measures could be greatly improved.

In the process of selecting users who are closely related to the privacy leakage of target users, we find that there exist timeliness loopholes, which means that after successfully eliminating the redundant users mentioned above, we found that there were still some users who have a very close relationship with the target user in the graph structure but who hold information about the target user that has lost its privacy value. This segment of users may have been very close to the target user, but now they are irrelevant. After trying multiple methods, we decided to use the user's behavior trace to solve this problem. In previous studies, researchers considered implementing user behavior traces to measure privacy; however, their utility in excluding redundant users was left unexplored.

Therefore, based on our early work on individual privacy scoring methods [41], we extend the method from individuals to all users in the entire network structure graph and incorporate the behavior of users to make up for the shortcomings of previous existing studies. Our contributions in this paper are as follows:

1) We propose the concept of structural similarity and find the problem of timeliness in the traditional node importance ranking algorithm.
2) We propose the concept of behavioral intimacy to add user behavior characteristics.
3) We solve the problem of timeliness and eliminate redundant users; the resulting method is faster and more efficient than the traditional algorithm.
4) We integrate the proposed structure similarity, behavior intimacy and attribute similarity with our previous personal privacy measurement method and then propose a more comprehensive method PMoB (Privacy Measurement of Behavior) for measuring a user's privacy status.

II. RELATED WORK

In the early stage of privacy measurement research, researchers focused on the information disclosed by users' profiles. Maximilien *et al.* [23] innovatively used profile information to quantify the sensitivity and visibility of attributes, then measured the user's privacy status through a Bayesian model. Based on their research, Liu and Terzi [24] provide a more intuitive, more mathematically reasonable method to calculate user privacy scores in OSNs by combining sensitivity and visibility with IRT (Item Response Theory); this study greatly promoted the research of privacy measurement. In subsequent research, Fang and LeFevre [25] designed a template to provide a privacy settings wizard for users to guide them to choose reasonable profile settings. Jain and Raghuwanshi [26] designed a naïve formula to calculate the sensitivity of profile items, then considered both the sensitivity and visibility of the information in the user's profile and computed an index value on this basis. Xu *et al.* [27] found the key factors affecting users' self-disclosure of personal information. Aghasian *et al.* [28] measured users' profiles published on multiple social networks and obtained the users' privacy disclosure status; they are the first to propose privacy measurement in multiple social networks.

Later, researchers found that the homogeneity of social networks results in the user's privacy being unable to be separated from the entire network environment. Users have a greater risk of privacy disclosure on social networks: the more friends there are around one person, sharing their lives and caring about their privacy, the greater the risk to the privacy of this person [29]. Zeng *et al.* [30] argued that an individual's risk of privacy breach depends on his friend's privacy protection and creatively proposed a trust-based framework to evaluate users' privacy disclosure. Alsakal *et al.* [31] introduced a unified information measure to a user's whole social network by using the theory of information entropy, then discussed the impact of individual identification information and its combinations on users' information leakage; it provides a new direction for future research. Pensa and Di Blasi *et al.* [32] innovatively designed a community group-based privacy

scoring method to measure users' privacy disclosure and proposed an online learning method to help users make changes to their privacy settings.

In recent years, some researchers have shifted their research direction to complex network graph structure analysis. Fan *et al.* [33] studied the similarity between users in heterogeneous information networks to detect a certain type of user, then use this similarity to link the relationship between users. Oukemeni *et al.* [11] aim to provide users as well as system providers with a measure of how much the investigated system is protecting privacy; this is the first study to measure privacy at the overall level of social websites. Yu *et al.* [34] analyzed various possible privacy leakages in social networks and constructed a model to divide the information in social networks into multiple categories in various situations, such as a user's grouping situation, the methods of controlling and reposting information, the situation after deleting information, and the situation after replying to information, then summarized them separately to measure the privacy of multiple social networking sites. Serfontein *et al.* [35] used ad hoc networks to explore threats to user information security in social networks; they proved that it is easier to identify risks in a self-organizing map. Shi *et al.* [36] defined the entropy value of a network graph innovatively, which quantifies the structure of a social network graph to obtain the user's privacy metric. The greater the entropy value is, the better the privacy protection of the user. Djoudi and Pujolle [37] used graph structure analysis to describe the trust index in social networks and to propose a contagious communication framework to test the impact of over-trust or low trust on the entire network and then help users to distinguish friends with low trust.

As the research progressed, researchers found that it was unrealistic to consider the network graph structure itself because users ultimately expose their private information through their own behavior. The high interaction of interpersonal communication in OSNs prompts us to regard privacy as a public affair. Users' private information is disclosed not only through their voluntary disclosure but also through their social activities [31]. A user's behavior information may be used by other services, but users do not consider security issues such as central node detection [38] and recommendation services [39]. Belanger & Xu argue that researchers should "focus on actual behavior rather than personal intentions" when researching information privacy, considering that users' behavior may not always be reasonable in regard to privacy [40]. That is, we should pay attention to what users do, rather than what they want to do, because users' privacy awareness and behavior are often divorced and inconsistent.

The paper is organized as follows: In Section III, we provide the problem description and notation used in our paper. Section IV proposes the privacy measurement method. In Section IV, part *A* describes the specific information in the datasets we collected and why we created a dataset independently. Part *B* introduces the method of calculating a user's structural similarity. We want to filter out the redundant users in the user's environment through structural similarity. However, we found that the results did not incorporate timeliness. The calculation of behavioral intimacy and relational degree is described in Section *C*. Through the introduction of behavioral characteristics, the problem of the timeliness of private information is solved, and then the total degree of the relationship is obtained by combining structural similarity and behavioral intimacy, which can more accurately filter a users' friends. *D* introduces attribute similarity, which measures information leakage in user profiles. In *E*, the relational degree and attribute similarity are used to calculate the metric value of the user's privacy disclosure. In F, we analyze the complexity of algorithm. Finally, Section V experimentally demonstrates the method we propose.

III. PROBLEM DESCRIPTION AND NOTATION

In this paper, our goal is to use the various information published by users and the network environment in which they are located to make an accurate and reasonable assessment of users' privacy status on social networking sites. To achieve this goal, we first regard a social network as a large complex network graph $G = (V, E, P, B)$, in which the node set V represents users in the social networks and the edge set E represents the relationships between users. In our research, in order to accurately express user relationships, we use a bilateral edge to distinguish the followed and follower relationship between users on social networks. $P = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7 \ldots\ldots\}$ represents the personal attribute information of all users, and each p in P represents the personal information filled in by a user, which is likely to map to a certain person in the real world. $p = \{a_1, a_2, a_3, a_4, \ldots\ldots\}$ represents a user's specific attribute information, and each a represents an attribute content. B represents a set of user behavior on a social website. We collected all the microblog and behavior information sent by users for a period of time and extracted all the behaviors of liking, reposting, commenting, @-mentioning, and providing topic content (i.e., content in hashtags). Specific symbols are shown in Table 1 below:

IV. METHODOLOGY

A. DATASETS

To verify the effectiveness of PMoB, we collected real experimental data from Sina Weibo for students, teachers, staff and their friends at our university. These data include all personal information completed in profile and microblog content published on Sina Weibo from October 2016 to April 2018. We cleaned up and performed a statistical analysis of these data and then obtained attribute content and statistics of behavior in the last six months, such as likes, reposts, comments, @ and topic tags.

We collect data from the people at our university is to ensure that there are a certain number of common friends between users and thus a more centralized small world system can be formed, similar to an entire social network. Another reason is that among existing public datasets, social network

TABLE 1. Notation.

Notation	Description	Notation	Description
V	User node set	N_{OAB}	Number of microblogs sent by other users containing behavior for A and B
E	Edge set	N_{ABO}	Number of microblogs sent by A and B users containing behavior for other users
P	User Personal Information Set	d_A	Degree of A (number of A's follow and follower)
B	User Behavior Set	d_{AB}	Number of identical friends of A and B
p	Single user attribute information	$\tilde{d}_{AB}$	Number of identical friends who interact with A and B
a	User attribute content	d_O	Number of Users with which A and B interact
N_l	Number of likes	$\tilde{N}_{OAB}$	Number of interactive behaviors from common friend to A and B
N_r	Number of reposts	γ_{AB}	behavior closeness from A to B
N_c	Number of comments	γ_{ABO}	behavior closeness from A and B to others
$N_@$	Number of @	γ_{OAB}	behavior closeness from others to A and B
$N_\#$	Number of same topics	$S(A,B)$	Structural similarity of A and B
N_A	Number of microblogs sent by User A	$R(A,B)$	Relational degree of A and B
N_{AB}	Number of microblogs sent by A containing behavior toward B	$F(A,B)$	attribute similarity of A and B
P_A	Friends with High Relational degree of A		

datasets are very scarce because such datasets involve users' personal privacy information. None of the few datasets available contain the user's network structure, personal information, and published text information. Therefore, to validate our approach, we had to collect and organize the dataset ourselves. In addition, to expand the dataset, we specially selected some users' roommates, classmates, tutors, etc. to collect data.

To increase the credibility of our experiments, we collected four datasets. We randomly selected 279 users of Sina Weibo at our school for the survey, including undergraduates, master's students, doctoral students, teachers, and staff. Due to space limitations in the present paper, seven users with large differences in privacy status were selected for the follow-up experiment. Dataset 1 contains all of the information of the 7 users and their lists of follows and followers. Dataset 1 contains 1385 user nodes and 2938 edges. To expand our dataset and the diversity of experimental users, we collected 9 other people who are the 7 users' roommates, classmates or friends to form Dataset 2. Dataset 2 contains 3244 user nodes and 6452 edges. At the same time, in order to increase the amount of data in our dataset, we crawled the data of the 7 users' friends' friends to form Dataset 3, with a total of 428614 user nodes and 929620 user relationship edges. Dataset 4 is crawled from friends and friends' friends of the 16 people in Dataset 2, with a total of 704,903 user nodes and 1,527,106 edges. Another reason for choosing users from our school instead of randomly selecting users from social networking sites is that we need to obtain accurate social relations between the target users and their friends in the subsequent experimental verification. Moreover, to determine if the target users filled in fake personal information, we confirmed everyone's profile separately. Later, experiments will be conducted on these four datasets.

TABLE 2. Datasets.

Dataset	Nodes	Edges
Dataset 1	1385	2938
Dataset 2	3244	6452
Dataset 3	428,614	929,620
Dataset 4	704,903	1,527,106

It should be noted that when crawling the data, we have carried out special processing on the user group with a huge number of friends, because most of these users are public figures or official institutions, and they have little impact on the privacy disclosure of ordinary users. If the corresponding processing is not carried out, the number of friends of a public figure account may exceed 100 million, which will lead to high computational complexity and large deviation in the analysis of target users. Therefore, for this part of users, we only crawl a part of their friends list. Meanwhile, among the 16 target users, two are doctoral students, eleven are master students and three are undergraduates. This composition is the key to the discovery of information timeliness.

To calculate the final privacy score, we need to obtain the sensitivity of various attributes. In previous studies, Srivastava *et al.* [42] and Liu and Terzi *et al.* [24] used Bayesian statistics to calculate the sensitivity of 11 attributes from profiles filled in by users on social networking sites.

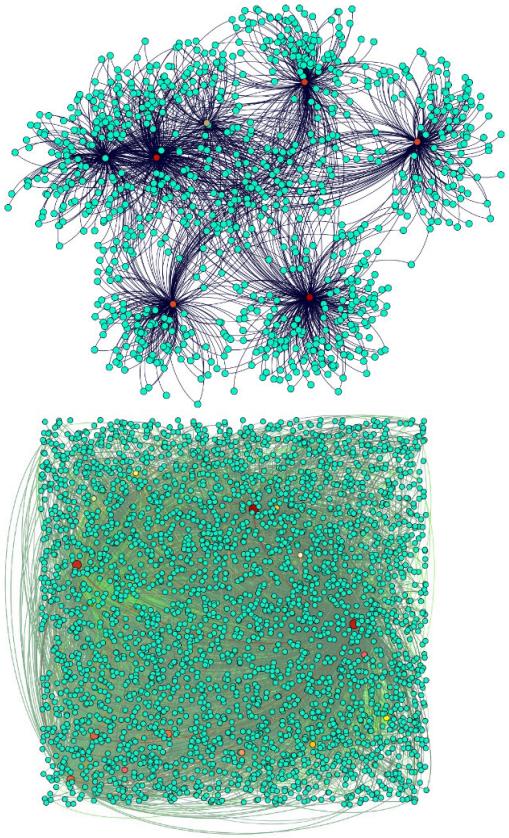

FIGURE 1. Dataset 1(up) and Dataset 2(down).

However, we think that the published attribute content on social websites does not truly reflect the user's perception of the sensitivity of the attribute because most users' privacy literacy cannot effectively support correct privacy behavior decisions. Therefore, we decided to use a questionnaire on the Internet so that users could record their most direct and true feelings (https://www.wjx.cn/jq/01647730.aspx). After a piece of certain private information is leaked, the user's extent of concern is set as an option, and users can choose the option according to their most direct thoughts. This allowed us to objectively and accurately calculate the privacy sensitivity of attributes. We designed five options: L1, not worried at all; L2, not worried; L3, not clear; L4, worried; and L5, very worried. The result of each coefficient is the percentage of the number of people who selected this option. We used the coefficient L3 as the benchmark to calculate the sensitivity of each attribute, and the calculation is shown in Formula (1). The larger the numerical value is, the more sensitive the attribute is, and the more worried the user is about the leakage of this attribute content. The results are shown in Table 3.

$$\theta = \frac{0.5 * L3 + L4 + 1.5 * L5}{1.5} \tag{1}$$

For example, if there are 100 users who participate in the questionnaire and choose the sensitivity of a certain attribute, the number of people who choose L1 is 10, L2 is 12, L3 is 25, L4 is 36, and L5 is 17; the final result is $(0 * 0.1 + 0 * 0.12 + 0.5 * 0.25 + 1 * 0.36 + 1.5 * 0.17)/1.5 \approx 0.4933$. In addition, 0, 0.5, 1 and 1.5 are adjustment coefficients, which are set to simply distinguish the worrying degree of different options, and 1.5 in the denominator is to ensure that the result is a number between [0,1]. However, these coefficients are not fixed. Other researchers can adjust them according to their own research. If these coefficients need to be applied to a general situation, they need to rely on the research of social psychology, which is not the focus of this paper, so we do not' provide an extensive description.

The flowchart of PMoB is shown in Fig. 2.

TABLE 3. Attribute sensitivity.

Attribute	Sensitivity
Username	0.2381
Avatar	0.3553
Phone number	0.5669
Email	0.3260
Hometown	0.2253
Birthdate	0.2748
Address	0.4212
Job Details	0.2024
Relationship Status	0.1731
Interests	0.1255
Education	0.1575

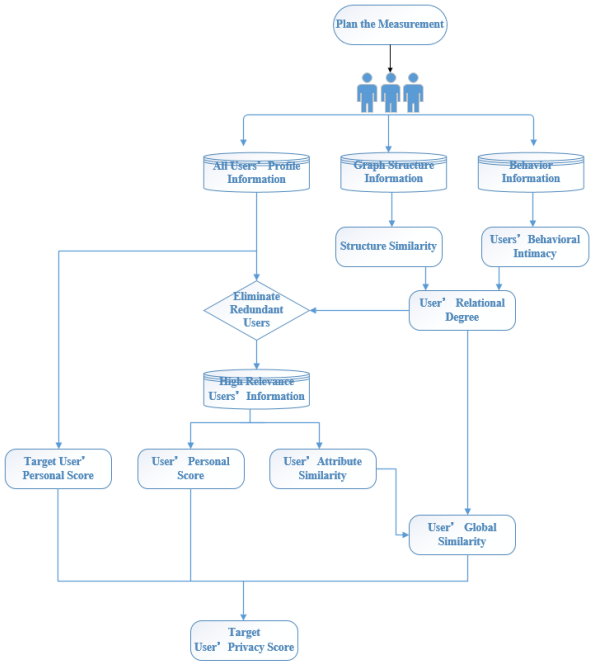

FIGURE 2. Overview of privacy measurement method.

B. STRUCTURE SIMILATITY

The purpose of calculating structural similarity is to screen out users who are closely related to the target users, as social networks are full of official accounts, marketing accounts, public figure accounts, zombie accounts and fake accounts that send spam, as well as some accounts that have been abandoned. Sina Weibo, for example, reached 340 million users by 2018, of which more than one third are estimated to be zombie accounts and multiple accounts for the same user. In 2019, Facebook released the Community Standards Implementation Report, which showed that more than 5% of Facebook's 2.4 billion monthly active users are fake accounts that spread spam, fake news and other inappropriate statements. From October to December 2018, the site blocked 1.2 billion fake accounts (https://transparency.facebook.com/community-standards-enforcement). These large numbers of abnormal accounts and public figure accounts have no impact on users' privacy disclosure, and it is not necessary to analyze these accounts. In fact, such accounts reduce the accuracy of the results. At the same time, as shown in Table 2, the number of users of social networks has been increasing exponentially with the further expansion of the step, requiring extensive amounts of computation, space and time consumption. Therefore, we first use the similarity relation in the network graph structure to eliminate redundant users.

To efficiently calculate the similarity between users in the network graph structure, we have tried many algorithms, such as degree centrality, closeness centrality, betweenness centrality, Laplace centrality, eigenvector centrality, and community identification algorithms. Finally, through the comparison of the time complexity and effect, we chose SimRank as the basis for our solution. SimRank is a topological information algorithm based on a complex network graph structure that is used to measure the degree of similarity between any two objects. The core idea is that if two objects are referenced by similar objects at the same time, the two objects are also similar. In recent years, the algorithm has received extensive attention in the field of information retrieval and has been successfully applied to web page sorting, collaborative filtering, outlier detection, network graph clustering, approximate query processing, etc. [43] Its core concept is highly consistent with the homogeneity of social networks. However, the algorithm also has some shortcomings which is the graph structure it was originally used for has no bidirectional edges. It cannot be applied to a bi-partite network structure graph. Therefore, we improved the original formula to allow its application to the bilateral relationship structure of social networks.

The improved formula is as follows:

$$S^{i}_{k+1}(a,b) = q_i \frac{C}{|I(a)|\,|I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} S_k(I_i(a), I_j(b))$$

$$S^{o}_{k+1}(a,b) = q_o \frac{C}{|O(a)|\,|O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} S_k(O_i(a), O_j(b))$$

$$S(a,b) = S^{i}_{k+1}(a,b) + S^{o}_{k+1}(a,b) \quad (2)$$

In Formula (2), C represents the damping coefficient, which has a value of 0.8, $I(a)$ represents the in-degree of node a, and when $I(a) = \emptyset, I(b) = \emptyset$, then $S(a,b) = 0$. $q_i = \frac{1}{1+e^{-(|I(a)|\cap|I(b)|-2)}}$ and $q_o = \frac{1}{1+2e^{-(|O(a)|\cap|O(b)|-5)}}$ represent the adjustment coefficients of the in-degree and out-degree. Two and five are the average numbers of the same in-degree and out-degree nodes between two nodes in a graph, which represent the average numbers of the same follows and followers between two users in the social network.

To verify the effectiveness of our improved algorithm, we validated and counted the social relationships with target users (if the algorithm outputs multiple users with the same similarity, we randomly select them according to the output order.)

TABLE 4. User analysis with higher structural similarity.

	Top 5	Top 10	Top 15
Colleague and friends	13	30	43
Middle school classmates	9	17	25
High school schoolmates	17	26	45
Undergraduate classmates	20	47	62
Postgraduate classmates	5	13	25
Kinsfolk	6	9	14
Others	10	18	26

As shown in Table 4, most of the users output by our improved algorithm are undergraduate classmates. These users have a certain intersection with the target users in real life and are relatively familiar with their private information. Our algorithm effectively excludes public figure accounts, marketing accounts, fake accounts and so on, greatly improving the efficiency of social network analysis. However, in the analysis of the results, we unexpectedly found a phenomenon that has been ignored by other researchers. Datasets 2 and 13 of the 16 target users are composed of graduate students, but the majority of users who are close to the target users in the results are undergraduates. That is, most of the users with the highest similarity of the algorithm output are not in the same environment as the target users. In the analysis of the results for master students, the most intimate users are not their classmates and friends around them in the current environment but their classmates and friends during the undergraduate course. Similarly, doctoral students are the most closely related to the master students in structural similarity. Analyzing the users who have graduated and entered the workplace revealed that their structural similarity is most closely related to their classmates and friends at school before graduation. In today's social networks, the main members are young people, the vast majority of whom are students and new workers. According to a Sina media white paper in 2018, 26.48% of Sina Weibo users are of the Post-90s generation, 22.1% are Post-95s, and 18.2% are Post-00s; that is, more than 66% of users are young people under 30 years of age, and 90% of these people are students or newcomers entering the workplace.

TABLE 5. User analysis of other algorithms with higher structural similarity.

	Colleague and friends	Middle school classmates	High school schoolmates	Undergraduate classmates	Postgraduate classmates	Kinsfolk	Others
Degree centrality	12	10	16	31	9	5	17
closeness centrality	18	9	11	15	18	3	26
betweenness centrality	15	7	11	18	21	3	25
eigenvector centrality	26	4	5	14	29	2	30
PageRank	26	3	9	16	21	2	23

With the development of the Internet, network users are becoming younger and younger, and this phenomenon will become more common. Therefore, the problems we have found are widespread, and the phenomenon represents the lack of information timeliness. This deficiency is reflected in the fact that some of the privacy information about the target user exposed by the person with the highest similarity has no timeliness. If the information loses its timeliness, it cannot be called privacy information. This part of the information is meaningless. For example, when an undergraduate enters a new campus after graduation, the addresses, educational information, email, relationship status, and even telephone numbers exposed by former friends lose their validity. Publishing this information will not result in leakage of privacy. If a student graduates and enters society, his work information, address and so on will change accordingly.

To test whether other existing algorithms are also subject to these problems, we ran the existing algorithms to calculate the importance of nodes in complex network graph structures in Dataset 2. We selected the degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, and PageRank algorithm and performed a statistical analysis of the top 100 nodes in the results of each algorithm to obtain the relationships between them and the 16 target users. The results are shown in Table 5.

As shown in Table 5, the results of these algorithms generally have two problems. The first is the timeliness we mentioned before. The output of the algorithm is mostly undergraduate classmates. However, most of our target users are graduate students. The second problem is that public figure accounts in the Others category occupy a large proportion because these users have a large number of followers. The closeness centrality algorithm considers that nodes through which other nodes can be reached more quickly have higher centrality and significance. In the eigenvector centrality, nodes are more important if they connect more of these significant nodes. The PageRank algorithm makes some compromise, but the effect is still not obvious.

Given the problems found above, we hope to find a way to solve the lack of timeliness while eliminating interference from public figures and other redundant users as much as possible to more accurately and reasonably select users who threaten other users' privacy leakage. After a variety of attempts, we find that the behavioral characteristics have strong time continuity, which can solve the problem of timeliness to a great extent. At the same time, most public figure users do not have any behavior toward ordinary users, and thus we decided to add traces of the user's behavior to solve the problem of timeliness.

C. BEHAVIORAL INTIMACY AND RELATIONAL DEGREE

In social websites, users can perform very limited behavior. Adding friends expands the relationship network, and browsing content and acquiring and sharing various information items is the fundamental purpose of users in using social networks. Among possible behaviors, liking, commenting, reposting, @-mentioning a specified user and participating in topic discussions are the most frequently performed actions of users. Therefore, we analyze and measure these behavior traces of users, then exclude redundant users by considering their structural similarity.

To accurately distinguish the differences between these behaviors, we divide them into three groups: the first group is behavior between A and B, including liking, reposting, commenting and @-mentioning. The second group is behavior of A and B directed toward others, including jointly liking, reposting, commenting on a microblog, @-mentioning the same user, and participating in the same topic. The third group is about other people's behavior toward A and B. Importantly, non-friend relationships can also repost and comment on social networking sites; however, it is too difficult to obtain these data, so we consider only the behavior of common friends toward A and B, including the number of times they like, repost and comment. It is worth noting that if they act only on one of A or B, this will not be included in the calculation; that is, the value of this item is 0.

In addition, we need to distinguish the effects of various behaviors on behavioral intimacy. For this reason, we set different influence coefficients for these behaviors. The size of the coefficients is determined according to the cost in time and effort of implementing these behaviors, which reflects the degree of attention to the target users. To show the effectiveness of the parameter, we counted all the likes, reposts, comments, @-mentions and topics in the original Weibo data in our dataset. The resulting statistics are consistent with our

TABLE 6. Impact of the coefficient of behavior.

Behavior	Number	Coefficient
like	13733841175	0.0123
repost	9038075110	0.0187
comment	2451047898	0.069
@	623387	0.0546
topic	748136	0.0454

hypothesis, that is, the less the energy spent on implementing a behavior is, the greater the number of instances of this behavior, the lower the importance of this behavior, and the less influence it has on the target users. We take the liking behavior as an example to obtain the final coefficients through the following formulas.

$$\zeta_l = \beta \frac{n_{repost} n_{comment}}{n_{repost} n_{comment} + n_{repost} n_{like} + n_{like} n_{comment}} \quad (3)$$

We take the three independent behaviors of liking, reposting, and commenting as a group, take the @-mentions and topics included in the Weibo content as a group, and quantify them separately, as shown in (3). The value n represents the number of behaviors, and the resulting coefficient is shown in Table 6. The β value adjusts the proportional relationship between behavioral intimacy and structural similarity. In this paper, its value is 0.1. Without this adjustment coefficient, structural similarity will be obscured by behavioral intimacy.

The formula for computing the behavioral intimacy of user B to target user A is:

$$\gamma_{BA} = \frac{N_{BA}}{N_B}(1 + \zeta_l N_l + \zeta_r N_r + \zeta_c N_c + \zeta_@ N_@ + \frac{N_{BA}}{d_B}) \quad (4)$$

In Formula (4), $\frac{N_{BA}}{N_B}$ represents the proportion of the number of microblogs containing B-to-A behavior to the total number of microblogs sent by B, the 1 in parentheses represents its proportion, and the second to fifth items represent the number and weight of each behavior. On the one hand, this represents the importance of B to A. On the other hand, it also distinguishes public figure users and fake accounts because these users will post a large number of microblogs, and even if they perform a certain number of behaviors on the target user, it will not distort the intimacy calculation. The last item is the number of microblogs about A sent by B multiplied by the inverse of the number of B's friends; this means that the more behaviors there are from B to A, and the fewer the friends of B, the stronger the concern B has for A, which means B may be able to reveal more private information about A. In addition, if B does not have any behaviors toward A, the numerator is 0, and γ_{BA} is 0.

The formula for computing the behavioral intimacy of A and B to other users is:

$$\gamma_{AB0} = \frac{N_{ABO}}{N_A + N_B}(1 + \zeta_l N_l + \zeta_r N_r + \zeta_c N_c + \zeta_@ N_@ + \zeta_\# N_\# + \delta \frac{N_{ABO}}{d_O}) \quad (5)$$

The numerator of $\frac{N_{ABO}}{N_A+N_B}$ means that as long as A and B act on the same user, the microblog containing these behaviors will be counted, and the denominator represents the sum of microblogs published by A and B. δ is an adjustment coefficient, which is 0.1 in this paper. The purpose of this coefficient is to adjust the proportion of each item. The last item in parentheses indicates the total number of microblogs published by A and B regarding the same users multiplied by the inverse of the number of these users. This item reflects the users that A and B are concerned about at the same time, even public figure users, and it can also reveal users' interests and hobbies. The larger the numerical value is, the more concentrated their interests are.

The formula for computing the behavioral intimacy of other users to A and B is:

$$\gamma_{OAB} = \frac{d_{AB}}{d_A + d_B}(1 + \zeta_l N_l + \zeta_r N_r + \zeta_c N_c + \zeta_@ N_@ + \zeta_\# N_\# + \frac{N_{OAB}}{\widetilde{d_{AB}}}) \quad (6)$$

In Formula (6), $\frac{d_{AB}}{d_A+d_B}$ represents the proportion of common friends of A and B to the total number of their friends. The last item in parentheses indicates the number of microblogs of common friends regarding A and B multiplied by the inverse of the number of these friends. The larger the numerical value is, the stronger the attention of common friends is toward A and B, and the greater the possibility that they will expose their private information.

Formulas (5) and (6) are the same when calculating the behavioral intimacy from A to B and from B to A. So far, we have obtained the behavioral intimacy between users in the relational network. Next, we combine structural similarity and behavioral intimacy to obtain the relational degree between each user and the target user, then sort them to obtain the users who are closely related to the target user. The relational degree is calculated as follows:

$$R(A, B) = (S^{\mathrm{i}}(A, B) + S^{o}(A, B))(\gamma_{BA} + \gamma_{ABO} + \gamma_{OAB}) \quad (7)$$

It is worth noting that if the relationship between user B and the target user A is unilateral, then the other item in the first parenthesis is 0; that is, if A pays attention to B, whereas B does not pay attention to A, then B is the out-degree node of A, so $S^{\mathrm{i}}_{k+1}(A, B) = 0$ and vice versa. Regarding the validity of the relational degree, we will verify it in the experiment.

D. ATTRIBUTE SIMILARITY

In social networks, the most direct way to expose users' privacy is the information they add to their profiles. Because social networks belong to a complex network that conforms to the phenomenon of homogeneity, even if users themselves do not expose too much personal information or fill in fake personal information, their true information can still be revealed by surrounding people with whom they are relatively intimate. Because homogeneity means that the attributes of connected users are more similar, the attribute content can

be obtained by inferring the attributes of friends around the target users. After eliminating a large number of redundant users through the relational degree, we can calculate the attribute similarity between the intimate users and the target users to obtain more accurate privacy measurements.

We list 11 attributes in Table 3. When calculating attribute similarity, birthday only matches the year, and hometown and address only matches the city. The reason for considering the avatar and username is that if avatars and usernames are very similar, the users tend to be very close to or multiple accounts of one user. Because the avatar image files are usually small, we use the perceptual hashing algorithm for judgment. In judging whether the usernames are consistent, we use the minimum edit distance algorithm, which outputs a decimal between 0 and 1. When the output value is greater than 0.8, it is considered consistent. We choose these two algorithms because they are simple, efficient, and suitable for a large amount of trivial data. In addition, when filling in addresses and other information on social networks, service providers often set the city as an option to choose rather than allowing it to be filled in freely, which also increases the readiness of the algorithm. For example, there are only three distances between No. 15 Xitucheng Road and No. 15 Beitucheng Road but six distances between No. 15 Xitucheng Road and No. 15 Xitucheng Garden, even though the similarity of the latter pair is greater than that of the former. Consequently, we just match the city. When matching user's educational information and interest, if the user fills in more than one piece of information, as long as one of them matches each other, we judge the attribute to be consistent. If the content of attribute a is consistent, the value is 1 and 0 otherwise. We set eleven matched attributes as a one-dimensional row vector κ with values of 0 or 1 and set the sensitivity of the eleven attributes as a one-dimensional column vector μ. Finally, we can obtain the attribute similarity $F(A, B) = \kappa(A, B) \cdot \mu$.

E. PRIVACY SCORE

After sorting the relational degrees to screen out users, we combine the relational degree and attribute similarity to obtain their global similarity, and then we use the global similarity to calculate the privacy scores of target users. The calculation method is as follows:

$$G(A, B) = \frac{R(A, B)(1 - e^{-F(A,B)})}{R'(A, B)} \tag{8}$$

where $R'(A, B) = \frac{\sum_{C \in P_A} R(A,C) + \sum_{D \in P_B} R(B,D) - R(A,B)}{|P_A| + |P_B| - 1}$

Consequently, we have obtained the users who are closely related to the privacy status of the target user and obtained the overall similarity between these users through structural similarity, attribute similarity and behavior intimacy. In our previous research [41], we proposed a method to quantify the extraction difficulty, reliability, accessibility and privacy awareness of users based on the attribute content in their profiles in multiple social networks and then quantify the personal privacy score by incorporating the attribute sensitivity.

Next, we will obtain a more accurate and reasonable privacy score based on the previous personal privacy score and the behavior in the user's entire network structure. The calculation method is as follows:

$$Score(A) = \frac{\eta(A) + \sum_{B \subseteq P_A} G(A, B)\eta(B)}{n + 1} \tag{9}$$

Here, $\eta(A)$ represents the user's personal privacy scoring method of multiple social networks, which was proposed in our previous study [41]. Because it rests on our previous work, and due to the limitations of space in this paper, we will briefly introduce it. Through the attribute information filled in regarding the user's profile, we quantify the extraction difficulty, accessibility, reliability and privacy awareness, then obtain the overall visibility through a machine learning algorithm, and finally obtain the user's personal privacy score combined with the sensitivity of the attribute. However, one thing to note is that in the calculation of $\eta(A)$ and $\eta(B)$, we cannot obtain the personal information of the target user's friends on multiple social networks, so we simplify the previous method to obtain the personal privacy scores of these users. In this paper, reliability and privacy awareness were quantified using the average values of 0.73 and 0.54. When quantifying extraction difficulty and accessibility, the settings in the Sina Weibo platform were used. As a result, we can obtain users' privacy scores based on their profiles on Sina Weibo. The inspiration for our calculation is that the user's privacy exposure depends on the user himself and the friends around him. Ultimately, the higher the privacy score is, the more serious the possibility of privacy leakage, and this should attract more attention.

F. COMPLEXITY

According to the flowchart in Figure 2, we analyze the complexity of the method. In the calculation of structural similarity, the complexity is $O(Kn^2)$. Where K represents the number of iterations and n represents the number of friends. In the survey of Sina Weibo users conducted by Lei K *et al.* [44], less than 1% of Sina Weibo users have more than 1000 friends. Meanwhile, as we said in Section IV A, we processed the data crawling for users with a large number of friends, and only randomly crawled 2000 friends for users with more than 2000 friends. Therefore, the maximum n value here is 2000. In the calculation of behavior intimacy, relational degree, attribute similarity, global similarity and privacy score, each of them has a complexity of $O(c)$, which represents constant time. Because they are all basic operations.

V. EXPERIMENTAL EVALUATION

A. EXPERIMENT 1

In experiment 1, we confirm the validity of the relational degree proposed above. To make a more intuitive and full comparison, we used Dataset 2 to conduct experiments, then compared the results with Tables 4 and 5. We analyzed and compared the top 5, top 10, and top 15 friends according to

TABLE 7. Analysis of users with higher structural similarity according to our algorithms.

	Colleague and friends	Middle school classmates	High school schoolmates	Undergraduate classmates	Postgraduate classmates	Kinsfolk	Others
Top5	16	6	12	13	23	4	6
Top10	34	13	22	24	46	7	14
Top15	47	21	29	43	69	11	20
Top100	25	4	10	14	34	4	9

16 target users' relational degrees and the top 100 friends in all target users' total rankings. The results are shown in Table 7:

Table 7 shows that the number of postgraduate classmates who are in the same environment as most target users, which is assessed by the relational degree, has been effectively increased. Because most of the target users we choose are graduate students, the proportion of undergraduate students has decreased in the result. There are three undergraduates in the target user group, so undergraduates still account for a certain proportion of the overall results. Meanwhile, the number of public figure users, marketing accounts and fake accounts decreased significantly. Compared with the results of other methods in Table 5, the results of PMoB are more effective.

To further demonstrate the effectiveness of our algorithm, we use Dataset 4 to perform the above experiments again in order to verify the stability of the algorithm in a larger network. The experimental results are shown in Table 8 below.

From the data in Table 8, we can see that when the network structure becomes more complicated, the resulting number of the Others category increases. This is because, in the two-step network, more follows and followers of public figure users are included, which leads to a significant increase in structural similarity. Furthermore, since the intimacy of behavior has not changed, the relational degree of some users in the previous environment has strengthened. Relatively speaking, the number of users in the current environment has decreased. Nonetheless, the effectiveness of the algorithm has not decreased significantly. From this, we can see that the algorithm has a certain stability in more complex and larger networks.

B. EXPERIMENT 2

In this section, we use the users in the groups Top5, Top10, and Top15 selected from Table 7 to calculate the privacy scores of 16 target users. The purpose of this experiment is to explore the impact of different numbers of close users around the target users on privacy disclosure. For convenience of comparison, we normalize the results.

As seen from Fig. 3, the overall trend of the experimental results is consistent. The difference is that in the calculation of Top5, the user's privacy score is relatively high; for Top10, the score is smaller; and for Top15, the score is generally smallest. However, there are special cases, and the differences

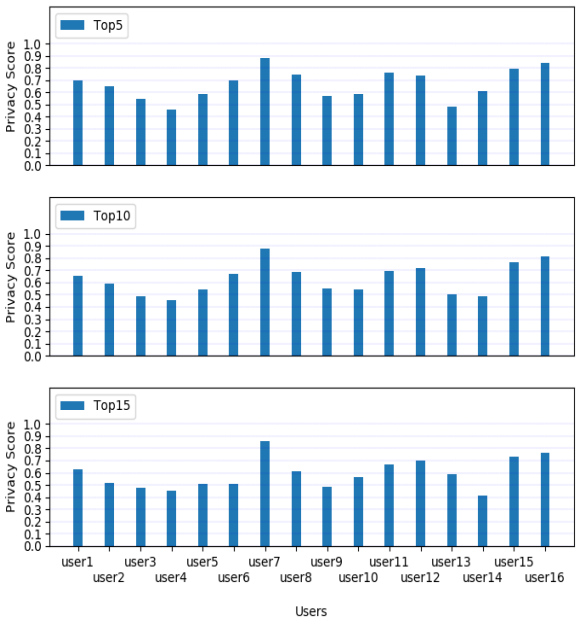

FIGURE 3. Comparison of privacy scores for different numbers of users.

between different users also vary. Compared with Top5, the latter two groups added more users with lower structural similarity and behavioral intimacy, which reduced the average privacy score. This situation also shows that users' privacy leakage has been somewhat reduced because their information is concealed within big data. However, due to the similarity of attributes, it is possible for the measurement scores of Top15 to exceed those of Top5, such as in the case of user10. Through analyzing this specific situation, we find that the middle school, high school, university and graduate school of user10 are all in the city in which he grew up, and the friends who are very close to him are always certain fixed people. Therefore, though the similarity of the selected users decreases, the similarity of attributes is generally high and the difference is not significant.

C. EXPERIMENT 3

In the current research on user privacy measurement in social networks, prior research has not implemented similarity to screen the large number of redundant users on social networking sites, nor has it considered the incorporation of

TABLE 8. User analysis of our algorithms in a larger network.

	Colleague and friends	Middle school classmates	High school schoolmates	Undergraduate classmates	Postgraduate classmates	Kinsfolk	Others
Top5	20	4	8	15	21	3	9
Top10	41	11	17	24	42	5	20
Top15	52	19	24	45	61	8	31
Top100	28	2	8	15	32	2	13

behavior traces into this measurement. Consequently, our work is the first to propose incorporating the timeliness of privacy information. However, in order to compare PMoB with different methods, we chose similar or recognized methods for comparison on Dataset 4. Liu *et al.* [24] were some of the earliest researchers to propose privacy measurement methods in social networks, which indicated a direction for later researchers that was widely recognized. Jain S *et al.* [26] proposed a new fine-grained method for calculating sensitivity and visibility to evaluate users' privacy metrics. Pensa and Di Blasi *et al.* [32] proposed calculating users' privacy metrics in a network structure graph, and they performed experimental analysis in a separate user community; although they did not use real social network scenarios, they also inspired later researchers. Shi *et al.* [36] defined a method of calculating the entropy of a graph structure, which measures the privacy of users in a network environment.

More importantly, detailed calculation methods and specific parameters are given in these studies. To fit the scenario of Ruggero G. Pensa *et al.*, we use the basic community discovery algorithm to divide the network into different communities, and then we calculate the final privacy measure in the target user's community. The final comparison results are as follows.

In Fig. 4, the method proposed by Liu *et al.* considers only the attributes in profiles, so the privacy scores of different target users show little difference. Jain S *et al.* propose a new method for calculating visibility and sensitivity, which does not consider the authenticity of user information and the network environment, so they overestimate users' privacy leakage, and the resulting scores are generally high. Ruggero G. Pensa *et al.* measure smaller communities and do not exclude redundant users, which exaggerates their effect, and the resulting privacy scores are generally low. The entropy value of a network structure proposed by Shi w *et al.* is generally very small in large and complex network structure graphs, because information entropy is a measure of the amount of information needed to eliminate uncertainty. However, with the increasing complexity of the user's environment, the information of the user is introduced in larger quantities, therefore generally resulting in a low entropy value.

In PMoB, we use structural similarity and behavioral intimacy to exclude redundant users, then select users who are closely related to the target users for the attribute similarity calculation. Since privacy measurement is subjective and there is no recognized public dataset, researchers often verify the datasets they collect themselves. Therefore, we cannot show for certain which method is superior or inferior but can reasonably explain only the final result in our datasets. Our approach quantifies a user by considering as many aspects as possible, taking into account aspects not considered in previous studies, including the user's privacy literacy and actual behavior on social networks; and these behaviors can be divided into profile settings, messages posted, interactive behaviors, etc. At the same time, we discovered the timeliness of privacy information and solved the timeliness problem by using behavioral characteristics. Based on the privacy values of different users from each method in Fig. 4, PMoB truly quantifies the difference in privacy between different users by considering the various reasons for the user's privacy leakage.

D. EXPERIMENT 4

In using social networks, users usually have individual needs. Ordinary users want to know about current events, public figures, and information of their interest, even simply for the sake of spending time. Public figure users and official users want to increase their influence through social networks and publish real-time dynamic messages and announcements. Marketing accounts and fake accounts focus on advertising, disseminating fake information and so on. Users have different purposes for privacy protection and can change their personal information according to the privacy score in our proposed methods, ultimately finding the privacy status that meets their expectations and thus controls privacy leakage.

In this part of the experiment, we propose some advice to help users mitigate privacy leakage based on the Sina Weibo platform and PMoB. The first piece of advice aims at a source of attribute similarity information that is used by most social networking sites: access control for profiles. In Sina Weibo, the attribute information in a profile can be selected to be visible to everyone, to the users I follow or only to myself. We recommend that the profile should be as visible as possible to the people the user follows or just visible to the user. If it is mandatory to fill out the profile and make it public in social networks, fake information can be used instead [45]. Using fake information can not only protect personal privacy but also effectively prevent attribute

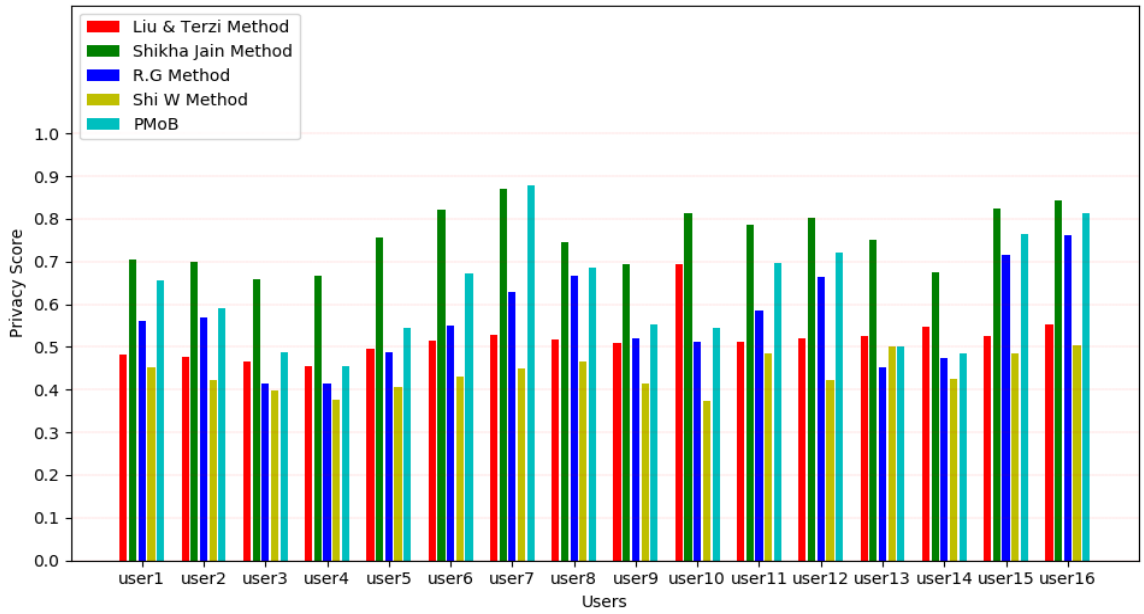

FIGURE 4. Comparison with other privacy measures.

inference and identity linkage. The second piece of advice is to use the function of silent attention. Users will not be listed in another user's follower list if they use silent attention regarding that user. Although they will not appear in the follower list of the other user, they can still receive the status and information sent by the other user. Therefore, this can have a good preventive effect against malicious attacks that use network graph structure analysis.

In terms of behavior, if the user's behavior is restricted, it will affect the normal use of social websites, which is not worthwhile. Therefore, we make some feasible suggestions according to the situations we found in the data collection and experiments. Among the five behaviors of liking, commenting, reposting, @-mentioning and topic creation, data acquisition regarding liking and commenting is more difficult, while the other three actions are completely public. The behavior of @-mentioning can be replaced by private messages in some cases. Users can use private messages to send other private messages. However, the best course of action is still to refrain from talking publicly about private information on social networking sites.

Next, we will simulate a user taking the advice we proposed above and verify whether the user's privacy leakage is effectively curbed. Using the experimental results of Top10 in Experiment 2 as the background, we randomly select five of the users in Top10 to use the silent attention function, and at the same time, we randomly select four of the 11 attributes in the target user's profile and replace them with fake content. The user's behavior still uses the data extracted earlier, and the change of the privacy score after taking these actions is shown in Fig. 5.

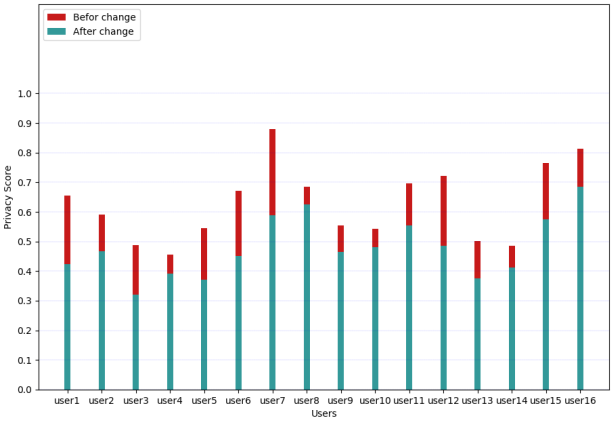

FIGURE 5. Comparison of privacy scores before and after improvement.

As seen from Fig. 5, after using the advice we proposed, the privacy scores are generally reduced, which indicates that the leakage of users' privacy is reduced. However, even if we use the function of silent attention, subsequent low-ranking friends will still expose a certain amount of information and behavior regarding the target users, so users should improve their privacy literacy and control their behavior on social networks. With the help of the privacy score we obtain, users can intuitively understand their privacy disclosure status based on their score, and through continuous adjustment, they can cultivate privacy literacy quickly and effectively.

VI. DISCUSSION

In this paper, we attempted to obtain the privacy scores of users in the whole social network. Initially, we consider the

attribute information of users in their profile and the graph structure information including their friend relationships. However, in the study, only 16 target users have more than 700,000 friends in two hops. If we analyze each user, it will consume a vast amount of resources. Therefore, we try to improve the existing algorithm to obtain a user group with a large correlation with the privacy leakage of the target user to analyze the privacy status of the target user. We use the graph structure relationship between these friends to acquire the structure similarity, and sort them to obtain some users with high similarity. In our dataset, 13 of the 16 target users are graduate students. However, there are only three postgraduate students (80 users in total) in the top 5 highest similarity users among the result of 16 target users. Among the top 10 users with the highest similarity, there are only eight postgraduate students (160 users in total), which is far from our expected results. It is found that this is due to the lack of timeliness. Therefore, we add behavioral characteristics between users. Finally, the number of graduate students in the top 5 of similarity increased to 23, and the number of top 10 increased to 46. This result shows that the timeliness problem we found is real, and the following experiments proved that solving the timeliness problem can greatly improve the accuracy of privacy measurement.

After culminating the above work, we were able to complete the contribution proposed in the Introduction: we found and solved the problem of timeliness, and the experiments show that our proposed PMoB framework can quickly and efficiently eliminate redundant users through the combination of structural similarity and behavioral features, ultimately providing more accurate privacy scores for the target users.

VII. CONCLUSION AND FUTURE WORK

To accurately quantify the privacy of users and reduce the interference of redundant users, we first proposed that we needed to identify the users who hold the private information of the target users, rather than all users around them. In the process of research, we found that the private information held by some user groups extracted by existing algorithms may lose its timeliness and have no privacy value. Users with such information should also be defined as redundant users. To solve this problem, we combined structural similarity and behavioral characteristics to accurately filter the user groups who have the current private information of the target user. Our experimental results demonstrated that this method could effectively filter the user groups that are in the same environment as the target user. These groups are more closely connected to the privacy leakage of the target user.

We had raised the issue of information timeliness in the measurement of privacy in social networks, which provided new ideas for future research work: we should consider not only a wider range of users but also the privacy relevance of users. However, PMoB in this paper still had great limitations: we needed to artificially select behavioral features and attribute features. The calculations of these different features were independent, ignoring the hidden relationships between features, and the calculation steps were tedious. Future research can be carried out to address these problems. We hope to introduce a deep learning framework that has a significant effect on feature extraction to solve these problems and to reduce human interference and the number of tedious steps.

APPENDIX

We take two users A and B as examples to show our privacy measurement process. B has the highest relational degree with A. We list the parameter values of each step in the table.

N_{BA}	N_B	N_l	N_r
32	106	18	6
N_c	$N_@$	d_B	
10	3	376	

$$\gamma_{BA} = \frac{N_{BA}}{N_B}\left(1 + \zeta_l N_l + \zeta_r N_r + \zeta_c N_c + \zeta_@ N_@ + \frac{N_{BA}}{d_B}\right) = 1.3471$$

N_{ABO}	N_A	N_B	N_l	N_r
64	73	106	24	11
N_c	$N_@$	$N_\#$	d_C	δ
5	7	16	10	0. 1

$$\gamma_{AB0} = \frac{N_{ABO}}{N_A + N_B}(1 + \zeta_l N_l + \zeta_r N_r + \zeta_c N_c + \zeta_@ N_@ + \zeta_\# N_\# + \delta\frac{N_{ABO}}{d_O}) = 1.2839$$

d_{AB}	d_A	d_B	N_l	N_r	N_c
18	141	376	2	1	4
$N_@$	$N_\#$	d_C	N_{OAB}	$\tilde{d}_{AB}$	
6	7	10	20	3	

$$\gamma_{OAB} = \frac{d_{AB}}{d_A + d_B}(1 + \zeta_l N_l + \zeta_r N_r + \zeta_c N_c + \zeta_@ N_@ + \zeta_\# N_\# + \frac{N_{OAB}}{\tilde{d}_{AB}}) = 0.3005$$

$S^i(A,B)$	$S^o(A,B)$
0.003551	0.002720

$$R(A,B) = (S^i(A,B) + S^o(A,B))(\gamma_{BA} + \gamma_{ABO} + \gamma_{OAB}) = 0.01838$$

User A and B are consistent in the attribute information in Education, Address, Interests and Relationship Status information.

$F(A,B)$	$R^{'}(A,B)$
0.8773	0.01085

$$G(A,B) = \frac{R(A,B)(1-e^{-F(A,B)})}{R'(A,B)} = 0.9895$$

In our previous research work [41], according to the user's attribute information and its visibility, accessibility, extraction difficulty, privacy awareness, we obtain the privacy score under the individual status. B, C, D, E, and F are the top five users with highest relational degree of A.

$\eta(A)$	$\eta(B)$	$G(A,C)$	$\eta(C)$	$G(A,D)$
1.173	1.016	0.9823	0.792	0.9731
$\eta(D)$	$G(A,E)$	$\eta(E)$	$G(A,F)$	$\eta(F)$
1.247	0.9439	0.866	0.9352	1.592

$$Score(A) = \frac{\eta(A) + \sum_{B \subseteq P_A} G(A,B)\eta(B)}{n+1} = 1.0793$$

REFERENCES

[1] D. Hayes, C. Snow, and S. Altuwayjiri, "A dynamic and static analysis of the uber mobile application from a privacy perspective," *J. Inf. Syst. Appl. Res.*, vol. 11, no. 1, p. 11, 2018.

[2] J. Isaak and M. J. Hanna, "User data privacy: Facebook, Cambridge analytica, and privacy protection," *Computer*, vol. 51, no. 8, pp. 56–59, Aug. 2018.

[3] Z. He, Z. Cai, and J. Yu, "Latent-data privacy preserving with customized data utility for social network data," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 665–673, Jan. 2018.

[4] Z. He, Z. Cai, and X. Wang, "Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks," in *Proc. IEEE 35th Int. Conf. Distrib. Comput. Syst.*, Columbus, OH, USA, Jun./Jul. 2015, pp. 205–214.

[5] M. Qiu, K. Gai, and Z. Xiong, "Privacy-preserving wireless communications using bipartite matching in social big data," *Future Gener. Comput. Syst.*, vol. 87, pp. 772–781, Oct. 2018.

[6] L. Xu, C. Jiang, Y. Chen, J. Wang, and Y. Ren, "A framework for categorizing and applying privacy-preservation techniques in big data mining," *Computer*, vol. 49, no. 2, pp. 54–62, Feb. 2016.

[7] I. F. Lam, K. T. Chen, and L. J. Chen, "Involuntary information leakage in social network services," in *Proc. Int. Workshop Secur.* Berlin, Germany: Springer, 2008 pp. 167–183.

[8] C. Patsakis, A. Zigomitros, A. Papageorgiou, and E. Galván-López, "Distributing privacy policies over multimedia content across multiple online social networks," *Comput. Netw.*, vol. 75, pp. 531–543, Dec. 2014.

[9] *Big Data. A European Survey on the Opportunities and Risks of Data Analytics*, Vodafone Inst. Soc. Commun., Berlin, Germany, 2016.

[10] Y. Wang and R. K. Nepali, "Privacy impact assessment for online social networks," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, Jun. 2015, pp. 370–375.

[11] S. Oukemeni, H. Rifa-Pous, and J. M. Marques Puig, "IPAM: Information privacy assessment metric in microblogging online social networks," *IEEE Access*, vol. 7, pp. 114817–114836, 2019.

[12] K. M. Altenburger and J. Ugander, "Monophily in social networks introduces similarity among friends-of-friends," *Nature Hum. Behav.*, vol. 2, no. 4, pp. 284–290, Apr. 2018.

[13] T. Khazaei, L. Xiao, R. E. Mercer, and A. Khan, "Understanding privacy dichotomy in Twitter," in *Proc. 29th Hypertext Social Media*. New York, NY, USA: ACM, Jul. 2018, pp. 156–164.

[14] S. Oukemeni, H. Rifà-Pous, and J. M. M. Puig, "Privacy analysis on microblogging online social networks: A survey," *ACM Comput. Surveys*, vol. 52, no. 3, Jul. 2019, Art. no. 60.

[15] H. Krasnova, O. Günther, S. Spiekermann, and K. Koroleva, "Privacy concerns and identity in online social networks," *Identity Inf. Soc.*, vol. 2, no. 1, pp. 39–63, Dec. 2009.

[16] M.-R. Ulbricht, "Privacy settings in online social networks as a conflict of interests: Regulating user behavior on Facebook," in *Computational Social Networks*. London, U.K.: Springer, 2012, pp. 115–132.

[17] M. Bartsch and T. Dienlin, "Control your Facebook: An analysis of online privacy literacy," *Comput. Hum. Behav.*, vol. 56, pp. 147–154, Mar. 2016.

[18] T. Dienlin and S. Trepte, "Is the privacy paradox a relic of the past? An in-depth analysis of privacy attitudes and privacy behaviors," *Eur. J. Social Psychol.*, vol. 45, no. 3, pp. 285–297, Apr. 2015.

[19] P. Suárez-Serrato, M. E. Roberts, C. Davis, and F. Menczer, "On the influence of social bots in online protests," in *Proc. Int. Conf. Social Inform.* Cham, Switzerland: Springer, 2016, pp. 269–278.

[20] M. Stella, E. Ferrara, and M. De Domenico, "Bots increase exposure to negative and inflammatory content in online social systems," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 49, pp. 12435–12440, Oct. 2018.

[21] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Proc. Intl. AAAI Conf. Web Social Media (ICWSM)*, 2017, pp. 280–289.

[22] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[23] E. M. Maximilien, T. Grandison, T. Sun, D. Richardson, S. Guo, and K. Liu, "Privacy-as-a-service: Models, algorithms, and results on the Facebook platform," in *Proc. Web 2.0 Secur. Privacy Workshop*, vol. 2, May 2009, pp. 1–4.

[24] K. Liu and E. Terzi, "A framework for computing the privacy scores of users in online social networks," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 1, 2010, Art. no. 6.

[25] L. Fang and K. LeFevre, "Privacy wizards for social networking sites," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, Raleigh, NC, USA, Apr. 2010, pp. 26–30.

[26] S. Jain and S. K. Raghuwanshi, "Fine grained privacy measuring of user's profile over online social network," in *Intelligent Communication and Computational Technologies*. Singapore: Springer, 2018, pp. 371–379.

[27] F. Xu, K. Michael, and X. Chen, "Factors affecting privacy disclosure on social network sites: An integrated model," *Electron. Commerce Res.*, vol. 13, no. 2, pp. 151–168, May 2013.

[28] E. Aghasian, S. Garg, L. Gao, S. Yu, and J. Montgomery, "Scoring users' privacy disclosure across multiple online social networks," *IEEE Access*, vol. 5, pp. 13118–13130, 2017.

[29] R. G. Pensa, G. Di Blasi, and L. Bioglio, "Network-aware privacy risk estimation in online social networks," *Social Netw. Anal. Mining*, vol. 9, no. 1, p. 15, Dec. 2019.

[30] Y. Zeng, Y. Sun, L. Xing, and V. Vokkarane, "Trust-aware privacy evaluation in online social networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 10–14.

[31] Y. Alsarkal, N. Zhang, and H. Xu, "Your privacy is your friend's privacy: Examining interdependent information disclosure on online social networks," in *Proc. 51st Hawaii Int. Conf. Syst. Sci.*, Jan. 2018, pp. 892–901.

[32] R. G. Pensa and G. Di Blasi, "A privacy self-assessment framework for online social networks," *Expert Syst. Appl.*, vol. 86, pp. 18–31, Nov. 2017.

[33] Y. Fan, Y. Zhang, Y. Ye, and X. Li, "Automatic opioid user detection from Twitter: Transductive ensemble built on different meta-graph based similarities over heterogeneous information network," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3357–3363.

[34] L. Yu, S. M. Motipalli, D. Lee, P. Liu, H. Xu, Q. Liu, J. Tan, and B. Luo, "My friend leaks my privacy: Modeling and analyzing privacy in social networks," in *Proc. 23nd ACM Symp. Access Control Models Technol.* New York, NY, USA: ACM, Jun. 2018, pp. 93–104.

[35] R. Serfontein, H. Kruger, and L. Drevin, "Identifying information security risks in a social network using self-organising maps," in *Proc. IFIP World Conf. Inf. Secur. Educ.* Cham, Switzerland: Springer, Jun. 2019, pp. 114–126.

[36] W. Shi, J. Hu, J. Yan, Z. Wu, and L. Lu, "A privacy measurement method using network structure entropy," in *Proc. Int. Conf. Netw. Netw. Appl. (NaNA)*, Oct. 2017, pp. 147–151.

[37] A. Djoudi and G. Pujolle, "Social privacy score through vulnerability contagion process," in *Proc. 5th Conf. Mobile Secure Services (MobiSecServ)*, Mar. 2019, pp. 1–6.

[38] S. Forouzandeh, A. Sheikhahmadi, A. R. Aghdam, and S. Xu, "New centrality measure for nodes based on user social status and behavior on Facebook," *Int. J. Web Inf. Syst.*, vol. 14, no. 2, pp. 158–176, Jun. 2018.

[39] S. Forouzandeh, "Health recommender system in social networks: A case of facebook," *Webology*, vol. 16, no. 1, Jun. 2019, Art. no. 178.
[40] F. Belanger and H. Xu, "The role of information systems research in shaping the future of information privacy," *Inf. Syst. J.*, vol. 25, no. 6, pp. 573–578, Nov. 2015.
[41] X. Li, Y. Yang, Y. Chen, and X. Niu, "A privacy measurement framework for multiple online social networks against social identity linkage," *Appl. Sci.*, vol. 8, no. 10, p. 1790, Oct. 2018.
[42] A. Srivastava and G. Geethakumari, "Measuring privacy leaks in online social networks," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Mysore, India, Aug. 2013, pp. 22–25.
[43] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: ACM, Jul. 2002, pp. 538–543.
[44] K. Lei, Y. Liu, S. Zhong, Y. Liu, K. Xu, Y. Shen, and M. Yang, "Understanding user behavior in Sina Weibo online social network: A community approach," *IEEE Access*, vol. 6, pp. 13302–13316, 2018.
[45] S. Sannon, N. N. Bazarova, and D. Cosley, "Privacy lies: Understanding how, when, and why people lie to protect their privacy in multiple online contexts," in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI)*. New York, NY, USA: ACM, 2018, p. 52.

XUEFENG LI was born in 1989. He received the B.Sc. degree in communication engineering from the Nanyang Institute of Technology, in 2009, and the M.Sc. degree in communication engineering from Henan Polytechnic University, in 2013. He is currently pursuing the Ph.D. degree with the School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include big data security and network security.

YANG XIN was born in 1977. He received the B.Sc. degree in signal and information system and the M.Sc. degree in circuits and systems from Shandong University, in 1999 and 2002, respectively, and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. He is currently an Associate Professor with the School of Cyberspace Security, BUPT. His research interests include big data security, cloud computing security, and network security.

CHENSU ZHAO is currently pursuing the Ph.D. degree in information security with the Beijing University of Posts and Telecommunications. His main research interests include information security, user cross-domain behavior analysis, and network security.

YIXIAN YANG was born in 1961. He received the M.Sc. degree in applied mathematics and the Ph.D. degree in electronics and communication systems from the Beijing University of Posts and Telecommunications, China, in 1986 and 1988, respectively. He is currently the Managing Director of the Information Security Center, Beijing University of Posts and Telecommunications. He has authored more than 40 national and provincial key scientific research projects and contributed to more than 300 high-level articles and 20 monographs. His main research interests include coding and cryptography, information and network security, and signal and information processing. He is a Yangtze River Scholar Program Professor and a National Teaching Master. He is a National Outstanding Youth Fund Winner.

SHOUSHAN LUO received the B.Sc. degree in mathematics from Beijing Normal University, in 1985, and the M.Sc. degree in applied mathematics and the Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1994 and 2001, respectively. He is currently a Professor with the School of Cyberspace Security, BUPT. His research interests include cryptography and information security.

YULING CHEN received the B.S. degree from Taishan University, Taian, China, in 2006, and the M.Sc. degree from Guizhou University, Guiyang, China, in 2009. She is currently an Associate Professor with the Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University. Her current research interests include cryptography and information safety.