2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable
Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)

A Multi-granularity Targeted Covert Collection Scheme for Internet Data

Yi Meng[1,2], Yangyang Li[2◇], Yashen Wang[2], Hao Jin[2],
Chun Wei[1], Xiaoyan Yin[1*]

[1]School of Information Science and Technology, Northwest University, Xi'an, China 710127.
[2]National Engineering Laboratory for Risk Perception and Prevention (RPP), China Academy
of Electronics and Information Technology of CETC, Beijing, China 100041.
◇co-first author *Corresponding author
Email: mengyi@stumail.nwu.edu.cn, liyangyang@cetc.com.cn, yswang@bit.edu.cn,
jh_cetc@163.com, weichun@stumail.nwu.edu.cn, yinxy@nwu.edu.cn

***Abstract*—The targeted covert collection of Internet data, which can not only effectively hide users' traces but also efficiently navigate to target data, plays an important role in three-dimensional social security. Due to big data generated by online social networks, large-scale network analysis usually confront with unbearable computational complexity, resulting in failing to achieve expected results within the specified response time. To retrieve and analyze the most important data within predetermined time, we study the problem of targeted covert collection with multiple granularity levels for Internet data in this paper. To find target data (or web users interchangeably), we use a tree indexes with the largest spanning tree to rank Internet data. Finally, we propose an effective collect algorithm for Internet data acquisition under different granularity levels. We validate our proposed scheme with the real data, which are gathered from the an online social platform. The experiment results verify that our algorithm can find the target data and collect data covertly with different granularity levels.**

I. Introduction

Compared with the real world, the Internet has a strong ability to release, exchange and share information. At the same time, the relationship between users in online social networks is mapping of offline user relationship to a certain extent. Sometimes, it takes only a few minutes for a hot event to spread among tens of thousands of web users. To quickly explore and control the trend of hot events, mass emotion and even the international situation, we can obtain the data of domestic and foreign news media, relevant organization websites and popular online social networks, and conduct online public opinion analysis. However, with hundreds of millions of web users, how to quickly find the most important representative is crucial.

There are many categories of topics on the Internet, and each user is interested in different topics because of the differences of personality, knowledge background and professional habits. To collect and analyze a specific topic, we need to find the set of influential users corresponding to the specific topic. Therefore, we need to capture sensitive target data according to the needs of comprehensive analysis. At present, the security settings of many websites restrict frequent network access. To visit any website freely, we must use trace hiding technology to hide the more concentrated information browsing and searching intention. Sometimes, there is a time limit for public opinion analysis on specific topics. While a detailed analysis should be given if time permits, only an overview can be given if time is limited. Therefore, it is of great significance to collect covetly target Internet data with different granularity.

In line with the reality, a few users have authority on certain topics and have lots of followers. Their influence and information propagation ability are significantly higher than other users. The interaction behaviors of web users includes reposting or comment one another's article, giving a like or dislike for one another's reviews on online social platforms. The type and frequency of interaction between web users are regarded as the strength of relationship among users, thus forming an online social network. Information propagation in online social networks is similar to the spread of diseases. Susceptible/Infectious/Removed (SIR) model and Independent Cascade (IC) model are two traditional models of disease propagation. The node with the strongest propagation ability is the node with the greatest influence. In general, the importance of a node is determined by the topology of the social networks (usage centrality, intermediate centrality, clustering coefficient, etc.).

The existing target user discovery algorithms only consider the user influence determined by the network topology. However, in real life, if users with less influence are more active in certain topics, they are also very important for such topics. In this paper, taking the ranking of users in online social networks and the activity levels of users participating in topics into account, we propose a novel algorithm to collect covertly Internet data by selecting target users. Furthermore, we use a tree indexes with the largest spanning tree to arrange web users in terms of aggregate influence. This structure makes full use of the hierarchical characteristics of the tree. It can preserve the importance ranking among users, on the other hand, it expresses the hierarchical levels of users under different topics. Finally, we retrieve Internet data with different granularity by index of the tree at a given response time.

The main contributions of our paper are summarized as

978-0-7381-3199-3/20/$31.00 ©2020 IEEE
DOI 10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00139

follows:

- We propose a novel aggregate influence analysis framework to effectively extract important information from online social networks. The existing research considers the ranking of individuals in static network structure. For the first time, we try to combine the ranking of users in social networks with the activity levels of users, so as to predict and find the set of influential users participating in specific topics or new topics.
- We propose a hierarchical retrieval algorithm to find the set of target users with multi-granularity within a given response time. The hierarchy of the tree not only preserves the ranking of users, but also reflects the activity levels of users on topic categories. According to the response time limit, users can be retrieved by the index of tree with different granularity.
- We evaluate our proposed algorithm with extensive experiments on real datasets that we crawl by ourselves. The results verify that the proposed algorithm can find target users and collect covertly Internet data effectively and efficiently. This shows that the proposed algorithm can be readily applied to online social networks to retrieve the desired Internet data.

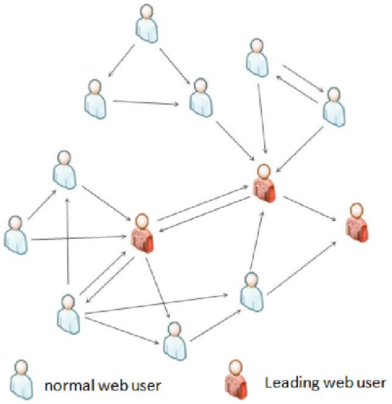

Fig. 1. **Weighted social network** $G(V, E)$**.** Each directed edge in E implies interactions between the two corresponding web users. The web users in red are the leading nodes and have better chances of getting involved in new topics with high probabilities.

The rest of this paper is organized as follows. We introduce our system model in Section 2. The target user discovery algorithm is presented in Section 3. The experimental results on the real dataset is illustrated in Section 4. Section 5 reviews related work, and Section 6 concludes the paper.

II. System Model

We consider a weighted directed graph $G(V, E)$ to model the web users and their relationships of a online social network, where V is the set of web users, E is the set of directed edges between web users. Here, the total number of web users is denoted by N, i.e., $|V| = N$. As shown in Fig. 1, the weight of each directed edge in E, $w_{i,j}$, which implies the potential influence of web user v_i on v_j, is calculated based on the interaction between the two users. According to the interaction rules of online social networks, these following behaviors between two web users should be taken into account when we compute the weight, e.g., reposting or comment one another's article, giving a like or dislike for one another's reviews on online social platforms. Because of the inherent asymmetry interaction between users, we assume that $w_{i,j} \neq w_{j,i}$ and $w_{i,j} + w_{j,i} = 1$, $w_{i,j} \in [0, 1]$, $w_{j,i} \in [0, 1]$. For a leading user v_i and one of his fans v_j, $w_{i,j}$ is infinitely close to 1, while $w_{j,i}$ is approximately equal to 0. Certainly, edges with a weight of 0 are not shown in Fig. 1. However, for two web users with equal influence, $w_{i,j} \approx w_{j,i} \approx 0.5$.

Due to the differences of personality, knowledge background and professional habits, each user is interested in different topics of online social networks. Let C denote the set of topic categories of all articles from online social networks, and c_i as the set of topic categories in which web user v_i is interested, i.e., $C = c_1 \cup c_2 ... \cup c_N$, where $|C|$ represents the total number of topic categories. To characterize the dynamic of users' interests, we use an finite time horizon that is composed of multiple time slots. We assume that the set of topic categories c_i that user v_i is interested in will not change significantly in a short period of time. Therefore, we use a sliding window $t = [t^s, t^f]$ to represent the time range in which the set of topic categories that the user is interested in remains unchanged, where t^s and t^f are the starting time and the ending time, respectively, $t^s \in [0, T]$, $t^f \in [0, T]$. In other words, for user v_i, $|c_i|$ is a constant during $[t_i^s, t_i^f]$.

Based on the historical data of the topics concerned by users, how to find the target data (i.e., target web users) involved in new topics is facing great challenges. Finding target web users is essentially equivalent to predicting which users will participate in new specific topics. In general, the categories of topics that users are interested in are very limited. The more topics they participate in, the more active the users are, and the more likely they are to participate in new topics. Therefore, when we predict the possibility of users participating in new topics, we take the historical information of users' activity in old topics into account. The total number of topic categories is expressed as m, i.e., $C = \{g_1, g_2, ..., g_m\}, |C| = m$, where g_i refers to a specific topic category. we create a one-dimensional vector s, $s = \{s_1, s_2, ..., s_m\}, |s| = m$, to characterize the intensity of specific users participating in each topic category. For user v_i, s is computed as follows: (1) obtain all articles published online by v_i; (2) classify articles by topic category; (3) assign a value to s_i, $i = 1, 2..., m$, s_i is equal to the number of articles in the specific topic category. Finally, we construct an activity representation matrix $\mathbf{A}$, whose element a_{ij} denotes the activity level of user v_i on the topic g_j.

The calculation of user activity matrix $\mathbf{A}$ is illustrated in Fig. 2. Firstly, based on the given three topic categories g_1, g_2

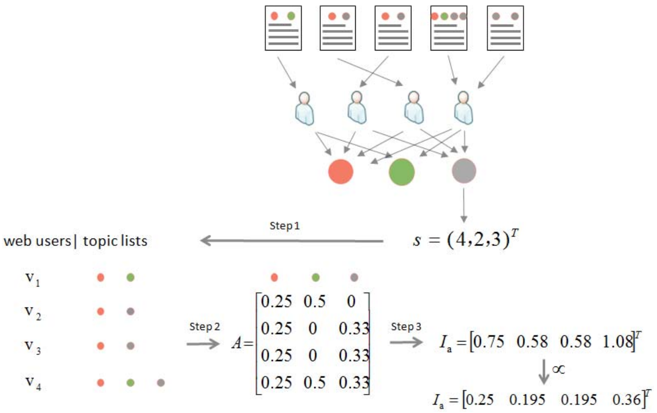
Step1
web users| topic lists
$s = (4,2,3)^T$
v_1
v_2
v_3
v_4
Step2
$A=\begin{bmatrix}0.25 & 0.5 & 0\\ 0.25 & 0 & 0.33\\ 0.25 & 0 & 0.33\\ 0.25 & 0.5 & 0.33\end{bmatrix}$
Step3
$I_a = [0.75 \quad 0.58 \quad 0.58 \quad 1.08]^T$
$\propto$
$I_a = [0.25 \quad 0.195 \quad 0.195 \quad 0.36]^T$

Fig. 2. **A toy example of the activity level calculation.** The online social network consists of four users, five articles published and three topic categories. Firstly, the activity level of each user participating in these three topic categories is computed. Secondly, the matrix A is obtained. Finally, the normalized activity levels for all users are achieved.

and g_3, as well as the articles published online by four users, we calculate the activity level of each user participating in these three topic categories, i.e., to obtain the corresponding vector s for each user. Secondly, by combining the user activity levels of four users, the matrix A is obtained. We assume that all users participating in the same topic category share equally the credit. Therefore, all four users have participated in g_1, $a_{i1} = 0.25$, $i = 1, 2, 3, 4$. Similarly, user v_1 and v_4 have participated in g_2, $a_{12} = a_{42} = 0.5$. Finally, the activity level of user v_i is denoted by $I_a(v_i|C)$, where $I_a(v_i|C) = \sum_j a_{ij}$.

In reality, the number of participants in hot topics can reach tens of millions, while the number of participants in unpopular topics is only a single digit. There is a big difference in the number of audience of topics. To measure this difference reasonably, we take the number of participants in topics into consideration. The activity level of user v_i in topic c_i can be recalculated as

$$I_a(v_i|c_i) = \frac{\sum_m I_a(v_i|C)\mu_m}{\bar{\mu}} \tag{1}$$

where μ_m is the number of messages under topic g_m, $\bar{\mu}$ is the average of all messages for all topics.

To rank the activity level for all users, we normalized $I_a(v_i|c_i)$ by

$$I_a(v_i|c_i) = \frac{I_a(v_i|c_i)}{\sum_{i=1}^{N} I_a(v_i|c_i)} \tag{2}$$

The key notations are summarized in Table 1.

III. Target web user discovery in a social network

To a certain extent, activity levels of users can indicate the possibility of users to participate in specific topics, especially new topics. However, the influences of users, which are determined by the topology of online social networks, is also crucial, and should be taken into account when we search for the target web users. Therefore, we calculate the aggregate influences of web users based on activity levels and ranks of users, then propose a multi-granularity target web user discovery algorithm with help of a tree structure, and an optimal solution is achieved by retrieving iteratively the data in a top-down manner by layer to find the most influential users finally.

TABLE I
Key notations

Notation	Definition
$G(V, E)$	a weight directed graph as an online social network.
V	the set of web users.
E	the set of directed edges.
N	the number of web users.
$w_{i,j}$	the potential influence of v_i on v_j.
C	the set of topic categories.
c_i	the set of topic categories in which v_i is interested.
T	the upper bound of the time slot.
s	user participation in the topics.
μ_m	the number of messages under topic g_m.
$\bar{\mu}$	the average of messages for all topics.
$I_a(v_i\|c_i)$	activity level of v_i in c_i.
$I_p(v_i\|c_i)$	the PageRank value of v_i.
$I(v_i\|c_i)$	the aggregate influence of v_i.
H_j^{in}	the set of in-degree neighbours of v_i.
H_j^{out}	the set of out-degree neighbours of v_i.
d	the damping coefficient.
$PR_{i,\gamma}$	the PageRank value of individual v_i at iteration γ.
β	the activity intensity coefficient.

A. Rank calculation of web users

The PageRank algorithm evaluates the importance of web pages according to the topological characteristics of the web graphs. Since the equivalence between the users of online social networks and the web pages of web graphs, the PageRank algorithm can be used to calculate the influence of users in online social networks based on the system model introduced in the previous section. Let $PR_{i,\gamma}$ denote the rank of web user v_i at iteration γ, H_j^{out} and H_i^{in} denote the set of out-degree neighbours of user v_j and the set of in-degree neighbours of user v_i, respectively. For each iteration, the $PR_{i,\gamma}$ can be update as follows:

$$PR_{i,\gamma+1} = d \cdot \sum_{v_j \in H_i^{in}} w_{i,j} \frac{PR_{j,\gamma}}{|H_j^{out}|} + \frac{1-d}{N}, \forall v_i \in V \tag{3}$$

where $w_{i,j}$ is the potential influence of v_i on v_j, N is the number of web users, and $d \in [0, 1]$ is the damping coefficient to prevent the ranks increasing indefinitely. Let $I_p(v_i)$ denote the rank of web user v_i, $I_p(v_i)$ can be obtain from the result of the final convergence of PageRank algorithm.

B. The aggregate influence calculation of web users

Do two users with the same ranks have the same activity level? We have known that the rank is decided by the topologies of web graphs, while the activity level is determined by the users' interest. Obviously, web users with high ranks and high activity levels are more likely to participate in specific topics or new topics. Therefore, such users are the target users we are looking for. Combing ranks and activity levels together, the aggregate influence of web users can be computed as follows:

$$I(v_i|c_i) = F(v_i) = \beta I_a(v_i|c_i) + (1-\beta) I_p(v_i) \qquad (4)$$

where $I_a(v_i|c_i)$ is the activity level of web user v_i, $I_p(v_i)$ is the rank of web user v_i, and β is a coefficient, $0 < \beta < 1$.

C. Aggregate influence-based model for online social networks

For an online social network with a large number of users, it is facing challenges to find the most influential users within a given response time for new topics, so as to obtain targeted information. The new topics in online social networks are constantly emerging, and large-scale network analysis usually confront with unbearable computational complexity, resulting in failing to achieve expected results within the specified response time. Without historical information, a comprehensive analysis of new topics is like looking for a needle out of the sea. Therefore, we propose a tree indexes with the largest spanning tree to rank web users based on the aggregate influence.

The indexes model consists of a set of nodes (leaf nodes and non-leaf nodes), in which each node is arranged according to users' aggregate influences in online social networks. For non-leaf node v in layer l, we have the following constrain:

$$I^l(v|c_v) \geq \max_{v_o \in \{\text{ the set of nodes in layer } l+1\}} I^{l+1}(v_o|c_{v_o}) \qquad (5)$$

where v is any node in layer l, v_o is any node in layer $l+1$, and $I^l(v|c_v)$ represents the aggregate influence of individual v of layer l. Here, c_v and c_{v_o} is the set of topics for node v and node v_o, respectively. Thus, Eq. (5) implies that the aggregate influence of node v in layer l is always greater than or equal to any node in layer $l+1$. As mentioned earlier, we assume that the set of topic categories in which users are interested remains unchanged over a fixed time window. In this way, the accuracy and time efficiency of the retrieval algorithm can be guaranteed even in the case of a large number of web users. Due to the particularity of the tree structure, searching for the target user corresponding to a specific topic or new topics is transformed into a breadth first traversal process starting from the root node.

We adopt the Prim algorithm to generate the tree indexes structure by selecting the node with the biggeset aggregate influence each time. Specifically, the indexes structure of each web user in V is determined jointly by their ranks and activity levels in a social network.

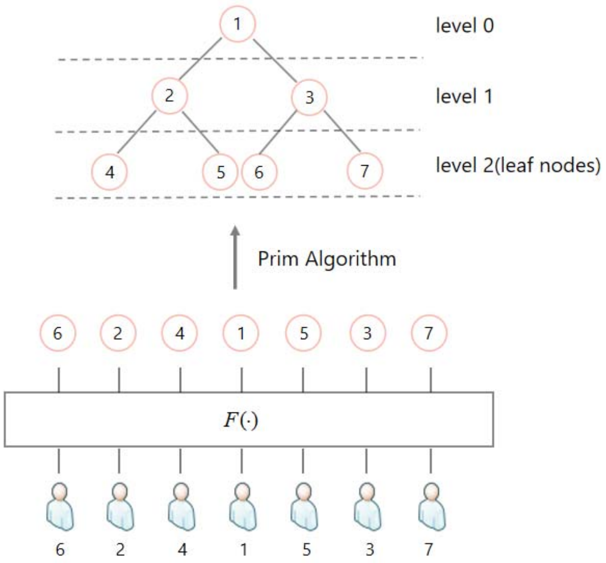

Fig. 3. The tree indexes architecture. The nodes (web users) are arranged based on the aggregate influences, and aggregate influences decline from the root node to leaf node.

D. Target user discovery Algorithm for an online social network

Our goal is to find the top-k influential web users in an online social network for given topics. According to the proposed tree-based indexes model, we introduce the hierarchical target user discovery algorithm, as shown in Algorithm 1.

Algorithm 1 Target User Discovery Algorithm

```
Input: The topology of a social network $G(V,E)$, The desired number of users $k$, The maximum time slot $T$, the set of topics $C$.
Output: The set of nodes $\mathcal{Q}$ and the tree $B$.
1:  Initialization $\mathcal{Q} = \emptyset$.
2:  for t=1,2,...,T do
3:      Calculate the activity level of user $v$ under different topics based on Eqs. (1) and (2).
4:      Compute the rank of user $v$ according to the topology of the social network.
5:      Calculate the aggregate influence of user $v$ based on the rank and the activity level of user $v$.
6:      Arrange nodes and build the tree $B$.
7:      for $l$=0,1,2,... do
8:          if Candidate set $|\mathcal{Q}| \neq k$ then
9:              Put all nodes from left to right in layer $l$ into $\mathcal{Q}$ until $|\mathcal{Q}| = k$.
10:     return $\mathcal{Q}$ and the tree $B$.
```

Our target user discovery algorithm first calculates the user's activity level(line 3), then calculate the user's rank by the PageRank algorithm based on the topology (line 4), and finally calculates the aggregate influence of the user (line 5). According to the aggregate influence, nodes are arranged to build the tree. Based on the given target number k and the set of target topics, all the target users are found.

As mentioned earlier, the set of topic categories in which one certain web user is interested in remains unchanged within a certain time window. Therefore, compared with other schemes, our proposed algorithm can find target users effectively and efficiently. Moreover, with the change of the number of targets and the required time, our algorithm can always provide high precision prediction.

IV. Evaluations

In this section, we evaluate the performance of our proposed algorithm with the real-world dataset that we crawl by ourselves. Meanwhile, compared with other existing algorithms, our proposed algorithm can find the target users effectively and efficiently. Furthermore, according to the given number of targets and response time, our method can obtain multi granularity target data.

A. Experiment Setup and Dataset

The experiments in this section uses datasets crawled from Microblog and Twitter. The datasets generated by crawling from these two platforms have similar properties. Microblog and Twitter are two of the largest online social platforms in the world. These two social platforms have the explosive network scale and involve hundreds of millions of web users. However, our computing and processing resources are limited. Therefore, we adopt the topic-based approach to crawl data. Accordingly, we select a few hot topics and crawl the users who participate in these topics. By deleting some isolated or unrepresentative web users, we retain about 2500 users and 100000 Blogs users have published, and build an online social network based on users' click, track, comment, forward and other interactive behaviors. The PageRank algorithm is used to calculate the user's rank based on the topology, and the historical data of the user's participation in topics is used to compute the activity levels for users. Finally, the aggregate influences for all users are achieved.

Due to different personality, professional background and life experience, users are interested in different topic sets. Furthermore, the topics can be divided into hot topics and unpopular topics. Therefore, it is necessary to differentiate topic categories for all the historical information of each user, and know about the topics or public opinions that users are interested in, so as to calculate the activity levels of users participating in topic categories. First of all, we extract the weight of each word in the Blogs, and then extract the weight of each word. Furthermore, through the threshold obtained by many experiments, the set of keywords of Blogs is extracted, and then the feature vector of each Blog is generated. Finally, we adopt a deep learning model based on the natural language processing (NLP) algorithm. By inputting the eigenvector of each Blog, the topic category of the Blog is obtained.

We believe that when a user is highly active on multiple topic categories and has high influence in the network topology, it is very likely that the user will have high influence under a new topic. Moreover, as mentioned above, the activity levels and ranks of users will not change significantly in a short period of time. Therefore, the target user discovery algorithm based on aggregate influence can achieve better performance.

B. The aggregate influence analysis

The relationship between three properties (number of followers, number of topics as well as number of Blogs) and aggregate influence. To verify effectiveness of our proposed analysis framework on online social network platforms, we test the relationship between aggregation influence and several parameters. As shown in Fig. 4, the figure contains three curves, which in turn show the relationship between the aggregate influence and the number of topics users participate in, the number of Blogs published, as well as the number of followers. The four group on the x-axis also have different quantitative intervals. For the number of topics, the four values on the x-axis represent the number of topics in the range of [0, 4], [5, 9], [9, 13], [13, $+\infty$]. For the number of Blogs, the four values on the x-axis represent the number of Blogs published by users in [0, 25], [25, 50], [50, 75], [75, $+\infty$]. For the number of followers, the four values on the x-axis represent the number of followers of the specific user in [0, 100], [100, 1000], [1000, 10000], [10000, $+\infty$]. Compared with the three lines in Fig. 4, we can know that with the increase of the number of topics, the number of Blogs published and the number of followers, the aggregate influence of users is significantly improved, which is highly consistent with our intuitive judgement. In addition, when the number of followers and the number of Blogs published are small, the aggregation influence is more affected by the number of Blogs published. We can also know that users who publish frequently are more important in the context of new topic content when the aggregate influence is small. At the same time, with the increase of the number of followers, the aggregation influence is more affected by the number of Blogs published.

The relationship between aggregate influence and two metrics (ranks and activity levels of users). In our target user discovery algorithm, tree is the key component. It can not only be used as the index of retrieval, but also can be used to achieve topic-based Internet data retrieval with different granularity levels. Fig. 5 illustrates the relationship between ranks as well as activity levels of users and aggregate influence at different layers in the tree model. All these three parameters decrease with the increase of the number of layers. For users at the first layer, their ranks are at the top of the list, and the corresponding aggregation influences are large. Surprisingly, even though users are ranked at the bottom of the list, thanks to their high activity levels, the corresponding aggregation influences can not be ignored. The proposed algorithm takes the activity levels of users into account, and can fully explore the users who contribute to specific topics. Therefore, our

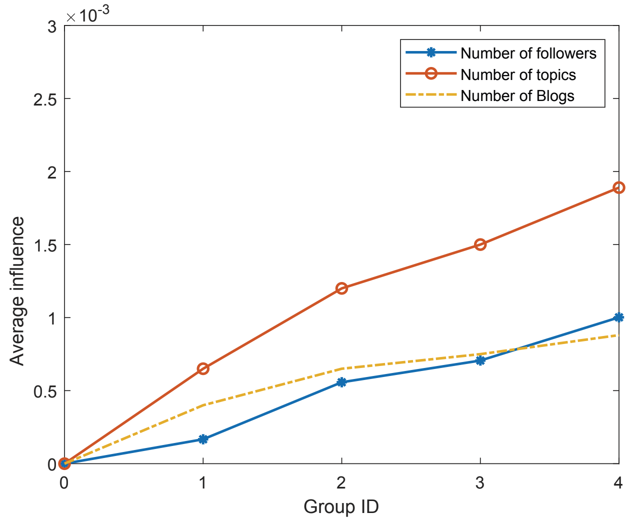

Fig. 4. **The aggregate influence comparison.**

proposed algorithm can find the target users effectively and efficiently.

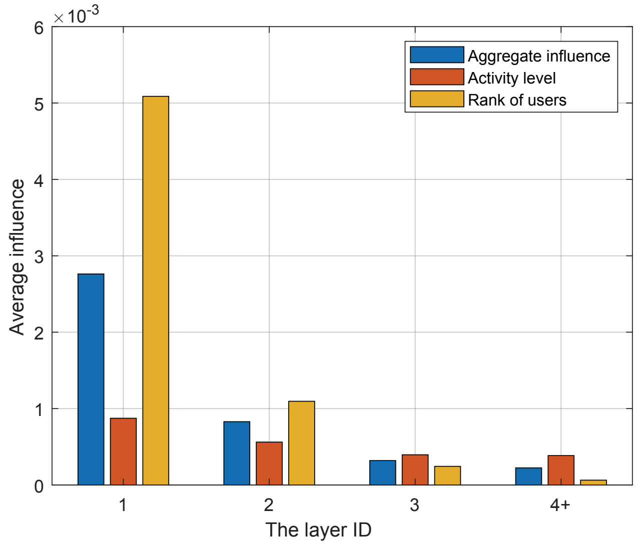

Fig. 5. **The relationship between ranks, activity levels and the aggregate influence.**

The relationship between topics and aggregate influence. In online social networks, the value of data will change dynamically with the change of social environment and time. The importance of users is not only related to the network topology, but also closely related to the set of topics they participate in. We have noticed that when a new topic category is generated, compared with users who participate in fewer topics, users who participate in more topics are more likely to participate in the new topic, and gain higher influence on such topics. Fig. 6 shows the relationship between the number of topics participated by different users and aggregate influence. As the number of topics increases, ranks, activity levels of uses and aggregate influence increase accordingly. Due to the big proportion of user activity in aggregate influence, even users with low rank in social networks will have higher aggregate influence if their activity levels are high. Therefore, topic plays an important role in the retrieval of important user information. This proves the correctness of our proposed algorithm again.

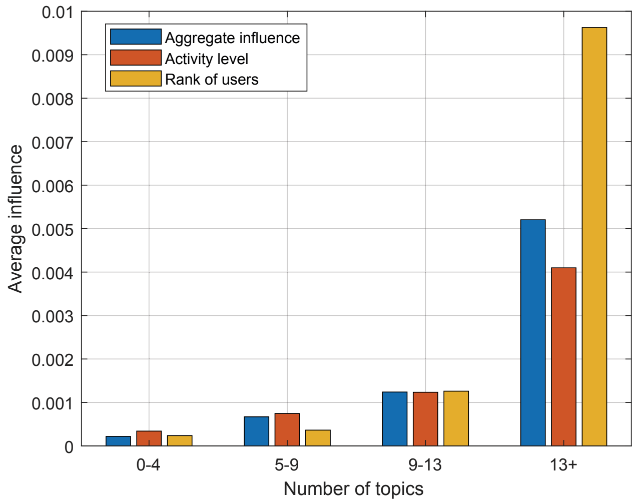

Fig. 6. **The relationship between topics and aggregate influence.**

V. Related Work

A. Target user discovery

Katz et al. [1] found that a fraction of individuals (opinion leaders) have an obvious guiding effect on the behavior of others for the first time. Since then, target user discovery has always been a hot topic, which aim to find the minimum number of users to maximize the influence coverage in social networks. In [2], Kempe et al. proved that the target user discovery problem is a NP-hard problem. Subsequently, more and more scholars evaluated the importance of users by analyzing the static statistics of the topology of social networks. Bonacich et al. [3] used the degree of centrality i.e., the number of direct neighbors of nodes in a network, to measure the influences of users. The two most classic algorithms based on link topology are PageRank algorithm [5] and HITS algorithm [6]. Several algorithms based on these two kinds of algorithms have been proposed. Duan et al. [4] proposed the ClusterRank algorithm, which used the degree and clustering coefficient of network nodes simultaneously to assess the importance of users. Weng et al. [7] proposed an algorithm, TwitterRank, which used the history records to infer the user's interest. The higher the interest similarity between users is, the faster the information spread between users. Lu et al. [8] proposed LeaderRank algorithm, which added a common node and two-way links with other nodes from a strongly connected graph to avoid the incorrect ranking. In recent years, signed social network has attracted the attention of scholars. Both positive influence and negative influence of the social network were considered in [9], [10], [11]. Compared with unsigned social networks, the signed social networks can provide a more detailed description of the relationship of users.

However, these existing schemes are based on the static network status in terms of network topology or user relationship, and can not meet the needs of dynamic public opinion analysis within a certain time limit.

B. Information propagation and traceability

In online social networks, the process of information spread is highly similar to the diseases propagation in biology. Furthermore, the representative model, i.e., Susceptible/Infectious/Removed (SIR) model and Independent Cascade (IC) model are often used to model information spread. Based on SIR model, Skaza et al. [12] studied the propagation model of twitter's topic tags and classify them. Moskowitz et al. [13] proposed the centrality effect of SIR propagation model. Advertising recommendations problem was discussed in [14], [15]. With the continuous research, some scholars found that there is Matthew effect in online social networks, because of the common interest, many users form a community. Thanks to the existence of the community, users interact more frequently with friends in the same community. Hao et al. [16] proposed a community detection algorithm based on the formal context, an improved version of the algorithm was proposed in [17], so as to satisfy the requirements of signed social network. Ding et al. [18] proposed a CSIR-based model based on community structure. Jain et al. [19] introduced the information propagation model based on the node of community center.

At the same time, as the reverse process of information dissemination, information traceability aims to find the source of information dissemination. Prakash et al. [20] proposed a minimum description length localization method based on the SI model. Zhu et al. [21] adopted the sample path method, which find the source node that is most likely to be the sample path in the snapshot as the traceability node according to the network snapshot. However, it is difficult to obtain the propagation status of all nodes on the large-scale online social networks. Thus, Pinto et al. [22] selected a part of nodes as the observation points in the network topology, recorded the propagation status of these nodes and used methods such as maximum likelihood estimation to find the source.

VI. Conclusion

In this paper, we study how to find the target users, so as to effectively collect Internet data covertly. We notice that the traditional user importance evaluation algorithms do not consider the users' activity levels to participate in topics. Therefore, taking ranks and activity levels of users, we propose a method to measure the possibility that users have high influence under new topics, and use tree index to reflect the hierarchical structure of users' aggregation influences. By selecting target users from top to bottom, data collection with different granularity can be realized. Our extensive experimental results have proved that the number of topics in which users are interested is an important factor for users' aggregate influence in online social networks. At the same time, experimental results verify the effectiveness and accuracy of our proposed algorithm.

VII. Acknowledgment

This research is supported in part by National Natural Science Foundation of China under Grant 61872295, the New Generation of Artificial Intelligence Special Action Project under Grant AI20191125008 and Shaanxi Natural Science Foundation under Grant 2020JM-416.

References

[1] E. Katz and P. F. Lazarsfeld, "Personal Influence: The Part Played By People in the Flow of Mass Communication," *The Canadian Journal of Economics and Political ence*, vol. 21, no. 6, 1957.

[2] D. Kempe, J. Kleinberg, and v. Tardos, "Maximizing the Spread of Influence through a Social Network," p. 137, 2003.

[3] P. F. Bonacich, "Factoring and Weighting Approaches to Status Scores and Clique Identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[4] C. Duan-Bing, G. Hui, L. Linyuan, Z. Tao, and P. Matjaz, "Identifying Influential Nodes in Large-Scale Directed Networks: The Role of Clustering," *Plos One*, vol. 8, no. 10, p. e77455, 2013.

[5] L. Page, "The PageRank Citation Ranking: Bringing Order to the Web," *http://google.stanford.edu/?backrub/pageranksub.ps*, 1998.

[6] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," in *proc. of ACN-SIAM Symposium on Discrete Algorithms, 1998*, 1998.

[7] J. Weng, E. P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding Topic-Sensitive Influential Twitterers," in *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, 2010.

[8] L. Lu, Q. Li, D. Chen, and T. Zhou, "Identifying Influential Spreaders by Weighted LeaderRank," *Physica, Statistical mechanics & its applications*, 2014.

[9] S. Chen and K. He, "Influence Maximization on Signed Social Networks with Integrated PageRank," 12 2015, pp. 289–292.

[10] D. Li, C. Wang, S. Zhang, G. Zhou, D. Chu, and C. Wu, "Positive Influence Maximization in Signed Social Networks Based on Simulated Annealing," *Neurocomputing*, vol. 260, no. oct.18, pp. 69–78, 2017.

[11] Y. Li, W. Chen, Y. Wang, and Z. L. Zhang, "Influence Diffusion Dynamics and Influence Maximization in Social Networks with Friend and Foe Relationships," p. 657, 2013.

[12] J. Skaza and B. Blais, "Modeling the Infectiousness of Twitter Hashtags," *Physica A Statal Mechanics & Its Applications*, vol. 465, no. Complete, pp. 289–296, 2017.

[13] I. S. Moskowitz, P. Hyden, and S. Russell, "Network Topology and Mean Infection Times," *Social Network Analysis and Mining*, vol. 6, no. 1, pp. 1–14, 2016.

[14] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model," in *IEEE International Conference on Data Mining*, 2012.

[15] S. Bharathi, D. Kempe, and M. Salek, *Competitive Influence Maximization in Social Networks*. Springer Berlin Heidelberg, 2007.

[16] F. Hao, S. S. Yau, G. Min, and L. T. Yang, "Detecting k-Balanced Trusted Cliques in Signed Social Networks," *IEEE Internet Computing*, vol. 18, no. 2, pp. 24–31, 2014.

[17] F. Hao, G. Min, Z. Pei, D. S. Park, and L. T. Yang, "K-Clique Community Detection in Social Networks Based on Formal Concept Analysis," *IEEE Systems Journal*, vol. 11, no. 1, pp. 1–10, 2015.

[18] F. Ding, B. He, C. Li, and L. Lan, "An Improved CSIR Information Propagation Model in Social Networks," pp. 312–315, 2017.

[19] S. Jain, G. Mohan, and A. Sinha, "Network Diffusion for Information Propagation in Online Social Communities," pp. 1–3, 2017.

[20] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Spotting Culprits in Epidemics: How Many and Which Ones?" in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 2012.

[21] K. Zhu and L. Ying, “Information Source Detection in the SIR Model: A Sample Path Based Approach,” *IEEE/ACM Transactions on Networking*, vol. 24, no. 1, pp. 408–421, 2016.

[22] P. Pinto, P. Thiran, and M. Vetterli, “Locating the Source of Diffusion in Large-Scale Networks,” *Physical Review Letters*, vol. 109, 08 2012.