electronics

MDPI

Article

Intelligent Prediction of Private Information Diffusion in Social Networks

Yangyang Li [1,*], Hao Jin [1], Xiangyi Yu [1], Haiyong Xie [1,2], Yabin Xu [3], Huajun Xu [3] and Huacheng Zeng [4]

[1] National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data, China Academy of Electronics and Information Technology, Beijing 100041, China

[2] School of Cyber Security, University of Science and Technology of China, Hefei 230052, China

[3] School of Computer, Beijing Information Science & Technology University, Beijing 100041, China

[4] Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292, USA

* Correspondence: liyangyang@cetc.com.cn

Received: 13 March 2020; Accepted: 23 April 2020; Published: 27 April 2020

Abstract: In the information age, leaked private information may cause significant physical and mental harm to the relevant parties, leading to a negative social impact. In order to effectively evaluate the impact of such information leakage in today's social networks, it is necessary to accurately predict the scope and depth of private information diffusion. By doing so, it would be feasible to prevent and control the improper spread and diffusion of private information. In this paper, we propose an intelligent prediction method for private information diffusion in social networks based on comprehensive data analysis. We choose Sina Weibo, one of the most prominent social networks in China, to study. Firstly, a prediction model of message forwarding behavior is established by analyzing the characteristic factors that influence the forwarding behavior of the micro-blog users. Then the influence of users is calculated based on the interaction time and topological structure of users relationship, and the diffusion critical paths are identified. Finally, through the user forwarding probability transmission, we determine the micro-blog diffusion cut-off conditions. The simulation results on Sina Weibo data set show that the prediction accuracy is 86.9%, which indicates that our method is efficient to predict the message diffusion in real-world social networks.

Keywords: social network; micro-blog; privacy information; diffusion forecast; key paths

1. Introduction

One key aspect about today's social network, such as Sina Weibo, is that they provide a major number of people with a wide range of public opinion space conveniently, users are increasingly accustomed to express their opinions and share information on these network [1]. However, information published and spread in social network can be as negative as it can be positive, as the online information is always coming with people's privacy. The spread of private information is different from the spread of other information. Its spreading speed and scope are far greater than that of ordinary information. Once the privacy information is widely spread, it may cause physical and mental harm to the privacy leakage related person, even resulting in a negative social impact. As a result, private information diffusion prediction in social networks has become an important issue to be concentrated on.

At present, most researches on information dissemination in social networks work with public information, there are not many literature studies on privacy information diffusion [2], and there are even less studies on privacy information diffusion in social networks. In terms of research on the influence factors of information diffusion, a social network analysis method is proposed to study how the relationship between users and their neighbors affected the spread time of their private information [3]. For users with different privacy concerns, the time rules of their privacy information diffusion in social networks are different. The scope and depth of information diffusion are defined to evaluate the effect of information diffusion. Based on threshold and forwarding probability, several approaches [4,5] are proposed to predict the forwarding behavior of users in Sina Weibo. These methods use large scale recursive loop of nodes in social networks to generate the prediction results, which causes high cost in the calculation of the algorithms.

Thanks to the advancement of big data analytics [6] and machine learning technology [7], a large number of studies have been carried out to analyze the influence of nodes in social networks. Graph theory based method is used to evaluate the importance of nodes in social networks, the importance of each node is denoted by the numbers of its connections [8]. In addition, more and more factors are taken into consideration to measure the influence of the node in social networks, such as the average distance of information diffusion [9]. An influence rate concept is proposed to describe the users' influence according to activeness, dissemination degrees and number of fans [10]. It would be beneficial if we could take key uses and critical paths into account to predict the private information diffusion.

In view of this, this paper proposes a method to predict the spread of private information by considering the users' privacy concern as a factor. In addition, the key forwarding nodes and crucial paths are calculated to make the prediction more efficient. The remainder of the paper is organized as follows. In Section 2, related works are described. The prediction model of user forwarding behavior is explained in Section 3. In Section 4, the privacy diffusion prediction model is formulated. The experiments are elaborated in Section 5. Finally, the conclusion of this paper is presented in Section 6.

2. Related Work

Thus far, most information diffusion predictions are based on infectious disease models [11–13], or classification models [14–16].

Due to the similarity between information diffusion and the spread of infectious diseases, information can be regarded as infectious diseases, and the process of information diffusion is equivalent to the process of disease infection. Therefore, many researchers directly apply SIR model [11,12] or SIS model [13] to model and predict the spread of information. However, these methods are only verified on the simulation data with uniform probability of forwarding behavior of each user, which is obviously unreasonable in real social networks.

Classification models are mainly used to analyze the key factors (e.g., forwarding behavior) that affect information diffusion. Based on the analysis of user group characteristics [14], it is pointed out that the forwarding behavior of users is not only related to the interest points of micro-blog content, but also related to the number of fans, attentions and micro-blogs. The forwarding behavior of micro-blog users is predicted based on the Bayesian prediction model. Similarly, the importance of different features in forwarding behavior is analyzed [15]. According to the weights of features, a feature-weighted forward prediction method is proposed to predict the forwarding behavior in micro-blogs. Based on user attributes, social relations and micro-blog contents, a three feature integration method [16] is proposed to predict users' retweet behavior by classification algorithms. Nevertheless, These studies just predict the forwarding of general micro-blog information, not the forwarding of private information. In addition, those studies do not delve deep into the trend prediction issues such as the range and depth of information diffusion in spreading.

In a nutshell, although the existing literature uses several features to predict user forwarding behavior, they do not consider user privacy concerns. Hence, it is not suitable for the prediction of private information forwarding. Moreover, the existing information diffusion prediction model predicts the user's large-scale forwarding behavior layer by layer based on the fan path, and then predicts the information diffusion. However, in the actual information diffusion process, there are key users and critical paths that can bring a lot of forwarding. As long as we can find out these key users and critical paths, it will be of great significance to predict the spread of private micro-blog. Therefore, we need to design a new prediction method of user forwarding behavior and privacy information diffusion.

3. Prediction of User Forwarding Behavior

3.1. Analysis of Influencing Factors of Forwarding

To predict the forwarding behavior of users, we first study and determine the factors that affect the forwarding behavior of users, and establish the prediction model of the forwarding behavior of users. After that, we find there are four factors that affect the forwarding of private information.

3.1.1. Weight of Privacy Information Entropy p_{ew}

According to the information theory, the diffusion of information indirectly reflects the value of information. Without introducing any additional value, the information that spreads more widely and spreads for longer is the information with greater value. It can be seen that the greater the privacy content of information, the easier it is to be widely spread. Therefore, information entropy, which reflects the content of privacy information, determines whether the user forwards the message to a certain extent.

If the probability of an event is $p(x)$, its information entropy calculation formula as shown in Equation (1):

$$H(x) = -\sum_{x \in X} p(x) \log p(x) \tag{1}$$

Here, we cite the four privacy categories given in reference [17], namely health category (0), location category (1), emotion category (2) and social category (3). In this paper, the user's privacy information is represented as $R_{m\times n}(r_{ij})$ matrix [18], as shown in Table 1.

Table 1. Privacy information matrix.

Privacy Information	Class 0	Class 1	Class 2	Class 3
privacy information 1	p_{10}	p_{11}	p_{12}	p_{13}
privacy information 2	p_{20}	p_{21}	p_{22}	p_{23}
......	......	......	......	......

The calculation steps of entropy weight value of privacy information of users are as follows:

1. The probability that the *i*th private information belongs to category *j* is:

$$p_{ij} = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}} \tag{2}$$

2. The entropy value of class *j* is:

$$e_j = -k \sum_{i=1}^{m} p_{ij} \cdot \ln(p_{ij}) \tag{3}$$

where, $k = \frac{1}{\ln m}$.

3. The entropy weight of class *j* is:

$$w_j = \frac{1 - e_j}{\sum_{j=1}^{n}(1 - e_j)} \tag{4}$$

Thus, the weight of users' privacy information entropy can be obtained:

$$p_{ew} = I_j * w_j \tag{5}$$

When $j \in (0,1,2,3)$, $I_j = 1$; otherwise, $I_j = 0$.

3.1.2. Forwarding Users' Privacy Concerns u_{fp}

References [3,19] points out that the higher the user's privacy concern is, the lower the probability of the privacy information being forwarded and the smaller the diffusion scope. It can be seen that users' privacy attention is also an important factor influencing their forwarding. In addition, users with high privacy attention generally do not share other people's privacy information or their friends' privacy information.

Suppose there is a list of privacy items $L_i = (a_{i_1}, a_{i_2}, \cdots, a_{i_n})$ and the corresponding privacy risk factors $S_i = (s_1, s_2, \cdots, s_n)$, even $\sum_{t=1}^{n} s_t = 1$. According to Equation (6), the privacy concerns value of user i can be obtained.

$$u_{fp}(i) = \sum_{t=1}^{n} s_t p_{it} \tag{6}$$

Based on reference [17], we denote the user's privacy concern factors include the following: user attributes, user privacy disclosure, friend privacy attitude and influence factors of the user as s_1, s_2, s_3 and s_4, respectively. The weights of these privacy concern factors can be set up according to Table 2:

Table 2. The setting of privacy concern factors.

s_1	s_2	s_3	s_4
0.1	0.5	0.3	0.1

3.1.3. Relation with Privacy Information Releasing User u_{rp}

Relation between users is also an important factor to influence user forward behavior. In this paper, the bidirectional attention of each user in the social network is defined as the user's first level of friends, and the friend of the user's friend is defined as the second level of friends, as well as other relations that do not belong to the above two relations, namely, the public. Finally, the values of the three user relationships mentioned above are set as: 2, 1, 0. We take u_{rp} as the relationship between users, the calculation is in Equation (7).

$$u_{rp} = 2 * rp_1 + 1 * rp_2 + 0 * rp_3 \tag{7}$$

where, rp_1 represents the relation between each user is first level of friends. The value is 0 or 1. rp_2 represents the relation between each user is second level of friends. The value is 0 or 1. rp_3 represents other relations. The value is 0 or 1.

3.1.4. Release Time of Private Information w_{ti}

The release time of private information is also an important factor affecting users' forwarding behavior. Reference [20] points out that users' demand and interest in information will decrease with the passage of time. In other words, the longer the information is released, the less likely the user is to forward the message. The time weight of the ith message is expressed with w_{ti}, the calculation is shown in Equation (8):

$$w_{ti} = \frac{1}{1 + \alpha_1 e^{(-\alpha_2 t)}} \tag{8}$$

In Equation (8), α_1 and α_2 are the time parameters, t represent time (in days), range from $-7 \le t \le 8$. That is, that day (0) represents 8, and the previous 15 days represent -7.

3.2. Prediction Model of User Forwarding Behavior

The basis of predicting privacy diffusion is to predict user forwarding behavior, but the problem of predicting user forwarding behavior cannot be simply attributed to the dichotomy problem in machine learning. Each node has the probability of forwarding or not, therefore the accuracy of judgment will be greatly reduced if each node is only considered according to forwarding or not forwarding. For example, in the process of spreading privacy information from the source node to the outside step by step, if a node is simply judged not to forward, all subsequent nodes of that node will be deemed not to forward. In practice, however, it is entirely possible for subsequent nodes in its diffusion path to be forwarded. Therefore, the forwarding probability of nodes must be considered.

Logistic regression model has a low complexity and is very suitable for large-scale data. Moreover, logistic regression model is an interpretable model, which can provide a certain degree of explanation for the importance of each feature through training. In addition, the key point is that the model predicts that when users forward some information, it will be presented in the form of probability value, which is very suitable for us to transform the prediction problem of user forwarding behavior into a probabilistic calculation problem. Therefore, after feature extraction, this paper adopts logistic regression model to predict user forwarding behavior, and the model is expressed by Equation (9).

$$F(u) = P_u(y_u = 1 \mid x) = \frac{1}{1 + e^{-(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n)}} \tag{9}$$

In Equation (9), $(x_1, x_2, \cdots x_n)$ expresses the feature vector which affects user u forwarding. In this paper, the feature vector is composed of p_{ew}, u_{fp}, u_{rp} and w_{ti}. y_u expresses the forwarding behavior of user u, when $F(u) = P_u(y_u = 1 \mid x)$ is beyond a certain threshold θ, user u will forward the private information; if not, user u will not forward it. θ is related to the prediction accuracy of the logistic regression model, which is determined by experiments. α_0 is the intercept, vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ is the weight of corresponding vector $(x_1, x_2, \cdots x_n)$, it can be obtained during model training.

4. Privacy Information Diffusion Prediction

4.1. Choice of Key Forwarding Nodes

The information diffusion in social networks is largely influenced by opinion leaders, who can promote the rapid diffusion of information and expand the scope of influence. The so-called opinion leaders are the users with a certain influence in the social network [21]. Users with high influence are the key to the information spread scale and continuity. In social network research, the problem of maximizing the information diffusion scope is also defined as the problem of maximizing of users influence [22]. Therefore, finding the users who have important influence in the information diffusion process and the corresponding nodes in the social network topology are called key forwarding nodes, and they can predict the information diffusion more accurately and effectively.

Based on the time interaction between nodes [23] and the topological connection structure between nodes, this paper presents a choice method of key forwarding nodes.

4.1.1. Calculate the Time Interaction Behavior Weight between User u and User v

$$r_u^v = \sum_{t=-\frac{T}{2}}^{t=\frac{T}{2}} A'(u,v,t)[\frac{1}{1+\alpha_1 e^{(-\alpha_2 t)}}] \tag{10}$$

In Equation (10), r_u^v is the interaction weight of user u and v; α_1 control time interaction curve steep. The longer the time of the interaction, the smaller the weight value. The closer the time of interaction, the greater the weight; α_2 is used to control the distribution of interaction weight. Through the adjustment of α_1, α_2 values, we can get the results as shown in Figure 1. Figure 1 shows that when α_1, α_2 were 0.1, 0.7, the α curve steep potential was the most obvious and the time effect was strongest.

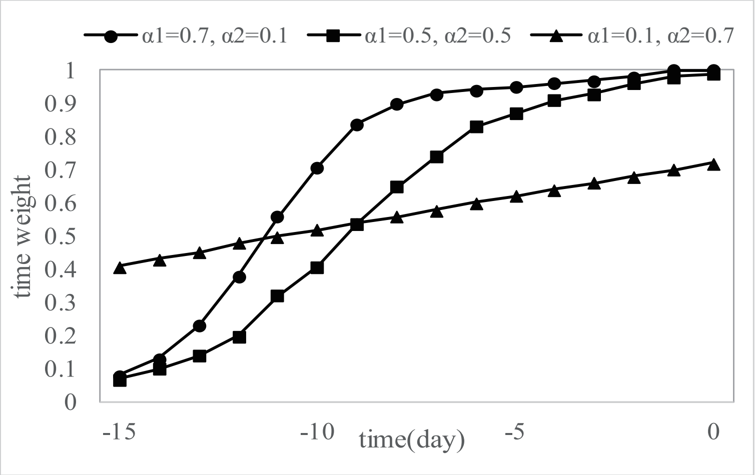


Figure 1. Time weight changes with time.

$A'(u,v,t)$ refers to the interaction between user u and v in time t, which can be represented by the probability of the user browsing information, the activity of fans and the intensity of fan interaction with concerned users. The calculation steps of $A'(u,v,t)$ are as follows:

(1) Calculate the probability of the information being viewed by the user's fans $p_{see}(u,v,t)$.

$$p_{see}(u,v,t) = \frac{T_{count}(u,t)}{\sum_{i \in Friends(v)} T_{count}(i,t)} \tag{11}$$

where, *Friends(v)* represents concerned users collection of fan v, even $u \in Friends(v)$. $T_{counts}(u,t)$ represents the total number of messages released by user u during the time period t.

(2) Calculate the activity $Act(v,t)$ of fan v

$$Act(v,t) = \frac{T_{count}(v,t) + T_{repost}(v,t) + T_{comment}(v,t)}{t} \tag{12}$$

where, $T_{count}(u,t)$ is the total amount of messages released by the user during time period t, $T_{repost}(v,t)$ is the number of times someone retweets someone else's message, $T_{comment}(v,t)$ is the number of times where v participated in comments.

(3) Calculate the interaction strength $Relation(u,v,t)$ between fan v and user u. It reflects the power degree of relation between users. The stronger the interaction relationship, the greater the influence on information recipient, and the more likely he/she will participate in forwarding. It can be calculated by the following Equation (13):

$$Relation(u,v,t) = \alpha \frac{T_{repost}(u,v,t)}{T_{repost}(v,t)} + \beta \frac{T_{comment}(u,v,t)}{T_{comment}(v,t)} \tag{13}$$

where, $T_{repost}(u,v,t)$, $T_{comment}(u,v,t)$ are respectively the total number of micro-blog and total number of comments which v forward u within time period. α, β are respectively the weight of retweet and comment, and $\alpha + \beta = 1$.

In conclusion, the formula for calculating the weight of interaction behaviors between users can be obtained as follows:

$$A'(u,v,t) = p_{see}(u,v,t) \cdot Act(v,t) \cdot Relation(u,v,t) \tag{14}$$

4.1.2. Measure the Contribution Rate of Fan v of User u and the Fan w to u

The calculation method is as follows:

$$P_u^v = r_u^v + r_u^v \sum_{w \in V \cap w \neq u} r_v^w \tag{15}$$

where, P_u^v is the influence v to u, v is the fan of u, w is the fan of v.

4.1.3. Influence Calculation of User u

Assume that user u_0 releases the information and the fan set of u_0 is F, the influence of user u ($u \in F$) is calculated by Equation (16).

$$NR(u) = r\frac{NI(u)}{max_{g \in F}NI(g)} + (1-r)\frac{d(u)}{max_{g \in F}d(g)} \tag{16}$$

where, $NI(u) = \sum_{v \in U} P_u^v$, refers to the influence of all fans v of user u, and fans of fans w to u, $MAX_{g \in F}NI(g)$ represents the most influential node among fans. *d(u)* is the out-degree of user u, meaning the number of followers. $MAX_{g \in F}d(g)$ represents the fans number of the node with the largest out-degree in the fan set. *r* is a parameter whose value is adjustable between [0, 1], when *r = 0*, it represents that the influence of u only considers the networking topology relations of the node, when *r = 1*, it represents that the influence of u only considers the social interaction of that node.

The impact of the nodes can be calculated from Equation (16). The greater the influence, the more critical the node.

4.2. Determination of Key Forwarding Path

According to the communication characteristics of social network, users who play a key role in the process of information transmission tend to drive large-scale forwarding, and information will be transmitted layer by layer through many key forwarding nodes, and then spread widely in the social network.

In order to obtain the critical forwarding path, firstly, we find the key forwarding node from the root level by level, in which the root means initial node of privacy information. Then, we can connect the key forwarding nodes to obtain the critical forwarding path.

According to the theory of small world effect in complex network [24], a message can reach any other user in the network after passing 6 levels at most, that is to say, the average shortest distance between users is 6 levels. We used 8630 original micro-blogs of 500 users in micro-blog to calculate the proportion of their average number of reposts at each level. The results are shown in Figure 2.

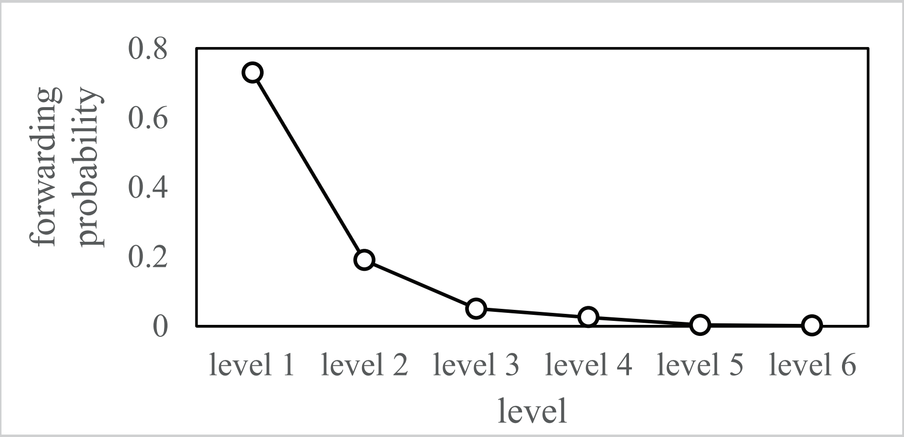


Figure 2. Micro-blog forwarding series.

As can be seen from Figure 2, the forwarding situation is mainly concentrated in the first two levels, and the forwarding rate is only 2.5% at the fourth level, and the forwarding rate after the sixth level is basically zero. Therefore, when looking for the critical path, we only need to set the number of calculation iterations to 4. The pseudo-code of the critical path finding algorithm is shown in Algorithm 1.

Algorithm 1 critical path search algorithm

```
Input: social graph $G(V, E)$, release user $u$, privacy micro-blog $T$
Output: critical path set $KPath$
Initialization data: critical $path$, critical path set $KPath$
FUNCTION GetKeyPath($u$)
for each $u_i$ in follower($u$) do
    calculate influence NR($u_i$)
    if NR($u_i$)>threshold then
        path ← path add $u_i$ //take $u_i$ add path
        if Level($path$)≤4 then
            $path \leftarrow GetkeyPath(u_i)$ // recursively find the path
        else
            return $path$
        end if
    else
        $KPath \cup path$ //add $path$ to key path set
    end if
end for
End Function
```

In Algorithm 1, threshold refers to the threshold of influence. Nodes with influence greater than the threshold can be identified as key nodes, and the determination of this value can be obtained through the following experiment 1.

The complexity analysis of Algorithm 1 is as follows: assumes that the number of users u is n, select a subset n_1 ($n_1 << n$), the influence of which is greater than the threshold of the node. In addition, select the subset of n_1 until the end of the iteration cycle. Therefore, the time complexity is $O(n * n_1^k)$. Where, k represents the number of recursions.

We can illustrates the key steps of Algorithm 1 through a simple example of privacy information dissemination (as shown in Figure 3). Assume that the initial node root is node 1. In Figure 3, each tagged dot represents the user, the gray circle represents the key node and the line connecting the two users represents the fan relationship. For the publishing user (i.e., the initial root node), after Algorithm 1 processing, the key node set is {2, 5, 7}. Then, according to Algorithm 1, key nodes are searched layer by layer, and the key node set of node 2 is 11, the key node set of node 5 is 20, and the key node set of node 7 is 25. By analogy, the final critical path is as follows: 1 → 2 → 11 → 13 → 17, 1 → 5 → 20 → 22, 1 → 7 → 25 → 30 → 31.

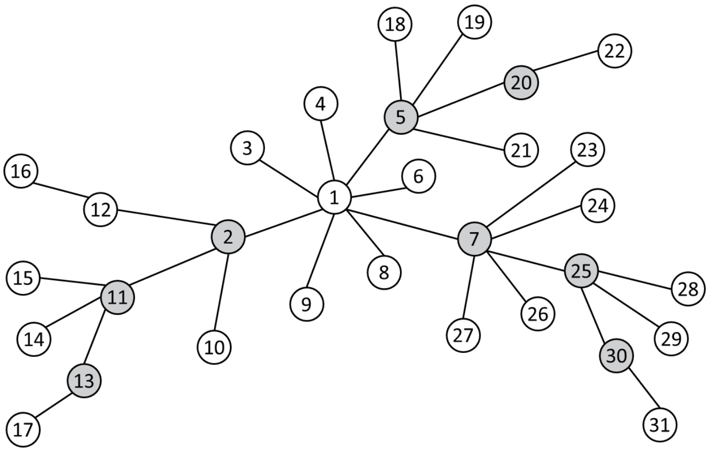


Figure 3. Simple privacy micro-blog transmission diagram.

4.3. Privacy Diffusion Prediction Model

Privacy information diffusion prediction means to predict the depth and breadth of privacy information dissemination in the future according to certain methods and rules on the basis of mastering the existing information diffusion forms, so as to realize the interference and control of privacy information dissemination.

As mentioned above, the user forwards the micro-blog of the superior user with a certain probability, so the forwarding probability of the user is affected by his previous level. For example, as shown in Figure 4 is a critical forwarding path, when u_1 forwards the information published by u_0, the forwarding probability of u_1 is $p_1 = p(u_1)$. Then the forwarding probability of u_2 is $p_2 = p_1 \times p(u_2|u_1)$ when they forward the information forwarded by u_1. Since u_1 and u_2 are independent of each other, p_2 can express $p_2 = p(u_1) \times p(u_2)$. By analogy, at the i level, the forwarding probability of u_i is $p_i = p(u_1) \times p(u_2) \times \cdots \times p(u_i)$.

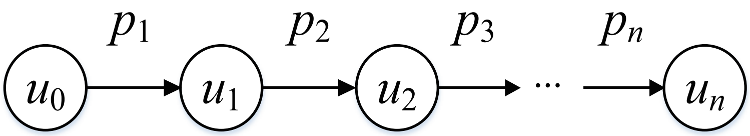


Figure 4. Forwarding path diagram.

In conclusion, the method through the size of forwarding probability (p_i) to determine the searching conditions of critical path keep searching down, is as follows:

$$R(u) = \begin{cases} +1, & p_i \geq \varepsilon \\ -1, & p_i < \varepsilon \end{cases} \tag{17}$$

When forwarding probability is greater than or equal to ε, we can downward along the critical path to search the forwarding users. Otherwise, stop searching. The value of ε can be determined by privacy diffusion scope predicting experiment.

The establishment of privacy diffusion prediction model mainly includes three processes: (1) the key forwarding node selection according to Algorithm 1; (2) determine the key forwarding path; (3) follow the key forwarding path and use logistic regression model to predict user forwarding behavior for each key forwarding node level by level. If the key user forwarding probability product on the critical path is less than a certain threshold or reaches the maximum number of iterations, the search is stopped. Otherwise, the forwarding probability or user forwarding behavior is calculated for the next layer node along the critical path. Finally, the number of forwarding users (fans) and the length of the critical forwarding path of root node and key forwarding node at each level are calculated to obtain the diffusion range and depth of privacy information.

Based on the above analysis, the design of DEPM-KP (Diffusion Effect Predict Model Based on Key Paths) is shown in Algorithm 2.

Algorithm 2 Diffusion prediction model based critical path DEPM-KP

```
Input: social graph G(V,E), released user $u$, privacy information $T$
Output: diffusion scale $R_s$, diffusion depth $R_d$
initialize $R_s = 0$, $R_d = 0$, $depth = 1$
select user $u$, create a forward queue $Q$, $u$ add queue
through Algorithm 1, the critical path set $KPath$ is obtained
for $\forall path_i \in KPath$ do
    for $\forall v_j \in path_i$ do
        if $L(v_j) = 1$ then
            obtain the set of fans of $v_j$: $follower(v_j)$
            for $\forall node \in follower(v_j)$ do
                Calculate the forwarding probability $F(node)$
                if $F(node) > \theta$ then//node forwarding probability meet $\theta$
                    $R_s \leftarrow R_s + 1$
                end if
            end for
            $depth \leftarrow depth + 1$
        end if
        if $L(v_j) = -1$ then
            $R_d \leftarrow max(R_d, depth)$// Select the longest path
            break
        end if
    end for
end for
```

The complexity analysis of Algorithm 2 is as follows: the algorithm is divided into two stages: (1) in the first stage, we find the critical path, the time complexity is $n * n_1^k$, in which n is the number of fans of user u, and k is the number of recursion; (2) in the second stage, for each node in the critical path of the forwarding probability calculation, the time complexity is $m * N * n$, in which m is the total number of critical path, $N(0 \leq N \leq 4)$ is the number of key nodes on the critical path. Finally, we can get the total time complexity is $O(n * n_1^k + m * N * n)$.

5. Experiment

5.1. Experimental Data and Evaluation Criteria

This paper carried out experiments on Sina Weibo, a typical social network. The experimental data set was obtained through crawler and Sina API. The data set contained a total of 6,541,203 micro-blog posts published by users, as well as a total of 54,629,532 following relationships among users. Among them, there are 104,530 private micro-blogs. After proper data cleaning, a total of 103,176 private micro-blogs were obtained, among which 59,706 were forwarded.

The experiment in this paper consists of two parts: prediction of user forwarding behavior and prediction of privacy micro-blog proliferation. The evaluation methods used to predict the accuracy of user forwarding behavior were accuracy, recall rate and F value. For the diffusion prediction experiment [4,21], this paper adopts the following two indicators as the evaluation criteria.

(1) Accuracy of prediction of privacy diffusion range: the forwarding number of micro-blog conforms to the characteristics of power law distribution, that is, the forwarding scale of most users is small, while only a small number of users have a large forwarding scale, and the forwarding scale is even different by several orders of magnitude. Therefore, the scope of privacy diffusion can be predicted by an order of magnitude, so that the evaluation index can be suitable for micro-blog forwarding of different sizes. The accuracy index of diffusion range can be calculated by the following process.

Assume the positive integer a, b, m meet a<b, and $10^a < m < 10^b$, diffusion scale $m(S_m)$ is defined as the area which take m as the midpoint and expand half to the left and right, that is:

$$S_m \in [m - \frac{10^b - 10^a}{2}, m + \frac{10^b - 10^a}{2}] \tag{18}$$

Then, when the actual diffusion scale N_f and predicting diffusion scale N_p satisfy the following formula, we judged to be right.

$$\mid N_p - N_f \mid < \frac{10^{\log_{10}\lceil N_f \rceil} - 10^{\log_{10}\lfloor N_f \rfloor}}{2} \tag{19}$$

where, $\lceil \cdot \rceil$ express integer upwards, $\lfloor \cdot \rfloor$ express integer down.

(2) The prediction accuracy of privacy diffusion depth: the diffusion depth of the micro-blog is usually very small, so this article selects deviation rate π to evaluate the accuracy of the diffusion depth. This index can be calculated by the following equation:

$$\pi = \frac{\mid d_p - d_m \mid}{max(d_p, d_m)} \tag{20}$$

where, d_p is the prediction depth, d_m is the actual diffusion depth. It can be seen that, when $\pi = 0$, indicates prediction is accurate; when $\pi \in (0, 1]$, the predicted depth is not equal to the actual diffusion depth and there is an error in the prediction.

5.2. User Forwarding Behavior Prediction Experiment

During model training, the data set is prepared according to 1:1 ratio for the training set and test set. By calculating various features, the prediction model of user forwarding behavior is established, and the model parameters are trained by micro-blogs that are forwarded or not forwarded. The threshold θ of logistic regression prediction model will be determined according to the forecast accuracy, accuracy under different θ values change trend is shown in Figure 5.

The Figure 5 shows that, when θ set to 0.51, the highest prediction accuracy. In the process of prediction, the trained classification model is used to carry out classification, so as to realize the prediction of user forwarding behavior. Under the same data set, naive Bayes, logistic regression and support vector machine are adopted to carry out the comparative experimental of the prediction effect, as shown in Figure 6.

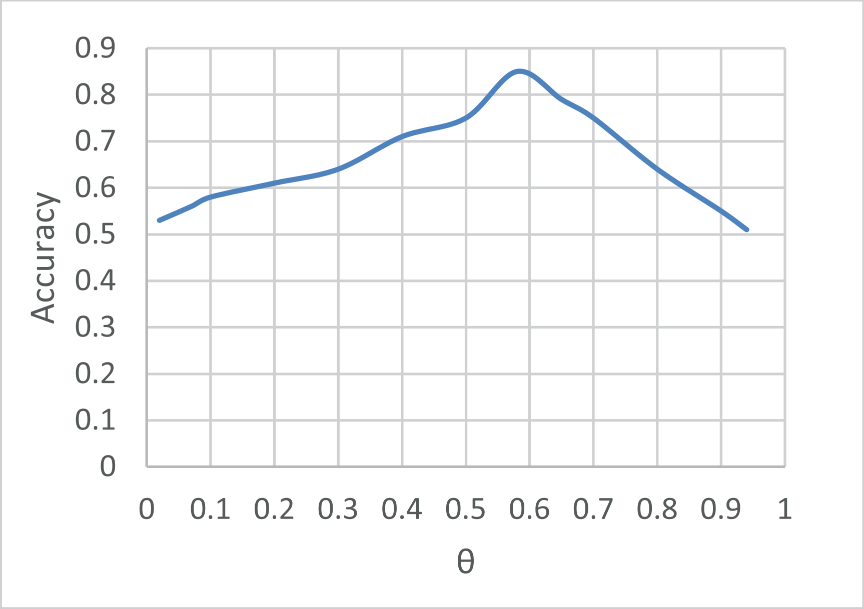
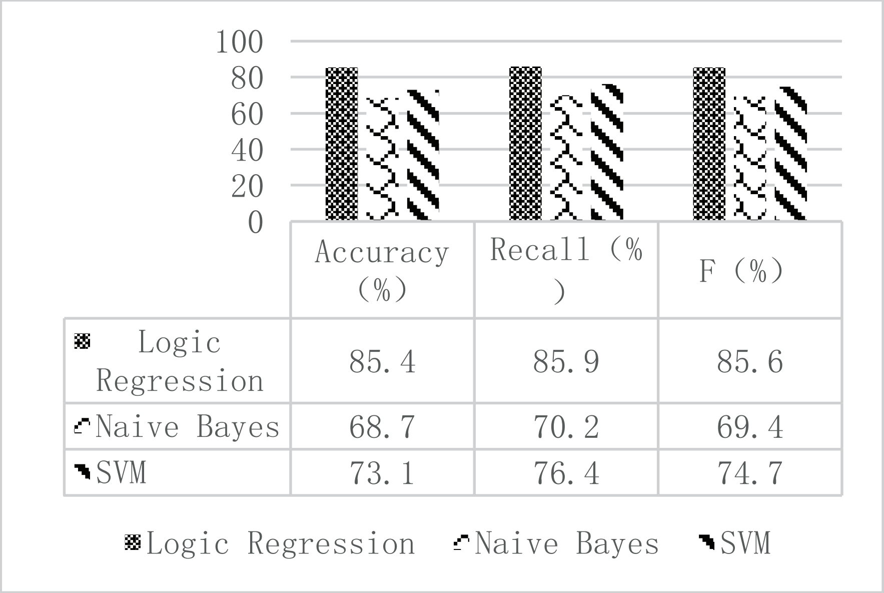


Figure 5. Accuracy changes with *θ*.



	Accuracy（%）	Recall（%）	F（%）
▩ Logic Regression	85.4	85.9	85.6
Naive Bayes	68.7	70.2	69.4
SVM	73.1	76.4	74.7

Figure 6. Comparison of experimental results of three different methods.

As can be seen from Figure 6, the accuracy rate and recall rate of logistic regression model used to predict user forwarding behavior reached 85.4% and 85.9%, respectively, which were significantly better than the prediction effect of naive Bayes and support vector machine models. Therefore, it is feasible to select logistic regression model for prediction.

5.3. Privacy Diffusion Prediction Experiment

The key to predict the spread of private micro-blog is to find influential users, that is, to select key forwarding nodes. How many influential users are crucial to the spread of private information? In this paper, all key forwarding nodes are selected by determining the threshold of influence in Algorithm 1, and the coverage of information diffusion is calculated for these key nodes [25]. Coverage is defined as the proportion of the number of nodes affected by key nodes to the actual diffusion nodes, and its calculation formula is shown in Equation (21). According to Equation (21), the coverage is related to the number of key nodes selected. The larger the coverage, the stronger the diffusion capacity of these key forwarding nodes is, and the more reasonable the threshold value is.

$$S = \frac{Total\ number\ of\ forwarding\ users\ affected\ by\ all\ key\ nodes}{Total\ number\ of\ forwarding\ users} \tag{21}$$

In Equation (21), the selection of key node is determined according to the threshold of user influence. If the influence of the node is greater than the threshold, then, the node is determined as the key node. In order to determine threshold value, we respectively select the private scale within the range of 0–100, 101–1000 and 1001–10,000 for the experiment. If a user forwards multiple times from the same user, only one forwarding relationship remains. The experimental results are shown in Figure 7.

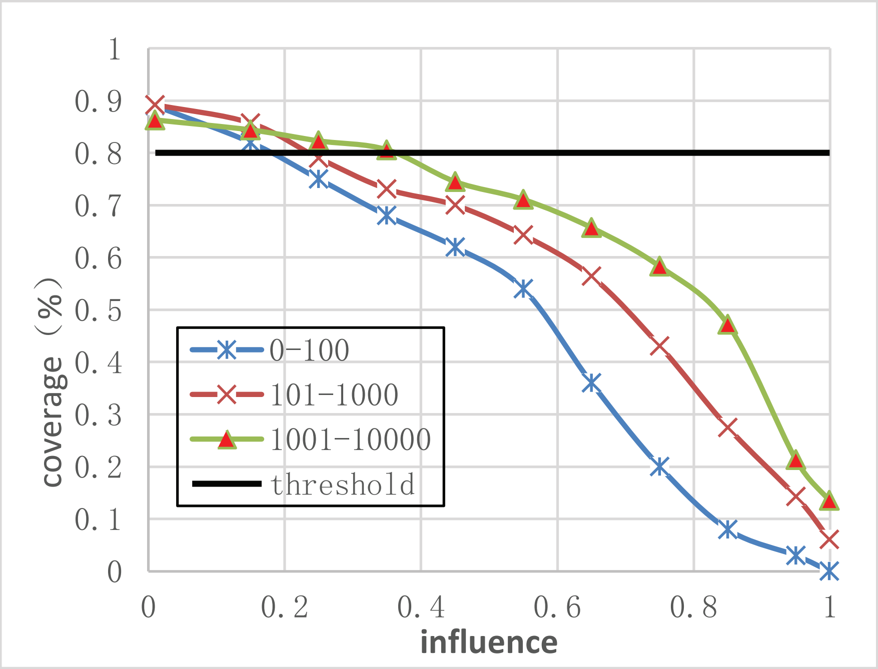


Figure 7. Coverage changes with impact.

As can be seen from Figure 7, when the coverage is determined to be above 80%, the user influence threshold of different forwarding sizes is also different. Among them, when the forwarding scale is between 1001 and 10,000, the threshold size is 0.35. In other words, the node with influence greater than 0.35 can be selected as the key node. It can also be seen from Figure 7 that when the selected influence is 0.01, no matter what the forwarding scale is, the coverage rate can reach about 88%. When the influence is 0.9999, the coverage of forwarding scale between 0 and 100 is close to 0, while the coverage of forwarding scale between 101–1000 and 10,001–10,000 is also very low. This shows that in the process of information diffusion, there are only a few nodes with great influence, and the information diffusion process is also promoted by many key nodes.

In order to verify the prediction effect of the model in this paper on the spread of privacy micro-blog, the spread effect selected in this paper mainly includes the two dimensions of spread scope and spread depth, and the accuracy index of spread scope is adopted as the evaluation standard.

In the experiment, we selected 6320 original private micro-blogs published by 500 users and used the method proposed in this paper to predict the diffusion range. The selection principles of these users are as follows:

(1) micro-blog is often forwarded;
(2) a certain amount of micro-blog forwarding.

The experimental results are shown in Figure 8.

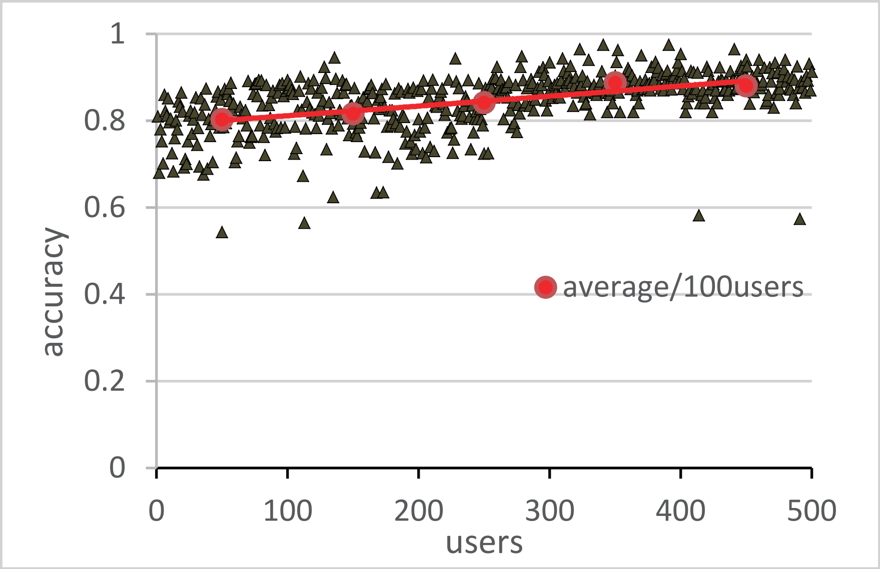
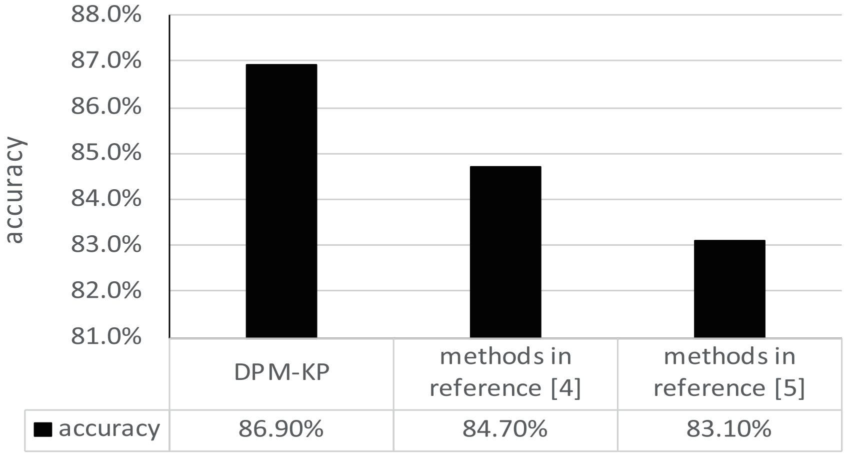


Figure 8. Prediction accuracy of diffusion range.

The experimental results show that, when ε = 0.25, the highest privacy diffusion range accuracy. From the prediction results, the prediction accuracy of 500 users' diffusion range was averaged, and the overall prediction accuracy was 86.6%. It can be seen that this method can better predict the spread range of private micro-blog.

In order to prove the effectiveness of the method in this paper, the method proposed in references [4,5] were selected for comparison. Reference [4] adopted a recursive method to predict the forwarding behavior of each user step by step. Reference [5] adopted a diffusion scale prediction model (RSPM-RPT) based on forwarding probability transmission. By conducting diffusion range prediction experiments on the same data set, the comparison results are shown in Figure 9.



	DPM-KP	methods in reference [4]	methods in reference [5]
■ accuracy	86.90%	84.70%	83.10%

Figure 9. Comparison experiment of diffusion range prediction.

The comparison of experimental results shows that the privacy diffusion prediction model proposed in this paper based on the key forwarding node and the key forwarding path is more accurate in predicting the privacy micro-blog diffusion range

When predicting the diffusion depth of private micro-blog, as long as the critical forwarding path is found, the diffusion depth of the target can be obtained according to the length of the critical forwarding path. In the stage of data acquisition, we retained the relatively complete forwarding path of private micro-blog. We extracted 9830 forwarding paths for prediction of diffusion depth and averaged the experimental results. The experimental data were expressed according to the forwarding path representation proposed in reference [16]. That is ($orgin\_mid$, mid_1, mid_2, $\cdots$, mid_n, $ignore\_mid$), but we take $orgin\_mid$ as the id of the released user, mid_i refers to the intermediate forwarding user, $ignore\_mid$ refers the last user to forward id. Taking deviation rate π as evaluation standard, the experimental results as shown in Table 3.

Table 3. Accuracy of diffusion depth.

value range of π	0	(0, 0.5)	(0.5,1)	1
prediction accuracy of privacy diffusion depth	89.3%	10.1%	0.5%	0.1%

The experimental results show that the prediction results of diffusion depth with high accuracy can be obtained according to the found critical forwarding path length. In fact, once the private micro-blog spreads, we can calculate the spreading depth of the micro-blog according to the key forwarding path. If the key forwarding path is cut off in time, the spreading range of the private micro-blog can be effectively suppressed.

6. Conclusions

To measure the diffusion of private micro-blog, we proposed an intelligent prediction model for the diffusion of private information based on comprehensive data analysis. By identifying key nodes, the model can find the users who play an important role in the diffusion trend, and predict users' forwarding behaviors along the critical path. The experimental results show that the accuracy of our intelligent prediction model can reach 86.9%, higher than that of using the logistic regression model. Analytical results based on real-world data show that our proposed method is effective and accurate to the predict the diffusion of private messages on social networks.

Author Contributions: Methodology, Y.L.; software, H.X.; writing–original draft preparation, H.J.; writing–review and editing, X.Y. and H.Z.; supervision, H.X.; project administration, Y.L.; resource, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the New Generation of Artificial Intelligence Special Action Project (Grant No. AI20191125008); The Major Special Science and Technology Project of Hainan Province (Grant No. ZDKJ2019008).

Acknowledgments: The author would like to thank the support of Director Foundation Project of National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data (PSRPC).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dong, G.; Gao, J.; Huang, L.; Shi, C. Online Burst Events Detection Oriented Real-Time Microblog Message Stream. *Comput. Mater. Contin.* **2019**, *60*, 213–225. [CrossRef]
2. Liu, Y.; Zhang, T.; Jin, X.; Cheng, X. Personal privacy protection in the era of big data. *J. Comput. Res. Dev.* **2015**, *52*, 230–232.
3. Li, Z. *Study on User Privacy in Mobile Internet*; Beijng University of Posts and Telecommunications Press: Beijng, China, 2019.
4. Hou, W.; Huang, Y.; Zhang, K. Research of micro-blog diffusion effect based on analysis of retweet behavior. In Proceedings of the IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing, Beijing, China, 6–8 July 2015; pp. 255–261.
5. Li, Y. Research on Microblog Communication Effect Prediction Technology. Ph.D. Thesis, PLA University of Information Engineering, Zhengzhou, China, 2013.

6. Zhou, H.; Sun, G.; Fu, S.; Jiang, W.; Xue, J. A Scalable Approach for Fraud Detection in Online E-Commerce Transactions with Big Data Analytics. *Comput. Mater. Contin.* **2019**, *60*, 179–192. [CrossRef]
7. Wang, B.; Kong, W.; Guan, H.; Xiong, N.N. Air Quality Forecasting Based on Gated Recurrent Long Short Term Memory Model in Internet of Things. *IEEE Access* **2019**, *7*, 69524–69534. [CrossRef]
8. Huang, Y.; Liu, Z.; Yu, M.; Gao, M. The Microblog Retweeting Prediction Evaluation System and Performance Comparation The Microblog Retweeting Prediction Evaluation System and Performance ComparationThe Microblog Retweeting Prediction Evaluation System and Performance Comparation. *J. Harbin Univ. Sci. Technol.* **2013**, *18*, 52–57.
9. Yu, H.; Yang, X. Studying on the node's influence and propagation path modes in microblogging. *J. Commun.* **2012**, *S1*, 96–102.
10. Wu, K. Information Transmission Modeling and Node Influence Research Based on Microblog. Ph.D. Thesis, PLA University of Information Engineering, Zhengzhou, China, 2013.
11. Gruhl, D.; Guha, R.; Liben-Nowell, D.; Tomkins, A. Information diffusion through blogspace. In Proceedings of the 13th International conference on World Wide Web, New York, NY, USA, 17–20 May 2004; pp. 491–501.
12. Yang, X.; Liu, Y. Application of Improved SIR Model on Information Diffusion in Microblog. *Sci. Mosaic* **2015**, *2*, 12–16.
13. Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; Hurst, M. Cascading behavior in large blog graphs: Patterns and a model. In *Society of Applied and Industrial Mathematics: Data Mining*; Carnegie Mellon University: Pittsburgh, PA, USA, 2007; pp. 551–556.
14. Xie, J.; Liu, G.; Su, B.; Meng, K. Prediction of User's Retweet Behavior in Social Network. *J. Shanghai Jiaotong Univ.* **2013**, *4*, 584–588.
15. Zhang, Y.; Lu, R.; Yang, Q. Predicting retweeting in microblogs. *J. Chin. Inf. Process.* **2012**, *26*, 109–114.
16. Cao, J.; Wu, J.; Shi, W.; Liu, B.; Zheng, X.; Luo, J. Sina Microblog Information Diffusion Analysis and Prediction. *Chin. J. Comput.* **2014**, *37*, 779–790.
17. Xu, H.; Xu, Y. Research on privacy disclosure detection in microblog. In Proceedings of the 3rd IEEE International Conference on Computer and Communications, Chengdu, China, 13–16 December 2017; pp. 1479–1486.
18. Hu, J.; Sun, J. A case retrieval method of hybrid data based on information entropy. In Proceedings of the 2nd IEEE International Conference on Computer and Communications, Chengdu, China, 14–17 October 2016; pp. 155–159.
19. Bioglio, L.; Pensa, R. Modeling the impact of privacy on information diffusion in social networks. In *International Workshop on Complex Networks*; Springer: Cham, Switzerland, 2017; pp. 95–107.
20. Li, X.; Croft, W. Time-based language models. In Proceedings of the Twelfth International Conference on Information and Knowledge Management, New Orleans, LA, USA, November 2003; Association for Computing Machinery: New York, NY, USA, 2003; pp. 469–475.
21. Ding, Z.; Jia, Y.; Zhou, B.; Tang, F. Survey of Influence Analysis for Social Networks. *Comput. Sci.* **2014**, *41*, 48–53.
22. Richardson, M.; Domingos, P. Mining knowledge-sharing sites for viral marketing. In Proceedings of the Eighth ACM SIGKDD International Conference On Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 61–70.
23. Ullah, F.; Lee, S. Identification of influential nodes based on temporal-aware modeling of multi-hop neighbor interactions for influence spread maximization. *Physica A* **2017**, *486*, 968–985. [CrossRef]
24. Chaoran, F.; Huang, S.; Li, Y. Study on microblog social network community detection. *Microcomput. Appl.* **2012**, *23*, 67–70.
25. Chen, Z.; Liu, X.; Li, B. Analyzing micro-blog users' propagation influence based on behavior and community. *Appl. Res. Comput.* **2018**, *7*, 37.