Hindawi
Security and Communication Networks
Volume 2020, Article ID 8014958, 11 pages
https://doi.org/10.1155/2020/8014958

Research Article

Abnormal User Detection Based on the Correlation Probabilistic Model

Xiaohui Yang and Ying Sun

School of Cyber Security and Computer, Hebei University, Baoding, China

Correspondence should be addressed to Xiaohui Yang; yxh@hbu.edu.cn

Received 25 October 2019; Revised 1 February 2020; Accepted 24 February 2020; Published 13 June 2020

Guest Editor: Wenjia Li

As an important part of the new generation of information technology, the Internet of Things (IoT), which is developing rapidly, requires high user security. However, malicious nodes located in an IoT network can influence user security. Abnormal user detection and correlation probability analysis are fundamental and challenging problems. In this paper, the probabilistic model of the correlation between abnormal users (PMCAU) is proposed. First, the concept of user behavior correlation degree is proposed, which is defined as two parts: user attribute similarity degree and behavior interaction degree; the attribute similarity measurement algorithm and behavior correlation measurement algorithm are constructed, respectively, and the spontaneous and interactive behaviors of users were analyzed to determine the abnormal correlated users. Second, first-order logic grammar is used to express the before and after connection of user behavior and to deduce the probabilistic of occurrence of the correlation of behavior and determine the abnormal user groups. Experimental results show that, compared with the traditional anomaly detection algorithm and Markov logic network, this model can identify the users correlated with anomalies, make probabilistic inferences on the possible associations, and identify the potential abnormal user groups, thus achieving higher accuracy and predictability in the IoT.

1. Introduction

Internet of Things (IoT) is the latest evolution of the Internet, including a great deal of connected physical devices and applications [1]. IoT allows object collection, data exchange, etc. [2], which can perform medical data management, medical information monitoring, and user information analysis. IoT is an open network, and there are a large number of malicious nodes in the network. These malicious nodes may tamper with the correct data and pass them to other nodes. The normal nodes will use the wrong data for information dissemination due to the lack of ability to verify the correctness of the messages received, resulting in the dissemination of false information on medical, social, and other networks. Network individuals form a "relationship structure" through various connection relationships, including virtual communities composed of various complex relationship associations; [3] based on the relationship structure, a large number of network individuals aggregate and influence each other around an event to form a "network group" with common behavior features; based on the relationship structure and network group, all kinds of information can be quickly released and spread to form social media, feedback and act on the real society, so that the interaction the real society has a great impact on the real world.

To detect abnormal user groups in the IoT, the probabilistic model of correlation between abnormal users (PMCAU) is proposed. PMCAU studies the daily behavior of malicious users, considers the behavior information and interaction information of malicious users, constructs the attribute similarity measurement algorithm and behavior correlation measurement algorithm, calculates various information of malicious users respectively, and finds abnormal correlated users. At the same time, probabilistic soft logic is used to predict the possibility of abnormal correlation in the future, to identify the potential abnormal user groups in the IoT.

The contributions of this paper are listed as follows:

(1) Construct attribute similarity measurement algorithm, read user attribute data, and calculate user attribute similarity, including geographic location similarity, user follower similarity, and personal information similarity.

(2) Construct behavioral interaction measurement algorithm, consider user interaction behavior information, calculate the degree of interaction between user interaction behaviors, and reflect the features of user interaction behavior.

(3) Propose the concept of behavioral correlation degree as an important distinguishing indicator between abnormal correlated users and nonabnormal correlated users. At the same time, the correlation threshold is defined, correlation threshold judgment based on behavioral correlation degree to identify abnormal correlated users.

(4) Use probabilistic soft logic to express the causes and consequences of abnormal behavior correlation in abnormal user groups and combine multiple factors such as geographic location, behavior information, interest and preference, and fans' attention information to propose a set of reasoning rules more suitable for predicting the probabilistic of correlation among abnormal correlated users and to identify the potential abnormal user groups.

(5) Analyze the performance of PMCAU in the real data set and compared it with other algorithm models. PMCAU has better performance in improving the accuracy, stability, and probabilistic reasoning of identifying abnormal correlated users.

The rest of this paper is organized as follows: In Section 2, we analyze and explain existing related work and theoretical basis. The model is described in Section 3. In Section 4, we introduce the proposed model PMCAU in detail. In Section 5, we present the experimental results. Finally, we conclude our work in Section 6.

2. Related Work

Users are the main part of IoT. Accurate analysis of their behavioral correlation and anomaly detection and inference prediction are necessary to maintain network security. For abnormal detection in the IoT, scholars mostly adopt analysis methods based on network traffic behavior [4, 5] and analysis methods based on host behavior anomalies, such as network behavior modeling through host audit command [6]. But this class of methods only considers the behavior of a single user, without considering the influence of the user's own attributes, and hence it has lower detection accuracy. Recently, research results have been obtained by analyzing users' daily behaviors to predict their later behaviors and detect abnormal behaviors of users. Cao et al. [7] statistically analyzed various factors that may affect user behaviors and used logistic regression, Bayesian network, and other methods to predict user forwarding behaviors by virtue of user attributes, social relations, and other characteristics. Xu et al. [8] used a hybrid implicit topic model to predict user forwarding behavior based on the features of the user's forwarding behavior being affected by factors such as breaking news and users' own interests. The above algorithms do not take into account the multiple behaviors of users and the correlation between behaviors, so it cannot comprehensively judge and deal with abnormal behaviors of users' attribute information and behavioral information, and its reliability and accuracy are often difficult to be guaranteed. The PageRank-based account anomaly detection algorithm [8] builds a social relation matrix based on the user relationship and ranks the account to detect malicious users through the iterative calculation of PageRank value. Because this method does not consider the user's own attribute characteristics, the ranking result of the user is affected by the time delay, so the accuracy rate is low in the heterogeneous network with complex structure and uneven scale. The Markov logic network is a statistical relation learning model combining Markov network and first-order logic [9, 10]. Although it has certain reasoning abilities for correlation probability of microblog users, its accuracy is low because it uses Boolean value {0, 1} instead of continuous value [0, 1] to determine whether there is an abnormal correlation. Sun et al. [11] proposed a joint fraud detection method based on abnormal user groups and mining abnormal user groups through the similarity adjacency graph. Yang et al. [12] proposed a step-by-step detection method to find the anomaly level, constructed a scoring matrix, and used the fast maximum margin matrix decomposition to perform level prediction to capture anomalies. These algorithms cannot accurately reflect the behavior features of users, cannot adapt to the features of microblog data information, and cannot make probabilistic reasoning on the occurrence of behavioral correlation, so the accuracy of detecting abnormal user groups is low.

In summary, existing anomaly group detection methods have two important defects. First, the methods based on the network popularity or behavior prediction model only consider the behavior information sent by the user. It is impossible to comprehensively judge and process abnormal behaviors of user attribute information and behavior information, and its reliability and accuracy are often difficult to guarantee. Second, the user's multiple behaviors and the relationship between behaviors are not taken into consideration, and anomaly detection and probabilistic reasoning of the group cannot be taken into consideration, and the accuracy of detecting potential abnormal user groups is low.

When objects connected to the Internet of Things continue to generate information and report to Internet users, a noteworthy development is that they will also join traditional social networks and interact with "people" in social networks. Social networks are not just person-to-person social, but person-to-person, person-to-thing, and thing-to-thing. Therefore, abnormal user groups in social networks will inevitably pose a threat to the security of the

Internet of Things. At this time, the Internet of Things to hardware also has social attributes. Therefore, to maintain the security of the Internet of things and detect abnormal user groups in the network, in response to the above problems, the probabilistic model of correlation between abnormal users (PMCAU) is proposed by taking the social platform of microblog with a large user volume as an example. PMCAU defines the concept of behavioral correlation degree and constructs an attribute similarity measurement algorithm and a behavioral interaction measurement algorithm to identify abnormal correlated users in microblog and to perform probabilistic reasoning in abnormal correlated users to complete the detection of abnormal user groups.

3. Model Description

To accurately identify abnormal users and find potential abnormal user groups before the occurrence of threat events, a probabilistic model of correlation between abnormal users (PMCAU) is proposed. The key of PMCAU is to analyze the behavior correlation between malicious users, calculate the behavior correlation degree of malicious users, determine the abnormal correlated user, and complete the probabilistic reasoning for users with different behavior correlation degree. $UM = \{um_i\}(i = 1, \ldots, n)$ represents a collection of malicious users, namely the zombie users, spam users, compromised users, etc. And $acu \in U$ represents an abnormal correlated user identified in the microblog. Here, the abnormal correlated user refers to a special group of junk users who publish specific information for specific content in the network, who are organized to publish, reply to, and forward blog posts or refer to others, to quickly spread bad and instructive error information. Not only let normal users cannot see the truth of the event but also will cause misleading to normal users, with adverse social consequences. The model framework is shown in Figure 1:

(1) *Data Layer*. Read the original data and preprocess the data. The user vector is constructed, and the valid user attribute information and user blog text information in the original data are selected.

(2) *Feature Layer*. Analyze user attribute information and behavior information based on user features. Extracting attribute features of the user's daily behavior based on the geographical distribution of the user's self-issued location, the number of followers, and personal data; the interaction behavior features of the user are extracted based on the occurrence object of the interaction behavior between users, the frequency of interaction, and the like.

(3) *Detection Layer*. Two algorithms are proposed to detect abnormal correlated users. The attribute similarity measurement algorithm calculates the attribute similarity between users' spontaneous behaviors, including three aspects: geographical similarity, follower's information, and personal information similarity. The behavioral interaction measurement algorithm constructs a user interaction degree formula, which indicates the strength of the user correlation process and is used to calculate the degree of correlation between user interaction behaviors.

(4) *Reasoning Layer*. The first-order logic grammar is used to express the causal connection before and after the abnormal behavior occurs and to reason the probabilistic of the connection and determine the abnormal user groups.

PMCAU detects potential malicious threats by determining abnormal correlated users *acu* and probabilistic reasoning of correlation degree between abnormal behaviors. Before the occurrence of a threat event, it can quickly mark abnormal user groups and find possible attack behaviors in advance to ensure the healthy and smooth operation of the network.

The abnormal behavior of users in the network seriously damages the network security. Due to the increasing diversification and concealment of user behaviors in microblog, PMCAU detects microblog users from the aspects of behavioral correlation analysis and abnormal correlated probabilistic reasoning. Behavioral correlation analysis is used to calculate the user's attribute similarity and behavioral interaction degree and to determine the abnormal correlated users *acu*; Abnormal correlation probabilistic reasoning is used to calculate the probability of future correlation between abnormal correlated users, and the probabilistic of occurrence of abnormal correlation is determined by the probabilistic, to discover potential groups threats.

4. Behavioral Correlation Analysis

Security of microblog users depends not only on their security but also is closely related to the behavior between users. By defining and calculating the behavioral correlation degree, users with a behavioral correlation greater than the correlation threshold φ are determined as abnormal correlated users *acu*.

Attribute similarity (AS) measures the similarity of three attributes, such as the user's geographical location, the number of user followers, and the integrity of personal information. Behavior interaction (BI) represents the interaction frequency of users, which are reflected by the interaction behavior between users. Behavior correlation degree (BC) represents the correlation strength of malicious user behaviors, which are obtained by calculating malicious user attribute similarity AS and behavior interaction BI. The calculation formula is shown in the following equation:

$$\mathrm{BC}\left(um_i, um_j\right) = \theta \mathrm{AS}\left(um_i, um_j\right) + (1 - \theta)\mathrm{BI}\left(um_i, um_j\right), \tag{1}$$

where θ represents the harmonic coefficient and the specific value is determined by the experiment. BC (u_i, u_j) represents the degree of correlation of users' behaviors.

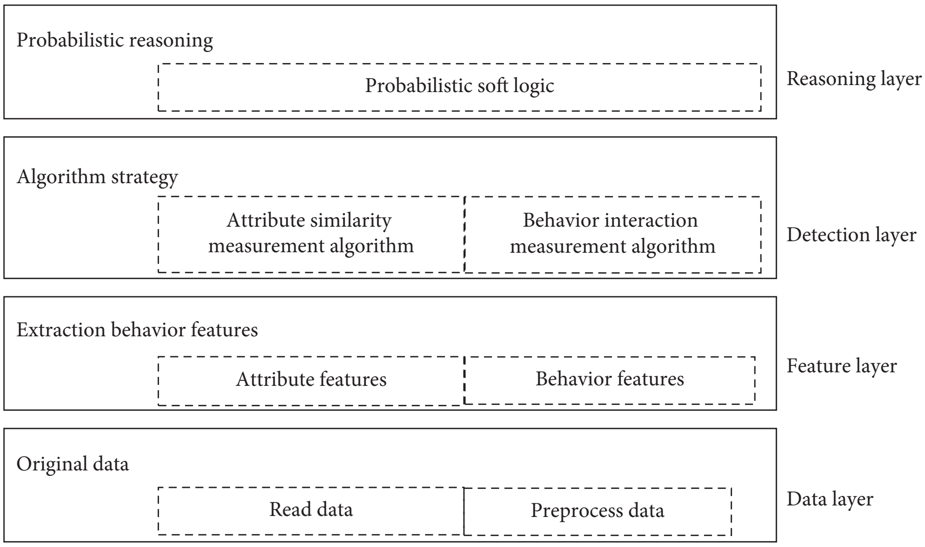

Figure 1: PMCAU framework.

4.1. Attribute Similarity. For malicious users in microblog, it is very important to study their geographical location, number of followers, and personal information. If malicious users living in the same area have a large number of overlapping users and similar personal information, it indicates that there may be some links between these malicious users. Jaccard coefficient is used to calculate the similarity of user attributes, and the relationship between users is obtained. The user attribute features are shown in Table 1.

The Jaccard coefficient is used to calculate the similarity between samples of a symbol metric or a Boolean metric and to compare similarities and differences between finite sample sets. Given two sets A and B, the similarity is measured by the ratio of the intersection of the two sets and the union. The larger the Jaccard coefficient, the higher the sample similarity. Since the Jaccard coefficient is suitable for calculating the similarity of discrete sets, the values of each element in its scoring matrix are expressed as 1 (with) and 0 (without) to determine whether there are common features between samples, which has a good calculation effect. The formula is shown in the following equation:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{2}$$

Users in microblog are affected by their attributes, such as users' geographical location, user followers' coincidence degree, and personal information, and their interactive relationship implies a certain real behavioral connection. The Jaccard coefficient was used to calculate the similarity of its three attributes of geographical location A_g, number of followers A_f, and personal information A_c and analyze the correlation strength of behaviors among malicious users. AS (um_i, um_j) is the attribute similarity of malicious users um_i and um_j, and the higher its value is, the higher the attribute similarity is; the formula is as shown in the following equations:

$$J_1\left(um_i\left(A_g\right), um_j\left(A_g\right)\right) = \frac{\left|um_i\left(A_g\right) \cap um_j\left(A_g\right)\right|}{\left|um_i\left(A_g\right) \cup um_j\left(A_g\right)\right|}, \tag{3}$$

$$J_2\left(um_i\left(A_f\right), um_j\left(A_f\right)\right) = \frac{\left|um_i\left(A_f\right) \cap um_j\left(A_f\right)\right|}{\left|um_i\left(A_f\right) \cup um_j\left(A_f\right)\right|}, \tag{4}$$

$$J_3\left(um_i\left(A_c\right), um_j\left(A_c\right)\right) = \frac{\left|um_i\left(A_c\right) \cap um_j\left(A_c\right)\right|}{\left|um_i\left(A_c\right) \cup um_j\left(A_c\right)\right|}, \tag{5}$$

$$\begin{aligned} \mathrm{AS}\left(um_i, um_j\right) = 1 - \Big[& J_1\left(um_i\left(A_g\right), um_j\left(A_g\right)\right) \\ & \times J_2\left(um_i\left(A_f\right), um_j\left(A_f\right)\right) \\ & \times J_3\left(um_i\left(A_f\right), um_j\left(A_f\right)\right)\Big]. \end{aligned} \tag{6}$$

4.2. Behavior Interaction. The interaction of microblog users mainly includes 7 behaviors, such as following, forwarding, commenting, thumbs up, collecting, mentioning @, and private letter. When calculating the behavioral interaction degree of malicious users, their login time and online time is different. Therefore, a time scale should be considered the maximum time malicious user interactions with other users of the interval rather than simply calculate the interaction between malicious users in a specified period.

Maximum time interval (MIT) represents the maximum time interval between the malicious user um_i and other users, and this is taken as the time label to measure the interaction frequency of malicious users.

In this regard, mutual behaviors of um_i and um_j in their respective MTI should be taken into consideration together to comprehensively analyze the behavioral interaction degree of um_i and um_j. Calculate the ratio between interaction frequency of um_i against um_j in MTI (um_i), and the interaction frequency of um_i against all malicious users in MTI

Table 1: User attribute features.

Feature symbol	Feature category	Feature name
A_g	User attribute	Geographical location
A_f	User attribute	Number of followers
A_c	User attribute	Personal information

(um_i, um_j). The ratio of um_i to um_j in MTI (um_j) to um_j to all malicious users in MTI (um_i, um_j) is calculated. And the two ratios are added together. The formula is shown in the following equation:

$$\mathrm{BI}\left(um_i, um_j\right) = \sum_{b_i} b_i\left(um_i,\ um_j\right), \tag{7}$$

where b_i represents 7 interactive behaviors among malicious users: follow, forward, comment, thumbs up, favorite, mention @, and private letter. BI (um_i, um_j) calculates the degree of interaction according to user interaction behavior and obtains the degree of association of behaviors sent by malicious users, as shown in the following equations:

$$\mathrm{inter}\left(um_i, um_j\right) = \frac{\mathrm{NBI}\left(um_i, um_j\right)/\mathrm{MTI}\left(um_i,\ um_j\right)}{\mathrm{NBI}\left(um_i\right)/\mathrm{MTI}\left(um_i\right)} + \frac{\mathrm{NBI}\left(um_i, um_j\right)/\mathrm{MTI}\left(um_i,\ um_j\right)}{\mathrm{NBI}\left(um_j\right)/\mathrm{MTI}\left(um_j\right)}, \tag{8}$$

$$b_i\left(um_i, um_j\right) = \frac{\mathrm{inter}\left(um_i, um_j\right)}{\mathrm{inter}_{\max}\left(um_i, um_j\right)}, \tag{9}$$

where NBI (um_i, um_j) represents the number of times the malicious user um_i interacts with um_j; NBI (um_i) and NBI (um_j) represent the total number of times that user um_i and um_j interact bi with all users. MTI (um_i, um_j) represents the maximum time interval in which a malicious user um_i interacts with um_j, MTI (um_i) and MTI (um_j), respectively, represent the maximum time interval for malicious user um_i and um_i to interact b_i with all users.

5. Abnormal Correlation Probabilistic Reasoning

Correlation of abnormal users involve a series of tedious processes such as behavior analysis, prediction, and judgment, and users have many uncertain factors such as incomplete login time, space, and behavior records. Using probabilistic soft logic (PSL) to express the cause and effect of behaviors, combining with many factors such as geographical location, behavioral information, interest preference, and user follower information, a set of reasoning rules which are more suitable to predict the possibility of correlation between abnormal behaviors are proposed and encoded in the PSL framework. PSL has efficient reasoning capabilities that allow the use of first-order logic to specify probabilistic models, which are an expression of if-then rules and support user groups.

5.1. PSL Grammar. Rule composition in PSL is shown in the following equation:

$$P_1\left(\alpha, \gamma\right) \wedge P_2\left(\gamma, \delta\right) \gg P_2\left(\alpha, \delta\right)\text{: weight.} \tag{10}$$

P_1 and P_2 are the predicates that define the relationship between random variables α, γ, and δ; "weight" represents the weight, representing the importance of each rule in reasoning; and ">>" represents the pointing relation of rule body to rule header in PSL [13].

For example, Relationship ($B(acu_i,\ acu_j)$, Fut) represents whether there is a relationship between $B(acu_i,\ acu_j)$ in Fut. Correlation ($B(acu_i,\ acu_j)$, Hist) represents that there will be a correlation between $B(acu_i,\ acu_j)$ in Hist. Based on the combination of the predicates Relationship and Correlation, the probability of a correlation between $B(acu_i,\ aca_i)$ in Fut can be expressed, where $B(acu_i,\ acu_j)$ represents the behavior of the exception correlated user acu_i to acu_j; Fut represents the future period; and Hist represents the historical period.

5.2. Rule Construction. In general, the probabilistic reasoning model uses explicit correlative reasoning and uses Boolean values {0, 1} to determine whether the two are related; "1" for correlation and "0" for no correlation. For example, if it is known that abnormal correlated users acu_i and acu_j have a correlation relationship in the historical period, then the relationship between acu_i and acu_j is still judged as "1" in the future period, that is, there is a correlation relationship in the future period. However, in reality, due to the differences in information such as microblog users' location, active time, interactive behavior, mutual attention, and mutual attention, the accuracy of reasoning is low if only relying on simple explicit correlation rules. Therefore, it is necessary to optimize the inference rules to carry out implicit correlation reasoning. This kind of implicit correlation inference rules make the expression of inference results more in line with the actual situation and have higher accuracy. According to each rule obtained by weight learning, confidence is allocated as weight built in PSL. The rules are as follows.

The more the same locations are located, the greater the likelihood of correlation will be as follows:

$$\begin{aligned}&\text{position}\left(B\left(acu_i, acu_j\right),\ Po\right) \wedge \text{position judge}\left(Po\right)\\&\quad\Longrightarrow \text{strong correlation}\left(B\left(acu_i, acu_j\right)\right).\end{aligned} \tag{11}$$

The higher the coincidence degree of online active time, the higher the possibility of correlation as follows:

$$\begin{aligned}&\text{time coincidence}\left(B\left(acu_i, acu_j\right), \mathrm{TC}\right) \wedge \text{time coincidenc judge}\left(\mathrm{TC}\right)\\&\quad\Longrightarrow \text{strong correlation}\left(B\left(acu_i, acu_j\right)\right).\end{aligned} \tag{12}$$

The higher the frequency of user interaction, the greater the likelihood of correlation:

$$\begin{aligned}&\text{behavior interaction}\left(B\left(acu_i, acu_j\right), \mathrm{BI}\right) \wedge \text{behavior interaction judge}\left(\mathrm{BI}\right)\\&\quad\Longrightarrow \text{strong correlation}\left(B\left(acu_i, acu_j\right)\right).\end{aligned} \tag{13}$$

The more users with the mutual concern, the greater the possibility of correlation:

$$\text{mutual concern}\left(B\left(acu_i, acu_j\right), \text{SF}\right) \wedge \text{mutual concernjudge}\,(\text{SF}) \Longrightarrow \text{strong correlation}\left(B\left(acu_i, acu_j\right)\right). \tag{14}$$

Focus on each other can create correlation:

$$\text{focus on each other}\left(B\left(acu_i, acu_j\right), \text{ MA}\right) \wedge \text{focus on each other}\,(\text{MA}) \Longrightarrow \text{strong correlation}\left(B\left(acu_i, acu_j\right)\right). \tag{15}$$

5.3. Weight Learning. PSL provides maximum probabilistic inference [13] to infer the most likely probabilistic of atoms in logical rules from existing data. Since the continuous value between [0, 1] is used for the probabilistic, MPE reasoning is transformed into a convex optimization process to find the optimal solution.

For the rules learned, the weight of each rule is assigned according to the confidence of each rule. For example, when the correlation between abnormal behaviors is inferred, there are three rules with confidence levels of 0.7, 0.8, and 1.0. When converting them into PSL rules, the confidence can be multiplied by multiple times as its weight. However, for manually defined rules, weight learning is required. In weight learning, the maximum likelihood estimation method [14] is used, and the gradient function is used to perform weight estimation. The formula is as shown in the following equation:

$$\frac{\partial}{\partial \lambda_w} \log_a f\,(Q) = -\sum_{r \in R_w} \left(D_r\,(Q)\right)^P + E\left[\sum_{r \in R_w} D_r\,(Q)\right], \tag{16}$$

where R_w represents all logical rules with weight λw being initialized, $E[\sum_{r \in R_w} D_r\,(Q)]$ is replaced by a $\sum_{r \in R_w} D_r\,(Q^*)$ approximation, and Q^* is the most likely correct interpretation of the atom.

5.4. Probabilistic Reasoning. PSL is different from other kinds of probabilistic models and is the feature of closed atoms using soft constraints, namely closed atomic probabilistic values is a continuum of values between [0, 1], rather than Boolean value {0, 1}, the record for $Q(r)$, usually adopt Lukasiewicz logic [13] the conjunction (∧), disjunction (∨), and negative (¬) as a logical connective to calculate $Q(r)$. The formulas are as shown in the following equations:

$$Q\left(l_1 \wedge l_2\right) = \max\{Q\left(l_1\right) + Q\left(l_2\right) - 1, 0\}, \tag{17}$$

$$Q\left(l_1 \vee l_2\right) = \min\{Q\left(l_1\right) + Q\left(l_2\right), 1\}, \tag{18}$$

$$Q\left(l_2\right) = 1 - Q\left(l_1\right). \tag{19}$$

A rule r in PSL can be described as $r$body $\longrightarrow$ $r$head. This rule is satisfied when $Q(r\text{body}) \leq Q(r\text{head})$ or $Q(r) = 1$. Otherwise, the degree of satisfaction of the logic rule is measured by calculating the distance satisfaction $D(r)$, as in the following equation:

$$D(r) = \max\{0, Q(r\text{body}) - Q(r\text{head})\}. \tag{20}$$

PSL defines the probabilistic value of the probabilistic distribution for all closed atoms, as in equation (21), and defines the probabilistic value of the closed atom as the distance satisfaction between the highest probabilistic value and the lowest probabilistic value, that is, the probabilistic satisfies all logic rules:

$$p(Q) = \frac{1}{Z} \exp\left\{-\sum_{r \in R} \lambda_r D(r)\right\}, \tag{21}$$

$$Z = \int_Q \exp\left\{-\sum_{r \in R} \lambda_r D(r)\right\}, \tag{22}$$

where Z is the normalization constant, λ_r is the weight of the rule r, and R is the set of all rules.

The PSL model input is the abnormal behavior data of malicious users, the input data set is used to initialize the logic rules in the PSL model, and the weight learning is performed. Then, distance satisfaction is defined in the PSL model, and the probabilistic of meeting the logic rules initialized every day is calculated. Finally, MPE reasoning mechanism is applied to calculate the probabilistic correlation between abnormal behaviors.

6. Experiments

6.1. Experimental Environment and Data. The environment used in the experiment was Intel(R) Core(TM) i5-7300HQ CPU @2.50 GHz, 8 GB of memory, the operating system is Windows 10, and model code is based on c++ implementation.

The data set published in literature [15] was used to verify the feasibility of the model. The data set contains 1,787,443 microblog user data, and each of the user data includes basic information of the users (such as user ID, gender, number of followers, number of fans, etc.), 1000 microblogs newly released by each user, and user interaction behavior data. Among them, there are nearly 4 billion relationships of mutual concern among users, and each user has an average of 200 followers. Due to a large amount of data in the data set, 10 groups are randomly selected from the data set, each group has 10,000 pieces of user data, and each piece of user data includes the basic information of the user, the newly published blog content, and user interaction behavior data, which is recorded as "Data1," "Data2," "Data3," "Data4," "Data5," "Data6," "Data7," "Data8," "Data9," and "Data10."

6.2. Evaluation Index. To solve the data imbalance problem, confusion matrix analysis experimental results are established [16]. In the matrix, TP stands for the number of users that are originally abnormal correlated users and are judged to be abnormal correlated users during detection; FN stands for the number of users that are originally abnormal correlated users but are judged to be nonabnormal correlated users during detection; FP stands for the number of users that are originally nonabnormal correlated users but are judged to be abnormal correlated users during detection; and TN stands for the number of users that are originally nonabnormal correlated users and are judged to be nonabnormal correlated users during detection, as shown in Table 2.

To evaluate the performance of PMCAU, three evaluation indexes, namely, precision rate (Pre), recall rate (Rec), and harmonic mean value $F1$_score, were selected. Among them, the precision rate and recall rate were used to evaluate the accuracy of the experiment, and the harmonic mean value $F1$_score was used to evaluate the comprehensive performance of the experiment. Pre refers to the proportion of the number of correctly identified abnormal correlated users of all detected abnormal correlated users. Rec refers to the proportion of correctly identified abnormal correlated users in the total number of truly abnormal correlated users. $F1$_score is the harmonic mean value of precision rate and recall rate, and the equation is as follows:

$$\begin{aligned}\mathrm{Pre} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\\ \mathrm{Rec} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\ F1\_\mathrm{score} &= 2\cdot\frac{\mathrm{Pre}\cdot\mathrm{Rec}}{\mathrm{Pre}+\mathrm{Rec}}.\end{aligned} \tag{23}$$

6.3. Parameter Settings. Parameters involved in the experiment include the linear coefficient θ (used to reconcile the user's attribute similarity and behavioral interaction) and the correlation threshold φ; their values are determined experimentally, and its value was determined by referring to the performance of the model evaluated utilizing harmonic mean $F1$_score.

For malicious users, the definition of their abnormal correlated users is the key to detect abnormal groups in microblog. Linear coefficient θ and correlation threshold φ are directly related to the detection of abnormal correlated users. According to formula (9), different linear coefficients θ correspond to different correlation degree thresholds. $F1$_score values of correlation threshold φ under different values of linear coefficient θ were calculated, respectively, as shown in Figure 2. According to experience, the degree of behavioral interaction between malicious users has a greater impact than the degree of attribute similarity. Therefore, the value range of the linear coefficient θ is set as [0.15, 0.55] and the value range of the correlation threshold is set as [0.1, 0.9].

Figure 2 represents different $F1$_score with the value range of linear coefficient θ set as [0.15, 0.55] and the correlation threshold φ with a value range of [0.1, 0.9]; when the linear coefficient $\theta = 0.35$ and the correlation threshold $\varphi = 0.7$, $F1$_score is the largest. Therefore, it can be considered that in the calculation of behavioral correlation degree, when the linear coefficient θ is set at 0.35, the correlation threshold φ for defining abnormal behaviors is set at 0.7. That is, when users with a behavior correlation degree greater than 0.7 in malicious users are defined as abnormal correlated users, PMCAU has better detection performance.

The distribution of user behavior correlation degree and abnormal correlation probabilistic are shown in Figures 3 and 4. After the parameter setting, the behavioral correlation degree of abnormal correlated users is distributed in [0.7, 1.0] and that of nonabnormally correlated users is distributed in [0.2, 0.6]. Meanwhile, when reasoning the abnormal correlated probabilistic of users, the correlation probabilistic of abnormal correlated users is [0.8, 1.0] and that of nonabnormal correlated users is [0.5, 0.7]. The above indicates that when the linear coefficient θ is set to 0.35 and the correlation threshold φ for defining abnormal behaviors is set to 0.7, PMCAU can better classify abnormal correlated users and nonabnormal correlated users to find potential abnormal groups in microblog.

6.4. Experimental Analysis. To test the performance of PMCAU in detecting abnormal correlated users and probabilistic reasoning of possible associations between abnormal correlated users, a comparative experiment was set up. DBSCAN-based clustering algorithm and PageRank-based abnormal detection algorithm of microblog account were compared with PMCAU. The detection performance of the three algorithms was compared by corresponding indexes of the experiment.

The DBSCAN-based clustering algorithm is an anomaly detection method based on density clustering, which can find abnormal points while clustering. The PageRank-based microblog account anomaly detection algorithm constructs a social relationship matrix according to the user relationships and ranks the account by iteratively calculating the PageRank value to detect malicious users. The two algorithms have good results in user anomaly detection, so the above two algorithms are selected for comparison experiments with PMCAU. PMCAU is divided into two parts: behavioral correlation analysis and abnormal correlation probabilistic reasoning. Behavioral correlation analysis detects and determines abnormal correlated users among malicious users by calculating their attribute similarity and behavioral interaction. Abnormal correlation probabilistic reasoning is used to analyze and predict the abnormal correlated users to get their correlation probabilistic.

Using these three algorithms, 10 groups of experiments were conducted on the data set of "data1–data10" in turn, which was recorded as "G1–G10." Pre, Rec, and $F1$_score were used as the evaluation criteria of the experiment, and the experimental results are shown in Figure 5.

The results show that the DBSCAN-based clustering algorithm clustering by calculating the Euclidean distance of each data point, because the calculation is numerical attributes, does not consider the user published blog content and other text-based information, and user interaction behavior and other behavioral information, so the Pre of detection is low. Although the PageRank-based abnormal detection algorithm for microblog account has a high Pre, it

Table 2: Symbol description.

Detection result	Actual situation	
	Abnormal correlated users	Nonabnormal correlated users
Abnormal correlated users	TP	FP
Nonabnormal correlated users	FN	TN

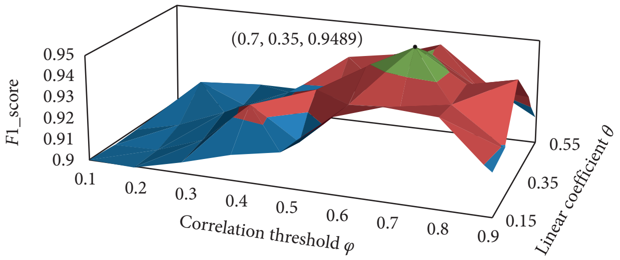

Figure 2: $F1$_score value under different parameters.

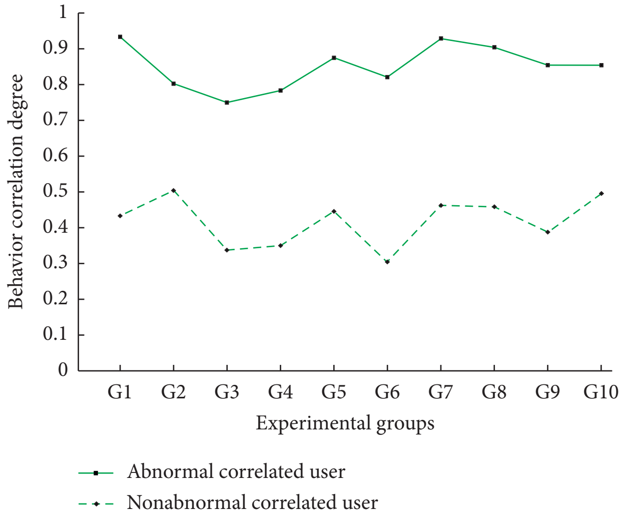

Figure 3: User behavior correlation degree distribution.

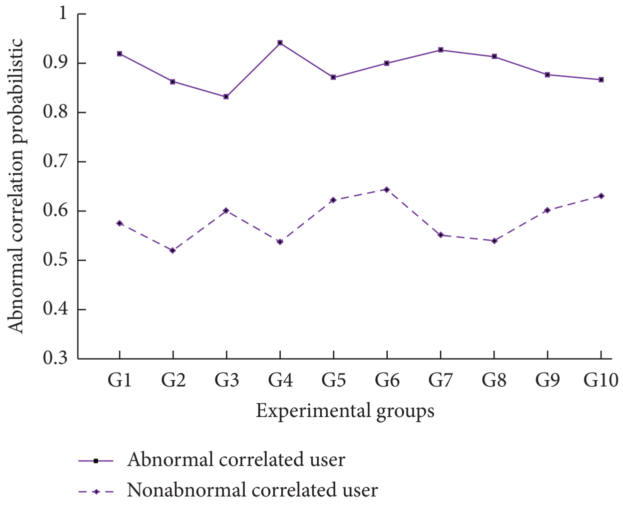

Figure 4: User abnormal correlation probabilistic.

only considers the social relationship between users and does not consider the attributes and features of users, and the ranking results of users are affected by a time delay, so the Rec and $F1$_score are lower.

When detecting abnormal correlated users and inferring abnormal correlation probabilities, PMCAU considers the interaction behavior between users and determines the abnormal correlation users by comprehensive numerical attributes, text type information, and interactive behaviors. Probabilistic soft logic reasoning is correlated with the probabilistic of occurrence of correlation behavior, so it has high Pre, Rec, and $F1$_score.

In order to check the stability of PMCAU, the average and variance of 10 groups of experimental results corresponding to the three algorithms were compared. The experimental results are shown in Figure 6.

It can be observed in Figure 6 that the 10 sets of experiments corresponding to the three algorithms are compared in terms of Pre, Rec, and $F1$_score. Among them, the average value of the three indexes of DBSCAN clustering algorithm is medium. In PageRank ranking algorithm, although the average Pre is 90%, its Rec is low, and the overall performance of the algorithm is medium. The Pre, Rec, and $F1$_score of the 10 groups of experiments corresponding to PMCAU are highest, and the model performance is outstanding, with the Pre reaching 97.35%.

As can be seen from Figure 7, the variance of DBSCAN clustering algorithm and of PageRank ranking algorithm on the four experimental evaluation indexes are large, indicating that the experimental results of the above two algorithms fluctuate greatly in the 10 groups of experiments respectively and the stability of the algorithm is poor. The variance of PMCAU corresponding to the three kinds of experimental evaluation indexes is small, indicating that the experimental results of the corresponding 10 groups fluctuate less and the algorithm is stable.

According to the average and variance of 10 groups of experimental results corresponding to each of the three algorithms, compared with the other three algorithms, PMCAU has good stability and adaptability when it comes to detecting abnormal correlated users of microblog and reasoning abnormal correlation probabilistic while guaranteeing accuracy.

6.5. Reasoning Performance Analysis. To compare the performance difference between PMCAU and the advanced reasoning models in existing research studies, a comparison experiment based on the Markov logic network (MLN) and PMCAU was set up to analyze the reasoning ability of two algorithms for abnormal correlation probabilistic.

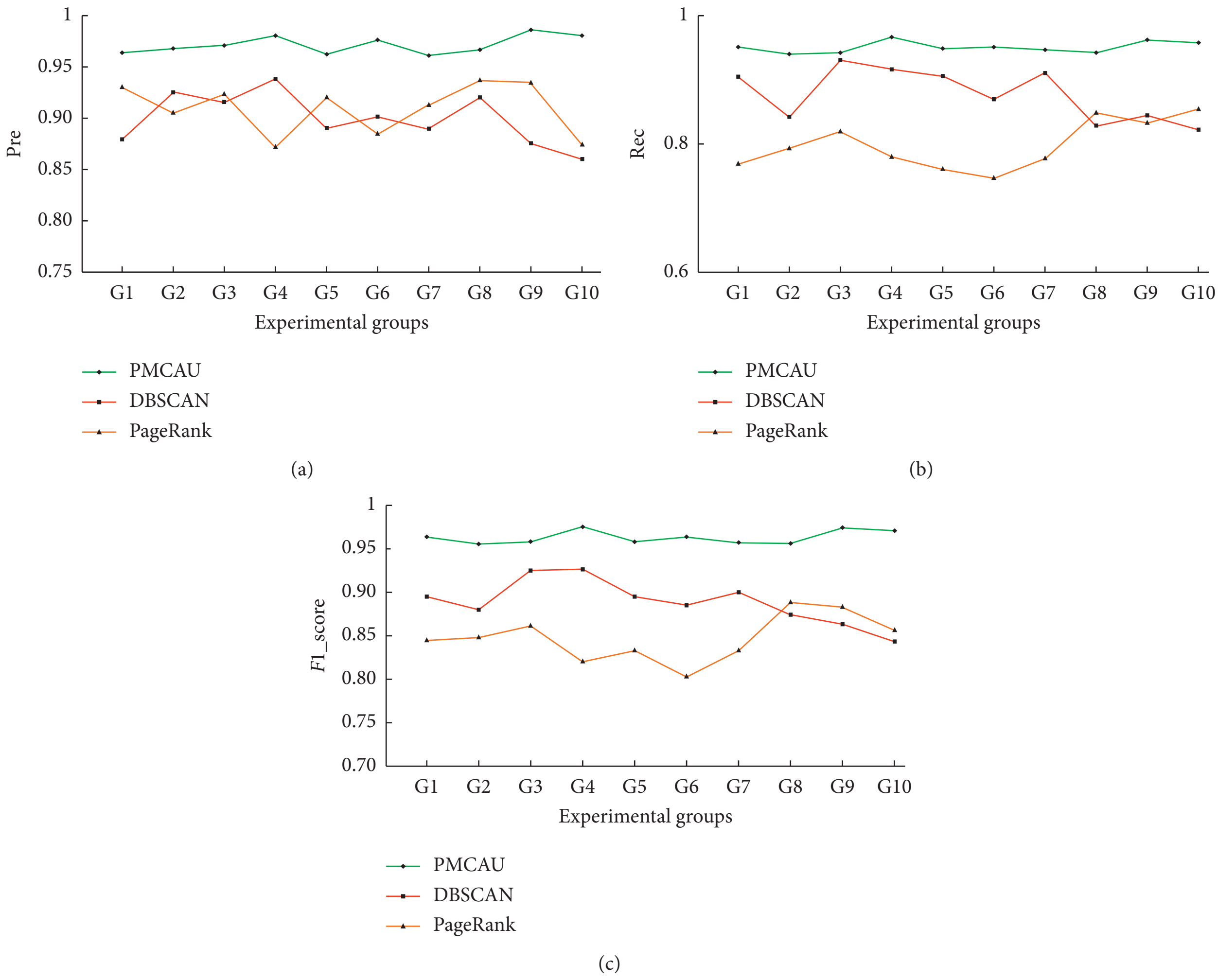

Figure 5: (a) Precision rate. (b) Recall rate. (c) *F*1_score.

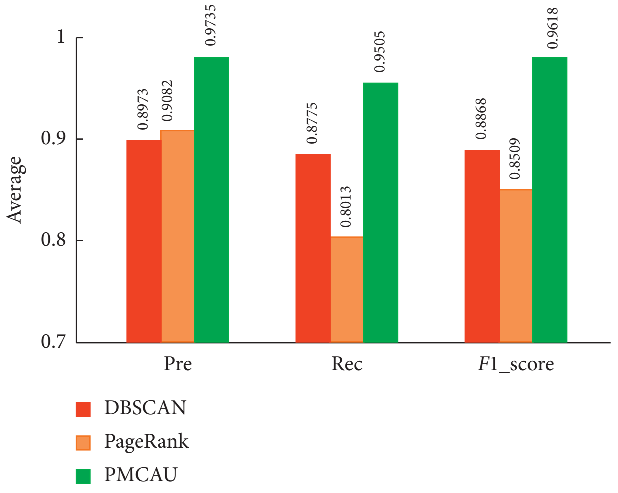

Figure 6: Average.

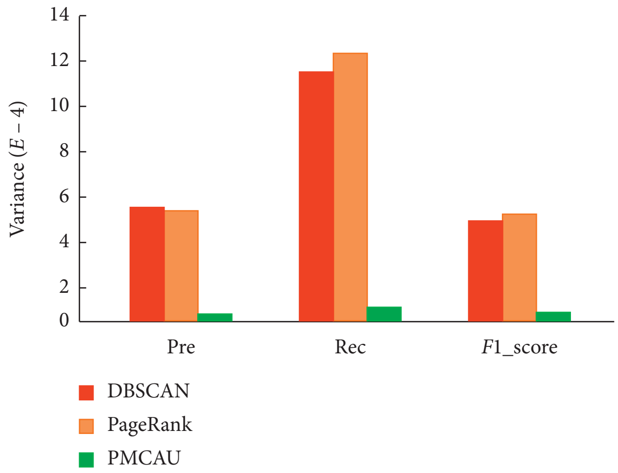

Figure 7: Variance.

Similarly, data groups "Data1–Data10" were taken as data samples for 10 groups of experiments, denoted as "G1–G10." MLN with the same logical reasoning ability was used for the experiment, and the experimental results were compared with PMCAU. During the experiment, 80% of user data of each group of samples were selected to train PMCAU and MLN models, and 20% of user data were taken as test samples. With *F*1_score as the evaluation index, experimental results are shown in Figure 8.

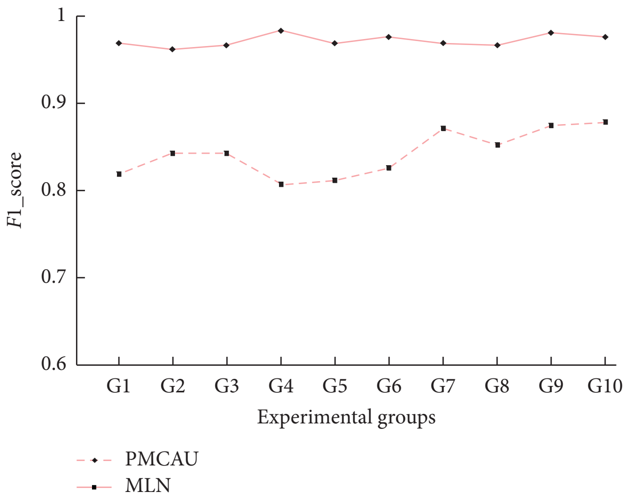

Figure 8: Comparison of *F*1_score between MLN and PMCAU.

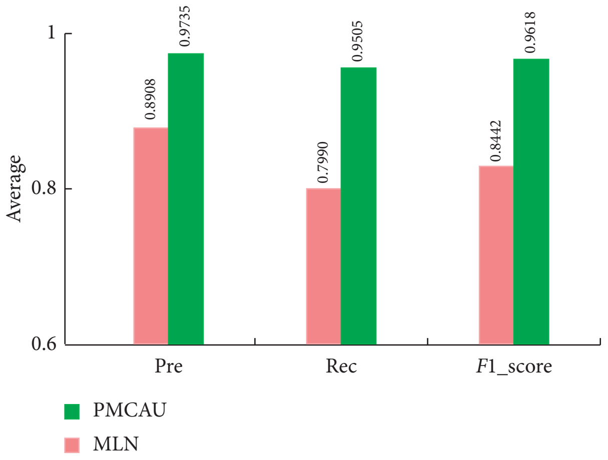

Figure 9: Average of MLN and PMCAU.

The results show that when MLN is used to infer the abnormal correlation probabilistic between abnormal correlated users in microblog, the Boolean value {0, 1} is used to determine whether there is an abnormal correlation, and the complex relationship between users in microblog cannot be better expressed, the corresponding *F*1_score in [0.8, 0.9]. PMCAU uses PSL reasoning—the use of [0, 1] between continuous soft true value as operation values—which makes the reasoning become a continuous convex optimization problem and can effectively express the cause and effect of user behavior and overcome the user's existence time. The limitations of factors such as space and incomplete records, the corresponding *F*1_score is higher, and *F*1_score is stable at [0.9, 1].

The average and variance of the 10 sets of experiments corresponding to MLN and PMCAU are shown in Figures 9 and 10. Among them, although MLN can perform probabilistic reasoning on users' complex relationships to a certain extent, its Pre, Rec, and *F*1_score evaluation standards are all lower than that of PMCAU. Experiments show that PMCAU has better logical reasoning ability and is more suitable for predicting the abnormal correlation probabilistic of abnormal correlated users in a microblog network and identifying potential abnormal user groups under the premise of ensuring accuracy.

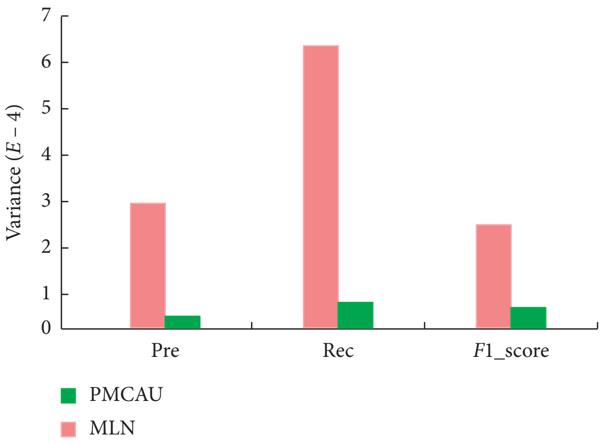

Figure 10: Variance of MLN and PMCAU.

7. Conclusion

Aiming at the problem of abnormal user group detection in the IoT, a PMCAU model for abnormal user correlation probabilistic inference is proposed. First, the concept of user behavior correlation degree is proposed for malicious user nodes in the IoT. The similarity measurement algorithm is applied to construct the correlation measurement algorithm. Attribute similarity and behavior interaction were calculated respectively, and user behavior interaction was determined to determine *acu*. Second, probabilistic soft logic is used to express the causal relationship among abnormal user groups. Combined with geographical location, behavioral information, interest preference, and user follower information, a set of reasoning rules more suitable for predicting the possibility of abnormal correlation with users is proposed, and the possibility of abnormal correlation is predicted to determine the potential abnormal group nodes in the IoT.

The experimental results show that PMCAU has high detection accuracy and beneficial stability, and it has a good effect on the correlation probabilistic inference prediction between abnormal behaviors. In the future, the behavior of abnormal correlated users in the IoT will be specifically analyzed, the types of abnormal behaviors will be determined, and the tracking of abnormal behaviors will be further discussed based on the medical IoT.

Data Availability

The data came from an article [15] by Zhang Jing of Tsinghua University, in which crawlers were used to construct a data set of microblog users. The microblogging network they used in this study was crawled from Sina Weibo.com, which, similar to Twitter, allows users to follow each other. Particularly, when user A follows B, B's activities such as tweet and retweet will be visible to A. A can then choose to retweet a microblog that was tweeted (or retweeted) by B. User A is also called the follower of B, and B is called the followee of A. After crawling the network structure, for each one in the 1,787,443 core users, the crawler collected her 1,000 most recent microblogs. At the end of the crawling, they produced in total 4 billion following relationships among them, with average 200 followees per user.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant no. 2017YFB0802300.

References

[1] K. Fan, S. Sun, Z. Yan, Q. Pan, H. Li, and Y. Yang, "A blockchain-based clock synchronization scheme in IoT," *Future Generation Computer Systems*, vol. 101, pp. 524–533, 2019.

[2] K. Fan, W. Jiang, L. Qi, H. Li, and Y. Yang, "Cloud-based RFID mutual authentication scheme for efficient privacy preserving in IoV," *Journal of the Franklin Institute*, 2019.

[3] Z. Ding, Y. Jia, and B. Zhou, "Survey of data mining for microblogs," *Journal of Computer Research and Development*, vol. 51, no. 4, pp. 691–706, 2014.

[4] Y. Zhou, "Behavior analysis based traffic anomaly detection and correlation analysis for communication networks," Doctoral dissertation University of Electronic Science and Technology of China, Chengdu, China, 2013.

[5] J. Zhao, H. Huang, and S. Tian, "Protocol anomaly detection based on hidden Markov model," *Journal of Computer Research and Development*, vol. 47, no. 4, pp. 621–627, 2010.

[6] X. Xiao and Q. Zhai, "Masqerade detection based on shell commands and high-order Markov chain models," *Acta Electronica Sinica*, vol. 39, no. 5, pp. 1199–1204, 2011.

[7] J. Cao and J. Wu, "Sina microblog information diffusion analysis and prediction," *Chinese Journal of Computers*, vol. 37, no. 4, pp. 779–790, 2014.

[8] Z. H. Xu, Y. Zhang, Y. Wu, and Q. Yang, "Modeling user posting behavior on social media," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Portland, OR, USA, August 2012.

[9] S. Li, X. Li, and H. Yang, "A zombie account detection method in microblog based on the pagerank," in *Proceedings of the 2017 IEEE International Conference on Software Quality, Reliability and Security Companion*, Prague, Czech Republic, July 2017.

[10] C.-F. Xu, C.-L. Hao, B.-J. Su, and J.-J. Lou, "Research on Markov logic networks," *Journal of Software*, vol. 22, no. 8, pp. 1699–1713, 2011.

[11] C. Sun, Z. Yan, Q. Li, Y. Zheng, X. Lu, and L. Cui, "Abnormal group-based joint medical fraud detection," *IEEE Access*, vol. 7, pp. 13589–13596, 2018.

[12] Z. Yang, Q. Sun, and B. Zhang, "Evaluating prediction error for anomaly detection by exploiting matrix factorization in rating systems," *IEEE Access*, vol. 6, pp. 50014–50029, 2018.

[13] A. Kimmig, S. Bach, and M. Broecheler, "A short introduction to probabilistic soft logic," in *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, Lake Tahoe, NV, USA, 2012.

[14] S. Tomkins, J. Pujara, and L. Getoor, "Disambiguating energy disaggregation: a collective probabilistic approach," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 2857–2863, Melbourne, Australia, August 2017.

[15] J. Zhang, J. Tang, J. Li, Y. Liu, and C. Xing, "Who influenced you? Predicting retweet via social influence locality," *Acm Transactions on Knowledge Discovery from Data*, vol. 9, no. 3, pp. 1–26, 2015.

[16] M. Yang, J. M. Yin, and G. L. Ji, "Classification methods on imbalanced data: a survey," *Journal of Nanjing Normal University (Engineering and Technology Edition)*, vol. 8, no. 4, pp. 7–12, 2008.