A new security model in p2p network based on Rough set and Bayesian learner

Wang Hai-Sheng[1] and Gui Xiao-Lin[2]
[1] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2] School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China
[e-mail: haisheng.wang@yahoo.cn]
*Corresponding author: Wang Hai-Sheng

Received April 17, 2012; revised July 21, 2012; accepted August 16, 2012; published September 26, 2012

Abstract

A new security management model based on Rough set and Bayesian learner is proposed in the paper. The model focuses on finding out malicious nodes and getting them under control. The degree of dissatisfaction (DoD) is defined as the probability that a node belongs to the malicious node set. Based on transaction history records local DoD (LDoD) is calculated. And recommended DoD (RDoD) is calculated based on feedbacks on recommendations (FBRs). According to the DoD, nodes are classified and controlled. In order to improve computation accuracy and efficiency of the probability, we employ Rough set combined with Bayesian learner. For the reason that in some cases, the corresponding probability result can be determined according to only one or two attribute values, the Rough set module is used; And in other cases, the probability is computed by Bayesian learner. Compared with the existing trust model, the simulation results demonstrate that the model can obtain higher examination rate of malicious nodes and achieve the higher transaction success rate.

Keywords: Degree of Dissatisfaction; Rough set; Bayesian learner; Information entropy

This work is supported by the National Natural Science Foundation of China under Grant No. 60873071
http://dx.doi.org/10.3837/tiis.2012.09.023

1. Introduction

In the social relation network, an individual in a transaction always hopes to choose trust objects, in particular, hopes to avoid encountering malicious objects. For instance, the banking system usually builds blacklists for malicious or dishonest customers, in order to control the unsatisfactory behaviors. The proposed model can help users in P2P network environment to select correctly the transaction object and to avoid malicious nodes.

Transaction history records are used as the training sample set. Through the learning from the training sample set, the Bayesian classifier is produced. Using the same training sample set as that used by the Bayesian classifier, without any extra apriori information, the Rough set can be trained to produce some precise, verifiable classification rules, i.e. the Rough set classifier. When we conduct the classification of a node, first, the Rough set classifier is used, according to the value of an attribute or values of two attributes the classification of the node can quickly be determined. If the classification of the node using the Rough set classifier is ended in failure (e.g., there are unrecognized conditions), the Bayesian classifier will be used. Using Rough set combined with Bayesian classifier the computation accuracy and the computation efficiency of the probability is improved.

According to the DoD, trading nodes are classified as trust nodes, strange nodes or malicious nodes. Trusted nodes are those that frequently trade with the local node with higher success rate and without malicious attack records. In the trusted node list, those nodes are classified as “trusted neighbor nodes” that are the most familiar with the local node. Malicious nodes are those that "the serious damage" occurred many times in transaction records. "The trusted node list" and "the malicious node list" are shared among trusted neighbor nodes.

When the local node needs to trade with a strange node, it will submit a request to other nodes and ask for them to evaluate the strange node[1,2], and the Rough set module and Bayesian learner will be called to make comprehensive evaluation of recommendations and to determine whether the strange node can be traded with.

Shannon took the thermodynamic entropy into the information theory and put forward the information entropy that was used to measuring the degree of the confusion of the information system. The more orderly an information system is, the less the information entropy is. Sometimes recommendations may be very centralized, but sometimes may be quite decentralized. Obviously, when a set of recommendations is centralized, its credibility will be high, and its information entropy is smaller. Otherwise, when a set of recommendations is decentralized, the set of recommendations will be with low credibility, and its information entropy is larger. Obviously, the credibility of a set of recommendations is inversely proportional to the information entropy. We have proposed the computational method of the information entropy of a set of recommendations. The calculation of the information entropy of a set of recommendations is equal to the calculation of its dispersion. The credibility of a set of recommendations is calculated based on the information entropy.

Through simulation experiments, the performance of the model was analyzed. In experiments, once a serious failure event happened, the corresponding transaction record would be updated immediately, and the Rough set module and Bayesian learner would be called to judge whether the node would be reclassified due to the failure event. If a malicious node was detected, the node would be immediately added to the malicious node list.

The rest of the paper is organized as follows: the related work is introduced in section 2. In section 3, the security model based on Rough set and Bayesian learner is described. Section 4 presents simulations and results analysis, and the conclusion is made in section 5.

2. Related Work

In peer to peer network environment, all nodes were connected directly to exchange data and services, and the system had properties of anonymity, dynamic and openness. In the environment, there existed malicious or selfish users and a lot of safety hazards [3][4]. There existed a variety of malicious behaviors, such as to provide malicious services strategically or to submit false evaluation (the false trust data).About the trust mechanism following models and algorithms were existed:

(1)EigenTrust and PowerTrust algorithms. In EigenTrust [5] algorithm, Global trust value was calculated by the direct trust value. The node with higher direct trust value had more credible, and bigger recommendation weight value was given. PowerTrust[6] improved EigenTrust from two aspects:

a) It proposed there existed a small proportion of the Power nodes whose evaluation weights were significantly higher than others. Through the Power nodes the trust node set was established.

b) It proposed the strategy of look forward random walk (LRW) to upgrade the speed of matrix iterative convergence.

The disadvantage of PowerTrust was that:

a) The size of transactions was not taken into account by the algorithm, which could allow a malicious user in small transactions to accumulate trust, and in large transactions to cheat;

(b) Malicious users were not punished.

(2)PeerTrust and GossipTrust algorithms [7][8]. PeerTrust algorithm had the following advantages:

a) It calculated the node's direct trust with the feedback, the influence of malicious nodes in the trust calculation could be reduced by evaluating the credibility of information;

b) collusion attacks of malicious nodes could be prevented by measuring the credibility of the evaluation according to the similarity of evaluations.

The GossipTrust model achieved fast gossip-based reputation aggregation algorithms and efficient reputation storage with Bloom filters.

(3)The fuzzy logic based algorithms [9][10]. They had the following advantages:

a) The uncertain evaluation was exploited;

b) The credibility of the recommendation was adjusted dynamically;

c) The punishment to malicious recommendations was considered;

d) A fuzzy logic-based reasoning process was presented.

Such algorithms had the following disadvantages:

a) When choosing the membership function the effectiveness could not be evaluated;

b) It was difficult for the effectiveness of the algorithm to be verified by simulation or experiment.

(4) The model based on reputation and risk evaluation [11]. The uncertainty between trust relationships and risk factors was considered in the model. In the paper [12], author presented a security management model based on group. In the model based on group, by using direct trust degree, group trust degree, trust degree between groups and multiple control factors the global trust degree of a node was computed.

In above mentioned models, they employed two concepts, namely, "reputation" and "trust", and failure events of transactions had not been classified, only the number of failed transactions was counted, so failure events could not been categorized into different levels by the severity. In the paper, unsatisfactory transaction events were quantified, categorized and managed according to the type and the severity of the damage. The proposed model in the paper employed the concept of "DoD", and we proposed a new security model based on the degree of dissatisfaction (DoD).

3. The Security Model Based On Rough Set And Bayesian Learner

3.1 The Classification Of Unsatisfactory Transaction Events

The model in the paper is a security management model in P2P network. A p2p file sharing application is used here as an example to describe the proposed approach. In the application, each node needs to play two roles: one, as a document provider to provide documents to other nodes; the other, as a document consumer to apply the documents provided by other nodes[13][14]. In the application, a document is found with the aid of the search engine. In most situations the user may receive a provider list in which normal nodes and malicious ones are included. If a malicious node is chosen, false documents or infected files will be provided, the time and the energy will be wasted, and even the user's computer system will be damaged. According to the type of unsatisfactory transaction events and the severity of the damage, unsatisfactory transaction events among the nodes are categorized, quantified and controlled. Unsatisfactory transaction events are divided into two categories: the serious damage and the general unsatisfactory event. The serious damage includes: 1) malicious attack files are downloaded, 2) there is the trading on a large scale with a very low quality (malicious or selfish nodes deliberately provide false documents and sizes of documents are equal to or greater than 5mb and divided into several sections: 5mb =< size =< 25mb, 25mb < size =< 100mb, 100mb < size);3) malicious feedback. The general unsatisfactory event includes: 1) there is the trading on a small scale with a low quality (the qualities of provided files are poor and sizes of files are less than 5mb and divided into several sections: size =< 500kb, 500kb< size =< 1mb, 1mb < size < 5mb); 2) download speed is too low or often offline; 3) the feedback is accidental error.

Trading nodes are classified as trusted nodes, malicious nodes or strange nodes. Here, C_{trust} stands for the trusted node set, C_{virus} stands for the malicious node set and $C_{stranger}$ stands for the strange node set.

Definition 1: The degree of dissatisfaction (DoD) is defined as the probability $P(C_{virus} \mid X_{ei})$ that the node X_{ei} belongs to the malicious node set.

Definition 2: The degree of satisfaction (DoS) is defined as the probability $P(C_{trust} \mid X_{ei})$ that the node X_{ei} belongs to the trusted node set.

3.2 The Training And The Applying Of Bayesian Learner

Suppose that $\boldsymbol{W}$ is the set for all finite or infinite observation objects in the given world scope, F: $\boldsymbol{X}\rightarrow\{0,1\}$ $\boldsymbol{X}\in\boldsymbol{W}$, which is the model for the world[15][16]. As a result of the limitation of observation ability, we can only obtain a finite subset of this world, $<<x_1,d_1,> \ldots <x_m, d_m>>$, where x_i is an example in $\boldsymbol{X}$, d_i is the goal value for x_i, that $d_i\in\{0,1\}$, which is called a sample set. The machine learning is based on the sample set to calculate the model on the world, f: $\boldsymbol{X}\rightarrow\{0,1\}$, which makes f to be an approximate of F ,that $d_i=f(x_i)$.

Here Bayesian learner is used to seek the classification supposition $h\in\boldsymbol{H}$ which has the biggest possibility in the candidate classification supposition set $\boldsymbol{H}$ when data set $\boldsymbol{D}$ is given, namely to seek maximum a posterior (MAP) supposition. Through calculating the posterior probability for each candidate supposition with Bayesian formula, when formula (1) is met, and h_{MAP} is the MAP supposition:

$$h_{MAP}\equiv\arg\max_{h\in H}\mathrm{P}(h\mid\boldsymbol{D})=\arg\max_{h\in H}\frac{\mathrm{P}(\boldsymbol{D}\mid h)\mathrm{P}(h)}{\mathrm{P}(\boldsymbol{D})} \tag{1}$$

Transaction history records are used as the training sample set for Bayesian learner. Suppose a node ei has the characteristic vector $\boldsymbol{X}_{ei}$, ($\boldsymbol{X}_{ei}$= (x_1, x_2, ... x_n)), here x_i is the characteristic value which corresponds to the variable X_i (i=1 to n). Here X_i (i=1 to n) corresponds to each column in "the node transaction record table"(see table 1).

Suppose there are m types of the classification: C_1, C_2,…C_m. To an unknown data sample $\boldsymbol{X}_{ei}$, the Bayesian learner will find out respectively the probability that the sample belongs to each type and $\boldsymbol{X}_{ei}$ belongs to the type that corresponds to the highest probability, when and only when formula (2) is met:

$$\mathrm{P}(C_i\mid\boldsymbol{X}_{ei})>\mathrm{P}(C_j\mid\boldsymbol{X}_{ei}),1\le j\le m,\ j\ne i \tag{2}$$

Here, C_i is the type that corresponds to the maximum posterior probability. According to the formula (1), the posterior probability can be calculated using the formula (3):

$$\mathrm{P}(C_i\mid\boldsymbol{X}_{ei})=\frac{\mathrm{P}(\boldsymbol{X}_{ei}\mid C_i)\mathrm{P}(C_i)}{\mathrm{P}(\boldsymbol{X}_{ei})} \tag{3}$$

Suppose there is conditional independence among the characteristic vectors (i.e. the Bayesian assumptions). $\mathrm{P}(\boldsymbol{X}_{ei}|C_i)$ can be calculated using the formula (4):

$$\mathrm{P}(\boldsymbol{X}_{ei}\mid C_i)=\prod_{k=1}^{n}\mathrm{p}(\boldsymbol{X}_k\mid C_i) \tag{4}$$

Through the training sample set, the Bayesian learner can estimate probability $\mathrm{P}(C_i)$ and $\mathrm{P}(\boldsymbol{X}_k|C_i)$. Here $\mathrm{P}(C_i)=S_i/S$, S is the total number of the training sample and S_i is the number of the training sample that the type number of the classification is equal to C_i. $\mathrm{p}(\boldsymbol{X}_k|C_i)=S_{ik}/S_i$, S_{ik} is the number of the training sample that the type number of the classification is equal to C_i and the value of the property $\boldsymbol{X}_k$ is equal to x_k .

$$\mathrm{P}(\boldsymbol{X}_{ei})=\sum_{i=1}^{m}\mathrm{P}(\boldsymbol{X}_{ei}\mid C_i)\mathrm{P}(C_i) \tag{5}$$

In order to avoid appearing of zero probability when computing the conditional probability $\mathrm{P}(X_k|C_i)$, the smooth factor is joined , which is estimated by the formula(6):

$$p(\boldsymbol{X}_k \mid C_i) = \frac{S_{ik} + 1}{S_i + 2} \tag{6}$$

In the paper C_i could be one out of three types: C_{trust}, C_{virus} and $C_{stranger}$. Here, formula (3) is used to meet the normalization requirements and to ensure that:

$$\sum_{i=1}^{m} P(C_i \mid \boldsymbol{X}_{ei}) = 1 \tag{7}$$

The LDoD of a node is calculated by the Bayesian learner based on transaction history records. Each record includes some transaction characteristics. In the paper selected characteristic items are divided into two parts: one is the statistical data during all periods of time (here recent 30 periods of time are taken); the other is the statistical data for recent 2 periods of time. The part of the transaction history record list is truncated (see **Table 1**). The meaning of symbols is described in the **Table 2**.

Table 1. the trading history record table

ID	the statistical data during all time periods(recent 30 periods of time)				..	the statistical data during recent 2 periods of time					The type of node
	X_1	X_2	X_3	X_4		X_5	X_6	X_7	X_8	X_9	C_i
100101	X_1>=223	X_2=1	X_3=3	X_4=2		X_5=0	X_6=1	X_7=1	X_8=2	X_9=16	C_{trust}
…	… …	… …	… …	… …		… …	… …	… …	… …	… …	… …

Table 2. The Meaning Of Symbols

symbols	the meaning of symbols	symbols	the meaning of symbols
X_1	the total number of trading between the local node and the node i in all time periods. X_1 is divided into several sections: X_1> = 100,100> X_1> = 50,50> X_1> = 20 and 20> X_1	X_6	the number of trading on a large scale and bad quality between the local node and the node i in the recent 2 periods of time. sizes of the downloaded files are equal to or greater than 5mb and divided into several sections: 5mb =< size =< 25mb, 25mb < size =< 100mb, 100mb < size
X_2	the number of malicious attacks on the local node from the node i in all time periods. X_2 is divided into several sections: X_2> = 4, X_2 = 3, 3> X_2> = 1 and X_2 = 0	X_7	the number of malicious feedbacks from the node i in the recent 2 periods of time. X_7 is divided into several sections: X_7> = 5,5> X_7> = 3, 3> X_7> = 1, X_7 = 0
X_3	the number of trading on a large scale and bad quality between the local node and the node i in all time periods. sizes of the downloaded files are equal to or greater than 5mb and divided	X_8	the number of other unsatisfactory trading events in the recent 2 periods of time. including: the trading on a small scale with a low quality, download speed is low, the

	into several sections: 5mb =< size =< 25mb, 25mb < size =< 100mb, 100mb < size		feedback deviation, etc.
X_4	the number of malicious feedbacks from the node i in all time periods. X_4 is divided into several sections: X_4> = 6, 6> X_4> = 4, 4> X_4> = 1 and X_4 = 0	X_9	The transaction failure rate. $X_9 = n_f / n$ n is the total transaction number in the recent 2 periods of time between the local node and the node i ; n_f is the total number of failed transactions.
X_5	the number of malicious attacks on the local node from the node i in the recent 2 periods of time. X_5 is divided into several sections: X_5> = 3, X_5 = 2, X_5 = 1, X_5 = 0	C_i	The type of node i

The initial training sample set is composed of the 120 transaction records between the local node and other 120 nodes, includes 40 trusted nodes, 40 malicious nodes and 40 strange nodes. In the proposed model, the training sample set will be improved at fixed period. Every 30 periods of time, some new samples will be added to the training sample set, or the section division of some characteristic columns will be adjusted. When the training sample set is changed, Bayesian learner will be produced again.

The processes of the training and the applying of Bayesian learner are described as follows:

1) Through the learning from the training sample set, the Bayesian learner is produced, which is called as "Bayesian learner A."

2) Using Bayesian learner A, according to the transaction history records between the local node and the node $\boldsymbol{X}_{ei}$, we can obtain three probabilities $P(C_{trust} \mid \boldsymbol{X}_{ei})$, $P(C_{stanger} \mid \boldsymbol{X}_{ei})$ and $P(C_{virus} \mid \boldsymbol{X}_{ei})$.

Definition 3: the local DoD (LDoD) is defined as the probability $P(C_{virus} \mid \boldsymbol{X}_{ei})$ that the node $\boldsymbol{X}_{ei}$ belongs to the malicious node set and that is calculated based on transaction history records between the local node and the node $\boldsymbol{X}_{ei}$.

Definition 4: the degree of satisfaction in the local (LDoS) is defined as the probability $P(C_{trust} \mid \boldsymbol{X}_{ei})$ that the node $\boldsymbol{X}_{ei}$ belongs to the trusted node set and that is calculate based on transaction history records between the local node and the node $\boldsymbol{X}_{ei}$.

3.3 The Training And The Applying Of Rough Set

Rough set theory is a kind of new mathematical tools handling the ambiguity problem. The Rough set can be used to conduct knowledge processing and to obtain the minimum expression of the knowledge on the premise of keeping the key information.

We use a four-tuple $\boldsymbol{S}=(\boldsymbol{U}, \boldsymbol{A}, \boldsymbol{V}, f)$ as a knowledge representation system. Here, $\boldsymbol{U}$ is a finite set of objects, called the domain; $\boldsymbol{A}$ is a non-empty finite set, called the attribute set; $\boldsymbol{V}$ is the value domain of the attribute set $\boldsymbol{A}$; f: $\boldsymbol{U}$ x $\boldsymbol{A} \rightarrow \boldsymbol{V}$ is the information function, in which any element of $\boldsymbol{U}$ has the property $\boldsymbol{A}$ with the value a which is uniquely determined value in $\boldsymbol{V}$.

Suppose $\boldsymbol{R}$ is a family of equivalence relations, $\boldsymbol{r} \in \boldsymbol{R}$, if relationship $ind(\boldsymbol{R})=ind(\boldsymbol{R}-\{\boldsymbol{r}\})$ exists, the attribute set $\boldsymbol{R}$ has the same classification as the attribute set $\boldsymbol{R}-\{\boldsymbol{r}\}$ on the object set $\boldsymbol{U}$, i.e.

$\boldsymbol{r}$ is a redundancy in R , otherwise $\boldsymbol{r}$ is a necessity in $\boldsymbol{R}$. If $\boldsymbol{Q}=\boldsymbol{R}-\boldsymbol{r}$ and $\boldsymbol{Q}\in\boldsymbol{R}$, $\boldsymbol{Q}$ is independent and satisfies to ind($\boldsymbol{Q}$)=ind($\boldsymbol{R}$), $\boldsymbol{Q}$ is a reduction of $\boldsymbol{R}$, that can be represented by $\boldsymbol{Q}$ = red($\boldsymbol{R}$). A family of equivalence relations $\boldsymbol{R}$ possibly has many reductions, the intersection of all reductions is defined as the Core of $\boldsymbol{R}$ and can be represented by core($\boldsymbol{R}$), core($\boldsymbol{R}$)=∩red($\boldsymbol{R}$). The smallest reduction is the smallest condition attribute set that can represent the same knowledge as the original information system. The smallest reduction is the simplest form of the information system that can keep the same classification ability. The classification rules are derived through the knowledge reduction.

The innovation of the paper is to improve the computation accuracy and the efficiency of the probability by using Rough set combined with Bayesian learner.

When we employ the Rough set classifier to conduct the classification of nodes, there are unrecognized conditions. When we employ the Bayesian learner to calculate the probabilities $P(C_{trust} \mid \boldsymbol{X}_{ei})$, $P(C_{stanger} \mid \boldsymbol{X}_{ei})$and $P(C_{virus} \mid \boldsymbol{X}_{ei})$, all attributes of the node are involved in the calculation, not one or several attributes can decide the calculation result. In practice, the probability can be obtained only by calculating one or several attributes, not all attributes. In other words, according to the value of an attribute can obtain the probability $P(C_{virus} \mid \boldsymbol{X}_{ei})$ that the node belongs to the malicious node set. For example, if X_2 (the number of malicious attacks in all time periods) is equal to 5 or greater than 5, the situation can independently determine the probability that the node belongs to the malicious node set is 100%, regardless of values of other attributes.Using the same training sample set as that used by the Bayesian learner, without any extra apriori information, the Rough set can be trained to produce some precise, verifiable classification rules (including the classification judgment rule that the classification result is determined by the value of an attribute or by the values of two attributes).Using these rules the classification judgment can be carried on before using the Bayesian learner, and the work load of the Bayesian learner will certainly be reduced. Using the Rough set combined with the Bayesian learner we can improve the calculation accuracy and efficiency. The process is described as follows:

1) Using the 120 training samples that used in "the Bayesian learner A" (described in section 3.2), the Rough set is trained. Through the attribute reduction and the attribute value reduction we can derive classification rules which are precise and easy to be inspected and confirmable.

2) Choose and retain a part of the classification rules according to the following requirements:

a) Only classification rules with confidence coefficient being 100% can be chosen and retained;

b) Only the classification rule in which decision attribute is "malicious node" can be chosen and retained;

c) Only the classification rule including 1 or 2 condition attributes can be chosen and retained.For example, IF (X_2)> = 5) THEN (classified as malicious node);

IF (X_2) = 3 AND (X_4) = 3) THEN (classified as malicious node).

3) These chosen and retained classification rules will be regenerated only when the training sample set is changed.

4) When we conduct the classification of a node:

a) First, the Rough set module is called, the chosen retained classification rules are used to conduct the classification of a node ,if we achieve success on the classification of the node , the probability for the node belongs to the malicious node set will be 100%, i.e. $P(C_{virus}|\boldsymbol{X}_{ei})=100\%$

b) Second, if the classification of the node using the Rough set module is ended in failure, "Bayesian learner" is called to calculate the probability of that the node respectively belongs to

the trusted node set or the strange node set or the malicious node set, i.e. $P(C_{trust} | \boldsymbol{X}_{ei})$, $(C_{stranger} | \boldsymbol{X}_{ei})$ or $(P(C_{virus} | \boldsymbol{X}_{ei})$.

c) The probability $(P(C_{virus} | \boldsymbol{X}_{ei})$ is the local DoD (LDoD) of the node.

5) When a change between the periods of time occurs, i.e. from a period of time to next period or when the serious damage occurs the Rough set module and the Bayesian learner will be called to calculate the probability of that the node respectively belongs to the trusted node set, the strange node set or the malicious node set.

3.4 The Collection And The Integration Of Recommendations

When transaction record data about a service provider is lacking or the service provider belongs to the strange node set, for the information of the service provider the local node may search in the trusted node lists and the malicious node lists that are provided by "trusted neighbor nodes". If the related information is still lacking, the local node will submit a request and ask for that other nodes to evaluate the service provider [17,18]. FBRs from other nodes are sent to the local node, FBRs are integrated, without adopting feedbacks that provided by malicious nodes. With different credibility for trusted nodes and strange nodes FBRs are integrated. The monitoring of malicious nodes includes the monitoring the quality of provided files, but also the quality of provided feedbacks. FBRs from other nodes are integrated, the service provider is evaluated, the feedback behavior is monitored.

3.4.1 The Comprehensive Evaluation Of The Service Provider

FBRs are provided in accordance with the format given in the **Table 1**,which includes: ID,X_1,X_2,……X_9,C_i. The meaning of symbols is described in the **Table 2**. In other words, a recommendation is a row of the **Table 1**.

Using different training sample set we can produce different classifier. In the model we employ comprehensive recommendation tables to produce a comprehensive training sample set. We employ the comprehensive training sample set to produce "Rough set module B" and "Bayesian learner B". The training sample set includes 120 recommendation lists. There are 40 recommendation lists according to that the conclusion of the recommendation is that the service provider is "a trusted node". There are other 40 recommendation lists according to that the conclusion is that the service provider is "a strange node". There are remaining 40 recommendation lists according to that the conclusion is that the service provider is "a malicious node".

Using all received recommendations about a service provider, we can obtain the comprehensive recommendation table (Table_overall). By adding together each column in the Table_overall (ignoring some columns), we can get a record (Record_overall). The Record_overall is a trade statistics between the service provider and all other nodes. Using 120 recommendation tables, we can obtain 120 records. Using these 120 records as the training sample set, we can produce "Rough set module B" and "Bayesian learner B".

In order to evaluate a service provider we employ all received recommendations about the service provider and obtain a comprehensive recommendation table (Table_overall). Using the Table_overall, by adding together each column in the Table_overall, we obtain a record (Record_overall), the Record_overall is a trade statistics. According to the Record_overall, using the Rough set module B and Bayesian learner B we can classify the service provider and calculate the probability that the service provider respectively belongs to the trusted node set , to the strange node set or to the malicious node set.

When the local node collected and integrated FBRs, a new table (Table_T) is comprised of FBRs provided by "trusted nodes". And another table (Table_S) is comprised of FBRs provided by "strange nodes".

By adding together each column in the Table_T, we may get a record (Record_T). The Record_T is a trade statistics between the service provider and all other trusted nodes. According to the Record_T, using the Rough set module B and Bayesian learner B we can classify the service provider and calculate the probability that the service provider respectively belongs to the trusted node set, the strange node set or the malicious node set，namely，Prob_T($P_t(C_{trust}|\boldsymbol{X}_{ei})$, $P_t(C_{stranger}|\boldsymbol{X}_{ei})$, $P_t(C_{virus}|\boldsymbol{X}_{ei})$).

We can do the similar process in Table_S as in Table_T to obtain the Record_S. And we classify the service provider to get Prob_S($P_s(C_{trust}|\boldsymbol{X}_{ei})$,$P_s(C_{stranger}|\boldsymbol{X}_{ei})$,$P_s(C_{virus}|\boldsymbol{X}_{ei})$).

According to experience, the credibility of recommendations from trusted nodes is higher than that from strange nodes. The credibility of recommendations from trusted nodes is expressed by T_t and the credibility of recommendations from strange nodes is expressed by T_s. (the calculation of T_t and T_s will be described in 3.4.2). The comprehensive evaluation of the service provider is expressed by Prob_G (which includes three probability values):

$$\begin{aligned} \text{Prob\_G}(& T_t*P_t(C_{trust}|\boldsymbol{X}_{ei}) + T_s*P_s(C_{trust}|\boldsymbol{X}_{ei}), \\ & T_t*P_t(C_{stranger}|\boldsymbol{X}_{ei}) + T_s*P_s(C_{stranger}|\boldsymbol{X}_{ei}), \\ & T_t*P_t(C_{virus}|\boldsymbol{X}_{ei}) + T_s*P_s(C_{virus}|\boldsymbol{X}_{ei})) \end{aligned} \quad (8)$$

Definition 5: The recommended DoD (RDoD) is defined as the probability that a service provider $\boldsymbol{X}_{ei}$ belongs to the malicious node set. It is calculated based on FBRs with the following formula:

$$\text{RDoD} = T_t* \text{Pt}(C_{virus}|\boldsymbol{X}_{ei}) + T\text{s}* \text{Ps}(C_{virus}|\boldsymbol{X}_{ei})$$

Definition 6: The degree of satisfaction based on recommendations (RDoS) is defined as the probability that a service provider $\boldsymbol{X}_{ei}$ belongs to the trusted node set. It is calculated based on FBRs with the following formula:

$$\text{RDoS} = T_t* \text{Pt}(C_{trust}|\boldsymbol{X}_{ei}) + T_s* \text{Ps}(C_{trust}|\boldsymbol{X}_{ei})$$

The comprehensive evaluation of a service provider (Prob_G) is treated as the correct result, the feedback behavior will be judged. If a feedback is consistent with the comprehensive evaluation result, the feedback behavior will be adjudged an honest feedback. Otherwise if the feedback is opposite to the comprehensive evaluation result, the feedback behavior will be judged as a reversal feedback. If within a time period, a node provides 1 or 2 reversal feedbacks, these feedback behaviors will be judged as "feedback accidental error", and if the node provides 3 or more than 3 reversal feedbacks, these feedback behaviors will be judged as "malicious feedback". These results of the evaluation of feedback behaviors will be added to "the trading history record table".

3.4.2 The Information Entropy Of Recommendations

The concept of the entropy comes from thermodynamics. The definition of the entropy in the thermodynamics is the logarithmic value of the possible status number of the system, called heat entropy. The entropy is a physical quantity used to measuring the degree of the disorderliness of the state of the molecular. *Shannon* took the thermodynamic entropy into the information theory and put forward the information entropy. The information entropy is used to measuring the degree of the confusion of the information system. The more orderly an information system is, the less the information entropy is. On the contrary, the more disorderly an information system is, the larger the information entropy is.

In information theory the entropy is a random variable which is measured with probability distribution. The formula for the information entropy H is:

$$H(\boldsymbol{X}) = H(p_1, p_2, ..., p_n) = -\sum P(x_i)\log(x_i) \quad (9)$$

Here, $P(x_i)$ is the probability that the information source takes the i_{th} symbol.(i=1,2,…,n.)

Through analyzing recommended tables (Table_T and Table_S), we discover that sometimes recommendations may be very centralized, but sometimes may be quite decentralized. Obviously, when a set of recommendations is centralized, they will be with high reliability,and the information entropy of them will be smaller. Otherwise, when a set of recommendations is decentralized, they will be with low reliability and the information entropy of them will be larger. Obviously, the credibility of recommendations is inversely proportional to the information entropy.

The calculation of the information entropy of a set of recommendations is equal to the calculation of the dispersion of the set of recommendations. Recommendations are divided into three cases: the evaluated node may be recommended as a trusted or strange or malicious node. According to the recommendation table (Table_T or Table_S), we respectively count the number of nodes who recommended the evaluated node for a trusted or strange or malicious node. Next, we respectively calculate the probability of that the evaluated node belongs to the trusted node set, or belongs to the strange node set, or belongs to the malicious node set. See **Table 3**.

Table 3.A example of the calculation of the information entropy

	the number of node who recommended the evaluated node for a trusted node	the number of node who recommended the evaluated node for a strange node	the number of node who recommended the evaluated node for a malicious node
the number of nodes	70	20	10
the probability	70%	20%	10%
the information entropy of recommendations	$H=-(0.7\log_2 0.7 + 0.2\log_2 0.2 + 0.1\log_2 0.1) = 1.157$		

If recommendations on the evaluated node are decentralized, for example, the percentage of nodes who recommended the evaluated node for a trusted node, or for a strange node, or for a malicious node is one third respectively. In this case, the uncertainty or the information entropy is the largest, the maximum entropy is: H = 1.58. If all recommendations on the evaluated node are centralized, one of the percentages is 1, other two are 0. In this case, the uncertainty or the information entropy was 0, the range of the information entropy of recommendations was: $0 \le H \le 1.58$

The credibility is in reverse proportional to the information entropy: $T_t/T_s = H_s/H_t$, the formula of the credibility which is based on the information entropy is as follows:

$$T_t=H_s/(H_t+H_s);\ T_s=H_t/(H_t+H_s)\ ;\ T_t + T_s =1 \quad (10)$$

Here,T_t is the credibility of recommendations from trusted nodes. H_t is the information entropy of recommendations from trusted nodes. T_s is the credibility of recommendations from strange nodes. H_s is the information entropy of recommendations from strange nodes.

The calculation of the credibility based on the information entropy is reasonable and

objective. In most cases, experimental results show that the credibility of recommendations from trusted node would be around 0.6, and that from strange nodes would be around 0.4.

Similarly, recommendations from all nodes constitute a recommendation table (Table_overall). H_overall stands for the information entropy of these recommendations from all nodes, and T_overall stands for the credibility of them. It is calculated as follows:

T_overall = 1 / H_overall and if(T_overall >= 1) than T_overall=1

3.5 The Calculation Of The Degree Of Dissatisfaction (DoD)

DoD will be calculated in accordance with following rules:

```
If  ( LDoD > 0.6 ) or ( LDoD < 0.2  && LDoS > 0.6 )
 then {
        DoD = LDoD     if  LDoD > 0.6;
        DoS = LDoS      if  LDoD < 0.2  and  LDoS > 0.6;
  }

  else If ((RDoD > 0.6 )  or  (RDoD < 0.2  and  RDoS > 0.6) )
  then {
          DoD = RDoD   if RDoD > 0.6;
          DoS = RDoS    if RDoD < 0.2  and  RDoS > 0.6;
 }
 else {
          DoD =  LDoD/(1+ T_overall)  +  RDoD * T_overall /(1+ T_overall);
}
```

Clearly, when all the recommendations are consistent, T_overall is equal to 1, the contribution to the DoD from recommendations is bigger. When all recommendations are scattered, T_overall is less than 1, the contribution to the DoD from recommendations is smaller.

3.6 The Classification Of Nodes

The nodes are classified according to the DoD.

Definition 7: A node with DoD greater than 0.6 is identified as that the node belongs to the malicious node set.

Definition 8: A node with DoS greater than 0.6 and DoD less than 0.2 is identified as that the node belongs to the trusted node set.

Definition 9: A node does not belong to the trusted node set and does not belong to the malicious node set is identified as that the node belongs to the strange node set.

Once the evaluated node is judged as a malicious node, it will be added to the malicious node list and will be under the control. If the evaluated node is judged as a trusted node, it will be added to the trusted node list and it will have a priority to do transaction. If the evaluated node is judged as a strange node, according to the rank sorted with their DoD, it will be made transaction properly. In each transaction the first partner is chosen from the trusted node list, only when needed contents can not be obtained with satisfied result from the current transaction group, the user have to choose new transaction partner from the strange node list.

When a change between the periods of time occurs, i.e. from a period of time to next period or when a serious damage occurs the Rough set module and the Bayesian learner will be called to recalculate the probability that the node respectively belongs to the trusted node set, the strange node set or the malicious node set and to recalculate the DoD of the related node and to reclassify the related node.

4. The Simulation And The Result Analysis

Simulation experiment is the most popular evaluation method for security model at present. We used the computer to simulate the concrete application scene and the interactive behavior between nodes in the p2p network. The security model is evaluated from many aspects [19][20]. We realized a simulation peer-to-peer network environment by integrating the PeerSim platform with the Rough set module and the Bayesian learner to achieve the performance analysis of the model. In the simulation experiment, the scale of the p2p network was set to 800 nodes, within single simulation cycles each node carried on the transaction with at least other 30 nodes.

Simulation experiments were divided into two stages: the preparatory stage (the initialization stage) and the detection stage. During the preparatory stage, at the beginning, all nodes (e.g., all 800 nodes) were in strange status. There wasn‘t any transaction record. In the preparatory stage, mutual recommendations among nodes would not be adopted. When the local node needed a file, it would by the help of a search engine get a list of candidate providers (e.g., there were 50 nodes that could provide the file). Some providers were random selected (e.g., there were 30 providers were selected) to make transactions. After n periods of time (e.g., n=5), for all nodes, transaction records were produced at the local. 30 simulation cycles were scheduled in every period of time. By calling the Rough set module and the Bayesian learner the classification of nodes was conducted and "the trusted neighbor node list", "the trusted node list" and "the malicious node list" were established.

During the detection stage, after getting a list of candidate file providers, we took following steps:

(1) Eliminate malicious nodes;

(2) Give preference to the selection of trusted nodes;

(3) For several strange nodes, ask for other nodes to evaluate these strange nodes and call the Rough set module and the Bayesian learner to make comprehensive evaluation. According to the inverse order of their DoD of these nodes, transaction nodes were selected.

In the beginning of every new periods of time, the Rough set module and the Bayesian learner were used to reclassify those nodes, which’s transaction records had changed. In the experimental process, once a serious damage happened, the corresponding transaction record would be updated immediately, and the Rough set module and the Bayesian learner would be used to judge whether the node should be reclassified due to the serious failure event. If a malicious node was detected, the node would be immediately added to the malicious node list. In the experiment, two configuration models were applied:

One security management model was that the Rough set module and the Bayesian learner was used to realize the node classification control (the proposed model). The trusted node list and the malicious node list were established in the model, and malicious nodes were examined and eliminated when the file provider was chosen.

Another model was for comparison purpose (the comparison model), in which a node had the higher reputation and the node had the higher priority for selection. In the comparison model, the classification management to unsatisfactory events was not carried out, only the

trade history records were established to choose the file provider with the reputation level in descending order.

4.1 Simulation experiments (1)

The goal of the experiment (1) was to verify whether the proposed security management model based on Rough set and Bayesian learner could help to find out the appropriate file provider effectively and could guarantee the file downloading success rate. The proposed security mechanism should have the strong examination ability for the malicious behavior. To express the ability of security management, the document downloading success rate and the examination rate of malicious nodes were used. The file downloading success rate reflected mainly the detection ability of the malicious attack when providing the document downloading service. The examination rate of malicious nodes reflected mainly the detection ability of the malicious feedback when providing FBRs service [21][22].

In the transaction process between the local node and another node,at some time t,the total number of the examined file downloading service was S(t), the total number of the success of the file downloading service was N(t), then the file downloading success rate SR(t) was defined: SR(t) = N(t)/S(t). Supposed that the percentage of malicious node in the initial setting was β, this value would immediately influence the initial success rate in the simulation experiment. If β = 20%, the initial success rate would be 1- 20% = 80%; along with increasing of the iteration number, the file downloading success rate was enhanced from 80% unceasingly.

Supposed when the local node collected and integrated FBRs, at some time t, the total number of the received FBRs was Fs(t),the total number of detected malicious feedbacks was Fm(t), the percentage of malicious node in the initial setting was β, the detection rate Fr(t) of malicious nodes was defined : Fr(t)= Fm(t)/ Fs(t)/β.

In the experiment (1), 5 different settings were used: β=0.2,0.25,0.3,0.35,0.4 . The proposed model and the comparison model were tested. Under each setting, the number of iterations was more than 30.

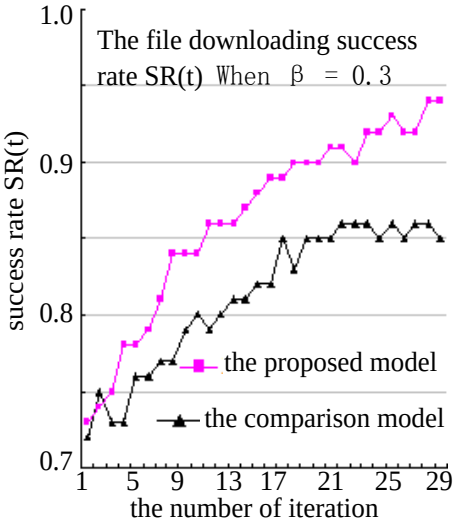

Fig. 1. The file downloading success rate SR(t) When β = 0.3

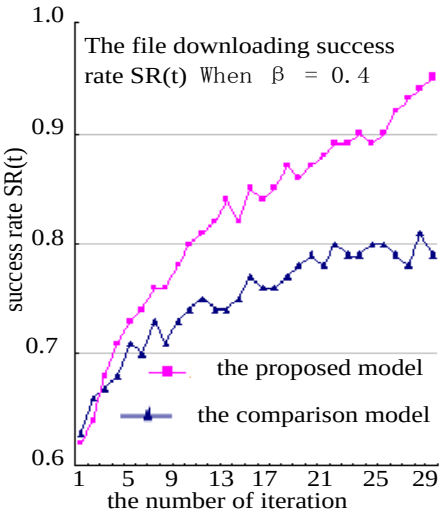

Fig. 2. The file downloading success rate SR(t) When β = 0.4

Result analysis:

Fig.1 and **Fig.2** gave the results of the file downloading success rate for two models respectively when $\beta = 0.3$ and when $\beta = 0.4$. From two figures, it could be concluded that two models performed closely at the beginning, along with increasing of the number of iterations, the proposed model in the paper gained a success rate 10%-15% higher than that of the comparison model. The reason was that in the proposed model the detailed analysis and classification management over transaction failure events were achieved and the classification control of trading nodes was realized.

4.2 Simulation experiments (2)

The goal for the second experiment was to verify the dynamic compatibility of two models. The dynamic compatibility was the ability for a model to provide the reliable service. A good security model could continuously provide the stable service in a complex dynamic environment. In the simulation experiment, two kinds of dynamic change factors had been considered: First, part of nodes might momentarily leave or join. Second, part of malicious nodes might change their behavior frequently, from time to time provided the malicious service, or from time to time provided the normal service. We used the following symbols:

(1) the percentage of the nodes that provided malicious file download service: β

(2) the percentage of the nodes that provided malicious feedback: β_2

(3) the percentage of the malicious nodes that dynamic changed malicious behavior way: β_3

(4) the percentage of the nodes that might momentarily leave or join: β_4

In the experiment three kinds of settings were used:

$\beta = 0.3$; $\beta_2 = 0.3$; $\beta_3 = 0.2$; $\beta_4 = 0.2$

$\beta = 0.25$; $\beta_2 = 0.25$; $\beta_3 = 0.3$; $\beta_4 = 0.1$

$\beta = 0.2$; $\beta_2 = 0.2$; $\beta_3 = 0.25$; $\beta_4 = 0.25$

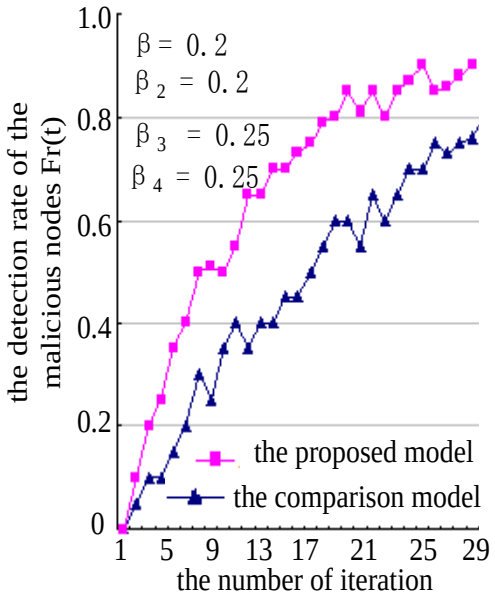

Fig. 3. the detection rate of malicious nodes in the complex dynamic environment $\beta = 0.2$; $\beta_2 = 0.2$; $\beta_3 = 0.25$; $\beta_4 = 0.25$

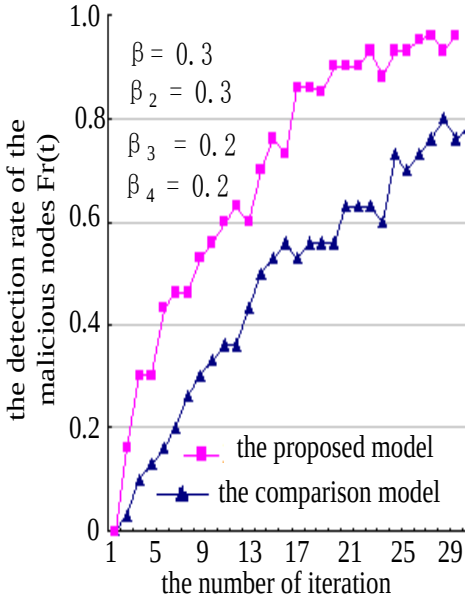

Fig. 4. the detection rate of malicious nodes in the complex dynamic environment $\beta = 0.3$; $\beta_2 = 0.3$; $\beta_3 = 0.2$; $\beta_4 = 0.2$

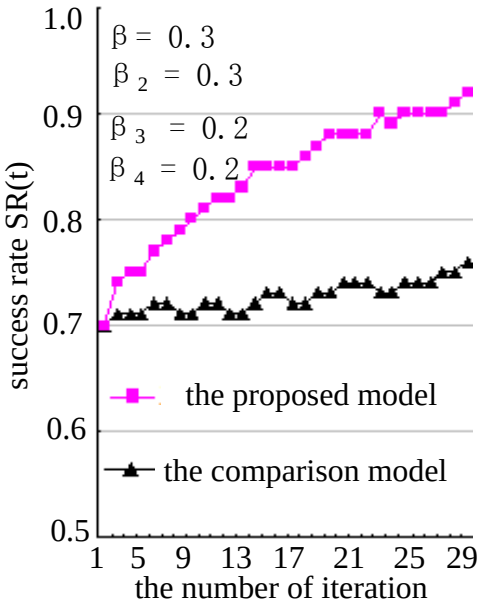

Fig. 5. The file downloading success rate SR(t) in the complex dynamic environment $\beta = 0.3$; $\beta_2 = 0.3$; $\beta_3 = 0.2$; $\beta_4 = 0.2$

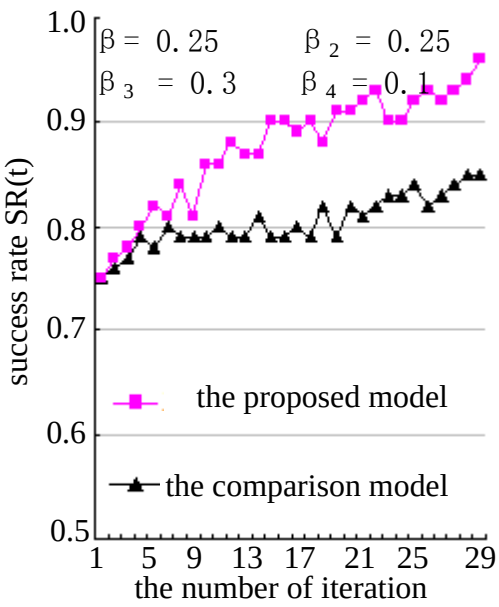

Fig. 6. The file downloading success rate SR(t) in the complex dynamic environment $\beta = 0.25$; $\beta_2 = 0.25$; $\beta_3 = 0.3$; $\beta_4 = 0.1$

Result analysis:

(1) **Fig. 3** and **Fig. 4** gave the comparison results of the examination rate of malicious nodes in the complex dynamic environment. What could be concluded was that at the beginning, the examination rate of malicious nodes of two models started from 0, along with increasing of iterations, the model based on Rough set and Bayesian learner gained a higher detection rate of malicious nodes.

(2) **Fig. 5** and **Fig. 6** gave the comparison results of the file downloading success rate in the complex dynamic environment. The results show that two models performed closely at the beginning, along with increasing of the number of the iterations, the model based on Rough set and Bayesian learner gained a success rate 10%-15% higher than that of the comparison model.

5. Conclusion

A new security management model based on Rough set and Bayesian learner is proposed in the paper. The small-scale system can be installed in any P2P node and managed by user. In the model, the definition and the computational method of the degree of dissatisfaction (DoD), the local DoD (LDoD) and the recommended DoD (RDoD) are provided. Using the Rough set combined with the Bayesian learner we have improved the calculation accuracy and efficiency. Based on the DoD, nodes are classified as trust nodes, strange nodes and malicious nodes. The model can help users in P2P network environment to select correctly the transaction object and to avoid malicious nodes. The simulation results show that compared with the existing security model, the model can improve the transaction success rate of 10%- 15% above and have the more stable dynamic adaptive ability. The paper is based on p2p network, how to apply the model to other area, such as wireless sensor networks (WSNs) is our future work.

References

[1] GUO Xiao-qiong and GUAN Hai-bing, "A Trust Management Model Based On Bayesian Network". *Information Security and Communications Privacy,* vol. 02,2008. Article (CrossRef Link)

[2] Tian jun-feng and Tian rui, "A fine-grain trust model based on domain and bayesian network for P2P E-Commerce system," *Journal Of Computer Research And Development,* vol.20, 2010.Article (CrossRef Link)

[3] Yu Zhihua, "Analysis of malicious behaviors in peer-topeer trust model". *Computer Engineering and Applications*, vol.43, no.13, pp.18-21 2007.Article (CrossRef Link)

[4] M. Srivatsa and L. Liu, "Securing decentralized reputation management using TrustGuard", *Journal of Parallel and Distributed Computing, v*ol.66, no.9, pp.1217–1232, 2006.
Article (CrossRef Link)

[5] Y. Sun, Z. Han, W. Yu and K.J.R. Liu, "A trust evaluation framework in distributed networks: vulnerability analysis and defense against attacks", in *Proc. of the 25th IEEE Conference on Computer Communication*, pp.230–236, 2006.
Article (CrossRef Link)

[6] A.A. Selcuk, E. Uzun and M.R. Pariente, "A reputation-based trust management system for P2P networks," *International Journal of Network Security,* vol.6, no.3, pp.235–245 2008.
Article (CrossRef Link)

[7] M. Mordacchini, R. Baraglia, P. Dazzi and L. Ricci, "A P2P recommender system based on gossip overlays PREGO", in *Proc. of the 10th IEEE International Conference on Computers and Information Technology*, pp.83–91, 2010.Article (CrossRef Link)

[8] Zhang Q, Zhang X, Wen X Z, Liu J R, et al. ,"Construction of peer-to-peer multiple-grain trust model". *Journal of Software*, vol 17, no.1, pp.96-107, 2006
Article (CrossRef Link)

[9] Song Shan-shan, Huang Kai and Zhou Run-fang, "Trusted P2P transactions with fuzzy reputation aggregation," *Internet Computing*, vol.9, no.6, pp.24-34 2005.Article (CrossRef Link)

[10] Li Xiao-yong,Gui Xiao-Lin and Zhao Juan, "Novel scalable aggregation algorithm of feedback trust information," *Journal of Xi'an JiaoTong University*, vol. 41, no.8, pp.142-146 2007
Article (CrossRef Link)

[11] Tian Chun-qi Zou Shi-hong and Tian Hui-rong, "A new trust model based on reputation and risk evaluation", *Journal of Electronics & Information Technology,* vol. 29, no.7, pp.1628-1632, 2007
Article (CrossRef Link)

[12] Shi Rong-Hua, Xin Jin-jin, "New P2P trust model based on group". *Application Research of computers v*ol.27, no.7, Jul.2010.Article (CrossRef Link)

[13] Chun-Ta Li, Cheng-Chi Lee and Lian-Jun Wang, "On the security enhancement of an efficient and secure event signature protocol for P2P MMOGs". *Lecture Notes in Computer Science*, vol.6016, pp.599-609 2010Article (CrossRef Link)

[14] Hou-Meng-shu,Lu Xian-liang, Zhou Xu, et al., "A trust model of p2p system based on confirmation theory" .*Operating Systems Review*, vol.39, no.1, pp.56-62, 2005.
Article (CrossRef Link)

[15] Cassio P. de Campos, Zhi Zeng and Qiang Ji, " Structure learning of Bayesian networks using constraints", in *Proc. of the 26th Annual International Conference on Machine Learning*, pp.113-120, 2009.Article (CrossRef Link)

[16] Nurmi, P. , "A bayesian framework for online reputation systems". In *Proc. of the Advanced Int. Conf. on Telecomm, Guadeloupe, French Caribbean, IEEE Computer Society,* 2006. Article (CrossRef Link)

[17] S. Lim, C. Keung and N. Griffiths, "Towards improved partner selection using recommendations and trust", in: R. Falcone et al. (Eds.), Trust in Agent Societies (TRUST), *Lecture Notes in CS*, vol. 5396, pp.43–64, 2008.Article (CrossRef Link)

[18] Jiang Shouxu, and Li Jianzhong, "A reputation-based trust mechanism for P2P e-commerce systems". *Journal of Software*, vol.18, no.10, pp.2551-2563, 2007.Article (CrossRef Link)

[19] Cheng Chien-Fu, Wang Shu-Ching and Liang Tyne, "File consistency problem of file-sharing in peer-to-peer environment". *International Journal of Innovative Computing Information And Control,* vol.6, pp.601-613, 2010 Article (CrossRef Link)
[20] Li Xiong and Ling Liu, "Peer Trust- supporting reputation-based trust for peer-to-peer election communities". *IEEE transactions on Knowledge and Data Engineering*, vol.16, no.7, pp.843- 857, 2004 Article (CrossRef Link)
[21] Chen Chyouhwa, Tsai Chia-Liang and Horng Shi-Jinn, "Exploiting attribute popularity distribution skew to enhance the performance of peer to peer publish/subscribe systems". *International Journal of Innovative Computing Information And Control,* vol.7, pp.4047-4066 2011
Article (CrossRef Link)
[22] Li Xiaoyong, Feng Zhou and Xudong Yang,"A multi-dimensional trust evaluation model for large-scale P2P computing ". *Journal of Parallel and Distributed Computing,* vol.71, pp.837-847, 2011Article (CrossRef Link)

Hai-Sheng Wang,born in 1978, received his M.S. degrees in Computer Engineering from Xidian University, Xi'an, China, in 2004, now he is a Ph.D. candidate in Xi'an Jiaotong University in China. As the first author, he has published more than ten journal papers. His current research interests mainly include networks computing and trusted system.

Xiao-Lin Gui ,born in 1966, is a professor and a Ph.D. supervisor in Xi'an Jiao-tong University. He has published more than eighty papers and obtained five patents and three soft-ware copyrights. In 2006, he is awarded New Century Excellent Talents in University (NCET). Now he is in charge of a project of the National High-Tech Research and Development 863 Program and a project of the National Nature Science Foundation of China. Currently, he leads the Trusted Computing Technology Research Center (TCT Lab) at Xi'an Jiaotong University. His research interests include networks computing and trusted system.