A Kind of Prediction Method of User Behaviour for Future Trustworthy Network[1]

Tian Liqin[1, 3], Lin Chuang[2], Sunjinxia[1]
([1]Information Engineering School, Beijing University of Science and Technology, Beijing, China)
([2]Department of Computer Science and Technology, Tsinghua University, Beijing China)
([3]North China Institute of Science and Technology, Beijing East China)
tianliqin@tsinghua.org.cn

Abstract—With the increasing development of the computer network application, network security is facing the heavy challenge. Currently, many kinds of the isolated network security measures, such as Fire Wall and Access Control, form a rampart around the protected system for defensive purposes. But these methods are not effective to deal with the network attack and destruction, due to these attacks and destructions are various, random, covert and sometimes epidemic. The international research shows that network security is on the way to Trustworthy Network. Apart from current security mechanism, the future Trustworthy Network adds network behaviour trust and user behaviour trust, so user behaviour prediction is important and significant for realization of the Trustworthy Network. The paper discusses how to based on past user behaviour to predict future user behaviour, including Bayesian Network modeling, prediction grading, computation of prior probabilities of user behaviour, computation of prior probabilities of user behavior attribute and prediction of future user behaviour. In order to meet needs of different prediction of user behaviour for different purpose, we also discuss that how to get and store user's past assorted behaviour statistical data. Finally, discuss how to predict user behaviour in real system and how to control user behaviour based on prediction result.

Keywords: Trustworthy Network; Network Security; User Behaviour; Bayesian Network, Method of prediction

I. INTRODUCTION

The computer network is facing increasing challenge. International research shows that network security is on the way to Trustworthy Network [1-4] and the future network security adds network behaviour trust and user behaviour trust. The definition of the Trustworthy Network is: network behaviour, user behaviour and their result are always predictable and controllable [4]. The Trustworthy Network is composed of three components, namely, network service provider, network and user. So user behaviour and its prediction is one of the important and significant things for realization of Trustworthy Network. User behaviour prediction is based on past user behaviour. Our idea is based on hierarchical and decomposition. User behaviour can be subdivided into some small behaviour unit named behaviour attributes, such as security behaviour attribute, dependability behaviour attribute and performance behaviour attribute. The behaviour attribute also can be subdivided into more small and specific behaviour evidence, such as User Lost packet rate, User Data transfer jitter, User Responding time and Establish-connection delay etc. This evidence can be gotten directly by the network detecting software, such as Bandwidthd[8], NetFlow Monitor[5,6,9].The paper discusses the prediction of user's future behaviour based on user's past behaviour evidence. The theory of Bayesian Network gives us robust prediction theoretical tool, which combines graph theory and probability theory. We can use it to flexibly predict user's future behaviour probability grade based on different qualification for different purpose. It also gives a compact, straight and effective expression for the associated relationship between the user behaviours.

1. This work is supported by the National Natural Science Foundation of China (No. 90412012，60429202, 60432030 and 90104002), NSFC and RGC (No. 60218003), the National Grand Fundamental Research 973 Program of China (No. 2003CB314804)

II. BAYESIAN NETWORK AND EVALUATION OF USER'S BEHAVIOUR

A. Evaluation of user's behaviour

The evaluation of user's behavior is comprehensive evaluation for user's behavior, which is reflected by user's past behaviors. In order to get evaluation result effectively, we first subdivide user's behavior into behavior attributes, such as security behaviour attribute and performance behaviour attribute. Then we subdivided behavior attributes once again into more small data unit, namely behavior evidences. We use the layered, subdivided and quantitative idea to convert the complicated and comprehensive evaluation of user's behavior into the measurable and computable evaluation of behavior evidences. Thus the evaluation of user's behaviour in the trustworthy Network [4] can be solved effectively. So this method is feasible in engineering.

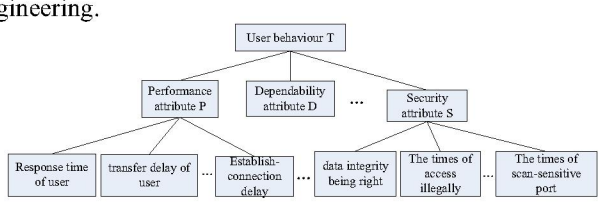

Fig. 1. Evaluation hierarchy architecture of user behaviour

We decompose user's behavior into three layers like Fig.1. Thus, the value of each behaviour evidences can be gotten by software and the weight of each behaviour evidences and behaviour attributes are acquired scientifically according to the Analytic Hierarchy Process (AHP) [7]. Then evaluation result of user's whole behavior can be acquired by "composing" the evidences effectively. We use matrix to compute evaluation value of user's behavior. Suppose n to be the number of attributes and k to be the maximum number of evidences included the attributes, $et_{ij} \in [0,1]$ to be the evidence value lying at i^{th} row and j^{th} column and $w_{ij} \in [0,1]$ to be the weight of et_{ij}, $E = \begin{bmatrix} et_{11} & \cdots & et_{1f} & \cdots & et_{1k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ et_{i1} & \cdots & et_{if} & \cdots & et_{ik} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ et_{n1} & \cdots & et_{nf} & \cdots & et_{nk} \end{bmatrix}$ to be the matrix of evidences and $WE = \begin{bmatrix} w_{11} & \cdots & w_{1f} & \cdots & w_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{i1} & \cdots & w_{if} & \cdots & w_{if} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ w_{k1} & \cdots & w_{kf} & \cdots & w_{kn} \end{bmatrix}$ to be the matrix of weight. Thus the formula to compute the attributes is $E * WE$. The value on main diagonal is the final result.

After the evaluation results of behavior attributes are acquired, we can evaluate whole user's behavior. Suppose behavior attributes vector to be A =[a_1,a_2...a_n], and the corresponding weighs vector to be WA =[w_1,w_2...w_n], then the formula to compute whole user's behavior is $A * WA$

B. *Establishment of Bayesian Network about user behaviour*

In order to use Bayesian Network to predict user behaviour, we need convert the evaluation network of user behaviour into Bayesian Network. A Bayesian Network is composed of two parts, one is its structure that is qualitative part and the other is function of Conditional probability that is quantitative part. We include following three steps to establish Bayesian Network model of user behaviour:

(1) Define domain variable that include user behaviour T, Performance P, Dependability D…Security S.

(2) Establish reliant relationship among the each variable according to the structure of the Network of user behaviour that is depicted in Fig.2.

(3) Compute each variable's Conditional probability according to the past interaction behaviour. This will be discussed more detail in section 4.4.

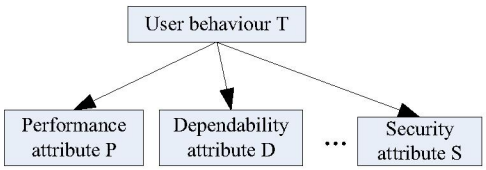

Fig.2. Bayesian Network mode for user behavior prediction

III. PREDICTION OF USER BEHAVIOUR

A. *Prediction grade and data structure*

In order to predict user behaviour effectively, we divide each node value, into n grades. They are $\left[1-\frac{0.4}{n-1},1\right]$, $\left[1-2\times\frac{0.4}{n-1},1-\frac{0.4}{n-1}\right],\cdots$, $\left[1-i\times\frac{0.4}{n-1},1-(i-1)\frac{0.4}{n-1}\right],\cdots$, $\left[th0,1-(n-2)\times\frac{0.4}{n-1}\right]$, $[0,th0]$, where $1 \le i \le n$, $th0$ is threshold, when the evaluation value is below the $th0$, the Service provider will deny the access.

After an each interaction of two entities, the node's value needs to be computed using the methods mentioned in the section 3.1. But for the sake of prediction's scalability, these evaluation data don't save all. The saved data is only the times that node value falls into certain range, which is our basic data for predicting user behaviour. After each interaction of two entities, if certain node's value falls into certain range, then the corresponding interaction times add one and the other statistical times remain immovability.

We also save the times that two or more different nodes' value fall into different certain range at the same time, which is used to flexibly predict future user behaviour probability grade based on more user behaviour attribute for specific purpose. We use two, three and four dimensions array to save the times that two, three and four nodes' value fall into different range at the same time. The array name denotes different node, for example, array TP denotes two nodes, namely, user behaviour node T and user performance attribute node P; the subscript denotes different node value range. Because each node value has n different scopes, we can sort the five ranges decreasingly, and number these sorted ranges from one to n. For example, let TP=$\begin{pmatrix} 5 & 4 & 3 & \cdots & 2 \\ 3 & 5 & 2 & \cdots & 0 \\ \cdots & & \cdots & & \cdots \\ 3 & 6 & 5 & \cdots & 1 \end{pmatrix}_{n\times n}$ and PS=$\begin{pmatrix} 3 & 5 & 3 & \cdots & 2 \\ 2 & 4 & 6 & \cdots & 0 \\ \cdots & & \cdots & & \cdots \\ 3 & 6 & 4 & \cdots & 1 \end{pmatrix}_{n\times n}$, then TP_{11} stores the times 5 that value of node T and node P are all within $\left[1-\frac{0.4}{n-1},1\right]$, array PS_{23} stores the times 6 that

value of node P and node S are within $\left[1-2\times\frac{0.4}{n-1},1-\frac{0.4}{n-1}\right]$ and $\left[1-3\times\frac{0.4}{n-1},1-2\times\frac{0.4}{n-1}\right]$ respectively.

For our user behaviour prediction mode, we let n=5, $th0$ =0.6, then the prediction grades are extreme satisfaction [0.9,1], satisfaction [0.8,0.9], comparative satisfaction [0.7.0.8], basic satisfaction [0.6,0.7], no satisfaction [0,0.6] respectively. We use T_i, P_i, D_i, T_i and S_i $(1 \le i \le 5)$ to denote these scopes respectively for perspicuous discussion in the rest of the paper. There are $C_4^4 + C_4^3 + C_4^2 = 11$ arrays in all, four of them are three dimensions array, six of them are two dimensions and one of them is four dimensions array.

B. *Prior probability of user behaviour*

Bayesian Networks theoretical foundation is Bayesian rule [5,6]:

$$p(\mathrm{h/e}) = \frac{p(e/h)p(h)}{p(e)} \qquad (1)$$

Where p(h) is the prior probability of hypothesis h; p(e) is the prior probability of behaviour evidence e; p(h|e) is the probability of h given e; p(e|h) is the probability of e given h.

From the part 4.1, we can know root node T value is divided into five ranges, namely, extreme satisfaction [0.9,1], satisfaction [0.8,0.9], comparative satisfaction [0.7.0.8], basic satisfaction [0.6,0.7], no satisfaction [0,0.6]. We use $p(T_i)$ $(1 \le i \le 5)$ to denote probability of extreme satisfaction, satisfaction, comparative satisfaction, basic satisfaction, no satisfaction respectively. Then we can use following formula to compute them:

$$p(T_i) = \frac{m_i}{n} \quad (1 \le i \le 5)$$

Where n denotes the total times that system interacts with the user and $\sum_{i=1}^{5} p(T_i) = 1$, m_i denotes the times that node T's value fall into T_i $(1 \le i \le 5)$. T's value can be computed using the method mentioned in the section 3.1

C. *Prior probability of user behavior attribute*

The leaf node is similar to the root node. Their values are possible in one of the five intervals, which denote "very satisfaction", "satisfaction", "comparative satisfaction", "basal satisfaction" and "dissatisfaction", respectively. $p(P_1)$ denotes the probability that performance attribute P is "very satisfaction". $p(P_i)$ $(2 \le i \le 5)$ is the same as $p(P_1)$ in meanings. They can be computed by the formula: $p(P_i) = \frac{a_i}{n}$ $(1 \le i \le 5)$, where a_i denotes the total interaction times that the value of performance attribute P in the P_i, $(1 \le i \le 5)$ and $\sum_{i=1}^{5} p(P_i) = 1$. The value of P has been computed in the foregoing user behavior evaluation above-mentioned in the Section 3.1. The method for computing prior probability of other leaf node is similar to the method of the leaf node P.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

D. *Conditional probability table*

Each leaf node has a conditional probability table. Table 1 shows conditional probability table of leaf node P.

Table 1. Conditional probability table of the node P

	T_1	T_2	T_3	T_4	T_5
P_1	$p(P_1/T_1)$	$p(P_1/T_2)$	$p(P_1/T_3)$	$p(P_1/T_4)$	$p(P_1/T_5)$
P_2	$p(P_2/T_1)$	$p(P_2/T_2)$	$p(P_2/T_3)$	$p(P_2/T_4)$	$p(P_2/T_5)$
P_3	$p(P_3/T_1)$	$p(P_3/T_2)$	$p(P_3/T_3)$	$p(P_3/T_4)$	$p(P_3/T_5)$
P_4	$p(P_4/T_1)$	$p(P_4/T_2)$	$p(P_4/T_3)$	$p(P_4/T_4)$	$p(P_4/T_5)$
P_5	$p(P_5/T_1)$	$p(P_5/T_2)$	$p(P_5/T_3)$	$p(P_5/T_4)$	$p(P_5/T_5)$

$p(P_i/T_j)$ denotes the probability that the value of performance attribute P is in P_i and the value of root node T is in T_j. It can be computed according to the following formula:

$$p(e/h) = \frac{p(h,e)}{p(h)} \qquad (2)$$

For instance, $p(P_i/T_j) = \frac{p(P_i,T_j)}{p(T_j)} = \frac{TP_{ij}/n}{m_j/n} = \frac{TP_{ij}}{m_j}$, m_i denotes the total interaction times that the root node is in T_i $(1 \le i \le 5)$. The conditional probability table of the other leaf nodes is similar to this table. For example, the conditional probability table of the leaf node D can be obtained by making node D displace node P in Table 1. Each conditional probability table includes five columns and the sum of values of each column is equal to 1.

E. *Prediction of User behavior*

Once getting the prior probability of all nodes and leaf nodes' conditional probability table, the system can predict the probability of user behaviour by using Bayesian rule [10] based on specific past behaviour attribute, the formula is:

$$p(T_1 / P_1) = \frac{p(P_1 / T_1)p(T_1)}{p(P_1)}$$

We can compute the probability that the user behaviour is "very satisfaction" given performance attribute is "very satisfaction". We also can predict the probability of user behaviour based on several different behaviour attributes. For instance, according to the following formula:

$$p(T_2 / P_1, S_1) = \frac{p(P_1, S_1 / T_2)p(T_2)}{p(P_1, S_1)} = \frac{p(P_1, S_1, T_2)}{p(P_1, S_1)}$$

we can compute the probability that user behaviour is "satisfaction", given performance behavior attribute is "very satisfaction" and security behavior attribute is "very satisfaction".

Now we can decide whether the user can interact with the system or restrict the user access system resource. Comparing the prediction probability with different threshold θi_i , if the probability is greater than certain threshold θ_{i0} , then the user can get corresponding right and access to resource; if the probability is smaller than the minimal threshold $\theta_{\min}$, then the system can cut the network connections with the user and refuse the user to accessing it. So we can predict, control and make warning of the user behaviour in advance according to the different needs of the system.

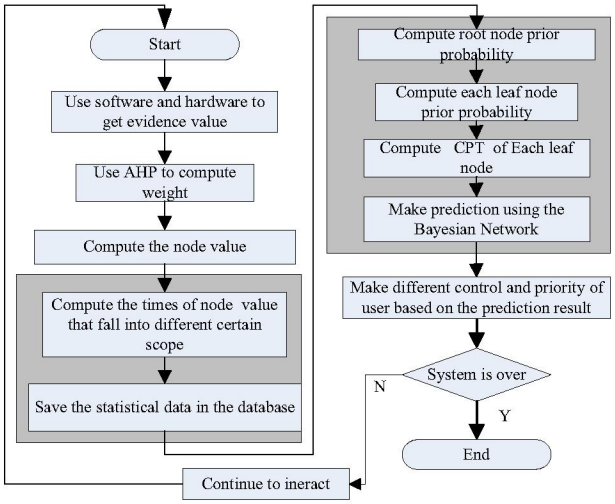

Fig. 3. Chart flow of evaluation, prediction and control of user behaviour

For single user, we can get probability distribution of different attribute through contrasting the prediction values of different attribute. Then the system can do corresponding control according to the prediction probability distribution. For multi-user, we can sort the prediction result at the same condition and set different PRI. Then the system can encourages "good behaviors" and punishes "bad behaviors" by network control mechanism based on prediction result. The whole computation flowchart about user behaviour evaluation, prediction and control is provided in Fig.3. The grey background parts are the primary content discussed in this paper.

IV. Analysis and Conclusions

In trustworthy network, the accuracy of prediction result mainly lies on following three aspects: the accuracy of evidence, the accuracy of comparison between every two evidences by the expert and the times of interaction that influences the prior-probability. When we predict the user behaviour we should consider these three aspects.

The user end not only is headwaters of producing and storing important data, but also headwaters from which most attack originates. So apart from current security systems, if we can evaluate, predict and control the user behaviour to make its behaviour accord with the regulation, we can avoid the attach accident and make more full security protection for the network. We provide a method, which can effectively predict future user behaviour based on past user behavior evidence for different qualification and different purpose. We can use the prediction result to control user behaviour and set up different priority for users with different prediction result, which can finally improve network security and user quality through encouraging the good deeds and punishing the bad deeds.

References

[1]. Cyber Trust, http://www.nap.edu/catalog/6161.html, 2005.

[2]. Cyber Trust Full Proposal http://www.nsf.gov/pubs/2005/nsf05518/nsf05518.htm, 2006.

[3]. Trusted Computing Group, https://www.trustedcomputinggroup.org/home 2005.

[4]. Lin Chuang, Peng Xue-Hai. Reseach on Trustworthy Networks.Chinese Journal of Computers. May.2005, 28(5): 751-758. (in Chinese)

[5]. Ayedemir, M., Bottomley, L. Two tools for network traffic analysis Computer Networks, Volume 36, Issue 2-3, July, 2001:169-179

[6]. Parra, Gilbert R., Martin, Christopher S. and Clark, Duncan B. An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks. Expert Systems With Applications, Volume 29, Issue 4, November, 2005: 713-722

[7]. Rabelo, L, Eskandari, H, Shalan, T. and Helal, M., Supporting Simulation-based Decision Making with the Use of AHP Analysis, 2005 Proceedings of the Winter Simulation Conference, Dec. 4, 2005:2042 – 2051.

[8]. Tracks usage of TCP/IP network subnets and builds html files with graphs to display utilization. http://bandwidthd.sourceforge.net/, 2006

[9]. NetFlow Monitor (NF) is tool for processing and evaluating NetFlow Exports from CISCO routers http://netflow.cesnet.cz/, 2006

[10]. Wenhui Liao, Weihong Zhang, Zhiwei Zhu ,Qiang Ji , A Real-Time Human Stress Monitoring System Using Dynamic Bayesian Network, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 2005,(3): 70 – 70.