Scientific Research and Essays Vol. 7 (10), pp. 1103-1112, 16 March, 2012
Available online at http://www.academicjournals.org/SRE
DOI: 10.5897/SRE11.1792
ISSN 1992-2248 ©2012 Academic Journals

Review

A Bayesian networks-based security risk analysis model for information systems integrating the observed cases with expert experience

Nan Feng and Jing Xie*

College of Management and Economics, Tianjin University, 300072 Tianjin, China.

Accepted 2 December, 2011

In the process of security risk analysis for information systems, establishing an appropriate model suitable for the target security risk problem is a crucial task that will ultimately influence the effectiveness of risk analysis results. For inducing a representative model for observed information systems, a security risk analysis model is proposed based on the knowledge from observed cases and domain experts. In this model, a Bayesian network (BN) is developed by integrating the database of observed cases with domain expert experience and knowledge. Based on the BN, the model facilitates the visibility and repeatability of the decision-making process of security risk analysis. Finally, the model is further demonstrated and validated via a case study.

Key words: information systems, risk analysis, Bayesian networks, probabilistic inference.

INTRODUCTION

Nowadays, security risk management is of vital importance for an enterprise to keep its information systems secure at an acceptable level, the key issues focus on both how to reduce the probability of risk occurrence and decrease the loss of risk consequence. The main tasks for the implementation of such requirements involve the determination of the causes of security risk, the estimation of risk occurrence probability, and the evaluation of risk consequence severity, which are all included in the security risk analysis. Nevertheless, the security risk analysis for information systems is a very critical challenge due to technical difficulties as well as, changes of environment.

In the process of security risk analysis for information systems, models are built in order to analyze and better understand the security risk factors and their causal relationships in real-world information systems. Establishing an appropriate model suitable for the target security risk problem is a crucial task that will ultimately influence the effectiveness of risk analysis results. In the existing literature, all the approaches either assumed that the structure of the model was provided by domain expert experience and knowledge, or assumed that the structure was chosen from some general well-known class of model structures, thus, the results of security risk analysis were relatively subjective (Cavusoglu et al., 2009).

To overcome these drawbacks, not only expert have the experience and knowledge that needs to be taken into account, but also, the database of observed cases from information systems should be utilized in the process of modeling.

Therefore, how to fuse the database of observed cases with domain expert experience and knowledge for inducing a representative model for observed information systems is a critical issue in security risk analysis. In this paper, we propose a security risk analysis model based on the knowledge from observed cases and domain experts. In this model, through structure learning and parameter learning, a Bayesian network (BN) is developed to simultaneously define the risk factors and their causal relationships. The effectiveness and accuracy of the model are demonstrated through a case study, which indicates that the model is able to improve the accuracy and efficiency of security risk analysis for information systems.

*Corresponding author. E-mail: fengnan@tju.edu.cn Tel: +86-22-27401021.

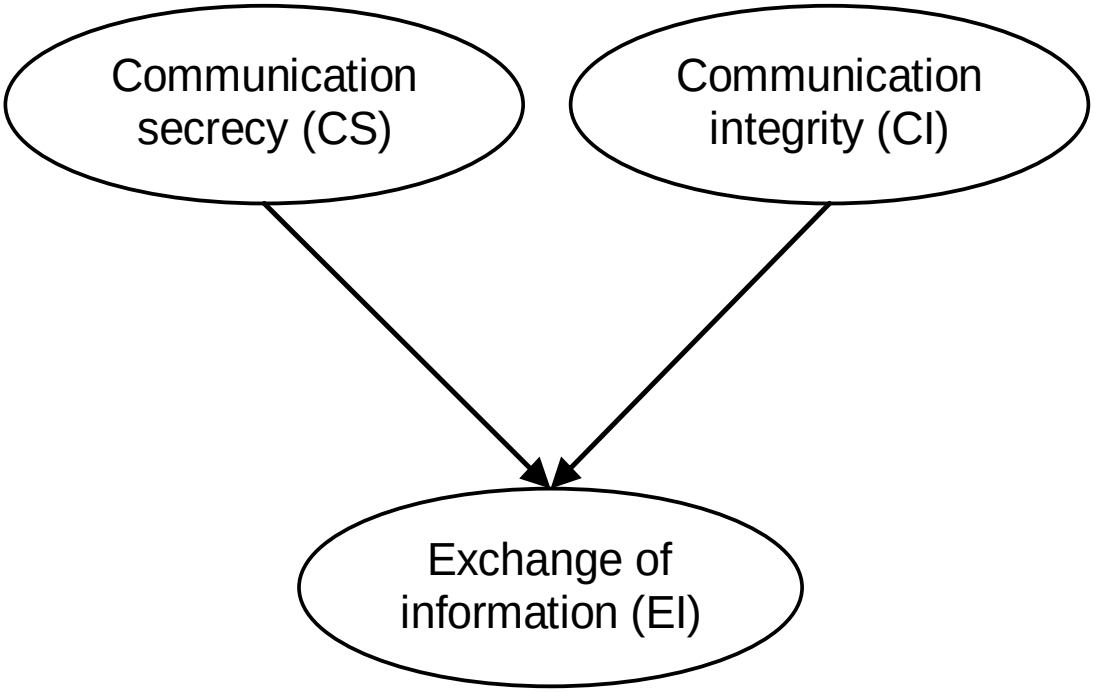

Figure 1. BN example.

LITERATURE REVIEW

Related work

In recent years, the security risk analysis for information systems has attracted much attention of researchers in the field of security risk management (Peltier, 2007; Karabacak and Sogukpinar, 2005). As information systems have become more complex in an organization, neither quantitative nor qualitative approaches can properly model the assessment process alone. Therefore, the comprehensive approaches combining both the quantitative and the qualitative approaches are needed (Alter and Sherer, 2004; Salmela, 2008). The approach based on the fuzzy comprehensive evaluation (FCE) (Yang et al., 2008; Ding and Chou, 2011) is a mathematical method to comprehensively evaluate the ISS risks by using fuzzy set theory of fuzzy mathematics. Although, this approach is good at processing the ambiguous information by simulating the characteristic of humans in making the judgment not capable to provide the graphical relationships among various ISS risk factors using flow charts or diagrams. Afterwards, Sun et al. (2006) proposed an evidential reasoning approach under the Dempster–Shafer theory for the risk analysis of information systems security. This approach provided a rigorous, structured manner to incorporate relevant security risk factors, related countermeasures, and their interrelationships when estimating security risk in information systems. In addition, sensitivity analyses were performed to evaluate the impact of important parameters on the model's results in this approach. Fan and Yu (2004) developed a Bayesian belief networks based procedure to provide an objective and visible support for risk analysis. This approach facilitated the visibility and repeatability of the decision-making process of risk management.

Bayesian networks

According to Jensen (2001), a BN N = (X, G, P) consists of:
1) A Directed Acyclic Graph (DAG) G = (*V*, *E*) with nodes $V = \{v_1, , v_n\}$ and directed links *E*.
2) A set of discrete random variables, X, represented by the nodes of G.
3) A set of conditional probability distributions, P, containing one distribution, $P\ (X_v\ |X_{parents(v)})$, for each random variable $X_v \in$ X.

To solve a BN N = (X, G, P) is to compute all posterior marginal probabilities given a set of evidence ε, that is; $P(X \mid \varepsilon)$ for all $X \in$ X. If the evidence set is empty, that is; $\varepsilon = \phi$, then the task is to compute all prior marginal probabilities, that is; $P(X)$ for all $X \in$ X. Figure 1 is a simple example, where Exchange of information (EI) is influenced by Communication secrecy (CS) and Communication integrity (CI). The Conditional Probability

Table 1. CPT of *P* (EI|CS, CI).

CS	CI	EI = secure	EI = insecure
Effective	Effective	1.0000	0
Average	Effective	0.8652	0.1348
Ineffective	Effective	0.3943	0.6057
Effective	Average	0.8199	0.1801
Average	Average	0.5581	0.4419
Ineffective	Average	0.1856	0.8144
Effective	Ineffective	0.3359	0.6641
Average	Ineffective	0.1092	0.8908
Ineffective	Ineffective	0	1.0000

Table (CPT) for Exchange of information is shown in Table 1.

In general, a BN models the constructor's belief. Based on this belief, it provides mathematical calculation and prediction. BNs have been widely applied in the field of medical diagnostics, classification systems, and software agents for personal assistants, multi-sensor fusion, and legal analysis of trials (Kjaerulff and Madsen, 2008).

SECURITY RISK ANALYSIS MODEL

The procedure of the proposed security risk analysis model is defined through three phases (Figure 2), which are the BN development, BN probabilistic inference, and security risk monitoring. In Figure 2, Database1 (DB1) contains the basic information about the BN nodes. Database2 (DB2) stores the case data of the BN nodes, and Database3 (DB3) has current observation data.

BN Development

It is assumed that some underlying process in the security risk management has generated a database of observed cases as well as domain expert experience and knowledge. The task of BN development is to fuse these information sources in order to induce a representative model of the underlying process. If the underlying process follows a probability distribution P_0, the goal of BN initialization is to identify a model representation of P_0. The probability distribution P_0 is assumed to be a DAG-faithful probability distribution with underlying DAG G_0. That is, we assume that the distribution P_0 can be represented as a BN.

The faithfulness assumption says that the distribution *P* induced by N = (X, G, P) satisfies no independence relations beyond those implied by the structure of G. A Bayesian network is faithful if and only if for every diverging connection there is a corresponding conditional dependence. The underlying probability distribution P_0 is assumed to be DAG-faithful with DAG G_0.

The database of cases generated by the underlying and unknown process is denoted D = $\{c^1, \ldots, c^N\}$ where *N* is the number of cases in the database. We assume D consists of independent and identically distributed data cases drawn at random from the probability distribution P_0, that is, we assume cases are drawn at random and independently from the same probability distribution P_0. Each case $c^i = (x_1^i, \ldots, x_n^i)$ in D specifies an assignment of a value x_j^i to each variable $X_j \in X$.

We consider learning a BN as the task of identifying a DAG structure G and a set of conditional probability distributions P with parameters Θ on the basis of D = $\{c^1, \ldots, c^N\}$ and possibly some domain expert background knowledge.

Greedy hill climbing (GHC) is one of the most competitive algorithms in terms of structural quality and runtime performance (Tsamardinos et al., 2006). In the implementation of GHC (Figure 3), it is conducted with a tabu list using the parameters of the original authors: the algorithm keeps a list of the last 100 structures and allows only changes that lead to a structure not contained in the list.

Based on the BN's structure, the set of conditional probability distributions of the BN can be defined. The parameters of this set of distributions may be set manually, but more often the parameters of the distributions will be estimated from the same database of cases as used by the structure learning algorithms.

In this study, we utilize the EM algorithm to realize the BN parameter learning, because it has been proven that it is an efficient approach for dealing with incomplete information when building statistical models (Little and Rubin, 1987). The EM algorithm is an iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. In ML estimation, we wish to estimate the model parameter(s) for which the observed data are the most likely. Each iteration of the EM algorithm consists of two processes: The E-step, and the M-step. In the expectation, or E-step, the missing data are estimated, given the observed data and current estimate of the model parameters. This is achieved using

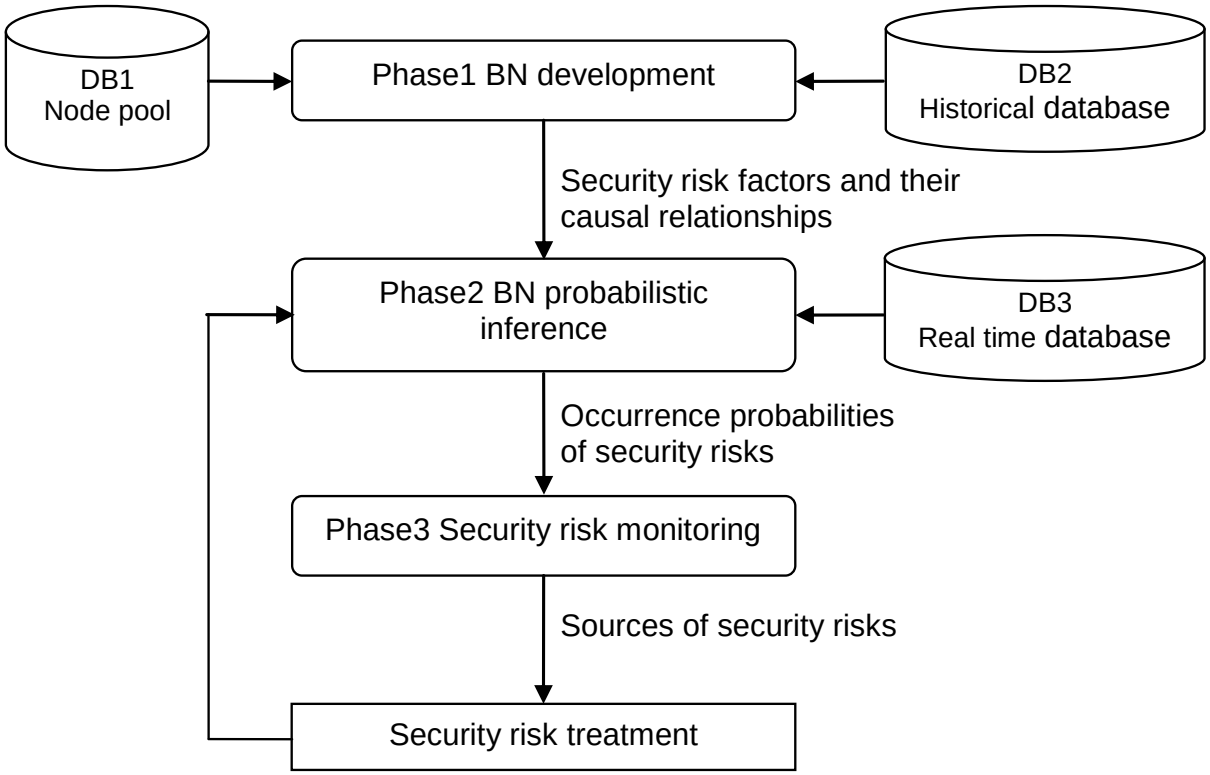

Figure 2. The procedure of security risk analysis mode

Algorithm GHC:

Input: Set of all/candidate edges

Output: Bayesian network

1. G_{total} = empty graph;
2. G = empty graph;
3. *TL* : tabu list (FIFO) with last 100 structures;
4. **repeat**
5. choose best action for G with resulting graph not in *TL* (possible actions: add, remove, revert);
6. apply best action to G;
7. add G to TL;
8. ΔS: score G - score G_{total};
9. **if** (ΔS>0) **then**
10. G_{total} = copy of G;
11. **until** G_{total} has not changed last 20 times;
12. **return** Bayesian network with structure G_{total}

Figure 3. The procedure of security risk analysis mode.

the conditional expectation, explaining the choice of terminology. In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The EM algorithm is discussed as shown: let *X* be random vector which results from a parameterized family. We wish to find θ such that *P* (*X*| θ) is a maximum. This is known as the Maximum Likelihood (ML) estimate for θ . In order to estimate θ , it is typical to introduce the log likelihood function defined as:

$$L(\theta) = \ln P(X \mid \theta) \quad . \tag{1}$$

The EM algorithm is an iterative procedure for maximizing *L* (θ). Assume that after the nth iteration the current estimate for θ is given by θ_n . Since the objective is to maximize *L* (θ), we wish to compute an updated estimate θ such that:

$$L(\theta) > L(\theta_n) \tag{2}$$

$$L(\theta) - L(\theta_n) = \ln P(X \mid \theta) - \ln P(X \mid \theta_n) \quad . \tag{3}$$

Considering the unobserved or missing variables, the Equation (4) can be rewritten as:

$$L(\theta) - L(\theta_n) = \ln\left(\sum_z P(X \mid z, \theta) P(z \mid \theta)\right) - \ln P(X \mid \theta_n) \quad , \tag{4}$$

Where, *z* is the hidden variables. According to McLachlan and Krishnan (1996), we have:

$$L(\theta) \geq L(\theta_n) + \Delta(\theta \mid \theta_n) \quad . \tag{5}$$

For convenience, we define

$$l(\theta \mid \theta_n) \square L(\theta_n) + \Delta(\theta \mid \theta_n) \quad , \tag{6}$$

So that the relationship between *L* (θ) and $l(\theta \mid \theta_n)$ can be made explicit as $L(\theta) \geq l(\theta \mid \theta_n)$.
In addition, we can also have:

$$\begin{aligned} l(\theta_n \mid \theta_n) &= L(\theta_n) + \Delta(\theta_n \mid \theta_n) \\ &= L(\theta_n) + \sum_z P(z \mid X, \theta_n) \ln \frac{P(X \mid z, \theta_n) P(z \mid \theta_n)}{P(z \mid X, \theta_n) P(X \mid \theta_n)} \\ &= L(\theta_n) + \sum_z P(z \mid X, \theta_n) \ln \frac{P(X \mid z, \theta_n)}{P(X \mid z, \theta_n)} \\ &= L(\theta_n) + \sum_z P(z \mid X, \theta_n) \ln 1 \\ &= L(\theta_n) \end{aligned} \tag{7}$$

So, for $\theta = \theta_n$ $l(\theta \mid \theta_n)$ and *L* (θ) are equal. Therefore, any θ which increases $l(\theta \mid \theta_n)$ will also increase *L* (θ). In order to achieve the greatest possible increase in the value of *L* (θ), the EM algorithm calls for selecting θ such that $l(\theta \mid \theta_n)$ is maximized. We denote this updated value as θ_{n+1} .
Formally, we have:

$$\begin{aligned} \theta_{n+1} &= \arg\max_{\theta} l(\theta \mid \theta_n) \\ &= \arg\max_{\theta} E_{z \mid X, \theta_n} \ln P(X, z \mid \theta) \end{aligned} \quad . \tag{8}$$

The EM algorithm can thus be conveniently summarized as:

1. E-step: Determine the conditional expectation $E_{z \mid X, \theta_n} \ln P(X, z \mid \theta)$.
2. M-step: Maximize this expression with respect to θ .

BN Probabilistic Inference

Whenever the new evidence is available in this phase, it should be plugged in the BN to update previous estimates by probabilistic inference. In BNs, probabilistic inference can be defined as the task of computing all posterior marginals of non-evidence variables given the evidence.

In this phase, we develop an inference engine based on junction tree (also known as a join tree or a Markov tree) (Jordan, 1999) to compute the posterior marginal *P* (*X* | ***ε***) of a variable *X* approximately, given the evidence ***ε***. A junction tree representation T of a Bayesian network N = (X, G, P) is a pair T = (C, S) where C is the set of cliques and S is the set of separators. The clique set C indicates the nodes of T, whereas the separators S annotate the links of the tree. Each clique *C* ø C represents a maximal complete subset of pair wise connected variables of X that is; *C* ⊆ X. Once the junction tree T = (C, S) has been constructed, a probability potential is associated with each clique *C* ø C and each separator *S* ø S between two adjacent cliques C_i and C_j where $S = C_i \cap C_j$. The inference engine is performed using a message passing algorithm on the junction tree (Kjaerulff and Madsen, 2008). Its process involves the following steps:

(1) Each item of evidence must be incorporated into the junction tree potentials. For each item of evidence, some potential containing the variable in target problem is modified to reflect the evidence.
(2) A clique of the junction tree is selected. This clique is referred to as the root of the inference.
(3) Then messages are passed towards the selected root.

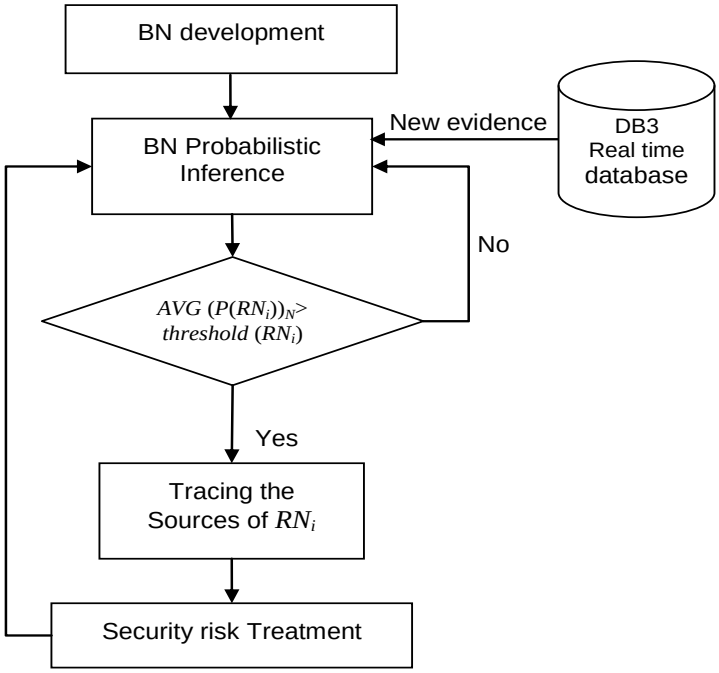

Figure 4. The procedure of security risk monitoring.

The messages are passed through the separators of the junction tree (that is along the links of the tree). These messages cause the potentials of the receiving cliques and separators to be updated.
(4) The messages are passed in the other direction (that is, from the root towards the leaves of the junction tree).
(5) At this point, the junction tree is said to be in equilibrium: The probability $P(X \mid \varepsilon)$ can be computed from any clique or separator containing X. The result will be independent of the chosen clique or separator.

Prior to the initial round of message passing, for each variable $X_v \in \mathbf{X}$ we assign the conditional probability distribution $P(X_v \mid X_{pa(v)})$ to a clique C such that $X_{pa(v)} \subseteq C$. Once all conditional probability distributions have been assigned to cliques; the distributions assigned to each clique are combined to form the initial clique potential.

Security risk monitoring

We categorize the BN nodes into two groups: Risk-Factor Nodes (RFN) and Risk Nodes (RN). Risk probability threshold for every risk node can be assigned by expert knowledge. In this phase, whenever new evidences are obtained, they are plugged into the BN to update estimates. A chronological record of such inputs and estimates are kept as a security risk profile saved in the security risk database. In other words, a risk profile can be viewed as historical snapshots of these BN's images.

In the process of security risk monitoring, if the average probability of the previous N time units of the profile records for RN_i exceeds the threshold for it, that is; $AVG(P(RN_i))_N > threshold(RN_i)$, the sources of RN_i will be traced. Since a BN's structure can visually model cause consequence relations, the observed RN_i can be traced to its ancestors interactively with the user to identify the sources of the security risk. The procedure of security risk monitoring is given in Figure 4. After the security risk treatment, new evidences are plugged into the BN to update previous estimates. Therefore, BN calculation can also be used to predict the effectiveness of risk treatment decisions.

CASE ANALYSIS

Here, the proposed model is applied to a real company's information systems, which has been in service for six years to manage its security risks.

The details of the case study are discussed next. Six domain experts, two of whom are also security managers of the Company were interviewed to select the security risk related variables, that is, the nodes in BN which are verified to significantly affect the security risk of information systems based on their experience. The information of risk nodes and risk factor nodes is described in Table 2.

Based on the application architecture of the proposed security risk analysis model, the BN structure was

Table 2. Information of nodes in the BN.

Risk node	State space	Risk factor node	State space
Physical and environment security risk	High; medium; low	Physical entry controls	High level; average; need to be improved
		Secure areas	Secure; average; insecure
		Cabling security level	High; medium; low
		Equipment maintenance	Regular; irregular
		Equipment security level	High; medium; low
	High; medium; low	Network connection control	Effective; average; ineffective
		Network routing control	Effective; average; ineffective
		Network access control	Effective; average; ineffective
		Network intrusion protection	Effective; average; ineffective
		Network security audit	Comprehensive; incomprehensive
Termination security risk	High; medium; low	Termination access control	Effective; average; ineffective
		Termination security audit	Comprehensive; incomprehensive
		Termination intrusion protection	Effective; average; ineffective
Operation security risk	High; medium; low	Documented operating procedures	Good; not good; bad
		Change management	Effective; average; ineffective
		Segregation of duties	Clear; unclear
		Operational procedures and responsibilities	Very standard; standard; non-standard
		Exchange of information	Secure; average; insecure

developed. In the BN structure learning, we tested different number of iterations and found that the performance of GHC did not improve significantly with the number of iterations larger than 150. Thus, the maximum number of iterations was set to 150. Taking the network security risk for example, the BN structure of network security risk is shown in Figure 5.

Based on the BN structure, we adopted the EM algorithm to perform the BN parameter learning. Then, we verified the validity of the BN parameter learning from two aspects of the algorithm calculation accuracy and the algorithm convergence. KL distance (Kullback et al., 1987), a natural distance function from a "true" probability distribution to a "target" probability distribution, is used to test the calculation accuracy of the EM algorithm. Take note of network access control as an example. Figure 6 illustrates the KL distance of the node's parameter learning.

In Figure 6, we observe that the KL distance is narrower when the sample number increases. Therefore, it proves that the result of parameter learning is precise.

For the algorithm convergence, the reaching relationship between the cycle number and the likelihood function is applied to analyze the convergence of the algorithm.

As presented in Figure 7, the maximum likelihood function is assumably estimated through 75 cycles in the algorithm, which indicates that the convergence of the algorithm is good. The new evidence was obtained from real time database, which gives updated information about each observable node in the BN as inference evidence. Based on the principle presented in the phase of BN probabilistic Inference, we compute the posterior probability of the nodes in the BN based on the evidence.

For the risk nodes in the BN, the probabilities of risk occurrence and the severities of risk consequence estimated by security risk assessment are shown in Table 3. In Table 3, the severity of each security risk consequence was investigated based on its influence to the customer, economy and internal environment, such as the duration of service interruption, economic loss, interference with users' work, and the cost of service recovery. Expert rating method and statistic analysis are adopted to determine the security risk consequences in information systems. From Table 3, we observed that the probability of network security risk is higher than the threshold 0.4 set by experts in advance. Based on the phase of security risk monitoring, the potential source of this problem was traced to be the nodes of network routing control and network connection control. And then, security risk treatment was performed to reduce the security risk level. After above activities, we performed the BN probabilistic inference again, and found that all the occurrence probabilities of the risk nodes in the BN

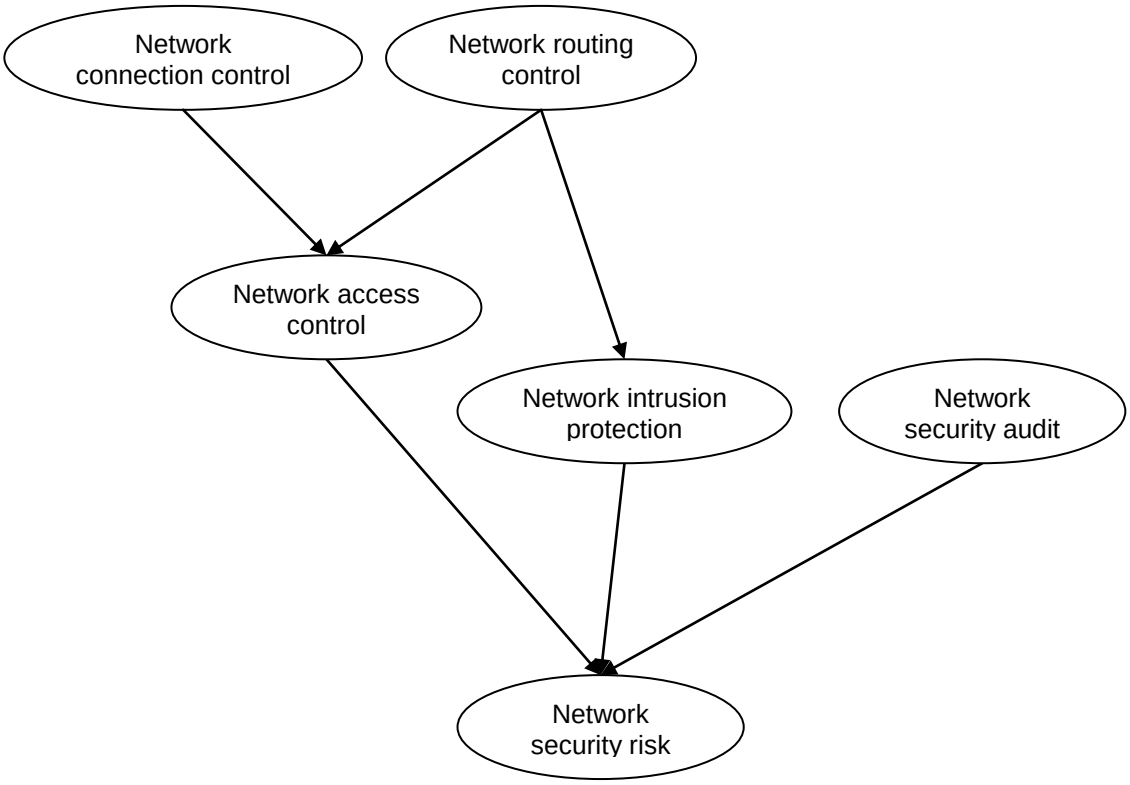

Figure 5. BN structure of network security risk.

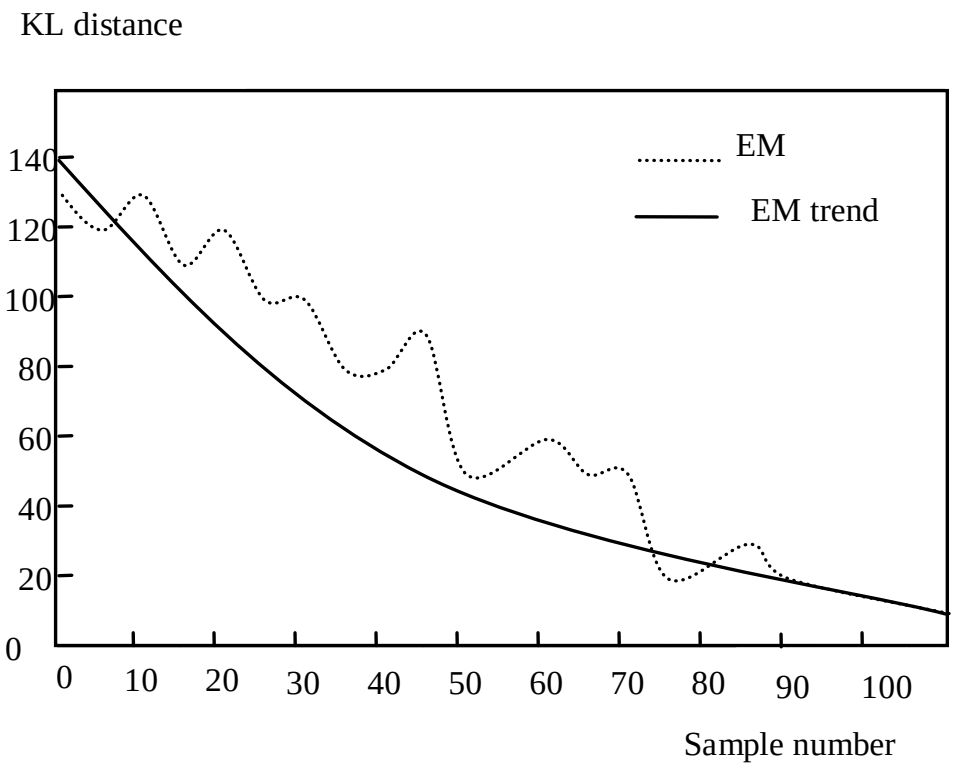

Figure 6. KL distance of parameter learning.

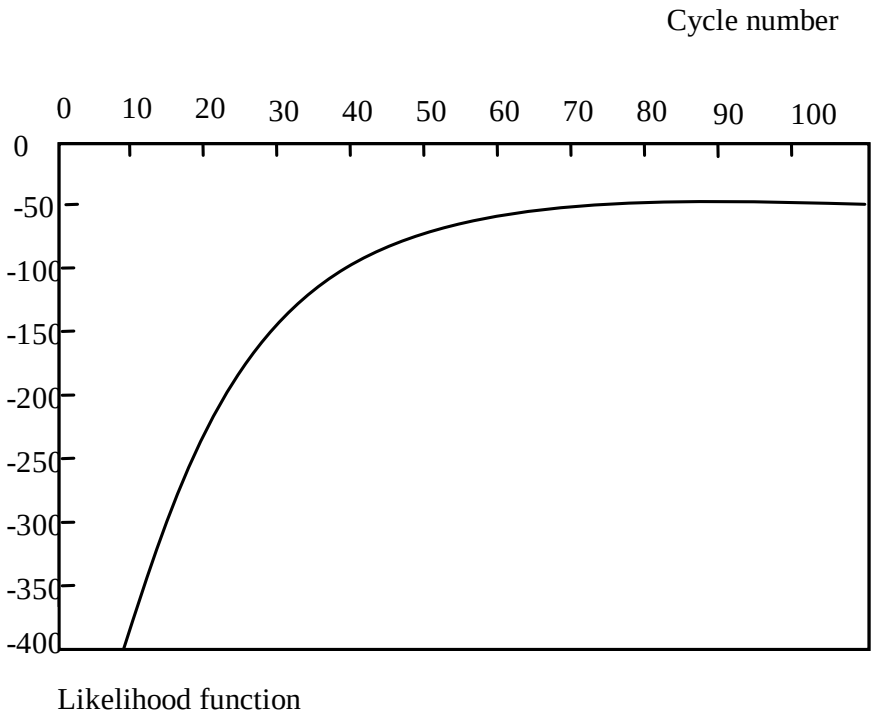

Figure 7. The reaching relationship between the cycle number and the likelihood function.

Table 3. The probabilities of risk occurrence and the severities of risk consequence.

Risk nodes	State	Probability	Severity
Physical and environment security risk	High	0.1988	0.9
	Medium	0.5159	
	Low	0.2853	
Network security risk	High	0.5816	0.8
	Medium	0.2933	
	Low	0.1251	
Termination security risk	High	0.2552	0.7
	Medium	0.4571	
	Low	0.2877	
Operation security risk	High	0.3201	0.8
	Medium	0.4798	
	Low	0.2001	

were lower than the threshold set by experts in advance. It is thus, verified that the model proposed in the study was valid on real data.

CONCLUSION

In this paper, a security risk analysis model is proposed to induce a representative model for observed information systems. In the model, a Bayesian network is developed by integrating the database of observed cases with domain expert experience and knowledge. Based on the BN, the model can facilitate the visibility and repeatability of the decision-making process of security risk analysis. The effectiveness and accuracy of the model are demonstrated through a case study, which

indicates that the model is able to improve the accuracy and efficiency of security risk analysis for information systems.

Future researches will focus on applying the proposed model to other practice situations, and incorporating more sophisticated constraints into the model to enhance the handling of more complex problems.

ACKNOWLEDGMENT

The research was supported by the National Natural Science Foundation of China (Grant No. 70901054). The authors are very grateful to all anonymous reviewers whose invaluable comments and suggestions substantially helped improve the quality of the paper.

REFERENCES

Alter S, Sherer S (2004). A general, but readily adaptable model of information system risk. Commun. AIS, 14(1): 1-28.

Cavusoglu H, Mishra B, Raghunathan S (2009). The effect of Internet security breach announcements on market value: Capital market reactions for breached firms and Internet security developers. Int. J. Electron. Comm., 14 (3): 69-104.

Ding JF, Chou CC (2011). A fuzzy MCDM model of service performance for container ports. Sci. Res. Essays, 6(3): 559-566.

Fan C, Yu Y (2004). BBN-based software project risk management. J. Syst Software, 73(2): 193-203.

Jensen FV (2001). Bayesian networks and decision graphs. Springer-Verlag, New York.

Jordan MI (1999). Learning in Graphical Models. Cambridge, MA: MIT Press.

Karabacak B, Sogukpinar I (2005). ISRAM: information security risk analysis method. Comput. Secur., 24(2): 147-159.

Kjaerulff UB, Madsen AL (2008). Bayesian networks and influence diagrams: a guide to construction and analysis. New York: Springer Science.

Kullback S, Burnham KP, Laubscher NF, et al. (1987). Letter to the Editor: The Kullback–Leibler distance. Am. Stat., 41(4): 340-341.

Little R, Rubin D (1987). Statistical analysis with missing data. New York, John WIley & Sons.

McLachlan G, Krishnan T (1996). The EM Algorithm and Extensions. New York, John Wiley & Sons.

Peltier T (2007). Information security risk analysis. 2nd ed., Boca Raton, FL: Auerbach Publications.

Salmela H (2008). Analysing business losses caused by information systems risk: a business process analysis approach. J. Inf. Technol., 23(3): 185-202.

Sun L, Srivastava RP, Mock TJ (2006). An information systems security risk assessment model under the Dempster-Shafer theory of belief functions. J. Manage. Infor. Syst., 22(4): 109-142.

Tsamardinos I, Brown F, Aliferis F (2006). The max-min hillclimbing BN structure learning algorithm. Mach. Learning, 65(1): 31-78.

Yang X, Luo H, Fan C, Chen M, Zhou S (2008). Analysis of risk evaluation techniques on information system security. J. Comput. Appl., 28(8): 1920-1924.