A Data-driven Assessment Model for Information Systems Security Risk Management

Nan Feng
Tianjin University, Tianjin, China
Email: fengnan@tju.edu.cn

Xue Yu*
Tianjin University, Tianjin, China
Email: tjufengnan@gmail.com

***Abstract*—In this paper, a data-driven assessment model for information systems security risk management is proposed based on the knowledge from observed cases and domain experts. In the model, genetic algorithm is applied to search the rules of security risk identification based on historical data. For identifying the causal relationships of risk factors and predict the occurrence probability of security risk, a Bayesian network (BN) is developed. Structure learning and parameter learning are utilized to integrate the database of observed cases with domain expert experience in the development of the BN. The significance of the work is that the model provides more objective and visible support for security risk assessment in the information systems.**

***Index Terms*—information systems, risk management, Bayesian networks, genetic algorithm**

I. INTRODUCTION

Information systems security risk assessment involves identifying and analyzing the security risks in the information systems. Risk identification is the basic step, on which other steps are based in security risk management. The main methods and tools used in risk identification are the decision-making tree, the fault tree, the risk source list, and so on. Nevertheless, these methods are mainly based on the experience of experts, so they are inappropriate for a large-scale application. To address this issue, we employ genetic algorithm (GA) [1] to search the rules of risk identification in the information systems based on historical data. Compared with other risk identification methods, GA is more suitable for large-scale parallel computation. As a result, it is more effective when applied in complicated systems like information systems. Meanwhile, GA starts to search from a group of initial points rather than a single point, which greatly increase the possibility of the acquisition of the global optimum solution.

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 [2]. Nevertheless, in the existing literature, all the approaches [3-5] 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.

To overcome these drawbacks, not only expert experience need 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 target information systems is a critical issue in security risk assessment. In this paper, we propose a security risk assessment model based on the knowledge from observed cases and domain experts. 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.

The rest of this paper is organized as follows: Section 2 is the literature review on risk assessment for information systems security. In the next section, we discuss the application architecture of the proposed security risk assessment model in detail. After that, the proposed approach is further demonstrated and validated in Section 4 via a case study. Finally, we summarize our contributions and present further research.

II. LITERATURE REVIEW

Gordon and Loeb [6] was the first to present a mathematical model that determined the optimal security investment level for the information systems. Their work and subsequent literatures on security risk analysis focused on a single system or a single type of protection technology. Yue et al. [7] extended those studies by formulating and solving the problem according to the risk management paradigm, and therefore provided manager

* Corresponding author. Tel.: +86 22 27401021.

doi:10.4304/jcp.7.12.3103-3109

with additional insights into making optimal decisions. Grunske and Joyce [8] proposed a risk-based approach that created modular attack trees for each component in information systems. These modular attack trees were specified as parametric constraints, which allowed quantifying the probability of security breaches that occurred due to internal component vulnerabilities as well as vulnerabilities in the component's deployment environment. The above quantitative approaches can be more efficient than qualitative approaches if mathematical models are established exactly. However, security risk assessment for information systems which utilizes intensive quantitative metrics would become infeasible for an organization because current information systems have more complex structure and widespread usages.

Chen et al. [3] applied the similarity measures of generalized fuzzy numbers to deal with fuzzy risk analysis problems. Although this approach is good at processing the ambiguous information by simulating the characteristic of human in making judgments, it is unable to provide the graphical relationships among various security risk factors using flow charts or diagrams. For representing the relationships among risk factors, Fan and Yu [5] developed a Bayesian networks (BNs) based procedure to provide risk analysis support. This approach facilitates the visibility and repeatability of the decision-making process of risk management. However, in Fan's approach, the BN is structured only based on domain experts' experience. So the results of security risk analysis are relatively subjective. Feng and Li [9] proposed an information security risk assessment model based on the improved evidence theory. In order to deal with the uncertain evidence found in the process of risk assessment, this model provides a new way to define the basic belief assignment in fuzzy measure. Moreover, the model also provides a method of testing the evidential consistency, which can reduce the uncertainty derived from the conflicts of evidence.

III. The Proposed Security Risk Assessment Model

The application architecture of the proposed security risk assessment model is defined through three phases, which are security risk identification, Bayesian network development, and security risk analysis. And, the architecture of the model is given in Fig. 1 (Security risk treatment is not involved in the system discussed in this paper).

A. Security Risk Identification

The security risk rule can be represented as follows:

$$R_j: \ IF \ \ rf_{1j} \ \& \ rf_{2j} \ \& \ldots \& \ rf_{nj} \ \ THEN \ \omega . \qquad (1)$$

The rf_{ij} (i=1,2,…,n) represents the attribute values of the security risk rule precondition; ω represents the conclusion of the security risk rule. The "attribute" corresponds with the security risk factor; the "conclusion" corresponds with the security risk in the process of security risk identification for information systems.

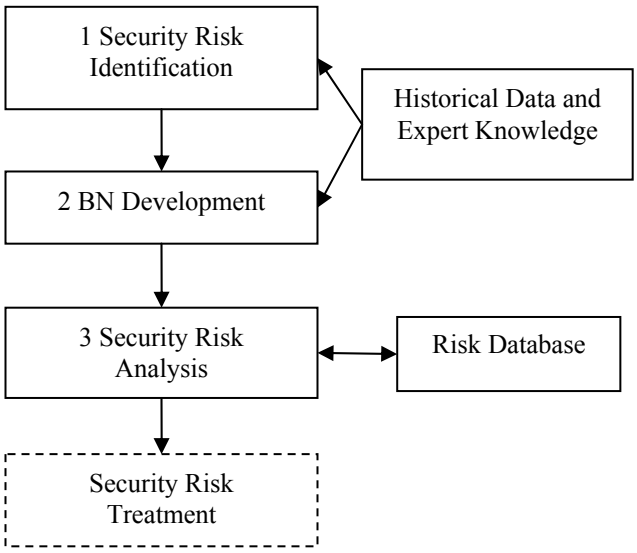

Figure 1. Architecture of the model

The process of security risk identification based on genetic algorithm is shown in appendix A.

B. Bayesian Network Development

For the representation of the causal relationships among security risk factors, a K2-based algorithm is developed to learn the BN structure.

Input: A set of n nodes, an ordering on the nodes, an upper bound u on the number of parents a node may have, and a database D containing m cases
Output: For each node, a printout of the parents of the node
1. **for** i: = 1 to n do
2. $\pi_i := \square$;
3. $P_{old} := f(i, \pi_i)$;
4. OKToProceed := **true**;
5. **while** OKToProceed and $|\pi_i| < u$ do
6. let z be the node in (Pred(x_i) -π_i) that maximizes $f(i, \pi_i \cup \{z\})$;
7. $P_{new} := f(i, \pi_i \cup \{z\})$;
8. **if** $P_{new} > P_{old}$ **then**
9. $P_{old} := P_{new}$;
10. $\pi_i := \pi_i \cup \{z\}$;
11. **else** OKToProceed := **false**;
12. **end** {**while**};
13. **write**('Node: ', x_i, 'Parent of x_i: ', π_i);
14. **end** {**for**};
15. **end** {**K2**};

Figure 2. K2-based algorithm

K2 algorithm is the most famous score-based algorithm for BN network structure construction from data in the last two decades. Specifically, it recovers the underlying distribution in the form of DAG efficiently. Fig. 2 is the procedure of K2-based algorithm [10].

In above algorithm, $f(i, \pi_i)$ is defined as:

$$f(i,\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} \alpha_{ijk}! \qquad (2)$$

where π_i is a set of parents of node x_i, $q_i = |\phi_i|$, $r_i = |V_i|$, V_i is a list of all possible values of the attribute of x_i, $\square_i$ is a list of all possible instantiations of the parents of x_i in database D, and α_{ijk} is a number of cases (i.e. instances) in D in which the attribute x_i is instantiated with its k^{th} value, and the parents of x_i in π_i are instantiated with the j^{th} instantiation in $\square_i$.

As for the BN parameters (i.e., the conditional probability tables), they can be determined by learning the parameters on historical data and expert's knowledge. In this paper, we use maximum likelihood estimation (MLE) [11] to calculate the conditional probability tables of each node in the BN from the complete data.

C. Security Risk Analysis

Once the BN of information systems is constructed, it serves as a tool for security risk analysis based on real time database, which provides updated information about each observable node in the BN as inference evidence. This phase finally comes up with the occurrence probability and the consequence severity of security risk in the BN. The results from the risk analysis will be used for the decision-making procedure: if the future estimated situation of information systems is a state considered ‘‘secure” or ‘‘successful”, no action should be taken. Otherwise, if the probability of one security risk node in the BN exceeds the threshold set in advance, the risk treatment should be started up.

In Bayesian Networks, probabilistic inference can be defined as the task of computing all posterior marginals of non-evidence variables given the evidence. In general, probabilistic inference is a NP-hard task. Therefore, the most critical task related to the risk assessment can be defined as identifying the posterior probability of each security risk based on the evidence obtained from real time database. In this paper, junction tree (also known as a join tree or a Markov tree) [12] is applied to compute the posterior marginal $P(X \mid \varepsilon)$ of a variable X approximately.

Based on the influence to information systems, the organization, and the system users, such as the duration of service interruption, economic loss, interference with customers' work, and the cost of service recovery, expert rating method and statistic analysis are adopted to determine the severity of security risk consequences for information systems.

IV. Case Analysis

In order to further validate the proposed model, we used it in assessing an actual company's information systems. This company is a Chinese financial services firm providing on-line services in securities trading and sales. In this section, we demonstrate the presented system via a case study according to the procedure of Section 3.

Table I and Table II are related 34 kinds of security risk factors extracted from the risk database and risk identification instances. The security risk rules would be found out by genetic algorithm based on these data.

TABLE I.
Security Risk Factors In The Information Systems

No.	Risk Factor	No.	Risk Factor
1	Protecting against external and environmental threats	18	Host/Server intrusion protection
2	Physical entry controls	19	Application access control
3	Physical security perimeter	20	Application security audit
4	Secure areas	21	Capability of fault tolerance
5	Supporting utilities	22	Data secrecy
6	Cabling security level	23	Data integrity
7	Equipment maintenance	24	Data back-up policy
8	Equipment security level	25	Documented operating procedures
9	Network connection control	26	Change management
10	Network routing control	27	Segregation of duties
11	Network access control	28	Operational procedures and responsibilities
12	User authentication for external connections	29	Communication secrecy
13	Network intrusion protection	30	Communication integrity
14	Network security audit	31	Exchange of information
15	User identification and authentication	32	Audit logging
16	Host/Server access control	33	Protection of log information
17	Host/Server security audit	34	Monitoring

During the process of the security risk identification, supposed that the population scale n = 34, crossover probability $p_c = 0.8$, mutation probability $p_m = 0.05$, times of iteration T = 100. For each kind of risk, choose the one that has the best function form the searched rule collection as the identification rule for the security risk. Table III is the search results.

TABLE II.
SECURITY RISK IDENTIFICATION INSTANCES

Risks	Risk Identification Instances
1.Physical and environment security risk	(1,4,8) (1,2,4,8) (3,6,8) (1,2,4,6) (2,4,8,11) (4,8) (1,8,34) (5,6,7,8) (2,4,5,6,8) (3,4,7,8) (1,4,6,8)
2.Network security risk	(9,11,13, 14) (13,14) (10,11,13,14,26) (11,13,14) (10,13,14) (11,14) (9,10,11,12) (10,11,12,13) (9,10,12,14) (10,11,34) (12,13,14)
3.Host/Server computer security risk	(15,16,17,18) (8,11,16,17) (8,17,18) (11,16,17,18) (11,17,18) (8,11,15,16,17) (16,17,18) (8,11)
4.Application security risk	(19,20,21) (20,21,31) (19,20) (20,21) (19,31) (15,19,20) (21,31,34) (19,21,31) (21,31) (19,31)
5.Data security and back-up risk	(22,23) (23,24) (22,24,25) (4,22,23,24) (22,24) (8,23,24) (22,23,24) (22,23,34) (23,24,31)
6.Communication and operation security risk	(25,28,29,31,34) (28,31) (26,28,31,34) (31,34) (26,28,31,32,34) (28,30,31,34) (27,28,29,31,34) (28,34) (26,28,34)

TABLE III.
SEARCH RESULTS

No.	The value of attribute	Risks
1	(10010101000000000000000000000000)	Physical and environment security risk
2	(00000000001011000000000000000000)	Network security risk
3	(00000001001000011100000000000000)	Host/Server computer security risk
4	(00000000000000000011100000000001000)	Application security risk
5	(00000000000000000000011100000000000)	Data security and back-up risk
6	(00000000000000000000000000001001001)	Communication and operation security risk

According to results of security risk identification, we built the BN for risk assessment based on the related historical data set including 200 cases from the Company. The K2-based algorithm and the MLE were utilized to identify the BN's structure and parameter. Due to space limitations, we take example for a local BN structure shown in appendix A. The node, "Communication and operation security risk", was a risk node, and other nodes were risk-factor nodes. The threshold of the risk node was set as 35% by expert experience.

From May 2011 to July 2011, the new evidence was obtained from the real time database is shown in Table IV, which gives updated information about each observable node in the BN as inference evidence.

TABLE IV.
THE EVIDENCE OBTAINED FROM REAL TIME DATABASE

Evidence	Related nodes
During the past two months several suspicious incidents related to network have not been logged.	Network security audit
There are a few other processes aside from normal browser caching, which store, alter or copy information	Host/Server access control
No restrictions on connection times to provide additional security for high-risk applications.	Application access control
There is no security policy document that details the procedure of changes to systems.	Change management
Several requests for data are not channeled through a DBA who then requests from operation staff.	Segregation of duties

We kept monitoring the security risk node, i.e. "Communication and operation security risk". After a period of time, it was found that the average probability of the previous *N* time units of the profile records for the risk node exceeded 35%. Based on the Bayesian network structure, the potential source of this problem was traced to be the security risk node of "Change management". And then, security risk treatment was performed to reduce the security risk level.

V. CONCLUSIONS

In this paper, a data-driven assessment model for information systems security risk management is proposed based on the knowledge from observed cases and domain experts. The proposed system utilizes genetic algorithm to search the rules of risk identification based on historical data in order to identify the security risks in information systems. Based on the results of risk identification, a BN is developed to predict security risks, identify sources of risks, and take proper measure to reduce risk occurrence probability in the information systems. Structure learning and parameter learning are utilized to integrate the database of observed cases with domain expert experience in the development of the BN. Finally, the proposed model is further demonstrated and validated via a case study, which indicates that the model is able to improve the accuracy and efficiency of security risk analysis for information systems.

APPENDIX A K2-BASED ALGORITHM AND BN STRUCUTRE

1. initialize population *P*(0);

Security risk rule precondition is represented by a binary string, where 1 indicates that it shows the attribute represented by the bit, and 0 indicates not. The ordered string constituted of 0 and 1 represents the attribute combination that can be applied to distinguish a certain classified type.

2. for each P_i in *P*(0)

3. compute Fitness(P_i);

The fitness function consists of the similarity function and the information measure function.

4. while (max Fitness(P_i) < Fitnessthreshold)

5. do Selection();

The expected number of each chosen individual relates to the ratio of its fitness value and population average fitness value. With regard to a given population $P = \{a_1, a_2, \cdots, a_n\}$ whose scale is n and an individual $a_j \in P$ whose fitness value is $f(a_j)$, the selection probability is $p_s(a_j) = f(a_j) / \sum_{i=1}^{n} f(a_j)$.

6. do Crossover();

Firstly, select a pair of individuals for copulation at random from the copulation pool; then, according to the length of bit string, *L*, randomly select one or more integer *k* in [1, *L*-1] as the crossover position from the selected pair; at last, do crossover according to the crossover probability $p_c (0 < p_c \le 1)$. Those individuals exchange part of each other's content at the crossover position to form a new pair of individuals.

7. do Mutation();

Mutation operators reverse the binary char value in the allele according to the mutation probability $p_m (0 < p_m \le 1)$. The mutation process for a given chromosome bit string $s = a_1 a_2 \cdots a_L$ is $O(p_m, x) : a_i' = \begin{cases} 1\text{-}a_i, & x_i \le p_m \\ a_i, & otherwise \end{cases}$ $i \in \{1, 2, \cdots, L\}$. The new individual generated is $s' = a_1' a_2' \cdots a_L'$, among which x_i corresponds to the even stochastic variables generated by each gene, $x_i \in [0,1]$.

8. end {while};

9. update population;

New risk rules replace the original ones, and then turn to the procedure of definition of fitness function to continue the optimization. At last, under the elite-reserved selection strategy, the terminate condition is determined by the alternation of the best individual in each generation.

10. for each h in population

11. compute Fitness(h);

12. return best fitness value;

Figure A1. The process of security risk identification

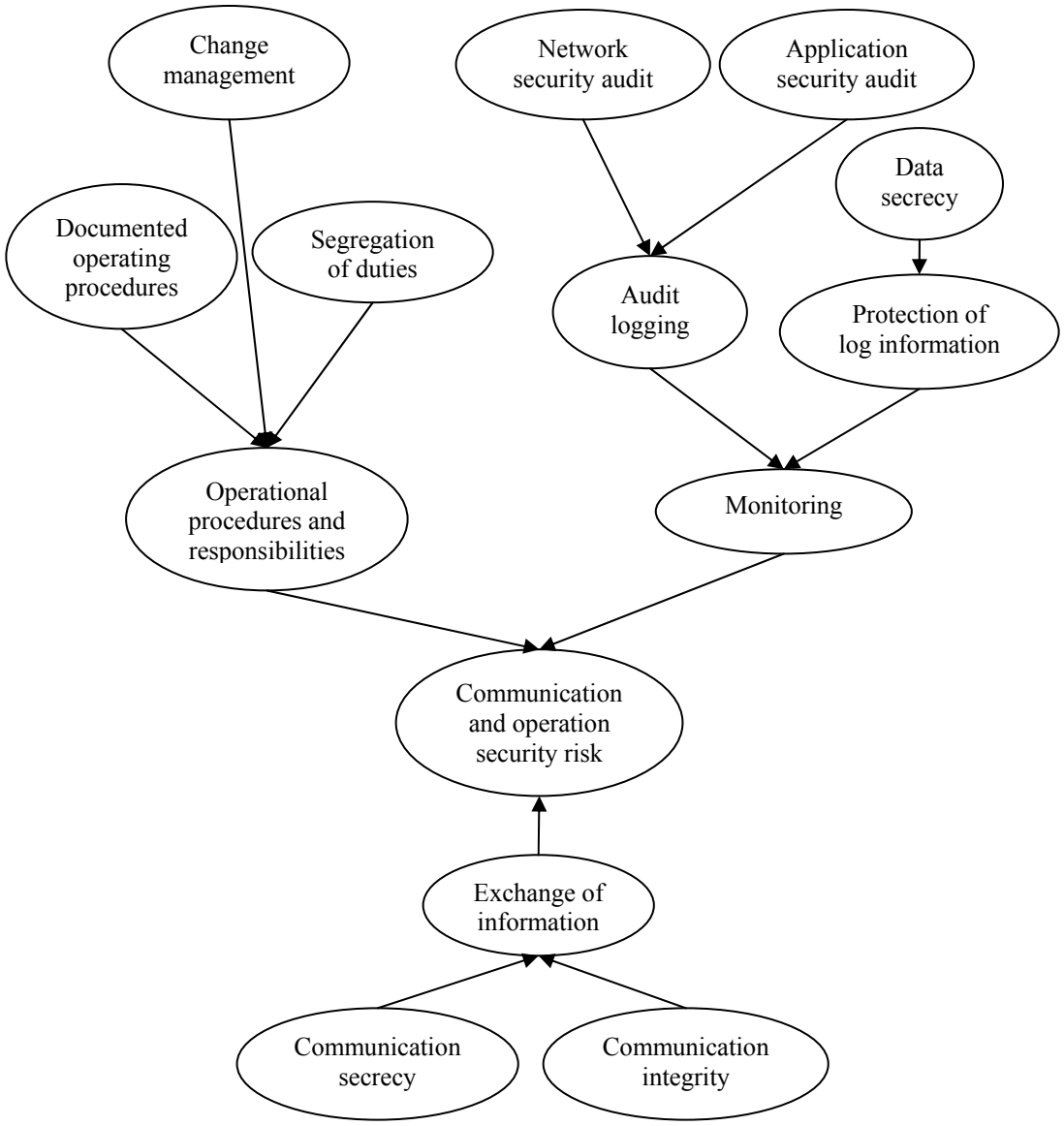

Figure A2. BN structure related to communication and operation security risk.

ACKNOWLEDGMENT

The research was supported by the National Natural Science Foundation of China (No. 70901054 and No. 71271149). The authors are very grateful to all anonymous reviewers whose invaluable comments and suggestions substantially helped improve the quality of the paper.

REFERENCES

[1] J. Kratica, T. Kostić, D. Tošić, D. Dugošija, and V. Filipović, "A Genetic Algorithm for the Routing and Carrier Selection Problem," *Computer Science and Information Systems*, Vol. 9(1), pp. 49-62, 2012.

[2] Y. Dong, J. Xu, Y. Xu, and W. Xu, "The on-line rental problem under risk-reward model with probabilistic forecast," *Information-An International Interdisciplinary Journal*, Vol. 14(1), pp. 89-96, 2011.

[3] S. J. Chen and S. M. Chen, "Fuzzy risk analysis based on similarity measures of generalized fuzzy numbers," *IEEE Transactions on Fuzzy Systems*, Vol. 11(1), pp. 45-56, 2003.

[4] L. Sun, R. P. Srivastava, and T. J. Mock, "An information systems security risk assessment model under the Dempster-Shafer theory of belief functions," *Journal of Management Information Systems*, Vol. 22(4), pp. 109-142, 2006.

[5] C. Fan and Y. Yu, "BBN-based software project risk management," *Journal of Systems and Software*, Vol. 73(2), pp. 193–203, 2004.

[6] L. A. Gordon, M. P. Loeb, and W. Lucyshyn, *CSI/FBI computer crime and security survey*, Computer Security Institute, San Francisco, USA, 2010.

[7] W. T. Yue, M. Çakanyildirim, Y. Ryu, and D. Liu, "Network externalities, layered protection and IT security risk management," *Decision Support Systems*, Vol. 44(1), pp. 1-16, 2007.

[8] L. Grunske and D. Joyce, "Quantitative risk-based security prediction for component-based systems with explicitly modeled attack profiles," *Journal of Systems and Software*, Vol. 81(8), pp. 1327-1345, 2008.
[9] N. Feng, M. Li, "An information systems security risk assessment model under uncertain environment," *Applied Soft Computing*, Vol. 11(7), pp. 4332-4340, 2011.
[10] G. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, Vol. 9(4), pp. 309-347, 1992.
[11] D. Heckerman, *A tutorial on learning Bayesian networks*, In. Redmond Washington: Microsoft Research, 1996.
[12] M. I. Jordan, *Learning in Graphical Models*, Cambridge, MA: MIT Press, USA, 1999.

Nan Feng was born on September 2, 1978, in China. He received the B.S. degree in Management Information Systems from Tianjin University, Tianjin, China, in 2001; and the M.S. and the Ph.D. degrees in Information Management and Information Systems from Tianjin University, Tianjin, China, in 2004 and 2007, respectively.

He is currently an Associated Professor in the Department of Information Management and Management Science, College of Management and Economics, Tianjin University, Tianjin, China. He has published several research articles in academic journals such as: A Hybrid Approach of Evidence Theory and Rough Sets for ISS Risk Assessment (Journal of Networks, 2012), An Information Systems Security Risk Assessment Model under Uncertain Environment (Applied Soft Computing, 2010), and Software Project Risk Management Model Based on Bayesian Networks (Computer Engineering, 2007). His research interests involve information systems security and software project risk management.

Dr. Feng is a member of the Association for Information Systems in China.

Xue Yu received the M.S. degree in Computer Science in 2003 from University of Wollongong, Wollongong, N.S.W., Australia, and Ph.D. degree in Information Management and Information Systems in 2009 from Tianjin University, Tianjin, China.

She is currently an Assistant Professor in the Department of Information Management and Management Science, College of Management and Economics, Tianjin University, Tianjin, China. She has published several research articles in academic journals such as: Effective hybrid collaborative filtering model based on PCA-SOM (Systems Engineering Theory & Practice, 2010), Collaborative filtering recommendation model based on effective dimension reduction and K-means clustering (Application Research of Computers, 2009), and Collaborative Filtering Recommendation Model Based on Local Principle Component Analysis (Computer Engineering, 2010). Her research interests are information filtering and business intelligence.