Modeling and Simulation for Security Risk Propagation in Critical Information Systems*

Young-Gab Kim[1], Dongwon Jeong[2], Soo-Hyun Park[3], Jongin Lim[1], and Doo-Kwon Baik[4]

[1] Graduate School of Information Management and Security, Center for Information Security Technologies (CIST), Korea University, 1, 5-ga, Anam-dong, SungBuk-gu, 136-701, Seoul, Korea
{always,jilim}@korea.ac.kr
[2] Dept.of Informatics & Statistics, Kunsan National University San68, Miryong-dong, Gunsan, Jeolabuk-do, 573-701, Korea
djeong@kusan.ac.kr
[3] School of Business IT, Kookmin University, 861-1, Chongnung-dong. SungBuk-gu, Seoul, Postfach 136-702, Korea
shpark21@kookmin.ac.kr
[4] Department of Computer Science & Engineering, Korea University 1, 5-ga, Anam-dong, SungBuk-gu, 136-701, Seoul, Korea
baikdk@korea.ac.kr

Abstract. Existing risk propagation models are limited and inadequate for the analysis of cyber attacks caused by various threats to information systems, because of their limited focus only on one specific threat, such as a single virus or worm. Therefore, we herein propose a risk propagation model based on the Markov process, which can be applied to diverse threats to information systems. Furthermore, simulations including in case a threat occurs related with other threats are performed using five scenarios to verify the proposed model.

1 Introduction

Security risk analysis (it is also called risk assessment) is a process of evaluating the systems assets, their vulnerability to various threats, and the cost or impact of potential losses. Precise security risk analysis provides two key advantages: supporting practical security policies for organizations by monitoring and effectively protecting the critical assets of the organization, and providing valuable analysis data for future estimation through the development of secure information management [1]. Despite the considerable research relating to risk analysis, little attention has focused on evaluating the security risk propagation [1, 2, 3]. Furthermore, the existing security

* "This research was supported by the MIC (Ministry of Information and Communication), Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Advancement)" (IITA-2006-(C1090-0603-0025)).

Y. Wang, Y. Cheung, and H. Liu (Eds.): CIS 2006, LNAI 4456, pp. 858–868, 2007.

risk propagation models are inadequate to apply this to the analysis of attacks caused by diverse threats because they can only be applied to specific threats such as a virus or worm. Furthermore, it is difficult to globally analyze the scope of risk propagation caused by such threats, using their interrelationships. Therefore, a previous work [4] proposed a probabilistic model for damage propagation based on the Markov process [6, 7] and on historical data collected over several years. Using the proposed model, the occurrence probability and frequency for each threat to information systems can be predicted globally, and applied to establish effective countermeasures. However, the previous work [4] only analyzed the approach method with a case study. Furthermore, simulations performed in the previous paper [5] only simulated security risk propagation for case of an independent threat. Therefore, this paper presents a modeling and simulation for security risk analysis. In addition, five scenario simulations are performed in this paper to verify the proposed model.

The sebsequent sections of this paper are organized as follows: In Section 2, the security risk propagation model that has been proposed in previous work [4] is explained. Section 3 shows the simulations for security risk propagation, including in case a threat occurs related with other threats. Section 4 shows the related work, including the worm and virus propagation model. Section 5 concludes this paper.

2 Modeling of Security Risk Propagation

In this section, the risk (or damage) propagation model based on the Markov process proposed in the previous work [4] is explained briefly. The model proposed in the previous work [4] is composed of 4 steps: Threat-State Definition, Threat-State Transition Matrix, Initial Vector, and Risk Propagation Evaluation. A more detailed description will be presented in the following subsections.

2.1 Definition of a Set of Threat-States (Step 1)

In Step 1, three tasks are performed to define the threat-states: the gathering of occurrence data of threats, threat analysis, and definition of a set of threat-states. That is, in this step, all kinds of threats are examined, the threat-occurrence data are collected and analyzed in information systems, and finally the possible threat-states can be defined. If S is a set of threat-states, S can be defined as formula (1).

$$\begin{aligned} &T = \text{a set of thresholds, } \{T_1, T_2, \ldots, T_n\} \\ &T_i = \text{a specific threat such as hacking, worm or virus} \\ &S = \text{a set of threat-states, } \{S_1, S_2, \ldots, S_i, \ldots, S_n\} \\ &S_i = \text{a pair of thresholds, } (T_\alpha, T_\beta, \ldots, T_\gamma), \\ &\qquad \text{where } \alpha, \beta, \text{ and } \gamma \text{ are each a different threat.} \end{aligned} \tag{1}$$

It is particularly important to collect reliable and numerous historical data related with the threats because such historical data is more important than other elements in the probability model based on the Markov process. Therefore, in the simulation results presented in Section 3 of this paper, statistics on hacking and virus propagation published by the Korea Information Security Agency (KISA) were used for 54 months, from January 2001 to June 2005, to ensure the reliability of the past data [8].

The definition of a threat-state task decides the threat-states by analyzing the threat-occurrence, and establishing a threshold indicating the frequency range of the threat-occurrences. Two methods are available to define the set of threat-states, according to the dependency among threats. That is, when a threat occurs independently of other threats, the set of threat-states is composed of a number of several thresholds. Conversely, when a threat occurs that is related with other threats, the set of threat-states is created with the combination of thresholds of each threat. Therefore, in the latter case, the number of threat-states and the complexity of transition matrix, which describes the probabilities of moving from one state to another, will increase in proportion with the number of threat-states.

2.2 Transition Matrix of Threat-State (Step 2)

In Step 2, the threat-state transition matrix is calculated, which is a square matrix describing the probabilities of moving from one threat-state to another. In order to obtain the transition matrix, three tasks are performed. First, threat-states are listed by mapping the threat-occurrence data of each threat into the threat-state defined in the previous step. Second, the number from one threat-state to another is counted. Finally, the matrix is constructed. The function mapping each state S to a set of thresholds is as follows:

$$\text{Threat-states: } S \rightarrow 2^T, \text{ a function mapping each state } S \text{ to a set of thresholds } T \tag{2}$$

As in Step 1, the creation of a transition matrix is divided into two methods, according to the dependency among threats. When a threat occurs independently, the transition matrix can be created simply with the two tasks mentioned previously. However, when a threat occurs that is related to others, the size and complexity of the threat-transition matrix are increased, depending on the number of related threats and the threat-state defined in Step 1. Therefore, in order to reduce the complexity and size of the transition matrix, it is very important to decide the proper number of threat-states in Step 1.

If P is the transition probability matrix created in this step, it is compactly specified as the form of matrix (3). Furthermore, the entries of the matrix P satisfy the property (4).

$$P = \begin{bmatrix} P_{11} & P_{12} & \ldots & P_{1n} \\ P_{21} & P_{22} & \ldots & P_{2n} \\ \ldots & \ldots & P_{ij} & \ldots \\ P_{n1} & P_{n2} & \ldots & P_{nn} \end{bmatrix} \tag{3}$$

$$\text{where } \sum_{j=1}^{n} P_{1j} = 1, \sum_{j=1}^{n} P_{2j} = 1, \ldots, \sum_{j=1}^{n} P_{nj} = 1 \text{ . That is, } \sum_{j=1}^{n} P_{ij} = 1, i = 1,2,\ldots,n \tag{4}$$

Each row shows the probabilities of moving from the state represented by that row, to the other states. The rows of the Markov transition matrix therefore each add up to one.

2.3 Initial Probability (π Vector) (Step 3)

Step 3 is a process to obtain the initial probability vector, which represents the occurrence possibility of each threat-state in the initial state. In order to obtain the initial probability, the most recent threat-occurrence data are used, which can be divided by the time period such as three, six, and nine months and one year. By analyzing the most recent data, the initial probability vector is calculated using formula (5) satisfied by condition (6).

$$P(S_1 \quad S_2 \quad \ldots \quad S_k \quad \ldots \quad S_n) = P(\frac{\alpha}{F} \quad \frac{\beta}{F} \quad \ldots \quad \frac{\gamma}{F} \quad \ldots \quad \frac{\delta}{F}) \tag{5}$$

$$F = \sum_{i=1}^{n} f_i = \alpha + \beta + \ldots + \gamma + \ldots + \delta \tag{6}$$

where α, β, γ and δ represent the number of threat-occurrences for each state, S_1, S_2, S_k and S_n, respectively. Furthermore, the initial probability $P(S_i)$ of each state S_i satisfies the formula (7) because the sum of the initial probabilities must add up to one.

$$\sum_{i=1}^{n} P(S_i) = 1 \tag{7}$$

2.4 Prediction of Threats (Step 4)

In Step 4, the probability and frequency of the threat-occurrence that will occur in the future are estimated, using the transition matrix created in Step 2 and the initial probability vector created in Step 3. Formula (8) depicts the computation of the probability of threat-occurrence.

$$P(S_1 \quad S_2 \quad \ldots \quad S_k \quad \ldots \quad S_n)\begin{bmatrix} P_{11} & P_{12} & \ldots & P_{1n} \\ P_{21} & P_{22} & \ldots & P_{2n} \\ \ldots & \ldots & P_{ij} & \ldots \\ P_{n1} & P_{n2} & \ldots & P_{nn} \end{bmatrix} = P'(S_1 \quad S_2 \quad \ldots \quad S_k \quad \ldots \quad S_n) \tag{8}$$

where n is the number of threat-states, $P(S_i)$ the initial probability of each threat-state, and $P'(S_i)$ the next probability of threat-occurrence.

Finally, the Expected Frequency (EF) of threat-occurrence is estimated using the probability of threat-occurrence and the median for each threat-state, as formula (9).

$$EF = \sum_{i=1}^{n} P(S_i) M(S_i) \tag{9}$$

where n is the number of threat-states, $P(S_i)$ the probability of threat-occurrence for each threat-state, and $M(S_i)$ the median of each threat-state.

Further details on the creation of the Markov process-based risk propagation model are available in [4].

3 Simulation for Security Risk Propagation

As described in section 2.1 above, simulation studies require the use of an organization's historical data for some period of time. First, threat-occurrence data is gathered and analyzed, and priority is given to threats. Second, the monthly frequency and statistics of threats are obtained, as presented in Tables 1 and 2.

Table 1. Frequency and statistics of threat T_1 for each month

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
2001	85	125	70	89	85	64	65	495	268	77	51	97	1571
2002	401	119	82	59	286	417	313	298	210	465	472	990	4112
2003	1148	557	1132	934	306	450	185	544	119	137	129	96	5837
2004	154	148	118	1066	493	181	72	22	16	24	125	90	2509
2005	29	20	15	3	15	36	-	-	-	-	-	-	118

T_1 is an 'illegal intrusion using malicious applications such as Netbus and Subseven' as one of the hacking threats to an information system. This threat leaks information and interrupts the normal process in information systems.

Table 2. Frequency and statistics of threat T_2 for each month

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
2001	1	1529	2429	625	684	520	6106	5965	10772	4795	4068	3024	40518
2002	2005	1384	1306	3165	2760	1774	1706	1458	1610	3566	3028	1684	25446
2003	1361	1320	2537	2350	3704	1854	1185	9748	19682	3999	11658	8949	68347
2004	4824	5750	9820	4233	19728	22767	15228	8132	3153	2658	2319	2117	100727
2005	1832	1205	1049	648	1302	1040	-	-	-	-	-	-	

T_2 is an 'Internet Worm' as an example of a virus threat. The Internet worm is a self-replicating computer program or executable program with rapid self-propagation. The incidence of this threat has recently increased greatly, and considerable research relating to the propagation of Internet worms is processing.

The proposed model is simulated using a statistical method for comparing real-world observations and simulation output data as in the inspection approach [9], which computes one or more statistics from the real-world observations and corresponding statistics from the model output data. The two sets of statistics are then compared without the use of a formal statistical procedure. An inspection approach may provide valuable insight into the adequacy of a simulation model for certain simulations. In this Section, 5 scenarios are investigated to verify the proposed, Markov process-based, risk propagation model.

First of all, in order to verify the proposed model, the elements of the risk propagation model (that is, threat-states, initial vector, and threat transition matrix) are defined using the statistics on hacking and virus attack reported by KISA for 42 months, from January 2001 to June 2004. Next, using this model, the frequency of

threat-occurrence for 1 year, from July 2004 to June 2005, is calculated. Finally, the one-year *EF* calculated from the proposed model is compared with the real frequency as presented by KISA.

Scenario 1. In *Scenario 1*, three different ranges are used to calculate the median: 1 month, an average of 2 months, and an average of 6 months. The simulation condition is as follows:

- ***Median:*** the ranges to calculate the median are divided into 3 cases: 1 month, an average of 2 months and an average of 6 months
- Initial vector: the most recent 6-month frequency data are used to calculate the initial vector. Furthermore, the initial vector is changed every month
- Threat-states transition matrix: the transition matrix is changed every 6 months

The simulation result of *Scenario 1* is presented in Fig. 1.

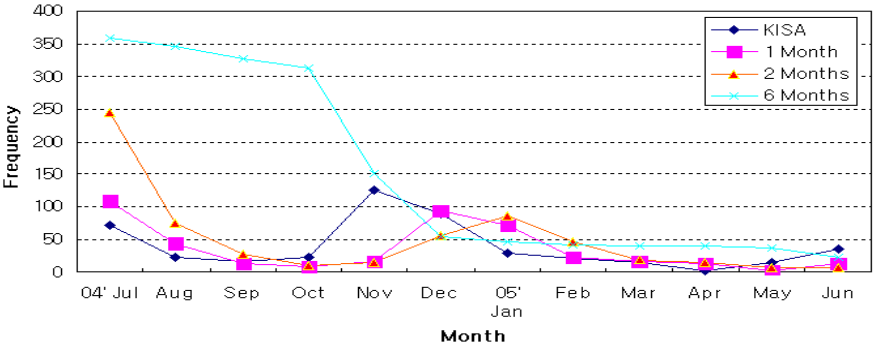

Fig. 1. Simulation result of Scenario 1

In the simulation result with 1 month set as the median range, the frequency of threat-occurrence is closer to the real frequency reported by KISA than using 2- and 6-month medians, i.e., a more precise result is obtained with a shorter range.

Scenario 2. In *Scenario 2*, three different ranges are used to calculate the initial probability vector: 3 months, 6 months, and 1 year. The simulation condition is as follows:

- Median: the most recent frequency data from the previous month are used to calculate the median. Furthermore, the median is changed every month
- ***Initial vector:*** the ranges to calculate the initial vector are divided into 3 cases: 3 months, 6 months and 1 year
- Threat-states transition matrix: the transition matrix is changed every 6 months

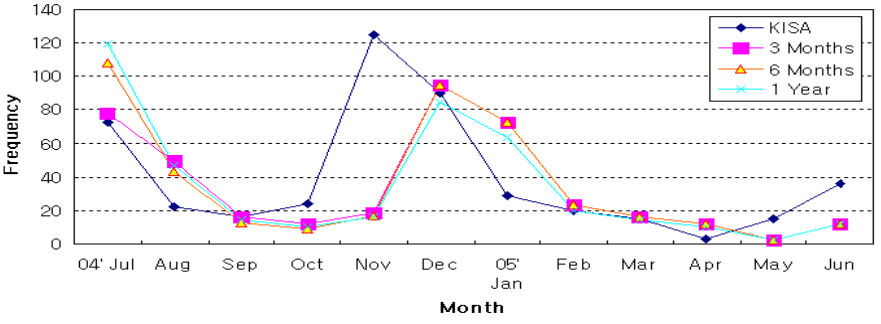

Fig. 2. Simulation result of Scenario 2

The simulation result of *Scenario 2* is presented in Fig. 2. As in *Scenario 1*, when the most recent frequency data is used as the range, the frequency of threat-occurrence is closer to the real frequency reported by KISA, i.e., a more precise result is obtained with a range of 3 months to calculate the initial vector.

Scenario 3. In this scenario, the period of changing the transition matrix is divided into 3 cases: 3 months, 6 months and 1 year. The simulation condition is as follows:

- Median: the most recent one-month frequency data are used to calculate the median. Furthermore, the median is changed every month
- Initial vector: the most recent 6-month frequency data are used to calculate the initial vector. Furthermore, the initial vector is changed every month
- ***Threat-states transition matrix:*** In order to construct the transition matrix, the periods of changing the matrix are divided into 3 cases: 3 months, 6 months and 1 year

The simulation result of *Scenario 3* is presented in Fig. 3.

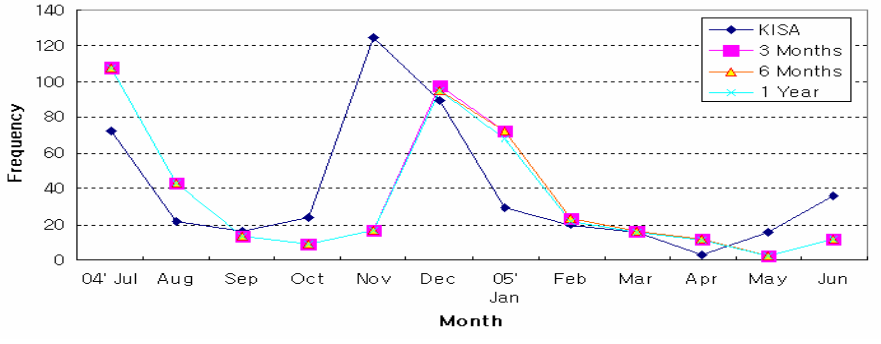

Fig. 3. Simulation result of Scenario 3

As shown in Fig. 3, the simulation results of the 3 cases are almost unaffected by the different period of changing the matrix. The period of changing the matrix hardly affects the frequency of threat-occurrence because the changes of the transition matrix are too small to create a new transition matrix, which is greatly different from the existing one.

Scenario 4. Six thresholds are applied in *Scenario 4*, unlike the previous three scenarios. The threat-states are divided into 2 cases: four threat-states and six threat-states. The simulation condition is as follows:

- Four threat-states : = S_1: 0~300, S_2: 301~600, S_3: 601~900, S_4: 901~1200
- Six threat-states := S_1: 0~200, S_2: 201~400, S_3: 401~600, S_4: 601~800, S_5: 801~1000, S_6: 1001~1200
- Median: the most recent one-month frequency data are used to calculate the median. Furthermore, the median is changed every month
- Initial vector: the most recent three-month frequency data are used to calculate the initial vector. Furthermore, the initial vector is changed every month
- Threat-states transition matrix: the transition matrix is changed every 6 months

The simulation result of *Scenario 4* is presented in Fig. 4.

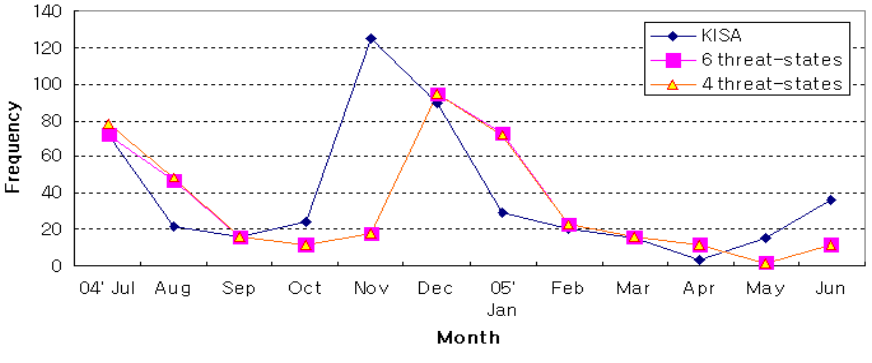

Fig. 4. Simulation result of Scenario 4

The simulation results of *Scenario 4* show a slight difference between the two cases. However, the amount of frequency data, which is applied to create the model proposed in this paper, was considered to be too small. As a result, a more precise result was obtained with a larger number of thresholds.

Scenario 5. In *Scenario 5*, the frequency of threat-occurrence is analyzed for cases of interrelated threats. The simulation condition is as follows:

- Thresholds of T_1 := H_1: 0~400, H_2: 401~800, H_3: 801~1200
- Thresholds of T_2 := W_1: 0~4000, W_2: 4001~8000, W_3: Over 8001

- Median: the most recent one-month frequency data are used to calculate the median. Furthermore, the median is changed every month
- Initial vector: the most recent 6-month frequency data are used to calculate the initial vector. Furthermore, the initial vector is changed every month
- Threat-states transition matrix: the transition matrix is changed every 6 months

The simulation result of *Scenario 5* is presented in Figs. 5 and Fig. 6.

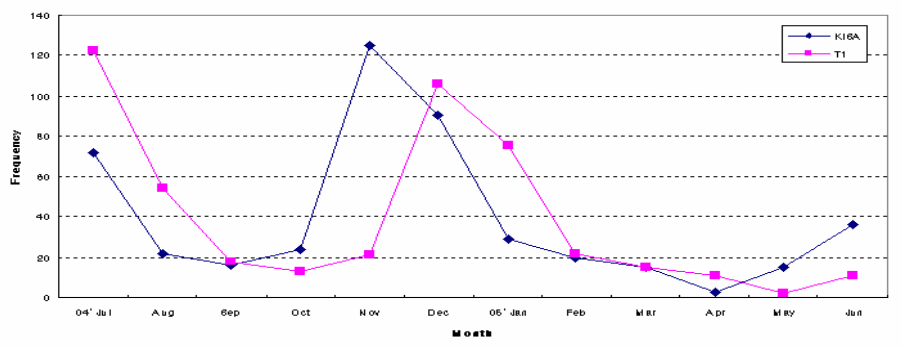

Fig. 5. Simulation result of Scenario 5 (T_1)

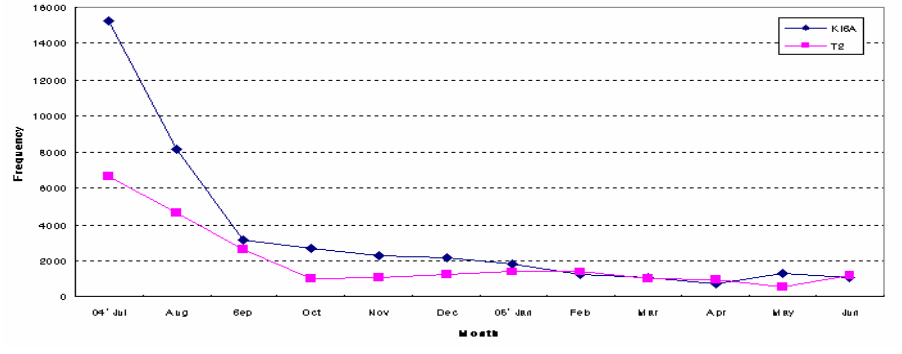

Fig. 6. Simulation result of Scenario 5 (T_2)

The number of thresholds is 4 for the independent threats of Scenarios 1 to 4, but is 3 in *Scenario 5*. That is, the simulation result of the *EF* for T_1 is different between *Scenario 5* and the previous 4 scenarios due to the different number of thresholds. From the simulation result of *Scenario 5* for T_2, the *EF* estimated by the proposed model is close to the real frequency presented by KISA.

Through the simulation result of the 5 scenarios, the *EF* estimated by the Markov process-based risk propagation model is generally close to the real frequency, except for specific months such as Nov. 2004 for T_1, due to the new emergence of malicious applications such as Netbus and Subsevens, and Jul. 2004 for T_2, due to the new emergence of an Internet worm. Further requirements are necessary to obtain a more precise

estimation in the proposed model [4]. First, the estimation, which is close to the real occurrence of a threat, is decided by subdivision of the threshold, i.e., more precise data can be obtained with a larger the number of thresholds. Second, the scope of the most recent data to define the Initial Probability should be considered. Third, statistically analysis is required. In this paper, although the past data of each month is used, a more precise result can be obtained than if past data is used relative to the date or week.

4 Related Work

Several research efforts have been made to model risk propagation, especially for viruses and worms. Two classical epidemic models are initially introduced. A simple epidemic model is a simple model of an epidemic of an infectious disease in a population [10, 11,12]. It is assumed that the population consists of two types of individuals, whose numbers are denoted by the letters S and I, which are susceptible individuals, who do not presently have the disease but are susceptible, and infective individuals, who have the disease and can infect others, respectively. That is, this model assumes that each host stays in only one of two states: susceptible or infective. These are, of course, functions of time. The second epidemic model is the Kermack-Mckendrick (KM) epidemic model [9, 11, 13], which adds a third state, R (removed), into the simple epidemic model. R is the number of removed individuals, who cannot be infected by the disease or infect others with the disease. This is called an SIR model due to the S→I→R possible state transition. Various propagation models extend from these two epidemic models. Although the KM model improves the simple epidemic model by considering the possibility for some infectious hosts to either recover or die after some time, this model is not suitable for modeling worm propagation because it does not consider human countermeasures. The two-factor worm model considers the effect of human countermeasures and the congestions caused by worm scan traffic [13, 14]. In the Internet, countermeasures such as cleaning, patching, and filtering against worms will remove both susceptible and infectious hosts from circulation in the KM model. Zou et al. and Moore et al. study the effect of quarantine on the Internet level to constrain worm propagation [14, 15]. They show that an infectious host has a number of paths to a target due to the high connectivity of the Internet. Therefore, they can prevent the wide spread of a worm on the Internet level by analyzing the effect of quarantine on the Internet. Chen et al. and Vogt present a discrete-time worm model that considers the patching and cleaning during worm propagation [16, 17]. As shown previously, most risk propagation models focus on viruses and worms and therefore cannot be applied to the diverse threats faced by modern information systems.

5 Conclusion

This paper has briefly presented a probabilistic model of security risk propagation based on the Markov process, which can estimate the spread of risk when attacks occur from diverse threats as well as viruses and worms. Furthermore, the proposed model was verified by running five scenario-based simulations. The simulation results confirmed the close agreement of the *EF* estimated by the Markov process-based, risk propagation model over a one-year period with the real frequency as presented by KISA, except for two specific months: Nov. 2004 for T_1, due to the new emergence of

malicious applications such as Netbus and Subsevens, and Jul. 2004 for T_2, due to the new emergence of an Internet worm. Future research will therefore need to focus on a suitable and effective method to deal with the regular appearance of a diverse range of threats to information systems.

References

1. In, H.P., Kim, Y.-G., Lee, T., Moon, C.-J., Jung, Y.-J., Kim, I., Baik, D.-K.: A Security Analysis Model for Information Systems. In: Baik, D.-K. (ed.) Systems Modeling and Simulation: Theory and Applications. LNCS (LNAI), vol. 3398, pp. 505–513. Springer, Heidelberg (2005)
2. Stoneburner, G., Goguen, A., Feringa, A.: Risk Management Guide for Information Technology Systems, NIST Special Publication 800-30, NIST (2002)
3. GAO: Information Security Risk Assetment-Practices of Leading Organizations. GAO/AIMD-00-33 (1999)
4. Kim, Y.-G, Lee, T., In, H.P., Jung, Y.-J., Kim, I., Baik, D.-K.: A Probabilistic Approach to Estimate the Damage Propagation of Cyber Attacks. In: Won, D.H., Kim, S. (eds.) Information Security and Cryptology - ICISC 2005. LNCS, vol. 3935, pp. 175–185. Springer, Heidelberg (2006)
5. Kim, Y.-G., Jeong, D., Park, S.-H., Baik, D.-K.: Simulation of Risk Propagation Model in Information Systems. In: Proc. of the 2006 International Conference on Computational Intelligence and Security (CIS 2006), pp. 1555–1558. IEEE Computer Society Press, Los Alamitos (2006)
6. Trivedi, K.S.: Probability and Statistics with Reliability, Queuing and Computer Science Applications. In: Probability and Statistics with Reliability, Queuing and Computer Science Applications, 2nd edn. WILEY Interscience, New York (2002)
7. Yates, R.D., Goodman, D.J.: Probability and Stochastic Process. 2nd edn. WILEY International, New York (2003)
8. KISA: Statistics and Analysis on Hacking and Virus, http://www.krcert.or.kr
9. Law, A., Kelton, W.: Simulation Modeling and Analysis, 3rd edn. McGraw-Hill Higher Education, New York (2000)
10. Frauenthal, J.C.: Mathematical Modeling in Epidemiology. Springer, New York (1980)
11. Deley, D.J., Gani, J.: Epidemic Modeling: An Introduction. Cambridge University Press, Cambridge (1999)
12. Staniford, S., Paxson, V., Weaver, N.: How to Own the Internet in Your Spare Time. In: Proc. of the 11th USENIX Security Symposium (Security02) (2002)
13. Zou, C.C., Gong, W., Towsley, D.: Worm Propagation Modeling and Analysis under Dynamic Quarantine Defense. In: Proc. of the ACM CCS Workshop on Rapid Malcode (WORM'03), ACM Press, New York (2003)
14. Zou, C.C., Gong, W., Towsley, D.: Code Red Worm Propagation Modeling and Analysis. In: Proc. of the proceedings of the 9th ACM Conference on Computer and Communications Security, pp. 138–147. ACM Press, New York (2002)
15. Moore, D., Shannon, C., Voelker, G.M., Savage, S.: Internet Quarantine: Requirements for Containing Self-Propagating Code. In: In: Proc. of the proceedings of IEEE INFOCOM, IEEE Computer Society Press, Los Alamitos (2003)
16. Chen, Z., Gao, L., Kwiat, K.: Modeling the Spread of Active Worms. In: Proc. of the proceedings of IEEE INFOCOM2003, IEEE Computer Society Press, Los Alamitos (2003)
17. Vogt, T.: Simulating and Optimising Worm Propagation Algorithms (2003), http://web.lemuria.org/security/WormPropagation.pdf