View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Xiamen University Institutional Repository

Markov Graph Model Computation and Its Application to Intrusion Detection*

ZENG Jian-ping (曾剑平), GUO Dong-hui (郭东辉)
Department of Physics, Xiamen University, Xiamen, Fujian 361005, China

Abstract: Markov model is usually selected as the base model of user action in the intrusion detection system (IDS). However, the performance of the IDS depends on the status space of Markov model and it will degrade as the space dimension grows. Here, Markov Graph Model (MGM) is proposed to handle this issue. Specification of the model is described, and several methods for probability computation with MGM are also presented. Based on MGM, algorithms for building user model and predicting user action are presented. And the performance of these algorithms such as computing complexity, prediction accuracy, and storage requirement of MGM are analyzed.

Key words: *Markov Graph Model; intrusion detection; probability computation*
CLC number: TP 393.08 **Document code:** A
Article ID: 1672-5220(2007)02-0272-04

Introduction

Intrusion detection is one of the main security enforcements for computer and network systems. By building a statistical profile of a user's normal activities and comparing observed activities of the subject with the profile, an alarm will be raised if the user's observed activities differ largely from its norm profile [1]. When constructing the statistical profile in an anomaly IDS, Markov model (we call it Conventional Markov Model (CMM) in the paper in order to distinguish from our MGM) is usually used[2-3]. However, the computing complexity should be concerned in order to put IDS into use. The chief complexity exists in building model, updating model and predicting user action. Especially when introducing k order Markov model, kth power of the one-step transition matrix has to be calculated according to the well-known Chapman-Kolmogorov [4].

This paper discusses how to tackle this problem. The basic idea of the approach taken in this paper is to divide the whole status space into several small spaces according to the WEB-based application. Graph is introduced into the organization of CMM distributed in these smaller spaces. Although, several models that incorporate graph into Markov model have been proposed [5-6], these models are not suitable for intrusion detection for WEB-based application, because they treat status as a node and the calculation for WEB-based application remains excessive. So, a new representation method MGM is proposed.

1 A Brief Background on CMM

Markov process can be used to describe many dynamic systems, such as independent stochastic process. However, Markov chain is more concerned when parameter set and status set are discrete in IDS.

Supposed $X(t)$ is a random variable and its corresponding status set is $E=\{i_0, i_1, \cdots, i_n\}$, then homogeneous Markov chain, which is an important model and is reference as CMM, should satisfied the following requirements [7]:

(1) $X(t+1)$ is only dependent on $X(t)$;

(2) The probability transition between status at time t and t+1 is independent on time.

2 Specification of MGM and Probability Computation

2.1 Specification of MGM

Markov graph model, which is a kind of directed graph, is denoted as MGM. And it can be represented as $MGM=(r, N, V, E)$, the four parameters are described as follows:

$V=\{V_i\}$ is the set of graph vertex, but not including initial vertex. N is the number of vertex, r is initial vertex, $E=\{E_{ij}\}$ is the set of edge between two vertexes, E_{ij} denotes that it is a edge which is from vertex V_i to V_j. An edge or no edge is allowed between any two vertexes.

Fig. 1 shows an MGM with three vertexes, and vertex 1 is the initial vertex r.

Each vertex has a corresponding CMM. And the initial vertex can be denoted as $\lambda_r=(N_r, A_r, \pi_0^r, \pi_0^{ir})$, $1\leqslant i\leqslant N$. However, other vertex can be denoted as $\lambda_r=(N_j, A_j, \pi_0^{ir})$, $1\leqslant i\leqslant N$, $1\leqslant j\leqslant N-1$. The parameters in the

Received date: 2006-08-20
Biography: ZENG Jian-ping (1973-), male, Dr. E-mail: zeng_jian_ping@hotmail.com

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

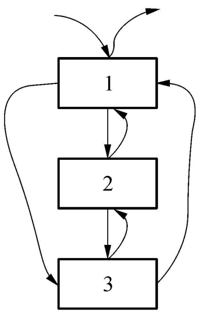

Fig. 1 MGM with three vertexes

model are described as follows: N_r is the number of status in the initial vertex, and N_j is the number of status in the vertex j, A_r is the matrix of probability transitions for initial vertex, and A_j is the matrix of probability transitions for vertex j, π_0^r is the initial probability distribution when entering the MGM, π_0^{ir} is the initial probability distribution when transits from vertex i to initial vertex, and π_0^{ij} is the initial probability distribution when transits from vertex i to j.

The status set in vertex j is denoted by $X_j = \{x_1 \; x_2 \; \cdots \; x_{N_j}\}$. For any two vertexes V_i, V_j, if exists E_{ij}, then there exists only one status $x_p \in X_i$ and it points to vertex j, we denote this as $x_p \rightarrow V_j$. As the same, if exists E_{ji}, then there exists $x_q \in X_j$ and $x_q \rightarrow V_j$.

For x_p, V_k, supposes that x_p is in the V_l, if $x_p \rightarrow V_k$ and $k \neq l$, then the transition probability a_{pi}, $1 \leqslant i \leqslant N_l$ is meaningless. So, a_{pi}, $1 \leqslant i \leqslant N_l$ is not included in the matrix A_l. Then, suppose the number of status that satisfied this relationship is M_l and the size of A_l is $m \times n$, then we can get,

$$m = N_l - M_l, \; n = N_l. \qquad (1)$$

2.2 Execution of MGM

The execution of MGM means as a process that starts from initial vertex, then, follows the directed edges and stops at certain vertex. We call the passing vertex sequence generated in the process of execution as ExecuteTrack, and it can be denoted as a set of sequence pairs, that is, ExecuteTrack = $\{(r, x_i), (V_i, x_i), \cdots, (V_j, x_j)\}$. For any pair (V_i, x_i), $1 \leqslant i \leqslant N_i$ is satisfied.

2.3 Probability of computation

MGM can be used to compute the probability transition distribution. We take one-order and two-order transition as an example in this article and suppose the execution is at status x_p in V_i at time t, x_q in V_j at time $t-1$ and V_k is a new vertex.

For one-order transition probability computation, there are two cases:

If $x_p \rightarrow V_k$ is not satisfied, then the transition probability can be calculated according to CMM, that is,

$$V(t+1) = V(t) \times A_i. \qquad (2)$$

If $x_p \rightarrow V_k$ is satisfied, then the next status will be in the vertex k, $\lambda_k = (N_k, A_k, \pi_0^{ik})$, so

$$V(t+1) = V(t) \times \pi_0^{ik}. \qquad (3)$$

For two-order transition probability computation, $V(t+1)$ is calculated based on the status at time $t-1$. There are four cases to be considered:

(1) If $V_i = V_j$ and $x_p \rightarrow V_k$ is not satisfied, then this means that the two transitions happen at the same vertex and the next state is also in the vertex, so, according to two-order CMM,

$$V(t+1) = V(t-1) \times A_i \times A_i. \qquad (4)$$

(2) If $V_i = V_j$ but $x_p \rightarrow V_k$ is satisfied, then the two transitions happen at the same vertex but the next state is transited to the new vertex k, so we can get

$$V(t+1) = V(t-1) \times A_i \times \pi_0^{ik}. \qquad (5)$$

(3) If $V_i \neq V_j$ and $x_p \rightarrow V_k$ is not satisfied, then the two transitions happen at different vertexes and the next state is remain at vertex i, so we can get

$$V(t+1) = V(t-1) \times \pi_0^{ji} \times A_i. \qquad (6)$$

(4) If $V_i \neq V_j$ and $x_p \rightarrow V_k$ is satisfied, then the two transitions happen at different vertexes and the next state is transited to the new vertex k, so we can get

$$V(t+1) = V(t-1) \times \pi_0^{ji} \times \pi_0^{ik}. \qquad (7)$$

3 Applying MGM to Intrusion Detection for WEB Application

3.1 Building the MGM

Typically, in a WEB-based application, the pages are organized as graph-like WEB, and each page is composed of many URLs (Universal Resource Locator), different pages and their links can be viewed as a directed graph with pages as vertex. So, by applying MGM, we can simply take page as vertex, URL as status, links from URL to page as edge and homepage as initial vertex. Before building the MGM, a larger record set containing ExecutionTracks, which compose of (Page, URL), should be collected. Then, the building algorithm for MGM can be described as follow.

Algorithm: Building MGM

Input: ExecutionTracks

Process:

```
For i = 1 to M      //M is the pages in ExecutionTracks
Initialize and allocate space for Markov model $\lambda_i$ of page$_i$
//L is the execution times in ExecutionTracks
For i = 1 to L do begin
Count the times for the first page and URL in ExecutionTracks
For each next record in ExecutionTracks do begin
Count the number of transiting between URL
```

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

```
    End
    End
    For i = 1 to M do begin
    $\pi_0^{ji} = \frac{\text{initial count for transition to URL}_j \text{ in page}_i}{\text{initial count for transition to page}_i}$
    $a_{ij} = \frac{\text{transfer count URL}_i \text{ to URL}_j}{\text{total transfer count for URL}_i}$
    End
```

Output: transition probability and initial distribution of MGM

3.2 Predicting user action

When predicting user action, the last two statuses should be saved in order to perform a 2- order prediction. Then we can get the prediction result as follows:

$$V(t+1) = \frac{w_1 \times V_1(t+1) + w_2 \times V_2(t+1)}{w_1 + w_2}, \quad (8)$$

where $V_1(t+1)$, $V_2(t+1)$ are the calculated results according to Eqs. (2)-(7), w_1 and w_2 are the weights for one and two-order predicting results.

4 Performance Analysis

4.1 Storage requirement

The required storage of a CMM with W statuses is the sum of transition matrix and initial probability distribution, that is W × W + W. However, for an MGM, the storage space required is the sum of all transition probability matrix and initial distribution vector in each vertex. According to the previous description, an MGM with total W number of status can be treated as the following matrix:

$$\boldsymbol{A} = \begin{bmatrix} A_1 & \pi_0^{12} & \cdots & \pi_0^{1n} \\ \pi_0^{21} & A_2 & \cdots & \pi^{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_0^{n1} & \pi^{n2} & \cdots & A_n \end{bmatrix}, \quad (9)$$

where the size of A_i is m × n, and it is satisfied with Eq. (1). π_0^{ij} is one dimension vector, and its size is N_1. So the storage space can be calculated as,

$$MG_{space} = T_{space} + V_{space},$$

where

$$T_{space} = N_1^1 \times (N_1^1 - M_1^1) + \cdots + N_1^n \times (N_1^n - M_1^n),$$

$$V_{space} = (N-1) \times (N_1^1 + N_1^2 + \cdots + N_1^N) = (N-1) \times W,$$

N_1^i, M_1^i is the N_1, M_1 for vertex i.

Suppose $N_1^1 = N_1^2 = \cdots = N_1^N = W/N$, then

$$T_{space} = N \times \left(\frac{W}{N}\right)^2 - \frac{W}{N} \times (M_1^1 + \cdots + M_1^n) = \frac{W}{N}(W - \sum_{i=1}^{N} M_1^i) = \frac{W}{N}(W - TL),$$

where $TL = \sum_{i=1}^{N} M_1^i$, and it means the total links in MGM.

By plotting the relation between the number of status and the size of store space in Fig. 2, we can see the difference in storage requirement clearly. As the number of status increases, the CMM will require much more space than MGM.

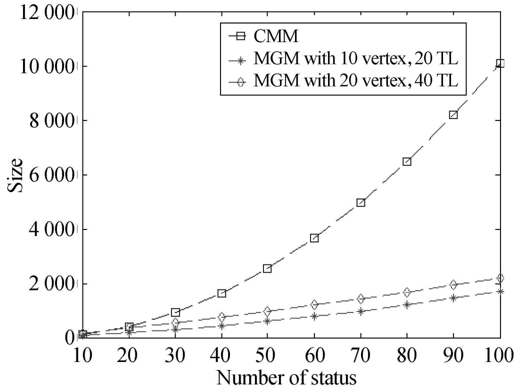

Fig. 2 Storage require for CMM and MGM

4.2 Time complexity

To predict user action with a k-order CMM, the maximum amount of time required happens at the time when status transits in the same vertex, so, k times of multiply of N dimension matrix should be done. So, the time complexity for 2-order CMM is $O(n^2)$. For MGM, because the status number in certain vertex is much smaller that the total status. So, the time required can be much lower.

4.3 Accuracy for prediction

It can be proved that the accuracy in prediction user action for MGM and CMM is the same. Suppose there are three vertexes and perform two-order predict. Take case Eq. (3) in two-order transition probability computation as example. Then according to Eq. (9), the whole transition probability can be given as

$$\boldsymbol{A} = \begin{bmatrix} A_1 & \pi_0^{12} \\ \pi_0^{21} & A_2 \end{bmatrix}.$$

Suppose the status of the current user is x_2, x_1 at time t- 2, t- 1. With respect to MGM, the x_2 item in vector $V(t-2)$ is 1 and its dimension is N_1. The same as this, the x_1 item in vector $V(t-1)$ is 1 and its dimension is N_2, the value of other items is 0. However, with respect to CMM, the x_2 item in vector $V(t-2)$ is 1 and the $x + x_1$ (x is the status number in vertex 1) item in vector $V(t-1)$ is 1, other items value is 0. The vector $V(t)$ for CMM is N-dimension, and it satisfies: $N = N_1 + N_2$.

So, according to Eqs. (2) and (6), the prediction result of MGM is

$$V_1(t) = V(t-1) \times A_2 = A_2(x_1), \quad (10)$$

$$V_2(t) = V(t-2) \times \pi_0^{12} \times A_2 = \pi_0^{12} \times A_2, \quad (11)$$

where, we denote $A(x)$ as the x line vector in $\boldsymbol{A}$.

However, according to CMM, the result is,

$$V'_1(t) = V(t-1) \times \boldsymbol{A} = (0 \cdots 0\ 1\ 0 \cdots 0) \times \boldsymbol{A} = A(x + x_1), \quad (12)$$

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\begin{aligned} V'_2(t) = {} & V(t-2) \times \boldsymbol{A} \times \boldsymbol{A} = (0 \cdots 1\ 0 \cdots 0) \times \boldsymbol{A} \times \boldsymbol{A} = \\ & A(x_2) \times \boldsymbol{A} = (0 \cdots 0 \mid \pi_0^{12}) \times \boldsymbol{A} = \\ & (0 \cdots 0 \mid \pi_0^{12} \times A_2), \end{aligned} \tag{13}$$

where the previous value of x in $A(x + x_1)$ is 0. By comparing Eq. (12) to (10), and Eq. (13) to (11), it can be seen that just a 0 vector is added to $V_2(t)$, $V_1(t)$. So, according to Eq. (8), the prediction results of $V_2(t)$, $V_1(t)$ and $V'_2(t)$, $V'_1(t)$ are the same.

5 Conclusion

To overcome the computation complexity in CMM when applying to WEB-based application system, a new model — Markov Graph Model (MGM) is proposed. The MGM divides the larger status space into smaller ones. A better performance in building detection model and predicting user action can be achieved by introducing MGM into the design. Also, MGM requires much smaller storage space with the same prediction accuracy as CMM. For future work, we will concentrate on improvement of the representation ability for complexity links between vertexes, and incorporate other intelligent technology to improve the detection performance.

References

[1] Kemmerer R A, Vigna G. Intrusion Detection: a Brief History and Overview [OL]. IEEE Security & Privacy, http: //computer. org/computer/sp/articles/kem/, 2002.

[2] Ye N, Chen Q, Borror C M. EWMA Forecast of Normal System Activity for Computer Intrusion Detection [J]. IEEE Transactions on Reliability, 2004, **53**(4): 557- 566.

[3] Tan K M C, Maxion R A. Determining the Operational Limits of an Anomaly-Based Intrusion Detector[J]. IEEE Journal on Selected Areas in Communications, 2003, **21**(1): 96- 110.

[4] Li Yuqi. Stochastic Process[M]. Beijing: Publishing House of National Defense Industry, 2003: 241- 242(in Chinese).

[5] Wolfertstetter F, Ruske G. Structured Markov Models for Speech Recognition[C]. Proc. International Conference on Acoustics, Speech, and Signal Processing, Detroit, USA, 1995: 544- 547.

[6] Matthias Eichner, Matthias Wolff, Sebastian Ohnewald, et al. Speech Synthesis Using Stochastic Markov Graphs[C]. IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, USA, 2001: 161- 164.

[7] Gong Guang-ru, Qian Min-ping. Course for Application of Stochastic Process [M]. Beijing: Publishing House of Tsinghua University, 2004: 86- 90(in Chinese).

© 1994-2013 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net