Journal of Automation and Information Sciences, 35(8), 2003

Adaptive Anomaly Detection in the Behavior of Computer Systems Users on the Basis of Markov Chains of Variable Order. Part II. Anomaly Detection Methods and Experimental Results

N.N. KUSSUL, A.M. SOKOLOV

Several ways to deal with behavior anomalies of computer system users and new methods of replay-attacks detection are suggested. They are based on the model developed in the first part of the paper.

Key words: Markov chain, anomaly detection, legal data, inserted data, pseudo-user.

The first part of our work [1] has suggested adaptive modification of model of Markov chains of variable order. In this paper this model is used for detecting anomalous activity of computer systems users. It has proposed some ways of detecting anomalies by means of the model developed as well as the new method of controlling replay-attacks.

1. Environment and technology of carrying out the experiments

To verify the approach we have conducted a series of experiments. Due to complexity of obtaining or modeling artificial data with intrusion traces and controversial extrapolation of results obtained in this way on non-artificial data the use was made of audit-files of sessions of real users.

The experiments were carried out with data obtained within a year from UNIX-server of physical & engineering faculty of NTU of Ukraine "KPI". Audit mechanisms of operation system FreeBSD traced all processes triggered on behalf of those registered in the system of users, including pseudo users (demons). Thus, there was no necessity in modifying the kernel of operation system so that to collect the necessary statistics, as it had been done in [2]. Altogether we obtained data for more than 800 users.

From the available information about the process we used only the name of the process taking no notice of the other parameters (time, characteristics of using resources of users, availability of superuser rights and so on). They are not of great importance in verifying our approach. However, they are necessary for construction of valuable detection system of anomalies.

Among numerous commands of users (more than 900 unique names of processes) registered while observations we kept only those whose use intensity exceeded 200 triggerings within the whole period (350 commands). All other rare commands were substituted by one special command RARE. Moreover, at the beginning and the end of each work session the indicators ALPHA and OMEGA were inserted correspondingly. Thus, learning was held with the examples of the form {ALPHA/tcsh/pine/who/tin/RARE} and started with an empty tree with one root node.

It is of interest that if we used "entire" Markov chains with the fixed order, for example $L = 5$, then to construct the appropriate automaton we would need $350^5 \approx 5 \cdot 10^{12}$ states, whereas by using PST-tree (probabilistic Suffix Tree) with the same parameter L the average number of states of the equivalent automaton among all users does not exceed some thousands.

ISSN 1064-2315

For each work session of a user its probability was calculated by formula (1) from [1]. With probability being a product of numbers smaller than unity, sessions with fewer number of commands have higher probability. Hence, as the "normality" coefficient we select the quantity which (with the equal multipliers) is independent of the sequence length $K = \sqrt[N]{P_T^N(r)}$ where $P_T^N(r)$ is the probability of the session r with the length N.

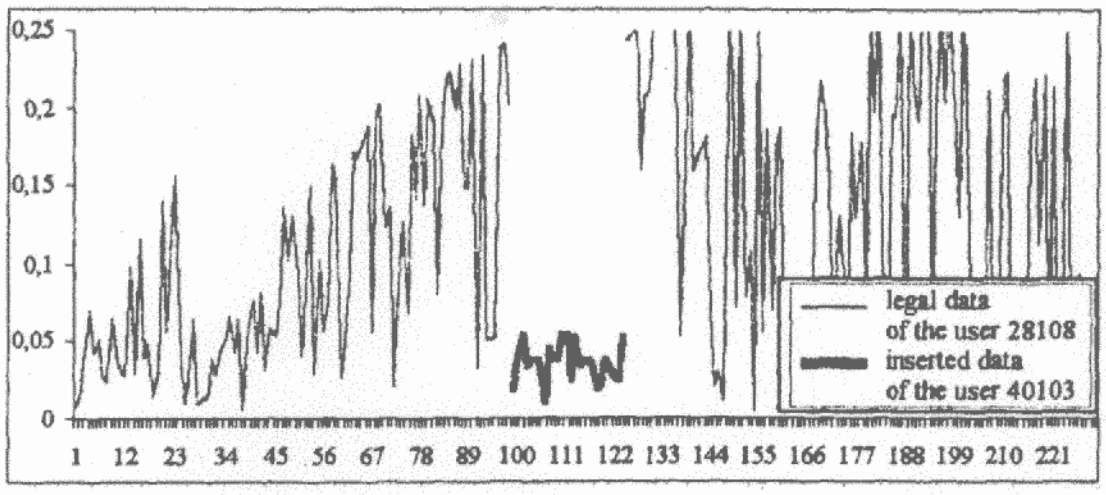
2. Intrusion detection

For each user by the modified algorithm we constructed the individual PST. Figures 1–4 depict the evolution process of the value of coefficient K within the definite time (half a year). Along the axis of abscissas the session number is put in order of entering, along the axis of ordinates — the values of K. For all plots the learning starts with the empty tree at $L = 5$ and $\alpha = 0.99$, excluding plots on Figures 3 and 4 where these parameters vary.

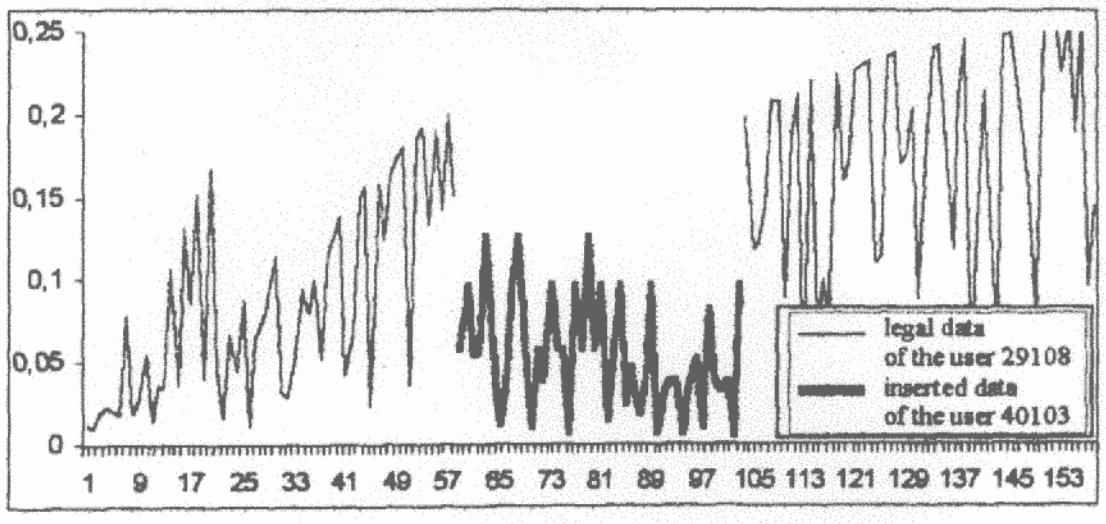
Behavior of the most users is of more or less regular character. Regularity can be explained by the use of scenarios performed on behalf of the user and in his absence — by performing a great number of identical actions or preferential use of his record in the system for verification of the mail entering and so on.

2.1. A user substitution

The detection system of anomalies on a user's level must respond if the current session was generated by a user who the data of the session audit are obtained from or some other person acts on his behalf. A popular example of such kind of intrusion is the situation with the stolen password. Figure 1 depicts results of intrusion simulation (the session processing with deliberately inserted parts of sessions of other users).



a



b

Figure 1

Note. Data of the user 40103 have been used correspondingly in the profile of the user 28108 (Figure 1, *a*) and 29108 (Figure 1, *b*).

It is seen that "alien" sessions are essentially different by values of K. Hence, we can introduce the threshold classification of the session: normal or anomalous. If the coefficient value K exceeds this threshold, the session is marked as normal, otherwise — as anomalous. However, if the both users belong to the same "class" then such classification might not be very reliable (see Figure 1, *b*, where the both users are students with identical habits).

2.2. Replay-attacks

One of the drawback of the detection system of anomalies is their vulnerability with respect to replay-attacks. They are characterized by the fact that a hacker upon achieving access to audit-files of a legal user "duplicates" his session by having substituted their insignificant part or inserting the appropriate commands. The conventional IDS will pass this session through as normal since the number of supplemented commands is insufficient. One of the approaches to this problem includes implementation of the upper threshold for probability of the current session. The session is considered to be anomalous if it is "too normal" [3]. It is not always correct, since, for example, systems pseudousers (demons) usually demonstrate very regular behavior with high session probabilities. Hence, it is pointless to signalize about appearance of replay-attack for each session with calculated high probability because this will lead to a large number of errors of the form "false positives". Moreover, if even such a form of the attack has been detected one cannot specify which of the commands are inserted or modified.

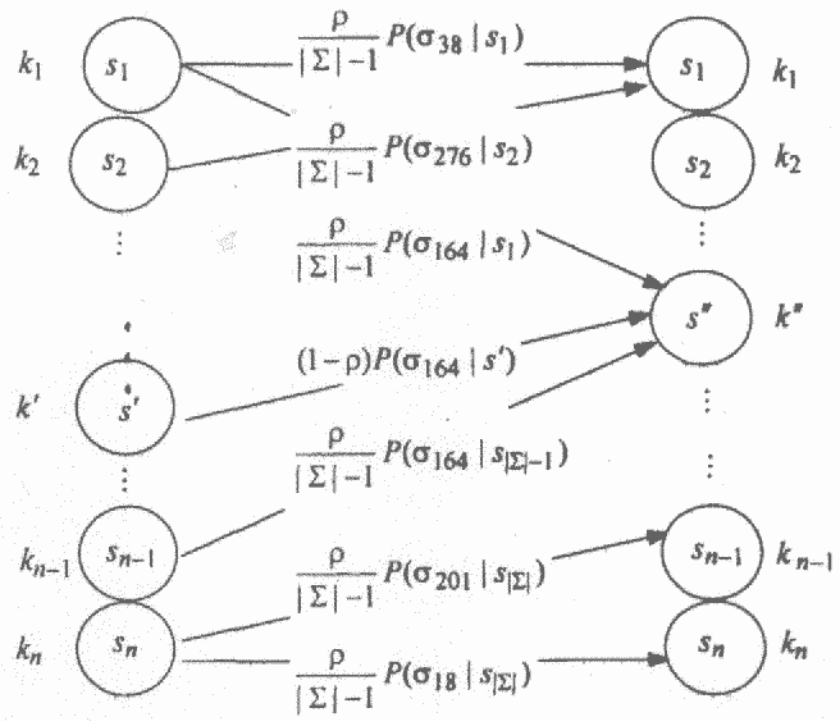


Figure 2

We propose a new way of analyzing and detecting such attacks based on possibility of transforming PST into automaton [4]. It not only detects a modified session but also singles out from it "alien" commands. The idea consists in considering the obtained session as a noisy one and trying to restore the most probable sequence of automaton states which it had passed so that to generate its noiseless variant.

We assume that in expert way we have obtained the information by which the probability of command substitution in the session is $0 < \rho < 1$. By interpreting PST as automaton generating sequences of commands we conclude that being in the state k with the label s it with probability $(1-\rho)P(\sigma^* | k)$ will pass into the state k^{**} with the label suffix $(s\sigma^*)$, where σ^* is the next command in the current session; or with probability $\frac{\rho}{|\Sigma|-1} P(\sigma' | k)$ — into some state k' with the label suffix $(s\sigma')$ where $\sigma' \in \Sigma \setminus \{\sigma^*\}$ (see Figure 2).

After construction of such automaton we set the problem: to find the most probable sequence of its states, i.e.

$$\bar{k} = \arg\max_{k \in \Sigma^N} \sum_{k_0 \in \Sigma} P_0(k_0) \prod_{i=1}^{N} P(x_i, k_i | k_{i-1}),$$

where $P_0(k_0)$ is the probability of the first command to be determined from transition probabilities of the node labeled by ALPHA.

By solving this problem by the known algorithms [5] we obtained the sequence of states $\bar{k}$ we have been seeking for. If for some i the state $\bar{k}_i$ has the mark s whose last symbol does not coincide with available for this place in session it implies the command substitution to have happened.

3. Assessments of the model parameters

The operation of the proposed adaptive model based on the Markov chains with variable size of memory depends on some parameters. The basic of them are the coefficient of learning α and the maximum order L to which PST grows up.

3.1. The coefficient of learning α

To determine the impact of the coefficient α upon the process of adaptation to users behavior we have carried out the experiments with different values α (see Figure 3). As it was expected the values α close to unity would lead to slow adaptation and slow growth of values of coefficient K, the adaptation process accelerating for smaller values α.

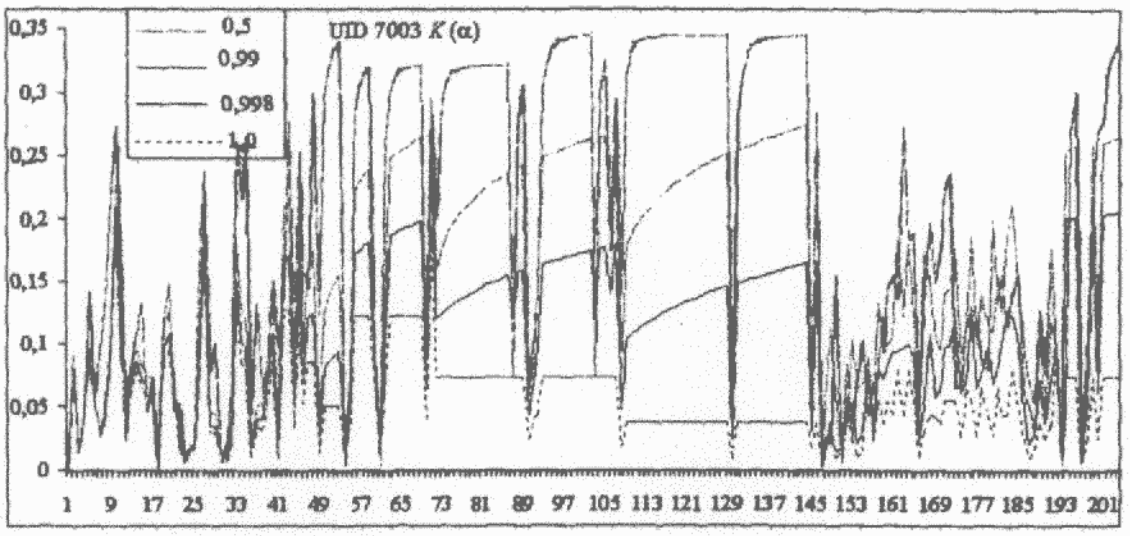


Figure 3

While selecting the specific value of the coefficient α one should employ individual approach since a various character of behavior variation fits to different people. It is important for α not to be given to small because it might increase the level of errors of the type "false negatives" although at that the learning will be held faster.

3.2. The maximal order parameter L

The parameter L specifies the possible length of the previous context The optimal value of L is considered to be such value for which the characteristics of PST constructed are not improved, i.e., with further L increasing the values of coefficient K do not increase.

Figure 4 illustrates how the gradual increase of L influences the values of K. It is clear that the worst value is $L = 1$ since at that for calculation the session probability account is taken only of the relative part of each command in audit-files. With L increasing one can see essential improvement of results of sessions processing since the longer context enables one to get statistical regularities of behavior more precisely. The experiment implies $L = 5$ to be the optimal value of the order for the given user since even for $L = 5$ and $L = 7$ the plots are almost similar.

Conclusions

The adaptive modification of the initial model of Markov chains of variable order has been proposed. It can be applied to various fields of modeling natural sequences where statistical characteristics vary in the

course of time. We presented the example of its application to the problem of detecting anomalies in computer systems. Adaptive adjustment of Markov chain with a memory of variable length allows one to adjust more precisely to peculiarities of active subjects of a computer system to specify the model permanently that fits in with the evident notstatic character of subjects behavior.

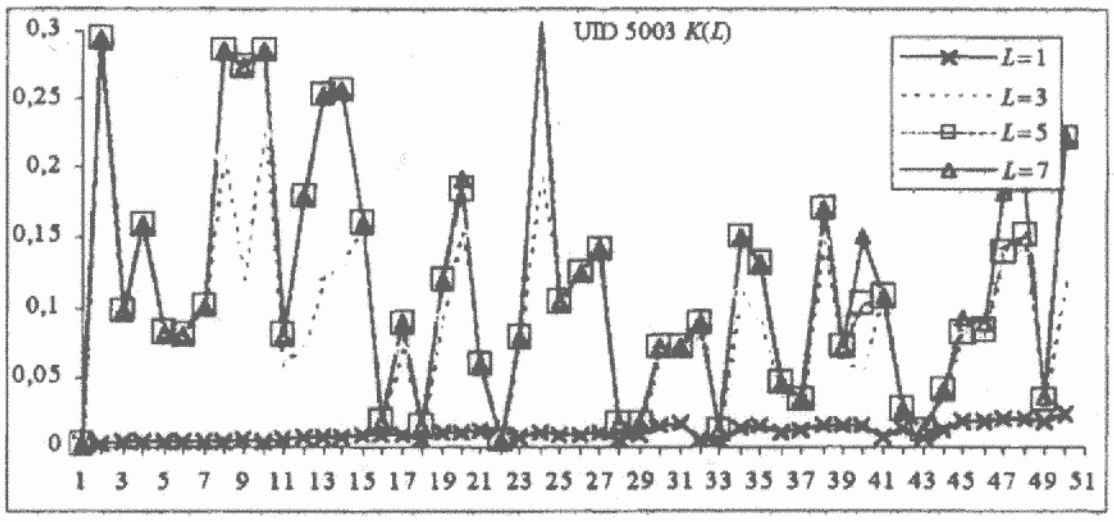


Figure 4

Due to accurate formalization of anomaly criteria being impossible one has to use the set of indices by whose collection one can draw a final conclusion about the presence of intrusion. This work has proposed some possible technologies for anomaly detection and their experimental substantiation, namely: Threshold classification method and approach enabling one to detect anomalous replay sessions and to specify what anomaly implies. All experiments and technologies proposed can be used in real systems of anomaly detection.

One of the challenging trends of further research is the construction of adaptive models of ensembles of probabilistic trees [6] analogous to those which have been used in this paper. The trees making up ensembles are taken into account with a definite weight while calculating the probability of the next event in audit-flow.

It is worth noting the method of anomaly detection at the level of systems calls of programs since it allows one to abstract from irregular behavior of human. However, at that one should not completely reject analyzing audit-data coming directly from users (for example, command session) since only analysis at this level will detect individual intrusions which at the level of systems calls do not reveal (for example, the use of a stolen password which is absolutely legal in terms of authentication subsystem).

References

1. Kussul N.N., Sokolov A.N., Adaptive anomalies detection in systems on the basis of Markov chains of variable order. Part 1. Adaptive Markov chain of variable order, *Problemy upravleniya i informatiki*, 2003, No.3, 83–93.
2. Somayaji A., Automated response using system-call delays, *Proc. Of USENIX Security Symposium Learning about Users*, 2000, 185-197.
3. Lane T., Hidden Markov models for human/computer interface modeling, *IJCAL-99 Workshop on Learning about Users*, 1999, 35–44.
4. Ron D., Singer Y., Tishby N., The power of amnesia: Learning probabilistic automata with variable memory length, *Machine Learning*, 1996, **25** (2–3), 117–149.
5. Schlesinger M.I., Hlaváč V. Deset prednasek z teorie statistickeho a strukturniho rozpoznavani, Praha, Vydavatelstvi. CVUT, 1999.
6. Eskin E., Anomaly detection over noisy data using learned probability distributions, *Proc. 17-th Intern. Conf. on Machine Learning*, San Francisco, CA, 2000, 255–262.