Detection of Data Anomalies at Network Traffic Analysis

S.G. Antipov, V.N. Vagin, M.V. Fomina
National Research University "Moscow Power-Engineering Institute"
Moscow, Russia
Email: antisergey@gmail.com, vagin@appmat.ru, m_fomina2000@mail.ru

***Abstract*—The paper is devoted to the problem of possibilities to recognition for unauthorized access to information transferred over networks on the basis of time series analysis. Algorithms of anomaly search in time series collections and the results of computer modeling are produced.**

***Keywords*—time series, anomaly detection, traffic analysis, dissimilarity function**

I. Introduction

The up-to-date rapid development of information technologies and implementation them to different spheres of human activities made very urgent an information security problem since informatization of the whole society caused by the growth of computer crimes the related plunders of the confidential and other information, subject to protection, and also the material losses [1].

At present, using components of foreign production is the cornerstone of production of technical means and the software for the majority of computing systems. However, at the same time there is a threat of information leakage due to use of the functional capabilities negatively affecting on safety of processed information (further in the text —a malicious code or malware).

Such malicious functional possibilities could use for organizing hidden channels in round of well-known protection means at passing defended information on computer networks.

One of the ways of information losses under discussion with 90 years [2] is possibility of malicious codes to use parameters of transfer protocols for coding defendable hidden information The channel capacity of information losses could be rather low, for example, on 2-4 orders low than the capacity for ordinary information channels. Nevertheless, under modern velocities of information transferring (gigabytes per second) such hidden channels possesses a high capacity. Even on low-speed links of communication, the method could be efficient, if it is necessary to draw a not great volume of important information.

In case a protection system implements encoding the information transferred according to the IP protocol, a malicious code can use such parameters as a lengths of packets, temporal intervals between packets for coding transferred unauthorized information. Taking account of a high complexity of up-to-date software and closing its program codes for researching by developers, often it is impossible to produce software researching for detecting similar malicious program functions.

II. Problem investigation

Transferring encoded data on communication links is brought about by conversion of bit sequences to electromagnetic signals. The data presented by bits or bytes are transferred with a velocity defined by the number of bits in time unit. Such parameters of the physical layer of network protocols as bit rate, the encoding method, the transfer scheme and a range of a signal are determined by standards that are developed by the competent organizations. The process of physical data transfer on a certain interval can be considered as time series

In general, the time series TS is an ordered sequence of values $TS =< ts_1, ts_2, .., ts_i, .., ts_m >$ describing the flow of a long process, where the index i corresponds to a time mark. ts_i values can be sensor indications, product prices, exchange rates and so on.

The analysis of a traffic will be a basis for detection of malicious program functions in an information exchange system: if parameters of a typical information exchange on certain protocol are known, then the abnormal template of behavior in a computer network obtained by traffic analysis in the case of exchange according to this protocol could speak that at analyzed system there is a malicious program function, and anomalies in a traffic are caused by actions of such programs.

The anomaly detection problem [3] is set up as the task of searching for templates in data sets that do not satisfy some typical behaviors. The anomaly, or "outlier" is defined as an element that stands out from the data set which it belongs to and differs significantly from the other elements of the sample. Let's consider, thus, the problem of search of anomalies in sets of time series.

The problem of anomaly detection in time series sets is formulated as follows. Let $TS\_STUDY = \{ts\_study_1, ts\_study_2, .., ts\_study_m\}$ be a set of objects where each object is time series. We call it a learning set. Each of the time series in a learning set represents some "normal" behavior or a process flow. Based on the analysis of TS_Study one needs to build a model to distinguish the instances of time series from $TS\_TEST = \{ts\_test_1, ts\_test_2, .., ts\_test_n\}$ to "normal" or "abnormal".

III. Methods of anomaly searching

We propose a method for the anomaly detection in the sets of time series, This method is a modification of an "exact exception problem" [4] that is described as follows: for the

given set of objects I, one needs to get an exception set I_x. To do this, there are introduced

1) the dissimilarity function $D(I_j)$, $I_j \in I$, defined on *P(I)* – the set of all subsets for I and receiving positive real values;
2) the cardinality function $C(I_j)$, $I_j \in I$,defined on *P(I)* – and receiving positive real values such that for any $I_1 \subset I$, $I_2 \subset I$ $I_1 \subset I_2 \Rightarrow C(I_1) < C(I_2)$;
3) the smoothing factor $SF(I_j) = C(I \setminus I_j) * (D(I) - D(I \setminus I_j))$, which is calculated for each $I_j \subseteq I$.

Then $I_x \subset I$ will be considered an exception set for *I* with respect to *D(I)* and *C(I)*, if its smoothing factor $SF(I_x)$ is maximal [4].

Informally, an exception-set is the smallest subset of I, that makes the largest contribution to its dissimilarity. A smoothing factor shows how much dissimilarity of a set *I* can be reduced, if from it to exclude a subset I_j.A dissimilarity function can be any function that returns a low value if elements of a set are similar to each other and a higher value if elements are dissimilar.

The algorithm *TS-ADEEP* that is based on this method was adapted for the anomaly detection problem in sets of time series. As a set *I* we use $TS\_STUDY \cup \{ts\_test_j\}$ for each $ts\_test_j \in TS\_TEST$. A dissimilarity function for time series is set up as follows:

$$D(I_j) = \frac{1}{|I_j|} * \sum_{i \in I_j} |i - \overline{I_j}|^2, \overline{I_j} = \sum_{i \in I_j} \frac{i}{|I_j|}$$

First, the average for the time series of I_j is calculated. The dissimilarity function is calculated as the sum of squared distances between the mean and vectors of I_j. The cardinality function is given by the formula $C(I \setminus I_j) = \frac{1}{|I_j|+1}$. The formula for calculating the smoothing factor is $SF(I_j) = C(I \setminus I_j) * (D(I) - D(I \setminus I_j))$.

If an exception set for $I = TS\_STUDY \cup \{ts\_test_j\}$ contains ts_test_j, then ts_test_j is an anomaly.

Based on this method, the algorithm *TS-ADEEP* for anomaly detection in sets of time series was introduced.

In this paper we propose the algorithm *TS-ADEEP-Multi* that is a generalization of the algorithm *TS-ADEEP* for the case of a study set contains several classes of time series. The generalization is quite obvious: splitting study set to single class subsets and consequently applying the *TS-ADEEP* algorithm, we can determine whether the considered time series is an anomaly. If time series is an anomaly for each subset or time series is not an anomaly for the only subset of the study set, the answer is quite obvious. However, there is a case where the time series is not an anomaly for several study set subsets. The algorithm *TS-ADEEP-Multi* is shown in the Table I.

A simulation of anomaly detection process was conducted on widely used datasets "cylinder-bell-funnel" [6] and "control chart" [7]. "Cylinder-bell-funnel" [6] contains three different classes – "cylinder", "bell", "funnel". "Control chart" [7] contains six different classes that describe the trends may be presented in the process: normal, cyclic, increasing trend, decreasing trend, upward shift, downward shift. The results of

Table I. THE ALGORITHM TS-ADEEP-MULTI

TS-ADEEP-Multi algorithm
input: (TS Study: learning set that contains time series of several classes; TSTEST: test set)
output: TS_ANOM_O – a set of anomaly time series of on the "optimistic" assessment; TSAnom_P - a set of anomaly time series on the "pessimistic" assessment
begin
$TS\_ANOM = \emptyset$
Let N be a number of classes containing in the learning set
$TS\_STUDY\_C = \{TS\_STUDY\_C_1, TS\_STUDY\_C_2, .., TS\_STUDY\_C_N\}$ – is a partition of TS_STUDY such that
$TS\_STUDY\_C_k$ containts only examples of class $k, k = 1..N$
for j from 1 to $|TS\_TEST|$ **begin**
  choose ts_test_j from TS_TEST
  for k from 1 to N **begin**
    $I = TS\_STUDY\_C_k \cup ts\_test_j$
    Find the exception set I_x in I
    IF $ts\_test_j \in I_x$, then ts_test_j is an anomaly for class k (doesnt belong to it)
  end
  If ts_test_j is an anomaly for all of the classes $TS\_STUDY\_C_k, k = 1..N$, then
    $TS\_ANOM\_O = TS\_ANOM\_O \cup ts\_test_j$
  If ts_test_j belongs to a single class $TS\_STUDY\_C_k$, then it is not an anomaly
  If ts_test_j belongs to r classes from TS_STUDY, 1<r<k then
    $TS\_ANOM\_P = TS\_ANOM\_P \cup ts\_test_j$
end
print TS_ANOM_O, TS_ANOM_P
end

program modeling given in [8] confirmed that these algorithms can successfully find anomalies among the time series relating to these datasets.

IV. EXPERIMENTAL RESULTS

The possibility of applying algorithms for anomaly searching in collections of time series to the problem of detection of cases of atypical information exchange on a network that assumes existence of malicious codes has been researched. This problem was complicated because it is very hard to receive a representative sample which would be rather exact and at the same time exactly describing all possible variants of behavior in an information system.

Also it is necessary to note that obtaining a sample for normal behavior of an information systems is enough easily than for abnormal one because normal behavior could be modeled in laboratory conditions while abnormal behavior happens extremely rare. Moreover, the abnormal behavior is dynamic by the nature, and there can be new types of anomalies which weren't represented in the original learning sample.

It is proposed the following problem solution. There are reference models presented by time series which reflect changes of parameters of the protocol depending on types of information exchange. For comparing, the time series are used, that present real behavior of an information system in case of a data interchange. The comparison of these two models of information exchange allows to look for behavior types in the case of information exchange, different from standard, i.e. the anomalies.

As an illustration of the method, the exchange protocol by the FTP files was chosen. The method can be distributed to other standard protocols of information exchange having

specifications in the form of standards or widely distributed de facto and having the description in open sources.

On the basis of analysis of a network traffic under transferring files in accordance to the FTP protocol in various conditions (including simultaneous transferring several files), the data set was obtained that represents a learning sample for creation of a model of reference data transferring.

For data acquisition, the special test bench was assembled. Data transferring on a network between two computers both according to the FTP protocol, and on a compound of protocols was carried out on this bench. For example, along with transmission of the file scanning of a network through the ICMP protocol (PING command), arbitrary exchange according to the UDP protocol – "information noise" for a reference traffic was carried out. Only length of the transferred data packet was fixed. Backward transferring accompanying exchange wasn't fixed.

Let's consider the example of the fixed data packet

```
No. Time Source Destination Protocol Info
4339 23.071158 10.10.10.50 10.10.10.100 FTP–DATA FTP
     Data: 1448 bytes
Frame 4339 (1514 bytes on wire, 1514 bytes captured)
Ethernet II, Src: 00:27:0e:2d:06:df (00:27:0e:2d:06:
    df), Dst: 00:27:0e:2d:06:17 (00:27:0e:2d:06:17)
Internet Protocol, Src: 10.10.10.50 (10.10.10.50),
    Dst: 10.10.10.100 (10.10.10.100)
Transmission Control Protocol, Src Port: 59022
    (59022), Dst Port: 9680 (9680), Seq: 288153, Ack
    : 1, Len: 1448
FTP Data
```

Here transferring was brought about from the computer with IP-address 10.10.10.50 (port 59022) to the computer with IP-address 10.10.10.100 (port 9680). The number of the packet 4339, accept time 23.071158, data are transferred in a packet (not the control footing), packet length 1448 bytes. Transferring was brought about on FTP protocol.

The following variants of data transferring were researched:

- transmission according to the FTP protocol (standard);
- simultaneous transmission according to the FTP protocols and ping (FTP-traffic was analyzed);
- simultaneous transmission according to the FTP protocols and UDP (FTP-traffic was analyzed).

The example of recorded data transferred simultaneously according to the FTP and UDP protocols is given in fig. 1. Having this information about data transferring on a network, it is necessary to define whether data transferring is "suspicious", what could testify about a possible compromise of network infrastructure, and malicious code existence.

As test data, were used specially generated time series imitating the unauthorized data transferring.

For the analysis of one of types of a traffic (the FTP protocol, the FTP and ping protocols, the FTP and UDP protocols), there was used the algorithm *TS-ADEEP* given in [4].

The obtained experimentally data were considered as a set of time series where ts_i values represent lengths of packets. These data have been previously exposed by preliminary treatment consisting of two stages: normalization and a subsequent

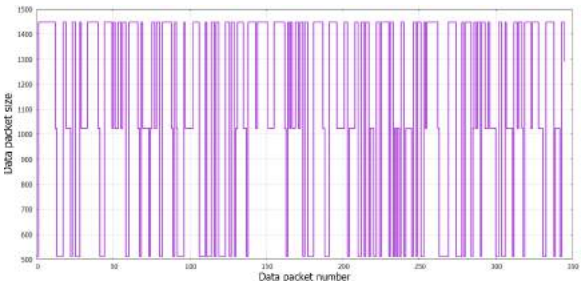

Figure 1.

discretization of the normalized time series with transition to symbolical data presentation, and moreover an alphabet size (the number of used symbols) has been varied depending on a task. The process of preliminary data conversion is based on the ideas of the *SAX* algorithm [9].

Changing parameters – an alphabet size and a time series dimensionality – it is possible to obtain an optimum representation of time series for their using by *TS-ADEEP* and *TS-ADEEP-Multi* algorithms.

Results of recognizing anomalies in the case of data transferring according to the above-named protocols are given in Table II. As it is seen from results, for the given task it was succeeded to reach the accuracy of anomaly classification of 100%.

Table II. THE ACCURACY OF ANOMALY DETECTION (%) IN DATA SETS "TRAFFIC" WITH ONE CLASS FOR THE *TS-ADEEP* ALGORITHM

		Time series size						
		210	150	100	50	30	20	10
Alphabet size	5	71.43	82.14	60.71	64.29	60.71	46.43	67.86
	10	92.86	96.43	100	96.43	82.14	85.71	64.29
	15	92.86	100	100	96.43	92.86	82.14	85.71
	20	92.86	100	100	96.43	92.86	82.14	82.14
	25	92.86	100	100	96.43	92.86	82.14	92.86
	30	92.86	100	100	96.43	92.86	92.86	82.14
	40	92.86	100	100	96.43	92.86	92.86	92.86
	50	92.86	100	100	96.43	92.86	92.86	92.86

By the simultaneous analysis of all traffic types considered in the experiment, the *TS-ADEEP-Multi* algorithm was used. Here the problem becomes complicated by that the normal behavior can correspond to one of several classes. Presented in the Table III results show, that for the case of several classes it is possible to reach the classification accuracy of anomalies up to 100% at selection of parameters of time series normalization.

Table III. THE ACCURACY OF ANOMALY DETECTION (%) IN DATA SETS "TRAFFIC" WITH SEVERAL CLASSES FOR THE *TS-ADEEP-Multi* ALGORITHM

		Time series size						
		210	150	100	50	30	20	10
Alphabet size	5	85.71	89.29	57.14	60.71	67.86	46.43	67.86
	10	96.43	96.43	100	96.43	82.14	85.71	67.86
	15	92.86	100	100	92.85	85.71	82.14	67.86
	20	96.43	100	100	96.43	96.43	96.43	75
	25	96.43	100	100	96.43	96.43	82.14	82.14
	30	96.43	100	100	96.43	96.43	96.43	92.86
	40	96.43	100	100	96.43	96.43	96.43	92.86
	50	96.43	100	100	96.43	96.43	96.43	96.43

In spite of the fact that by using *TS-ADEEP* and *TS-ADEEP-Multi* algorithms it was succeeded to reach the high classification precision by search of anomalies, these algorithms demand the large volume of calculations, therefore it is difficult to use them in the analysis of large volume of the new data. In that case, the approach, when a big calculation volume is required at the stage of constructing (learning) model, is perspective. Using of already the learned model requires a small calculation volume. Approach on the basis of the ideas of supervised learning is applied in many generalization algorithms, this approach is used also in the case of decision tree building [10].

We offer to use the approach on the basis of temporal decision trees [11]. On the basis of analysis of the learning sample containing descriptions of "normal" samples of data transfer, the temporal decision tree is built. This decision tree can classify further again entering examples. A new example is an anomaly if a temporal decision tree can't classify it.

To build a decision tree, the *Temporal ID3* algorithm was used. This algorithm was described in detail in [12]. The decision tree classifies the next example after several checks in internal nodes of a tree, therefore decision-making velocity is high. Results of anomaly detection by the *Temporal ID3* algorithm for data sets "Traffic" are given in Table IV.

Table IV. THE ACCURACY OF ANOMALY DETECTION (%) IN DATA SETS "TRAFFIC" WITH SEVERAL CLASSES FOR THE *Temporal ID3* ALGORITHM

		Time series size						
		210	150	100	50	30	20	10
Alphabet size	5	64.29	64.29	32.14	32.14	32.14	78.57	67.86
	10	78.57	75.00	64.29	67.86	60.71	67.86	57.14
	15	82.14	46.43	60.71	57.14	71.43	82.14	57.14
	20	85.71	89.29	82.14	67.86	60.71	96.43	82.14
	25	96.43	71.43	64.29	67.86	89.29	78.57	78.57
	30	96.43	82.14	60.71	96.43	71.43	85.71	67.86
	40	89.29	92.86	82.14	100.00	60.71	96.43	82.14
	50	71.43	89.29	92.86	100.00	71.43	85.71	82.14

As shown in Table IV, the temporal decision tree built by using time series for which dimensionality has been reduced to 50 and the alphabet size contains 40-50 symbols, most successfully copes with searching of anomalies.

V. CONCLUSION

Methods of anomaly detection at the solution of the problem of the network traffic analysis for the purpose of detecting malicious functional capabilities have been viewed. The algorithms implementing anomaly detection were simulated. Results of the program experiment have shown the high precision of anomaly detection what demonstrates good prospects of using the suggested methods and software.

Research has been conducted with financial support from Russian Foundation For Basic Research grant (project N 16-51-00058 Bel_a and N 15-01-005567a).

REFERENCES

[1] Shangin V.F. Information security of computer systems and networks –Moscow ID-Forum. –INFRA, .: 2011 (in Russian).

[2] https://en.wikipedia.org/wiki/Covert_channel

[3] Varun Chandola, Arindam Banerjee, Vipin Kumar. Anomaly Detection - A Survey // ACM Computing Surveys. — 2009. — Vol. 41(3). — Pp. 1–72.

[4] Andreas Arning, Rakesh Agrawal, Prabhakar Raghavan. A Linear Method for Deviation Detection in Large Databases // In Proceedings of KDD'1996. — 1996. — Pp. 164–169.

[5] Antipov S., Fomina M. Problem of anomalies detection in in time series sets. / Program Products and Systems, 2012, No.2, pp. 78-82 (in Russian)

[6] Saito Naoki. Local feature extraction and its application using a library of bases: Ph.D. thesis / Yale University. — 1994.

[7] D. T. Pham, A. B. Chan. Control Chart Pattern Recognition using a New Type of Self Organizing Neural Network // Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering. — 1998. — Vol. 212(2). — Pp. 115–127.

[8] Marina Fomina, Sergey Antipov, Vadim Vagin. Methods and algorithms of anomaly searching in collections of time series // Proceedings of the first International Scientific Conference Intelligent Information Technologies for Industry (IITI'16), Vol.1, pp.63-73 In Series Advances in Intelligent Systems and Computing, Volume 450. — 2016.

[9] A Symbolic Representation of Time Series, with Implications for Streaming Algorithms / Jessica Lin, Eamonn Keogh, Stefano Lonardi, Bill Chiu // In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. — 2003. — Pp. 2–11.

[10] V. Vagin, E. Golovina, A. Zagoryanskaya and M. Fomina. Exact and Plausible Inference in Intelligent Systems./V. Vagin and D. Pospelov, Eds., Moscow: FizMatLit, 2008, p. 714 (in Russian).

[11] Luca Console, Claudia Picardi, Daniele Theseider Dupre. Temporal decision trees: model-based diagnosis of dynamic systems on-board // Journal of Artificial Intelligence Research. — 2003. — Vol. 19(1). — Pp. 469–512.

[12] Antipov S.G., Fomina M.V. The method of forming generalized notions with using temporal decision trees // Artificial intelligence and decision-making, 2010, N 2, pp. 64-76. (in Russian)

ОБНАРУЖЕНИЕ АНОМАЛИЙ В ДАННЫХ НА ОСНОВЕ АНАЛИЗА СЕТЕВОГО ТРАФИКА

Антипов С.Г., Вагин В.Н., Фомина М.В.

Статья посвящена проблеме обнаружения возможного несанкционированного доступа к информации, передаваемой по сетям, на основе анализа временных рядов. Описаны алгоритмы поиска аномалий в наборах временных рядов, также приводятся результаты компьютерного моделирования.