Model-based Algorithm for Statistical Intrusion Detection

Petar Čisar

Telekom Srbija, Subotica, Serbia, petarc@telekom.yu

Sanja Maravić Čisar

Subotica Tech, Subotica, Serbia, sanjam@vtssu.rs

Abstract: Intrusion detection is used to monitor and capture intrusions into computer and network systems which attempt to compromise their security. Many intrusions (attacks) manifest in changes in the intensity of events occuring in computer networks. A lot of different approaches exist for statistical intrusion detection. One of them is behavioural analysis, thus in accordance with this, a model-based algorithm is presented. The research is realized on historical traffic data of authentic network users.

Keywords: traffic curve, modelling, statistics, contol limits

1 Introduction

The paper deals with statistical analysis of network traffic curves of major users. Based on the identification of common characteristics, the general model will be formed and defined the function of traffic, which represent the necessary condition for creation of algorithm for statistical detection of network anomalies.

2 Characteristics of Network Traffic

The research uses daily, weekly and monthly traffic curves of major users (Internet service providers and enterprises) which derive from the software MRTG (Multi Router Traffic Grapher) version 2.10.15. Without the loss of generality, here is given the graphical presentation of curves from three users, noting that the observed traffic curves of other users do not deviate significantly from here shown forms.

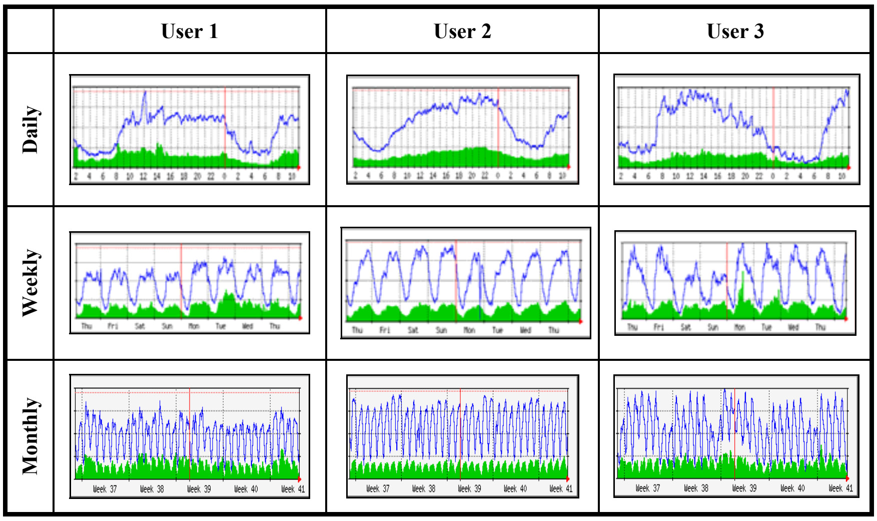

Figure 1
Traffic curves of different users

Having in mind the previous figure, it can be noticed a general periodic curve trend (the curve of average value of traffic in certain time intervals) with period T=24 h, which appears in the following way:

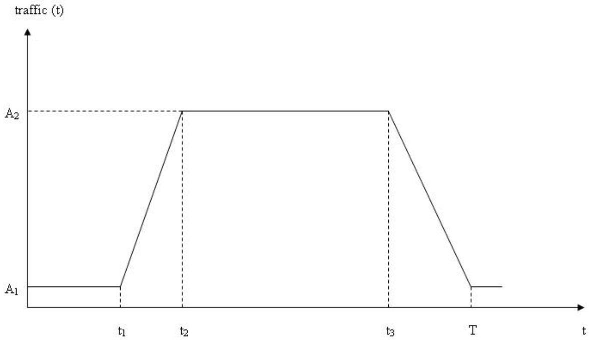

Figure 2
Model of traffic curve

In accordance with the previous figure, periodic traffic trend curve (with period T=24 hours) can be defined as follows:

- for the interval 0 – t_1 (night traffic): y(t) = A_1
- for the interval t_1 – t_2 (increase of morning traffic):

$$y(t) = A_1 + (A_2 - A_1) \cdot \frac{t - t_1}{t_2 - t_1}$$

- for the interval t_2 – t_3 (daily traffic): y(t) = A_2

- the interval t_3 – T (fall of night traffic): $y(t) = A_2 - (A_2 - A_1) \cdot \frac{t - t_3}{T - t_3}$

where A_1, A_2, t_1, t_2, t_3 and T represent values whose meaning is shown in the figure and vary from user to user.

In case of some users, for example, the user 3 in Figure 1, in days of weekend the fall of average daily traffic is about 25%. In this sense, the value of A_2 for a same user can not be considered as constant in all periods T.

In order to determine the range of expected values of traffic in a certain time during the day, 39 samples were taken from the curve of traffic from all segments of time: 0 – t_1, t_1 – t_2, t_2 – t_3 and t_3 – T. Then, descriptive statistics is applied on them, with aim of calculating the lower and upper control limit. In this sense, the arithmetic mean and standard deviation of samples are caculated, and on the basis of them, with confidence interval around 99% (i.e. 3σ), the maximum and minimum of expected value of traffic is determined. Each value that falls outside the specified interval, in statistical terms is a network anomaly and doubt on attack. The user 1 is taken for an example, and obtained are the values shown the following table.

Table 1

Samples and control limits

Sample	09-24 h (Mb/s)	24-03 h (Mb/s)	03-07 h (Mb/s)	07-09 h (Mb/s)
1	19	18	6	7
2	21,5	17,5	6,5	7,5
3	18,5	16	7	8
4	21	15,5	7,5	8,5
5	25	15	8	9
6	18	14,5	8,5	9,5
7	20,5	14	9	10
8	18,5	13,5	9,5	10,5
9	22	13	10	11
10	25	14	10,5	11,5
11	28	13	11	12
12	30	12	11	12,5
13	33,9	16	11,5	13
14	30	16	11,5	13,5
15	27	15,5	11	14
16	26	15	10,5	14,5
17	24	14,5	10	15
18	22	14	9,5	15,5
19	19	13,5	9	16
20	24	13	8	16,5
21	20	12,5	7	17
22	25	12	6,5	17,5
23	21	11,5	6	18
24	22	11	7	18,5

25	23	10,5	6,5	19
26	25	10	6	19,5
27	27	9,5	5	20
28	28	9	6	20,5
29	18,5	8,5	7	21
30	22	8	7,5	21,5
31	23	7,5	8	22
32	24	7	7,5	22,5
33	19	7,5	7	23
34	21	8	7,5	23
35	23	8,5	8	22
36	24	8	8,5	21
37	21	7,5	6	20
38	20	7	5,5	19
39	23,5	6	7	19

Avg	23,15128205	11,87179487	8,076923077	15,87179487
St.dev.	3,654188066	3,365298823	1,833670684	4,93206206
Max (99%)	34,11384625	21,96769134	13,57793513	30,66798105
Min (99%)	12,18871785	1,775898403	2,575911024	1,075608691

In the same way it is possible to determine the range of expected values for the traffic curve of any user. In order to check the calculated values, two measurements were made in the range of a month and obtained were the following maxima:

Table 2
Maxima in two measurements

	Daily (Mb/s)	**Weekly (Mb/s)**	**Monthly (Mb/s)**
1st measurement	33,9	29,7	30,9
2nd measurement	33,1	33,4	32,4

Analizing the results from the previous table it can be concluded that nor in one case is not exceeded the calculated maximum value of traffic (34,1 Mb/s). This fact justifies the used method.

The research also dealt with establishing the size and variation of characteristic values of the traffic in different time periods. In this regard, the observed values are the average and maximum traffic of several users, in daily, weekly and monthly periods. The results are as follows:

By comparison of data from table it can be concluded that the changes in maxima and average values of traffic are relatively small - the average value of difference in maximum traffic is 3,26% while in average value is 8,44%.

Table 3
Differences in characteristic values of traffic

	Daily 1 [Mb/s]	**Daily 2 [Mb/s]**	**Diff. [%]**	**Weekly 1 [Mb/s]**	**Weekly 2 [Mb/s]**	**Diff. [%]**	**Monthly 1 [Mb/s]**	**Monthly 2 [Mb/s]**	**Diff. [%]**
User 1									
Max.	33,9	33,1	-2,4	29,7	33,4	12,4	9,7	9,8	1
Average	16,5	19,1	15,8	17,0	21,1	24,1	6,01	6,6	9,8
User 2									
Max.	3,94	3,63	-7,8	3,98	3,68	-7,5	48,2	49,2	2
Average	2,35	2,09	-11	2,28	2,09	-8,3	30,9	30	-3
User 3									
Max.	9,31	10,0	7,4	9,71	9,99	2,9	9,9	9,7	-2
Average	5,71	6,01	5,2	5,63	6,64	17,9	5,4	4,9	-9,2
User 4									
Max.	9,69	9,99	3,1	10,0	9,91	-0,9	10	10	0
Average	4,96	5,14	3,6	5,2	4,94	-5	7,4	7,6	2,7
User 5									
Max.	48,2	46,3	-3,9	48,5	45,2	-6,8	1,8	1,8	0
Average	29	24,4	-15,9	30,4	26,4	-13,1	0,14	0,14	0
User 6									
Max.	10,1	10,1	0	10,0	10,0	0	3,94	3,66	-7,1
Average	7,78	7,95	2,2	7,43	8,14	9,6	1,9	2,03	6,8
User 7									
Max.	3,98	3,97	-0,02	3,94	3,99	1,2	3,9	3,9	0
Average	1,74	1,79	2,9	1,88	1,99	5,9	1,9	2	5,2

3 Time Factor in Intrusion Detection

Considering the variety of attacks, it is rather difficult to precisely define the starting part of the attack timeline. Some attacks are immediately recognizable, perhaps taking the form of one or more packets operating over a short time period Δt – e.g. less than one second [9]. Others are active for a much longer period (e.g. hours, days or even weeks) and may not even be identified as attacks until a vast collection of event records are considered in aggregate. Thus, while every attack has a definite beginning, this starting point is not always discernible at the time of occurence.

The main idea of this paper is the detection of such type of attacks that are recognizable in real time (real-time detection) – true time zero plus some arbitrary small time slice beyond that – i.e. less than a few seconds [9]. Real–time, according to industry definitions, can be expressed like time interval 5s – 5min.

Reaction time of security systems on the appearance of attacks is changing over time. So in the eighties and nineties of the last century the reaction time was about twenty days, from 2000 – 2002 about two hours, while from 2003 and later this time is needed to be less than 10 seconds.

Conclusions

In research proposed way of determining the maximum expected (allowed) value of the network traffic is checked in case of different users, in different time periods and in cases of time-distant measurements. No false alarm is generated. Algorithm based on the model has fixed input parameters: A_1, t_1, t_2, t_3 and T and a variable parameter A_2. If the periodically update of variable parameter is enabled in a short enough time intervals following the actual traffic, this algorithm gets the feature of adaptivity, which further reduces the possibility of generating false alarms.

References

[1] S. Sorensen: Competitive Overview of Statistical Anomaly Detection, White Paper, Juniper Networks, 2004

[2] F. Gong: Deciphering Detection Techniques: Part II Anomaly – Based Intrusion Detection, White Paper, McAfee Security, 2003

[3] SANS Intrusion Detection FAQ: Can you explain traffic analysis and anomalydetection?
Available: http://www.sans.org/resources/idfaq/ anomaly_detection.php

[4] Engineering Statistics Handbook – Single Exponential Smoothing,
http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc431.htm

[5] D. Montgomery: Introduction to Statistical Quality Control, 5th Edition, John Wiley & Sons, 2005

[6] Statistical Quality Control, Available:
ww.wiley.com/college/reid/0471347248/samplechapter/ch06.pdf

[7] V. Siris, F. Papagalou: Application of Anomaly Detection Algorithms for Detecting SYN Flooding Attacks. Available:
http://www.ist-scampi.org/publications/papers/siris-globecom2004.pdf

[8] CAIDA, the Cooperative Association for Internet Data Analysis: Inferring Internet Denial-of-Service Activity, University of California, San Diego, 2001

[9] M. Roesch: Next-generation intrusion prevention: Time zero (during the attack). Available: http://searchsecurity.techtarget.com/tip/