Abstracts of the workshop "Networking games and resource allocation"

The detection of computer network anomalies using the method of cumulative sums.

D. Zhuravlev

Petrozavodsk, July 12-15, 2002

We consider the problem of optimal disorder time detection for computer networks. We derived an analytical solution for the main parameters of the cumulative sums method with Pareto distribution function.

Let θ be an integer random variable (r.v.) with the countable state space $0, 1, \ldots$. If we suppose that $\theta = n$, then independent r.v.'s $\xi_1, \xi_2, \ldots, \xi_{n-1}$ have a common distribution $F_0(x)$, and r.v.'s $\xi_n, \xi_{n+1}, \ldots$ have another distribution $F_1(x)$ (we suppose that the densities of these distributions are $p_0(x), p_1(x)$, respectively). R.v. θ may be considered as the moment of disorder in the observation process. The observer's aim is to determine by the incoming observations $x_1, x_2, \ldots$ the point of change τ as accurately as possible. There is a number of methods to detect the point of change. Here we use the method of cumulative sums [1-4]. Let S_0 be any value in the interval $[1, \mu]$, where $1 < \mu$, and

$$S_n = \max\left\{1, S_{n-1}\frac{p_1(x_n)}{p_0(x_n)}\right\}, n \geq 1.$$

Define the moment τ in the form

$$\tau = \inf\left\{n \geq 1 : S_n \geq \mu\right\}.$$

The task is to find by practical consideration the optimal level μ.
The main parameters of the cumulative sums method are:
1) mean delay time to detect disorder

$$E_0\tau = \mathbf{E}\left\{\tau|\theta = 0\right\};$$

2) mean time between "false alarms"

$$E_\infty\tau = \mathbf{E}\left\{\tau|\theta = \infty\right\},$$

should not be less than some value R.

Consider the case when before the disorder θ observations with a Pareto distribution

$$F_0(x) = 1 - \left(\frac{k}{x}\right)^{\alpha},$$

where $0 < \alpha < 2, 0 < k, k < x$ enter, and after the disorder - with another Pareto distribution

$$F_1(x) = 1 - \left(\frac{k}{x}\right)^{\beta},$$

where $0 < \beta < 2, 0 < k, k < x$.

Denote $\pi_n(x) = \mathbf{P}\left(\tau = n | \theta = 0, S_0 = x\right)$, where $1 \leq x < \mu$. After calculating the variables $\pi_n(x), n \geq 1$, it is possible to find $E_0\tau = \sum_{n=1}^{\infty} n\pi_n(1)$.

For any $n \geq 1$ and $\alpha > \beta$

$$\pi_n(x) = a_n + b_n x^{\frac{\beta}{\alpha-\beta}}, 1 \leq x < \mu$$

is correct. If parameter μ satisfies the condition

$$\left(\frac{\beta}{\alpha}\right)^{\frac{\beta}{\alpha-\beta}} \left(\frac{\beta}{\alpha-\beta}\ln(\mu) - 1 + \mu^{\frac{\beta}{\beta-\alpha}}\right) \leq 1$$

then the sequence

$$E_0\tau = \sum_{n=1}^{\infty} n\pi_n(1) = \frac{1 - \left(\frac{\beta}{\alpha}\right)^{\frac{\beta}{\alpha-\beta}} \left(\frac{\beta}{\alpha-\beta}\ln(\mu) - 1\right)}{\left(\frac{\beta}{\alpha\mu}\right)^{\frac{\beta}{\alpha-\beta}}} - 1$$

converges. For $\alpha < \beta$ we have

$$E_0\tau = 1 + \frac{\left(\frac{\alpha\mu}{\beta}\right)^{\frac{\beta}{\beta-\alpha}}}{1 - \left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\beta-\alpha}} \left(1 + \frac{\beta}{\beta-\alpha}\ln(\mu)\right)},$$

if the condition

$$\left|\left(\frac{\alpha}{\beta}\right)^{\frac{\beta}{\beta-\alpha}} \left(1 + \frac{\beta}{\beta-\alpha}\ln(\mu)\right)\right| < 1$$

is true. In the same way we are able to find $E_\infty\tau$.

Using the method

Consider an idealized strictly alternating ON/OFF source model which allows for packet trains ("ON"time) and inter-train distances ("OFF"time). There are two kinds of OFF time for web object transferring. Inactive OFF time corresponds to the time between transfers of web objects, this is user "think time". Active OFF time corresponds to the time between transfer of components of a single web object. Statistical investigations have shown, that observations of inactive OFF time of web traffic are independent identically distributed r.v.'s with the Pareto distribution function [6]. The traffic Mixing deviate heavy-tail index [8]. Using the method of cumulative sums we can determine the point of difference from a legitimate user's behavior (web robots traffic).

The TCP SYN flooding is the most commonly used attack. This attack exploits the TCP's three-way hand-shake mechanism. The number of TCP connection arrivals per Web session are heavy-tailed with indications of infinite variance behavior [9]. The Pareto distribution function is the mostly used for modeling heavy-tailed distributions. Hence, we can determine anomalous arrival behavior of WWW clients (SYN Flood).

Research was supported by Russian Foundation for Basic Research (RFBR), grant 01-01-00113.

References

1. Shiryaev A.N. Detection of Spontaneous-arising Effects // Dokladyi USSR Academy of Sci, 1961, №4, p.173-174.
2. Shiryaev A.N. Statistical Sequential Analysis, Nauka, 1969, p.232.
3. Page E.S. Continuous inspection schemes // Biometrika, 1954, N41, p.100-114.
4. Mazalov V.V. Method of cumulative sums and problems of optimal control by technological processes // Software Engineering and Knowledge Engineering, 1997, №2, p.219-229.
5. Crovella M., Bestavros A. Self-similarity in world wide web traffic: evidence and possible causes // IEEE/ACM Transactions on networking. 1997, №6, p.835-847.
6. Barford P. Modeling, measurement and performance of world wide web transactions, http://www.welltall.com/ymc/academics/thesis.ps.gz
7. Wang H., Zhang D., Shin K. Detecting SYN flooding attacks. 2002, http://citeseer.nj.nec.com/cleveland00ip.html
9. Feldmann A., Gilbert A., Willinger W., Kurtz T. The changing natureof network traffic: Scaling phenomena, http://citeseer.nj.nec.com/feldmann98changing.html.