基于统计的异常流量发现检测技术

1.1 引言

大规模网络流量异常发现技术的本质是在寻找网络自身特征的的异常变化，即在一定的时间和空间范围内，网络自身的特征会发生明显特性的变化，甚至是剧烈的改变。由于恶意代码的在互联网上的传播具有相当的复杂性和行为不确定性，所以对其监测有一定的难度。

鉴于网络流量具有自相似性，符合一定周期函数。虽然网络流量随时间在不断增长，但是其曲线仍具有自相似性，特别是在广域网上进行较长周期的统计流量更能显现出自相似性的特点，本章通过统计分析正常的网络流量数据，提出正常情况下描述网络流量的理想曲线的方法，运用统计学方法建立大规模网络流量的数学模型。通过对任意一个时间周期内的流量曲线和正常流量曲线的对比，来发现当前流量是否发生异常，当异常流量达到某个临界量的时候，即可以判断出蠕虫疫情暴发。为此在哈尔滨工业大学的教育网出口上进行半年多的统计分析，验证了上述结论。

1.2 网络流量模型

1.2.1 网络流量模型建立的前提条件

如果把网络上所有可用资源视为节点 P，用 P 的数量 M 近似代表网络规模，假设本模型研究的网络对象为 N，则本模型适用的 N 必须符合以下条件：

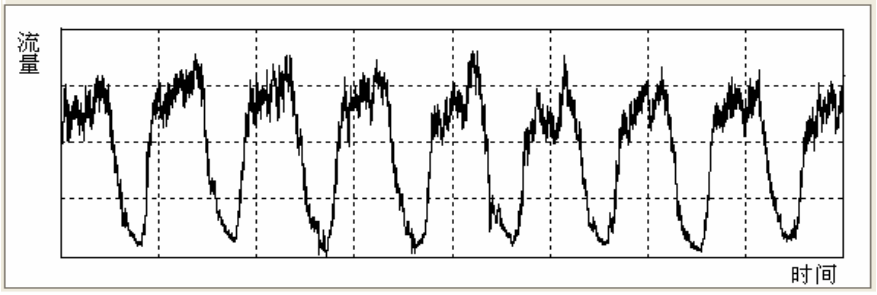

图 0-1 哈尔滨工业大学校园网流量曲线形状

Fig. 0-1 The curve shape of network flow in HIT

第一，单个或少数节点 P 的状态对整个网络 N 不会造成很大的影响；第二、N 的规模 M 不会产生突变，整个网络资源利用规律从总体上看在一定时间段内保持相对稳定，网络流量具有周期性。第三、整个网络流量是可检测的，检测方法是对流经出口的流量作镜像，进行旁路监听。

以上假设的条件是为了保证 N 的流量情况是有规律并且是可以被检测的，一般来讲对于一个企业内部网、校园网或是甚至一个省的总出口来讲都符合这种假设。图 0-1 是哈尔滨工业大学校园网在一个时间段内出如口总流量的统计数据。

1.2.2 模型建立过程中数据的预处理

用函数 $f(t)$ 代表 N 在 t 时刻的流量。根据前提条件 $f(t)$ 是一个周期函数。考

虑函数$f(t)$在其周期T_0中的变化，建立原始周期曲线，图 2 为原始周期曲线。

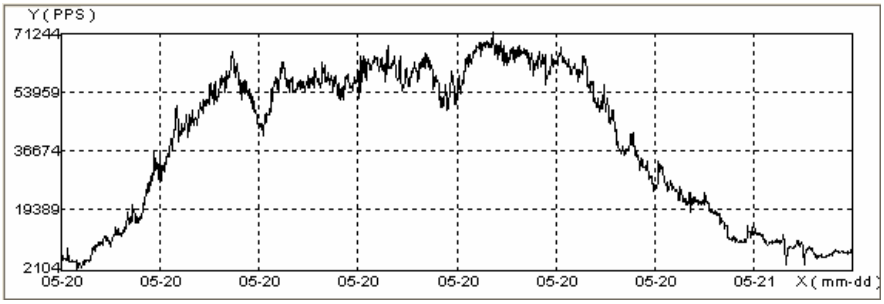

图 0-2 原始周期曲线

Fig. 0-2 The original curve shape in one period

由于原始曲线上的点可能会产生突变(例如监测系统突然断电使统计数据量为 0 等原因)，在对原始数据处理时，利用最小二乘法对原始流量曲线作一下平滑处理。

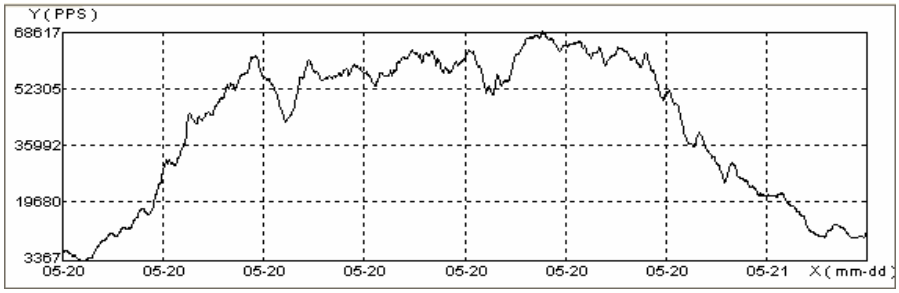

图 0-3 利用最小二乘法进行线性回归后的曲线

Fig. 0-3 The curve shape after linear regression with the least-squares procedure

1.2.3 建立正常流量模型

在建立正常情况下网络数据流模型时，主要从两个方面考虑：流量大小和流量曲线形状。分别构造函数来描述这两个特征量，通过对特征量分析来实现对网络数据流异常的发现。

流量的大小

在一个周期$[T,T+T_0]$内网络流量$f(t)$的平均值为Avg,如果把T取为时间原点($T=0$),则：

$$Avg=\frac{\int_{T}^{T+T_0} f(t)\cdot dt}{T_0}=\frac{\int_{0}^{T_0} f(t)\cdot dt}{T_0} \tag{2-1}$$

在该周期上致密地取 m 个样本点，Avg的离散化形式可表示为：

$$Avg=\frac{\sum_{i=0}^{m-1} f(t_i)\cdot(t_{i+1}-t_i)}{T_0} \tag{2-2}$$

其中：$t_i\in[T,T+T_0]$ 且 $t_0=T,t_m=T+T_0$

同时：$t_0<t_1<\ldots<t_m$

Avg的值从总体上反映了 N 在某一周期上的流量大小情况，它对于局部的

流量变化反应不是很灵敏，然而如果网络上出现大规模异常事件(例如蠕虫暴发等)，变化最剧烈的就是 Avg 值，根据近几次蠕虫暴发时积累的资料数据显示，Avg 的值可以改变为平时的 50 倍以上甚至更高。

曲线形状差异

从 Avg 值上是很难观测出来局部的流量异常。为了确定某一周期内局部流量是否出现了异常，关键要判断出该周期内的函数分布规律是否和正常情况下一样，由于网络绝对流量大小会由于某些正常事件的发生而改变，这种改变和蠕虫暴发时的变化有质的区别，它不会造成流量产生数量级变化，所以在考虑曲线形状的时候，应排除网络绝对流量产生的影响。

把流量函数的定义稍作修改，

令：
$$f^*(t) = \frac{f(t) * AVG_0}{Avg} \tag{2-3}$$

其中：AVG_0 为理想网络的平均流量，这样就可以消除网络流量的大小产生的影响了。

设 $F^*(t)$ 是理想情况下正常网络流量函数，$F^*(t)$ 的曲线形状就是理想网络曲线形状，考虑任意周期 $[T^*, T^* + T_0]$ 内网络流量 $f^*(t)$，为了表示 $f^*(t)$ 和 $F^*(t)$ 之间的形状差异，定义函数 $Q^*(x)$。

令:
$$Q^*(x) = \sqrt{\frac{\int_{T^*}^{T^*+T_0} (f^*(t) - F^*(t+x))^2 \cdot dt}{T_0}} \tag{2-4}$$

取内部积分项为平方的原因是因为考虑到有可能产生正负抵消的因素，x 是相位偏移量，$x \in [0, T_0)$。当 x 发生变化时相当于调节 $f^*(t)$ 和 $F^*(t)$ 之间的相位差。当 $Q^*(x)$ 取得最小值时，说明 $f^*(t)$ 和 $F^*(t)$ 之间达到了最佳吻合。

令:
$$G^*(x) = \min(Q^*(x) \mid x \in [0, T_0)) \tag{2-5}$$

则 $G^*(x)$ 就反映了 $[T^*, T^* + T_0]$ 内的流量和正常流量之间的形状差别。

推广：考虑时间间隔 ΔT 小于 T_0 时的情况，如果有以下条件：不存在 $t_1, t_2 \in (0, T_0)$ 使得在 $\forall t \in (t_1, t_2)$ 有 $F^*(t) = F^*(t + \Delta T)$ 成立，也即在 ΔT 的时域范围内 $F^*(t)$ 的函数曲线不存在两段相同的情况。

那么在 $(T, T + \Delta T)$ 上同样可以定义：

令:
$$G^*(x) = \min(Q^*(x) \mid x \in (0, \Delta T)) \tag{2-6}$$

其中：
$$Q^*(x) = \sqrt{\frac{\int_{t^*}^{t^*+\Delta T} (f^*(t) - F^*(t+x))^2 \cdot dt}{\Delta T}} \tag{2-7}$$

流量均值 AVG_0 和函数 $F^*(t)$ 的确定

为了确定正常数据流模型中的 AVG_0 和 $F^*(t)$，在网络处于正常状态时对网络流量进行采样，每次取起始点为时间零点。把 $[0, T_0]$ 分成 n 个区间，为了便于

计算，采样的时间间隔选为均匀变化。

即：
$$\Delta T = \frac{T_0}{n}$$

对 K 个周期进行采样，并保证网络流量在这 K 个周期内是正常的，采样结果如下：

$$\begin{array}{l}(t_0, f_1(t_0))\ (t_1, f_1(t_1))\ ...(t_n, f_1(t_n)) \\ ..... \\ (t_0, f_i(t_0))\ (t_1, f_i(t_1))\ ...(t_n, f_i(t_n)) \\ ..... \\ (t_0, f_k(t_0))\ (t_1, f_k(t_1))\ ...(t_n, f_k(t_n))\end{array} \tag{2-8}$$

其中：$t_i = i * \Delta T\ \ (i = 0,1,2...n)$

$f_i(t)$ 代表第 i 个周期的采样结果

对以上采样数据进行综合处理，可以得出 $f(t)$ 在理想状态下各时间点的样本估计值：

$(x_0, y_0)....(x_n, y_n)$

其中：$x_i = t_i \quad y_i = \frac{1}{k}\sum_{j=1}^{k} f_j(t_i) \qquad i = 0,1...n$

于是可以得出 AVG_0 的估计值：

$$AVG_0 = \frac{1}{T_0}\sum_{i=0}^{n-1} y_i \cdot \Delta T = \frac{1}{n \cdot k} \cdot \sum_{i=0}^{n-1}\sum_{j=1}^{k} f_j(t_i) \tag{2-9}$$

由于 $(x_0, y_0)....(x_n, y_n)$ 被认为是理想状态下流量函数的估计样本点，所以 $(x_0, y_0)....(x_n, y_n)$ 也即 $(x_0^*, y_0^*)....(x_k^*, y_k^*)$。

在正常情况下，网络数据流量在某一时间点不会发生突变，也即流量函数是平滑的，对以上数据点采用 3 次样条插值的方法得出 $F^*(t)$。

构造 3 次样条函数：

$$F^*(t) = \begin{cases} S_1(t) & t \in (t_0, t_1) \\ ..... & \\ S_i(t) & t \in (t_{i-1}, t_i) \\ ..... & \\ S_k(t) & t \in (t_{n-1}, t_n) \end{cases} \tag{2-10}$$

$S_i(t)$ 都是不高于 3 次的多项式且 $S(x_i^*) = y_i^*$，在解这个方程时，需要确定边界条件。

由于 $F^*(t)$ 的周期性，取边界条件为：

$$\begin{cases} F^{*'}(t_0) = F^{*'}(t_n) \\ F^{*''}(t_0) = F^{*''}(t_n) \end{cases} \tag{2-11}$$

即可解出在整个区间$[0,T_0]$上$F^*(t)$的表达式。

1.2.4 异常数据流的确定

在建立了正常数据流模型后，就可以考虑在任意一个周期T_0内的流量是否产生了异常。确定任意一个周期内流量是否产生了异常，实际上就是要确定这个周期内流量曲线和理想流量曲线之间是否有很大的差距，同样从流量大小和形状两方面来考虑差距。设定临界量G^*_0、Avg_0，当$Avg > Avg_0$时认为流量大小产生了异常，当$G^*(x) > G^*_0$时认为该周期内流量形状产生了异常。

对于流量大小，情况比较简单，根据(2-2)式可以很方便地计算出Avg。为了确定G^*，应把$Q^*(x)$离散化。考虑一个周期时间段$(T, T+T_0)$，在该时间段的样本数据点如下：

$$(t_0, f(t_0)),(t_1, f(t_1)),.....(t_m, f(t_m))\ ,$$

其中：$t_0 = T, t_m = T + T_0$

则：当样本空间 m 足够大，并且足够致密时有:

$$Q^*(x) \approx \sqrt{\frac{\sum_{i=0}^{m-1}(f^*(t_i) - F^*(t_i + x))^2 \cdot (t_{i+1} - t_i)}{T_0}} \qquad (2\text{-}12)$$

$$G^* = \min(Q^*(x) \mid x \in (0, \Delta T))$$

其中：$f^*(t) = \dfrac{f(t) \cdot AVG_0}{\dfrac{1}{m}\sum_{i=0}^{m} f(t_i)} = \dfrac{m \cdot f(t) \cdot AVG_0}{\sum_{0}^{m} f(t_i)}$

G*(x)的计算机求解

在实际计算时，利用以上公式把$(0, \Delta T)$区间平均划分为 n 小段，分别取$x_i = \dfrac{\Delta T \cdot i}{n}$，计算出对应的$Q_i^*$值最后取$G^* = \min(Q_1^*, Q_2^*, ...Q_n^*)$。

临界量的确定

下面来确定临界量Avg_0和G_0^*，对于样本数据点(6)，由于它是理想状态下流量函数的样本估计点，相位差为 0，所以$G_i^* = Q_i^*(0)$。为了简化计算，把区间$(T, T+T_0)$均匀地分成 m 份，使$t_0 = T, t_i = T + i \cdot \dfrac{T_0}{m}$，如果取 T 为时间零点，则：

于是：$$t_0 = 0, t_i = i \cdot \frac{T_0}{m}$$

$$Avg_i = \frac{1}{m}\sum_{j=0}^{m-1} f_i(j\cdot\frac{T_0}{m})$$
$$G_i^* = \sqrt{\frac{\sum_{j=0}^{m-1}(f^*(j\cdot\frac{T_0}{m}) - F^*(j\cdot\frac{T_0}{m}))^2}{m}} \tag{2-13}$$

根据以上公式，分别计算出 $Avg_1, Avg_2 ... Avg_k$，和 $G_1^*, G_2^* ... G_k^*$。

经过大量实验统计发现 Avg_i 近似服从正态分布，而 G_i^* 具有比较复杂的分布规律，为了简化计算可以利用近似的方法去估计 G_0^*，分别以 $Avg_1, Avg_2 ... Avg_k$ 和 $G_1^*, G_2^* ... G_k^*$ 为样本，求出其均值和方差的无偏估计值：

$$\hat{\mu}_G = E(G^*) = \frac{1}{k}\sum_{i=1}^{k} G_i^*$$
$$\hat{\sigma}_G{}^2 = D(G^*) = \frac{\sum_{i=1}^{k}(G_i^* - E(G^*))^2}{k-1}$$

$$\hat{\mu}_{Avg} = E(Avg) = \frac{1}{k}\sum_{i=1}^{k} Avg_i$$
$$\hat{\sigma}_{Avg}{}^2 = D(Avg) = \frac{\sum_{i=1}^{k}(Avg_i - E(Avg))^2}{k-1} \tag{2-14}$$

根据概率论知识，Avg_1 在区间 $(\hat{\mu}_{Avg} - 2\hat{\sigma}_{Avg}, \hat{\mu}_{Avg} + 2\hat{\sigma}_{Avg})$ 分布的概率为 0.9544，而 G_i^* 的取值范围也总是在 $\hat{\mu}_G$ 周围波动。这里统一取估计值：

$$Avg_0 = \hat{\mu}_{Avg} + 2\hat{\sigma}_{Avg}$$
$$G_0^* = \hat{\mu}_G + 2\hat{\sigma}_G \tag{2-15}$$

这里的 Avg_0 和 G_0^* 只是一个参考值，可以根据具体网络的情况稍作调整。

1.3 模型的验证

以上建立的模型中所用的数据流是广义上的概念，具体它可以是任何一种实际的数据流，例如 HTTP 数据流、SMTP 数据流、SYN 包数据流等等。，在网络出现异常时利用上述模型可以很快就发现该异常。以哈尔滨工业大学校园网为例，在校园网的总出口处利用旁路监听的方式，以 $\Delta t = 60$ 秒为时间间隔，对整个校园网流量进行统计，经过半年以上的统计总结出了哈尔滨工业大学校园网正常情况下的理想网络曲线。

这里分别以总流量、SYN 数据包流量和 ICMP 数据包流量为例。

1.3.1 正常网络流量参数

总流量参数

流量函数周期 $T_0 = 24$ 小时

网络总流量：$AVG_{0-Total} = \hat{\mu}_{Avg} = 39756.4$ pps　$\hat{\sigma}_{Avg} = 694.2$　$Avg_0 = 41144.8$

$\mu_{G-Total} = 3581.6$　　$\sigma_{G-Total} = 573.1$　　$G_{Total}^{*} = 4727.8$

总数据流 $F_{Total-flow}^{*}(t)$ 在一个周期内的曲线形状如图 0-4 所示：

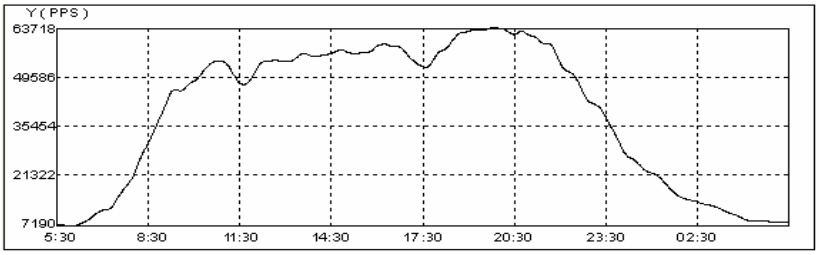

图 0-4 理想情况下总流量的 $F^{*}(t)$ 曲线

Fig. 0-4 The Curve shape of all flow in ideal conditions

SYN 数据包流量参数

流量函数周期 $T_0 = 24$ 小时

SYN 数据包流量：$AVG_{\mathrm{SYN}} = \hat{\mu}_{Avg} = 525\,\mathrm{pps}$　$\hat{\sigma}_{Avg} = 38.8$　$Avg_0 = 602.6$

$\mu_{\mathrm{SYN}} = 81.3$　　$\sigma_{SYN} = 16.6$　　$G_{\mathrm{SYN}}^{*} = 114.5$

SYN 数据流量 $F_{\mathrm{SYN}}^{*}(t)$ 在一个周期内的曲线形状如图 0-5 所示：

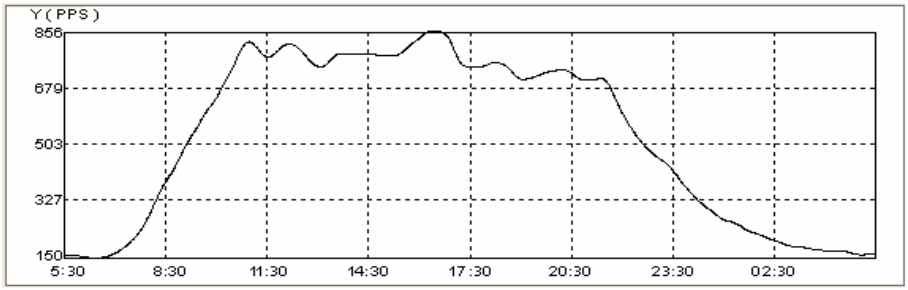

图 0-5 理想情况下 SYN 数据包流量的 $F^{*}(t)$ 曲线

Fig. 0-5 The curve shape of SYN flow in ideal conditions

ICMP 数据包流量参数

流量函数周期 $T_0 = 24$ 小时

ICMP 数据包流量：$AVG_{\mathrm{ICMP}} = \hat{\mu}_{Avg} = 38.6\,\mathrm{pps}$　$\hat{\sigma}_{Avg} = 2.82$　$Avg_0 = 44.24$

$\mu_{\mathrm{ICMP}} = 6.4$　　$\sigma_{ICMP} = 2.9$　　$G_{ICMP}^{*} = 12.2$

ICMP 数据流量 $F_{\mathrm{ICMP}}^{*}(t)$ $F_{\mathrm{SYN}}^{*}(t)$ 在一个周期内的曲线形状如图 2-6 所示：

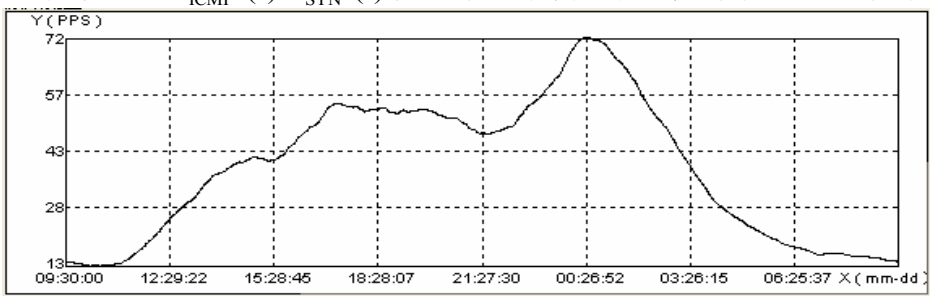

图 0-6 理想情况下 ICMP 数据包流量的 $F^{*}(t)$ 曲线

Fig. 0-6 The curve shape of ICMP flow in ideal conditions

1.3.2 蠕虫暴发时的情况

冲击波蠕虫(Worm.Welchia)暴发时的网络情况：

在冲击波蠕虫(Worm.Welchia)暴发时，由于冲击波蠕虫会发出大量的 ICMP 数据包，考虑冲击波暴发时间段内 ICMP 数据包的变化曲线。

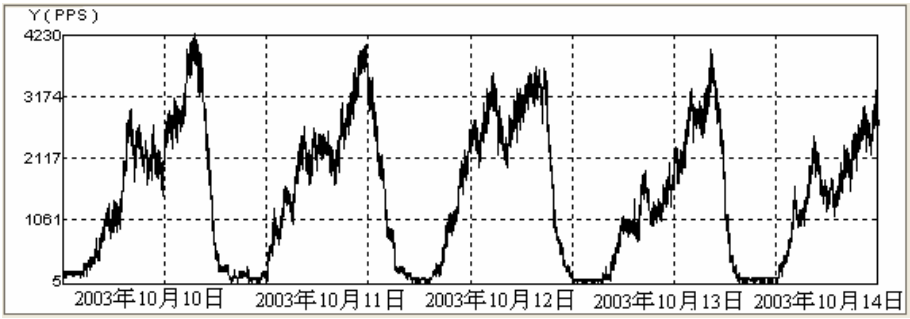

图 0-7 冲击波暴发时 ICMP 数据流量

Fig. 0-7 The curve shape of ICMP flow when weichia was bursting

表 0-1 冲击波蠕虫暴发时的 ICMP 流量参数

Table 0-1 The parameters of ICMP flow when the worm welchia was bursting

月 / 日	10/10	10/11	10/12	10/13	10/14
AVG_{ICMP}	1628.241	1616.976	1693.515	1318.632	1509.7
G^*_{ICMP}	14.6	12.5	12.7	18.6	14.8

可以看出，冲击波蠕虫暴发时，ICMP 数据包流量猛增为平时的 50 倍以上，其数量已经达到了对整个网络流量产生影响的程度，而且曲线形状差异超出了预定的范围。由此可以判定在该网络 N 内有相当一部分机器感染了冲击波蠕虫。

震荡波蠕虫(Worm.Sasser)被清除时的网络情况

根据上述模型，看一下在震荡波蠕虫(Worm.Sasser)清除时流量的变化过程，由于震荡波蠕虫会发出对网络的 SYN 扫描包，分析震荡波蠕虫清除过程中的 SYN 流量曲线。

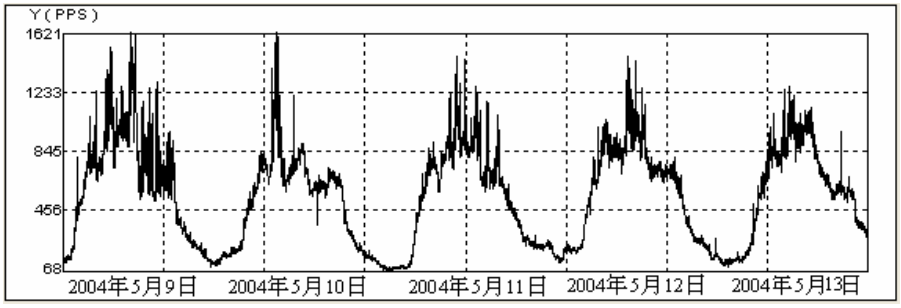

图 0-8 震荡波清除过程中的 SYN 流量曲线

Fig. 0-8 The curve shape of SYN flow when the sasser was being destroyed

表 0-2 震荡波清除过程的 SYN 流量参数

Table 0-2 The parameters of SYN flow when the worm sasser was being destroyed

月 / 日	5/9	5/10	5/11	5/12	5/13
AVG_{SYN}	581.6	489.6	526.9	548.7	538.8

G^*_{SYN}	147.7	138.9	116.5	101.2	112.6

通过以上分析可知震荡波蠕虫在校园网上的传播并不明显，对整个校园网流量没有造成太大的影响，校园网在 5 月 8 日左右对全校进行了处理，从上图可以明显看出，到了 5 月 11 日以后，校园网基本已经恢复了正常。