网络流量的自相似性

1994 年 Leland[1] 等对 Bellcore 的局域网测试与分析的结果显示, 实际网络流量模型具有统计自相似性. 这完全不同于以往的通信领域的传统的业务量模型—基于泊松(连续时间)或贝努利(离散时间)过程, 这些模型是短时相关的. 1995 年, Beran 等通过对大量的不同类别的可变比特率视频流数据的统计发现, 它们也同样表现出一种长相关特性[2,4]. 另外, 对 WAN[3]、FASTPAC[5]等网络的测量, 同样发现这些网络业务量表现出长相关的特性. 而 A. Veres 等人通过模拟产生单个 TCP 对话流量的自相似现象[6], 得出 TCP 流量控制机制也是一个能产生自相似现象的确定性因素 网络流量的这种特性 不仅使得传统的模型无法准确对其描述、分析 同时 也对实际的网络性能有着重要的影响

那么网络流量的自相似特性是怎样产生的呢& 通过对自相似现象成因进行分析后可知 在网络终端用户及多种业务的共同作用下 网络业务流量表现出的突发性是造成自相似性的主要原因'6 7(例如 网络用户的个体行为的突发性与随意性 文件的重尾分布等等原因 而 A Veres 等人通过模拟产生单个TCP 对话流量的自相似现象[8], 得出 TCP 拥塞控制机制也是一个能产生自相似现象的确定性因素

目前有着多种对自相似过程定义 且它们并不是完全等价的 这里采用了文献' 2 9(的描述) 考察一个广义平稳过程 * = * = 0 1 2 3 …% * 表示第 个单位时间内到达的网络流量单元的数目 如到达的数据包的个数 或者到达的字节数% 记 = +' * (2= +' * - %(自相关函数 ,%- +' * - %* + , - %(- 2 令 * , % = * , - + 1 + … + * , % , = 1 2 3 … 称为. * /的 阶的聚合过程 对每个 * · /都表示一个广义平稳随机过程 %为其对应的自相关函数

定义 如果对所有的 阶聚合过程* · /都具有与原过程 * 同样的相关函数结构 即)

% , %- , %- , - 0< < 1 当 , → ∞ 1%

成立 则称 * 为精确二价自相似过程 并称 0 = 1- -2 为其自相似参数 也就是说至少 * · /与 * 直到二阶统计特性是不可分的

定义 如果)

% , %→ , %- , - 当 → ∞ , = 0 1 2 … 2%

则称 * 为渐近二价自相似过程 且具有自相似系数 0 = 1- -2

参数 0 是表述自相似特性的唯一参数 其值越大过程的自相似程度越高 取值范围是 1-2 1%

对自相似过程参数的求解 前人已经提出了许多的方法 其中有 R/S 法'1]、Whittle 估计法'10(、基于小波分析的 EM 法'11(基于周期图的半参数估计法以及方差- 时间图 Variance Time, VTP%分析方法'12(

参数的 分析法

由定义 1、2 我们可以得出相等价的式 3%)

1 * %% - 当 → ∞ 3%

式中 是一正常数 的含义与定义 1、2 中相同

对式 3%)两边取对数得出)

log 1 * %%- log %- 当 → ∞ 4%

由式 4%可以看出 当序列 * 是自相似过程时 如果 → ∞ 则可以通过式 4%求出参数 再通过 0 = 1 - -2 就可以解出 Hurst 参数 具体的求解过程如下步骤所示)

1%假设有一个自相似过程的 * = . * = 0 1 2 3 … / 将 * 中的元素按聚合度 进行聚合 形成新的时间序列 * %= . * 1 % * 2 % * 3 % …/ 其中) * , % = * , - + 1 + … + * , % , = 1 2 3 … 从中也可看出 原始的时间序列 * 其实也就是 = 1 的序列 * 1%

2%对每个时间序列 * % 我们计算它的方差 1 * %)

$$1\ *\ \%\%2\ \frac{1}{-}\sum_{,21}^{-} *\ ,\ \%\%3\left(\frac{1}{-}\sum_{,21}^{-} *\ ,\ \%\right)^{2} \quad 5\%$$

3%根据计算出的 1 * %%与 值 我们可以作出 log %log 1 * %%%图 则可以根据曲线的斜率 得出 值 = - 而 0 = 1- -2 这样就可以估算出自相似时间序列 * 的0 参数值

实际分析过程中 不可能通过绘图进行测量 因此 可以采用回归分析方法 来估算 0 参数值

再来对 VTP 方法进行性能分析 看看其性能瓶颈主要在什么地方 首先 假设序列 * 共有 个数据源需要分析 那么 在计算 Hurst 参数过程中 主要有两个过程需要耗费较大的代价) 1%对每个数据进行聚合形成新的序列 并计算这个新序列的平均值 定义该过程中的计算代价为 2%对每一个聚合形成的新序列 计算方差 定义该过程中的计算代价为

在过程一中 主要有两种类型的操作 加法与除法 如上面的步骤 1%所示 对每一个聚合度 = 2 3 … r -2 计算加法和除法的代价为)

“+”) - 1% - “÷”r - 其中r 表示向下取整

则过程 1%的总的代价为)

$$= 2r\ -2 + 3r\ -3 + \cdots + \frac{-1}{2}\left\lceil \frac{}{-1\%2} \right\rceil + \frac{}{2}$$

$$\cdot\left\lceil\frac{N}{N/2}\right\rceil=\sum_{m=2}^{N/2}k\left\lceil\frac{N}{k}\right\rceil \tag{6}$$

再看过程(2),其中主要计算代价有:①计算所有元素的平方和再求其平均值;②计算所有元素的和求平均值并计算其平方值;③计算 1 和 2 的结果的差.则过程(2)总的计算代价为:

$$P_{vr}=(3N+1)+\{3\lceil N/2\rceil+1\}+\{3\lceil N/3\rceil+1\}+\cdots+\left\{3\left\lceil\frac{N}{N/2}\right\rceil+1\right\}=3\sum_{k=1}^{N/2}\left\lceil\frac{N}{k}\right\rceil+\frac{N}{2} \tag{7}$$

由式(6)和式(7)可以看出,相对来讲,计算 P_{av} 比计算 P_{vr} 的复杂度要更高,因此,在整个 VTP 算法中,计算 P_{av} 是瓶颈.为了提高 VTP 的性能,使之能适应真实网络流量的实时检测,我们将着重提高 P_{av} 的计算速度.

4　基于 VTP 方法的实时估算 Hurst 参数技术

在 VTP 分析法的基础上,文献[13]提出了一种在线实时求解 Hurst 参数的方法.正是在此基础上,我们得出了通过 Hurst 参数的在线实时求解,来在线实时检测 DDoS 攻击的发生.

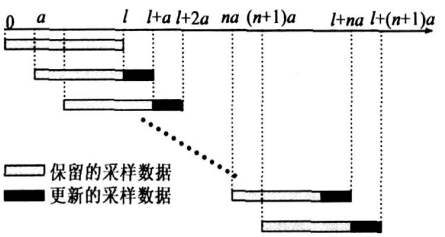


图 1　实时计算 Hurst参数示意图

实时求解真实网络流量的Hurst 参数方法如图 1 所示 在图中 我们以　代表一次计算 Hurst 参数的时间序列长度 每计算一次 Hurst 参数以后 都将这个时间序列的前面长度为　的数据丢弃 而在后面重新添加新采样的长度为　的时间序列 再次计算 Hurst 这样 每一次计算Hurst 参数 仅需要重新采样长度为　的数据即可 而不需要每次都要重新采样　时间长度的数据 不仅解决了计算 Hurst 参数所需的一定长度的时间序列问题 又解决了 VTP 方法中每次计算 Hurst 参数不能准确代表网络流量实时变化的问题

也就是说 选择合适的　与　我们可以方便地调整新采样数据与过去保留数据之间的相关性 如 选择序列总的采样时间　= 100s 每次新添加数据的后移采样时间为　= 10s 同时 我们选择聚合度　的最大值为) -　向下取整%则 计算　的代价由原 VTP 方法的计算所有的 100s 时间内的序列所有的数据变为仅仅计算　= 10s 内的　也因此 计算　的代价仅仅约为 VTP 方法的 1-10 正是因为如此 使得 VTP 方法能实时的应用到真实网络流量的Hurst 参数的检测中

通过对实际网络流量数据分析 我们可以证明该技术的实时检测的高效性 在实际测量中 通过读文件方式 读取从真实网络上采集下来的数据包 进行数据分析 分析的结果显示 除了第一个 Hurst 参数的计算需要较长时间以外 以后的每次计算都能在这次的采样时间结束后 大约 1~ 2s 的时间内计算完成 也就是说 完全能满足实时检测的需要 因为我们每次更新的采样间隔时间为 10s 那么只要能在 10s 内计算完成 就可以满足实时检测的需要 实时检测的耗费时间如表 1 所示

表　实时检测技术计算 Hurst 参数耗费时间

序列	0	1	2	3	4
计算时间	21 5s	2 2s	1 8s	1 8s	1 8s

从表中可以看出 在 VTP 方法不能满足实时性的情况下,该技术完全能满足实际检测的需要.因为,VTP 方法每次耗费时间需要几十秒甚至更多,而且,由我们后面的分析可以知道,仅仅一次计算出的 Hurst 参数值是不能判断是否真正有 DDoS 攻击发生的,所以,使用 VTP 方法,计算出几个Hurst 参数之后,判断出攻击的存在时,已经是攻击发生的几分钟甚至更久以后了.而采用实时检测技术,可以在 20~ 30 秒的时间内完成检测判断.甚至,我们可以调整更新采样时间　的值,取一个更短的采样更新时间,从而在更短的时间内,例如在几秒内根据 Hurst 参数的变化判断攻击的发生

试验数据分析

此次分析所采用的数据是麻省理工大学林肯实验室的入侵检测系统数据库的数据.在这次试验中,整个网络模型被划分为三个域: inside(代表美国空军研究所的内部网络) , outside(代表美国空军研究所的外网部分) , dmz(连接 inside 和 outside 的子网) . 其中 outside 域包括有 Linux、Solaris、SunOS 及 MacOS 等多台主机,以及 SNMP Monitor、外网网关、外网 web 服务器和思科 2514 路由器. dmz 域包括有多台主机与嗅探器以及防火墙、路由器等. 在 inside 域中包括近 40 台的主机以及防火墙等.

在这次试验数据采集时间,被划分为五个阶段: (1) 黑客探测主机 mill. eyrie. af. mil, 这台主机是公用 DNS 服务器; (2) 入侵主机 mill. eyrie. af. mil; (3) 通过 FTP 上传 DDoS 攻击工具 Mstream 以及攻击脚本,并且侵入更多的主机,上传 Mstream 攻击工具的被控制端; (4) 通过 Telenet 登录 mill. eyrie. af. mil, 并初始化 Mstream 攻击工具的被控制端; (5) 通过Telnet 登录到 mill, 并且Telnet 本地的 6723 端口, 连接到 Mstream 的攻击工具控制端,发动了一段时间的 DDoS 攻击. 整个过程的试验数据如图 2 所示. 将图 2 中的 DDoS 流量放大如图 3 所示.

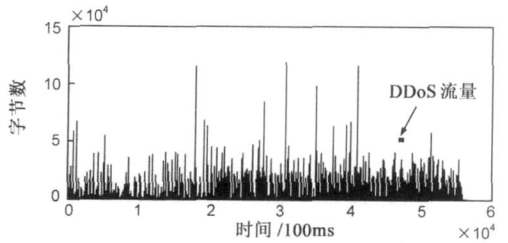


图 2 麻省理工大学林肯实验室数据示意图 (100ms)

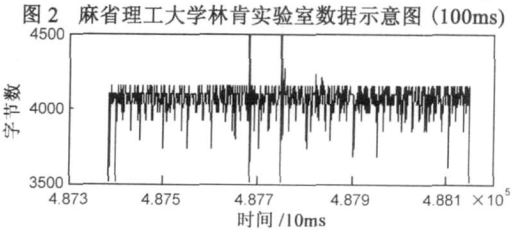


图 3 DDoS 攻击流量放大示意图 (10ms)

黑客在发动 DDoS 攻击前的网络流量, 都是黑客在做 DDoS 攻击准备的网络行为所产生网络流量, 相对于 DDoS 攻击的流量来说, 这都是正常的网络流量. 在对发生 DDoS 攻击前的数据流量分析可知, 此时的网络流量模型的 Hurst 参数在正常的范围内小幅度波动, 如图 4 所示.

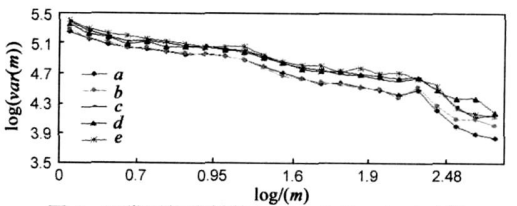


图 4 正常网络流量的 log(*var*(*m*)) - log(*m*)图

图 4 中, 系列的数据曲线是图 1 中所示长度 (此处 = 10000) 数据根据 VTP 分析方法计算出 log(%)与 log %的对应值绘制的, 聚合度 取值为 1, 2, 3, …, 9, 10, 20, 30, …, 100, 200, …, 500. # 系列是 系列的数据丢弃图 1 中所示的长度为 (此处 = 1000) 的数据, 并在后面添加长度为 的新采集数据所构成的新的序列, 计算出的 log(%)与 log %的对应值绘制的. 同理, 序列的曲线都是依次处理所得. 从图中可以看出, 这些曲线的斜率变化趋势基本是一致的, 也就是说 值基本是差不多的, 只是在小幅度内波动. 由图 4 所得的各个序列的 H 值如表 2 所示.

表 正常流量各序列所对应的 Hurst 参数值

序列		#			
0 参数值	0.796	0.785	0.807	0.823	0.815

也就是说, 在正常网络行为下所产生的网络流量, 其 Hurst 参数值是基本稳定的. 并且其值也是在正常的自相似模型范围内.

再来看看当发生 DDoS 攻击时的网络流量对 值的影响. 图 5 是从新添加数据中含有 DDoS 攻击流量开始的第一个序列计算出的一些 log( )) log( )图.

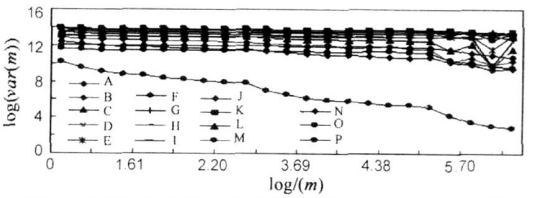


图 5 含有 DDoS 攻击流量的序列的 variance-time 图

为了更清楚地了解 DDoS 攻击流量对 Hurst 参数的影响, 在此次的数据分析中, 开始的 10 个序列, 每次更新的数据长度 为总序列长度的 1% 左右, 而不是如前面的图 4 中所示序列的更新长度约为 10%. 图 5 中, 各个序列中 DDoS 流量占整个序列的流量的比例情况: A 序列 0%, B 序列约为 1%, C 序列约为 2%, D 序列约为 3%, E 序列约为 4%, F 序列约为 5%, G 序列大约为 6%, H 序列约为 7%, I 序列约为 8%, J 序列约为 9%, K 序列约为 10%, L 序列约为 75%, M 序列约为 97%, N 序列约为 98%, O 序列约为 99%, P 序列是完全由 DDoS 攻击流量构成. 可以发现 P 序列的曲线斜率变化趋势明显不同于其他的曲线. 把 P 序列去掉, 再把图 5 上面部分的曲线进行放大, 如图 6 所示, 来看看当 DDoS 攻击流量占有比例逐渐加大时的各序列的曲线斜率变化情况.

从图 6 中可以看出, 从 A 序列(A 序列为正常的网络流量序列) 的曲线开始, 随着 DDoS 攻击流量的逐步增大, variance time 图的曲线渐趋平缓, 亦即 值越来越小 0 值越来越大. 而且还可以发现, 在开始的 1%、2% 的 DDoS 攻击流量对正常流量的 0 参数的影响是很大的. 然后, 随着 DDoS 攻击流量逐步增大, 其对 0 参数的影响逐渐变小, 如表 3 所示. 当 DDoS 攻击流量占到 10% 时, 0 参数已经达到了 0.976, 如表 3 的 K 序列所示 0 值. 也就是说 DDoS 的攻击流量, 在此时实际上增加了网络流量突发性. 实际上, 这也可以很直观地认识到, 当 DDoS 攻击流量加入正常网络流量时, 相对于正常的网络流量, 这是一批突发的网络流量, 因而加大了网络流量的突发特性.

表 含有 DDoS 流量各序列的 Hurst 参数值

序列	A	B	C	D	E	F	G	H
0 值	0.812	0.871	0.928	0.949	0.954	0.965	0.969	0.974
序列	I	J	K	L	M	N	O	P
0 值	0.973	0.975	0.976	0.994	0.990	0.972	0.960	0.443

再从表 3 的 L、M、N、O 这几个序列的 0 值变化情况来看, 在 DDoS 攻击流量占有绝大多数时, 当 DDoS 攻击流量进一步加大时, 0 值在逐渐变小, 也就是说此时, DDoS 流量的比例逐渐增大, 在削弱网络流量的突发性. 然而, 也可以看出来, 当 DDoS 比例已经很大了时,

如O 序列, DDoS 流量已经占有整个序列的99%, 0 值仍然有 0. 960, 依然表现出极大的突发特性. 这是因为, 此时那部分的正常的网络流量, 相对于 DDoS 攻击流量来说, 是一种负方向的突发, 因而, 此时整个序列依然表现出很大的突发特性. 当网络流量完全是 DDoS 攻击流量时, 其表现出了另外一种特性, 已经不再具有突发性, 其0 值也从 O 序列的 0 960 突降到 0 443.

从以上序列的分析过程中, 我们可以看出, 当正常序列中含有 DDoS 攻击流量时, 随着 DDoS 攻击流量所占比例的逐步加大, Hurst 参数值是逐步增大的, 并且是在初期变化明显, 而后变化逐渐趋小. 当 DDoS 攻击流量所占流量比例进一步加大到一定程度时, Hurst 参数值开始跟 DDoS 参数成反比例变化, 但 Hurst 参数值依然很大, 表现依然是网络流量的突发特性. 但当网络流量完全是 DDoS 攻击的流量时, Hurst 参数会有一个突降的过程, 而此时的网络流量特性也就不再具有突发性了.

如果按照图 4 所示的网络流量模型, 每次更新的数据长度占到总的数据序列的长度的 10%, 则如表 3 中A 序列直接到 K 序列, Hurst 参数变化有 0 164 之多, 比正常流量的 Hurst 参数小幅度变化要大得多.

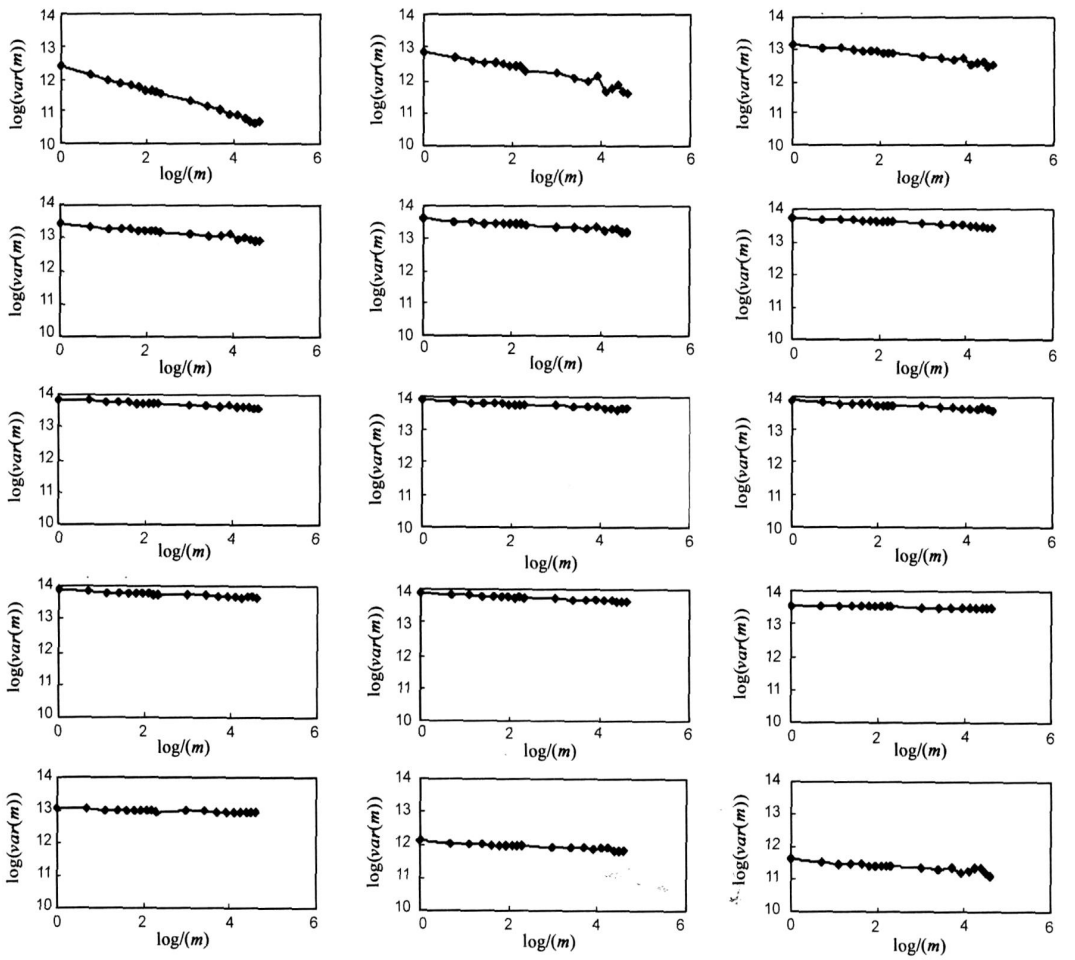


图 6　图 5 部分序列的分解图

结论

从第4 节对 Hurst 参数的在线实时检测技术的分析论证中, 可以得出在线实时技术的时效性满足在线检测的需要; 而从第 5 节的试验数据分析中可以看出, 采用第 3 节所述的 VTP 分析法及第 4 节在线实时计算 Hurst 参数技术, 可以从计算结果中发现 DDoS 攻击流量对 Hurst 参数的影响: 在 DDoS 攻击刚刚开始时, Hurst 参数有增大的趋势, 且变化较大, 不同于正常网络流量的 Hurst 参数小幅度变化; 当序列中的正常流量从有到无时, Hurst 参数有一个从很大到很小的突变, 如表 2 中, 从 O 序列到 P 序列的 Hurst 参数突变. 因而, 从 Hurst 参数的变化上, 可以判断出 DDoS 攻击的发生.

该实时检测 DDoS 攻击技术, 有以下几个优点: (1) 相对于传统的 DDoS 攻击检测技术, 由于采用了保留大部分的已有数据, 而只需要计算更新时间内的数据, 所以计算量更少, 所耗费的计算时间更短, 所以能满足实时检测的需要; (2) 在前面我们分析了 DDoS 攻击过程

中对网络流量的 Hurst 参数的影响, 并得出了 Hurst 参数在 DDoS 攻击过程中的变化变化规律, 从而能更加准确地判断 DDoS 攻击的发生; (3) 采用 Hurst 参数的检测技术, 更能避免传统的基于特征包的检测技术的漏报情况, 提取了所有 DDoS 攻击的共性. 当然, 这种技术也还有许多需要继续研究的地方, 例如, 如何确定实时技术中的 、以及 值, 对计算的复杂度有一定的影响, 有必要做更多的研究; 还有, 如何设计一个准确的智能判断机制, 也有很多的工作需要做.

参考文献:

[1] W E Leland, M S Taqqu, W Willinger, D V Wilson, On the self similar nature of Ethernet traffic (extended version) [J]. IEEE/ ACM Trans on Networking, 1994, 2(1) : 1- 15.

[2] J Beran, R Sherman, M S Traqqu, W Willinger. Long range dependence in variable bit rate video traffic[J]. IEEE Trans on Communication, 1995, 43(2/ 3/ 4) : 1566- 1579.

[3] Paxson V, Floyd S. Wide area traffic: the failure of poisson modeling[A]. Proc ACM Sigcomm' 94[C]. 1994. 257- 268.

[4] M W Garrett, W Willinger. Analysis, modeling and generation of self similar VBR video traffic[A]. Proc ACM Sigcomm' 94 [C]. 1994. 269- 280.

[5] Addie R, et al. Fractal traffic: Measurements, modeling and performance evaluation [A]. In: Proc of INFOCOM' 95[C]. Boston, MA, 1995. 977- 984.

[6] Crovella M E, Bestavros A. Self similarity in World Wide Web traffic evidence and possible cause[A]. Proceedings of ACM Sigmetrics' 96[C]. 160- 169.

[7] W Willinger, M S Taqqu, R Sherman, D V Wilson. Self similarity through high variability: statistical analysis of Ethernet LAN traffic at the source level[J]. IEEE/ ACM Transactions on Networking, 1997, 5(1) : 71- 86.

[8] A Veres, M Boda. The chaotic nature of TCP congestion control[A]. Proceedings of the IEEE Infocom' 2000[C].

[9] B Tsybakov, N D Georganas. On self similar traffic in ATM queues: definitions, overflow probability bound, and cell delay distribution[J]. IEEE/ ACM Trans on Networking, 1997, 5(3) : 397- 408.

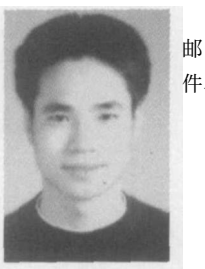
[10] M Garrett. Contribution toward real time service on packet switched networks[D], Columbia University. 1993.

[11] G W Wornell, A V Oppenheim. Estimation of fractal signals from noisy measurements using wavelets[J]. IEEE Trans on Signal Processing. 1992, 40(3) : 611- 623.

[12] Zhang H F, Shu Y T, Yang O. Estimation of Hurst parameter by variance time plots communications[A], Computers and Signal Processing, 1997' 10 Years PACRIM 1987- 1997 Networking the Pacific Rim'[C]. 1997 IEEE Pacific Rim Conference on, 20 22 Aug. 1997(2) : 883- 886

[13] Hagiwara T, Doi H, Tode H, Ikeda H. High speed calculation method of the Hurst parameter based on real traffic[A]. Local Computer Networks, 2000. LCN 2000[C]. Proceedings. 25th Annual IEEE Conference on, 8 10 Nov. 2000. 662- 669.

作者简介:



李金明 男, 1972 年出生于安徽望江, 南京邮电学院博士研究生. 主要研究方向为计算机软件、计算机网络、信息安全等.

王汝传 男, 1943 年出生于安徽合肥, 教授、博士生导师. 主要研究方向是计算机软件、计算机网络和网格、信息安全、移动代理和虚拟现实技术等.