网络流量宏观行为分析的一种时序分解模型[1]

程 光　龚 俭　丁 伟

（东南大学计算机系　江苏 南京　210096）

摘要：大规模网络中的流量行为体现为一个相当复杂的非线性系统，目前国内外对它的研究还没有成熟的方法。文章考虑网络流量非线性的特点，通过不同的数学模型将流量时间序列分解成趋势成分、周期成分、突变成分和随机成分。根据分解，利用相应的数学工具分别建模四个相对简单的子成分以仿真复杂流量。使用分解模型分析CERNET主干网络和NSFNET主干网络的长期流量行为，并将分析结果同传统的 ARIMA 季节模型比较。通过比较仿真自相关函数和预报误差，发现分解模型在描述流量宏观行为时具有简单和高精度的优点。

关键词：非线性，流量宏观行为，分解模型

1．引言

大规模网络管理、规划、设计以及新一代网络体系结构的设计等均离不开对网络流量行为（traffic behavior）的理解。研究网络流量行为首先是要直接对测量流量数据进行统计分析，寻找统计规律，这方面代表性的工作是 MCI 的 Thompson[1]，他完成对 MCI 网络两个测量点的 24 小时、7 天的流量进行详细分析。其次是在流量统计分析的基础上建立流量模型，如：94 年 Leland[2]等人对以太网测量数据进行统计分析发现以太网流量具有自相似性，并建立网络流量自相似模型。对网络流量行为特征的研究还可在不同测量时间粒度上展开。Paxson 和 Floyd[3]的研究发现，不同时间粒度流量服从不同的行为规律：毫秒级的超细时间粒度的流量行为由于主要受网络协议的影响，不体现自相似特征；小时级以上粗时间粒度的流量行为由于主要受外界因素的影响，也不具有自相似性，是一种非线性复杂的过程；只有秒级细时间粒度的流量行为体现出自相似性。本文的研究是粗时间粒度下流量时间序列模型，其结果更多地体现网络行为的宏观特征，因此也称为宏观流量行为。

在描述网络流量行为的模型中，时间序列模型起着相当重要的作用。由于传统的宏观流量时序模型只能处理平稳过程和特殊的非平稳过程，所以描述流量行为误差较大。如：AR[4]模型、MA 模型和 ARMA[5]模型用于解决平稳过程，ARIMA[5]模型、ARIMA 季节模型[6]、小波分解法[7]等处理非平稳过程。由于大规模网络本身是复杂非线性系统，同时又受多种复杂外界因素的影响，其宏观流量行为往往复杂多变，数据中既含有多种周期类波动，又呈现非线性升、降趋势，还受到未知随机因素的干扰，而这些特点难以用传统模型来描述。

考虑网络流量宏观行为的特点，本文使用不同的数学工具将网络流量时序分解成结构相对简单的子成分，通过对各子成分的分别研究来获得复杂流量总体行为的认识，并描述和预测流量行为的非线性规律。文章的最后还通过一组实测数据验证理论模型并同传统的 ARIMA 季节模型比较。

2．网络流量分解模型

[1]本文受国家自然科学基金重点项目 90104031、国家 863 项目 2001AA112060 资助

由于产生网络流量的用户行为受各种外界因素的影响，其上网行为具有一定的规律性和偶然性，同时大规模网络本身是一个非线性系统，因而产生的非线性宏观网络流量具有一定的规律性、突发性和偶然性。根据流量行为的这些特性，表示宏观非线性流量的时间序列 X（t）可以分解为趋势成分 A（t）、周期成分 P（t）、突变成分 B（t）和随机成分 R（t）组成，宏观流量时序表达式可分解如下：

X（t）= B（t）+A（t）+P（t）+R（t）　　（1）

其中趋势成分 A（t）反映的是流量行为因网络用户或环境因素而引起的长期变化趋势。周期成分 P（t）反映的是流量现象的周期性变化。突变成分 B（t）是表示流量行为受到外部突变影响而形成的变化。趋势成分、周期成分和突变成分反映了流量时间序列变化中的确定性成分，随机成分 R（t）又可进一步分解为：平稳时间序列成分 S（t）和噪声 N（t）。

R（t）=S（t）+N（t）　　（2）

在流量时间序列的五个组成成分中，突变成分和噪声属于“无记忆”成分；而 A（t）、P（t）和 S（t）是有“记忆”的成分，它们分别反映 X（t）的长期趋势、周期和平稳过程等三方面的客观行为规律。三种“记忆”可以分别建立数学模型：a（t）、p（t）和 s（t），如果我们忽略影响建模的“无记忆”成分，则根据流量 X（t）的分解模型可以合成模型 x（t）：

x（t）=a（t）+p（t）+s（t）　　（3）

根据流量分解模型（1）和（2），分别采用不同的数学工具将流量分解成五个组成子成分，将“记忆”子成分分别建模，并根据合成模型（3）建立原始序列的时序模型。

2.1 流量突变成分分解

流量突变成分分解模型利用“中位数”是均值的鲁棒估计的事实。流量序列 X（t）剔除突变成分算法如下：

（1）用流量序列 X（t）构造一个新的序列 X'（t），即：

X'（t）＝middle（X（t-2），X（t－1），X（t），X（t＋1），X（t＋2））　　（4）

其中，middle（）是求括号列表序列中位数的函数，t∈[2, n-2]；

（2）同（1）步从 X'（t）的相邻的三个数据中选取中位数构成序列 X"（t），即：

X"（t）＝middle（X'（t－1），X'（t），X'（t＋1））　　（5）

其中，middle（）函数意义同（1），t∈[3, n-3]；

（3）最后使用（6）式由序列 X"（t）构成 X"'（t），

X"'（t）＝X"（t－1）/4+X"（t）/2+X"（t＋1）/4　　（6）

其中，t∈[4, n-4]；

（4）分析序列 X（t）－X"'（t），if　|X（t）－X"'（t）|>k，then　用线性内插值法代替 X（t），其中，t∈[4,n-4]，k 为预先确定值。

对[4,n-4]集合中的每个测量点按以上 4 步处理，即可实现分离流量序列 X（t）中的突变成分 B（t），得到不包含 B（t）的新序列 X1（t）。

2.2 流量趋势成分分解

趋势成分 A（t）分解模型借鉴灰色系统的 GM（1，1）理论模型。GM（1，1）模型能从包含趋势成分、周期成分、随机成分复杂的流量序列 X1（t）中分离出趋势成分 A（t），

趋势成分分解模型具体算法如下：

（1）建立累加方程。设流量序列 X1（t）表示为下列流量序列：

$$X1^{(0)}=\{X1_0^{(0)},X1_1^{(0)},\cdots,X1_i^{(0)},\cdots,X1_n^{(0)}\} \quad (7)$$

其中：$X1^{(0)}$为分离突变成分后的流量序列 X1（t），$X1_i^{(0)}$为第 i 时刻的流量速率，i∈[0, n]。对（7）式 1 次累加生成 $X1^{(1)}$为：

$$X1^{(1)}=\{X1_0^{(1)},X1_1^{(1)},\cdots,X1_i^{(1)},\cdots,X1_n^{(1)}\} \quad (8)$$

式中：$X1_i^{(1)}=\sum_{t=0}^{i}X1_t^{(0)}=X1_{i-1}^{(1)}+X1_i^{(0)}$，i∈[1, n]，$X1_0^{(1)}=X1_0^{(0)}$，$X1_i^{(1)}$实际是从 0 到 i 时刻这段时间内网络流量吞吐量。

由于序列 $X1^{(1)}$接近指数曲线，故认为是光滑离散系数，可用微分方程描述。一阶带系数的微分方程表达式为：

$$\frac{dX1^{(1)}}{dt}+aX1^{(1)}=b \quad (9)$$

其解可写成

$$X1_t^{(1)}=(X1_0^{(1)}-\frac{b}{a})e^{-at}+\frac{b}{a} \quad (10)$$

式中：参数 a、b 为待估参数，

（2）a、b 参数估计。参数估计应用最小二乘法来求解，写成矩阵形式：

$$Y=XB \quad (11)$$

其中

$$X=\begin{bmatrix}-\frac{1}{2}\left|X1_0^{(1)}+X1_1^{(1)}\right| & 1\\ -\frac{1}{2}\left|X1_1^{(1)}(2)+X1_2^{(1)}\right| & 1\\ \vdots & \vdots\\ -\frac{1}{2}\left|X1_{n-1}^{(1)}+X1_n^{(1)}\right| & 1\end{bmatrix}\quad Y=\begin{bmatrix}X1_1^{(0)}\\ X1_2^{(0)}\\ \vdots\\ X1_n^{(0)}\end{bmatrix}\quad B=\begin{bmatrix}a\\ b\end{bmatrix}$$

则：$B=(X'X)^{-1}X'Y$

（3）趋势成分模型 a（t）

确定了参数 a、b 之后，可计算出流量 X（t）中的 A（t）的模型 a（t）为：

$$a(t)=X1_t^{(1)}-X1_{t-1}^{(1)} \qquad t\in[1,n] \quad (12)$$

根据式（10）和（12），流量模型可表示为式（13）

$$a(t)=(e^{-a}-1)(X1_0^{(1)}-\frac{b}{a})e^{-a(t-1)} \qquad t\in[1,n] \quad (13)$$

令 X2（t）=X1（t）－a（t）=P（t）＋R（t）为分离出趋势成分后的剩余序列，X2（t）实际是 X1（t）以 A（t）为轴心的新序列，其优点是新的序列能突出 P（t）成分的作用。

2.3 流量周期成分分解

周期成分分解模型将 X2（t）序列看成是由不同周期的规则波动态叠加而成，因而在分离周期时，逐步分解出一些比较明显的周期波，然后叠加起来作为该时间序列的周期成分。

（1）列出 X2（t）中可能存在的周期。在分析周期之前，事先并不知道这一序列存在多少个周期，所以要根据序列长度，列出可能存在的周期，逐个试验，即：

$$K=\left\{\frac{n}{2}\right\}=\begin{cases}\dfrac{n}{2}，n为偶数\\ \dfrac{n+1}{2}，n为奇数\end{cases} \tag{14}$$

式中： n+1 为 X2（t）序列长度；K 为最大可能周期数。

（2）计算离均差平方和。将时间序列按每一试验周期排列，计算离均差平方和，包括排列组内离均差平方和（式 15）及组间离均差平方和（式 16）。

$$Q_2^2=\sum_{j=1}^{k}\sum_{i=1}^{m}(y_{ij}-\bar{x}_j)^2，\qquad \bar{x}_j=\frac{1}{m}\sum_{i=1}^{m}y_{ij} \tag{15}$$

$$Q_3^2=\sum_{j=1}^{k}m(\bar{x}_j-\bar{x})^2，\qquad \bar{x}=\frac{1}{m}\sum_{i=1}^{m}\bar{x}_i \tag{16}$$

式中：k 为选择的周期长度，m 为组内项数；y_{ij} 为时间序列 X2（t）按试验周期分组排列后的序列值，其自由度 f_2=n-k，其自由度 f_3=n-1。

（3）计算试验周期的方差比。试验周期的方差比计算公式为：

$$F=\frac{Q_3^2/f_3}{Q_2^2/f_2} \tag{17}$$

（4）检验方差。选定某一信度 α =0.05，查 F 分布表得 F_α，if F> F_α，then 试验周期存在；else，不存在该周期，执行第 5 步。

（5）k 从 2 直至 K，逐一取值进行（2）－（4）步骤测试，直至无显著周期为止。

2.4 流量随机成分分解

令 X3（t）=X2（t）-p（t）=R（t）=S（t）+N（t），我们关心的是从随机成分中提取 S（t）成分，建立 s（t）模型。对平稳时间序列成分可建立自回归模型 AR(P)，

$$x(t)=\beta_{p,1}x(t-1)+\beta_{p,2}x(t-2)+\cdots+\beta_{p,p}x(t-p) \tag{18}$$

式中：$\beta_{p,j}$(j=1,2,⋯,P)为自回归系数，P 为模型阶数。算法描述如下：

（1）计算模型系数。自回归系数利用最小二乘法，建立 Yule-Walker 方程组：

$$\begin{cases}\beta_{1,1}=\gamma_1\\ \beta_{k,k}=\dfrac{\gamma_k-\sum_{j=1}^{k-1}\beta_{k-1,k}\gamma_{k-j}}{1-\sum_{j=1}^{k-1}\beta_{k-1,j}\gamma_j}(k=2,3,\cdots)\qquad ,\gamma_k=\dfrac{\sum_{t=1}^{n-k}X3_tX3_{t+k}}{\sum_{t=1}^{n}X3_t^2}\\ \beta_{k,j}=\beta_{k-1,j}-\beta_{k,k}\beta_{k-1,k-j}(j=1,2,\cdots,k-1)\end{cases} \tag{19}$$

式中，$\beta_{i,j}$为自回归系数，γ_k为 X3(t)的 k 阶样本自相关系数。

（2）计算模型阶数。可通过 AIC 准则来确定：

$$AIC = \min\left\{ n \ln \frac{\sum (X3(t) - \overline{X3})^2}{n - P - 1} \right\} \qquad (20)$$

式中：$\overline{X3}$为 X3（t）的均值；n 为 X3（t）序列长度；P 为模型阶数。

（3）平稳序列模型 s（t）。计算 s（t）之前首先要定义 s（0）之前的 P 个数据，根据的反向 X3（t）赋值给 s（t），得到观测点之前的数据 s（－1），s（－2），…，s（－p）。使用这些数据作为初始值，得到模型（21）

$$s(t) = \beta_{p,1} s(t-1) + \beta_{p,2} s(t-2) + \cdots + \beta_{p,p} s(t-p) \qquad t \in [0, n] \qquad (21)$$

3．网络流量分析

为了验证上述分解模型，下面对三组不同时间粒度的实测网络流量进行分析并建模。一组来自于 2001 年上半年通过 CERNET 地区主干某路由器 11 天的流量（cernet01，时间粒度为小时，见图 1），一组来自 2001 年上半年通过 CERNET 国家主干某路由器 121 天的流量（cernet02，时间粒度为日，见图 2），另一组来自于 1988 年 8 月 1 日至 1993 年 6 月 30 日通过 NSFNET 国家主干流量[8]（nsfnet，时间粒度为周，见图 3）。

图1 cernet01时间序列

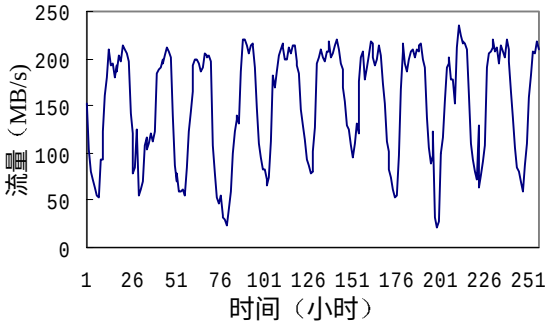


图2 cernet02时间序列

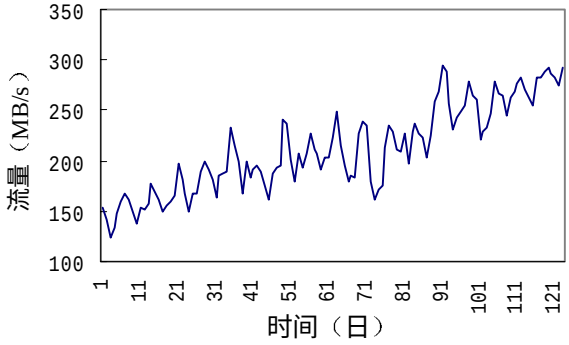
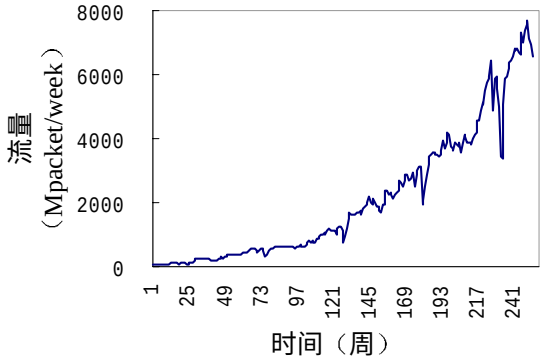


图3 nsfnet时间序列



3.1 各 trace 的分解模型参数

使用分解模型算法分别对三种 trace 建模，模型参数见表 1 所示，

表 1：cernet01、cernet02 和 nsfnet trace 的分解模型参数表

Trace	分解模型参数								
	A（t）		P（t）			S（t）			
	A	B	周期	A	B	AR（p）			
Cernet01	-0.0009	121.109	24	-0.0056	152.802	0.6863	0.0352		
Cernet02	-0.0052	149.592	7	0.0003	50.804	0.6842	0.1016	-0.0090	-0.1706
						0.0639	0.0711	-0.0011	
Nsfnet	-0.018	108.962	26	-0.4057	84.470	0.7024	0.0082	0.0394	-0.1438
			52	－0.6547	95.734	0.0196	-0.1535		

3.2 建模分析

网络流量长期行为基本特性可用自相关函数 ACF（i）来反映。图 2 的 cernet02 trace 时间序列图，由于趋势成分在总流量序列中起决定优势，将周期成分掩盖了，因此在图 2 中流量序列的周期行为不明显。图 4 为剔除 B（t）和 A（t）后剩余流量序列的自相关函数图，由于残余流量序列中周期成分占较大比例，因此图 4 中的序列明显具有 7 天周期行为。

图 5 为将趋势成分、周期成分剔除后的自相关函数图，图 5 的残余序列自相关函数图具有托尾性质，当 lag 增大时，ACF 趋近于 0，这种被负指数控制的衰减形式，其图象如一条越来越小的尾巴，这种行为符合 AR（p）自回归模型的性质。流量序列自相关函数图表明流量分解模型的合理性。

3.3 与 ARIMA 季节模型的比较

将分解模型同 ARIMA 季节模型进行仿真建模和预测比较。cernet01 和 nsfnet 两组 trace 进行预报比较，使用预报误差 error 比较两种模型效果，error 定义为式（22）。cernet02 trace 进行仿真效果比较，使用自相关函数无偏估计样本方差 SSD 比较两种模型的效果，SSD 定义为（23）。

$$error = \sqrt{\frac{\sum_{i=n+1}^{n+1+r}(X_i - \hat{X}_i)^2}{r}} \qquad (22)$$

式（22）中，n 为序列中用于建模的时间长度，r 为预测的长度。cernet01 中 n 为 240，r 为 24。nsfnet 中 n 取 253，r 取 52。

$$SSD(m) = \frac{1}{n-1}\sum_{i=1}^{n}(ACF_m(i) - ACF_s(i))^2 \qquad (23)$$

SSD（m）表示模型 m 的自相关样本方差量，ACF_m（i）表示模型 m 的第 i 阶自相关函数，ACF_s（i）表示实测样本序列的第 i 阶自相关函数。该统计量反映了模型对序列自相关函数的描述效果，值越小，效果越好。

图4：cernet02周期、随机自相关

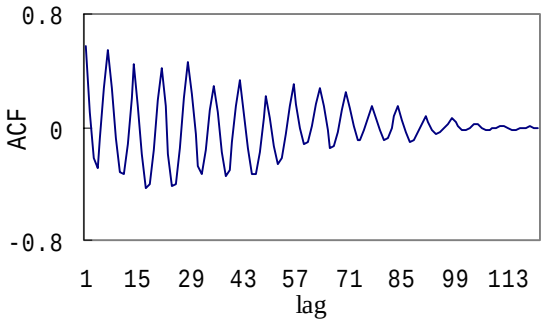


图5：cernet02随机成分自相关

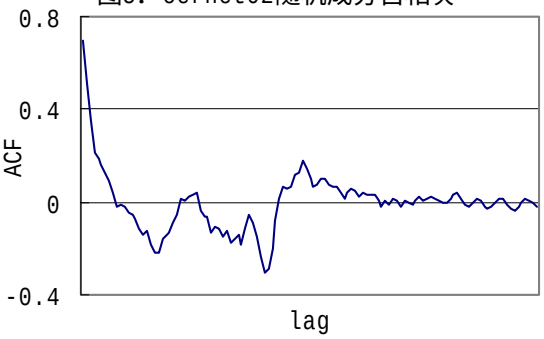


图6： 分解模型与cernet02的自相关函数

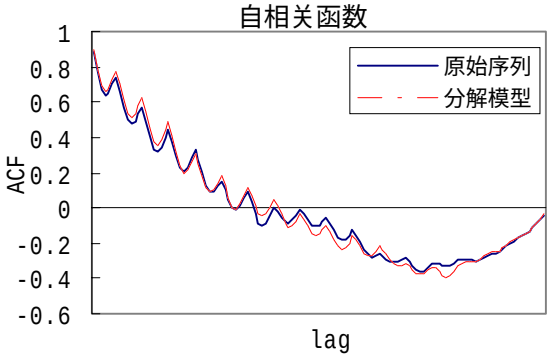
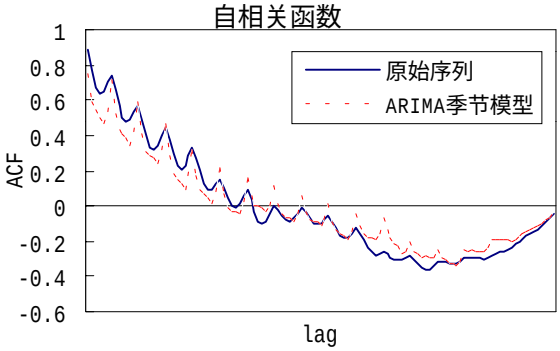


图7： ARIMA季节模型与cernet02的自相关函数



三种 trace 的 ARIMA 季节模型分别为：

cernet01 预报模型：ARIMA（2，0，2）×（0，1，0）24：参数（β_1，β_2，θ_1，β_2）＝（0.1652，-0.676，0.8705，-1294）；

nsfnet 预报模型[6]：ARIMA（2，2，1）×（2，2，0）52：参数（β_1，β_2，θ_1，β_3，β_4）＝（-0.176822，-0.000685，0.993894，-0.273511，0.653531）。

cernet02 仿真模型：ARIMA（7，0，0）×（0，1，0）7：参数（β_1，β_2，…，β_7）＝（0.6606，0.1631，-0.0805，-0.1232，-0.0085，0.1721，-0.2153）。

对 cernet02 采用仿真技术，分别产生 2 种模型的样本，并与原始序列进行比较。图 6 是分解模型与 cernet02 自相关函数的比较，图 7 是 ARIMA 季节模型同 cernet02 自相关函数的比较。从图 6 和图 7 可知分解模型的仿真精度远高于 ARIMA 季节模型的仿真精度。

表 2 为 cernet02 仿真模型中的 SSD 统计量，表 3 为 cernet01 和 nsfnet 预报 error 统计量。从表 2 和表 3 可知，本文的分解模型描述宏观流量序列是非常有效的。理论上分解模型的误差等于流量中的 B（t）和 N（t）成分，所以在描述流量时精度高于传统模型。

表 2：cernet02 各模型的 SSD 统计量

模型	SSD 统计量
分解模型	0.000984
ARIMA 季节模型	0.005471

表 3：各模型的 error 统计量

模型	cernet01 error	nsfnet error
分解模型	6.81	209.31
ARIMA 模型	10.26	421.92

4．结论

本文研究大规模宏观网络流量的分解建模方法，并分别使用小时、日、周三种不同时间粒度的实测数据和仿真数据的自相关函数对该方法进行了验证。研究发现 CERNET 用户行为存在以日为周期和以小时为周期的行为，日周期行为的语义背景是比较清楚的，它反映了用户对上网时间的选择。而周为周期流量行为所反映的用户行为语义则较为复杂，它一定程度上反映了 CERNET 用户对网络的依赖性和使用的后效性。通过分解模型我们发现 nsfnet 流量不仅存在以年[6]为周期的特性，同时还存在半年为周期的特性。当然，对于流量周期行为产生原因的更深入研究需要进一步分析流量中不同网络应用组成成分及会话数等统计量的变化等因素。

分解模型同传统 ARIMA 季节模型的实验结果比较发现分解模型精度高于 ARIMA 模型一倍左右。分析其原因发现分解模型具有以下优点：首先：分解模型从多角度、多侧面、多个不同类型的子模型描述流量行为，而传统模型仅从某一方面描述流量行为性质；同时分解模型描述流量行为需要比传统模型更多的参数，分解模型的仿真误差只有与流量行为规律无关的“无记忆”成分 B（t）和 N（t），因此分解模型能比传统模型更精确、更完整地在描述流量行为。其次：由于分解模型是由各模块组成，每个模块描述流量的某一方面性质，同传统模型相比，对每个模块分别建模较为简单，便于计算机软件实现。

参考文献

[1] Kevin Thompson, Gregory J. Miller, and Rick Wilder. Wide-Area Internet Traffic Patterns and Characteristics [J]. IEEE Network, November/December 1997, 5(6): 10-23.

[2] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic [J].

IEEE/ACM Transaction on Networking, Feb. 1994, 2(1): 1-15.

[3] V. Paxson, S. Flod. Wide-area traffic: The failure of poisson modeling. IEEE/ACM Transactions on Networking [J], June 1995, 3(3):226～244.

[4] Rich Wolski. Forecasting Network Performance to Support Dynamic Scheduling Using the Network Weather Service [DB/OL]. UCSD Technical Report, TR-CS96-494 (1996), http://citeseer.nj.nec.com/wolski98dynamically.html.

[5] S. Basu and A. Mukherjee. Time series models for internet traffic [J]. Proc. IEEE INFOCOM'96, San Francisco, CA., March, 1996, 2: 611-620.

[6] N. Groschwitz, G. Polyzos. A Time Series Model of Long-term Traffic on the NSFnet Backbone [j]. In Proceedings of the IEEE International Conference on Communications(ICC'94), May 1994.

[7] 徐科，徐金梧，班晓娟．基于小波分解的某些非平稳时间序列预测方法[J]．电子学报，2001, 29(4):566-568

[8] K. Claffy, G. C. Polyzos, and H. W. Braun. Traffic Characteristics of The T1 Nsfnet Backbone [J]. proceedings IEEE INFOCOM'93, March 28- April 1, 1993: 885-892.

A Time-series Decomposed Model of Network Traffic Macro-Behavior Analysis

Cheng Guang　Gong Jian　Ding Wei

(Computer Department of Southeast University　Nanjing 210096)

Abstract: Traffic behavior in a large-scale network is very perplexing and can be viewed as a complicated non-linear system. So far the research on traffic behavior doesn't have a well-rounded method. In the paper, according to the character of non-linear network traffic, the traffic time series is decomposed into trend component, period component, mutation component and random component. With such the decomposition, a complicated traffic can be simulated by compound of four simpler sub-series with different mathematical tools. In order to check our model, the long-term traffic behavior of the CERNET backbone network and NSFNET backbone network are analyzed using the decomposed model, and the results are compared with ARIMA model. According to the autocorrelation function value and prediction error function value, the decomposed model has the advantage of simplicity and high precision to describe the traffic marco-behavior.

Keywords: non-linearity, Traffic Macro-Behavior, Decomposed Model

程光，男，1973 年 2 月，安徽黄山人，东南大学计算机系博士研究生，研究方向：计算机网络行为学，email：gcheng@njnet.edu.cn

龚俭，男，1957 年 8 月，上海人，东南大学计算机系博士生导师，研究方向：网络行为学，网络安全，email: jgong@njnet.edu.cn