基于端到端的单播层析网络时延分布①

范献国，黎文伟

(湖南大学 软件学院，长沙 410082)

摘 要：推断网络内部各链路的特性，已成为管理和评估大型电信网络的重要条件。通过某个特定路径直接监测每个链路是不现实的，所以一般通过发送端到端的探测包，利用网络的终端节点来收集网络链路的特征信息。通过单播探测包方法来推断链路的时延特性。针对网络内部链路时延累积量生成函数(CGF)，提出一种基于端到端的单播探测包时延测量的偏差校正估计法。通过仿真，所提出的方法获得的链路延迟特性与直接测量所得的链路延迟特性相比，具有更小的均方误差。

关键词：网络；单播；层析；时延分布

End-to-End Unicast Tomography Inference Network Delay Distribution

FAN Xian-Guo, LI Wen-Wei

(School of Software, Hunan University, Changsha 410082, China)

Abstract: Inference of network internal link statistics has become an important condition for operating and evaluating large-scale telecommunication networks. Since it is not realistic to directly monitor each link along some specific path, so end-to-end probes are used to collect the network link statistics at terminal nodes of the network. This paper uses an unicast probing method to infer the link delay statistics. This paper proposes a bias corrected estimator for the internal link delay cumulant generating function (CGF) based on unicast probe end-to-end delay measurements. This paper shows that the proposed estimator obtains the smaller mean square error comparable to link delay CGF estimates obtaines from directly measured link delay statistics.

Key words: network; unicast; tomography; CGF

1 引言

网络的监管、预测及故障的诊断是网络的运营者和设计者重点考虑的几个因素。一般情况下，无法对传输的数据包进行直接测量，这是因为网络路由器等节点可能不支持这种测量，即使支持但费用开销比较大而无法实施，另外局外人是无法获取由互联网服务提供商所掌握的网络内部各链路的特性。正是基于这些原因使得网络的测量更富有挑战性。根据所使用的网络监测算法可将网络测量分为被动方法和主动方法。被动方法，不主动发送探测包，而是被动监测终端节点或网络中的路由器、主机来收集数据以获得链路参数特征。主动方法是在源节点通过发送专门的探测包(即包对)到目的终端节点来测量网络的时延分布，丢包概率以及包的错误率等参数，且不需要网间节点的协作。虽然被动方式不主动发送探测包避免增加网络负载，但依据被动方法所推断的链路数据不如主动方法的准确。因此，基于主动的方法符合网络层析技术发展的趋势，近些年来，国内外越来越多的研究者开始关注主动方法的单播测量层析技术。

Vardi 提出了一种通过监测网络链路中的数据，来估计网络中从源点到目的节点流的性能参数的网络层析方法[1]。Cao 等人利用上述方法对时变网络传输的特性做了进一步研究[2]。上述文献中所使用的方法都是被动方式，不主动发送探测包。主动发包方法是收集

① 基金项目:国家自然基金(60703097)
收稿时间:2011-03-25;收到修改稿时间:2011-04-29

网络链路的特征的一个可供选择的方法。虽然主动发送探测包的方法在一定程度上会加重网络的负载，但是它能获得比采用被动方法更可靠、更准确的网络链路的性能参数。主动发送探测包有两种方法：一是主动发送多播探测包，二是主动发送单播探测包。本文主要研究单播探测包，现有的研究中提出了单播的多种形式的探测包[3-5]。

在本文中，重点研究估计各链路的时延累积量生成函数(CGF)。包的延迟主要由四部分组成：介质访问时延、传播时延、传输时延及在每个路由器内的排队时延。所有这些延迟的总和，可以通过基于端到端的单播探测包方法来进行测量。在收集足够数量的探测包之后，就能构成有关时延 CGF 的超定系统方程。基于近似最小二乘法，本文提出了有关各链路时延 CGF 的偏差校正估计方法。通过 NS2 网络仿真实验对该方法进行了性能评估。

2　网络延迟模型

假设通信网络由 m 条链路和 n 条路径组成，通过源节点向网络中的 n 条路径，发送同样的探测包来覆盖整个网络。假设每次在测量的时候，网络的路由矩阵 A 是已知的，由 $m\times n$ 构成，a_{ij} 是矩阵 A 中的元素，当路径 i 与链路 j 有重合时 a_{ij} 的值为 1，否则就为 0。设 M_i 表示组成第 i 条路径所有链路的集合，$i=1,...,n$。然后用 $Y_i=\sum_{j\in Mi}X_{ij}$ 表示端到端的探测包经过路径 i 的时延，其中的 X_{ij} 表示第 i 个探测包通过链路 j 所产生的时延，$i=1,...,n$。定义 $K_{Yi}(t)=\log E[e^{tYi}]$ 为端到端时延累积量生成函数，$K_{Yi}(t)=\log E[e^{tYi}]$ 为第 j 条路径的时延累积量生成函数，其中 $j\in M_i, t\in(-\infty,+\infty)$。假设网络是平稳的且链路时延 $X_{ij}(i=1,...,n, j\in M_i)$ 之间相互独立。如果一个路径的探测包 i 与另一个路径的探测包 k 都包含链路 j，那么 X_{ij} 和 X_{kj} 具有相同的时延累积量生成函数，用 K_{Xj} 来表示。因此关于 Y 的 CGF 可以用下面的式子来表示：

$$\begin{aligned}K_{Yi}(t)&=\log E[e^{tYi}]\\&=\log E[e^{t\sum_{j\in Mi}Xij}]\\&=\log\left\{\prod_{j\in M_i}E[e^{tXij}]\right\}\\&=\sum_{j\in M_i}\log E[e^{tXij}]\\&=\sum_{j=1}^{m}a_{ij}\cdot K_{Xj}(t)\\&=A_{(i)}\cdot K_X(t)\end{aligned}\tag{1}$$

$A(i)$ 表示矩阵 A 的第 i 行元素，$K_X(t)=[K_{X1}(t),...,K_{Xm}(t)]^T$ (T 为转置)。因此也可以用 $K_Y(t)=[K_{Y1}(t),...,K_{Yn}(t)]^T$ 来表示端到端的时延累积量生成函数的向量。由此即得线性关系式：

$$K_Y(t)=A\cdot Kx(t)\tag{2}$$

当 $n\geq m$ 且矩阵 A 为满秩时，式(2)是可逆的。因此 $K_X(t)$ 是可以通过有关 $K_Y(t)$ 的方程来唯一确定的。即 $K_X(t)=\left(A^TA\right)^{-1}A^TK_Y(t)$。设 $B=\left(A^TA\right)^{-1}A^T$，则得下式：

$$K_{Xj}(t)=\sum_{i=1}^{n}b_{ji}K_{Yi}(t)\tag{3}$$

当满足 $n\geq m$ 就能确保矩阵 A 是满秩，通过选择发送探测包来覆盖网络。一旦矩阵 A 不为满秩时，网络链路的时延累积量生成函数就无法从式(2)中得出确定的值。

3　网络时延累积量生成函数(CGF)

用 N_i 表示从给定路径 i 中所收集到的探测包数，$i=1,...,n$。定义：

$$\hat{M}_{Yi}(t)=\frac{1}{Ni}\sum_{k=1}^{Ni}e^{tYik}\tag{4}$$

Y_{ik} 表示终端节点所收集的第 k 个探测包沿着路径 i 的端到端延迟。根据 $\hat{M}_Y(t)=[\hat{M}_{Y1}(t),...\hat{M}_{Yn}(t)]^T$ 利用最小二乘法，可以求解 $K_X(t)$。上式中的 $\hat{M}_{Yi}(t)$ 是端到端的矩生成函数 $M_{Yi}(t)=e^{KYi}(t)$ 无偏差函数，对于 $K_{Xj}(t)$ 通过式(3)可以给出其矩估计：

$$\hat{K}_{Xj}(t)=\sum_{i=1}^{n}b_{ji}\log\left(\hat{M}_{Yi}(t)\right)\tag{5}$$

然而，由于 log 的非线性，所以存在偏差。为了获得 $K_{Xj}(t)$ 的偏差估计，这里利用类似文献[6]提出的关于有效带宽估计的方法，来估计网络的时延累积量生成函数，该方法的数学形式类似于网络时延累积量生成函数。

利用 $\log(1+u)=u-\frac{1}{2}u^2+o(u)$，得式(6)如下：

$$\begin{aligned}\hat{K}_{Xj}(t)&=\sum_{i=1}^{n}b_{ji}\log(\hat{M}_{Yi}(t))\\&=\log\{\prod_{i=1}^{n}(\hat{M}_{Yi}(t))^{bji}\}\\&=\log\{\prod_{i=1}^{n}E^{bji}[\hat{M}_{Yi}(t)]-(\prod_{i=1}^{n}E^{bji}[\hat{M}_{Yi}(t)]\\&\quad-\prod_{i=1}^{n}(\hat{M}_{Yi}(t))^{bji})\}\\&=\log\{\prod_{i=1}^{n}E^{bji}[\hat{M}_{Yi}(t)](1-(1-\frac{\prod_{i=1}^{n}(\hat{M}_{Yi}(t))^{bji}}{\prod_{i=1}^{n}E^{bji}[\hat{M}_{Yi}(t)]}))\}\\&\approx K_{Xj}(t)-w_j-\frac{1}{2}w_j^2\end{aligned}\tag{6}$$

其中 $w_j=\dfrac{\prod_{i=1}^{n}(\hat{M}_{Yi}(t))^{b_{ji}}}{\prod_{i=1}^{n}E^{b_{ji}}[\hat{M}_{Yi}(t)]}$。从以上分析及结合式(3)提出用来校正网络时延累积量生成函数的偏差估计：

$$\hat{K}_{Xj}(t)=\sum_{i=1}^{n}b_{ji}\log(\hat{M}_{Yi}(t))+\hat{E}[w_j]+\tfrac{1}{2}\hat{E}[w_j^2] \tag{7}$$

$\hat{E}[\cdot]$ 表示通过式(5)所计算得出的平均经验值，$\hat{E}[w_j]$ 和 $\hat{E}[w_j^2]$ 分别表示：

$$\hat{E}[w_j]=1-\frac{\prod_{i=1}^{n}\hat{E}[(\hat{M}_{Yi}(t))^{b_{ji}}]}{\hat{M}_{Xj}(t)} \tag{8}$$

$$\hat{E}[w_j^2]=1-\frac{2\prod_{i=1}^{n}\hat{E}[(\hat{M}_{Yi}(t))^{b_{ji}}]}{\hat{M}_{Xj}(t)}+\frac{\prod_{i=1}^{n}\hat{E}[(\hat{M}_{Yi}(t))^{2b_{ji}}]}{\hat{M}^{2}_{Xj(t)}} \tag{9}$$

上式中的 $\hat{M}_{Xj}(t)\hat{E}[(\hat{M}_{Yi}(t))^{2b_{ji}}]$ 表示在链路 j 上的时延矩生成函数的估计，可通过下式求得其值：

$$\hat{M}_{Xj}(t)=\prod_{i=1}^{n}(\hat{M}_{Yi}(t))^{b_{ji}} \tag{10}$$

通过采用滑动窗口的方法来获得 $\hat{E}[\hat{M}_{Yi}(t)])^{b_{ji}}$ 的经验平均值，定义窗口的增量 $N_w=\dfrac{N_i-W}{S}$，其中 W 为窗口的大小，S 为步长。

$$\hat{E}[\hat{M}_{Yi}(t)])^{b_{ji}}=\sum_{l=1}^{N_w}\frac{1}{N_w}\left(\frac{1}{W}\sum_{k=(l-1)S+1}^{(l-1)S+W}e^{tY_{ik}}\right) \tag{11}$$

通过同样的方法可得到 $\hat{E}[(\hat{M}_{Yi}(t))^{2b_{ji}}]$ 经验平均值。

4 实验仿真与分析

利用 NS2 来完成拓扑结构如图 1 所示，以 TCP/UDP 为背景流来进行仿真。通过发送 5 个不同的路径探测包来估计 4 条链路的时延累积量生成函数。相对应路由矩阵 A 如下：

$$\begin{vmatrix}1&1&0&0\\1&0&1&1\\1&0&1&0\\0&0&1&1\\0&1&0&0\end{vmatrix} \tag{12}$$

测试环境类似于文献[7]所有的带宽为 8Mb/s，时延为 100 ms，网络的每个链路采用去尾队列管理(带有有限缓存的先进先出队列)。队列的缓存大小为 50 个包，产生的探测包为 40 字节的 UDP 包。每个源节点发送的探测包与探测包之间相互独立且服从泊松分布，其发包间隔为 15ms，速率为 20Kb/s。背景流为 ON-OFF 的 UDP 流和 FTP 流。每个路径发送 N 个探测包，总共的探测包数就为 $5\times N$ 个探测包。通过 N 次试验后，去掉少部分极小的延迟，然后来估计每个包的延迟。当然去掉了极少部分的探测包延迟有可能引起误差，但随着 N 的增大，其误差可忽略不计。为了估计式(8)和(9) 的期望值，我们设置窗口的尺寸 W 的大小为 N 的三分之一，步长 S 取 10 个探测包的延迟作为样本。

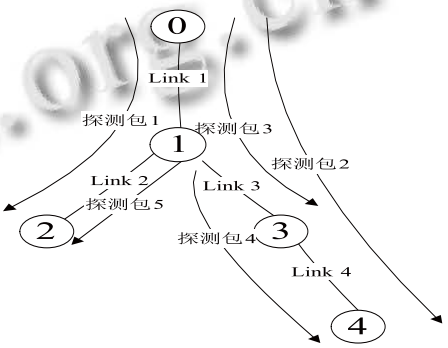

图 1 探测包发送路径图

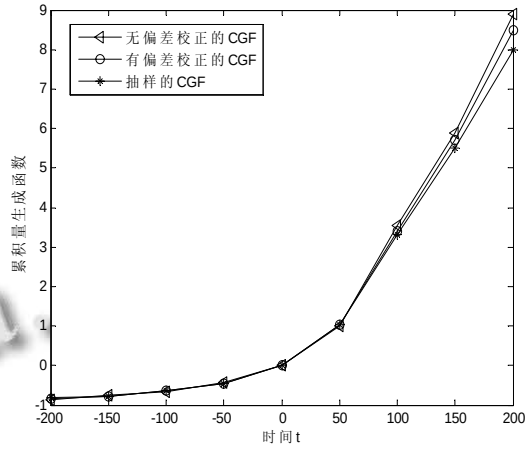

图 2 链路 4 的时延累积量生成函数图

对于 $K_X(t)$ 估计，用本文提出的式(5)来进行估计偏差的校正。当 t 的取值在-200 和 200 之间时，对链路的时延累积量生成函数进行了评估，链路 4 相应的时延累积量生成函数如图 2 所示，其中的 N 的取值为 1500。从图中可看出，有偏差校正时延的 CGF 与没有偏差校正时延的 CGF 相比，更接近于真实的情况，这就说明了带偏差修正的时延的 CGF 比没有偏差修正的 CGF 更准确。其它的几个链路时延累积量生成函数大体情况与此类似。图 3 中有偏差校正 CGF 的均方误差表示

$\hat{K}_{Xj}(t)$，没有偏差修正 CGF 的均方误差表示 $\hat{K}'_{Xj}(t)$。从图中可看出，带有校正的时延累积量生成函数的均方误差更小。 链路的时延累积量生成函数表征着时延的统计特性，因为它是链路时延的概率密度函数经傅里叶变换而来的对数函数。从时延累积量生成函数中我们可以准确的估算出链路的许多时延分布特性。这里利用时延累积量生成函数，给出有关瓶颈链路的诊断[8]，定义链路延迟的概率超过某个事先定义的延迟阈值的概率 P 就为瓶颈，由切尔诺夫界有下式：

$$P(X_j \le \delta) \le e^{-t\delta} \quad E[e^{tX_j}] = P_j \tag{13}$$

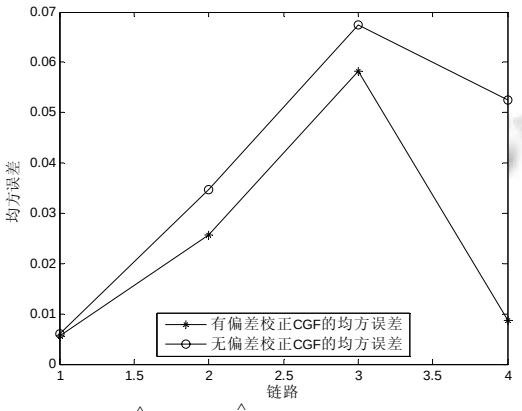

图 3　$\hat{K}_{Xj}(t)$ 和 $\hat{K}'_{Xj}(t)$ 的均方误差对比图

通过选择合适的阈值 δ 和概率 P，P 接近于 1，那么只要存在 P_j 的值满足 $\max_{j=1,\ldots,m} P_j > P$,就能诊断出哪个是瓶颈链路。表 1 中的切尔诺夫界 $P(X_j \le \delta = 0.02s)$ 由仿真模拟给出。通过设置阈值 P=0.90 可以看出链路 3 为链路瓶颈。

表 1　各链路的切尔诺夫界和经验估计值

链路	1	2	3	4
P_j	0.7636	0.5050	0.9483	0.8916
$P(X_j \le \delta)$	0.2413	0.1821	0.4547	0.2845

5　结语

本文利用单播的方法对网络内部的时延特性进行了推断。针对时延偏差，我们提出了网络内部链路时延累积量生成函数偏差的估计校正方法，该估计校正法是基于近似最小二乘法的方法。该方法在 NS2 平台上，以 TCP/UDP 为背景流，在先进先出的有限缓冲的队列管理模式下进行性能评估。通过仿真结果可看出，该偏差校正法与没有偏差校正的方法相比,具有更小的均方误差。

未来工作主要在以下几个方面：本文提出的时延偏差校正法是假设整个测量阶段网络是平稳的，这有悖于真实的网络环境。为了掌握网络链路的实时延迟特性，该校正方法还需要做进一步的改进。此外如果网络内部的链路时延不服从空间和时间独立的话，那么就要提出更复杂的网络模型。

参考文献

1 Vardi Y. Network tomography: Estimating sourcedestination traffic intensities from link data. J. Amer. Statist. Assoc. , 1996,91:365−377.

2 Cao J, Davis D, Wiel SV, Yu B. Time-Varying network Tomography: Router link data. Journal of the American Statistical Association, 2000,95(452): 1063−1075.

3 Coates M. Nowak R. Network tomography for internal delay estimation. Proc. IEEE Int. Conf. Acoust, Speech,and Signal Proc,2002.

4 Duffield N, Lo F, Paxson V. Network Loss Tomography Using Striped Unicast Probes. IEEE/AC M Trans. Networking, Aug. 2006,14(4): 697−710.

5 Qian F, Hu GM, Yao XM. Unicast Network Loss Tomography Using#R-Cast Probes Scheme. cmc, 2009 WRI International Conference on Communications and Mobile Computing, 2009, 3: 113−117.

6 Gibbens RJ. Traffic characterisation and effective bandwidths for broadband network traces.Stochastic Networks, Theory and Applications, Oxford Science Pub.1996:169−179.

7 Presti FL, Duffield NG. Multicast-based inference of network-internal delay distributions. Preprint from MINC, http://wwwnet.cs.umass.edu /minc/.

8 Johnson N, Thompson J, McLaughlin S. Network tomography delay estimation&bottleneck link discover. 2008, IAPR. Workshop on Cognitive Information Processing, Santorini, Greece June9−10, 2008.