

改进的基于最大似然的快速拓扑估计方法

王育红, 费高雷, 胡光岷

(电子科技大学 宽带光纤传输与通信网技术教育部重点实验室, 成都 611731)

(wyhwxh6515@uestc.edu.cn)

摘要:基于最大似然的网络拓扑估计方法能够获得全局最优的估计结果, 优于一般局部最优化和节点对融合方法, 但在网络规模较大时存在计算复杂度较高的缺点。首先证明了网络拓扑估计似然函数是单峰的且峰值为最大值; 然后利用单峰特征, 改进了现有基于最大似然的拓扑估计方法, 降低了计算复杂度。最后, Matlab 和 NS-2 仿真结果证明, 在不降低拓扑估计准确率的情况下, 改进算法将计算复杂度减少了 30%~40%。

关键词:拓扑估计; 最大似然; 网络层析成像; 三明治包; 马尔可夫链蒙特卡洛算法

中图分类号: TP393.02 **文献标志码:** A

Improved fast topology estimation method based on maximum likelihood

WANG Yu-hong, FEI Gao-lei, HU Guang-min

(Key Laboratory of Broadband Optical Fiber Transmission and Communication Networks of Ministry of Education,
University of Electronic Science and Technology of China, Chengdu Sichuan 611731, China)

Abstract: Maximum likelihood based topology estimation method can obtain globally optimal estimation results, and its performance is better than that of other methods such as general local optimization and node fusion method. However, when network scale is large, the topology estimation is computationally intensive. To solve this problem, firstly, the authors proved that the network topology estimation likelihood function is a single-peaked function and the peak value is maximum value. Using the single-peaked property of likelihood function, the current maximum likelihood based topology estimation method had been improved. Finally, the results of Matlab and NS-2 simulations show that the improved method cuts computational complexity by 30%~40% without reducing topological estimation accuracy.

Key words: topology estimation; maximum likelihood; network tomography; sandwich packet; Markov Chain Monte Carlo (MCMC) algorithm

0 引言

在互联网遍及全球的今天,为了更好地控制和管理网络,必须高速感知网络状态参数(如:网络拓扑、流量、时延、丢包率等)及其变化。目前普遍使用的各种网络状态参数估计方法都是在网络拓扑已知的前提下进行的,当网络拓扑发生变化甚至遭受破坏时,需要及时掌握最新的网络拓扑信息,进而得到整个网络的状态参数,因此网络拓扑估计是网络测量的关键问题之一。

网络拓扑估计是利用各种测量手段对所关心网络的逻辑拓扑进行推测,常用的方法主要有两类。一类是传统网络拓扑识别方法,该方法最早由 Siamwal 等人^[1]提出,它要求所有的中间节点都是可协作的。但是随着网络规模的不断扩大和对安全性的要求越来越高,节点间协作变得越来越困难,因此该类方法的实施也受到挑战。另一类方法是基于网络层析成像拓扑估计方法^[2-3],该类方法假定所有中间节点均不协作,通过发送端到端的探测包,收集并分析探测包的统计特征(如时间差的平均值和方差,以及丢包率等)来推测网络拓扑。Ratnasamy 和 Duffield 等人^[4-5]在基于网络层析成像的拓扑估计研究上做出了开创性的贡献,他们从一个源节点向一系列目的节点发送多播探测包,估计源节点到两个目的节点

共享路径上的丢包率,然后利用 DBT 算法(Deterministic Binary Tree Classification Algorithm)实现二元逻辑树状拓扑的识别。由于现有多数网络采用单播方式进行数据传输,单播网络拓扑估计方法的研究对实际应用具有十分重要的意义。Coate 等人^[6-7]提出了一种基于最大似然估计的 MLT(Maximum Likelihood Network Topology Identification)算法。首先用单播的“Sandwich”包进行端到端测量,然后引进马尔可夫链蒙特卡洛(Markov Chain Monte Carlo, MCMC)算法,通过插入和删除节点两个步骤产生一系列候选树,选择这些候选树中似然函数上界值最大的树(最大似然树)作为估计出的拓扑。Shih 等人^[8]用有限混合模型对“Sandwich”包在共享路径的时延差建模,进而采用图聚类的方法获得网络拓扑结构。

基于最大似然的网络拓扑估计方法能够获得全局最优的估计结果,优于一般局部最优化和节点对融合方法,但是在网络规模较大时该算法的计算复杂度很高,原因有以下两个方面:1)每棵树似然函数上界值的计算实际上是求解一个非线性最优化问题,树叶节点对的数目即为最优化问题的维数。对于大规模网络,由于节点数较多,因此每个非线性最优化问题的维数很高,使得求解每棵树的似然函数上界值计算量很大;2)为了能够搜索到全局最大似然树,现有的基于最大似然的拓扑估计算法 MLT 算法在用 MCMC 算法搜索最大似然

收稿日期:2010-06-28;修回日期:2010-08-06。

基金项目:国家自然科学基金资助项目(60872033);新世纪优秀人才支持计划资助项目(NECT-07-0148)。

作者简介:王育红(1986-),女,山东威海人,硕士研究生,主要研究方向:网络层析成像、计算机通信网; 费高雷(1982-),男,浙江诸暨人,博士研究生,主要研究方向:计算机通信网、网络层析成像; 胡光岷(1966-),男,四川眉山人,博士生导师,主要研究方向:计算机通信网、网络行为学与安全。

树的过程中,会以一定的概率转到似然值较小的拓扑上,这使得搜索到的相应可能树的数目很大,从而求解非线性最优化的次数也很多,计算量很大。

经过以上分析,本文对 MLT 算法进行了改进。首先证明了网络拓扑估计似然函数是单峰且峰值为最大值,由此得到似然函数只存在全局最大值而不存在局部最大值的结论。然后利用单峰特性,在搜索最大似然树的过程中可以一直沿着似然值增大的方向进行,从而改进了 MLT 算法。改进后的算法有效的降低了算法的计算复杂度。

1 网络拓扑估计问题

1.1 基本定义

R 为目的节点集合,其中目的节点总个数为 n 。 K 为发向每一对目的节点对的三明治探测包的总数。 $\delta_{i,j}(k)$: $\{r_1(k) = i, r_2(k) = j\}$, 发向目的节点对 $\{i, j\}$ 的第 k 个探测包中,两个小包发送到的目的节点为 $r_1(k)$, 大包发送到的目的节点为 $r_2(k)$ ($k = 1, \dots, K$), $\delta_{i,j}(k)$ 表示两个小包到达 $r_1(k)$ 的时延差,与时钟没有关系,不需要保持时间同步。假设这些测量是独立同分布的,这一假设在探测包的间隔时间足够大的情况下是合理的。 $\sigma_{i,j}$ 为发向目的节点对 $\{i, j\}$ 的两个小包的平均时延差, $x_{i,j} = \frac{1}{K} \sum_{k=1}^K \delta_{i,j}(k)$ 为发向目的节点对 $\{i, j\}$ 的两个小包时延差的测量方差。 $f^l(i)$ 为节点 i 的 l 级祖先节点。 i 的 0 级祖先节点 $f^0(i)$ 就是节点 i 本身, i 的 1 级祖先节点 $f(i)$ 是节点 i 的父节点,则 $f^l(i)$ 的递归定义为 $f^l(i) = f(f^{l-1}(i))$ 。

1.2 最大似然估计函数

假设在探测过程中拓扑保持不变, F 表示所有可能的拓扑树的集合(又被称为拓扑树空间或森林), T^* 表示估计出的最佳拓扑。

拓扑识别的目的是在整个拓扑树空间 F 中利用端到端测量数据 x 寻找真实的拓扑树。测量数据 $x = \{x_{i,j}, i, j \in R\}$, 每一条链路 e_l 上的延时为 μ_l , 则测量值 $x_{i,j}$ 的真实值为 $r_{i,j} = \sum_{l \in s_{i,j}} \mu_l$, 其中 $s_{i,j}$ 为目的节点对 $\{i, j\}$ 共享链路的集合, 共享路径越长则 $r_{i,j}$ 越大。由于 $x_{i,j} = \frac{1}{K} \sum_{k=1}^K \delta(k)$, 并且 $\delta(k)$ 是独立同分布的, 根据中心极限定理:

$$x_{i,j} \sim N(r_{i,j}, \sigma_{i,j}^2), \quad (1)$$

其中: $N(r_{i,j}, \sigma_{i,j}^2)$ 是均值为 $r_{i,j}$, 方差为 $\sigma_{i,j}^2$ 的高斯分布。令 $\mu(T) = \{\mu_l\}$, l 包含拓扑树 T 中的所有链路, 测量数据 x 的概率密度分布用 $p(x|T, \mu(T))$ 表示。

定义拓扑树 T 的对数似然函数为:

$$L(x|T) = \lg p(x|T, \mu(T)) \quad (2)$$

在拓扑空间中有些树和真实拓扑树很接近但是拥有更多的链路,从而和测量数据更加吻合,最大似然值比真实拓扑树更大,造成误判。为了解决这一问题,Coates 等人引入了链路惩罚参数机制,以 λ 为惩罚参数的惩罚对数似然函数如下:

$$L_\lambda(x|T) = \lg p(x|T, \mu(T)) - \lambda n(T) \quad (3)$$

$n(T)$ 是拓扑树 T 中的总链路数,对于惩罚参数 λ 的选取详见文献[6], 本文将 λ 取为 3。惩罚最大似然树为:

$$T_\lambda^* = \max_{T \in F} L_\lambda(x|T) \quad (4)$$

2 改进的 MLT 算法

基于最大似然的拓扑估计算法——MLT 算法,在用

MCMC 算法搜索最大似然树的过程中,为了避免收敛于局部极大值,会以一定的概率转到似然值较小的拓扑上。本章首先证明拓扑估计似然函数是单峰的且峰值为最大值,因此 MCMC 算法在搜索最大似然树的过程中可以一直沿着似然值增大的方向进行。利用这一特性改进了 MLT 算法,减小了计算复杂度。似然函数单峰性的证明如下。

求证:式(2)中的似然函数 $L(x|T) = \lg p(x|T, \mu(T))$ 是单峰的且峰值为最大值。

证明 对于一组测量值有 $x_{i,j} \sim N(r_{i,j}, \sigma_{i,j}^2)$, 似然函数写为:

$$L(r) = \lg \left(\prod_{i=1}^{n-1} \prod_{j=i+1}^n \frac{1}{\sqrt{2\pi}\sigma_{i,j}} e^{-\frac{(x_{i,j}-r_{i,j})^2}{2\sigma_{i,j}^2}} \right) \quad (5)$$

要证明式(5)的 $L(r)$ 是单峰的(即只有一个极值)且峰值为最大值,只需证明式(6)中的函数 $g(r)$ 是单峰的且峰值为最大值即可。

$$g(r) = \prod_{i=1}^{n-1} \prod_{j=i+1}^n \frac{1}{\sqrt{2\pi}\sigma_{i,j}} e^{-\frac{(x_{i,j}-r_{i,j})^2}{2\sigma_{i,j}^2}} \quad (6)$$

式(6)两边同时取对数得:

$$\ln g(r) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \ln \frac{1}{\sqrt{2\pi}\sigma_{i,j}} - \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{(r_{i,j} - x_{i,j})^2}{2\sigma_{i,j}^2} \quad (7)$$

对于一组测量值, $\sum_{i=1}^{n-1} \sum_{j=i+1}^n \ln \frac{1}{\sqrt{2\pi}\sigma_{i,j}}$ 为常数,所以只需证式(8)中的函数 $f(r)$ 是单峰的且峰值为最小值即可。

$$f(r) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{(r_{i,j} - x_{i,j})^2}{2\sigma_{i,j}^2} \quad (8)$$

令 $r_m = r_{i,j}$, $x_m = x_{i,j}$, $\sigma_m = \sigma_{i,j}$ ($j > i$ 且 $i, j \in \{1, 2, \dots, n\}$, $m = 1, 2, \dots, \frac{1}{2}n(n-1)$), 则

$$f(r) = \sum_{m=1}^{\frac{1}{2}n(n-1)} \frac{(r_m - x_m)^2}{2\sigma_m^2} \quad (9)$$

只需证明式(9)的 $f(r)$ 是单峰的且峰值为最小值即可。根据多元函数无条件极值的充分条件^[9], 可以求出 $f(r)$ 的极值点, 步骤如下。

1) 令 $f(r)$ 的梯度 $\text{grad } f(r) = 0$, 得到唯一的驻点 $r_0 = (x_1, x_2, \dots, x_m)$ 。

2) 函数 $f(r)$ 在 r_0 点的海森矩阵为:

$$H_{f(r_0)} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & \cdots & 0 \\ 0 & 1/\sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & 1/\sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/\sigma_m^2 \end{bmatrix}$$

3) $H_{f(r_0)}$ 的所有顺序主子式均大于零, 所以 $H_{f(r_0)}$ 是正定矩阵。

4) 根据多元函数无条件极值的充分条件可以得到: 点 r_0 是 $f(r)$ 的极小值点。由于 $f(r)$ 只有一个极小值点 r_0 , 所以 r_0 也一定是 $f(r)$ 的唯一最小值点, 由此可得函数 $f(r)$ 是单峰的且峰值为最小值, 进一步可得似然函数 $L(x|T)$ 也是单峰的且峰值为最大值。证毕。

利用似然函数单峰性改进的 MLT 算法如下:

步骤 1 任意选择一棵有 n 个目的节点的树作为初始状态树, 该状态设为 $S_0 = (T_0, \mu_0)$, 根据测量数据 x 计算该状态

下的似然值记为 $p(T_0, \mu_0 | x)$, 设置最大迭代次数为 t 。

步骤2 用 S' 表示当前状态, $p(T', \mu' | x)$ 表示状态 S' 下的似然值; 初始化 S' 和 $p(T', \mu' | x)$, 令 $S' = S_0, p(T', \mu' | x) = p(T_0, \mu_0 | x)$ 。

步骤3 从 S' 开始任意地删除或者增加 1 个内部节点得到下一个状态 $S_1 = (T_1, \mu_1)$, 如果该状态已经被访问过就跳过, 直到找到一个没有被访问过的状态 S_1 , 计算在 S_1 状态下的似然值记为 $p(T_1, \mu_1 | x)$ 。

步骤4 如果 $p(T_1, \mu_1 | x) > p(T', \mu' | x)$, 则当前状态转到 S_1 , 有 $S' = S_1, p(T', \mu' | x) = p(T_1, \mu_1 | x)$; 如果 $p(T_1, \mu_1 | x) < p(T', \mu' | x)$, 当前状态 S' 不变。

步骤5 重复步骤 3、4, 直到迭代次数为 t , 结束迭代。

步骤6 估计出的最佳拓扑树 $T^*(S')$ 。

3 仿真实验

为了验证改进的 MLT 算法的性能, 分别用 Matlab 和 NS-2 进行了仿真。

3.1 Matlab 实验

分别产生 7 种不同目的节点数 (7, 8, ..., 13) 的树状拓扑, 对每种节点数的拓扑都生成 1000 个拓扑。对于每个拓扑的每对叶节点, 用 Matlab 生成 100 个节点对相似度样本, 生成方式如下: 给定叶节点对 $\{i, j\}$, 其相似度样本是共享路径上所有链路度量之和。每个链路度量样本是随机产生的, 服从均值为 1、方差为 0.25α 高斯分布, 其中 α 是倍数因子。在所有共享链路的集合中随机选取一条链路, 将 α 设为 3, 其余链路 α 等于 1。这与实际情况是符合的, 因为网络流量在某些链路上会呈现突发性, 使得经过这些链路的探测包具有较大的测量方差。实验中, 将原 MLT 算法、改进的 MLT 算法的最大迭代次数均设为 5000, 惩罚参数均设为 3。

利用原 MLT 算法和改进的 MLT 算法进行拓扑估计, 分别统计两种算法的拓扑估计正确率以及估计每一个拓扑时搜索过的平均拓扑数。

由图 1 可知, 两种算法的拓扑估计准确率的差不超过 2%, 说明在迭代次数足够大的情况下, 改进的 MLT 算法在拓扑估计准确率方面和原 MLT 算法性能差别不大。然而, 在计算量方面, 改进的 MLT 算法有效降低了原 MLT 算法的计算量。如图 2, 每估计一个拓扑, 改进的 MLT 算法搜索的平均中间拓扑个数比原 MLT 算法减少 30%~40%, 由于计算量主要集中在计算拓扑树的似然函数上界值上, 因此搜索过程中搜索到的中间拓扑树越少则计算量越小。

3.2 NS-2 仿真实验

为了更好地模拟真实网络环境, 用 NS-2^[10] 搭建了一个 12 节点的拓扑 (如图 3)。总共做了 6 次实验, 向目的节点对发送不同数目 (50, 100, 150, 200, 250, 300) 的三明治探测包, 每次实验进行 30 次仿真, 一次仿真时间为 300 s。拓扑搜索的最大迭代次数设为 5000, 惩罚系数为 5。

拓扑估计正确率方面: 由表 1 知, 在相同发包数下, 两种算法的拓扑估计正确率基本一样, 说明改进的 MLT 算法与原 MLT 算法的拓扑估计正确率相差不大, 和 Matlab 实验得到的结论一样。

计算量方面: 每估计一个拓扑, 改进的 MLT 算法搜索的中间拓扑的个数明显要少于原 MLT 算法, 如图 4, 由于算法的计算量主要集中在中间拓扑树的似然函数上界值的计算上,

因此改进的 MLT 算法的计算量要小于原 MLT 算法。

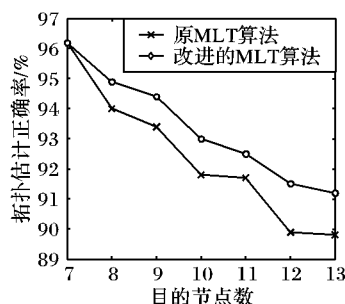


图1 Matlab 仿真两种算法的拓扑估计正确率

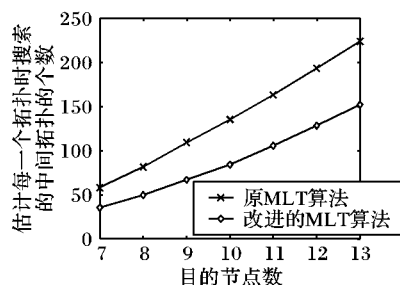


图2 Matlab 仿真每估计一个拓扑, 两种算法搜索到的中间拓扑数

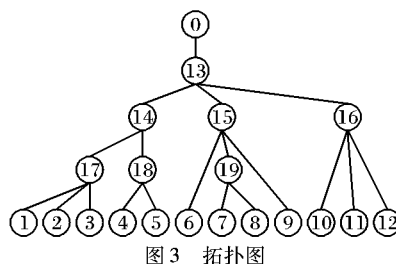


图3 拓扑图

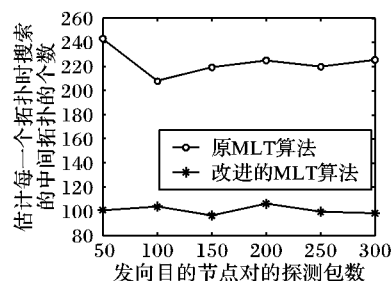


图4 NS-2 仿真每估计一个拓扑, 两种算法搜索到的中间拓扑数

表1 在不同发包数下两种算法的拓扑估计正确率

探测包个数	原 MLT 算法/%	改进的 MLT 算法/%
50	73.3	73.3
100	83.3	83.3
150	90.0	86.7
200	73.3	73.3
250	80.0	80.0
300	76.7	76.7

4 结语

本文通过证明 MLT 算法的似然函数是单峰的且峰值为最大值, 改进了 MLT 算法。改进后的算法在保证拓扑估计准确率不变的情况下, 有效降低了计算复杂度。然而, 本文假设所有内部节点都不协作, 但是在现实网络中会有部分内部节点协作, 因此我们的下一步工作是将可协作节点的协作信息加入到网络拓扑估计中, 以进一步降低拓扑估计算法的计算复杂度。

(下转第 238 页)

置为8、16、32、64对系统BER数值进行仿真实验。得到不同强度混沌信号扩频时,系统BER如图8所示。

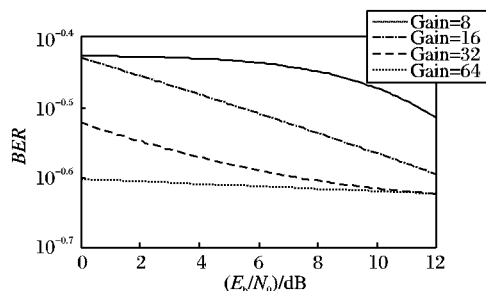


图8 不同强度混沌序列扩频时系统BER曲线

3.6 多径信号干扰下的BER仿真设计

灾害现场由于建筑物倒塌、山体滑坡或大气严重恶化等原因,信号大多通过多条路径进行传输,在接收端其合成信号与原信号存在较大失真,从而影响了应急通信的传输质量。为了测试本系统对多径效应的抑制能力,现对图1中信道模型进行修改,只保留选用“AWGN Channel”模拟白噪声,“Random Number”模拟随机信号干扰,用不同时延和增益的三路信号叠加模拟多径效应。再先后以混沌序列、PN序列、和Walsh序列作为扩频序列,得到系统的BER仿真曲线如图9所示。

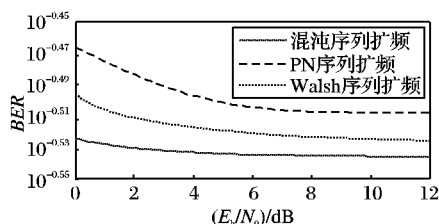


图9 多径信号干扰下系统BER曲线

4 仿真数据分析

从图4、图5可以看出,大功率脉冲干扰和大功率邻频干扰下采用混沌序列扩频时的系统BER整体下要比其他序列扩频时低得多,而且随着信道条件的改善而急速下降,体现出混沌序列扩频系统对大功率脉冲干扰和大功率邻频干扰的抑

制能力;从图6可以看出,4用户系统的BER值在相同信道条件下要比2、3用户系统低,体现出了在一定范围内本系统具有抗拥塞和邻频干扰的能力;从图7可以看出,虽然采用各种调制方式时系统BER均随着信道条件的改善而下降,但在强噪声干扰(信噪比低于7 dB)时,8PSK的抗噪声干扰能力最强;从图8可以看出,Gain为64的混沌信号进行扩频调制时,系统BER值整体最低;从图9可以看出,多径信号干扰作用下,混沌信号进行扩频调制时的系统BER值最低,体现出对多径干扰的良好抑制能力。

5 结语

深入研究并论述了应急通信中信道的噪声与干扰本质特性,构建了应急信道的仿真模型;根据应急通信的实际需要,提出了多用户应急通信系统的系统结构,并且针对典型信道噪声干扰环境和系统典型应用需要,进行了一系列多用户应急通信系统BER数值仿真(以4用户为例)。仿真实验数据证明,本系统具有良好的抗强噪声干扰能力,能够有效地提高应急通信的可靠性。

参考文献:

- [1] 焦方源,李佳,李薇.基于MANET+MDS的自适应灾害应急通信系统结构及其应用研究[J].电信快报,2010,464(2):17-20.
- [2] 兰燕红.城市交通应急系统的能力评价及预案研究[D].北京:北京交通大学,2008.
- [3] 崔维嘉,陈曦,陆杰青,等.多跳蜂窝网络代理选取问题研究[J].计算机工程与应用,2007,43(6):114-116,155.
- [4] 于银辉.混沌扩频通信技术研究及其应用[D].长春:吉林大学,2005.
- [5] 孙克辉,周家令,牟俊.多用户混沌扩频通信系统的设计与仿真[J].计算机仿真,2007,24(8):112-115.
- [6] 柳平,闫川,黄显高.改进的基于Logistic映射混沌扩频序列的产生方法[J].通信学报,2007,28(2):134-140.
- [7] LIU CHONGXIN, LIU TAO, LIU LING, et al. A new chaotic attractor[J]. Chaos Solitons and Fractals, 2004, 22(5): 1031-1038.
- [8] WANG FAQIANG, LIU CHONGXIN. A new criterion for chaos and hyperchaos synchronization using linear feedback control[J]. Physics Letters A, 2006, 360(2): 274-278.
- [9] SIAMWALLA R, SHARM R, KESHAV S. Discovering Internet topology [C/OL].// IEEE INFOCOM'99. Washington, DC: IEEE, 1999 [2010-03-02]. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1.3717&rep=rep1&type=pdf>.
- [10] CASTRO R, COATES M, LIANG G, et al. Network tomography: recent developments [J]. Statistical Science, 2004, 19(3): 499-517.
- [11] COATES M, HERO A O, NOWAK R, et al. Internet tomography [J]. IEEE Signal Processing Magazine, 2002, 19(3): 47-65.
- [12] RATNASAMY S, MCCANNE S. Inference of multicast routing trees and bottleneck bandwidths using end-to-end measurements [C].// INFOCOM'99: 18th Annual Joint Conference of the IEEE Computer and Communications Societies. Washington, DC: IEEE, 1999. 353-360.
- [13] DUFFIELD N G, HOROWITZ J, PRESTI F L, et al. Multicast topology inference from end-to-end measurements [EB/OL]. [2010-01-06]. <http://www.ce.uniroma2.it/~lopresti/Articoli/topologyITC.pdf>.
- [14] CASTRO R, COATE M, NOWAK R. Likelihood based hierarchical clustering [J]. IEEE Transactions on Signal Processing, 2004, 52(8): 2308-2321.
- [15] COATE M, CASTRO R, NOWAK R, et al. Maximum likelihood network topology identification from edge-based unicast measurements [C].// Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. New York: ACM, 2002: 11-20.
- [16] SHIN M-F, HERO A O, III. Hierarchical inference of unicast network topologies based on end-to-end measurements [J]. IEEE Transactions on Signal Processing, 2007, 55(5): 1708-1718.
- [17] 马知恩,王锦森.工科数学分析:下册[M].北京:高等教育出版社,2002:55-57.
- [18] 于斌,孙斌,温暖,等.NS2与网络模拟[M].北京:人民邮电出版社,2007:46-133.

(上接第214页)

参考文献: