基于全源 NT 的链路时延分布推断技术

段 琪[1], 王备战[2*], 蔡皖东[1]
(1. 西北工业大学计算机学院, 陕西 西安 710072; 2. 厦门大学软件学院, 福建 厦门 361005)

摘要: 互联网链路性能具有非对称性, 但是目前 NT 技术采用单源和多源测量, 只能推断从源节点到目的节点一个路径方向上的链路性能, 因此提出了推断功能更强的全源 NT 测量方法并解决了关键技术. 证明基于包对测量方法和交叉汇合测量方法, 严格全源网络结构的链路时延分布是可辨识的; 提出采用期望最大化(EM)算法的链路时延分布的极大似然估计方法; 最后通过模型仿真和网络仿真对推断方法的有效性进行了验证.

关键词: 全源 NT; 时延分布; 网络推断; EM 算法
中图分类号: TP 393 文献标志码: A 文章编号: 0438-0479(2011)04-0707-07

NT(network tomography)是近年来国际上提出的一种新的网络测量技术[1-2], 在网络拓扑固定的条件下, 根据网络外部的测量信息来分析和推断网络内部的性能以及网络拓扑, 如 AT&T 和马萨诸塞州立大学的 MINC 项目[3], 以及莱斯大学研究的单播 NT 项目为主等. 相对于内部测量, NT 技术具有不需要被测网络内部节点的配合, 且不依赖于特定的网络协议等优点; 但 NT 技术也有一些缺点, 其中主要一点就是推断能力有限.

首先, NT 技术研究成果中主要采用单个测量源节点(简称单源 NT), 而单源 NT 只能推断树型拓扑结构中的逻辑链路的性能, 如图 1(a). 链路时延(文中所指时延是排队延迟, queuing delay)分布推断的主要研究成果: 采用多播测量的包括 Lo Presti 等[4]提出的多项式启发式搜索算法, 该算法推断精度较低; Liang 等[5]提出的伪斯然估计算法和 Lawrence 等[6]的提出期望最大化(EM)算法, 来降低极大似然估计的复杂度且保留了较好的推断精度; Arya 等[7]研究了时延的时间相关性. 采用单播测量的包括 Coates 等利用单播包对的测量方法来估计链路延迟; Shih 等[8]使用点分布和高斯分布的混合模型描述时延分布, 并利用累积分布函数来推断.

其次, 多源测量采用多个源节点对多个目的节点的端到端测量(简称多源NT), 而多源NT技术研究

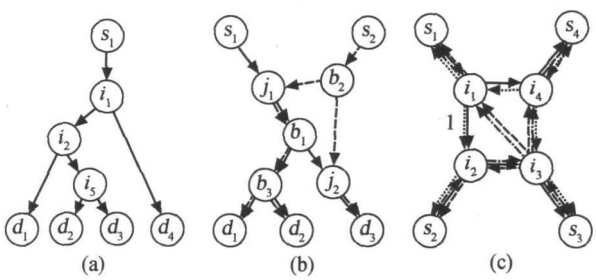

图 1 网络测量结构示意图
Fig. 1 The sketch map of the network measurements structure

成果主要集中在网络拓扑推断上. 如图 1(b). 设源节点和目的节点的个数分别为 m、n, 则测量覆盖的网络结构也被称为 m-by-n 网络结构. Bestavros[9] 等指出如果解决 2-by-1 的链路性能推断, 则可以推断出多源拓扑, 但没有给出解决方法. Coates 等[10] 和 Rabbat 等[11] 避免了 2-by-1 的链路性能推断难题, 采用区分共享 2-by-2 结构和非共享 2-by-2 结构来合并多个树型逻辑拓扑的方式推断多源拓扑, 但只能推断部分网络拓扑. 在链路性能推断问题上, Bu[12] 研究了多树网络拓扑的丢包率的推断的可识别性和出现在多个树中的链路丢包率的综合推断问题.

Internet 的网络结构是网状的, 且链路性能具有非对称性, 采用单源或多源测量方式, 只能推断从源节点到目的节点路径方向的性能, 而无法推断从目的节点到源节点路径方向的性能. 从推断功能上, 目前的 NT 技术没有完全解决 Internet 的链路性能推断问题. 因此, 本文提出全源测量模式, 即每个部署节点都向其他部署节点发送测量包(简称全源 NT), 如图 1(c). 可以看出多源 NT 技术是全源 NT 技术的特例,

收稿日期: 2010-05-13
基金项目: 国家高技术研究发展计划(863)项目(2009AA01Z424); 教育部博士点基金项目(200806990030)
* 通信作者: wangbz@xmu.edu.cn

所以以前研究的成果不适用于全源 NT 技术.

全源 NT 技术相对于单源 NT 和多源 NT 技术具有如下优点: 首先, 在相同数量的部署节点情况下, 全源测量不仅可以覆盖更多的链路, 还推导出最多的链路性能信息. 其次, 由于网络测量的费用主要是花费在部署节点上, 全源测量具有最好的经济性; 更主要的是, 全源测量可以推断链路上行和下行两个方向上的性能, 解决网络性能的非对称问题. 目前尚未见全源 NT 拓扑推断和链路性能推断研究文献. 因此研究全源 NT 技术具有重要研究意义和应用价值. 本文重点研究全源 NT 链路时延分布推断技术.

全源 NT 链路时延分布推断首先要解决如下 4 个问题: “如何进行网络测量?”, “理论上证明能否推断?”, “采用什么推断算法?”以及“推断算法的效率如何?”. 针对以上 4 个关键技术本文提出探测包的时延之间具有相关性的 2-by-1 网络结构多源测量方法, 并证明基于此测量方法和包对测量方法, 严格全源网络链路时延分布是可辨识的; 提出采用 EM 算法的链路时延分布的极大似然估计方法, 最后通过模型仿真和网络仿真进行了验证.

1 网络模型

1.1 拓扑模型

用 $G(V, L)$ 来描述全源网络拓扑, 其中 V 是节点集合, L 是连接节点的链路集合. 节点集合 S 和 I 分别表示源节点集合(目的节点集合) 和中间节点集合, $V = S \cup I$. 其中, 中间节点的入度和出度大于等于 1, 且不能同时为 1. 入度大于 1 的中间节点叫汇合节点, 出度大于 1 的中间节点叫分叉节点. 节点 i 到节点 j 的路径用 $P[i,j]$ 表示, 简记为 p_{ij} (或 $p_{i,j}$) . $p_{k,i}$ 和 $p_{k,j}$ 的共享路径用 $p_{k;i,j}$ 表示, 分叉节点用 $b(k;i,j)$ 表示; $p_{i,k}$ 和 $p_{j,k}$ 的共享路径用 $p_{i,j;k}$ 表示, 汇合节点用 $j(i,j;k)$ 表示, 由源节点 k 和目的节点 i, j 组成的 1-by-2 子网记为 $G^k_{i,j}$; 由源节点 i, j 和目的节点 k 组成的 2-by-1 子网记为 $G^{i,j}_k$.

1.2 延时模型

链路 l 上的时延用随机变量 $X(l)$ 表示. 离散化 $X(l)$ 到集合 $Q = \{0, q, 2q, \cdots, Bq, \infty\}$, 其中 q 表示离散化的粒度. $X(l) = iq(i \in [1, B])$ 表示链路的时延属于范围 $[iq - q/2, iq + q/2]$, 如果 $X(l) = \infty$, 就表示报文在链路 l 的时延大于 $Bq + 1/2q$ 或者发生丢失, 如果 $X(l) = \infty$, 就表示报文在链路 l 的时延大于等于 0, 小于 $1/2q$. 路径 p 上的时延表示为 $X(p)$, $X(p) \in \{0, q, 2q, \cdots, |p| \times Bq, \infty\}$. 令 $\alpha_p(i) = P\{X(p) = iq\}$, 记 $\alpha_p = (\alpha_p(1), \alpha_p(2), \cdots, \alpha_p(|p| \times B), \alpha_p(\infty))$, $\alpha = (\alpha)_{l \in L}$.

2 测量方法

在单源 NT 技术中, 单播测量的方法是采用发送单播包对, 对各个 1-by-2 子网, 依次进行测量, 如图 2(a) 所示. 包对中两个报文间隔很小, 可以近似认为在公共路径 p_{14} 上两个报文的时延相等. 利用此假设, 文献[7] 证明 1-by-N 网络链路时延分布是可辨识的.

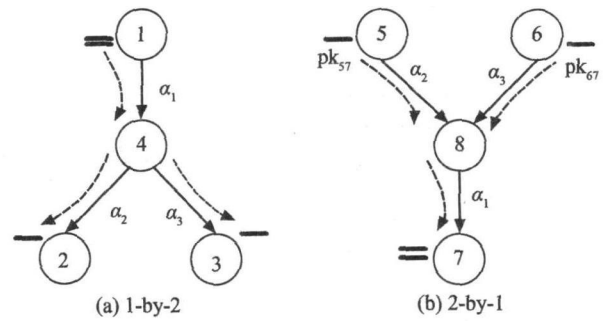

图 2 1-by-2 子网和 2-by-1 子网的测量方法

Fig. 2 Measurements methods of the 1-by-2 network and the 2-by-1 network

针对 2-by-1 子网链路时延推断问题, 采用交叉汇合测量方法, 2-by-1 子网是可辨识的. 如图 2(b) 所示的 2-by-1 子网, 源节点 5 和 6 分别发送探测包 pk_{57} 和 pk_{67}, 其中 $pk_{57}^{(m)}$ 和 $pk_{67}^{(m)}$ 表示第 m 个探测包, 由于 pk_{57} 和 pk_{67} 在路径 pk_{58} 和 pk_{68} 上的时延是随机的, 无法保证 $pk_{57}^{(m)}$ 和 $pk_{67}^{(m)}$ 同时到达汇合点而形成包对. 设 5 和 6 发送到 7 的探测报文的发送时间和接收时间分别为 $st_{57}^{(m)}$, $st_{67}^{(m)}$ 和 $rt_{57}^{(m)}$, $rt_{67}^{(m)}$, 固定时延分别为 δ_{57} 和 δ_{67}, 则 $X_{p_{57}}^{(m)} = rt_{57}^{(m)} - st_{57}^{(m)} - \delta_{57}$, $X_{p_{67}}^{(m)} = rt_{67}^{(m)} - st_{57}^{(m)} - \delta_{67}$, 令 $\delta_{5,6;7} = \delta_{57} - \delta_{67}$. 依次调整发送节点 5 的发送时间, 使 $st_{67}^{(m)} = st_{57}^{(m)} + nq + \delta_{5,6;7}$ 其中 n 分别为 $-(|p_{67}| \times B + 1/2), \cdots, 0, 1/2, 1 + 1/2, \cdots, |p_{57}| \times B + 1/2$, 对于不同 n 的取值都进行 M 次测量, 将这种测量方法称为交叉汇合测量法. 定理 1 表明该测量方法可以解决 2-by-1 子网链路时延分布推断问题.

引理 路径 $p_{i,j}$ 包含路径 $p_{i,k}$ 和 $p_{k,j}$, 已知其中两个路径的时延分布, 则可以推导另外一个路径的时延分布.

证明 因为 $X(p_{i,j}) = X(p_{i,k}) + X(p_{k,j})$, 且 $\alpha(p_{i,k})$ 和 $\alpha(p_{k,j})$ 相互独立, 根据卷积定理已知其中两个路径

的时延分布, 可以求另外一个路径的时延分布.

定理 **1**　采用交叉混合测量法, 对于任意的 2-by-1 子网链路时延是可辨识的.

证明　设发送节点分别是 s_i、s_j, 目的节点是 d_k, 中间汇合节点为 j.

1) 当 $n=\lfloor P[s_i,j]\rfloor\times B+1/2$, 即去除固定时延的差的因素后, 从 s_j 发送到 a_k 的报文 $\mathrm{pk}_{i,k}$ 比 s_i 发送到 d_k 的报文 $\mathrm{pk}_{j,k}$ 在双方都没有经历排队延迟的情况下, 到达汇合节点 j 晚 $(\lfloor P[s_i,j]\rfloor\times B)\,q$, 所以当 $X(P[s_j,j])=0$, 当且仅当 $X(P[s_j,j])=\infty$, 才使 $\mathrm{pk}_{i,k}$ 比 $\mathrm{pk}_{j,k}$ 晚到达点 j, 即 $rt_{i,k}>rt_{j,k}$. 因此, $P(rt_{i,k}>rt_{j,k}\mid X(P[s_j,k])=0, n=\lfloor P[s_i,j]\rfloor\times B+1/2)=\alpha_{i,k}(\infty)$. 进行一般化则有

$$P(rt_{i,k}>rt_{j,k}\mid X(P[s_j,d_k])=0;n=m+1/2,\ m\in[0,\lfloor P[s_i,j]\rfloor\times B)=\alpha_{s_i,j}(\infty)+\sum_{r=m+1}^{P[s_i,k]\rfloor\times B}\alpha_{i,j}(r). \tag{1}$$

所以可以获取 $X(P[s_i,j])$ 的分布率为:

$$\alpha_{i,j}(x)=\begin{cases}P(rt_{i,k}>rt_{j,k}\mid X(P[s_j,d_k])=0;\\ \quad n=x+1/2),\ \text{当 } x=\infty\\ P(rt_{i,k}>rt_{j,k}\mid X(P[s_j,d_k])=0;\\ \quad n=x-1/2)-P(rt_{i,k}>rt_{j,k}\mid X(P\\ \quad [s_j,d_k])=0;n=x+1/2),\\ \quad \text{当 } x\in[1,\lfloor P[s_i,j]\rfloor\times B]\\ P(rt_{i,k}>rt_{j,k}\mid X(P[s_j,d_k])=0;n=0)\\ \quad -P(rt_{i,k}>rt_{j,k}\mid X(P[s_j,d_k])=0;\\ \quad n=1/2),\ \text{当 } x=0\end{cases} \tag{2}$$

2) 同理可以获得 $X(P[s_j,j])$ 的分布.

3) 根据测量到的 $X(P[s_i,d_k])$ 的分布和 $X(P[s_i,j])$ 的分布, 利用引理可以推导出 $X(P[j,d_k])$ 的分布. 综上, 定理 1 得证.

3　全源网络链路性能的可辨识性

定理 **2**　1-by-N 网络链路性能是可以辨识的充要条件是 1-by-2 子网链路性能是可辨识的.

文献[6] 给出了简要的证明.

定理 **3**　M-by-1 网络链路性能是可以辨识的充要条件是 2-by-1 子网链路性能是可辨识的.

由于 M-by-1 网络和 1-by-N 具有对称性, 根据定理 1 利用和定理 2 类似的证明方法可以证明.

定理 **4**　M-by-N 网络链路性能是可以辨识的的充要条件是 1-by-2 子网链路性能和 2-by-1 子网链路性能是可辨识的.

证明　1) 证明充分性. 设任意一条路径 $P[s,d]$ 上的中间结点(包括分叉结点和汇合结点) 为 i_k, $k=1,2,\cdots,\lfloor P[s,d]\rfloor-1$, 根据文献[6] 和定理1可以得到 $\alpha(P[s,i_k])$, $\forall k=1,2,\cdots,\lfloor P[s,d]\rfloor-1$, 根据引理, 利用 $\alpha(P[s,i_k])$ 和 $\alpha(P[s,i_{k+1}])$ 可以得到路径上任意一个链路的性能 $\alpha(P[i_k,i_{k+1}])$, 所以整个 M-by-N 网络链路性能是可辨识的.

2) 证明必要性. 采用举反例法, 由于 1-by-2 网络和 2-by-1 网络也属于 M-by-N 网络, 所以如果 1-by-2 子网和 2-by-1 子网不能辨识, 则 M-by-N 网络也不能辨识.

定理 **5**　全源网络结构链路性能是可以辨识的充要条件是: 1) 1-by-2 子网链路性能和 2-by-1 子网链路性能是可辨识的; 2) 对于任意链路 $P[i_a,i_k]$, 存在链路 $P[i_b,i_k]$ 和 $P[i_k,i_{k+1}]$ 使得 $P[i_a,i_k]\subset P[s_z,s_x]\,(s_z,s_x\in S)$、$P[i_b,i_k]\subset P[s_y,s_n]\,(s_y,s_n\in S)$ 和 $P[i_k,i_{k+1}]\subset P[s_z,s_x]\cap P[s_y,s_n]$; 或者存在链路 $P[i_k,i_{k+1}]$ 和 $P[i_k,i_{k+2}]$ 使得 $P[i_k,i_{k+1}]\subset P[s_z,s_x]\,(s_z,s_x\in S)$, $P[i_k,i_{k+2}]\subset P[s_y,s_n]\,(s_y,s_n\in S)$ 和 $P[i_a,i_k]\subset P[s_z,s_x]\,(s_y,s_n\in S)$.

证明　1) 证明充分性, 采用递归法证明. 全源网络结构 $G(V,L)$ 中链路 $P[i_a,i_k]$, 如果存在链路 $P[i_b,i_k]$ 和 $P[i_k,i_{k+1}]$ 使得 $P[i_a,i_k]\subset P[s_z,s_x]\,(s_z,s_x\in S)$、$P[i_b,i_k]\subset P[s_y,s_n]\,(s_u,s_n\in S)$ 和 $P[i_k,i_{k+1}]\subset P[s_z,s_x]\cap P[s_y,s_n]$, 由于 2-by-1 子网是可辨识的, 所以 $G(V,L)$ 的可辨识性等价于 $G(V,L-\{P[i_a,i_k],P[i_k,i_{k+1}]\}+\{P[i_a,i_{k+1}],P[i_b,i_{k+1}]\})$. 同理, 对于链路 $P[i_a,i_k]$, 如果存在链路 $P[i_k,i_{k+1}]$ 和 $P[i_k,i_{k+2}]$ 使得 $P[i_k,i_{k+1}]\subset P[s_z,s_x]\,(s_z,s_x\in S)$, $P[i_k,i_{k+2}]\subset P[s_y,s_n]\,(s_y,s_n\in S)$ 和 $P[i_a,i_k]\subset P[s_z,s_x]\cap P[s_y,s_n]$, 由于 1-by-2 子网是可辨识的, 所以 $G(V,L)$ 的可辨识性等价于 $G(V,L-\{P[i_a,i_k],P[i_k,i_{k+1}],P[i_k,i_{k+2}]\}+\{P[i_a,i_{k+1}],P[i_a,i_{k+2}]\})$. 依次类推, 通过减少链路和中间节点的方法, $G(V,L)$ 等价于 $G(S,\{P[s_1,s_2]\mid s_1,s_2\in S\})$, 而 $G(S\mid,\{P[s_1,s_2]\mid s_1,s_2\in S\})$ 显然是可辨识的, 所以 $G(V,L)$ 是可辨识的.

2) 证明必要性. 条件 1) 的必要性证明, 采用举反例法. 由于 1-by-2 网络和 2-by-1 网络也属于全源网络, 所以如果 1-by-2 子网和 2-by-1 子网不能辨识, 则全源也不能辨识. 条件 2) 的必要性证明, 也采用举反例法, 如图 3(a), 链路 $P[s_1,i_1]$ 在节点处没有“分叉”, 也没有“汇合”, 通过 NT 测量和推断, 只能得到路径 $P[s_1,i_2]$

的性能, 而无法无偏估计 α_1 和 α_2, 链路性能是无法辨识的. 同样, 网络中也无法无偏估计 α_1' 和 α_{31}.

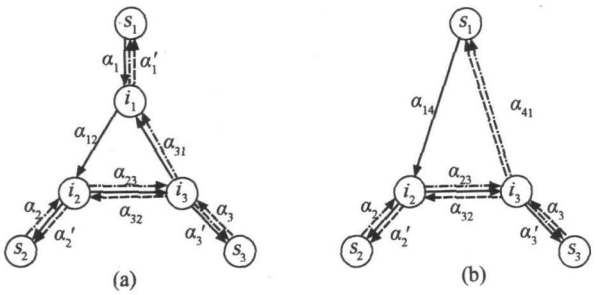

图 3 一种 3 个源节点的全源网络结构(a) 及与其对应的关联严格全源网络结构(b)

Fig. 3 One full source network structure with three souces (a) and the corresponding strict full source network structure of it (b)

定理 5 给出了全源网络结构可辨识的充要条件, 但是定理中条件 2) 在实际使用中并不好理解. 条件 2) 等价于任意一条链路在每个中间节点或是“分叉”, 或是“汇合”. 为此, 引入严格全源网络结构的概念.

用 $\overline{G}(V, L)$ 来描述严格全源网络拓扑, 首先, 严格全源网络结构是全源网络结构, 即 $\overline{G}(V, L) \subset G(V, L)$; 其次对于任意链路 $P[i_a, i_b]$, 存在链路 $P[i_b, i_k]$ 和 $P[i_k, i_{k+1}]$ 使得 $P[i_a, i_k] \subset P[s_z, s_x]$ $(s_z, s_x \in S)$、$P[i_b, i_k] \subset P[s_y, s_n]$ $(s_y, s_n \in S)$ 和 $P[i_k, i_{k+1}] \subset P[s_z, s_x] \cap P[s_y, s_n]$; 或者存在链路 $P[i_k, i_{k+1}]$ 和 $P[i_k, i_{k+2}]$ 使得 $P[i_k, i_{k+1}] \subset P[s_z, s_x]$ $(s_z, s_x \in S)$, 和 $P[i_a, i_k'] \subset P[s_z, s_x] \cap P[s_y, s_n]$.

根据定义, 严格全源网络结构是可以辨识的充要条件是, 1-by-2 子网和 2-by-1 子网是可辨识的. 一个非严格全源网络结构 $g(V, L)$, 通过最少次数“合并”链路, 得到的严格全源网络结构称为 $G(V, L)$ 的关联严格全源网络结构, 记作 $g(V, L)$. 例如, 图 3(a) 的关联严格全源网络结构为图 3(b).

4 测量子图的选取

根据定理 1 和文献[6], 可知严格全源网络结构的链路时延分布率是可以辨识的. 但对严格全源网络中每个 1-by-2 子网和 2-by-1 子网都进行测量, 将使测量流量和推断算法复杂度随被测网络节点的增加呈指数倍增加.

$$|G| = C_{N-1}^2 \times N \times 2 = P_N^3. \tag{3}$$

定理 6 给出了使严格全源网络链路时延分布率是可以辨识的, 测量子图选择的充分条件.

定理 6 任意的严格全源网络 $\overline{g}(V, L)$, 测量子网集合记为 G, 链路时延分布是可辨识的充分必要条件是: 1) 对于任意一个分叉节点 b, 至少存在一个分叉节点为 b 的 1-by-2 子网属于; 2) 对于任意一个汇合节点 j, 至少存在一个汇合节点为 j 的 2-by-1 子网属于 G; 3) 对于任意一个发送节点 s 至少存在一个发送节点为 s 的子网属于 G; 4) 对于任意一个目的节点 d 至少存在一个目的节点 d 为的子网属于 G. 充分性的证明类似定理 5, 此处不再证明.

在满足充分性的条件下, 可以选取尽量少的测量子网, 例如对于图 1(c) , 所有的测量子网有 24 个. 由于 $G_{s_2, s_4}^{s_3}$ 和 $G_{s_3}^{s_2, s_4}$ 及 $G_{s_2, s_3}^{s_4}$ 和 $G_{s_4}^{s_2, s_3}$ 是结构上对称的, 去除具有对称性的, 还有 22 个测量子网. 但是根据定理 6, 最小的测量子网可以采用的 $G = \{G_{s_2, s_3}^{s_1}, G_{s_2, s_4}^{s_1}, G_{s_3, s_1}^{s_2}, G_{s_2, s_4}^{s_3}, G_{s_2, s_3}^{s_4}, G_{s_2, s_1}^{s_4}\}$, 大大减少了测量规模及推断算法的复杂度.

5 链路时延分布推断算法

本文采用极大似然估计来推断全源网络结构链路时延分布. 对于 1-by-2 子网 $G_{i,j}^k \in G$, 采用包对进行测量, 测量值记为 $\vec{y}_{k;i,j} = (y_{ki}, y_{kj})$. 对于 2-by-1 子网 $G_k^{i,j} \in G$ 采用交叉汇合测量方法, 在 $st_{j,k} = st_{i,k} + nq + \varepsilon_{,k} - \varepsilon_{,k}$ 条件下, 引入变量 $\beta_{i,j;k}^n$ 则

$$\beta_{i,j;k}^n = \begin{cases} 1 & \text{当 } st_{j,k} = st_{i,k} + nq + \varepsilon_{,k} - \varepsilon_{,k}, \text{且 } rt_{i,k} > rt_{j,k} \\ 0 & \text{当 } st_{j,k} = st_{i,k} + nq + \varepsilon_{,k} - \varepsilon_{,k}, \text{且 } rt_{i,k} < rt_{j,k} \end{cases}$$

$N(y)$ 表示测量值为特定值 y(y 类型包括 $y_{i,k}$, $y_{k;i,j}$, $\beta_{i,j,k}^m$) 的次数, 记 $\Phi_{,j;k} = \{-(|p_{jk}| \times B + 1/2), \cdots, 0, 1/2, 1 + 1/2, \cdots, |p_{ik}| \times B + 1/2\}$. 测量值概率记为 $g(y; \alpha)$, 则所有的观测值对数似然函数为

$$\begin{aligned} \Lambda(Y; \alpha) &= \log g(Y; \alpha) = \\ &\sum_{\forall G_{i,j}^k \in G} \sum_{y_{k;i,j} \in Y_{k;i,j}} N(y_{k;i,j}) \log g(y_{k;i,j}; \alpha) + \\ &\sum_{\forall G_k^{i,j} \in G} \Big(\sum_{y_{k,j} \in Y_{k,j}} N(y_{k,i}) \log g(y_{k,i}; \alpha) + \\ &\sum_{y_{k,j} \in Y_{k,j}} N(y_{k,j}) \log g(y_{k,j}; \alpha) \\ &\sum_{\forall n \in \Phi_{i,j}} \sum_{\beta \in \{0,1\}} N(\beta) \log g(\beta; \alpha) \Big). \end{aligned} \tag{4}$$

利用 MLE 的方法求 α 的估计值为

$$\hat{\alpha} = \arg\max_\alpha \Lambda(Y; \alpha). \tag{5}$$

为方便计算, 采用 EM 算法求解. 首先引入不可见数据, 即测量包在各个链路的时延, 对于 $\forall G_{i,j}^k \in G$, 用 $D_{k;i,j}$ 表示探测包每条链路所经历的时延集合; 对于

$\forall G_k^{i,j} \in G$ 用 $D_{i,j;k}$ 表示探测包每条链路所经历的时延集合. 为方便表示引入 $D = \{D_{i,j;k}, D_{k;i,j};\ \forall G_{i,j}^k \in G, \forall G_k^{i,j} \in G\}$. 因此引入了 D 数据的对数似然函数, 可以表示为

$$\mathcal{L}(Y, D; \alpha) = \log g(Y, D; \alpha) = \log g(Y \| D; \alpha) + \log g(D; \alpha). \tag{6}$$

在已知 D 的条件下, 通过测量必然获得相应的观测值的 Y, 即 $g(Y \| D; \alpha) = 1$, $\log g(Y \| D; \alpha) = 0$. 令 $N_l(d)$ 表示链路 l 上的时延等于 d 的报文数量. 因此,

$$\mathcal{L}(Y, D; \alpha) = \log g(D; \alpha) = \sum_{\forall G_{i,j}^k \in G} \log g(D_{k;i,j}; \alpha) + \sum_{\forall G_k^{i,j} \in G} \log g(D_{i,j;k}; \alpha) = \sum_{l \in L(t)} \sum_{d \in Q} N_l(d) \log \alpha_l(d). \tag{7}$$

$\alpha_l(d)$ 的 MLE 估计值 $\hat{\alpha}_l(d)$ 可以通过上式进行求导获得, 因此

$$\hat{\alpha}_l(d) = \frac{N_l(d)}{\sum_{d \in Q} N_l(d)}. \tag{8}$$

由于时延集合 D 和 $N_l(d)$ 是引入的不可见数据, 利用EM 算法进行求解. 该算法的核心思想是使用 $\alpha_l(d)$ 和 $n_l(d)$ 的前次估计值推断它们当前的估计值, 通过有限步的迭代来得到收敛的 $\hat{\alpha}_l(d)$ 值. 令 $\hat{\alpha}_l^{(e)}(d)$ 表示 $\alpha_l(d)$ 第 e 步的概率估计值. 其求解过程如下: 1) 合理选择所有链路的初始时延分布 $\hat{\alpha}_l^{(0)}$. 在缺乏 α_l 的先验信息的情况下可假设 $\hat{\alpha}_l^{(0)}$ 服从[0, 1]上的均匀分布. 2) 在已知完整的测量值集合 Y 和当前第 e 步估计值 $\hat{\alpha}_l^{(e)}$ 的条件下, 计算对数似然函数的条件期望 $\hat{\alpha}_l''$.

$$Q(\hat{\alpha}_l'', \hat{\alpha}_l^{(e)}) = E_{\hat{\alpha}_l^{(e)}}[\mathcal{L}(Y, D; \hat{\alpha}_l'') \mid Y] = \sum_{l \in L(t)} \sum_{d \in Q} \hat{N}_l(d) \log \hat{\alpha}_l''(d). \tag{9}$$

其中,

$$\hat{N}_l(d) = E_{\hat{\alpha}_l^{(e)}}[N_l(d) \mid Y] = \sum_{\forall G_{i,j}^k \in G} \sum_{y_{k;i,j} \in Y_{k;i,j}} N(y_{k;i,j}) \log g(X(l) = d \mid Y_{k;i,j} = y_{k;i,j}; \hat{\alpha}^{(e)}) + \sum_{\forall G_k^{i,j} \in G} \Big(\sum_{y_{k,i} \in Y_{k,i}} N(y_{k,i}) \log g(X(l) = d \mid Y_{k,i} = y_{k,i}; \hat{\alpha}^{(e)}) + \sum_{y_{k,j} \in Y_{k,j}} N(y_{k,j}) \log g(X(l) = d \mid Y_{k,j} = y_{k,j}; \hat{\alpha}^{(e)}) \Big) + \sum_{\forall n \in \Phi_{i,j;k}} \sum_{\beta \in B_{i,j;k}^n} \big(N(\beta) \times \log g(X(l) = d \mid \beta_{i,j;k}^n = \beta; \hat{\alpha}^{(e)}) \big). \tag{10}$$

3) 根据式(6~ 13) 估计的 $\hat{N}_l(d)$, 从而获得第 $e+1$ 步时延分布概率的估计值为

$$\alpha_l^{(e+1)}(d) = \arg\max_{\alpha_l''} F(\hat{\alpha}_l'', \hat{\alpha}_l^{(e)}) = \frac{\hat{N}_l(d)}{\sum_{d \in Q} \hat{N}_l(d)}. \tag{11}$$

4) 交替地使用式(2)、(3) 进行计算, 直到估计的延迟分布概率达到收敛状态.

6　仿　真

6.1　模型仿真

首先采用模型仿真验证算法的有效性, 模型仿真采用图 1(c) 所示网络结构. 实验采用测量子网集 $G = \{G_{s_2,s_3}^{s_1}, G_{s_2,s_4}^{s_1}, G_{s_3,s_1}^{s_2}, G_{s_2,s_4}^{s_3}, G_{s_2,s_3}^{s_4}, G_{s_2,s_1}^{s_4}\}$. 在 G 中, 每个 1-by-2 测量子网, 产生 50 000 测量值. 设置推断算法初始链路延迟分布为均匀分布, 门限 threshold = 0.001. 所有链路的时延具有相同离散分布或连续分布.

1) 链路时延分布采用离散分布: $\alpha(0) = 0.40$, $\alpha(1) = 0.30$, $\alpha(2) = 0.15$, $\alpha(3) = 0.10$, $\alpha(4) = 0.5$. 图 4 为图 1(c) 所示网络结构中链路 $P[i_1, i_2]$ 的时延分布 α_1 的收敛过程, 仿真结果表明, 对于离散的时延分布, 推断算法可以无偏差地推断出各个链路的时延分布率.

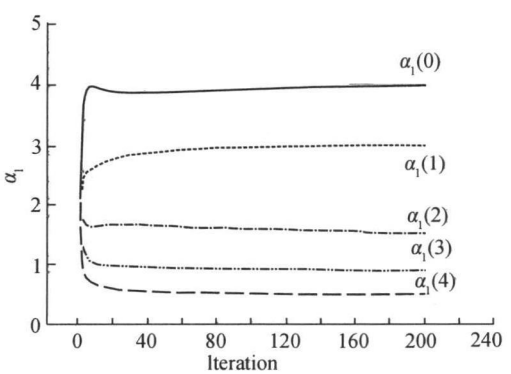

图 4　在离散分布下, $\alpha_{P[i_1, i_2]}$ 的收敛过程

Fig. 4　The constringency process of $\alpha_{P[i_1, i_2]}$ under the discrete distribution

2) 链路分布采用均值为 1 的截尾指数连续时延模型, 且 $q = 0.2$, 分别采用 $B = 5, 10, 15, 20$. 仿真结果如图 5 所示. 仿真结果随离散化精度的增加, 推断误差不断减少, 验证了算法的有效性, 同时推断算法的在相同计算机平台上运行时间分别为5, 14, 35, 52 s; 表明推断算法的时间复杂度随离散化精度的增加, 快速增加.

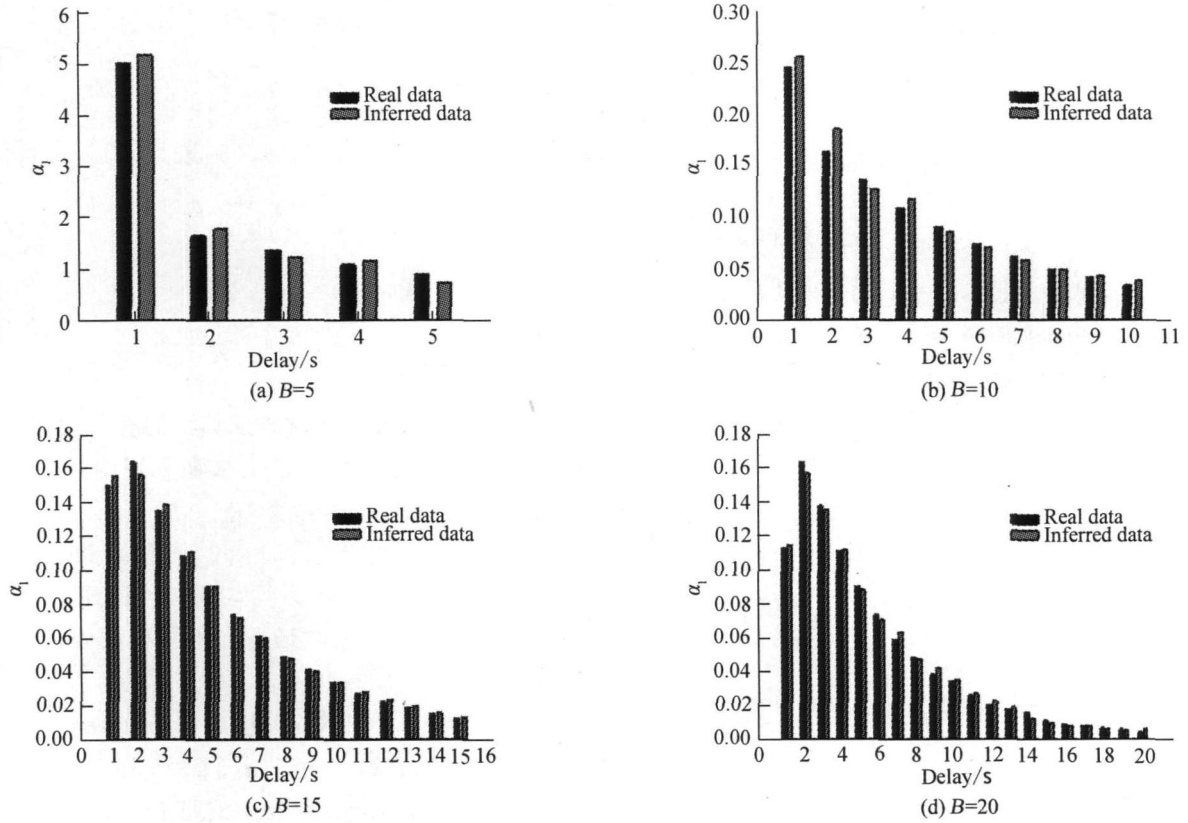

图 5 B 取不同值时真实分布和推断分布对照图

Fig. 5 The comparison diagram of real distribution and the inffered distribution when B take different values

6.2 网络仿真

采用网络仿真工具 ns-2 对图 3(b) 所示网络结构进行仿真. 设置每条逻辑链路上有 5 个物理链路, 物理链路带宽为 5~ 10 Mb/s, 每个物理链路背景流量为泊松流或有 3 个参数为 α= 1.9 的开关数据流构成的自相似流, 链路利用率为30%~ 70%, 数据包的大小为56 ~ 1 500 byte. 对于 1-by-2 测量子网, 探测包对两个间隔为 0.01 ms、周期为 0.2 s 的 CBR 数据流, 每个 1-by-2 测量子网产生 500 个测量值, 对于 2-by-1 测量子网, 两个源节点分别发送周期为 0.2 s 的 CBR 数据流, 且根据固定时延和调整源节点的发送时刻, 每个不同的 n 产生 200 测量值. 设置推断算法初始链路时延分布为均匀分布, q= 0.002 s, B= 10, 门限 threshold = 0.01. 网络结构运行 100 次, 得到的网络结构平均相对误差为 2.4%.

7 结 论

本文在的单源 NT 和多源 NT 技术的基础上, 提出了全源 NT 的概念, 并且可以证明采用“包对”和“交叉汇合”测量方法, 全源 NT 技术可以推断出严格全源网络结构的链路时延分布, 同时本文给出的测量子网的选取的充分条件, 可以减少测量子网集合. 最后通过仿真验证了全源 NT 技术是有效的, 并给出推断误差以及计算的时间复杂性. 本文的研究扩展了 NT 技术推断链路的能力和适用范围.

参考文献:

[1] Coates A, HeroIII A O, Nowak R, et al. Internet tomography[J]. IEEE Signal Processing Magazine, 2002, 19(3): 47-65.

[2] Castro R, Coates M, Liang G, et al. Internet tomography: recent developments[J]. Statistical Science, 2004, 52(3): 499-517.

[3] AT&T Labs-Research. Multicast-based inference of network-internal characteristics[EB/OL]. [2009-12-20]. http://www-net.cs.umass.edu/minc/.

[4] Lo Presti F, Duffield N, Horowitz J, et al. Multicast-based inference of network internal delay distributions[J]. IEEE/ACM Transactions on Networking, 2002, 10(6): 761-775.

[5] Liang G, Yu B. Maximum pseudo likelihood estimation in

network tomography[J]. IEEE Transactions on Signal Processing, 2003, 51(8): 2043-2053.

[6] Lawrence E, Michailidis G, Nair V. Network delay tomography using flexicast experiments[J]. Journal of Royal Statistical Society Series B, 2006, 68(5): 785-813.

[7] Arya V, Duffield N, Veitch D. Temporal delay tomography[C]//IEEE Infocom 2008. Phoenix: IEEE, 2008: 276-280.

[8] Meng F S, HeroIII A O. Unicast-based inference of network link delay distributions with finite mixture models [J]. IEEE Transactions on Signal Processing, 2003, 51 (8): 2219-2228.

[9] Bestavros A, Byers J W, Harfoush K A. Inference and labeling of metrie-induced network topologies[J]. IEEE Transactions on Parallel and Distributed Systems, 2005, 16(11): 1053-1065.

[10] Coates M, Rabbat M, Nowak R. Merging logical topologies using end-to-end measurements[C]//ACM SIGCOMM Conference on Internet Measurement. Miami: ACM, 2003: 192-203.

[11] Rabbat M, Coates M, Nowak R. Multiple source internet tomography[J]. IEEE Journal on Selected Areas in Communications, 2006, 24(12): 2221-2234.

[12] Bu T, Duffield N, Lo Presti F, et al. Network tomography on general topologies[J]. ACM Sigmetrics Performance Evaluation Review-Measurement and Modeling of Computer Systems, 2002, 30(1): 21-30.

Research on Delay Inference Technology Based on the Full Source Network Tomography

DUAN Qi[1], WANG Bei-zhan[2*], CAI Wan-dong[1]

(1. School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China;
2. Software School, Xiamen University, Xiamen 361005, China)

Abstract: The link performance of Internet is asymmetrical. At present, only the one-way link performance from source to destination can be inferred by the network tomography technology based on single source and multiple source measurements. Therefore, the full source measurement pattern is proposed in this paper. The link delay distribution of strict full source NT network structure is proved identifiable by using the cross joining probe method and the back-to-back pair probe method. Furthermore, the sufficient condition of the measurement sub-network selection which makes the link identifiable is proposed. The measurement traffic and the computational complexity could be reduced observably with the minimum measurement sub-network set. At last, the maximum likelihood estimation of link delay distribution computed by the EM algorithm is derived and the effectiveness is validated by the model simulation and network simulation results.

Key words: full source network tomography; delay distribution; network inference; EM algorithm