ISSN 1000-1239/CN 11-1777/TP
50(3): 509-516, 2013

因特网时延空间中 TIV 与接入时延的研究

王占丰　陈　鸣　邢长友　白华利　魏祥麟
(解放军理工大学指挥信息系统学院　南京　210007)
(hehengw@hotmail.com)

TIV and Access Delay in the Internet Delay Space

Wang Zhanfeng, Chen Ming, Xing Changyou, Bai Huali, and Wei Xianglin
(*Institute of Command Automation, PLA University of Science & Technology, Nanjing* 210007)

Abstract　Many researches demonstrate that triangle inequality violation (TIV) is a universal phenomenon in the Internet delay space, which is the result of routing inefficiency. However, there are still no investigations on the TIV's position and its relationship with the access delay. Firstly, the Internet is divided into two parts: access network and core network, by different definitions. A delay model of the Internet delay space is proposed to analyze the TIV's position. Based on this model, it is found that TIV appears in the core of the Internet rather than the access networks and the access delay can reduce the TIV number and alleviate the TIV's severity. Then, a series of experiments are carried out on the PlanetLab to measure the end-to-end delay matrix and its corresponding topology. Afterwards, a TIV searching algorithm ScoutTIV is designed to count the TIV number in the measurement dataset by different definitions of edges between the access network and core network. In the experiments, the datasets are divided into three small datasets based on the country attributes of hosts and one random dataset. The experimental results on different datasets accord with our conclusion, which can help network coordinate systems to achieve better prediction accuracy.

Key words　delay space; triangle inequality violation (TIV); network coordinate system; network measurement; access delay

摘　要　大量网络测量研究证实了违反三角不等式(TIV)是因特网时延空间存在的一种普遍现象,是影响网络坐标系统准确性的重要原因之一.通过将因特网分为接入网和核心网两部分,引入了时延空间模型来分析接入时延对于 TIV 的影响.理论分析表明 TIV 产生于网络的核心,接入时延可以使得在端到端路径中观察到的 TIV 数目会减少,并减轻 TIV 的严重程度.然后,在 PlanetLab 测试平台设计了一组网络测量实验,来测量端到端的时延矩阵和相应的拓扑信息.之后,设计了 ScoutTIV 算法来统计时延数据集中的 TIV 比例.在实验中,根据主机的 IP 属性将其分为 3 个子集,并生成了 1 个随机数据集来进行分析.在所有子集上的实验结果与理论分析结论一致,为网络坐标系统进一步提高预测精度提供了重要依据.

关键词　时延空间;违反三角不等式;网络坐标系统;网络测量;接入时延

中图法分类号　TP393

收稿日期:2011-01-20;**修回日期**:2012-06-29

基金项目:国家"九七三"重点基础研究发展计划基金项目(2012CB315806);国家自然科学基金项目(61070173,61103225);江苏省自然科学基金项目(2010133)

分组时延是IP网络最重要的性能指标之一.发展迅猛的新型多媒体网络应用如VoIP,IPTV和在线游戏等,其服务质量与网络时延密切相关,人们迫切希望理解IP网络的时延分布规律.然而,由于IP网络是一种尽力而为的统计复用网络,分组时延因网络资源配置和使用情况的不同而发生变化,建立准确的网络时延模型是一个巨大的挑战.

近年来,人们先后提出了多种网络坐标系统或者时延估计系统来为因特网时延空间建模[1-5].借助于网络坐标系统,人们希望通过少量的网络测量就可以准确地预测网络中任意两个节点之间的端到端时延.网络坐标系统的原理是将因特网时延空间视为一个度量空间,将其映射到一个网络坐标系统中.这样,通过计算节点的坐标就可以得到两节点之间的时延.因此,时延空间必须满足度量空间的3个基本约束条件:非负性、对称性和三角形不等式.显然,网络时延能够满足非负性的要求,但它却不能满足对称性和三角不等式的约束.由于因特网时延空间通常研究的是节点间的往返时延(round trip time, RTT),这就使其能够满足对称性的约束.网络时延违反三角不等式(triangle inequality violation, TIV,即三角形两边时延之和不大于第3边时延)的现象被认为是导致时延坐标系统不够准确的主要原因之一[1,4,6].

TIV现象已被许多网络测量研究所证实,且普遍认为TIV的生成是由于非高效路由策略(routing inefficiency)和网络结构而导致的[4,6-7].伴随着光纤技术不断进步,分组传输成本持续下降,在商业利益的驱动下ISP和内容提供商(如Google, Microsoft等)不断地增加光纤链路使得因特网的核心链路日益丰富[8],带来了核心网性能的提升.与之相比,接入网络的性能相对较差,引入了较大的接入时延,往往成为性能瓶颈.文献[9]指出接入时延使得网络在局部TIV比例较高.但是现有研究成果却从未关注接入时延对于TIV的影响,而多数的网络坐标系统使用者端用户往往具有较大的接入时延.本文通过对TIV建模,分析了接入时延对于TIV的影响,经理论分析发现接入时延可以使得在端到端路径中观察到的TIV数目有所减少,并减轻了TIV的严重程度.

1 相关工作

在1994年,Hotz[10]在其博士论文中对时延空间中的三角不等式约束进行了研究.在首个网络距离坐标系统GNP被提出后[11],人们试图将网络节点嵌入到不同的度量空间中.由于在度量空间中理论上是不能出现TIV现象的,因此网络时延空间中出现的TIV现象就成为困扰人们的难题.

有关TIV的研究可以分为两个方向.一是选择合适的网络坐标系统以减小TIV对空间嵌入的影响.文献[2-3]试图通过改变嵌入维数或选择一个凸或凹的嵌入空间,来减少TIV的影响.然而这些坐标系统都以出现较少的TIV或者TIV的程度较轻为前提,它们都无法忽视TIV的影响.

TIV的另一个研究方向是分析其成因及其特点[6-7,12-13].文献[6]分析了TIV严重性对于网络坐标系统的影响,发现TIV严重的数据集误差较大,并认为TIV与ISP的选路协议及策略和网络结构有关.文献[7]通过收集GREN研究网络的数据,发现TIV不是因网络测量偏差而导致的假象,而是持久地、广泛地存在于因特网中的普遍现象,认为TIV生成是因特网选路协议所致.多个研究也印证了TIV的广泛存在,且端到端时延的TIV比例约在10%到20%之间[9].文献[12]研究了BGP协议和TIV之间的关系,发现TIV所显现出的捷径(shortcut path)中有25%与BGP策略一致,而BGP协议却倾向于选择时延更长的路径.文献[12]进一步发现这些捷径的中继路由节点往往靠近源主机或者目的主机.文献[13]详细研究了TIV的生存期,发现80%的长生命TIV(long-lived TIV)生存期小于5 h,同时TIV的比例会受到处理数据方式的影响.

可见,目前的研究主要以测量数据分析为主,尚无关于TIV的生成位置以及其与接入时延关系的研究.

2 时延空间中的TIV与接入时延

2.1 核心网与接入网的划分

为研究接入时延对于TIV的影响,将因特网在逻辑上划分为核心网和接入网两部分.然而,如何确定核心网和接入网的边界是一个未决问题.文献[14]将端到端的路径分为非路由部分(non-routable section)和路由部分(routable section).非路由部分是指由于接入终端接入位置发生变化而在端到端路径中发生变化的前几跳或者后几跳,路由部分则指端到端路径中的其余部分.上述划分方法适用于移动的接入终端,对于固定节点无法判断其非路由部

分和路由部分.因此,本文拓展了这种划分方法.设端节点发送的分组先经一侧接入网进入核心网,然后再经过另一侧接入网到达另一个端节点,如图 1(a)所示.由于在因特网中,从某个端节点发出到达不同端节点的分组,必然会在沿某条路径前行时在某个路由器分开.在本文中,分组分离的路由器称为接入点.通过定义接入点,我们将接入点靠近端主机的链路称为接入链路,而在两台端主机接入点之间的链路称为核心链路.由此,可以将图 1(a)抽象为图 1(b),并进一步引入核心网和接入网边界的定义.

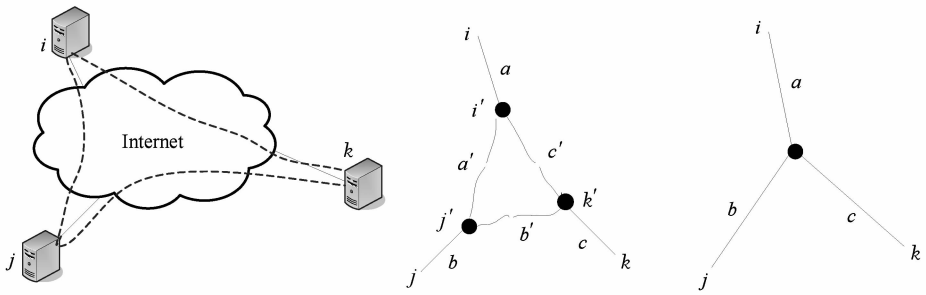


Fig. 1 The delay model of the internet.

图 1 因特网时延模型

用无向图 $G(N,E)$来表示因特网,N 表示因特网的节点集合,E 表示节点间的链路集合.N 又可以分为端节点集 Ne 和中继节点(intermediate nodes)集 Ni,显然 $Ne\cap Ni=\varnothing$.

定义 1. 路径.假定节点 i 和 j 之间存在一条路径,记为 $p(i,j)=\langle i,R_1^{(i,j)},R_2^{(i,j)},\cdots,R_{n-1}^{(i,j)},j\rangle$,其中 $R_n^{(i,j)}$ 表示该路径上的路由器,其下标表示是路径的第几跳,所有路径的集合记为 $P=\{p(i,j)|i,j\in N\}$.

定义 2. 接入点.如图 1(a)所示,$\forall i,j,k\in Ne$,节点 i 到节点 j,k 的路径分别为 $p(i,j)=\langle i,R_1^{(i,j)},R_2^{(i,j)},\cdots,R_{n-1}^{(i,j)},j\rangle$,$p(i,k)=\langle i,R_1^{(i,k)},R_2^{(i,k)},\cdots,R_{n-1}^{(i,k)},k\rangle$,则必存在某路由器 $R_i^{(i,j)}=R_i^{(j,k)}$ 使得 $p(R_i^{(i,j)},j)\cap p(R_i^{(j,k)},k)=\varnothing$.将 $R_i^{(i,j)}$ 用其下标简记为 i',则称节点 i' 为 i 相对于节点 j,k 的接入点,记为 $i'=div(i,\{j,k\})$,简称相对接入点(relative access point)(参见图 1(b)所示),节点 i 的所有接入点集记为 $DIV(i)$.如果一个端节点 i 有多个接入点的话,则取最靠近节点 i 的接入点,该接入点称为临近接入点(adjacent access point),记为 i'',在特殊情况下 $i''=i'$.

定义 3. 边界.由所有相对接入点构成的集合称为核心网与接入网的边界(edge),记为 $DIV(Ne)=\{DIV(i)|i\in Ne\}$;由所有临近接入点构成的集合称为核心网与接入网的临近边界(adjacent edge).

边界的定义给出了划分接入网与核心网的一种方法.只要能够找到接入点集合 $DIV(Ne)$,就能划分出核心网和接入网了.由此将网络划分为核心网 $G(Nc,Ec)$和接入网 $G(Na,Ea)$.显然,核心网节点 $Nc\subset Ni$,接入网节点 $Ne\subset Na$,$G(Na,Ea)$是个不连通的图,而 $G(Nc,Ec)$是个连通图.

上述关于接入网和核心网的定义与通常网络工程中的定义有所差别,是为了 TIV 建模和分析接入时延对 TIV 的影响而提出的.

2.2 TIV 与接入时延

下面依次定义时延空间中的三角形和 TIV.

定义 4. 三角形.$\forall i,j,k\in N$,若 $p(i,j),p(i,k),p(j,k)\in P$,则称节点 i,j,k 构成一个三角形 $\Delta(i,j,k)$.若 $\forall i,j,k\in Ne$,则称 $\Delta(i,j,k)$为端三角形;若 i,j,k 存在接入点 $i',j',k'\in Nc$,则 $p(i',j'),p(i',k'),p(j',k')\in P$,称 $\Delta(i',j',k')$为 $\Delta(i,j,k)$的内接三角形,如图 1(b)所示.

定义 5. 距离.节点 i,j 之间的距离是指节点间的往返时延,记为 $d(i,j)$,其中 $d(i,i)=0,d(i,j)=d(j,i)$;所有节点的距离矩阵表示为 $\boldsymbol{D}=[d_{ij}=d(i,j)]_{|Ne|\times|Ne|}$,$i,j\in Ne$.

在三角形和节点间距离的定义上给出时延空间中 TIV 的定义.

定义 6. TIV 三角形.$\forall i,j,k\in N$,构成的三角形为 $\Delta(i,j,k)$,设 $d(i,k)>d(i,j)$且 $d(i,k)>d(j,k)$,若 $d(i,k)\geqslant d(i,j)+d(j,k)$,则称 $\Delta(i,j,k)$具有 TIV 特性,简称 TIV 三角形.

定理 1. 任意一个端三角形 $\Delta(i,j,k)$都对应一个内接三角形 $\Delta(i',j',k')$或一个公共接入点.

证明. 设节点 i,j,k 为 $\Delta(i,j,k)$的三个顶点,由定义 2 得知端节点 i,j,k 必有接入点 $i',j',k'\in N$,则 $p(i',j')\subset p(i,j)\in P$,同理 $p(i',k'),p(j',$

$k')\in P$. 由定义 4,得知 $\Delta(i',j',k')$构成内接三角形;显然 $i'=j'=k'$时,i,j,k 具有公共的接入点,此时可以看作一个变为 0 的三角形. 证毕.

由定理 1 可知内接三角形 $\Delta(i',j',k')$三条边的取值只有 2 种情况:3 条边都不为 0 或者均为 0.

首先考虑内接三角形 3 边不为 0 的情况. 设 $d(i,k)$为 3 边中最长的边,若 $\Delta(i,j,k)$满足三角不等式,必须满足式(1):

$$d(i,k)<d(i,j)+d(j,k). \tag{1}$$

由定义 2 和定理 1 可以得到 i,j,k 的接入点 i', j',k'. 如图 1(b)所示,设 $d(i,i')=a, d(i',j')=a', d(j,j')=b, d(j',k')=b', d(k,k')=c, d(i',k')=c'$,将其代入式(1)得式(2):

$$c'<a'+b'+2b. \tag{2}$$

从式(2)可见,如果内接三角形 $\Delta'(i',j',k')$满足三角不等式,则由于接入时延 b 为非负,则式(2)的右边得到了加强,该三角不等式必成立. 反之,若 $c'>a'+b'$,$\Delta(i',j',k')$出现了 TIV,然而当 $b>(c'-a'-b')/2$ 时,式(2)依然能够成立,即此时从端到端的角度看 $\Delta(i,j,k)$不构成 TIV.

下面考虑 3 边均为 0 的情况. 此时,该内接三角形收缩为一个点,如图 1(c)所示. 若将各段时延代入式(1)得式(3). 由于 b 非负,式(3)恒成立,

$$0<2b. \tag{3}$$

实际上,3 边均为 0 是前者的一种特殊情况,即 3 个节点接入点重合,成树型拓扑.

通过上述分析,可以得到如下结论接入时延可以使得在端到端路径中观察到的 TIV 数目会减少,并减轻 TIV 的严重程度. 如图 1,设 $d(i,i')=5$, $d(i',j')=50, d(j,j')=4, d(j',k')=20, d(k,k')=9, d(i',k')=15$,此时其内接三角形 $\Delta(i',j',k')$构成 TIV,但是从端到端的角度看 $\Delta(i,j,k)$并不违反三角形不等式,从而减少了端到端 TIV 三角形的数目. 若 $d(k,k')=5$,则 $\Delta(i,j,k)$仍为 TIV 三角形,但与标准的三角形相差不大.

3 网络测量实验设计

为验证 TIV 模型分析得到的结论,基于 PlanetLab 平台设计了一组测量实验. 本节说明测量因特网时延的方法和数据预处理方法.

3.1 测量时延的方法

一般而言,测量网络端到端时延 T_{e2e} 是从网络边缘的一台端节点向位于网络另一侧边缘的端节点发起测量得到. 而要想获得核心网的时延 T_{core},可首先测量出端到端的时延 T_{e2e},再减去两台端节点的接入时延 T^1_{access} 和 T^2_{access}. 因此,本研究所需的网络测量实验难点在于除了需要测量各端节点之间的时延外,还需要知道其对应链路上的时延分布情况. 因此,选用测量时延的工具为 Tcptraceroute,其原因如下:1)在测量时延的同时也得到路径信息,然后再根据 DNS,ISP 和 BGP 等信息,推测因特网拓扑结构;2)基于 TCP 的测量工具能够提高被测节点的响应率,而基于 ICMP 的测量报文往往不被响应[15]. 这样在获得测量数据后,就可以根据核心网与接入网的边界提取出核心网的时延.

PlanetLab 实验平台是由分布于全世界各地的测量服务器组成,涵盖了世界主要国家、地区和 ISP 网络. 我们共选取了 314 台服务器作为测量端节点来参与实验,测量服务器所属国家及其数量情况参见表 1. 这些服务器以美国、欧洲、日本的节点为多,分别为 36.3%,31.5%和 5.4%,其中拥有 8 台以下服务器的国家未在表中列出. 许多机构和组织在 PlanetLab 中拥有不止一台服务器,在分析数据时对数据相似的一组服务器只选其中一台的数据进行分析.

Table 1 The Server Numbers of Different Countries

表 1 不同国家服务器的数目

Datasets	Count
US	144
EU	99
JP	17
PL	9
CA	9
CN	8

实验时,测量服务器的守护进程通过脚本每隔 10 min 进行一轮测量. 每轮测量时顺序地测量该服务器到其他节点的 RTT. 测量程序采用多线程并行方式运行,一次启动 15 个线程进行测量,每隔 300 s 检查并删除超时的线程. 经过统计,所有端主机之间的平均跳数为 15,平均端到端时延为 150 ms. 这样进行一次测量所花费的总时间一般会小于(150×15×314/15)ms=47.1s. 考虑到测量中引入的处理时延一般会在 10 倍以内,一个测量周期时间约为 7.85 min.

按照上述测量方法,在 PlanetLab 平台上从 2010 年 3 月 25 日 14:00 到 2010 年 3 月 26 日 14:00

进行了一天的测量,获得了原始测量数据达 561 MB 的数据集. 节点间时延受长链路的影响较大[16],一条跨洋的长链路很可能会大大增加数据集中 TIV 的比例. 为了避免跨洋链路的影响,在分析数据时,将数据集分为欧洲、美国和日本 3 个子集. 此外,我们又从数据集中随机选择了 114 个节点构成了一个随机数据集和整个测量数据集,共计 5 个数据集,分别记作 EU,US,JP,RAN 和 ALL 数据集.

3.2 数据预处理方法

分析测量数据时,考虑到有些路由器对测量没有响应,有些测量时延值抖动偏大,需要对这些原始数据进行预处理. 设对应每条路径序列 $p(i,j)=\langle i,R_1^{(i,j)},\cdots,R_{n-1}^{(i,j)},j\rangle$ 的逐跳时延序列为 $dp(i,j)=\langle d(i,R_1^{(i,j)}),d(R_1^{(i,j)},R_2^{(i,i)}),\cdots,d(R_{n-1}^{(i,i)},j)\rangle$,由此,可以得到路径集 P 对应的逐跳时延集 $DP=\{dp(i,j)|i,j\in Ne\}$. 对于 RTT,它应当满足递增性和对称性等性质,如式(4),(5):

$$\forall\, d(i,R_k^{(i,j)})\in dp(i,j), d(i,R_{k-1}^{(i,j)})<d(i,R_k^{(i,j)})<d(i,R_{k+1}^{(i,j)}), \tag{4}$$

$$|\,d(i,j)-d(j,i)\,|\leqslant\sigma. \tag{5}$$

对于不满足式(4),(5)的少数不完整数据要进行数据融合,其包括 2 个部分,即 IP 路径融合和时延融合. IP 路径融合是指利用前后几次测量结果,得到中间测量丢失的 IP 地址,并得到测量中一条完整路径. IP 路径融合首先通过扫描所有路径中出现频率最高的路径作为完整路径,其他路径称为临时路径. 在分析测量结果中,一般仅有少量出现临时路径的情况需要处理. 时延融合则是指利用前后几次测量结果获得丢失时延的过程. 在 Tcptrarceroute 测量过程中,到每一跳都测量 3 次,并选取第 3 次测量作为有效数据. 在经过数据融合后,根据式(4),(5)过滤掉不合理的数据.

由此,就能够得到一个包括所有端节点的距离矩阵 $\boldsymbol{D}$、路径集 P 以及逐跳时延集 DP. 利用 D,P 和 DP 可以分别求得在边界和临近边界定义下的出现 TIV 三角形比例的方法,具体方法如下.

如图 1(b)所示,设网络中任意 3 个端节点节点 i,j,k 接入点分别为 i',j',k' 和临近接入点分别为 i'',j'',k'',可以在边界定义下得到节点间距离:

$$d(i',j')=d(i,j)-d(i,i')-d(j,j'), \tag{6}$$

在临近边界定义下得到节点间距离:

$$d(i'',j'')=d(i,j)-d(i,i'')-d(j,j''). \tag{7}$$

同理,可以求得 $d(i',k')$,$d(j',k')$,$d(i'',k'')$,$d(j'',k'')$,然后可按照定义 6 判断出 $\Delta(i',j',k')$ 或 $\Delta(i'',j'',k'')$ 是否具有 TIV 特性. 基于此,我们在算法 1 中给出了一个计算数据集中 TIV 比例的算法 ScoutTIV.

算法 1 中 ScoutTIV 算法给出了在边界定义下计算 TIV 比例的过程,若是计算临近边界定义下 TIV 的比例,只要将算法中的相应步骤⑤按照临近接入点的定义进行替换就可以了. 分析可知,ScoutTIV 算法的时间复杂度为 $O(N^3)$、空间复杂度为 $O(N^2)$.

算法 1. ScoutTIV 算法.

输入: $\boldsymbol{D}$,P,DP;

输出: PR. /* TIV ratio */

① Initialize $\boldsymbol{D}$,P,DP;
② for(i to size($\boldsymbol{D}$))
③ for($j=i+1$ to size($\boldsymbol{D}$))
④ for($k=j+1$ to size($\boldsymbol{D}$))
⑤ $i'=div\{i,\{j,k\}\}$; $j'=div\{j,\{i,k\}\}$;$k'=div\{k,\{i,j\}\}$;
⑥ $d(i',j')=d(i,j)-d(i,i')-d(j,j')$; $d(i',k')=d(i,k)-d(i,i')-d(k,k')$;
 $d(j',k')=d(j,k)-d(j,j')-d(k,k')$;
⑦ if (Validate($d(i',j')$,$d(i',k')$,$d(j',k')$)
⑧ $TCount$++; /* the total number of triangles */
⑨ if($IsTIV(\Delta(i',j',k'))$
⑩ $TIVCount$++; /* the TIV number */
⑪ endif
⑫ endif
⑬ endfor
⑭ endfor
⑮ endfor
⑯ return $PR=TIVCount/TCount$.

4 TIV 模型的验证

本节使用在因特网中的实测数据来验证第 3 节的理论分析结果. 4.1 节首先分析接入点的位置;4.2 节分析接入时延对 TIV 的影响.

4.1 接入点位置

根据 3.1 节点定义本文计算了相对接入点和临近

接入点的位置,并给出了两者的累计分布概率(cumulative distribution function, *CDF*),如图 2 所示:

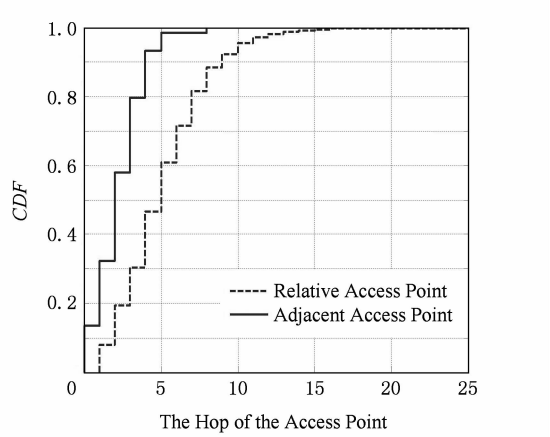


Fig. 2 The positions of the access points.

图 2 接入点的位置

临近接入点和相对接入点的位置可以作为反映网络连接丰富性的一个测度. 从定义看,接入点是分组路径分岔的起点,意味着在接入点之后存在一个提供穿越服务的网络,而非一条链路. 因而,接入点越靠近边缘,则说明网络的连接的丰富性. 在这种意义上,临近接入点反映了节点到核心网络到的距离;相对接入点则反映了位于临近接入点之后网络连接的丰富性. 文献[17]认为 86%节点接入网络的跳数为 4 至 5 跳. 从图 2 来看,在临近接入点定义下 95%以上节点的接入点位置位于 5 跳以内,这与文献[17]的结果类似,说明了测量数据的合理性;采用相对接入点定义,接入点的位置只有 60%位于 5 跳以内,80%的节点位于 7 跳以内,说明网络内部的连接是比较丰富的. 这就为时延空间中 TIV 的出现提供了必备条件.

4.2 接入时延对 TIV 的影响

为了考察接入时延对 TIV 的影响,本文使用 TIV 比例来分别定义数据集中 TIV 的数目多少和严重性. TIV 比例(prevalence ratio, *PR*)是指数据集中具有 TIV 特性的三角形数目与所有三角形数目之比,如式(8)所示:

$$PR=\frac{\sum_{i,j,k\in Ne} IsTIV(\Delta(i,j,k))}{\sum_{i,j,k\in Ne}\Delta(i,j,k)}. \tag{8}$$

其中,若 $\Delta(i,j,k)$ 具有 TIV 特性,则 $IsTIV(\Delta(i,j,k))=1$,否则 $IsTIV(\Delta(i,j,k))=0$. TIV 比例越大说明该数据集中 TIV 三角形的比例越大,反之,说明 TIV 三角形比例较小.

表 2 给出了在不同定义下按照式(10)计算出来的 TIV 比例,前两列是按照边界和临近边界计算的核心网中的 TIV 比例,第 3 列是端到端网络中的 TIV 比例. 从表 2 可以发现,采用临近边界定义计算获得的 TIV 比例要低于采用边界定义的 TIV 比例,这是由于临近边界更加靠近网络的端节点,引入了一定的接入时延. 然而,无论是采用边界还是临近边界的定义,核心网中的 TIV 比例都要高于端到端网络中的 TIV 比例. 这说明接入时延在很大程度上减少了 TIV 的数目.

Table 2 The PR Values in the Core Network and End-to-End Networks

表 2 核心网与端到端网络中的 TIV 比例%

Datasets	Core Network		End-to-End
	Edge	Adjacent Edge	Network
EU	43.1	15.4	11.9
JP	57.3	30.6	20.5
US	47.6	18.7	14.0
RAN	44.1	15.3	13.7
ALL	46.4	16.8	13.9

在第 3 节的分析中得知:1)若内接三角形满足三角形不等式,由于接入时延的作用,这种性质会加强,从端到端看也不会出现 TIV 三角形;2)若在核心网中存在 TIV 三角形,由于接入时延的作用,则从端到端看出现 TIV 三角形的比例将减少.

先验证第 1 种情况. 我们统计了核心网中所有不构成 TIV 三角形的节点,并将这些节点构成的子网看作一个核心网不存在 TIV 的网络. 通过计算各个数据集中的一致比(consistency ratio, *CR*)来分析接入时延对于三角形性质的加强作用. 一致比被定义为所有不具有 TIV 特性的内接三角形所对应的端三角形中不具有 TIV 特性的端三角形所占的比例,如式(9)所示:

$$CR=\frac{\sum_{i,j,k\in Ne} NoTIV(\Delta(i,j,k))}{\sum_{i',j',k'\in Nc} NoTIV(\Delta(i',j',k'))},\quad \text{if } NoTIV(\Delta(i',j',k'))=1. \tag{9}$$

其中,若 $\Delta(i,j,k)$或 $\Delta(i',j',k')$不具有 TIV 特性,则 $NoTIV(\Delta(i,j,k))=1$,否则 $NoTIV(\Delta(i,j,k))=0$.

当核心网中不存在 TIV 三角形时,对应的端到端网中也不应当有 TIV 三角形,即一致比在理论上应当等于 1,但实际测量值并非如此. 不过,从表 3 可以看出,这几个数据集的一致比都比较接近于 1,

这验证了结论 1 的正确性. 出现误差的可能原因是无法在同一时刻测得所有节点间的时延,网络在不同时刻表现出不同的性能,使分析存在一定的误差.

Table 3 The CR Values of Different Datasets
表 3 不同数据集的一致比

Datasets	CR/%
EU	98.3
JP	98.6
US	94.2
RAN	97.9
ALL	98.2

下面验证第 2 种情况,即接入时延会减少端到端 TIV 的数目. 为此,定义了转化比(transform ratio, TR)以说明具有 TIV 特性的内接三角形在接入时延的影响下,其对应的端三角形不再具有 TIV 特性. TIV 的转化比定义为所有具有 TIV 特性的内接三角形对应的端三角形中不具有 TIV 特性的三角形数目与所有具有 TIV 特性的内接三角形数目之比,如式(10)所示:

$$TR=\frac{\sum_{i,j,k\in Ne} NoTIV(\Delta(i,j,k))}{\sum_{i',j',k'\in Nc} IsTIV(\Delta(i',j',k'))},$$
$$\text{if } IsTIV(\Delta(i',j',k'))=1. \tag{10}$$

其中,若 $\Delta(i',j',k')$对应的端三角形 $\Delta(i,j,k)$具有 TIV 特性,则 $NoTIV(\Delta(i,j,k))=1$,否则:

$$NoTIV(\Delta(i,j,k))=0.$$

根据式(10)得到了表 4,从表 4 可见,各个数据集中均有 70%以上的 TIV 三角形在端到端的角度看不再具有 TIV 特性. 这说明接入时延的确减少了从端到端观察到的 TIV 三角形的数目,验证了结论 2 的正确性.

Table 4 The TR Values of Different Datasets
表 4 具有 TIV 特性的内接三角形的转化比

Datasets	TR/%
EU	81.6
JP	72.1
US	69.5
RAN	70.1
ALL	72.3

5 结论和下一步工作

准确地把握因特网时延空间特性对于保障网络服务质量至关重要,其中 TIV 的分析与建模是难点. 本文首先对因特网时延空间中 TIV 的位置进行了建模,经理论分析、实验测量发现接入时延可以使得在端到端路径中观察到的 TIV 数目会减少,并减轻 TIV 的严重程度. 这些发现对于深入理解因特网时延空间的性质和构建更为准确的模型具有十分重要的指导意义.

本文分析的实验数据来源于 PlanetLab 实验平台,其节点由大学和企业提供的服务器构成,它们靠近网络核心,具有较小的接入时延. 在未来工作中,将设法测量不同接入方式的端主机来进一步丰富和完善 TIV 模型,并提出克服 TIV 的影响的方法.

参 考 文 献

[1] Dabek F, Cox R, Kaashoek F, et al. Vivaldi: A decentralized network coordinate system [C] //Proc of ACM SIGCOMM 2004. New York: ACM, 2004: 15-26

[2] Venugopalan D, Malkhi F, Kuhn, et al. On the treeness of Internet latency and bandwidth [C] //Proc of ACM SIGMETRICS 2009. New York: ACM, 2009: 61-72

[3] Lumezanu C, Spring N. Measurement manipulation and space selection in network coordinates [C] //Proc of ICDCS 2008. Piscataway, NJ: IEEE, 2008: 361-368

[4] Wang G, Zhang B, Ng T S E. Towards network triangle inequality violation aware distributed systems [C] //Proc of the 7th ACM SIGCOMM Conf on Internet Measurement. New York: ACM, 2007: 175-188

[5] Wu Guofu, Dou Qiang, Ban Dongsong, et al. A novel passive-landmark based network distance prediction method [J]. Journal of Computer Research and Development, 2011, 48(1): 125-132 (in Chinese)
(吴国福, 窦强, 班冬松, 等. 一种基于被动路标的网络距离预测方法[J]. 计算机研究与发展, 2011, 48(1): 125-132)

[6] Kaafar M A, Gueye B, Cantin F, et al. Towards a two-tier Internet coordinate system to mitigate the impact of triangle inequality violations [C] //Proc of IFIP Networking. Berlin: Springer, 2008: 397-408

[7] Zheng H, Lua E K, Pias M, et al. Internet routing policies and round-trip-times [C] //Proc of PAM 2005. Berlin: Springer, 2005: 236-250

[8] Labovitz C, Iekel-Johnson S, McPherson D, et al. Internet inter-domain traffic [C] //Proc of ACM SIGCOMM 2010. New York: ACM, 2010: 75-86

[9] Lee S, Zhang Z, Sahu S, et al. On suitability of Euclidean embedding for host-based network coordinate systems [J]. IEEE/ACM Trans on Networking, 2010, 18(1): 27-40

[10] Hotz S. Routing information organization to support scalable interdomain routing with heterogeneous path requirements [D]. Los Angeles: University of Southern California, 1994

[11] Ng T S E, Zhang H. Predicting Internet network distance with coordinates-based approaches [C] //Proc of IEEE INFOCOM 2002. Piscataway, NJ: IEEE, 2002: 170-179

[12] Lumezanu C, Baden R, Spring N, et al. Triangle inequality and routing policy violations in the Internet [C] //Proc of PAM 2009. Berlin: Springer, 2009: 45-54

[13] Lumezanu C, Baden R, Spring N, et al. Triangle inequality variations in the Internet [C] //Proc of the 9th ACM SIGCOMM Conf on Internet Measurement. New York: ACM, 2009: 177-183

[14] Schwartz Y, Shavitt Y, Weinsberg U. A measurement study of the origins of end-to-end delay variations [C] //Proc of PAM 2010. New York: Springer, 2010: 21-30

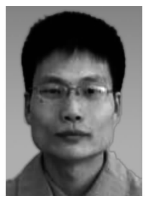
[15] Shaun C. Teptraceroute [EB/OL]. [2010-03-20]. http://michael. toren. net/code/tcptraceroute/tcptraceroute-1. 5beta7. tar. gz

[16] Zeitoun A, Chuah C, Bhattacharyya S. An AS-level study of Internet path delay characteristics [C] //Proc of GLOBECOM 2004. Piscataway, NJ: IEEE, 2004: 1480-1484

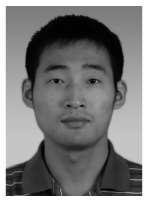
[17] Hu N. Network monitoring and diagnosis based on available bandwidth measurement [D]. Pittsburgh: Carnegie Mellon University, 2006

Wang Zhanfeng, born in 1982. PhD. His main research interests include network measurement, network performance analysis and modeling.

Chen Ming, born in 1956. PhD, professor and PhD supervisor. His main research interests include network architecture, network performance analysis and modeling, network measurement, and distributed system design and optimization.



Xing Changyou, born in 1982. PhD, and lecturer. His main research interests include network management, and distributed system design.



Bai Huali, born in 1981. PhD. His main research interests include network measurement, network performance analysis and modeling, and Internet topology discovery.

Wei Xianglin, born in 1985. PhD. His main research interests include distributed system design and optimization, peer-to-peer, and network security.