软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(5):1296−1314 [doi: 10.13328/j.cnki.jos.005148]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种 IP 网络拥塞链路丢包率范围推断算法*

陈 宇[1,2], 周 巍[1], 段哲民[1], 钱叶魁[3], 赵 鑫[3,4]

[1](西北工业大学 电子信息学院,陕西 西安 710072)

[2](郑州航空工业管理学院,河南 郑州 450015)

[3](解放军防空兵学院,河南 郑州 450052)

[4](中国电子科技集团公司 第五十四研究所 通信网信息传输与分发技术重点实验室,河北 石家庄 050081)

通讯作者: 陈宇, E-mail: chenyu3440@gmail.com

摘 要: 针对大规模IP网络拥塞链路丢包率范围推断算法中存在的不足,提出一种贪婪启发式拥塞链路丢包率范围推断算法.借助多时隙路径探测,避开单时隙探测对时钟同步的强依赖;通过学习各链路拥塞先验概率,借助贝叶斯最大后验定位拥塞链路;提出了聚类拥塞链路相关、性能相近路径集合的策略,通过对聚类路径集合中性能相似系数求解,循环推断拥塞链路丢包率范围.实验验证了算法的准确性及鲁棒性.

关键词: IP 网络;拥塞链路推断;丢包率范围;贝叶斯最大后验概率;贪婪启发算法

中图法分类号: TP393

中文引用格式: 陈宇,周巍,段哲民,钱叶魁,赵鑫.一种 IP 网络拥塞链路丢包率范围推断算法.软件学报,2017,28(5):1296−1314. http://www.jos.org.cn/1000-9825/5148.htm

英文引用格式: Chen Y, Zhou W, Duan ZM, Qian YK, Zhao X. Congested link loss rate range inference algorithm in IP network. Ruan Jian Xue Bao/Journal of Software, 2017,28(5):1296−1314 (in Chinese). http://www.jos.org.cn/1000-9825/5148.htm

Congested Link Loss Rate Range Inference Algorithm in IP Network

CHEN Yu[1,2], ZHOU Wei[1], DUAN Zhe-Min[1], QIAN Ye-Kui[3], ZHAO Xin[3,4]

[1](Institute of Electronic Information, Northwestern Polytechnical University, Xi'an 710072, China)

[2](Zhengzhou University of Aeronautics, Zhengzhou 450015, China)

[3](Air Defence Forces Academy of PLA, Zhengzhou 450052, China)

[4](Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory, No.54 Research Institute, China Electronics Technology Group Corporation, Shijiazhuang 050081, China)

Abstract: Addressing the shortcomings of existing link congestion loss rate range inference algorithms in large scale IP network, a new link congestion loss rate range inference algorithm based on greedy heuristic method is proposed. The strong dependency on the clock synchronization of single slot E2E path measurements is avoided through using multiple slots E2E path measurements. Each congested link can be located through adopting the link congestion Bayesian maximum a-posterior (BMAP) after learning prior probabilities of the link congestion. The set consisting of paths with related congested links and similar performance is constructed. Through solving the performance similarity coefficient dynamically, loss rate range of each congested link can be recurrently inferred. The accuracy and robustness of the algorithm proposed in this paper is verified by experiments.

* 基金项目: 国家重点基础研究发展计划(973)(2013CB329104); 国家自然科学基金(61103225); 通信网信息传输与分发技术重点实验室基金

Foundation item: National Program on Key Basic Research Project of China (2013CB329104); National Natural Science Foundation of China (61103225); Science and Technology on Information Transmission and Dissemination in Communication Networks Laboratory Research Fund

收稿时间: 2016-06-16; 修改时间: 2016-07-28, 2016-09-22; 采用时间: 2016-09-29; jos 在线出版时间: 2017-01-20

CNKI 网络优先出版: 2017-01-20 16:06:39, http://www.cnki.net/kcms/detail/11.2560.TP.20170120.1606.015.html

Key words: IP network; congested link inference; loss rate range; Bayesian maximum a-posteriori (BMAP); greed heuristic algorithm

随着 IP 网络规模的不断扩大,IP 网络多链路拥塞现象时有发生[1],甚至涉及违反 SLA(service-level agreement)等相关服务等级协定.因此,及时发现 IP 网络内部拥塞链路并采取相应处理,对流量工程、流量调度等网络管理方面有重要的指导意义.

IP 网络断层扫描(tomography)技术[2]通过少量端到端(end-to-end,简称 E2E)路径性能主动探测,利用统计学等相关技术间接推断 IP 网络内部各链路性能(如时延、带宽、丢包率等),不需要直接访问 IP 网络内部各路由器/交换机,不涉及用户隐私,虽然优点较多,但如何实现却极具挑战.早期的网络 tomography 技术借助单时隙 E2E 路径性能探测,构建系统线性方程组,计算各链路性能值(如丢包率),推断拥塞链路的方法被称为 analog tomography[3−6].由于 tomography 技术本身借助少量路径探测,覆盖尽可能多的待测 IP 网络链路,易造成方程组系数矩阵奇异,为了得到链路丢包率唯一解,需对方程组系数矩阵满秩扩展.因此,要求各 E2E 路径性能主动探测时保证严格的时钟同步.但是由于 IP 网络中大部分路由器的单播支持度高于多播,时钟同步难以保证,算法推断性能较差.另外,因方程组求解涉及复杂的求逆计算,大规模 IP 网络多链路拥塞极易引发维数灾难,实时性无法保证,甚至导致算法失效.为了简化链路性能推断过程,借助布尔(Boolean)代数模型推断拥塞链路的方法称为 Boolean tomography[7].其中,CLINK 算法[8]借助多时隙 E2E 路径性能探测,有效地避开了单时隙路径探测对时钟同步的依赖.但是该类方法仅能推断出链路“好”与“坏”两种状态,拥塞链路性能分辨率低,且基于最小覆盖集(smallest coverage set,简称 SCS)理论定位拥塞链路,如拥塞路径中存在多链路拥塞,仅能推断出其中 1 条链路而导致推断误差.为了提高拥塞链路性能分辨率,Range tomography[1]算法借助单时隙 E2E 路径性能探测,以“E2E 路径中共享数目最多的瓶颈链路为最有可能发生拥塞的链路”这一经验知识推断拥塞链路及其丢包率范围.但是 Range tomography 算法假设 IP 网络中发生拥塞的链路数较少,在复杂的 IP 网络环境中,特别是在多链路拥塞场景下,仅以共享数最多的瓶颈链路推断拥塞链路及其丢包率范围亦存在较大误差.

针对现有 IP 网络内部链路性能推断算法中存在的不足,本文提出一种贪婪启发式拥塞链路丢包率范围推断(congested link loss rate range inference,简称 CLLRRI)算法.该算法综合 Boolean tomography 技术中 Boolean 代数的实用性,借助多时隙 E2E 路径探测,学习各链路拥塞先验概率 p_k,作为定位拥塞链路经验知识,代替传统专家经验知识,基于各链路拥塞贝叶斯最大后验(Bayesian maximum A-posteriori,简称 BMAP)得出推断时刻各链路拥塞最大后验代价值 C_p,定位拥塞链路 l_m,提高拥塞链路定位准确性;提出聚类途经拥塞链路 l_m 且性能相近路径到聚类路径集合 $\boldsymbol{\Omega}$,并提出动态计算路径集合 $\boldsymbol{\Omega}$ 中的性能相似系数 δ,作为推断拥塞链路 l_m 丢包率范围的依据,借助贪婪启发式搜索算法循环推断出整个待测 IP 网络中各拥塞链路丢包率范围.

本文的创新在于:(1) 提出了一种拥塞链路丢包率范围贪婪启发式推断算法,该算法通过学习链路拥塞先验知识,基于 BMAP 准则定位拥塞链路,提出拥塞链路相关且性能相近路径聚类的策略,循环推断各拥塞链路丢包率范围;(2) 通过与经典算法 CLINK 及 Rang tomography 算法的模拟实验与仿真实验比较,验证了本文提出 CLLRRI 算法的准确性及鲁棒性.

本文第 1 节综述相关工作.第 2 节提出存在的问题并构建推理模型.第 3 节详细介绍 CLLRRI 算法.第 4 节进行实验评价.第 5 节进一步讨论.第 6 节总结全文.

1 相关工作

自 1996 年 Vardi 首次提出在 IP 网络性能推断中使用类似医学层析扫描的 tomography 技术[9]以来,借助网络 tomography 技术推断 IP 网络内部链路性能主要包括 3 类方法.早期的 analog tomography 方法[3]涉及线性方程组求逆计算,对各 E2E 路径的性能探测需要保持较高的时钟同步,Adams,Bu 等人[4,5]利用多播方式探测各 E2E 路径性能,推断链路丢包率.但是由于网络中路由器对单播的支持度远高于多播,Duffield 等人[10,11]提出利用类似多播的单播背靠背包测量来获取报文级的相关性以及借助模拟多播的包群探测方法.但因时间相关性问题,易导致算法推断性能变差.另外,因为线性方程组涉及复杂的求逆问题,在大规模 IP 网络多链路拥塞场景下易引

发维数灾难,实时性难以保证,甚至导致算法失效.Duffield 等人[7]将前期工作加以完善,基于 Boolean 代数域[12]建立用于诊断拥塞链路的网络层析成像新框架,Boolean tomography 算法根据各 E2E 路径性能,推断拥塞路径途经链路的性能状态,0 代表正常,1 代表拥塞.Boolean tomography 也称作 Binary tomography(二进制网络层析成像).Nguyen 等人[8]提出了 CLINK 算法,将概率模型引入到拥塞链路推断中,通过多时隙路径性能探测,有效地避开了单时隙路径探测对时钟同步[13]的强依赖,拥塞链路推断性能较之不使用先验概率的 SCFS 算法[10]及使用一致先验概率的 MCMC 算法[14]均有较大幅度的提高.但是 Boolean tomography 仅能推断出链路拥塞与否,无法推断出链路的拥塞程度,分辨率低.CLINK 算法在大规模 IP 网络多链路拥塞场景下,特别是 E2E 路径中存在多条拥塞链路时,以 SCS 理论作为拥塞链路集合推断准则,导致推断性能下降明显.Netscope[15]及 LIA[16]算法基于 SCS 准则,辅以各 E2E 路径多次丢包率测量值推断拥塞链路丢包率,但因路径性能测量时间相关性较难保证,导致算法鲁棒性较差.Zarifzadeh 等人提出的拥塞链路丢包率范围推断算法 Range tomography[1]认为,IP 网络中发生拥塞的链路数较少,且以瓶颈链路共享数多少作为拥塞链路推断的经验知识,在复杂的 IP 网络多链路拥塞环境下,算法推断性能显著下降.

本文提出的 CLLRRI 算法基于 BMAP 准则推断拥塞 E2E 路径中最有可能发生拥塞的链路,代替传统以瓶颈链路作为拥塞链路定位的经验方法[1,8];提出一种路径集合聚类新方法,通过对聚类集合中路径性能相似系数的动态求解方法,贪婪启发式推断各拥塞链路丢包率范围.模拟实验及仿真实验均验证了 CLLRRI 算法不仅在拥塞链路定位性能上优于 CLINK 算法及 Range tomography 算法,而且在拥塞链路丢包率范围推断性能上也优于 Range tomography 算法.

2 问题提出及模型构建

2.1 问题提出

利用主动检测方法推断网络内部各链路性能的 tomography 方法,首先,各探针部署路由器节点向其他节点发送 ICP/UDP 包(ping 及 traceroute),尽可能地覆盖整个 IP 网络,Ping 获取到各探测 E2E 路径的性能(丢包率、可用带宽、时延等),traceroute 获取到各探测 E2E 路径途经的链路关系;然后,借助获取到的信息,构建待测 IP 网络内部各链路性能与探测 E2E 路径性能之间的关系方程或者数学模型,对 IP 网络内部链路性能进行计算或推理,如公式(1)所示.

$$E_i=F(\{e_j|\forall l_j\in P_i\}) \tag{1}$$

其中,E_i 为已知路径 P_i 的性能,e_j 为未知链路 l_j 的性能,F 代表路径性能与链路性能之间的函数关系.

如,某 E2E 路径 P_i 的丢包率实际值为 L_i,则由“路径传输率等于途经各链路传输率的乘积”及“丢包率=1−传输率”,可将公式(1)转化为路径丢包率及链路丢包率之间的关系表达式(2).

$$L_i=1-\prod_{j=1}^{n_c}(1-\xi_j),\forall l_j\in P_i \tag{2}$$

其中,ξ_j 为路径 P_i 途经的第 j 条链路丢包率,n_c 为路径 P_i 途经的链路数.故可对公式(2)两边取对数构建线性方程组,求解各链路丢包率值.但是,此方法因无法保证各 E2E 路径探测时严格的时钟同步等原因,求解精度较差.另外,由于 E2E 路径数较少,易造成方程组系数矩阵欠定,无法求得各链路丢包率唯一解.由于不同 E2E 路径共享的同一条链路在短时间内,其丢包率变化不会太大[1].本文将丢包率作为链路性能推断目标,但是为了简化拥塞链路定位过程,借鉴 Boolean tomography 中对性能进行的实用性表述,E2E 路径及途经链路的状态变量定义如下.

定义 1. IP 网络中,各 E2E 路径的状态变量集合 $\boldsymbol{Y}=(y_1,...,y_i,...,y_{n_\theta})$,各 E2E 路径途经链路的状态变量集合 $\boldsymbol{X}=(x_1,...,x_j,...,x_{n_\varepsilon})$.当某路径 P_i 拥塞时,其状态变量 y_i=1;反之,y_i=0.同理,当某链路 l_j 拥塞时,其状态变量 x_j=1;反之,x_j=0.其中,n_θ为待测 IP 网络 E2E 路径总数,n_ε为 E2E 路径途经的链路总数.

在对待测 IP 网络进行 tomography 时,通常以尽可能少的 E2E 路径性能探测,覆盖网络中尽可能多的链路.因此,E2E 路径途经链路通常不止 1 条,当不少于 1 条途经链路发生拥塞时,路径拥塞;当途经链路均正常时,路径

正常.而在对实际 IP 网络各 E2E 路径进行探测时,同一条路径在不同时刻的状态会发生改变,虽然 E2E 路径状态能够通过 ping 获取,但由于网络流量的动态特性下,链路状态也在实时发生改变,因此,是哪条(些)链路拥塞造成的路径拥塞不易获取,链路的拥塞程度更是难以把握.

由于网络流量具有很强的动态特性,在实际网络中,造成路径拥塞的情况亦非常复杂,无论是路径中共享数目最多的瓶颈链路,还是其他瓶颈链路,或是非瓶颈链路,都可能造成该路径拥塞.不同时刻,同一条路径发生拥塞,途经的拥塞链路位置也可能不同.因此,以“共享瓶颈链路作为最容易发生拥塞的链路”这一专家经验知识定位拥塞链路可能存在较大误差.研究发现,拥塞路径与拥塞链路之间存在着一定的关联规律[8].因此,本文根据网络拓扑结构和 N 次 E2E 路径探测结果获取拥塞链路推理时刻前各链路的拥塞先验概率,通过引入贝叶斯推理模型,完成推理时刻各拥塞 E2E 路径中最有可能发生拥塞的途经链路定位,并提出一种丢包率范围循环推断方法,获取各拥塞链路的丢包率范围.通过仿真实验发现:在对待测 IP 网络各 E2E 路径进行 $N\geqslant 30$ 次探测时,能够有效学习路径与途经链路间的性能关系规律.

CLINK 算法虽然借助于贝叶斯理论学习各链路拥塞先验概率,但在定位拥塞链路时基于 SCS 理论,当拥塞链路数较多时,特别是一条路径中存在多条拥塞链路时,仅能定位其中最有可能发生拥塞的一条链路,而将该拥塞路径途经的其他拥塞链路视为正常链路,导致多链路拥塞时拥塞链路的定位性能下降,且链路性能分辨率低.为了提高链路性能分辨率,Range tomography 算法[1]根据 E2E 路径丢包率测量结果推断途经可能发生拥塞的链路及其丢包率范围.但该算法的前提假设是各 E2E 路径中发生拥塞的链路数不多,且仅以拥塞路径中共享数最多的瓶颈链路作为拥塞链路,推断该链路的丢包率范围.而当大规模 IP 网络多链路拥塞时,拥塞链路定位及其拥塞程度推理性能均不能有效保证.

2.2 贝叶斯网模型构建

本文利用图论中的有向无环图模型 $G=(\nu,\varepsilon)$,其中,ν为节点,ε为有向边.首先,对待测 IP 网络进行贝叶斯网模型建模.待测 IP 网络中,E2E 路径状态变量 $\boldsymbol{Y}$ 中的各元素构成贝叶斯网模型中的观测变量(证据节点),E2E 路径途经各链路的状态变量 $\boldsymbol{X}$ 中的各元素构成隐藏变量(隐藏节点),E2E 路径与该路径途经链路的关系为模型有向边.则构建的 IP 网络贝叶斯网模型如图 1 所示.

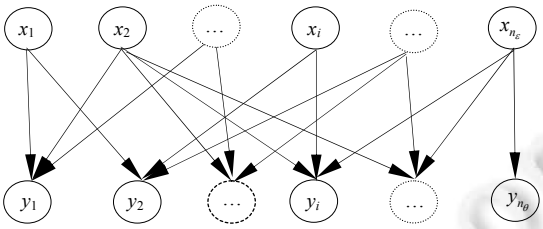


Fig.1 Bayesian network model of IP network

图 1 IP 网络贝叶斯网模型

在定位拥塞链路时,由于拥塞链路必存在于拥塞路径中,正常路径途经的各链路均正常,故可将正常路径对应的观测变量以及该路径途经各链路对应的隐藏变量从图 1 所示的贝叶斯网模型中移除.因此,在链路拥塞先验知识学习过程中以及当前时刻拥塞链路及其性能范围推断过程中,分别对待测 IP 网络构建的贝叶斯网模型 G 进行二次简化.

1) 在链路拥塞先验知识学习过程中,根据 N 次 E2E 路径性能探测,将拥塞次数小于设置阈值的路径视为正常(比如,设置阈值为 0,表明对路径性能要求极高,多次路径性能探测中出现 1 次拥塞,该路径即视为拥塞路径),对正常路径及其途经链路对应的节点以及有向边进行移除,移除后的剩余贝叶斯网模型为 G'.
2) 在当前时刻拥塞链路及其丢包率范围推断过程中,将 E2E 路径性能探测中正常路径及其途经链路对

应的节点以及有向边从 G'中移除,再次移除后的剩余贝叶斯网模型为 G''.

3 CLLRRI 算法

本文提出一种贪婪启发式拥塞链路丢包率范围推断算法 CLLRRI.算法原理框图如图 2 所示.

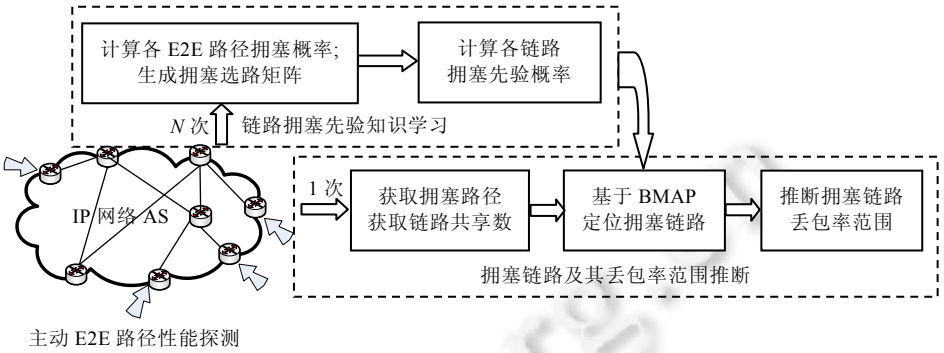


Fig.2 Principle diagram of CLLRRI algorithm

图 2 CLLRRI 算法原理框图

算法主要包括两个部分.

1) 链路拥塞先验知识学习.通过在待测 IP 网络各叶子节点部署探针,借助 tomography 技术,对各 E2E 路径进行 N 次性能探测[8],获取各 E2E 路径性能及各 E2E 路径途经链路,构建 Boolean 线性方程组,求解各链路拥塞先验概率,以获取各链路拥塞先验知识.
2) 拥塞链路定位及其丢包率范围推断.当前拥塞链路及其性能范围推断时刻,根据 1 次 E2E 路径性能探测结果,基于 BMAP 准则及拥塞路径丢包率更新值,贪婪启发式循环定位各 E2E 拥塞路径途经的拥塞链路,并推断各拥塞链路的丢包率范围.

3.1 链路拥塞先验知识学习

在 IP 网络中,各拥塞路径传输率等于该路径途经各链路传输率的乘积[17],如公式(3)所示.

$$\Psi_i = \prod_{j=1}^{n_{\varepsilon'}} \varphi_j^{d_{ij}} \tag{3}$$

其中,Ψ_i 为第 i 条拥塞路径传输率;φ_j 为该路径途经第 j 条链路传输率;d_{ij} 为选路矩阵 D 第 i 行 j 列元素值;d_{ij}={0,1};$n_{\varepsilon'}$为 E2E 路径性能探测中,拥塞路径途经的链路数.对 IP 网络选路矩阵 D 的构建方法详见文献[8].借助于二进制最大化操作“∨”,可将路径拥塞与链路拥塞关系表示为

$$y_i = \bigvee_{j=1}^{n_{\varepsilon'}} x_j \cdot d_{ij} \tag{4}$$

为了对 IP 网络链路拥塞先验概率进行推理,对公式(4)两边取数学期望 E 转换后,可得路径拥塞与链路拥塞关系表达式:

$$E[y_i] = P\left(\bigvee_{j=1}^{n_{\varepsilon'}} x_j \cdot d'_{ij} = 1\right) = 1 - P\left(\bigvee_{j=1}^{n_{\varepsilon'}} x_j \cdot d'_{ij} = 0\right) = 1 - \prod_{j=1}^{n_{\varepsilon'}} (1-p_j)^{d'_{ij}} \tag{5}$$

其中,d'_{ij} 为根据 N 次 E2E 路径性能探测,去除正常路径及途经链路后得到的拥塞选路矩阵.路径拥塞期望值 $E[y_i]$ 可通过 N 次路径探测得到的每次路径拥塞变量 y_i={0,1}求和取平均值后得出,用 $\overline{y}_i$ 表示.为了方便求解,对公式(5)两边同时取对数,可得剩余贝叶斯网模型 G'下拥塞链路先验概率求解 Boolean 代数方程组,如公式(6)所示.

$$\lg(1-\overline{y}_i) = \lg(1-P_i) = \sum_{j=1}^{n_{\varepsilon'}} d'_{ij} \cdot [\lg(1-p_j)] \tag{6}$$

将各 E2E 路径拥塞概率 P_i 及拥塞选路矩阵 d'_{ij} 带入公式(6),即可求得拥塞路径途经各链路的拥塞先验概率 p_j.但是,如果公式(6)中方程组系数矩阵 d'_{ij} 不满秩,则无法求出线性方程组唯一解.本文可通过文献[18]中提出的借助于 SSOR 迭代求解算法进行大规模 IP 网络下各链路拥塞先验概率 p_j 的近似唯一解求解.

3.2 拥塞链路及其丢包率范围推断

3.2.1 基于 BMAP 定位拥塞链路

在当前时刻推断拥塞链路丢包率范围时,根据当前 1 次 E2E 路径性能探测结果,移除正常 E2E 路径及其途经链路,剩余的拥塞 E2E 路径定义为集合 $\boldsymbol{P}$,在剩余拥塞路径集合 $\boldsymbol{P}$ 中找到具有最小丢包率测量值的路径 P_b;在路径 P_b 中基于 BMAP 准则确定最容易发生拥塞的链路 l_m.即基于当前剩余贝叶斯网模型 G'',根据各拥塞 E2E 路径性能集合 $\boldsymbol{Y}|y_i=1$,推理贝叶斯网模型中的隐藏变量 $\boldsymbol{X}$ 最有可能的一组取值(拥塞链路集合 $\boldsymbol{\chi}|x_j=1$).由贝叶斯原理以及 BMAP 准则可得公式(7).

$$\arg\max P(\boldsymbol{X}\mid\boldsymbol{Y})=\arg\max\frac{P(\boldsymbol{X},\boldsymbol{Y})}{P(\boldsymbol{Y})} \tag{7}$$

其中,$P(\boldsymbol{Y})$仅与路径性能探测结果有关,而与链路选取无关.故公式(7)的求解过程可用公式(8)表示.

$$\arg\max P(\boldsymbol{X}|\boldsymbol{Y})=\arg\max P(\boldsymbol{X},\boldsymbol{Y})=\arg\max\{P(x_j)\cdot P[y_i|p_a(y_i)]\} \tag{8}$$

其中,$p_a(y_i)$为 y_i 的父节点.根据路径 E2E 性能测量结果为正常状态,途经的所有链路状态均正常;拥塞时,路径途经链路至少有 1 条发生拥塞.于是,路径性能观测节点变量与链路性能隐藏节点变量之间存在着如公式(9)所示的概率关系.

$$P(y_i=0|p_a(y_i)=\{0,\ldots,0\})=1,P(y_i=1|\exists x_i=1\cap x_i\in p_a(y_i))=1 \tag{9}$$

为了最大化目标函数,由公式(9)可知,$P(y_i|p_a(y_i))$[19].由于 IP 网络中各链路状态是概率独立的随机变量,对当前 IP 网络进行 1 次 E2E 路径性能探测,推断各链路拥塞概率分布律的过程,服从贝努利概率模型中的二项概率公式 $P_n(k)=C_n^k p^k(1-p)^{(n-k)}$,$n$=1,故待测 IP 网络各链路拥塞全概率公式如公式(10)所示.

$$P(x_j)=\prod_{j=1}^{n_{\varepsilon''}}p_j^{x_j}\cdot(1-p_j)^{(1-x_j)} \tag{10}$$

其中,$n_{\varepsilon''}$为当前时刻E2E拥塞路径集合$\boldsymbol{P}$中途经的链路数.因此,推断最有可能发生拥塞的链路集合,即求解公式(10)中 $P(x_j)$取得最大值时对应 x_j=1 的拥塞链路 l_j,即

$$\arg\max P(\boldsymbol{X}\mid\boldsymbol{Y})=\arg\max P(\boldsymbol{X})=\arg\max\prod_{j=1}^{n_{\varepsilon''}}p_j^{x_j}\cdot(1-p_j)^{(1-x_j)} \tag{11}$$

对公式(11)两边取对数,可得公式(12).

$$\arg\max\sum_{j=1}^{n_{\varepsilon''}}[x_j\cdot\lg p_j+(1-x_j)\cdot\lg(1-p_j)]=\arg\max\sum_{j=1}^{n_{\varepsilon''}}\left[x_j\cdot\lg\frac{p_j}{1-p_j}+\lg(1-p_j)\right] \tag{12}$$

其中,$\lg(1-p_j)$的取值与链路状态 x_j 取值无关,故求解 $\arg\max P(\boldsymbol{X}|\boldsymbol{Y})$即求解 $x_j\cdot\lg[p_j/(1-p_j)]$的最大值.而取得最大值时对应的链路集合,即各拥塞 E2E 路径中最容易发生拥塞的链路组成的集合.因此,公式(12)可简化为公式(13).

$$\arg\max P(X\mid Y)=\arg\max\sum_{j=1}^{n_{\varepsilon''}}\left(x_j\cdot\lg\frac{p_j}{1-p_j}\right) \tag{13}$$

借助于 BMAP 代价值 C_p 的拥塞链路定位策略,引入 C_p 表达式,如公式(14)所示.

$$C_p=\left(\lg\frac{1-p_j}{p_j}\right)\Big/score(l_j) \tag{14}$$

其中,p_j 为链路拥塞学习过程中学习到的各链路 l_j 的拥塞先验概率,$score(l_j)$为当前时刻链路 l_j 在 E2E 拥塞路径中的路径共享数.根据 BMAP 准则,C_p 值最小的链路即 E2E 拥塞路径最容易发生拥塞的链路 l_m.

3.2.2 推断拥塞链路丢包率范围

(1) 路径集合聚类

在推断出 E2E 路径中最容易发生拥塞的链路 l_m 后,CLLRRI 算法提出一种拥塞链路丢包率范围推断新方法.首先,对拥塞路径集合 $\boldsymbol{P}$ 中包含链路 l_m 的路径进行性能(路径丢包率)相近路径聚类,得到聚类路径集合$\boldsymbol{\Omega}$.在确定性能相近路径时,算法做出如下定义.

定义 2. 若包含同一条链路 l_m 的两条路径 P_1 与 P_2、两路径丢包率差值的绝对值$|\Phi(P_1)-\Phi(P_2)|<0.05$,则路径 P_1 与 P_2 相关且性能相似.

根据公式(2),可得出路径丢包率与途经各链路丢包率之间的关系表达式(15).不难看出,E2E 路径中每增加一条拥塞链路(通常,以丢包率≥0.05 的链路界定为拥塞链路[1]),路径丢包率至少增加约 0.05.

$$\Phi(P_i)=1-\prod_{j=1}^{n_c}[1-\varphi(l_j)] \tag{15}$$

其中,$\Phi(P_i)$为路径丢包率,$\varphi(l_j)$为路径途经各链路的丢包率,n_c为某E2E路径P_i途经链路数.为了确保推断出拥塞链路 l_m 后,包含链路 l_m 路径中其他拥塞链路能被继续推断,而不以最小链路覆盖集理论[1,8]移除该链路所在路径,带来推断误差,本文提出的CLLRRI算法中,将途经链路 l_m 的各路径丢包率与路径 P_b 的丢包率数值之差大于阈值 0.05 的路径存放至另一个路径集合$\boldsymbol{\Omega}'$中.

另外,若聚类路径集合$\boldsymbol{\Omega}$中存在不止一条具有相同最小 C_p 值的链路,则根据初始待测 IP 网络中各链路 l_j 的共享路径数 $num(l_j)$值做进一步判断,将具有较大 num 值的链路确定为最容易拥塞的链路 l_m.

(2) 性能相似系数确定

为了避免 Range tomography 算法中需要通过大量实验确定 a-similar 系数的繁琐过程[1],CLLRRI 算法中提出一种利用聚类路径集合$\boldsymbol{\Omega}$中各 E2E 路径丢包率值,在线动态计算路径性能相似系数δ的方法.δ的求解方法如公式(16)所示.

$$\delta=\frac{\arg\max_{P_i\in\Omega}\Phi(P_i)-\arg\min_{P_i\in\Omega}\Phi(P_i)}{\arg\min_{P_i\in\Omega}\Phi(P_i)} \tag{16}$$

其中,$\arg\max_{P_i\in\Omega}\Phi(P_i)$ 和 $\arg\min_{P_i\in\Omega}\Phi(P_i)$ 分别为聚类路径集合$\boldsymbol{\Omega}$中丢包率的最大值和最小值.

(3) 链路丢包率范围推断

聚类路径集合$\boldsymbol{\Omega}$中,链路 l_m 的丢包率推断范围 $range(\varphi_{lm})$满足公式(17)所示不等式.

$$\arg\min_{P_i\in\boldsymbol{\Omega}}\Phi(P_i)\leqslant range(\varphi_{lm})\leqslant\arg\max_{P_i\in\boldsymbol{\Omega}}\Phi(P_i) \tag{17}$$

聚类路径集合$\boldsymbol{\Omega}$中,各 E2E 路径性能平均值 $\bar{\Phi}_{P_i}=avg\{\Phi_{P_i}\mid l_m\in P_i,P_i\in\boldsymbol{\Omega}\}$,且满足公式(18).

$$\arg\min_{P_i\in\boldsymbol{\Omega}}\Phi(P_i)\leqslant\bar{\Phi}_{P_i}\leqslant\arg\max_{P_i\in\boldsymbol{\Omega}}\Phi(P_i) \tag{18}$$

将公式(17)、公式(18)代入公式(16),可得公式(19)所示不等式关系.

$$\frac{|range(\varphi_{lm})-\bar{\Phi}_{P_i}|}{\arg\min(range(\varphi_{lm}),\bar{\Phi}_{P_i})}\leqslant\delta \tag{19}$$

由公式(19),拥塞链路 l_m 的丢包率范围可表示为公式(20).

$$range(\varphi_{lm})=\left[\frac{\bar{\Phi}_{P_i}}{1+\delta},(1+\delta)\bar{\Phi}_{P_i}\right] \tag{20}$$

证明:

(a) 当 $range(\varphi_{lm})\geqslant\bar{\Phi}_{P_i}$ 时,由公式(19),$\frac{range(\varphi_{lm})-\bar{\Phi}_{P_i}}{\bar{\Phi}_{P_i}}\leqslant\delta$, 整理后得 $range(\varphi_{lm})\leqslant(1+\delta)\bar{\Phi}_{P_i}$;

(b) 当 $range(\varphi_{lm})\leqslant\bar{\Phi}_{P_i}$ 时,由公式(19),$\frac{\bar{\Phi}_{P_i}-range(\varphi_{lm})}{range(\varphi_{lm})}\leqslant\delta$, 整理后得 $range(\varphi_{lm})\geqslant\frac{\bar{\Phi}_{P_i}}{1+\delta}$.

综合上述(a),(b),可得 $\frac{\bar{\Phi}_{P_i}}{1+\delta}\leqslant range(\varphi_{lm})\leqslant(1+\delta)\bar{\Phi}_{P_i}$, 公式(20)得证. □

(4) 路径及其丢包率更新

由于大规模 IP 网络存在多链路拥塞现象,为了能够推断拥塞 E2E 路径是否存在其他拥塞链路,对集合$\boldsymbol{\Omega}$中的路径进行判断,方法如下:1) 若集合$\boldsymbol{\Omega}$中的路径丢包率落在推断出的链路 l_m 丢包率范围内,则意味着此路径不再包含其他拥塞链路,将此路径从集合$\boldsymbol{\Omega}$中移除,不再进入下一轮推断,并移除此路径途经的所有链路;2) 对集合$\boldsymbol{\Omega}$及$\boldsymbol{\Omega}'$中路径丢包率超出推断出链路 l_m 的丢包率范围的路径,进行丢包率更新,更新策略如公式(19)所示.

$$\Phi_{P_i}=\Phi_{P_i}-\bar{\Phi}_{P_i},l_m\in P_i,P_i\in\boldsymbol{P} \tag{21}$$

其中,$\bar{\Phi}_{P_i}$ 为当前路径集合$\boldsymbol{\Omega}$中各 E2E 路径性能平均值.通过路径移除以及路径丢包率更新后,再次推断剩余路径集合 $\boldsymbol{P}$ 中的拥塞链路及其丢包率范围.如此循环,直到路径集合 $\boldsymbol{P}$ 为空集∅时,推断过程结束.

由于 CLLRRI 算法在进行拥塞链路 l_m 丢包率范围推断时引入了性能相似路径聚类策略,因此避免了拥塞链路丢包率范围推断过大造成对多链路拥塞 E2E 路径的直接移除,保证了拥塞 E2E 路径中不止一条链路拥塞时,拥塞链路及其性能范围能够被再次推断,有效避免了基于最小链路覆盖集理论造成的拥塞链路推断误差,提高了算法推断性能.

3.3 时间复杂度分析

本文提出的 CLLRRI 算法与 CLINK 及 Range tomography 算法均为 NP-hard[1,20,21].CLLRRI 算法主要包括链路拥塞先验概率求解运算和拥塞链路定位及其丢包率范围推断两部分.如对待测 IP 网络先验概率学习过程中,拥塞路径总数为 n_θ,拥塞路径途经的链路总数为 $n_{\varepsilon'}$,即需要获取的链路拥塞先验概率 p_j 的总数为 $n_{\varepsilon'}$.通常,由于拥塞 E2E 路径探测数小于所有途经这些拥塞路径的链路数,先验概率求解线性方程组系数矩阵欠定,通过拥塞路径两两关联补满秩,可得到 $n_\theta(n_\theta+1)/2$ 个约束条件.由于 IP 网络拓扑结构具有幂率特性,$n_\theta(n_\theta+1)/2>>n_{\varepsilon'}$,则求解各链路拥塞先验概率 p_j 的时间复杂度为 $O(n_\theta\log n_\theta)$.另外,由于 CLLRRI 算法在拥塞链路定位及其丢包率范围推断中采用贪婪启发式搜索算法,是一种 $\log(n_{\varepsilon''}+1)$近似算法,故其计算时间复杂度为 $O(n_{\varepsilon''}n_{\delta''})$.其中,$n_{\delta''}$为推理时刻拥塞路径数,$n_{\varepsilon''}$为拥塞路径途经的链路数[1,8].在 300 个节点(527 条链路)的网络模型下,CLLRRI 算法拥塞链路定位及其丢包率范围推断,算法整体运行时间不超过 2s.

3.4 算法流程及伪代码

CLLRRI 算法流程如下.

1) 借助 N 次 E2E 路径性能探测结果,学习待测 IP 网络各链路拥塞先验概率 p_j.
2) 根据当前时刻各 E2E 路径丢包率,在最小丢包率路径 p_b 中,基于 BMAP 准则,推断 p_b 中最有可能发生拥塞的链路 l_m.
3) 聚类包含 l_m 且性能相近的路径集合Ω.
4) 利用Ω中的路径性能相似系数δ推断Ω丢包率范围.
5) 判断Ω中各路径丢包率是否落在 l_m 推断的丢包率范围内:如果在,则移除此路径;如果路径丢包率不在 l_m 的丢包率范围内,则对该路径途经的其他链路继续进行下一轮拥塞推断,不移除此路径,进入第 6)步路径丢包率更新.
6) 更新包含 l_m 的路径丢包率.
7) 返回第 2)步,循环推断拥塞链路及其丢包率范围,直到拥塞路径集合为空,算法结束.

CLLRRI 算法伪代码如下.

CLLRRI 算法.

输入:学习过程中得到的各链路拥塞先验概率 p_j;

当前时刻待测 IP 网络中各 E2E 路径途经链路 l_j 所在的路径数 $num(l_j)$;

当前时刻各拥塞 E2E 路径集合 $\boldsymbol{P}$,正常 E2E 路径集合 $\bar{\boldsymbol{P}}$;拥塞 E2E 路径途经的链路集合$\boldsymbol{\Gamma}$;

当前时刻各拥塞 E2E 路径丢包率观测值 m_{p_i}.

输出:推断出的拥塞链路及其性能范围集合$\boldsymbol{\chi}$.

1: Initialize $\boldsymbol{\chi}$=∅; {初始化拥塞链路性能集合}

2: Initialize $\Phi_{p_i}=m_{p_i}$, $P_i\in\boldsymbol{P}$; {各拥塞路径丢包率赋初值(当前时刻各 E2E 路径丢包率测量值)}

3: Update $\boldsymbol{P}=\boldsymbol{P}-\bar{\boldsymbol{P}},\boldsymbol{\Gamma}=\boldsymbol{\Gamma}-\{l_j\mid l_j\in\bar{\boldsymbol{P}}\}$; {移除推断过程中正常路径及途经的链路}

4: **while** $\boldsymbol{P}\neq\varnothing$ **and** $\boldsymbol{\Gamma}\neq\varnothing$ **do**

5: $P_b=\arg\min_{\forall P_i\in\boldsymbol{P}}\{\Phi_{P_i}\}$; {$P_b$ 为 $\boldsymbol{P}$ 中具有最小丢包率的路径}

6: **for each** link $l_j\in P_b$ do {判断路径 P_b 中所有链路}

7: $score(l_j)=number(P_i)|_{l_j\in P_i,P_i\in\boldsymbol{P}}$; {当前时刻包含链路 l_j 的路径数}

8: **end for**

9: $C_p(l_k)=\log\left(\frac{1-p_k}{p_k}\right)\Big/score(l_k)$; {根据学习过程中各链路 l_k 拥塞先验概率 p_k,计算当前时刻各链路 l_k 的 BMAP 代价值 $C_p(l_k)$}

10: $\boldsymbol{\Omega}=\{P_i|P_i\in\boldsymbol{P},P_i||P_b,l_j=\arg\min C_p(l_k),l_j\in P_b,P_i-P_b<\varepsilon\}$; {找到与路径 P_b 相关且性能相近(路径丢包率差小于阈值ε=0.05)的路径集合}

11: $\boldsymbol{\Omega}'=\{P_i|P_i-P_b>\varepsilon\}$; {将路径丢包率差大于阈值$\varepsilon$=0.05 的路径存入集合$\boldsymbol{\Omega}'$}

12: $L_m=\arg\min_{l_k\in\Omega}C_p(l_k)$; {集合$\boldsymbol{\Omega}$中 $C_p(l_k)$最小值的链路 l_k 即为待推断丢包率范围的链路}

13: $l_m=\arg\min_{l_k\in L_m}num(l_k)$; {如有多条链路 $C_p(l_k)$最小值相同,根据 $num(l_k)$确定}

14: $\delta=\dfrac{\arg\max_{P_i\in\boldsymbol{\Omega}}\Phi(P_i)-\arg\min_{P_i\in\boldsymbol{\Omega}}\Phi(P_i)}{\arg\min_{P_i\in\boldsymbol{\Omega}}\Phi(P_i)}$; {在共享链路 l_m 的路径集合$\boldsymbol{\Omega}$中,求解各条路径的性能相似系数δ}

15: $\bar{\Phi}_{P_i}=avg\{\Phi_{P_i}\mid l_m\in P_i,P_i\in\boldsymbol{\Omega}\}$; {求解途经链路 l_m 的路径丢包率平均值}

16: $range(\varphi_{lm})=\left[\bar{\Phi}_{P_i}\left(\frac{1}{1+\delta}\right),\bar{\Phi}_{P_i}(1+\delta)\right]$; {链路 l_m 的丢包率范围被确定}

17: $\boldsymbol{\chi}=\boldsymbol{\chi}+range(\varphi_{l_m})$; {将已确定丢包率范围的链路放入集合$\boldsymbol{\chi}$中}

18: **if** $\boldsymbol{\Omega}'\neq\varnothing$ {若存在路径相关但性能不相似的路径}

19: Update $\Phi_{P_i}=\Phi_{P_i}-\bar{\Phi}_{P_i},l_m\in P_i,P_i\in\boldsymbol{P}$; {更新路径丢包率}

20: Update $P_i=P_i-l_m$, $l_m\in P_i$, $P_i\in\boldsymbol{P}$; {移除路径中已推断出的链路 l_m}

21: **end if**

22: $\boldsymbol{P}=\boldsymbol{P}-\{P_i\mid P_i\in\boldsymbol{P},\Phi_{P_i}\in range(\varphi_{lm})\}$; {从路径集合 $\boldsymbol{P}$ 中移除部分途经链路 l_m 的路径(路径丢包率在推断链路丢包率范围内)}

23: $\boldsymbol{\Gamma}=\boldsymbol{\Gamma}-\{l\},l\in\{domain(l_m\mid l_m\in P_i,P_i\in\boldsymbol{P},\Phi_{P_i}\in range(\varphi_{lm}))\}$; {更新剩余链路集合,移除链路 l_m 及路径途经的其他链路}

24: $\Phi_{P_i}=\max\{0,\Phi_{P_i}-\bar{\Phi}_{P_i}\},\forall P_i\in\boldsymbol{P},l_m\in P_i$; {更新包含链路 l_m 的路径丢包率}

25: **end while**

4 实验评价

通常,评价算法性能的实验方法有 3 种:模拟实验、仿真实验和实测实验.其中,模拟实验方法标准答案(benchmark)已知,实验细节可以完全掌握,但缺点是不够真实;仿真实验方法兼顾实验的可控性与真实性,操作灵活;实测实验环境真实,但较难获取 benchmark.因此,为了客观地评价算法的性能表现,本文分别采用模拟实验及 Emulab 仿真实验,对 CLLRRI 算法与 CLINK 以及 Range tomography 算法进行性能比较评价.

4.1 模拟实验评价

对 Internet 拓扑结构的研究,前人已经做了很多工作,从最初的 Waxman 随机图模型[22]到 Albert 等人[23]提

出的无标度网络模型 BA,再到 Bu 等人提出的基于节点度的幂率分布特征模型 GLP[24],人们都试图去发现和解析 Internet 拓扑演化的规律.因此,为了验证推理算法的有效性及准确性,本文借助 Brite 拓扑生成器[25],分别生成 3 种不同类型、规模的 IP 网络拓扑模型.其中,Waxman 模型中的节点度数值随着节点数量的增加而增加,但无法生成节点众多但节点平均度值较小的网络.因 IP 网络规模的不断扩大,新路由器节点加入通常倾向于与具有高度数值的"大节点"连接.BA 及 GLP 正是基于这两个特征构造的具有度分布呈幂率特征的无标度网络模型.3 种拓扑网络模型均体现了 Internet 特性.为了更好地验证本文提出算法在不同网络环境中的拥塞链路推理性能,在 Eclipse MARS.1 平台下,将 3 种不同网络模型拓扑文件导入构建模拟待测 IP 网络.

4.1.1 验证方法及场景设置

用丢包率模型 LM1[9]仿真 IP 网络中各链路拥塞事件.由于 IP 网络中各拥塞链路的丢包率通常不超过 0.2[1],因此在模拟真实 IP 网络环境的实验中,设置各拥塞链路的丢包率在[0.05,0.2]范围之间波动[1],正常链路丢包率范围为[0,0.01].一旦链路被分配丢包率,则模拟实际 IP 网络链路丢包服从 Bernoulli 随机过程产生丢包事件.

由于实际 IP 网络单时隙路径性能探测中,各 E2E 路径性能测量无法保证时钟同步,因此,同一链路在不同 E2E 路径中的丢包率也不尽相同.但是由于链路状态有一定的持续性[26],不同 E2E 路径中的同一条链路,其丢包率在较短时间内变化不大.因此,本文利用随机数发生器,根据拥塞链路比例 f,在 IP 网络拓扑模型中随机产生一定数目的拥塞链路,并给拥塞链路赋丢包率,丢包率数值为[0.05,0.2]之间任意一个随机数值.由公式(2)可得出各 E2E 路径的丢包率数值,以此来模拟 E2E 路径性能测量结果.

在进行拥塞链路推断时,由于对各 E2E 路径发包探测的间隔时间较短,如某链路为拥塞状态,在包含此链路的路径中,其丢包率相差不大.因此,为了模拟同一条链路在不同 E2E 路径中的丢包情况,在同一次拥塞事件中,E2E 路径中的共享链路丢包率在[–0.02,0.02]之间随机发生改变.

为了验证本文提出的 CLLRRI 算法在拥塞链路定位及其丢包率范围推断中的有效性及准确性,利用公式(22)所示检测率(detection rate,简称 DR)、误报率(false positive rate,简称 FPR)以及丢包率范围推断准确率(accuracy)对不同算法的推断性能进行评价,结果均为实验场景及设置参数不变的情况下,10 次实验取平均值后得出.

$$DR=\frac{F\cap X}{F},FPR=\frac{X\setminus F}{X},Accuracy=\frac{Q}{F\cap X} \tag{22}$$

其中,F 代表实际拥塞链路(benchmark),X 代表算法推断出的拥塞链路,Q 代表算法准确推断出拥塞链路丢包率范围的链路数目.

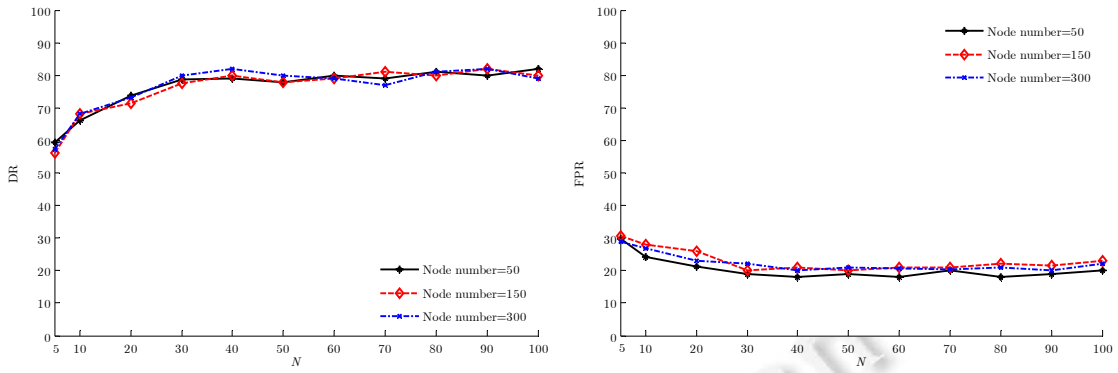
4.1.2 学习过程中 N 值的选取

链路拥塞先验概率学习过程中,为了能在保证推断性能的基础上尽量减少算法开销,N 值的取值非常重要.为了验证不同的 E2E 路径探测次数对本文提出 CLLRRI 算法性能的影响,分别在相同拥塞链路比例 f、不同网络规模(node number)以及相同网络规模、不同 f 条件下进行实验,对算法中不同 N 值选取时的拥塞链路定位性能进行评价,从而确定 CLLRI 算法中参数 N 的取值大小.

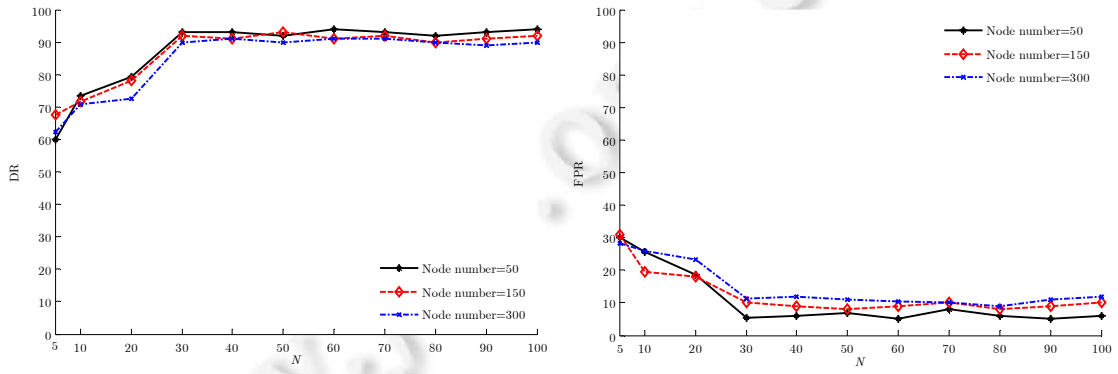
(1) 不同网络规模下的 N 值选取(f=0.2)

利用 Brite 拓扑生成器模拟不同网络规模,节点数(node number)分别为 50,150,300 的 Waxman,BA 及 GLP 网络拓扑模型,选取 N=5~100,以拥塞链路比例 f=0.2 随机产生拥塞链路.利用本文提出的 CLLRRI 算法,通过改变不同 N 值,得到 N 次 E2E 路径探测结果,借助公式(6)获取各链路拥塞先验概率,并对当前时刻拥塞链路进行定位.在 3 种不同类型的拓扑模型下,CLLRRI 算法拥塞链路定位 DR 及 FPR 如图 3 所示.

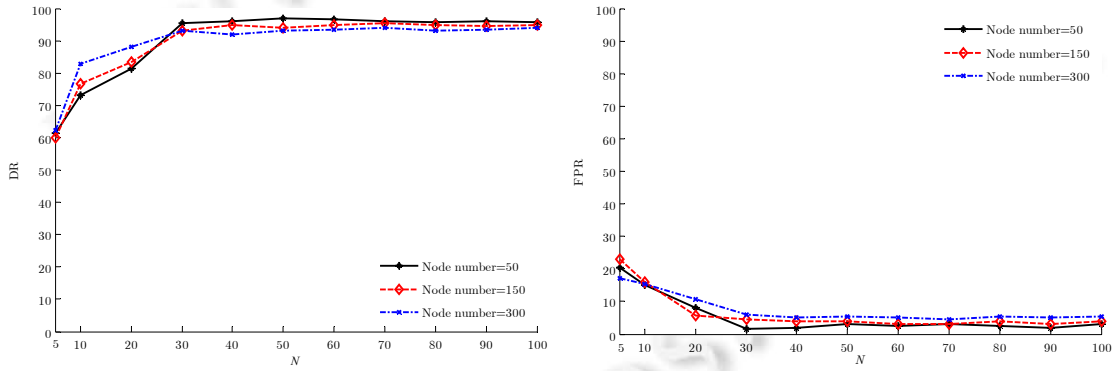
如图 3 所示,在不同拓扑模型下,当 N>30 后,CLLRRI 算法的推断性能比较稳定,说明通过多次 E2E 路径的性能探测,待测网络各链路的拥塞规律能够被较好地学习,作为贝叶斯推理的先验知识,实现对拥塞链路的准确定位.



(a) Waxman 模型



(b) BA 模型



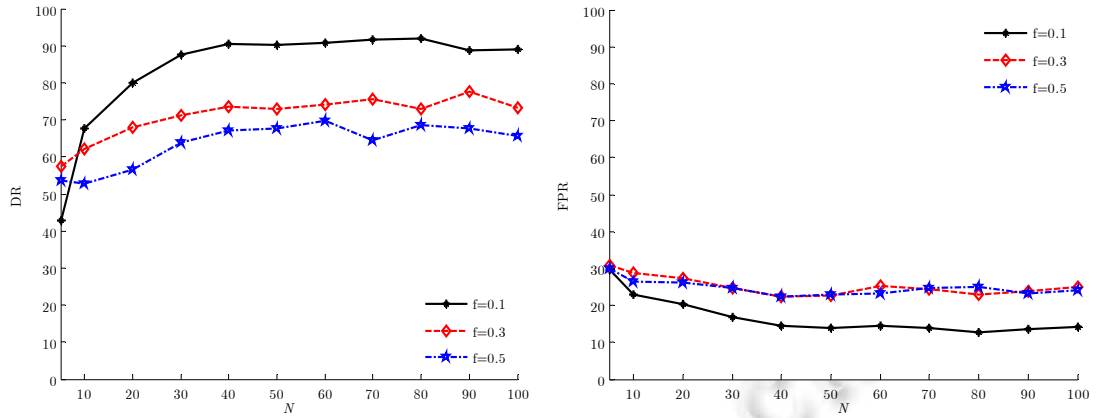
(c) GLP 模型

Fig.3 Congested link inference performance comparisons of CLLRRI algorithm under different values of N (different network scale)

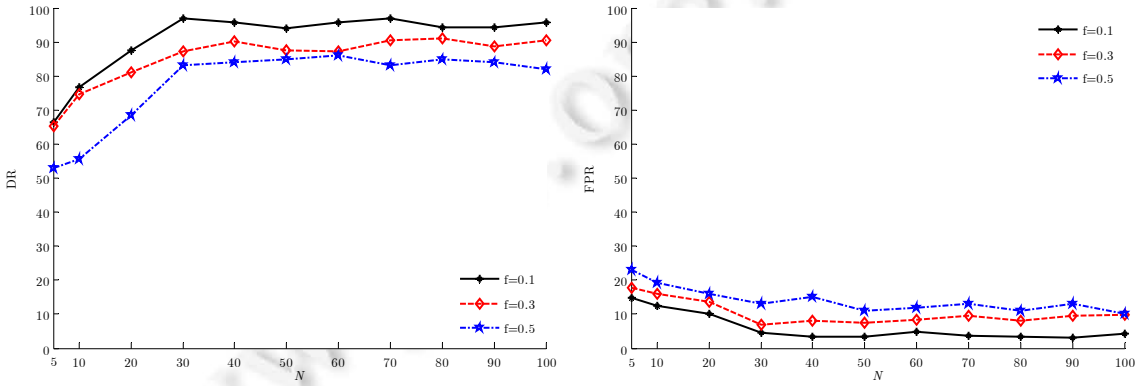
图 3 不同 N 值选取下,CLLRRI 算法的拥塞链路推断性能比较(不同网络规模)

(2) 不同 f 下的 N 值选取(*node number*=100)

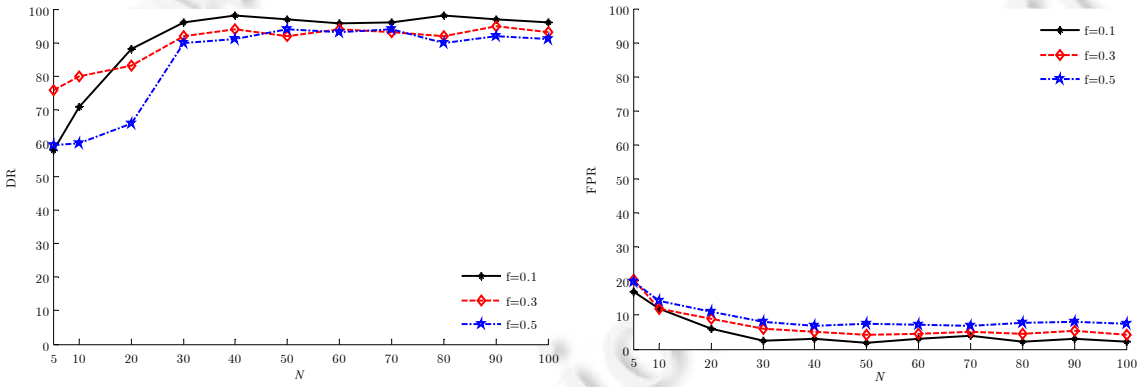
利用 Brite 拓扑生成器模拟 *Node number*=100 网络规模下的 Waxman,BA 及 GLP 拓扑模型,选取 N=5~100,f 分别以 0.1,0.3,0.5 的比例随机发生链路拥塞事件.利用本文提出的 CLLRRI 算法,通过改变不同 N 值,得到 N 次 E2E 路径探测结果.同样,借助公式(6)获取各链路拥塞先验概率,对当前时刻拥塞链路进行定位.在 3 种不同类型的拓扑模型下,CLLRRI 算法拥塞链路定位 DR 及 FPR 分别如图 4 所示.



(a) Waxman 模型



(b) BA 模型



(c) GLP 模型

Fig.4 Congested link inference performance comparisons of CLLRRI algorithm under different values of N (different congested link ratio f)

图 4 不同 N 值选取下,CLLRRI 算法的拥塞链路推断性能比较(不同拥塞链路比例 f)

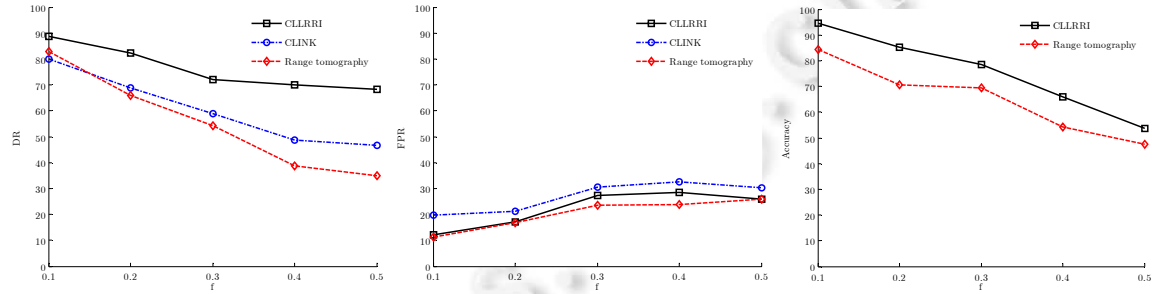
如图 4 所示,在不同拓扑模型下,N>30 均能有效地学习各链路拥塞规律,实现对推理时刻各拥塞链路的定位.因此,在后续的算法性能比较中,本文提出的 CLLRRI 算法以 N=30 作为学习过程中各 E2E 路径性能探测次数.

4.1.3 算法推断性能比较

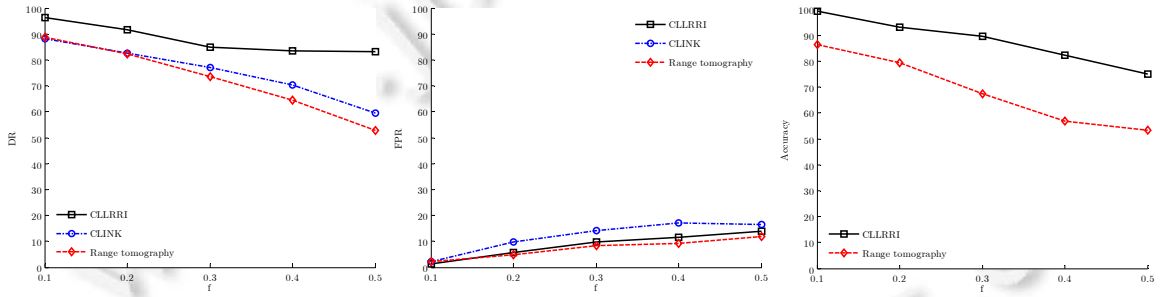
(1) 不同拥塞链路比例下算法推断性能比较

为了验证本文提出的 CLLRRI 算法在不同拥塞链路比例 f 下的推断性能,利用 Brite 拓扑生成器分别模拟 100 个节点的 Waxman,BA 及 GLP 网络拓扑,设置 f=[0.1~0.5]随机产生拥塞链路.

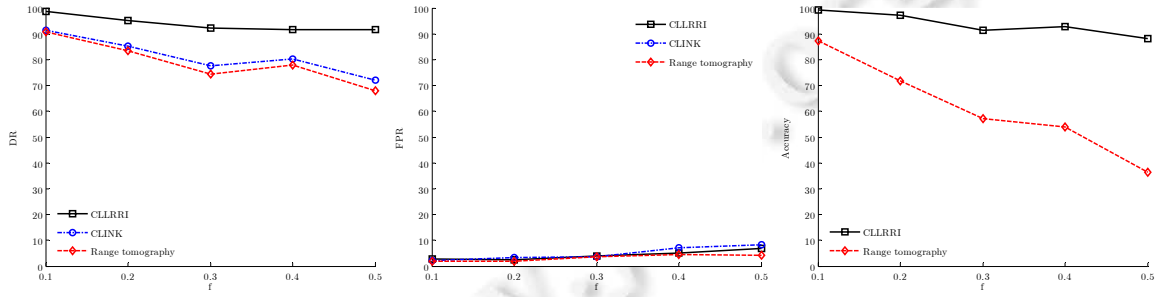
首先,根据连续 30 次 E2E 路径探测的链路拥塞事件,得到每次各 E2E 路径的拥塞情况,求出待测 IP 网络中各链路的拥塞先验概率.然后,根据当前链路拥塞事件产生的 E2E 路径拥塞测量值,分别利用 CLINK,Range tomography 及本文提出的 CLLRRI 算法进行拥塞链路及其丢包率范围推断,得出各算法性能,如图 5 所示.由于 CLINK 算法仅能定位出拥塞链路而无法进行拥塞链路丢包率范围推断,因此,在拥塞链路丢包率范围推理准确率 Accuracy 性能曲线图中缺少 CLINK 算法的性能曲线.



(a) Waxman 模型



(b) BA 模型



(c) GLP 模型

Fig.5 Inference performance comparisons among algorithms of CLLRRI, CLINK, and Range tomography under different congested link ratio f (*node number*=100)

图 5 不同拥塞链路比例 f 下,CLLRRI 算法与 CLINK 及 Range tomography 算法的推断性能比较 (*node number*=100)

1) 拥塞链路定位性能分析

由图 5 中的 DR 及 FPR 可以看出,对 3 种不同的拓扑网络模型 Waxman,BA 及 GLP,3 种算法都是在 GLP 模型下的推断性能最好,其次是 BA 模型,Waxman 模型最差.这与网络拓扑模型的结构有很大关系,由于 GLP 及 BA 均为幂率特性模型,且 GLP 为强幂率特性模型,E2E 路径长度较短,共享路由器的节点数较多(部分路由器度值较大).而 Waxman 模型路径长度较长,因途经链路数较多导致算法推断性能下降明显.本文提出的 CLLRRI 和 CLINK 算法均对待测 IP 网络各链路进行了拥塞先验知识学习,而 Range tomography 算法仅以“E2E 路径中共享数目最多的瓶颈链路为最有可能发生拥塞的链路”这一经验知识进行拥塞链路定位,未充分考虑各待测 IP 网络实际的链路拥塞情况.因此,多链路拥塞时,Range tomography 算法的推断性能下降明显.在 Waxman 模型中,CLLRRI 算法 DR 比 Range tomography 算法高 10%以上,并且随着 f 的增大,CLLRRI 算法鲁棒性较强,当 f 达到 0.3 时,比 Range tomography 高出 20%以上;而当 f 达到 0.5 时,高出近 40%.CLLRRI 算法在 BA 及 GLP 模型下的推断性能中,鲁棒性较强,未出现明显下降,特别是在 GLP 模型下,当 f 达到 0.5 时,DR 始终保持在 92%以上,与 f=0.1 时(平均 DR=98%)相比,下降不超过 10%,显示出 CLLRRI 算法具有较高的鲁棒性.同样,3 种算法的 FPR 均在 GLP 模型下最低,始终保持在 10%以下,在 BA 模型下不超过 20%,在 Waxman 模型下不超过 30%.

CLLRRI 算法及 CLINK 算法均通过多时隙路径性能探测,获取了各链路拥塞先验知识,对拥塞链路进行定位.但是由于 CLINK 算法基于最小链路覆盖集理论(在推断出 E2E 路径中的一条拥塞链路后,即将此路径移除,不再推断该路径是否还存在其他拥塞链路),而 CLLRRI 算法借助 E2E 路径丢包率,通过贪婪启发式定位 E2E 路径中可能发生拥塞的链路,推断性能与 CLINK 算法相比有显著提高.

2) 拥塞链路丢包率推断性能分析

为了验证 CLLRRI 算法对拥塞链路丢包率范围推断的准确性(accuracy),在不同类型的网络模型下,CLLRRI 及 Range tomography 算法的 Accuracy 性能比较结果如图 5 中的 Accuracy 所示.CLLRRI 算法在不同 f 下,Accuracy 始终高于 Range tomography 算法,特别是在 GLP 模型下,CLLRRI 算法的 Accuracy 始终保持在 95%左右.随着 f 的增加,Range tomography 算法推断性能下降明显,当拥塞链路比例达到 0.5 时,Range tomography 算法的 Accuracy 不足 50%.而本文提出的 CLLRRI 算法始终保持了较强的鲁棒性,特别是在 GLP 模型下,随着 f 的增大,Accuracy 降低不明显,当 f 达到 0.5 时,Accuracy 仍高达 90%左右.

CLINK 和 Range tomography 算法均基于最小链路覆盖集理论,CLINK 算法在确定了某 E2E 路径中的拥塞链路后,不再推断该路径中其他链路;Range tomography 算法仅以“E2E 路径中共享数目最多的瓶颈链路为最有可能发生拥塞的链路”这一专家先验知识作为拥塞链路推断的唯一依据,认为具有最小丢包率数值的拥塞路径,是由其中一条链路拥塞造成的[1],当拥塞链路数较多时,推断性能下降明显.CLLRRI 算法基于 BMAP 准则,定位 E2E 拥塞路径中最容易发生拥塞的链路,推断丢包率范围,避免了 E2E 路径中存在多条拥塞链路,CLINK 算法无法推断以及 Range tomography 算法错误移除导致的推断误差.

(2) 不同网络规模下算法推断性能比较

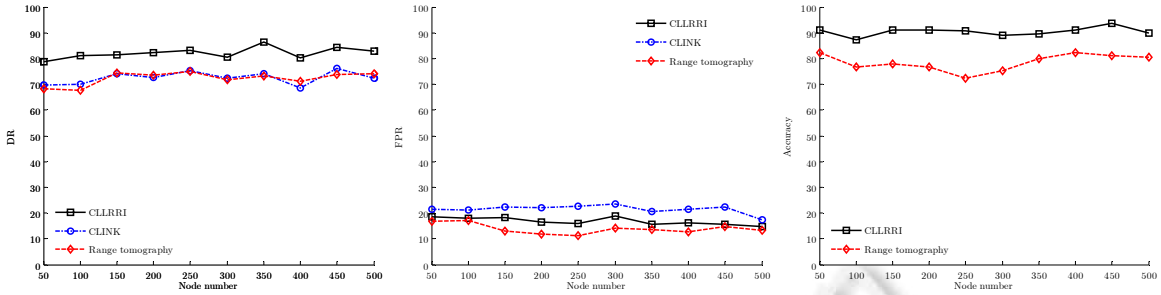
为了验证 CLLRRI 算法在不同网络规模下的推断性能,利用 Brite 拓扑生成器分别生成 50~500 个节点网络规模的 Waxman,BA 及 GLP 模型,并设置 f=0.2,随机产生网络中各拥塞链路.利用 CLINK,Range tomography 以及本文提出的 CLLRRI 算法分别进行拥塞链路推断,不同网络拓扑模型下,3 种算法的 DR,FPR 分别如图 6 中的 DR,FPR 所示.CLLRRI 和 Range tomography 算法拥塞链路丢包率范围推理准确率如图 6 中的 Accuracy 所示(不同网络规模下).

由图 6 中的 DR,FPR 可以看出,随着网络规模的增大,在不同网络拓扑模型下,3 种算法推断性能虽然有一定的下降趋势(DR 降低,FPR 升高),但从总体来看,变化不明显.这说明网络规模增大对各算法推断性能影响不大.

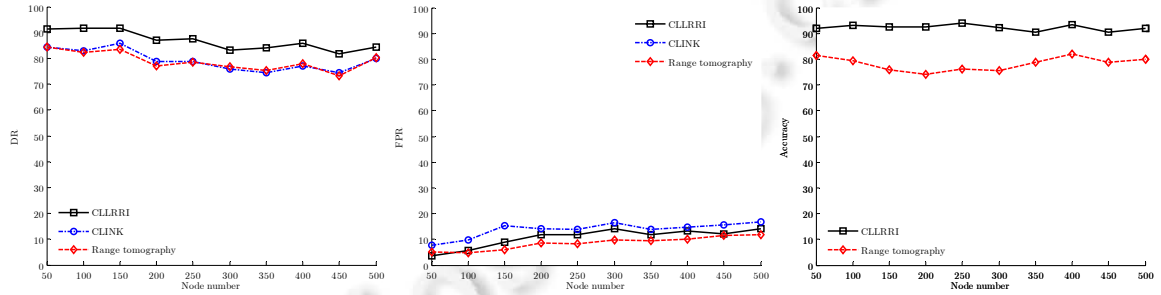
在不同的网络拓扑模型下,Range tomography 算法和 CLLRRI 算法推断拥塞链路丢包率的范围准确率如图 6 中 Accuracy 所示.CLLRRI 算法的 Accuracy 均高于 Range tomography 算法,且随着网络规模的增大,算法 Accuracy 始终保持稳定,验证了 CLLRRI 算法在不同网络规模下,特别是在大规模 IP 网络中,有较高的推断性能.

通过不同网络节点规模下的模拟实验发现,待测网络节点规模的改变对算法性能的影响较小.限于篇幅,这

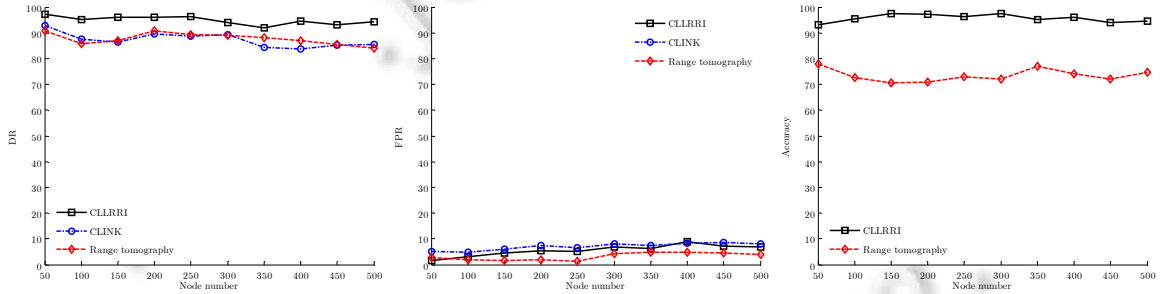
里不再对不同网络规模、不同拥塞链路比例 f 下的算法推断性能进行比较实验验证.



(a) Waxman 模型



(b) BA 模型



(c) GLP 模型

Fig.6 Inference performance comparisons among algorithms of CLLRRI, CLINK, and Range tomography (different network scale)

图 6 CLLRRI 算法与 CLINK 及 Range tomography 算法推断性能比较(不同网络规模)

4.2 仿真实验评价

4.2.1 实验场景设置

实际 IP 网络 AS(autonomous systems)内部链路性能很难获知,研究发现,大多数 IP 网络拓扑服从幂率规则[27].因此,为了验证 CLLRRI 算法在实际网络中的拥塞链路推断性能,在 Emulab 仿真实验平台上设计了 IP 网络仿真实验场景:(1) 使用 Brite 拓扑生成器生成 20 个节点的服从幂率规则的 GLP 拓扑模型 20GLP.brite;(2) 在 Emulab 仿真实验平台上导入此 Brite 文件,搭建被测 IP 网络,对网络中的各叶子节点路由器部署探针,并将性能监控台接入被测 IP 网络;(3) 由性能监控台给各探针下达测量任务,利用 traceroute 测出各 E2E 路径途经各链路,利用 ping 测出各 E2E 路径丢包率值,并上传至性能监控台.

分别利用本文提出的 CLLRRI 算法、CLINK 算法和 Range tomography 算法对当前时刻拥塞链路及其丢包率范围进行推断.仿真实验流程如图 7 所示.

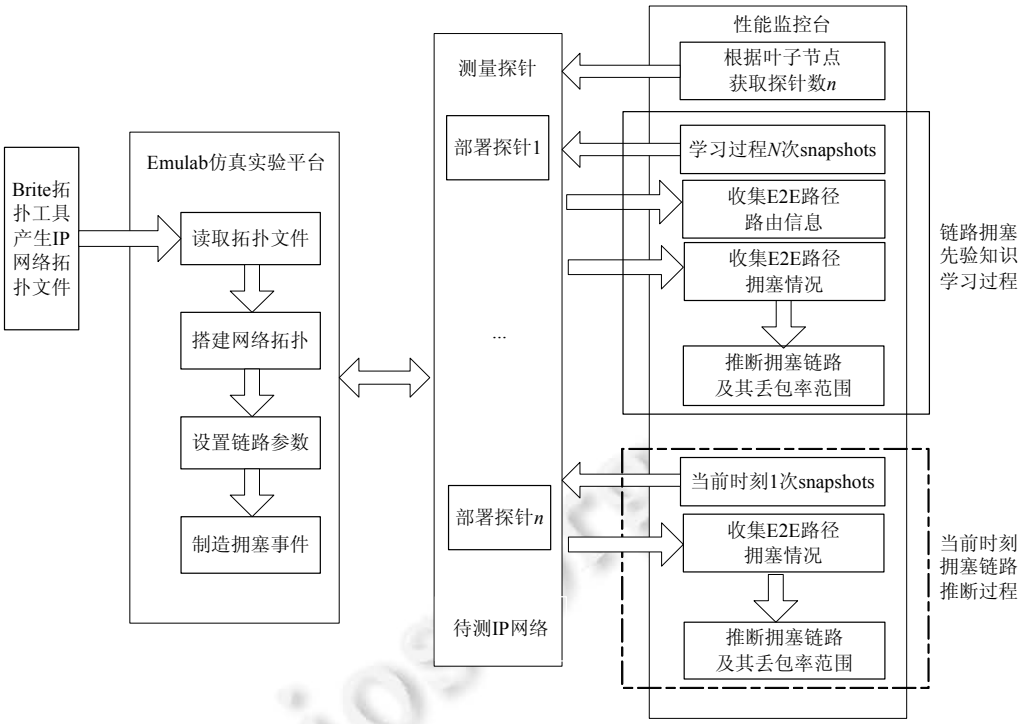


Fig.7 Flow chart of simulation experiment

图 7 仿真实验流程图

首先将拓扑文件读入 Emulab 仿真实验平台生成 IP 网络拓扑结构,设置各链路带宽 100Mbps,时延 15ms;AS 内部为 OSPF 协议,E2E 路径服从最短路径优先选路规则;链路丢包采用 LM1 丢包模型[9],当链路拥塞时,丢包率数值服从[0.05,0.2]均匀分布;当链路正常时,丢包率数值服从[0,0.01]均匀分布.为每条链路赋初始丢包率后,链路丢包服从 Bernoulli 随机过程,每隔 2min 以比例 f 随机产生链路丢包事件.

4.2.2 实验结果分析

为了验证拥塞链路推断算法在 Emulab 仿真实验平台下的推断性能,设置 f=[0.1,0.5].利用 3 种算法分别推断当前时刻拥塞链路及其丢包率范围,并与此时的 benchmark(即链路丢包率大于 0.05 的链路集合)进行比较,验证算法性能.每组实验结果均在相同参数设置情况下,实验 10 次取平均后得出,实验比较结果如图 8 所示.

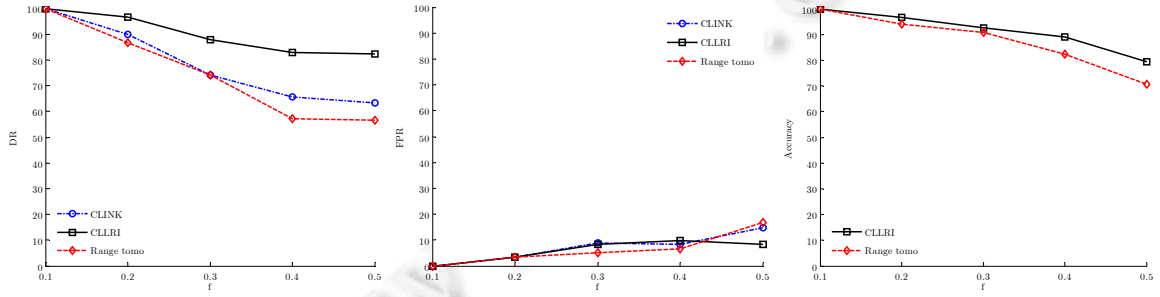


Fig.8 Inference performance comparisons among algorithms of CLLRRI, CLINK, and Range tomography on the Emulab platform under different congested link ratio f (*node number*=20)

图 8 Emulab 仿真实验平台,不同拥塞链路比例 f 下,CLLRRI 算法与 CLINK 及 Range tomography 算法推断性能比较(*node number*=20)

由图 8 中的 DR 可以看出,在 Emulab 仿真实验中,3 种算法的 DR 均随着 f 的增大呈下降趋势.其中,CLLRRI 算法的 DR 最高,在 f 不超过 0.2 的情况下,DR 始终保持 97%以上,CLINK 和 Range tomography 算法 DR 也均在 90%左右,但随着 f 的增大下降明显;当 f=0.5 时,CLLRRI 算法的 DR 仍保持在 85%以上,而 CLINK 算法不足 65%,Range tomography 算法不足 60%.如图 8 所示,本文提出的 CLLRRI 算法的 FPR 不超过 10%,且随着 f 的增大保持较高的稳定性.由图 8 中的 Accuracy 可以看出,CLLRRI 和 Range tomography 算法均实现了较高的丢包率范围推理精度,但当 f 增大到 0.5 时,CLLRRI 算法的 Accuracy 仍能保持在 80%以上.Emulab 仿真实验结论与模拟实验基本一致.由于篇幅有限,不再比较不同节点网络规模及不同拥塞链路比例下算法的推断性能.

在单次仿真实验中,实际拥塞链路结果(benchmark)已知,为了直观地比较不同算法的推断性能,对 f=0.3 时的某单次实验结果进行分析.实验中,首先通过 30 次 E2E 路径性能探测,借助公式(6)获取各链路拥塞先验概率,在链路拥塞先验知识学习过程中,根据 E2E 路径性能情况,拥塞路径途经的链路共 18 条(为了便于算法编写,链路用“路由器 id/路由器 id”表示),分别为 6/0,6/1,6/3,6/2,0/15,1/10,1/17,3/9,3/12,3/13,3/14,3/19,2/8,2/18,0/5,0/4,5/16,4/11.借助于学习过程中 traceroute 获取的拥塞矩阵及各 E2E 路径性能探测值,计算得出各链路的拥塞先验概率分别为 p_0=0.2561,p_1=0.5853,p_2=0.3098,p_3=0.2466,p_4=0.5715,p_5=0.3748,p_6=0.3343,p_7=0.2501,p_8=0.2001,p_9=0.4469,p_{10}=0.2156,p_{11}=0.4002,p_{12}=0.1759,p_{13}=0.222,p_{14}=0.516,p_{15}=0.4172,p_{16}=0.3291,p_{17}=0.257.

对待测网络中各路径进行一次 E2E 性能探测,获取各拥塞路径,此时,造成这些拥塞路径的 benchmark 为链路:3/19,6/1,6/2,2/8,0/4,其实际丢包率值分别为{0.06},{0.12,0.13},{0.07,0.08},{0.07},{0.19}.由于链路 6/1,6/2 为两条 E2E 路径中的共享瓶颈链路,为使仿真实验更接近实际网络,设置共享瓶颈链路的丢包率在[−0.02, 0.02]之间随机发生改变,故链路 6/1,6/2 有两个不同的丢包率数值.利用本文提出的 CLLRRI 算法及文献[1,8]提出的算法分别对当前时刻拥塞链路进行定位,结果见表 1.其中,算法漏检链路前加“−”,误检链路前加“+”进行表示.

Table 1 Results comparison of congested link position among different algorithms

表 1 不同算法的拥塞链路定位结果比较

算法	定位拥塞链路	漏检(−)链路,误检(+)链路	DR (%)	FPR (%)	Accuracy (%)
CLINK	6/1,0/4,3/19,6/2	−2/8	80	0	/
Range tomography	3/19,6/2,2/8,6/1,4/11	−0/4,+4/11	80	20	75
CLLRRI	3/19,6/1,6/2,2/8,0/4	−	100	0	100

由表 1,CLINK 算法漏检了拥塞链路 2/8,无误检链路;Range tomography 算法漏检了链路 0/4,误检了链路 4/11;本文提出的 CLLRRI 算法在此次拥塞事件推断中,拥塞链路定位结果与实际情况一致,无漏检及误检链路.并且,实际拥塞链路的丢包率数值均包含在算法推断的各自拥塞链路丢包率范围内;Range tomography 算法中对拥塞链路 2/8 的丢包率范围推断结果为[0.02,0.06],而推理时刻该链路实际丢包率值为 0.07.由单次仿真实验结果表 1,本文提出的 CLLRRI 算法拥塞链路定位 DR=100%,FPR=0,而 CLINK 算法及 Range tomography 均存在误差;且 CLLRRI 算法拥塞链路丢包率范围推断准确率 $Accuracy$=100%,同样优于此次仿真实验中 Range tomography 算法的推断性能.

仿真实验中,对网络中各 E2E 路径的路由信息及性能获取方法与模拟实验不同,但拥塞链路性能计算及推理过程与模拟实验基本一致.因此,仿真实验中不再对不同类型、规模网络拓扑模型下进行算法仿真实验结果的分析与比较.

5 结束语

本文针对由路由器/交换机组成的 IP 网络存在多链路拥塞等问题,提出了一种拥塞链路及其丢包率范围推断算法 CLLRRI.通过构建待测 IP 网络贝叶斯网模型,借助链路拥塞贝叶斯 MAP 准则,定位 E2E 拥塞路径中最容易发生拥塞的链路,代替传统以瓶颈链路作为拥塞链路选取的经验方法;通过聚类与拥塞链路相关且性能相近的路径集合,在集合中动态地计算路径性能相似系数,借助相似系数贪婪启发式循环推断各拥塞链路及其丢包率范围.模拟实验及仿真实验均验证了算法的准确性及鲁棒性.

虽然动态路由 IP 网络中,E2E 路径途经链路因带宽受限等因素可能发生重路由,理论上,重路由将给算法性能带来影响,但是发生在学习过程与推断过程中的链路变化几率较小,对算法性能影响不大.另外,算法在推断拥塞链路时,将正常 E2E 路径及途经链路移除,理论上会减小算法精度,但是由于实际 IP 网络各 E2E 路径 hop 通常不超过 6,移除正常 E2E 路径途经的各链路对算法性能并未带来影响.此外,算法虽未充分考虑各链路拥塞相关性,但通过多次实验验证,拥塞链路间是否存在相关性,对算法性能也均未带来较大影响.

References:

[1] Zarifzadeh S, Gowdagere M, Dovrolis C. Range tomography: Combining the practicality of Boolean tomography with the resolution of analog tomography. In: Bullard C, ed. Proc. of the 12th ACM Internet Measurement Conf. Boston: ACM Press, 2012. 385−398. [doi: 10.1145/2398776.2398817]

[2] Pan SL, Zhang ZY, Fei GL, Qian F, Hu GM. Survey on network tomography for link performance parameter evaluation. Ruan Jian Xue Bao/Journal of Software, 2015,26(9):2356−2372 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4867.htm [doi: 10.13328/j.cnki.jos.004867]

[3] Coates M, Nowak R. Network loss inference using unicast end-to-end measurement. In: Proc. of the ITC Conf. on IP Traffic, Modeling and Management. Monterey, 2000. 28.1−28.9.

[4] Adams A, Bu T, Friedman T, Horowitz J, Towsley D, Caceres R, Duffield N, Presti FL, Moon SB, Paxson V. The use of end-to-end multicast measurements for characterizing internal network behavior. IEEE Communications Magazine, 2000,38(5):152−159. [doi: 10.1109/35.841840]

[5] Bu T, Duffield N, Presti FL, Towsley D. Network tomography on general topologies. ACM Sigmetrics Performance Evaluation Review, 2002,30(1):21−30. [doi: 10.1145/511334.511338]

[6] Lawrence E, Michailidis G, Nair V, Xi B. Network tomography: A review and recent developments. In: Proc. of the Fan & Koul Editors Frontiers in Statistics. 2006. 345−364. [doi: 10.1142/9781860948886_0016]

[7] Duffield N. Simple network performance tomography. In: Proc. of the ACM/Usenix IMC 2003. ACM Press, 2003. 210−215. [doi: 10.1145/948205.948232S]

[8] Nguyen HX, Thiran P. The Boolean solution to the congested IP link location problem: Theory and practice. In: Proc. of the IEEE INFOCOM 2007. Anchorage: IEEE, 2007. 2117−2125. [doi: 10.1109/INFCOM.2007.245]

[9] Vardi Y. Network tomography: Estimating source-destination traffic intensities from link data. Journal of the American Statistical Association, 1996,91(433):365−377. [doi: 10.1080/01621459.1996.10476697]

[10] Duffield N. Network tomography of binary network performance characteristics. IEEE Trans. on Information Theory, 2006, 52(12):5373−5388. [doi: 10.1109/TIT.2006.885460]

[11] Duffield NG, Presti FL, Paxson V, Towsley D. Inferring link loss using striped unicast probes. In: Proc. of the INFOCOM 2001. Anchorage: IEEE, 2001. 915−923. [doi: 10.1109/INFCOM.2001.916283]

[12] Zhang ZY, Fei GL, Yu FC, Hu GM. Improving the accuracy of Boolean tomography by exploiting path congestion degrees. In: Proc. of the IEEE Symp. on Computers and Communications. Cappadocia: IEEE, 2012. 725−731. [doi: 10.1109/ISCC.2012.6249384]

[13] Tsang Y, Yildiz M, Barford P, Nowak R. Network radar: Tomography from round trip time measurements. In: Proc. of the ACM SIGCOMM Conf. on Internet Measurement 2004. Taormina: ACM Press, 2004. 175−180. [doi: 10.1145/1028788.1028809]

[14] Padmanabhan VN, Qiu L, Wang HJ. Server-Based inference of internet performance. In: Proc. of the IEEE INFOCOM 2002. San Francisco: IEEE, 2002. [doi: 10.1109/INFCOM.2003.1208667]

[15] Ghita D, Nguyen HX, Kurant M, Argyraki K, Thiran P. Netscope: Practical network loss tomography. In: Proc. of the Conf. on Information Communications. IEEE, 2010. 1262−1270. [doi: 10.1109/INFCOM.2010.5461918]

[16] Nguyen HX, Thiran P. Network loss inference with second order statistics of end-to-end flows. In: Proc. of the IMC 2007. ACM Press, 2007. [doi: 10.1145/1298306.1298339]

[17] Shavitt Y, Sun X, Wool A, Yener B. Computing the unmeasured: An algebraic approach to internet mapping. IEEE Journal on Selected Areasin Communications, 2004,22(1):67−78. [doi: 10.1109/JSAC.2003.818796]

[18] Chen Y, Zhou W, Duan ZM, Qian YK, Zhao X. IP network congested link inference based on variable structure discrete dynamic Bayesian. Journal of Communications, 2016,37(8):13−23 (in Chinese with English abstract). [doi: 10.11959/j.issn.1000-436x. 2016151]

[19] Lian K, Long B, Wang HJ. A fault diagnosis approach of the large complex system based on Bayes theory. Acta Armamentarii, 2008,29(3):352−356 (in Chinese with English abstract). [doi: 10.3321/j.issn:1000-1093.2008.03.021]

[20] Dhamdhere A, Teixeira R, Dovrolis C, Diot C. NetDiagnoser: Troubleshooting network unreachabilities using end-to-end probes and routing data. In: Proc. of the CoNEXT 2007. New York: ACM Press, 2007. 1−12. [doi: 10.1145/1364654.1364677]

[21] Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. New York: W.H. Freeman and Company, 1979. 77−118.

[22] Waxman BM. Routing of multipoint connections. IEEE Journal on Selected Areas in Communications, 1989,6(9):1617−1622. [doi: 10.1109/49.12889]

[23] Albert R, Barabási AL. Topology of evolving networks: Local events and universality. Physical Review Letter, 2000,85(24): 5234−5237. [doi: 10.1103/PhysRevLett.85.5234]

[24] Bu T, Towsley D. On distinguishing between Internet power law topology generators. In: Proc. of the IEEE INFOCOM 2002. New York: IEEE, 2002. 638−647. [doi: 10.1109/INFCOM. 2002.1019309]

[25] Medina A, Lakhina A, Matta I, Byers J. BRITE: An approach to universal topology generation. In: Proc. of the Int'l Symp. on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. Washington: IEEE, 2001. 346−353. [doi: 10.1109/MASCOT.2001.948886]

[26] Zhang Y, Duffield N, Paxson V, Shenker S. On the constancy of internet path properties. In: Proc. of the SIGCOMM Internet Measurement Workshop. San Francisco: ACM Press, 2001. 197−211. [doi: 10.1145/505202.505228]

[27] Chen Y, Bindel D, Song H, Katz RH. An algebraic approach to practical and scalable overlay network monitoring. In: Proc. of the ACM SIGCOMM Computer Communication Review. 2004. 55−66. [doi: 10.1145/1015467.1015475]

附中文参考文献:

[2] 潘胜利,张志勇,费高雷,钱峰,胡光岷,网络链路性能参数估计的层析成像方法综述.软件学报,2015,26(9):2356−2372. http://www.jos.org.cn/1000-9825/4867.htm [doi: 10.13328/j.cnki.jos.004867]

[18] 陈宇,周巍,段哲民,钱叶魁,赵鑫.基于变结构离散动态贝叶斯 IP 网络拥塞链路推理.通信学报,2016,37(8):13−23. [doi: 10.11959/j.issn.1000-436x.2016151]

[19] 连可,龙兵,王厚军.基于贝叶斯最大后验概率准则的大型复杂系统故障诊断方法研究,兵工学报,2008,29(3):352−356. [doi: 10.3321/j.issn:1000-1093.2008.03.021]

陈宇(1978－),男,河南浚县人,博士生,副教授,主要研究领域为数据采集与信号处理,网络信息安全.

钱叶魁(1980－),男,博士,副教授,主要研究领域为网络信息安全.

周巍(1979－),男,博士,副教授,CCF 专业会员,主要研究领域为视频信息处理及其 VLSI 设计.

赵鑫(1979－),男,博士,教授,博士生导师,主要研究领域为网络信息安全.

段哲民(1953－),男,教授,博士生导师,主要研究领域为电路与系统,集成电路分析设计.