AS级Internet拓扑幂律和节点时效分析

付大愚 赵 海 张 君 葛 新
(东北大学信息科学与工程学院 沈阳 110004)

摘 要 Internet拓扑,尤其是AS级拓扑,是目前研究的热点问题。研究Internet拓扑的演化趋势,可以更好地了解网络的内在连接机制。基于CAIDA项目授权的海量数据(数据采集时间为2004年1月至2008年6月),首先介绍了必要的基本概念,然后给出了CCDF(d)-degree幂律分析、degree-rank幂律分析、节点时效分析。结果表明,随着网络拓扑结构的演化,AS级Internet的高度值节点部分较为稳定,保持了网络的聚集性与幂律性,但这部分节点随时间变化逐渐丧失有效连接,网络拓扑呈缓慢均匀化趋势。

关键词 AS级,Internet拓扑,幂律,节点时效
中图法分类号 TP393.4 **文献标识码** A

AS-level Internet Topology Power-law and Node Aging Analysis

FU Da-yu ZHAO Hai ZHANG Jun GE Xin
(College of Information Science and Engineering, Northeastern University, Shenyang 110004, China)

Abstract The Internet topology, especially the AS-level topology, is the hotspot issue of the research. We can comprehend well the inner connective mechanism of the network by researching the evolvement trend of the Internet topology. In this thesis, the research task is based on the massive data authorized by CAIDA(The Cooperative Association for Internet Data Analysis) Skitter project and the data's time span is from January 2004 to June 2008. This paper introduced the essential basic conceptions first, then carried out the CCDF(d)-degree power-law analysis, the degree-rank analysis and the node aging analysis. It is shown by the evolvement of the network topology structure that the top degree nodes of the AS-level Internet are stable and maintenance the clustering and the power-law of network. But these nodes lost the efficiency connection gradually by time and the network topology presents the trend of the laggard evenly.

Keywords AS-level, Internet topology, Power-law, Node aging

Internet作为当今人类社会信息化的标志,其规模正以指数速度高速增长。文献[1]的研究表明,Internet的节点数大约每两年翻一番。如今Internet的"面貌"已与其原型ARPANET大相径庭,成为一个由计算机构成的"复杂自组织生态系统"。虽然Internet是人类亲手建造的,但却没有人能说出这个庞然大物看上去到底是个什么样子,运作得如何。Internet拓扑研究就是探求在这个看似混乱的网络之中蕴含着哪些还不为我们所知的规律。发现Internet拓扑的内在机制是认识Internet的必然过程,是在更高层次上开发利用Internet的基础。

1 AS级拓扑分析介绍和若干定义

对Internet的研究自Internet的诞生开始[2-4],一直是层出不穷、经久不衰的。在早期,人们更多关注的是Internet的体系结构、网络协议[5]、计算机互联以及Internet所提供的服务等方面的研究。特别是近几十年来人们在复杂性科学和复杂网络等领域取得的研究成果,使国内外的研究者认识到Internet也是复杂网络之一,由此人们开始从复杂性和复杂网络的角度对Internet进行研究。

Internet的路由选择结构是一种层次式的选择结构,由若干路由器汇集成一个AS(Autonomous System,自治系统),不同AS之间再通过边界路由器而彼此互连[6]。所以,Internet拓扑的研究工作也主要集中在AS级和路由级两个层面开展。由于与路由级拓扑相比,AS级拓扑位于网络的更"上"一层,其特征与变化对Internet的影响更为巨大,相关研究对网络未来的发展意义更为重大。同时,AS级拓扑数据集规模小,以现有计算能力来说,可以更加有效执行深层次、高时间复杂度的计算分析,探究更为深层次的客观特性。根据AS拓扑分析的结果,可以总结更加系统的分析策略以及可靠的分析手段,以利进一步发现路由级拓扑中的隐含规律,为超大规模网络统计分析提供可用手段。

Internet是动态变化的,其拓扑也是随时间而改变的,所以研究Internet拓扑的演化趋势,可以更好地了解网络的内在连接机制。下面介绍本文的几个重要概念[7,8]。

到稿日期:2008-10-27 返修日期:2009-01-21 本文受国家自然科学基金项目(69873007)资助。
付大愚(1972-),男,博士研究生,主要研究方向为复杂网络,E-mail:fdymail@yeah.net;赵 海(1959-),男,教授,博士生导师,主要研究方向为嵌入式技术、复杂网络、数据融合等;张 君(1967-),女,博士研究生,主要研究方向为复杂网络;葛 新(1982-),男,博士研究生,主要研究方向为复杂网络。

定义 1(图,Graph)　一个图 G 是一个三元组,这个三元组包括一个顶点集 $V(G)$、一个边集 $E(G)$和一个关系,该关系使得每一条边和两个顶点(不一定是不同的点)相关联,并将这两个顶点称为这条边的端点。

从定义中可以看到,从任意顶点 x 到 y 不能连接两条或以上边。本文所讨论的图,均符合上述要求,即均为不含多重边的图。

定义 2(网络,Network)　$N=(V,E,c,X,Y)$为一个网络,如果

(1)$G=(V,E)$是一个有向图;

(2)c 是 E 上正整数,称为容量函数,对于每条边 e,$c(e)$称为边 e 的容量;

(3)X 与 Y 是 V 的两个非空不相交子集,分别称为 G 的发点集与收点集,$I=(X|\overline{X\cup Y})$称为是 E 的中间点集,X 的顶点称为发点(源),Y 的顶点称为收点(汇),I 的顶点称为中间点。

在图论中网络 $G=(V,E)$是指由点集 $V(G)$ 和边集 $\vec{E}(G)$组成的图,且 $E(G)$中的每条边 e_i 有 $V(G)$中的一对点(u,v)与之对应。记顶点数为 $N=|V|$,边数为 $L=|E|$。

定义 3(度,Degree)　设 N 是一个网络,$V(N)$是所有顶点的集合,$E(N)$是所有边的集合。顶点 v 的度 k_v 是指与此顶点 v 连接的边的数量,$v\in V$。即

$$k_v=\sum_{l\in E}\delta_l^v$$

其中,当边 l 包含顶点 v,δ_l^v 取值为 1,否则为零,即

$$\delta_l^v=\begin{cases}1, & \text{if } l \text{ include } v\\ 0, & \text{if not}\end{cases}$$

定义 4(度分布,Degree Distribution)　对于无向图 $G(V,E)$,记度为 k 的顶点数目为 $P(k)$,则 $p(k)=\frac{P(k)}{\nu(G)}$给出了图 G 的顶点度分布。

关于复杂网络拓扑图中顶点度分布常使用幂律形式的分布函数研究,如 frequency-degree 幂律分布、degree-rank 幂律分布、eigenvalue-rank 幂律分布以及 CCDF(d)-degree 幂律分布等。

2　AS 级 Internet 拓扑幂律分析

CAIDA(The Cooperative Association for Internet Data Analysis)是一个对全球范围 Internet 结构及数据进行研究的国际合作机构,CAIDA Skitter 是由全世界范围的主要研究机构与高等学府参与并涉及多个领域和交叉学科的大型科研项目,研究的主要内容包括 Internet 网络的产生、发展及演化趋势,以及 Internet 网络行为、动力特征和 Internet 宏观拓扑结构的变化规律。Skitter 通过跟踪从一个源地址到多个目的地址的前向 IP 地址的方式来获取 Internet 的拓扑结构,通过收集 ICMP 协议的 TTL 值(生存时间)数据来绘制 Internet 节点间关系。东北大学嵌入式技术实验室经该组织授权,成立 CAIDA 中国第一节点(Neu node),成为该组织在中国的首家合作伙伴,共享研究成果,并保持长期经验技术交流。本文选取 2004 年 1 月至 2008 年 6 月 CAIDA Skitter 项目提供的 AS 级拓扑的最新测量数据进行分析,对网络特征量进行大时间跨度演化分析,观察网络生长过程中各项指标的变化,并分析其未来趋势。选取多种统计量进行分析,尽可能避免由于观察角度单一而导致的特征遗漏。

节点度的幂律分布是 Internet 的一个重要规律,可以体现网络中边的分布情况,即少数节点拥有大量连接,而大多数节点的连接数很低。幂指数可以量化说明这种拓扑的扭曲现象。1999 年,Faloutsos 等人[9]首次采用幂律(power-law)来刻画 Internet 拓扑结构特征,并提出了 degree-rank 幂律、frequency-degree 幂律以及 eigenvalue-rank 幂律,人们对 Internet 节点度分布的研究有了新的认识。2003 年,Siganos 等人在文献[10]研究中,发现由于 frequency-degree 幂律相当于考察的是频率 f_d 的概率密度函数(probability density function,简称 PDF),而累积分布比概率密度的统计鲁棒性更好。文献[10]考察了频率 f_d 的补累积分布函数(complementary cumulative distribution function,简称 CCDF)D_d 与度 d 的关系,发现 D_d 与 d 也呈幂律关系,称该幂律为 CCDF(d)-degree 幂律。由于 CCDF(d)-degree 幂律能够更好地体现数据的拓扑特征,因此本文仅考察测量数据集的 CCDF(d)-degree 幂律与 degree-rank 幂律。

2.1　CCDF(d)-degree 幂律分析

为考察 AS 级 Internet 静态拓扑的幂律性,本文选择 2008 年 6 月份的数据,计算其补累积函数与度值的关系,并在对数坐标系下拟合分布曲线斜率。分布曲线及拟合曲线如图 1 所示。

图 1 中横坐标为节点度值 k,纵坐标为度数大于 k 的节点补累积函数值,二者均取自然对数值。拟合补累积幂指数为 1.159。在节点度较高时拟合曲线偏差较大,说明拓扑中"热"节点部分不完全符合幂律分布,其吸引连接的能力与理想幂律模型相比要弱很多。同时,叶子节点部分的拟合情况也较差,说明叶子节点数目较之理想模型要多。

图 2 为补累积幂指数的时间变化规律。补累积幂指数呈现两阶段分布,且第二阶段较之第一阶段有明显下降。与文献[10]统计的 2002 年之前的数据分布情况相比,下降趋势有所变缓。幂指数的下降说明网络拓扑呈现平面化趋势,节点度分布逐渐均匀,度值较低的节点获得的连接有所增加,而"热"节点的连接数相应下降。

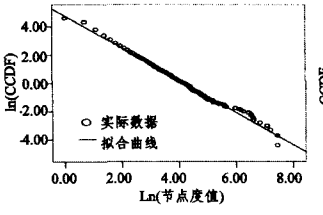

图 1　静态拓扑的 CCDF(d)-degree 幂律分布拟合曲线图

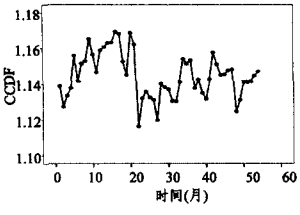

图 2　补累积幂指数随时间的统计变化图

2.2　degree-rank 幂律分析

选择 2008 年 6 月份数据,计算其中秩与度值的关系,同样在对数坐标系下拟合分布曲线斜率。分布曲线及拟合曲线如图 3 所示,图中横坐标为度值 k 的秩,纵坐标为节点度值 k,二者均取自然对数值。拟合秩幂指数为 0.874。与图 1 类似,图 3 在叶子节点部分以及节点度较高时拟合曲线偏差较大。

图 4 为秩幂指数的时间变化规律。由图中可以看出秩幂指数随时间变化下降规律明显。这与文献[10]根据早期数据

得出的规律一致，与 2.1 小节补累积幂指数分析结果也一致，且更加明显地体现了 Internet 度分布平面化的趋势。

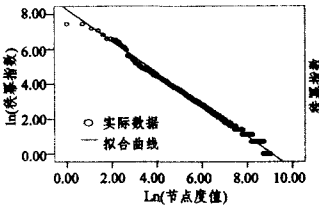

图 3 静态拓扑的 degree-rank 幂律分布的拟合曲线图

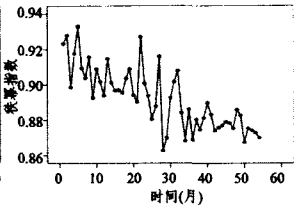

图 4 秩幂指数随时间的统计变化图

对测量数据的幂律分析，表明了 AS 级 Internet 拓扑的度分布呈现明显的幂律规律，网络中边的分布倾向性严重，少数节点长期拥有大量连接，可以看作是 Internet 的核心。而大多数节点连接很少，且这部分节点变化频繁，是网络变化的主要部分。

与早期数据分析结果相比，Internet 更加趋于稳定，幂指数变化愈加放缓，但仍在下降中，网络的连接均匀化过程仍在继续。两种幂律分布曲线的拟合情况均表明，在节点度值极大和极小的区域，幂律规律符合情况较差，网络度分布呈现一定的层次性，说明网络的拓扑复杂性不能只由幂指数衡量，需要按度值数的大小做进一步分级量化分析。

3 节点时效分析

本文所用数据覆盖 2004 年 1 月至 2008 年 6 月，共计 54 个月，共涉及 15609 个不同 AS。在随时间变化的 Internet 中，AS 不断地出现、消亡，其生存时间(时效)也是描述网络的一个重要特征量，可以说明网络各部分的动荡程度。

3.1 节点时效分布

图 5 和图 6 给出的是以月为单位划分的生存时间不同的节点数目分布情况。其中图 5 是节点时效与其相应数目的关系图，图 6 是节点时效的累计分布，图中数据点表示时效小于某一数值的节点数占节点总数的百分比。从图 5 中可以看出，除去个别点，大多数的生存时间跨度都有少量节点存在，且不同时效的节点数目分布均匀。数目最多的是长效节点，即在数据集的时间范围内均有出现的节点，其数量为 3291 个，而网络中同一月份存在的节点数约为 8000 至 9000 个，说明 AS 网络中长期稳定存在的节点并不占主要地位。由图 6 可进一步看出，网络中时效少于 20 个月的节点约占 40%，时效少于 30 个月的节点约占 60%，可见短时效的节点具有相当大的比例。

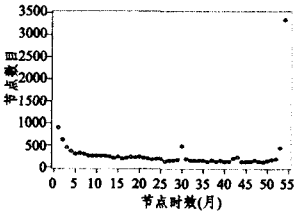

图 5 不同时效节点的节点数目图

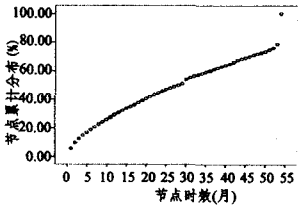

图 6 不同时效节点的累计分布图

图 7 给出的是节点度值与其时效间的关系。图中横轴为节点在其生存周期内各月份度值的平均值，纵轴为平均度值相等的节点其时效的平均值。除去右下角的奇异点外，大多数节点表现出一致的规律，即度值较大的节点，其时效也较长。该规律的逆否命题同样体现在图中，即时效短的节点，度值都较小。因此，AS 网络的整体稳定性可以得到进一步的解释：网络中较为重要的点(度值大的“热”节点)很少是突然出现或很快消亡的，维持了网络的整体稳定；而变化较多的节点多为度值较小的边缘节点，其新生或消亡对网络整体影响相对较小。

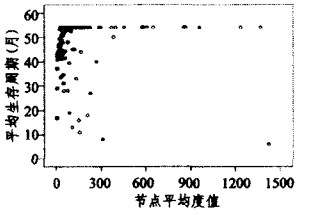

图 7 不同度值节点的时效图

3.2 节点的时效相似性

时效不同的节点在网络中具有不同的作用，长效节点使网络趋于稳定，相反的短效节点则使网络变化频繁，更加动荡。图 8 给出了时效不同的节点度值的频度分布情况三维示意图，3 个坐标分别为节点时效、节点平均度值(该节点时效内其度值的平均值，因取值范围较大，故使用了对数坐标)、度值频数(某度值在时效一定的节点集中出现的次数)。图 9 是图 8 的一个切面图的变形，为节点时效为 54 个月时平均度值与度值频数的关系，其中横纵轴均为对数坐标。

图 8 中显示，各时效下的节点集规律十分相似，频数较高的节点其度值也较小，同时存在少量度值很大的节点。为使全图更利于观察，图 8 中频数轴使用的是线性坐标。而由图 9 的对数坐标可以更清楚地看到，时效为 54 的节点集呈现出明显的 frequency-degree 幂律规律。由已观察到的相似性，可知任一时效下的节点集度值均为幂律分布。

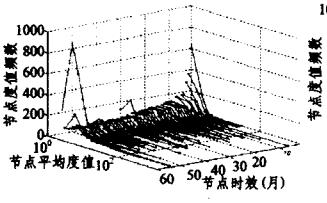

图 8 不同时效节点三维示意图

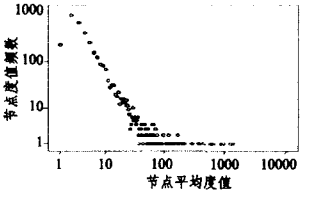

图 9 不同时效节点切片图

我们看到，除去时效为 54 的节点集之外，低度值节点的频数随节点时效降低而增加，所以可知在低时效节点集当中低度值节点比例较大，与图 3 结论一致。尤其是节点时效很小的时候，低度值频数极大，即网络中变化最为剧烈的部分大多是一些度值很小的节点。

结束语 AS 级网络拓扑的变化对 Internet 网络性能有很大影响，尤其是其主要度量指标的波动。对不同拓扑的网络，可以通过进一步优化网络协议或路由策略，充分利用网络中存在的幂律性质及聚集性质，减少不必要的数据转发，有效改善网络性能。本文通过对 CAIDA Skitter 实测数据的幂律分布、节点时效等多项分析的结果，论证了 AS 网络高度值节点部分较为稳定，保持了网络的聚集性与幂律性，但这部分节点随时间变化逐渐丧失有效连接，网络拓扑呈缓慢均匀化趋势。

(下转第 62 页)

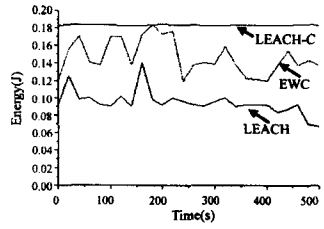

图 4　成簇过程中的能量消耗

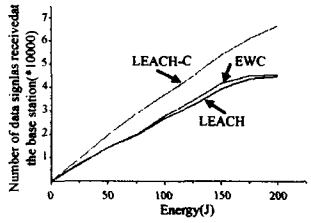

图 5　单位能量内基站所接收的数据量

模拟显示了 LEACH-C 协议在给定能量条件下能够接收到最多的数据。这是由于 LEACH-C 协议采用中心控制模式，由基站通过收集各个节点的信息来作出节点选择的判断，因此 LEACH-C 协议能够通过分析所有节点的信息，做出最优判断。但是与 LEACH 协议相比较，优化算法考虑了距离、剩余能量以及节点度等诸多因素，其性能优于 LEACH 协议。与 LEACH 协议随机选取簇头的机制相比，优化算法通过考虑本地信息等各方面因素，从而做出了更优的簇头选择。

图 6 显示了不同算法的网络生存周期。

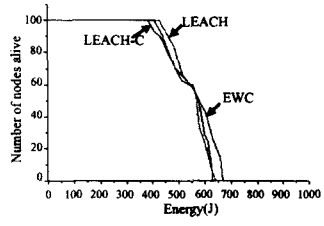

图 6　网络生存周期

从图 6 中可以看出，基于综合因素的分布式优化算法具有最长的网络生存周期。这是由于优化算法更合理地进行了簇头选择，从而减少了数据传输中的能量损耗，延长了网络生存周期。

结束语　提出了一种基于综合因素的分布式簇头选举算法。算法通过在簇头选举过程中考虑距离、剩余能量等综合因素，改善了 LEACH 协议簇头选举过程中的缺点，均衡了网络负载并且延长了网络生存周期。在此算法中，每个影响因子都分配有一个系数。通过对系数的调整能够使算法适应于不同的网络需求，从而增加了灵活性。仿真试验表明，与 LEACH 协议相比，此优化算法具有更好的性能表现、更长的网络存活时间。但是，该优化算法因为是分布式的，节点之间需要交互信息从而做出决定，所以会增加信息冗余。此外，如何有效地对系数进行分配，从而使优化算法性能更优等方面还需要更多的研究。在今后的工作中，将会在这方面进行更进一步的研究，使系数的分配更加合理，从而进一步优化系统效率。

参 考 文 献

[1] Heinzelman W, Kulik J, Balakrishnan H. Adaptive protocols for information dissemination in wireless sensor networks[C]// Proceeding of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking. Seattle, WA, August 1999

[2] Intanagonwiwat C, Govindan R, Estrin D. Directed diffusion: A scalable and robust communication paradigm for sensor networks[C]// Proceeding of the 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking. Boston, MA, August 2000

[3] Heinzelman W R, Chandrakasan A P, Balakrishnan B. Energy-Efficient Communication Protocol for Wireless Microsensor Networks[C]// Proc. 33rd Hawaii Int. Conf. System Science. Maui, HI, Jan. 2000

[4] Xu Y, Heidemann J, Estrin D. Geography-informed energy conservation for ad hoc routing[C]// Proceedings of the 7th Annual ACM/IEEE International Conference on Mobile Computing and Networking. Rome, Italy, July 2001

[5] Yu Y, Estrin D, Govindan R. Geographical and Energy-Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks[R]. UCLA-CSD TR-01-0023. UCLA Computer Science Department, May 2001

[6] Lindsey S, Raghavendra C S. PEGASIS: Power Efficient Gathering in Sensor Information System[C]// Proceedings of the IEEE Aerospace Conference. Big Sky, Montana, March 2002

[7] Younis O, Fahmy S. HEED: A Hybrid, Energy-efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks[J]. IEEE Transaction on Mobile Computing, 2004, 1(3): 366-379

[8] Estrin D, et al. Next centurychallenges: Scalable Coordination in Sensor Networks[C]// Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking. Seattle, WA, August 1999

(上接第 23 页)

参 考 文 献

[1] Willinger W, Doyle J. Robustness and the Internet: Design and evolution [EB/OL]. 2002. http://netlab.caltech.edu/Internet/

[2] Li J, Sung M, Xu J, et al. Large-scale IP traceback in high-speed Internet: practical techniques and theoretical foundation[C]// Proceedings of the IEEE Symposium on Security and Privacy. California, USA, 2004

[3] Subramanian L, Padmanabhan V N, Katz R H. Geographic properties of Internet routing[C]// Proceedings of the USENIX Annual Technical Conference. 2002, 6

[4] Akella A, Seshan S, Balakrishnan H. The impact of false sharing on shared congestion management[C]// Proceedings of the 11th IEEE International Conference on Network Protocols. 2003, 11

[5] Jose M, Barcelo, Juan I, et al. Study of Internet autonomous system interconnectivity from BGP routing tables[J]. Computer Networks: The International Journal of Computer and Telecommunications Networking, 2004, 45(3): 333-344

[6] 王大东，王洪君，王瑞军，等. 一种基于 AS 的 Internet 拓扑模型[J]. 计算机工程，2005，31(4)：23-25

[7] West D B. 图论导引[M]. 北京：机械工业出版社，2006：1-47，339-348

[8] Munkress J R. 拓扑学[M]. 北京：机械工业出版社，2006：2-52

[9] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology[J]. ACM SIGCOMM Computer-Communication Review, 1999, 29(4): 251-262

[10] Siganos G, Faloutsos M, Faloutsos P, et al. Power laws and the AS-level Internet topology[J]. IEEE/ACM Trans. on Networking, 2003, 11(4): 514-524