软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(10):2279−2290 [doi: 10.3724/SP.J.1001.2011.03924]

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

适合复杂网络分析的最短路径近似算法*

唐晋韬+, 王 挺, 王 戟

(国防科学技术大学 计算机学院,湖南 长沙 410073)

Shortest Path Approximate Algorithm for Complex Network Analysis

TANG Jin-Tao+, WANG Ting, WANG Ji

(College of Computer, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: E-mail: tangjintao@nudt.edu.cn

Tang JT, Wang T, Wang J. Shortest path approximate algorithm for complex network analysis. *Journal of Software*, 2011,22(10):2279−2290. http://www.jos.org.cn/1000-9825/3924.htm

Abstract: The tremendous scale of the social networks mined from Internet is the main obstacle of a social network analysis application. The bottleneck of many network analysis algorithms is the extortionate computational complexity of calculating the shortest path. Real-World networks usually exhibit the same topological features as complex networks such as the “scale-free” and etc, which indicate the intrinsic laws of the shortest paths in complex networks. Based on the topological features of real-world networks, a novel shortest path approximate algorithm which uses an existent short path passing through some local center nodes to estimate the shortest path in complex networks, is proposed. This paper illustrates the advantage and feasibility of incorporating the proposed algorithm within the network properties, which suggests a new idea for complex social network analysis. The proposed algorithm has been evaluated both on synthetic network stage and real world network stage. Experimental results show that the proposed algorithm can largely reduce the computational complexity and remain highly effective in complex networks.

Key words: social network; approximation algorithm; network property; shortest path problem

摘 要: 基于互联网抽取的社会网络往往具有较大的规模,这对社会网络分析算法的性能提出了更高的要求.许多网络性质的度量都依赖于最短路径信息,社会网络等现实网络往往表现出“无标度”等复杂网络特征,这些特征指示了现实网络中最短路径的分布规律.基于现实网络的拓扑特征,提出了一种适合于复杂网络的最短路径近似算法,利用通过局部中心节点的一条路径近似最短路径,该算法能够方便地用于需要最短路径信息的社会网络性质的估算,为复杂网络的近似分析提供了一种新的思路.在各种生成网络与现实网络上的实验结果表明,该算法在复杂网络上能够大幅降低计算复杂性并保持较高的近似准确性.

关键词: 社会网络;近似算法;网络性质;最短路径问题

中图法分类号: TP301 **文献标识码**: A

* 基金项目: 国家自然科学基金(60873097); 国家重点基础研究发展计划(973)(2005CB321802); 新世纪优秀人才计划(NCET-06-0926)

收稿时间: 2009-08-19; 修改时间: 2010-06-09; 定稿时间: 2010-07-28

随着 Web 2.0 的蓬勃发展,互联网正逐步融入人们的日常生活之中,深刻地改变着人们工作和生活的方式,这使得面向互联网的社会网络分析成为新的研究热点.随着数据挖掘技术的进步,研究者从互联网上挖掘的数据规模也在快速地增长.与其他领域的网络研究一样,社会网络分析也面临着数据规模急剧扩张的挑战.研究者通常将社会网络视为一个图,利用基于图的算法来分析社会网络的性质.而这些分析算法大都依赖于一个基本问题——节点间的最短路径的计算.如 20 世纪 60 年代 Milgram[1]提出的“六度分离”性质,就是对社会网络最短路径长度的假设;而近年在 Internet 中流行的“Bacon 数”和“Erdös 数”[2]游戏,以及对 Internet 等大规模网络的网络直径的研究[3],都是典型的最短路径查找问题;社会学家提出的度量网络元素重要性的两个性质——接近中心性、介数中心性[4]——也是通过元素对最短路径的贡献程度来度量的;许多聚类算法也需要节点之间的距离或最短路径信息[5],如 Girvan-Newman 算法[6]等.但最短路径的计算具有很高的复杂性,使得这些分析方法在面向规模较大的现实网络时存在性能问题.在串行计算机上,目前主要有两种解决思路:利用优化算法结构等方法降低算法的计算复杂性;利用启发式等方法限定搜索空间,近似计算最短路径.研究者们提出了各种不同形式的最短路径算法[7−9],由于问题本身的复杂性,目前最快的串行最短路径算法只能将计算复杂性降到 $O(n^{2.376})$[10].因此,如何快速而高效地近似最短路径成为研究者们关注的热点.

在社会网络的研究中,研究者们发现了不同于规则网络或随机网络的一些拓扑特征.如社会网络中往往具有较短的平均路径长度和较高的顶点集聚系数,Watts[11]提出了小世界模型来刻画这种现象,利用该模型可以较好地解释“六度分离”性质;Albert 等人[12]则发现了在大规模现实网络中度分布的无标度现象.研究者们将具有这些特性的现实网络称为复杂网络,并对复杂网络的拓扑特性、构造模型、传播动力学等方面都进行了深入的研究[13],也将其成果广泛应用于软件工程等领域[14].现实网络往往都具备复杂网络特征,本文利用复杂网络的拓扑特征推导现实网络中最短路径的可能分布.在此基础上,本文提出了基于区域中心点距离的最短路径(centers distance of zone,简称 CDZ)近似方法,利用复杂网络拓扑特征寻找一条实际存在的路径作为可能的最短路径,能够有效地应用于介数中心性等需要最短路径信息的社会网络分析算法的近似计算.实验结果表明,在具有复杂网络特征的网络上,我们的近似分析方法极大地提高了计算速度,而且保持了较高的有效性.

本文第 1 节回顾相关工作.第 2 节分析复杂网络拓扑特征对最短路径的影响,在此基础上提出针对复杂网络的最短路径近似算法 CDZ.第 3 节将 CDZ 算法应用到计算机生成网络和真实网络上,以检验 CDZ 算法的性能和正确性.第 4 节阐述如何结合 CDZ 算法近似计算中心性等社会网络性质的方法,并检验网络性质近似的有效性.最后总结全文,提出一些有待探讨的问题,并展望未来的工作.

1 相关工作

目前,面向互联网的社会网络挖掘和分析成为了一个新的研究热点.研究者们针对人们在互联网的活动进行挖掘和分析,如邮件交互[15]、科研合作[16]等;也研究了面向 Web 2.0 的社会网络挖掘方法[17,18];并将社会网络分析广泛应用于恐怖袭击分析[19]、犯罪核心挖掘[20]等问题.这些研究与其他现实网络分析一样,面临着网络规模急剧增大的挑战.中心性等网络性质度量方法由于复杂性太高而不能在大规模现实网络上有效应用,其中,最短路径的计算带来的复杂性最为显著.因此,如何快速而有效地近似最短路径,从而高效地分析网络性质,成为目前研究的一个热点.Chow 等人[21]提出了利用构建 A*算法的启发式来快速搜索最短路径的方法.Slivkins 等人[22]提出了 Rings of Neighbors 方法搜索最邻近节点,并在此基础上提出了基于环的距离近似算法.Rattigan 等人[23]则提出了图结构索引(network structure indices,简称 NSI)算法,利用保存的预处理结果快速近似节点间的距离,文献[24]将该算法用于图聚类算法的近似,取得了较好的结果.Zwick[25]提出了 t-最短路径近似算法的定义,一个算法被称作是 t-近似的,是指该算法估算的最短路径都不超过实际最短路径长度的 t 倍.Cohen 等人[26]设计了一种面向加权无向图的近似算法,该算法在 2-近似的情况下仅需 $O(n^{3/2}\sqrt{e})$ 的预处理时间.Thorup 等人[27]提出了 Approximate Distance Oracle 数据结构,对于(2k−1)-近似需要 $O(kn^{1+1/k})$的空间和 $O(n^{1/k})$的预处理时间,并能在 $O(k)$的时间内给出任意节点对之间的近似最短路径.Baswana 等人[28]设计了两种随机近似算法,在非加权无向图最短路径的 2-近似问题上具有 $O(e^{2/3}n\log n+n^2)$的时间复杂度.

上述方法的设计目标往往面向所有类型的网络,并没有针对复杂网络的拓扑特征进行优化.研究结果[29,30]表明,最短路径算法在具有不同拓扑特征的网络上效率差异很大,通用算法很难在各种网络形式下均有较好的效率.本文在分析复杂网络拓扑性质的基础上,提出了一种适用于现实网络的近似算法.该算法利用符合复杂网络拓扑特征的一条可能的最短路径估算距离,适用于现实世界中基于最短路径的网络分析方法的近似.

2 CDZ 最短路径近似方法

大规模网络中的所有节点对之间最短路径(all-pairs shortest paths,简称 APSP)问题一直是研究者面临的一个挑战.本节分析了现实网络的复杂网络特性,由此推导了关于复杂网络节点间最短路径特征的一个假设.在此基础上,提出了基于区域中心点的最短路径近似算法,并研究了适合于复杂网络的区域中心点选择策略.

2.1 CDZ算法

现实世界中,较大规模的网络在结构上往往表现出复杂网络特征,如社会网络、生物网络、Internet 物理网络等等[16].复杂网络特征包括小世界特征和无标度特征;小世界特征[11]说明,在复杂网络中,任意节点之间的距离均较短,且具有较高的聚集系数.而无标度特征[12]是指度分布较为符合幂律(power-law)定律,即复杂网络中存在着少量的占支配地位的中心节点,以及大量的不够活跃的普通节点.如在 Internet 网页链接结构中,绝大部分网页的链接数不超过 4 个,而不到 0.01%的网页占有了 80%以上的链接[12].本文统计了实验采用的两种现实网络是否同样具有复杂网络拓扑特征,包括科研文献引用网络 Cora 和 Blogger 社会网络(参见第 3.1 节).Cora 网络为 30 000 节点规模,Blogger 网络则包含 1 000 多个节点.本文利用 E-R 模型[31]分别生成与这两种网络规模基本相同的随机图进行比较,统计结果表明,这些现实网络的确具有复杂网络拓扑特征.如图 1 所示,不同于随机网络的正态分布,Blogger 网络和 Cora 网络的度分布都表现出了一定的幂律分布特征.而 Blogger 网络和 Cora 网络平均距离仅为 3 和 5,明显小于相同规模随机网络的平均距离(6 和 13).

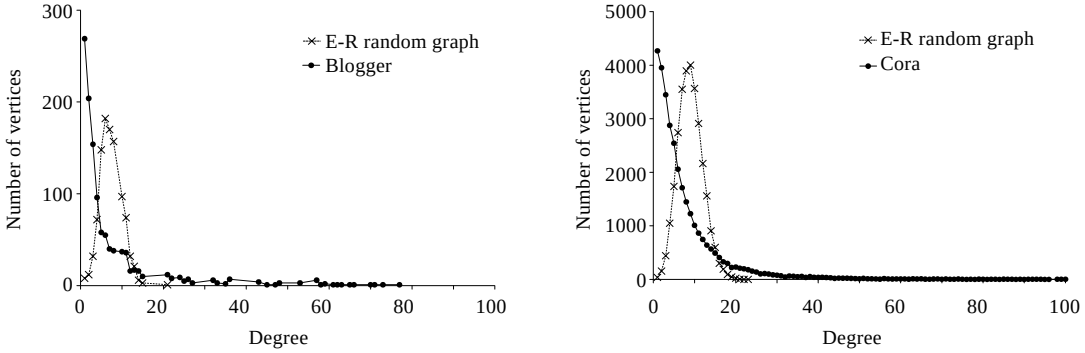


Fig.1 Degree distribution curve of real-world networks and random network

图 1 现实网络与相同规模的随机网路的度分布曲线

复杂网络拓扑特征体现了现实网络中最短路径的一些内在规律.不同于随机网络,复杂网络的大部分节点都只在小范围内相互连接,呈现出一定的高聚集系数特性.不同于规则网络,复杂网络任意两个节点间的距离都较短.复杂网络直径较小的原因在于少量连接不同簇的“长边”,社会学家 Granovetter 提出了弱链接的强度[32],描述了在人际关系中少量跨越交际圈的较弱关系对交际网络的巨大贡献.这说明,最短路径通过“长边”的可能性非常大.而复杂网络的无标度特征说明节点的度数分布不均衡,存在少量具有很高度数的中心节点以及大量度数很低的普通节点.这说明“长边”属于少数的中心节点的概率更大,任意节点之间的最短路径很有可能通过中心节点.因此,本文提出了在现实网络中关于最短路径规律的一个假设:

假设 1. 在具有复杂网络特征的现实网络中,大部分节点都是经过中心节点连接在一起的,任意节点之间的最短路径有较大的可能会经过中心节点.

该假设在很多现实网络中可以找到佐证,如在 Internet 网络中,绝大部分的节点是个人电脑终端,它们之间

主要通过路由器连接.在性接触网络[33]和科研合作网络[16]中也发现了类似的特点.根据假设 1,在 NSI 方法[23]的基础上,本文提出了一种基于区域中心点的最短路径近似方法——CDZ 方法.在预处理时,该方法查找中心节点并记录各个节点到最邻近中心节点的最短路径;在估算最短路径时,该方法利用通过中心节点的一条路径近似最短路径.在无向图情况下,CDZ 方法的预处理包括以下几个步骤:

(1) 根据某种节点选择策略(本文策略见第 2.2 节)选取 d 个可能的中心节点,表示为 $C_1,C_2,\ldots,C_d$;

(2) 从中心节点集合开始,宽度优先遍历整个图,记录每个节点到最临近中心节点的最短路径、相邻区域(两个区域有节点直接相连)中心节点之间的最短路径.该步骤将图划分为 d 个区域,记为 $Z_{C_1},Z_{C_2},...,Z_{C_d}$.

(3) 新构造一个仅包含 d 个中心点的图,如果原图中两个中心点所在区域相邻,在新图中添加一条连接这两个中心节点的边,边的权值为步骤 2 得到的在原图中两个相邻区域的中心节点之间的距离.计算新图中所有节点之间的最短路径,通过映射可以得到原图任意两个中心节点之间的最短路径.

CDZ 方法利用通过中心节点的一条实际存在的路径近似跨区域节点之间的最短路径.任意两个属于不同区域的节点 s,t 之间的距离,可以近似为这两个节点到各自区域中心节点的距离与中心节点之间的距离之和.

$$d(s,t)=d(s,C_s)+d(C_s,C_t)+d(C_t,t) \tag{1}$$

其中,C_s,C_t 分别为距节点 s,t 最近的中心节点,函数 $d(s,t)$为节点 s 和节点 t 之间的距离.属于不同区域的两个节点 s,t 之间的距离 $d(s,t)$,CDZ 方法利用一条通过中心节点的路径长度来近似:从节点 s 到其中心节点 C_s 的距离 $d(s,C_s)$、从节点 t 所在区域中心节点 C_t 到 t 的距离 $d(C_t,t)$以及中心节点之间的距离 $d(C_s,C_t)$之和.

如果两个节点属于同一区域,通过两个节点到中心节点的路径信息可以快速地发现它们的最近公共祖先(least common ancestors).本文利用两个节点通过其最近公共祖先的一条路径,近似它们之间的最短路径.对于属于同一区域的节点 s,t,C_{st} 为区域的中心节点,LCA_{st} 为 s,t 的最近公共祖先,s,t 之间的距离可以近似为

$$d(s,t)=d(s,LCA_{st})+d(LCA_{st},t) \tag{2}$$

如果 s,t 的最近公共祖先是它们中间的一个,那么近似公式(2)给出的近似路径即为最短路径.

与常见的随机方法不同,CDZ 方法利用复杂网络的高聚集特性来划分图的区域.在最短路径的近似上,该算法利用了复杂网络的小世界特性和无标度特性,认为复杂网络中最短路径有较大可能会经过中心节点.近似公式(1)、近似公式(2)在给出节点距离的一个近似值的同时,也提供了一条图中存在的可能最短路径.

2.2 区域中心点的选择策略

影响 CDZ 近似方法有效性的一个核心因素是区域划分,CDZ 方法的准确性很大程度上取决于划分的区域是否符合网络的拓扑结构.CDZ 方法利用基于中心节点集合的宽度优先算法划分区域,如何构造适合复杂网络的中心点选择策略成为 CDZ 方法需要解决的一个核心问题.在社会网络的常用分析手段中,局部中心性(local centrality)[34]通过统计与节点相连的边数来度量节点的重要性.该性质能够度量在局部环境中处于“核心”位置的点,测量方法也只需要利用到节点的度数,计算复杂度较低.因此,我们选取局部中心性最高的一系列节点作为 CDZ 算法的区域中心点.在此度量方法下,中心点选择策略还需要考虑区域的划分粒度,即中心点的数目.选择合适的中心点数目能够将网络划分为符合实际拓扑结构的区域,进而能够有效地拟合节点间的最短路径,对 CDZ 方法的性能也会产生较大的影响.

复杂网络一般都具有无标度特征,即度的分布较为符合幂律分布.在具有无标度特征的网络中,存在少量有着较高度数的活跃节点,而大部分节点往往只有一两个邻居节点;如在科研文献引用网络 Cora 中,度数最高的 20%的节点的度数之和就达到了网络总度数的 75.6%.因此,在复杂网络中,只需选择局部中心性最高的少量节点就能使划分的区域很好地符合网络拓扑特征.节点的度数是由与之相连的边的数目来决定的,我们认为,在所选取的中心点的度数之和接近或达到全图度数之和的 50%时,绝大部分边都会与中心点相连接.如在 Cora 网络中,当选择的中心点集合度数达到全图的 50%时,与中心点集合相连的边占到了全部边的 83%以上;而在图 1 所示的随机网络中,这个比例仅仅达到 51%.此时,CDZ 算法在遍历中心点时就覆盖了复杂网络的绝大部分边,构建的结构索引更符合实际的网络结构.根据复杂网络的无标度特性,大部分复杂网络的度数均符合或接近幂律分布,即 $y=cx^{-r}$,而社会网络的幂率 r 往往在 2~3 之间[12].如果中心点集合的度数大于总度数的 50%,则需要选择

的节点数目一般不超过总节点数的 10%.因此,CDZ 方法的中心节点选择策略如下:

(1) 将节点按照局部中心性大小进行排序,得到排序后的节点队列 Q;

(2) 取出队列 Q 中的第 1 个元素,加入中心点集合 C;

(3) 如果中心点集合 C 中所有节点的度数之和大于全图度数之和的 50%,或中心点集合的大小超过了全图节点总数的 10%,则终止循环;否则,继续执行第 2 步;

(4) 输出中心点集合 C,此时,C 中所有元素就是本文策略选择的中心节点.

在近似节点间的距离时,CDZ 方法仅需要常数时间.但预处理步骤中的图遍历需要 $O(e+n\log n)$的开销,其中,e 为网络边的数目,n 为节点数目.预处理步骤 3 将所有中心点抽取成一个加权网络,利用经典的最短路径算法计算中心点之间的距离需要 $O(d^3)$的时间,其中,d 为中心点数目;选择中心节点需要 $O(n\log n)$的时间.因此,CDZ 方法预处理的计算复杂性为 $O(e+n\log n+d^3)$,其中,e 为网络边的数目,n 为节点数目,d 为选择的中心节点数目.在具有无标度特性的网络中,高度数节点的数目非常少,即 d 远远小于 n.因此,在具有复杂网络特征的网络中,CDZ 方法的计算复杂性仅为 $O(e+n\log n)$.如果图的规模较大或无标度特征不明显,则仍可能由于 d 较大而使得精确计算所有中心点之间最短路径的时间开销过大.在中心点规模大于 500 的情况下,我们同样采用 CDZ 方法近似计算中心点之间的最短路径,以取得性能和正确性之间更好的平衡.

3 最短路径近似评测

作为很多网络性质的度量基础,近似算法对最短路径估算的准确性和效率直接影响到网络近似分析的性能.在本节中,我们在各种网络上评估了 CDZ 近似算法的有效性,并分析了影响算法正确性与性能的因素.

3.1 数据集

本文使用了科研文献引用网络 Cora[35]和搜狐博客(http://blog.sohu.com/)社会网络 Blogger 来评测算法在现实网络中的有效性.作为一种常用的关系数据,Cora 包含了在信息处理等领域的近 37 000 篇科研文献以及文献之间的 100 000 多个引用关系.我们将每篇文章作为一个节点,如果文章之间有引用关系,则在对应节点之间添加一条边.通过这些步骤,我们得到了一个包含 30 751 个节点及 134 996 条边的科研文献引用网络,本文称其为 Cora 网络.此外, 我们也抓取了搜狐博客的部分数据,共收集到了 1 327 个博客的 20 166 篇文章.将每个博客表示为图中的一个节点,如果博客之间存在着表示阅读、评论、引用、交友的链接,就在对应的节点之间添加一条边.去除了孤立节点后,生成了规模为 1 113 个节点和 3 685 条边的无向图,本文称其为 Blogger 网络.

为了全面评测 CDZ 方法的近似效果,本文利用经典的网络模型生成了 3 种具有不同结构特征的网络:根据 Erdös 模型[31]生成的随机网络,根据 Eppstein 幂律模型[36]生成的幂律网络以及根据 Barabási 模型[12]生成的无标度网络.随机网络属于简单网络,而后两种生成网络都有着显著的无标度特征.为了便于统计性质的直观比较,3 个生成网络的规模与 Cora 网络基本相当.这些网络的统计性质见表 1.

Table 1 The statisitical features of the networks used in the expermiments

表 1 实验所用网络的统计性质

Network	Cora	Blogger	Random	Power-Law	Scale-Free
Nodes	30 751	1 113	29 999	30 000	30 000
Average degree	8.78	6.62	9.01	6.0	8.36
Diameter	19	10	32	24	26
Average distance	5.42	3.36	12.6	5.98	7.51

3.2 评测方法

本文采用了 Dijkstra 算法的 Fibonacci 实现[30]作为距离的基准算法,该实现具有 $O(en+n^2\log n)$的时间复杂度.本文采用该算法计算的距离作为标准答案评测近似算法的准确率,根据它的计算时间评测近似算法的实际性能.作为对比,本文还实现了两种知名的最短路径近似方法,包括 Baswana 提出的随机化近似方法[28](简称 Baswana 方法)和 NSI 近似方法中近似效果最好的 DTZ 方法[23],以验证我们提出的近似方法的有效性.

平均路径比(PathRatio)[23]常用于度量距离近似方法的正确性.给定图 G,从中随机选择 r 对节点,对于任意

节点对 i,分别计算出精确距离 P_{o_i} 和近似距离 P_{f_i} .在此基础上,可以定义平均路径比 P:

$$P=\sum_{i=1}^{r}P_{f_i}\Big/\sum_{i=1}^{r}P_{o_i}.$$

由定义可知,平均路径比 P 的值越接近 1,最短路径近似算法的准确性就越高.P 的值与 1 的差距越大,说明近似结果和实际距离的偏差越大.为满足实际应用需求(如在路径导航问题中,用户能够接受比最短路径偏大的实际路径,而不能接受小于最短路径的不存在的一个近似解),常见最短路径近似方法(包括本文提到的几种方法)都要求近似结果不小于实际距离,所以平均路径比不会小于 1.CDZ 方法使用通过中心节点的一条实际存在的路径近似最短路径,这条路径的长度也不会小于最短路径的长度,其平均路径比同样不小于 1.

3.3 结果与分析

首先,我们测试了几种近似方法的准确性.CDZ 方法利用复杂网络结构特性划分区域,不需要输入参数.DTZ 方法需要确定划分的区域数目和区域划分次数,本文中设定为 10%和 5.Baswana 方法需要指定概率 p 将节点加入样本集,本文中,p 设定为 0.2.几种近似方法在各种类型网络上的正确性评测结果见表 2.

Table 2 The PathRatio measure of the approximate algorithms in different networks

表 2 近似算法在不同网络上的 PathRatio 值

Algorithm	Random	Power-Law	Scale-Free	Cora	Blogger
CDZ	1.454	1.103	1.172	1.018	1. 020
DTZ	6.412	6.246	6.600	5.549	5.648
Baswana	1.124	1.211	1.222	1.160	1.092

如表 2 所示,CDZ 近似方法和 Baswana 方法相对于 DTZ 方法更为精确.如在估算随机网络中节点的距离时,CDZ 方法和 Baswana 方法的平均路径比分别为 1.454 和 1.124,而 DTZ 高达 6.412.这是因为 DTZ 方法利用泛洪算法随机划分区域;为了避免随机划分引入的距离偏置,DTZ 方法需要多次划分区域,并叠加每一种划分情况下的距离近似值.这在消除了随机偏置的同时,也使得估算距离远大于实际的最短路径长度.CDZ 近似方法和 Baswana 方法在各种网络上均达到较高的近似水平;但 CDZ 方法在 Cora 网络和 Blogger 网络中的近似准确性明显高于 Baswana 算法;而在具有无标度生成网络中,CDZ 方法也要好于 Baswana 算法.如在 Barabási 无标度网络上,CDZ 方法的平均路径比达到了 1.172,而 Baswana 算法为 1.222.这是因为 Baswana 利用随机化的方法挑选“Sample”节点,具有较好的通用性,却没有针对现实网络的特征进行优化;而 CDZ 方法是以复杂网络的结构特征为假设基础,认为在复杂网络连通性方面起重要作用的“长边”往往与中心节点相连.在具有无标度特征的网络中,少量的中心节点占有了大部分的边,符合 CDZ 方法的假设.而现实网络大都具有无标度等复杂网络特征,所以 CDZ 方法更适合于现实复杂网络的最短路径的近似.

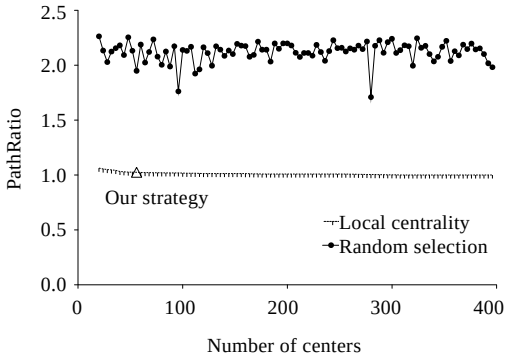


Fig.2 Influence of center selection strategy on CDZ

图 2 中心节点选择策略和对 CDZ 方法的影响

本文也在 2 000 个节点规模的 Cora 子网络上测试了 CDZ 方法中心节点选择策略对近似正确性的影响.本文采用随机方法作为对比策略,以 2 为间隔,从 20~400 设置中心点数目,计算不同的中心点情况下 CDZ 方法的准确性.如图 2 所示,基于局部中心性的选择策略在 CDZ 方法上整体表现出较好的近似准确性,而随机选择策略则有着较大的偏置.这说明,基于局部中心性划分的区域更符合现实网络的拓扑结构特征.而选择不同的中心点数目,对距离估算的准确性也有较大的影响.在随机选择策略中,CDZ 算法的近似准确率随着中心点数目的变化出现了较大的随机波动.基于局部中心性的选择策略在中心点数目很少时,对距离的估计偏大;而随着中心点数目的增加,估算距离快速地逼近实际距离.在中心节点数

目为 40~50 时,估算距离基本与实际距离一致,本文第 2.2 节设计的中心点数目选择策略选择的节点数目为 56,对应的平均路径比为 1.026.随着节点数目的增长,偏置程度仍会缓慢地缩小.这是因为,随着中心节点数目的增加,每个区域的规模变小,节点之间的近似路径越来越接近于区域中心点之间的最短路径;而区域中心点之间最短路径是通过 Dijkstra 算法得到的.因此,随着中心节点数目的增加,近似准确性会缓慢地得以提升,但计算代价也会急剧增大.而本文中心点选择策略使得中心点数目较小且估算结果较为正确,同时满足了计算距离的性能和准确性的要求.

在性能方面,几种最短路径算法的计算复杂性见表 3.CDZ 近似方法和 DTZ 方法在时空开销上都要远远小于经典的距离计算方法,但 Baswana 方法的预处理复杂性非常高,仅比 Dijkstra 方法的 Fibonacci 实现略好;Baswana 方法的时间复杂性依赖于节点数目 n 以及边的数目 e.CDZ 方法的时间复杂性线性地依赖于节点数目 n、边的数目 e 以及选择的中心点数目 d.DTZ 方法的时间复杂性则依赖于图中边的数目 e 与选择的区域数目 d 以及区域划分次数 k 的乘积.正如第 2.2 节所分析的,如果网络具有无标度特征,则本文策略选择的中心点数目 d 远远小于节点数目 n.在这种情况下,CDZ 方法仅仅需要 $O(e+n\log(n))$的计算时间.

Table 3 The computational complexity of the (approximate) shortest-paths algorithms

表 3 距离计算与近似算法的时空复杂性

Algorithm	Time	Space
Dij.	$O(n^2\log n+en)$	$O(n^2)$
CDZ	$O(n\log n+e+d^3)$	$O(n+d^2)$
DTZ	$O(ekd)$	$O(n+d^2)$
Baswana	$O(e^{3/2}n\log n+n^2)$	$O(n^2)$

在普通台式机上(2GHZ CPU,2G 内存),几种算法的 Java 实现在 Cora 网络和 E-R 随机网络不同规模子网络上的实际所用时间如图 3 所示.

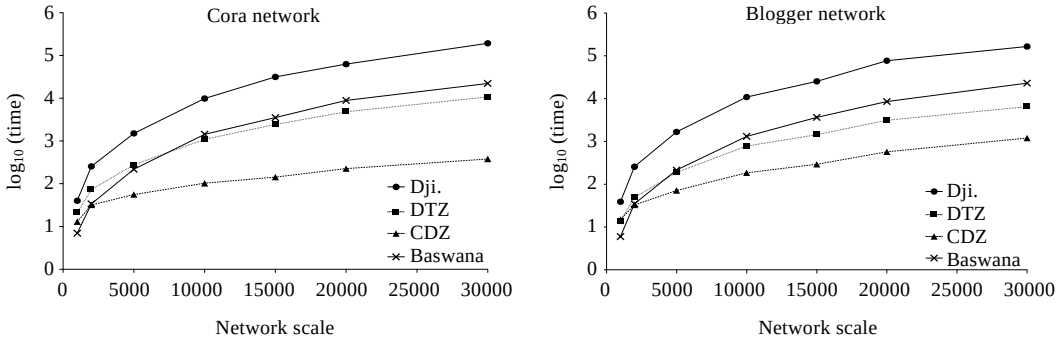


Fig.3 Run time of different approximate algorithms on Cora and Blogger network

图 3 几种距离计算方法在 Cora 和 Blogger 两种网络上的实际执行时间

通过图 3 可以发现,与其他几种算法相比,CDZ 方法具有非常明显的性能优势. Dijkstra 方法的 Fibonacci 实现的性能曲线表明,即使面对中等规模的网络,传统的精确算法也不能在工程应用可以接受的时间内计算出结果.Baswana 方法在网络规模较小时,近似计算性能最好;但随着网络规模的增大,该方法所需计算时间急剧增加;在网络规模大于 10 000 个节点后,该方法的计算时间开销在几种近似方法中最大.这是因为,Baswana 方法需要做多次的宽度优先遍历保存图的结构信息,以确保近似路径长度不超过实际路径长度的 2 倍.DTZ 方法的预处理时间开销也要高于 CDZ 方法;为了避免随机划分的影响,DTZ 方法需要多次划分区域并根据每次的近似值综合估算距离;多次区域划分大大地增加了 DTZ 方法的时间开销.而 CDZ 方法利用局部中心性划分区域,能够较好地符合实际复杂网络的拓扑结构,因此不需要多次划分区域来避免结果的偏置.在 Cora 等具有无标度特征的网络中,根据本文的中心点选择策略,我们仅需选择非常少的中心点,这也使得 CDZ 方法预处理时间大为降低.而在随机网络上,由于节点度分布较为均衡,CDZ 方法会选择较多的节点作为中心节点,使得 CDZ 方法的性

能有所降低.如在 20 000 个节点的随机网络上,CDZ 方法需要近 60s;而在同样规模的 Cora 子网络上,CDZ 方法仅需 20 余秒.综合性能和准确率考虑,我们认为,CDZ 方法更加适合具有无标度特征的复杂网络的最短路径近似;如果对近似精度要求较高而网络规模又不大,则可以使用 Baswana 方法,但该方法不适合较大规模现实网络的最短路径近似计算.

4 社会网络性质近似

个人或组织在其社会网络的中心度是社会网络分析的初始研究目标之一,也一直是研究者们关注的重要领域.但中心性质的度量通常都以距离或最短路径为基础,具有很高的计算复杂性.本节通过将 CDZ 方法与中心性度量方法相结合,近似计算现实网络的中心性质,并在不同网络上评价了近似方法的有效性.

4.1 接近中心性的近似

接近中心性(closeness centrality)[4]是度量节点在网络中是否处于核心位置的一个重要性质.该性质根据节点到其他所有节点的距离来判断节点的重要程度.对于任意节点 v,它的接近中心性 $C_v(v)$定义如下:

$$C_v(v)=\sum dis(v,s),$$

其中,s 为图中不等于 v 的其他节点,$dis(v,s)$为节点 v 和节点 s 之间的距离.

度量任意一个节点的接近中心性,需要计算该节点到所有其他节点的距离.我们可以结合 CDZ 方法估算的距离值得到节点接近中心性的近似值.接近中心性常用于节点之间的重要性排序,因此,近似结果的准确程度不仅在于中心性近似值与实际值的接近程度,更在于依据近似值排列的节点序列和实际排序的吻合程度.本文使用两种经典的排序评测方法对近似排序的正确性进行评测:斯皮尔曼(Sperman)秩相关系数和肯德尔(Kendell)秩相关系数[37].对于任意两个序列 Q_1,Q_2,度量其相似程度的斯皮尔曼系数ρ定义如下:

$$\rho=1-6\sum_i d_i^2\Big/(n^3-n),$$

其中,n 为元素个数,d_i 为第 i 个元素在序列 Q_1,Q_2 中的排名差距.

肯德尔秩相关系数τ定义如下:

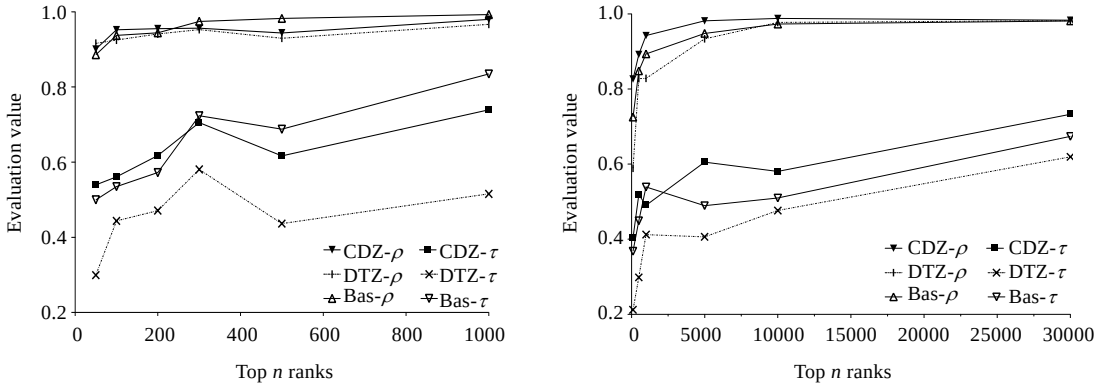
$$\tau=4\sum_i P_i\Big/n(n-1)-1,$$

其中,n 为元素个数,P_i 为第 i 个元素在两个序列中的同序对(concordant pair)数目.

由定义可知,ρ和τ取值都在[−1,1]之间.值越大,说明排序越相似;值为 1,说明两个序列排序完全一致.

本文利用第 3 节提到的 3 种方法估算的距离计算网络中节点的接近中心性的近似值,并根据近似值对节点进行排序,然后利用斯皮尔曼秩相关系数ρ以及肯德尔秩相关系数τ度量与真实排序的相似程度.

图 4 给出了几种近似方法在 Blogger 和 Cora 两种网络上的斯皮尔曼系数ρ和肯德尔系数τ的秩相关曲线.在考虑所有节点的排序时,基于 CDZ 方法的接近中心性在 Cora 网络上的ρ值和τ值达到了 0.983 和 0.733,高于 Baswana 方法的 0.981 和 0.674,更是显著地高于基于 DTZ 方法估算的接近中心性排序.而在 Blogger 网络上,若考虑所有节点,利用 Baswana 方法估算接近中心性能得到较好的排序.第 3 节的实验结果已经表明,Baswana 方法在距离估算上有着较好的准确性;但第 3 节也证明了 Baswana 方法有着非常高的时间复杂性和空间开销.这使得基于 Baswana 方法的接近中心性近似算法在 Cora 网络上的运行时间仅比精确方法略快,远慢于 CDZ 近似方法.因此,Baswana 方法不适合较大规模的复杂网络接近中心性的近似.在利用接近中心性度量节点重要性时,往往只会用到排序最高的一些节点,对 TOP N 排序的准确性往往更为重要.我们评价了接近中心性最高的前 n 个节点的排序准确性,发现 CDZ 方法在基于接近中心性估算最重要节点的排序时相对于其他方法有着明显的优势,如在度量 Cora 网络的 TOP 500 时,CDZ 方法的ρ值和τ值分别达到了 0.892 和 0.519,而 DTZ 方法仅为 0.827 和 0.298,Baswana 方法对应的值也只有 0.848 和 0.449.在 Blogger 网络上也有着同样的特点,基于 CDZ 方法得到的 TOP 100 排序的准确性(ρ为 0.952,τ为 0.561)明显优于 Baswana 方法(ρ为 0.937,τ为 0.535)和 DTZ 方法(ρ为 0.925,τ为 0.443).



(a) Blogger (b) Cora

Fig.4 Correlation coefficient curve of top n rank of closeness centrality approximation

图 4 不同近似算法接近中心性估算 Top n 排序准确性曲线

4.2 介数中心性的近似

介数中心性质(betweenness centrality)[4]来源于社会网络中对个体重要性的评估.介数中心性既能刻画节点和边的重要性,又能用于构建基于介数的聚类算法,发现网络社区结构.因此,它一直是研究网络结构性质的一种重要的量化手段.对于任意节点 v,它的介数定义如下:

$$C_B(v)=\sum_{s\neq t\neq v}\delta_{st}(v),$$

其中,$\delta_{st}=\partial_{st}(v)/\partial_{st}$,$\partial_{st}$ 为节点 s 到 t 的最短路径数目,$\delta_{st}(v)$为节点 s 到 t 的最短路径中经过节点 v 的最短路径数目.相应地,我们也可以导出任意边 e 的介数计算公式.

由定义可知,计算介数中心性需要查找图中任意节点对之间的最短路径,具有非常高的计算复杂性.目前,最快的介数计算方法是 2001 年提出来的 Ulrik Brandes 算法(简称为 UB 算法)[38].设加权图中包含 n 个节点和 e 条边,则 UB 算法的时间复杂度为 $O(en+n^2\log(n))$.CDZ 方法估算的距离对应了通过中心节点的一条可能最短路径.如果利用这条路径近似最短路径,我们可以高效地近似计算复杂网络的介数值.

在 CDZ 方法中,任意节点对(s,t)之间的最短路径可以近似为由点 s 经过其中心点 C_s 和点 t 所在区域的中心点 C_t 到点 t 的一条路径.利用这条路径可以得到∂_{st} 的近似值: $\partial_{st}\approx\partial_{sC_s}\cdot\partial_{C_sC_t}\cdot\partial_{C_tt}$ [38].于是,$\partial_{st}(v)$可近似为

$$\partial_{st}(v)\approx\partial_{sc_s}(v)\times\partial_{c_sc_t}\times\partial_{c_tt}+\partial_{sc_s}\times\partial_{c_sc_t}(v)\times\partial_{c_tt}+\partial_{sc_s}\times\partial_{c_sc_t}\times\partial_{c_tt}(v).$$

此时,节点 v 的介数值为

$$C_B(v)=\sum_{s\neq t\neq v}\partial_{st}(v)\Big/\partial_{st}\approx\sum_{s\neq t\neq v}\left(\partial_{sc_s}(v)\Big/\partial_{sc_s}+\partial_{c_sc_t}(v)\Big/\partial_{c_sc_t}+\partial_{c_tt}(v)\Big/\partial_{c_tt}\right)=\sum_{s\in V}\sum_{t\in V}(\partial_{sc_s}(v)+\partial_{c_sc_t}(v)+\partial_{c_tt}(v)) \tag{3}$$

定义 $\delta'_{c_s*}(v)=\sum_{s\in Z(c_s)}\delta_{sc_s}(v)$, 在无向图时易得 $\delta'_{c_s*}(v)=\delta'_{*c_s}(v)$.代入公式(3),可得:

$$C_B(v)=\sum_{s\in V}\left(n\delta_{sc_s}(v)+\sum_{c_i\in C}d(c_i)\delta_{c_sc_t}(v)+\sum_{c\in C}\delta_{c*}(v)\right).$$

$d(c_i)$为区域 Z_{c_i} 包含的节点数.

我们假设每个区域的大小基本相同,因此可以用 n/d 来对 $d(c_i)$进行近似估算.此时,

$$C_B(v)\approx\sum_{s\in V}\left(n\delta_{sc_s}(v)+(n/d)\sum_{c_i\in C}\delta_{c_sc_i}(v)+\sum_{c\in C}\delta_{c*}(v)\right)=n\sum_{s\in V}\delta_{sc_s}(v)+(n/d)^2\sum_{c_i\in C}\sum_{c_j\in C}\delta_{c_sc_j}(v)+n\sum_{c\in C}\delta_{c*}(v).$$

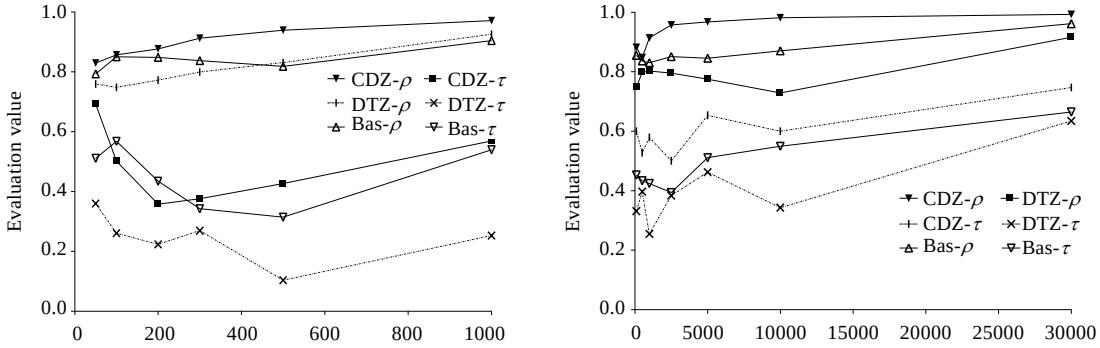
当图为无向图时,简化为

$$C_B(v) \approx (n/d)^2 \sum_{c_i \in C} \sum_{c_j \in C} \delta_{c_s c_j}(v) + 2n \sum_{c \in C} \delta_{c*}(v).$$

因此,在无向图的情况下,我们只需计算出任意区域 $\delta'_{c_s*}(v)$ 的值以及任意两个中心点之间的最短路径,就可以求得 $C_B(v)$的一个近似.文献[38]介绍了如何通过宽度优先的搜索求得 $\delta'_{c_s*}(v)$ 值,我们可以通过一次宽度优先的搜索将节点归入最近中心点所在区域,求得所有区域的 $\delta'_{c_s*}(v)$;并在这一次搜索中,发现相邻区域的中心点之间的所有最短路径.这次搜索所需时间开销仅为 $O(e+n\log(n))$.而通过相邻区域的中心节点之间的最短路径求得所有不相邻区域的中心节点之间的最短路径的时间开销为 $O(d^3)$,其中,d 为中心节点数目.第 3 节已指出,在复杂网络中,本文的节点选择策略选择的中心节点数目 d 要远远小于节点总数 n,选择策略需要 $O(n\log(n))$的时间开销.因此,在很多面向现实网络的应用中,算法的时间复杂性仅为 $O(e+n\log(n))$.

介数中心性主要用于对节点重要性进行排序,因此,同样可以使用秩相关系数评测基于 CDZ 的介数近似方法的有效性.与 CDZ 方法不同,DTZ 方法并不能给出一条实际存在的路径作为最短路径近似,因此不能直接用于估算介数中心性.文献[23]利用 DTZ 构建的索引使用最佳优先搜索(best first search)来查找节点间的可能最短路径,并由此估算介数值.由于复杂性较高,文献[23]只采样图中部分节点对进行介数值的估算.为了便于比较,本文采用与文献[23]同样的方法构造了基于 Baswana 方法的介数中心性近似方法.我们得到了几种方法在 Cora 网络和 Blogger 网络上计算的介数近似值,并评价了通过不同方法得到的前 n 个节点排名的准确性.

如图 5 所示,在考虑不同的 TOP n 的情况下,基于 CDZ 方法近似得到的介数中心性排序更符合实际排序.在考察所有节点的排序时,基于 CDZ 方法的介数中心性排序在 Cora 网络上的ρ值和τ值分别是 0.993 和 0.747,而基于 DTZ 方法的相应指标仅为 0.916 和 0.635,基于 Baswana 方法得到的排序准确性也低于 CDZ 方法(ρ为 0.962,τ为 0.665).在仅考虑最重要的前 n 个节点的排序时,基于 CDZ 方法的介数中心性排序相对于其他方法有更明显的优势.在度量 Cora 网络中 TOP 500 节点的排序时,基于 CDZ 的介数中心性近似算法的ρ值和τ值分别是 0.847 和 0.528,明显高于 DTZ 方法(ρ为 0.798,τ为 0.397),也高于 Baswana 方法(ρ为 0.836,τ为 0.434).



(a) Blogger (b) Cora

Fig.5 Correlation coefficient curve of top n rank of betweenness centrality approximation

图 5 不同近似算法在估算介数中心性 Top n 排序准确性曲线

5 结论与展望

本文在分析复杂网络的结构性质的基础上,提出基于区域中心点距离(CDZ)最短路径近似方法,在快速估算距离的同时,也给出符合复杂网络拓扑特征的一条可能最短路径作为最短路径的近似.此外,本文将 CDZ 方法与社会网络性质度量方法相结合,用以解决社会网络的性质快速估算问题.与现有近似方法相比,CDZ 方法利用经过中心节点的一条路径对节点间的最短路径进行近似,更符合现实网络中最短路径的实际特点,也能方便地用于基于最短路径的网络性质的估算.本文在随机网络、小世界网络等生成网络和 Blogger,Cora 等实际复杂网络上设计了一系列实验,实验结果证明了 CDZ 方法的有效性.在 Blogger 社交网络以及 Cora 科研文献引用网

络等复杂网络的实验中,CDZ 方法在距离、接近中心性、介数中心性等的近似都表现出了很好的有效性,这也支持了本文关于复杂网络最短路径的假设.本文提出的方法为面向较大规模具有复杂网络特征的现实网络近似分析提供了新的思路.尽管如此,作为一个新兴的研究领域,复杂网络的结构特征仍然需要更深入的研究.目前,已经发现一些现实网络并未完全表现出复杂网络特征,如高速公路网络的度分布并未表现出无标度特性.因此,如何在各种类型的现实复杂网络上进一步评估和改进 CDZ 算法,并深入研究复杂网络的最短路径特征,是我们进一步工作的目标.此外,如何将本文提出的 CDZ 方法应用到其他基于图的算法的快速近似,使得面向较大规模现实网络的近似分析更为可行,也是我们的一个研究目标.

References:

[1] Milgram S. The small world problem. Psychology Today, 1967,1(1):60−67.

[2] Wikipedia contributors. Six Degrees of Kevin Bacon. Wikipedia: The Free Encyclopedia, 2006.

[3] Xu Y, Zhao H, Su WJ, Zhang WB, Zhang X. Analysis on traveling diameter of Internet. Chinese Journal of Computers, 2006,29(5):690−698 (in Chinese with English abstract).

[4] Freeman LC. Centrality in social networks: Conceptual clarification. Social Networks, 1979,1(3):215−239. [doi: 10.1016/0378-8733(78)90021-7]

[5] Yang B, Liu DY, Liu JM, Jin D, Ma HB. Complex network clustering algorithms. Journal of Software, 2009,20(1):54−66 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3464.htm [doi: 10.3724/SP.J.1001.2009.03464]

[6] Girvan M, Newman MEJ. Community structure in social and biological networks. National Academy of Sciences of the United States of America, 2002,99(12):7821−7826. [doi: 10.1073/pnas.122653799]

[7] Dijkstra EW. A note on two problems in connextion with graphs. Numerische Mathematik, 1959,1(1):269−271. [doi: 10.1007/BF01386390]

[8] Floyd RW. Algorithm 97 (shortest path). Communications of the ACM, 1962,5(6):345. [doi: 10.1145/367766.368168]

[9] Deo N, Pang CY. Shortest path algorithms: Taxonomy and annotation. Networks, 1984,14(2):275−323. [doi: 10.1002/net.3230140208]

[10] Cherkassky BV, Goldberg AV, Radzik T. Shortest paths algorithms: Theory and experimental evaluation. Mathematical Programming, 1996,73(2):129−174. [doi: 10.1007/BF02592101]

[11] Watts DJ, Strogatz SH. Collective dynamics of small-world networks. Nature, 1998,393(6684):440−442. [doi: 10.1038/30918]

[12] Barabási AL, Albert R. Emergence of scaling in random networks. Science, 1999,286(5439):509−512. [doi: 10.1126/science.286.5439.509]

[13] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: Structure and dynamics. Physics Reports, 2006,424(4-5):175−308. [doi: 10.1016/j.physrep.2005.10.009]

[14] Lü JH,Wang HC, He KQ. Complex dynamical networks and their applications in software engineering. Journal of Computer Research and Development, 2008,45(12):2052−2059 (in Chinese with English abstract).

[15] Stolfo SJ, Hershkop S, Wang K, Nimeskern O, Hu CW. Behavior profiling of email. In: Proc. of the 1st NSF/NIJ Symp. on Intelligence, Security Informatics. Tucson: Springer-Verlag, 2003. 74−90.

[16] Newman MEJ. The structure of scientific collaboration networks. National Academy of Sciences of the United States of America, 2001,98(2):404−409. [doi: 10.1073/pnas.021544898]

[17] Ishida K. Extracting latent weblog communities: A partitioning algorithm for bipartite graphs. In: Proc. of the 2nd Annual Workshop on the Weblogging Ecosystem. 2005. http://www.blogpulse.com/www2005-workshop.html

[18] Lerman K, Jones L. Social browsing on Flickr. In: Proc. of the Int'l Conf. on Weblogs and Social Media. 2006. http://www.icwsm.org/papers/paper27.html

[19] Yang YB, Li N, Zhang Y. Networked data mining based on social network visualizations. Journal of Software, 2008,19(8):1980−1994 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/1980.htm [doi: 10.3724/SP.J.1001.2008.01980]

[20] Qiao SJ, Tang CJ, Peng J, Liu W, Wen FL, Qiu JT. Mining key members of crime networks based on personality trait simulation email analysis system. Chinese Journal of Computers, 2008,31(10):1795−1803 (in Chinese with English abstract).

[21] Chow E. A graph search heuristic for shortest distance paths. Technical Report, UCRL-JRNL-202894, Livermore: Lawrence Livermore National Laboratory, 2004.

[22] Slivkins A. Distance estimation and object location via rings of neighbors. In: Proc. of the ACM Symp. on Principles of Distributed Computing. Las Vegas: ACM, 2005. 41−50. http://dl.acm.org/citation.cfm?doid=1073814.1073823 [doi: 10.1145/1073814.1073823]

[23] Rattigan M, Maier MJ, Jensen D. Using structure indices for efficient approximation of network properties. In: Proc. of the 12th ACM Int'l Conf. on Knowledge Discovery and Data Mining. Philadelphia: ACM, 2006. 357−366. http://dl.acm.org/citation.cfm?doid=1150402.1150443 [doi: 10.1145/1150402.1150443]

[24] Rattigan M, Maier MJ, Jensen D. Graph clustering with network structure indices. In: Proc. of the 24th Int'l Conf. on Machine Learning. Corvallis: ACM, 2007. 783−790. http://dl.acm.org/citation.cfm?doid=1273496.1273595 [doi: 10.1145/1273496.1273595]

[25] Zwick U. Exact and approximate distances in graphs—A survey. In: Proc. of the 9th Annual European Symp. on Algorithms. London: Springer-Verlag, 2001. 33−48. http://dl.acm.org/citation.cfm?id=740642 [doi: 10.1007/3-540-44676-1_3]

[26] Cohen E, Zwick U. All-Pairs small stretch paths. Journal of Algorithms, 2001,38(2):335−353. [doi: 10.1006/jagm.2000.1117]

[27] Thorup M, Zwick U. Approximate distance oracle. Journal of the ACM, 2005,52(1):1−24. [doi: 10.1145/1044731.1044732]

[28] Baswana S, Goyal V, Sen S. All-Pairs nearly 2-approximate shortest paths in $O(n^2$ polylog $n)$ time. Theoretical Computer Science, 2009,410(1):84−93. [doi: 10.1016/j.tcs.2008.10.018]

[29] Zhan FB, Noon CE. Shortest path algorithms: An evaluation using real road networks. Transportation Science, 1998,32(1):65−73. [doi: 10.1287/trsc.32.1.65]

[30] Fredman ML, Tarjan RE. Fibonacci heaps and their uses in improved network optimization algorithms. Journal of the ACM, 1987,34(3):596−615. [doi: 10.1145/28869.28874]

[31] Erdös P, Rényi A. On the evolution of random graphs. Publ. Math. Inst. Hungar. Acad., 1960,Sci.5:17−61.

[32] Granovetter MS. The strength of weak ties. American Journal of Sociology, 1973,78(6):1360−1380. [doi: 10.1086/225469]

[33] Bearman PS, Moody J, Stovel K. Chains of affection: The structure of adolescent romantic and sexual networks. American Journal of Sciology, 2004,110(1):44−91. [doi: 10.1086/386272]

[34] Eppstein D, Wang J. A steady state model for graph power laws. In: Proc. of the Int'l Workshop on Web Dynamics. 2002. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.6295

[35] Nieminen J. On centrality in a graph. Scandinavian Journal of Psychology, 1974,15(1):332−336. [doi: 10.1111/j.1467-9450.1974.tb00598.x]

[36] Brandes U. A faster algorithm for betweenness centrality. The Journal of Mathematical Sociology, 2001,25(2):163−177. [doi: 10.1080/0022250X.2001.9990249]

[37] McCallum A, Nigam KA, Rennie JK, Seymore K. A machine learning approach to building domain-specific search engines. In: Proc. of the 16th Int'l Joint Conf. on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1999. 662−667. http://dl.acm.org/citation.cfm?id=687598

[38] Myers JL, Well AD. Research Design and Statistical Analysis. 2nd ed., Lawrence Erlbaum, 2003.

附中文参考文献:

[3] 徐野,赵海,苏威积,张文波,张昕.Internet 网络的访问直径分析.计算机学报,2006,29(5):690−698.

[5] 杨博,刘大有,LIU Jiming,金弟,马海宾.复杂网络聚类方法.软件学报,2009,20(1):54−66. http://www.jos.org.cn/1000-9825/3464.htm [doi: 10.3724/SP.J.1001.2009.03464]

[14] 吕金虎,王红春,何克清.复杂动力网络及其在软件工程中的应用.计算机研究与发展,2008,45(12):2052−2059.

[19] 杨育彬,李宁,张瑶.基于社会网络可视化分析的数据挖掘.软件学报,2008,19(8):1980−1994. http://www.jos.org.cn/1000-9825/19/1980.htm [doi: 10.3724/SP.J.1001.2008.01980]

[20] 乔少杰,唐常杰,彭京,刘威,温粉莲,邱江涛.基于个性特征仿真邮件分析系统挖掘犯罪网络核心.计算机学报,2008,31(10):1795−1803.

唐晋韬(1981－),男,湖南安化人,博士生,CCF 学生会员,主要研究领域为信息抽取,社会网络分析,信息检索.

王戟(1969－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为高可信软件技术,软件方法学,软件工程.

王挺(1970－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为语义 Web,信息抽取,信息检索.