利用 KSN 算法发现网络中有影响力的结点

田 艳 刘祖根
(云南财经大学信息学院 昆明 650221)

摘 要 准确高效地发现网络中有影响力的传播者具有非常重要的理论和现实意义。近年来,结点影响力排序受到了多领域学者的广泛关注。K-shell 是一种较好的结点影响力评价指标;然而,仅仅依赖结点自身 K-shell 值实现的算法通常具有评估结果精确度不高、适用性较差等缺陷。针对此问题,提出 KSN(the K-shell and neighborhood centrality)中心性模型,该算法综合考虑了结点本身及其所有二阶以内邻居结点的 K-shell 值。实验结果表明,所提出算法度量结点传播的能力比度中心性、介数中心性、K-shell 分解、混合度分解等方法更准确。

关键词 复杂网络,有影响力结点,中心化测量,KSN 中心性
中图法分类号 TP393.02 **文献标识码** A

Detecting Most Influential Nodes in Complex Networks by KSN Algorithm

TIAN Yan LIU Zu-gen
(School of Information, Yunnan University of Finance and Economics, Kunming 650221, China)

Abstract It is very important in theory and practice to detect influential spreaders in networks accurately and efficiently. Recently, scholars from various fields have paid their attention to the study of ranking nodes. The K-shell index is a relatively powerful indicator to estimate the spreading ability of nodes. However, due to only attributes of the node itself being considered, limitation in accuracy and universal applicability will exist for K-shell decomposition. To solve this problem, this paper proposed a novel algorithm called KSN (the K-shell and neighborhood centrality) to estimate the spreading influence of a node by its K-shell value and the K-shell indexes of its nearest and next nearest neighbors. Experimental results demonstrate that this proposed algorithm acts more precisely in detecting the most influential nodes than degree centrality, betweenness centrality, K-shell decomposition and the mixed degree decomposition, et al.

Keywords Complex networks, Influential nodes, Centrality measure, KSN centrality

1 引言

现实世界中的复杂系统均可使用网络表示,如社会网络、生物网络、文献引用网等。最近几年新兴的网络实例更贴近人们的生活,例如微博传播网络模型、城市公交网络模型及图书借阅网络模型等[1]。网络中的基本元素是结点(Node)和边(Edge);其中,结点表示各系统的构成要素,边代表两个要素之间的关系。准确度量复杂网络中结点的传播能力具有十分重要的理论和现实意义[2,3],如有效防控疫病、抑制谣言扩散、预防网络攻击[4]、解决电力网络相继故障[5]等。

社会网络分析中,结点的重要性也称为"中心性",其主要观点是结点的重要性等价于该结点与其它结点的连接使其具有的显著性[6]。最简单直接的度中心性[7,15](Degree Centrality, DC)只考虑了结点的近邻信息,它指出结点的邻居数越多,则影响力越大。介数中心性[8](Betweenness Centrality, BC)描绘了结点对沿网络中最短路径传输信息流的控制程度。然而,BC 只考虑了最短路径这种极端情况。Yan 等人[9]的研究表明:在很多情况下选择最短路径传输网络信息流,可能会有适得其反的效果。紧密度中心性[10](Closeness Centrality, CC)通过计算结点与网络中其它所有结点距离的平均值来确定结点的重要性。BC 和 CC 从全局视角较好地量化了结点的影响力,但存在计算复杂度过高而很难应用于大规模网络的缺点。另外,CC 也无法计算两个完全不连通结点之间的距离。Kitsak 等人[11]研究表明,由 K-shell 分解(K-shell Decomposition, KS)划分出的网络核心结点才是最具影响力的结点。KS 能刻画网络结点的结构特性,但它无法区分同层结点的影响力。

与度中心性和介数中心性等指标相比较,传统的 K-shell 分解能较好地识别最具影响力结点,却在较多场合(如树形网络、星型网络)无法发挥其优势。为弥补 KS 方法的不足,研究者提出了一批改进算法。Zeng 等人[12]提出混合度分解法(Mixed Degree Decomposition, MDD),进一步提高了结点重要性的区分度,但是算法中的可变参数难以确定。Liu 等人[13]改进了 KS 算法(KS_+),旨在区分同层结点的传播影响力,又因为涉及到目标结点到网络核心结点的距离计算,所以其时间复杂度也较高。

本文受云南省社科规划项目(YB2012080),云南省自然科学基金项目(2011FZ148),云南财经大学研究生创新基金项目(云财研创(2014)24),云南财经大学校级项目(YC2012D11),云南财经大学研究生创新基金项目(2015YUFEYC004)资助。

田 艳(1989 —),女,硕士生,主要研究方向为社会网络,E-mail:tianyan0820@126.com;刘祖根(1971 —),男,博士,副教授,主要研究方向为社会网络、隐写分析等,E-mail:jzlzg@163.com。

发现网络中重要结点的传统研究都强调结点自身的属性影响(如度、介数等),却忽视了由结点近邻凝聚出的影响力,这理应成为非常重要的影响因子。K-shell 也不例外,同此前算法相比,虽然它能够更恰当地评价结点的影响力,但是仅仅依赖结点自身 K-shell 值实现的算法存在评估精度不够高、算法适用性较差等缺陷。为此,本文提出 KSN(the K-shell and Neighborhood Centrality)中心性模型,认为结点的影响力不仅由结点自身的 K-shell 属性决定,而且同其二阶以内邻居结点的 K-shell 值息息相关,其正是针对这些缺陷而设计的。

本文第 2 节介绍相关算法模型以及所采用的算法评估模型;第 3 节通过真实网络数据的 SIR (Susceptible-Infected-Recovered)模型传播仿真,分析比较 DC、BC、KS、MDD、KS_+ 和 KSN 指标的结果;最后总结全文,并展望今后的研究方向。

2 相关研究及评估模型

2.1 算法模型

假设 $G=(V,E)$ 为一无向、无权网络,由 $n=|V|$ 个结点和 $m=|E|$ 条边组成。

度中心性[10](DC)是描述网络结点影响力最简单、最直观的方法。结点 v 的度是指网络中与 v 直接相连的结点个数,定义如下:

$$k(v)=\sum_{w}^{n}|a_{vw}|,|a_{vw}|=\begin{cases}1, & \text{若 } v \text{ 与 } w \text{ 相连}\\ 0, & \text{否则}\end{cases} \tag{1}$$

DC 仅局限于结点最基本、最局部的静态指标。如果结点的度很高,而其邻居结点的度都极低,那么该结点的影响程度也不会高。

介数中心性[8](BC)以经过结点的最短路径数目来衡量该结点的重要程度。结点 v 的介数指标定义为:

$$b(v)=\sum_{w,z\neq v}\frac{|g_{wz}(v)|}{|g_{wz}|} \tag{2}$$

其中,$|g_{wz}|$ 是从结点 w 到 z 的所有最短路径的数目,$|g_{wz}(v)|$ 表示结点 w 经过 v 到达 z 的最短路径数。

Kitsak 等[11]指出,网络传播中的最具影响力的结点并非高度数或高介数的 Hub 结点,而是经过 K-shell 分解所得的具有最大 K-shell 值的结点。K-shell 分解[11](KS)是一种基于网络结点度的结点重要性的粗粒划分,它通过递归地剥除网络边缘所有度 $k(k\leqslant K)$ 的结点来揭示网络的层次特征,位于核心层的结点具有高影响力。结点 v 的 K-shell 值记作 $ks(v)$。详细分解步骤如下:首先去除网络中度 $k=1$ 的结点,若剩余的网络中仍存在 $k=1$ 的结点,那么再去掉这些结点。循环此过程,直到剩余网络中不再有 $k=1$ 的结点。那么,此时所有被删除结点的 K-shell 值记作 $ks=1$。然后,依次令 $k=2,3,4,\cdots$,即可完成网络的 K-shell 分解。

与度中心性、介数中心性相比,传统的 K-shell 分解能较好地识别网络中最具影响力的结点,却在很多场合(如树形网络,所有结点的 ks 值均为 1)无法发挥其优势。因此,仅仅依赖结点自身 ks 值实现的算法通常具有评估结果精确度不高、普遍适用性较差等缺陷。混合度分解(MDD)[12]在 K-shell 分解的础上,同时考虑结点的剩余度(Residual Degree)和结点的移除度(exhausted degree),定义如下:

$$MDD(v)=kr+\lambda * k_e \tag{3}$$

其中,变量 $\lambda\in[0,1]$,k_r 是结点的剩余度,k_e 是结点的移除度。MDD 在一定程度上优化了结点影响力的区分度,然而,如何根据不同的网络结构确定参数 λ 以获得最优解,却缺乏行之有效的方案。

Liu 等人[13]改进的 K-shell 算法(KS_+)如下:

$$KS_+(v)=(ks_{\max}-ks(v)+1)\sum_{w\in Sc}d(v,w) \tag{4}$$

其中,Sc 是网络的核心结点集,$ks_{\max}$ 是网络结点的最大 ks 值,也就是集合 Sc 中的结点 ks 值。$d(v,w)$ 是从结点 v 到结点 w ($w\in Sc$)的最短路径。这种方法仅适用于连通网络,并且计算复杂度甚高。

综上所述:针对结点按照其影响力排序这一难题,缺乏简单而高效的解决方案。显然,结点的重要性不仅受其自身 ks 制约,其邻居结点的 ks 属性也是决定结点影响力的关键因素。基于这种思路,本文提出 KSN 中心性模型,结点 v 的 KSN 值定义如下:

$$KSN(v)=ks(v)+\sum_{w\in\Gamma(v)}ks(w) \tag{5}$$

其中,$\Gamma(v)$ 是结点 v 的二阶以内邻居结点的集合,即从 v 出发 2 步内可到达的邻居结点集合。显然,KSN 方法以更加全面的网络拓扑特征来度量结点的重要程度。若结点本身距离网络核心越近,其近邻结点亦是如此,显然,结点的重要性越高。K-shell 分解算法的时间复杂度为 $O(m)$[14](m 为网络总边数),因此,KSN 的时间复杂度比 BC、KS_+ 等算法的时间复杂度更低。

2.2 评估模型

在社会网络中,SIR[15-19]模型已广泛应用于疫病、信息及谣言传播的研究。为验证所提出 KSN 算法模型的有效性,本文采用 SIR 模型来模拟仿真网络结点传播过程。该模型包含 3 类状态:易感状态 S、感染状态 I、免疫状态 R。感染结点以概率 β 传染易感状态的邻居个体,被感染结点以概率 γ(简单起见,本文取 $\gamma=1$)恢复为免疫状态 R,此过程循环进行,直到网络中不再存在处于感染状态 I 的结点。初始结点 v 被感染后,导致网络中最终存在的免疫结点的平均数量即为该结点的传播能力,记作 $\sigma(v)$。

本文采用 Kendall's tau[20] 系数比较不同算法之间的相关性。该系数用来揭示两个序列之间相关性的指标,其计算依赖于对两个随机变量 X 和 Y(本文中 X 指任一中心性方法结果,Y 指网络所有结点的 SIR 仿真值)的观测值集合的统计。对于任意两个观测值对 (x_i,y_i) 和 (x_j,y_j):当 $x_i>x_j$ 且 $y_i>y_j$ 或 $x_i<x_j$ 且 $y_i<y_j$ 时,它们被认为是一致的;当 $x_i>x_j$ 且 $y_i<y_j$ 或 $x_i<x_j$ 且 $y_i>y_j$ 时,它们被认为是不一致的;当 $x_i=x_j$ 或 $y_i=y_j$ 时,它们既非一致也非不一致。两个向量 $\boldsymbol{R}_1$、$\boldsymbol{R}_2$ 的 Kendall's tau 系数 τ 定义如下:

$$\tau(R_1,R_2)=\frac{N_c-N_d}{\frac{1}{2}n(n-1)} \tag{6}$$

其中,N_c、N_d 分别是一致和非一致结点对的个数,n 是序列大小。那么,τ 越大则两序列 $\boldsymbol{R}_1$、$\boldsymbol{R}_2$ 的一致性越高。

为进一步量化排序列表的单调性,定义排序向量 $\boldsymbol{L}$ 的单调性如下:

$$M(\boldsymbol{L})=\left|1-\frac{\sum_{r\in R}Nr(Nr-1)}{N(N-1)}\right| \tag{7}$$

其中,N 是排序向量 $\boldsymbol{L}$ 的大小,Nr 是具有相同排序 r 的结点数量。$M(\boldsymbol{L})\in[0,1]$,若 $M(\boldsymbol{L})=1$,则向量 $\boldsymbol{L}$ 单调性最好;若 $M(\boldsymbol{L})=0$,则排序列表 $\boldsymbol{L}$ 中所有结点都具有相同的排序,即算法排序效果差。

3 实验与分析

选用 4 个网络拓扑结构完全不同的真实社会网络数据集进行仿真实验：

① Zachary 空手道俱乐部成员关系网络[21]（Karate Club）；

②Netscience 科学家合著关系网络[22]；

③URV 电子邮件网络[23]（Email）；

④Blogs 交流关系网络[24]。

本文所有实验结果均取 1000 次独立仿真实验的平均值。

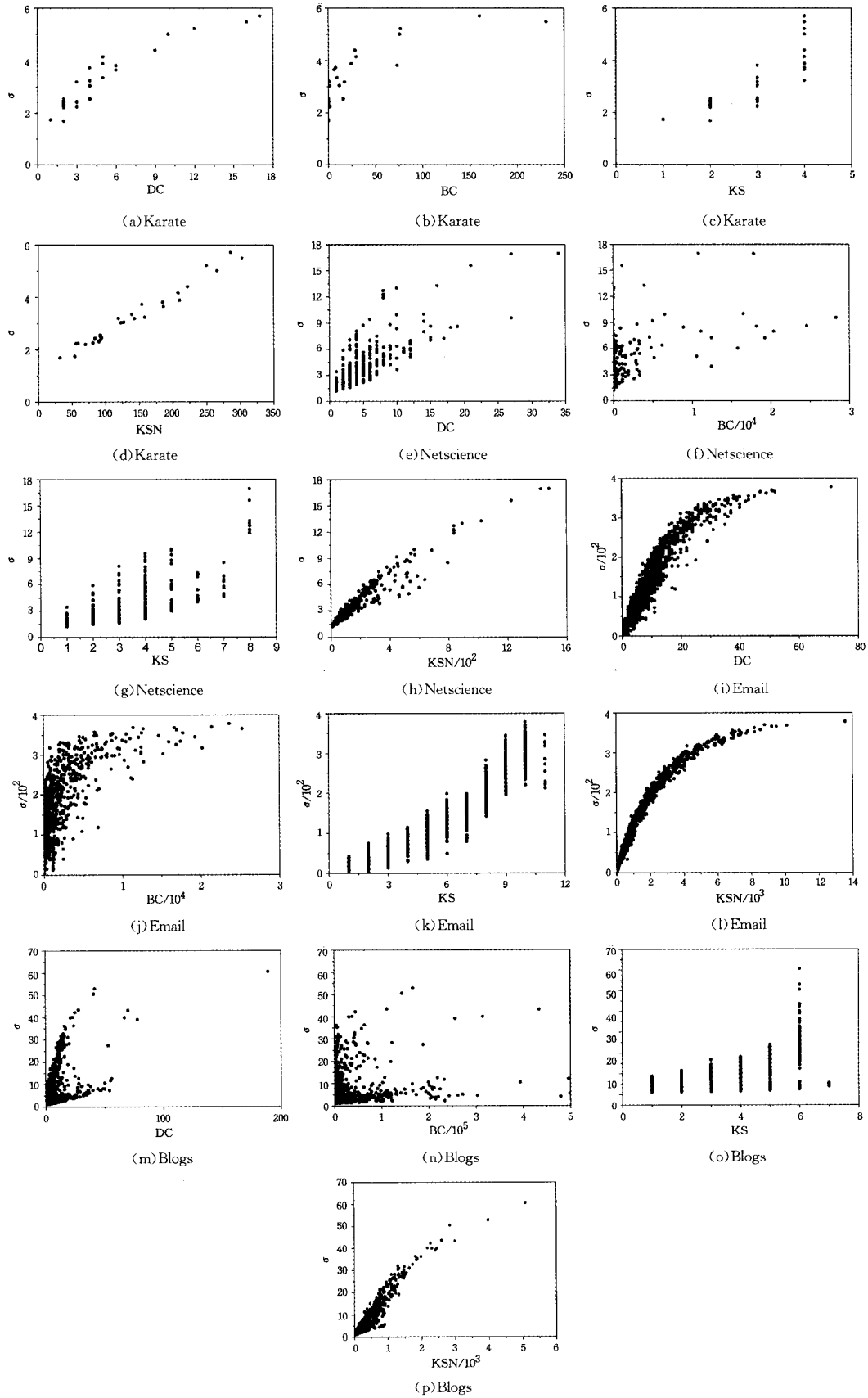


图 1　结点影响力对比(其中横坐标表示各算法指标值，纵坐标表示结点 SIR 仿真得到的结点传播影响力)

实验一　验证算法指标与结点真实影响力的相关程度。实验将4种算法的结点影响力与SIR仿真结果σ进行对比，如图1所示。观察发现，与DC、BC、KS指标相比，不同网络中KSN得到的结点感染范围与真实结点影响力σ的相关性均更强：当结点的KSN指标较小时，对应的σ值较小；反之，亦成立。如在网络Email中，4种中心性方法效果都较好，但KSN表现更为理想，DC仅次于KSN；在网络Blogs中，DC、BC、KS与传播影响力σ并没有太明显的关系。如图1(o)所示，$ks=7$这种结点的影响力却小于那些ks值更小的结点的影响力，这种不合理的现象暴露出KS的缺陷。因此，传统中心性指标各有千秋，它们能在一定程度上发现有影响力的结点，其总体表现却均比KSN略逊一筹。采用KSN算法更容易准确识别网络中的重要结点。

实验二　验证算法排序列表的单调性。算法排序单调性是结点影响力排序区分度的评判指标之一，实验采用式(7)对各算法排序结果的单调性进行评估。从表1可以看出，KSN模型的单调性在4种网络中表现均非常突出而且特别稳定。同其它算法相比较，排序单调性KSN＞KS_+＞MDD，其远超过DC、BC、KS 3种指标的表现。在不同网络环境中，BC的单调性波动最为明显，如在网络Email中，单调性值为0.9695，而在网络Netscience中，仅有0.5820。因此，KSN算法的结点影响力区分度十分理想。

实验三　验证算法相关性τ随感染概率β的变化情况。图2所示为：DC、BC、KS、MDD、KS_+及KSN刻画了结点影响力与真实影响力之间的相关性τ因感染概率β不同而产生的变化情况。从曲线全局走势分析，在网络Karate Club中，各算法相关性波动较大，仅有KSN随感染概率β的增大呈上升趋势；在网络Karate Club、Netscience、Blogs中，除KSN外，其余算法的相关性均呈下降趋势；MDD和KS_+走势相近，DC、BC、KS在4种不同网络中相关性不太稳定，并且BC较其它算法相关性始终较差。因此，KSN表现更令人满意，其相关性受感染概率β和网络结构的影响较小，能更有效地预测结点的传播影响力。

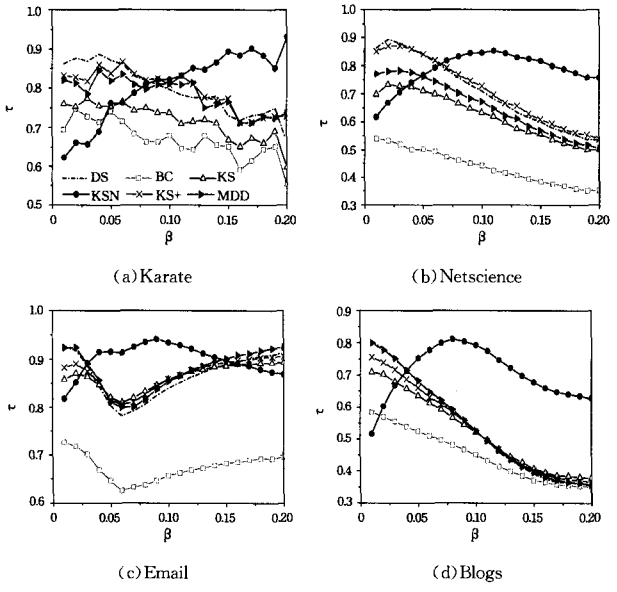


图2　不同感染概率β下，各算法指标与结点真实影响力的相关性

上述实验表明：KSN算法的准确性及普遍适用性均优于传统算法。

表1　排序列表单调性对比

网络	M (DC)	M (BC)	M (KS)	M (MDD)	M (KS+)	M (KSN)
Karate Club	0.8414	0.8788	0.7041	0.8681	0.9376	**0.9768**
Netscience	0.8742	0.5820	0.8013	0.9064	0.9808	**0.9953**
Email	0.9420	0.9695	0.8993	0.9607	0.9891	**0.9996**
Blogs	0.7519	0.6328	0.6834	0.7685	0.9377	**0.9939**

结束语　准确高效地发现网络知识社区中有影响力的传播者，具有非常重要的理论和现实意义。近年来，结点影响力排序受到了多领域学者的广泛关注。与度中心性、介数中心性相比较，传统的K-shell分解能较好地识别最具影响力结点，却在很多场合(如树形网络、星型网络)无法发挥其优势。本文通过研究发现，仅仅依赖结点自身K-shell值实现的算法通常具有评估结果精确度不高、适用性较差等缺陷。众多改进算法也是各有利弊，例如混合度分解法(MDD)中的参数确定就是一大难点。KSN中心性模型综合考虑了结点本身的K-shell值及其所有二阶以内邻居结点的K-shell属性。在4种真实网络数据上通过SIR模型得到的仿真实验结果表明，KSN算法在准确性和稳定性方面胜过度中心性、介数中心性、K-shell分解等传统算法。KSN指标与结点真实影响力高度相关，由它得到的结点排序列表具有非常理想的单调性。因此，本算法能够更加精确地鉴别结点的传播重要性。

KSN算法所得实验结果表明：结点自身K-shell值和二阶以内邻居结点的K-shell值在决定结点影响力方面均具有非常重要的作用。在不同的网络结构中，两者的影响因子可能会有所差异。因此，下一步研究将对结点自身K-shell值和二阶以内邻居结点的K-shell值的确切关系进行实验研究，并尝试从理论上开展一定的工作。另外，本文工作是针对无向、无权的简单网络进行的研究，进一步工作将扩展到有向、加权网络的研究方面。

参考文献

[1] 马杰良，宋艳，潘贞贞，等. 图书漂流网络模型实证研究[J]. 计算机科学，2015，42(3)：51-54

[2] 周涛，傅忠谦，牛永伟，等. 复杂网络上传播动力学研究综述[J]. 自然科学进展，2005，15(5)：513-523

[3] 李翔，刘宗华，汪秉宏. 网络传播动力学[J]. 复杂系统与复杂性科学，2010，Z1：41-45

[4] Liu J G，Wang Z T，Dang Y Z. Optimization of scale-free network for random failures[J]. Mod. Phys. Lett. B，2006，20：815-820

[5] Brummitt C D，D'Souza R M，Leicht E. Suppressing cascades of load in interdependent networks[J]. Proc Natl Acad Sci USA，2012，109：E680-E689

[6] Burt R S，Minor M J，Alba R D. Applied network analysis：A methodological introduction [M]. Sage Publications Beverly Hills，1983

[7] Sabidussi G. The centrality index of a graph[J]. Psychometrika，1966，31：581-603

[8] Freeman L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40: 35

[9] Yan G, Zhou T, Hu B, et al. Efficient routing on complex networks[J]. Phys Rev E, 2006, 73: 1-5

[10] Freeman L C. Centrality in social networks conceptual clarification[J]. Soc Netw, 1979, 1: 215-239

[11] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nat Phys, 2010, 6: 888-893

[12] Zeng A, Zhang C J. Ranking spreaders by decomposing complex networks[J]. Phys. Lett. A, 2013, 377(14): 1031-1035

[13] Liu J G, Ren Z M, Guo Q. Ranking the spreadinginfluencein complex networks[J]. Physica A, 2013, 392: 4154-4159

[14] Newman M E J. A measure of betweenness centrality based on random walks[J]. Soc. Netw., 2005, 27 (1): 39-54

[15] Bonacich P. Factoring and weighting approaches to status scores and clique identification[J]. J Math Sociol, 1972, 2: 113-120

[16] Hethcote H W. The mathematics of infectious disease[J]. Soc. Industr. Appl. Math, 2000, 42: 599-653

[17] Blower S, Bernoulli D. An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it[J]. Rev. Med. Virol, 2004, 14: 275-288

[18] Anderson R M, Robert M. Infectious Diseases of Humans: Dynamics and Control[M]. New York: Oxford University Press, 1992: 66

[19] Diekmann O, Heesterbeek J A P. Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation [M]. New York: Wiley Series in Mathematical & Computational Biology, 2001

[20] Kendall M. A new measure of rank correlation[J]. Biometrika, 1938, 30: 81-93

[21] Zachary W W. An information flow model for conflict and fission in small groups[J]. J. Anthropol. Res., 1977, 33: 452-473

[22] Newman M E J. Finding community structure in networks using the eigenvectors of matrices[J]. Phys. Rev. E, 2006, 74 (3): 036104

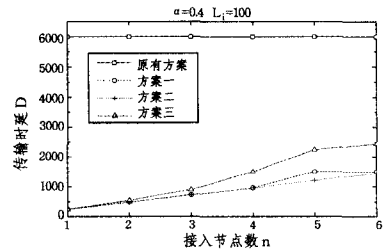
[23] Guimera R, Danon L, Diaz-Guilera A, et al. Self-similar community structure in a network of human interactions[J]. Phys. Rev. E, 2003, 68: 065103

[24] Xie N. Social network analysis of blogs[D]. UK: University of Bristol, 2006

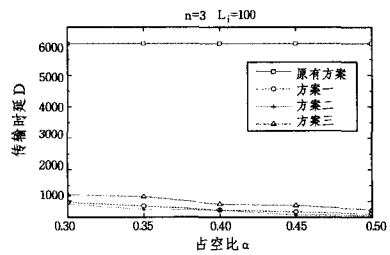
(上接第 276 页)

于该方案中优先级低的那些节点在每个周期中发送的比特数目很少，从而有很大的包传输时延。不过，当方案三中各个接入节点的优先级相同时，该方案退化为方案一，这时的传输时延也与方案一相同。

图 7(b)给出了不同占空比下的各种方案的传输时延，其中 $n=3$，$L_i=100$。同上，方案三的包传输时延为各个接入节点包传输时延的平均值。可以看出，原有方案的传输时延与占空比无关，而其他改进方案都随着占空比的增大而减少。总的来说，所提出的方案的包传输时延远小于已有 TS-OOK 方案的传输时延。



(a)不同接入节点数下的传输时延



(b)不同占空比下的传输时延

图 7 传输时延性能比较

结束语 本文提出了 3 种无线纳米传感器网络改进型 TS-OOK 接入控制方案，这 3 种方法都通过接入节点与中继节点之间简单控制包交互来实现冲突避免，并且每个发送周期内会发送多个比特。与原有的 TS-OOK 方案的性能比较表明，本文提出的改进方案不仅避免了发送冲突的发生，而且有效提升了网络吞吐量，大大降低了单个接入节点的传输时延。

参 考 文 献

[1] Rao F, Fan Z, Dong L, et al. Molecular nanosensors based on the inter-sheet tunneling effect of a bilayer grapheme[C]// Proceedings of IEEE International Conferences on Nano/Molecular Medicine & Engineering. Hong Kong: IEEE, 2010: 172-175

[2] Sorkin V, Zhang Y. Graphene-based pressure nano-sensors[J]. Journal of Molecular Modeling, 2011, 17(11): 2825-2830

[3] Atakan B, Akan O. Carbon nanotube-based nanoscale ad hoc networks[J]. IEEE Communications Magazine, 2010, 48 (3): 129-135

[4] Akyildiz I, Jornet J. Electromagnetic wireless nanosensor networks[J]. Nano Communication Networks, 2010, 1(1): 3-19

[5] Jornet J, Akyildiz I. Low-weight channel coding for interference mitigation in electromagnetic nanonetworks in the terahertz band[C]// Proceedings of IEEE International Conference on Communications. Kyoto: IEEE, 2011: 1-6

[6] Kocaoglu M, Akan O. Minimum energy channel codes for nanoscale wireless communications[J]. IEEE Transactions on Wireless Communications, 2013, 12(4): 1492-1500

[7] Chi K, Zhu Y, Jiang X, et al. Energy-efficient prefix-free codes for wireless nano-sensor networks using OOK modulation[J]. IEEE Transactions on Wireless Communications, 2014, 13(5): 2670-2682