软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(Suppl.(1)):131−141

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

K-CLOSE:基于不确定图挖掘技术的传感器网络紧密区域发现算法*

韩 蒙[1], 李建中[1,2+], 邹兆年[2]

[1](黑龙江大学 计算机科学技术学院,黑龙江 哈尔滨 150080)

[2](哈尔滨工业大学 计算机科学与技术学院,黑龙江 哈尔滨 150001)

K-CLOSE: Algorithm for Finding the Close Regions in Wireless Sensor Networks Based Uncertain Graph Mining Technology

HAN Meng[1], LI Jian-Zhong[1,2+], ZOU Zhao-Nian[2]

[1](School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China)

[2](School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: E-mail: lijzh@hit.edu.cn

Han M, Li JZ, Zou ZN. *K*-CLOSE: Algorithm for finding the close regions in wireless sensor networks based uncertain graph mining technology. *Journal of Software*, 2011,22(Suppl.(1)):131−141. http://www.jos.org.cn/1000-9825/11014.htm

Abstract: Due to the instability of wireless links and the complexity of geographical environment where wireless sensor networks are deployed, sensor nodes can not guarantee communication with their neighboring nodes with high probabilities. Resolving the problem of finding the sensor nodes that communicate well in practical settings can play an important role in node clustering and the optimization of routing protocols. It is important to note the discovery which nodes in a region are more close to each other in actual movement. A new algorithm called *K*-CLOSE is proposed in this paper to solve the problem which finding the most *k* close regions. First, *K*-CLOSE abstracts wireless sensor networks into uncertain graphs in a distributed manner. Then, the closeness threshold is determined by an approximation algorithm proposed in this paper, which has approximate rate 2. Finally, the *k* close regions where sensor nodes communicate with high probabilities are discovered using tree searching and branch-and-bound methods. Moreover, the experimental results show that the proposed algorithm is efficient in practice.

Key words: wireless sensor networks; uncertain graph; data mining; close region

摘 要: 由于无线传感器网络通信的不稳定性及地理环境的复杂性,传感器节点间经常出现地理位置相近但连通概率却很低的情况.在网络中快速发现通信质量好的节点集以及内部相互联系紧密的子区域,对于传感器网络中的节点分簇、路由优化等具有重要作用.使用不确定图挖掘技术研究如何从一个不确定的无线传感器网络拓扑结构中,快速发现联系紧密且存在概率高的不重叠连通区域问题.提出 *K*-CLOSE 算法,首先,使用分布式方法将无线传感器网络的拓扑结构构建为不确定图;然后,提出一种近似比为 2 的近似算法来计算紧密阈值;最后,通过构建搜索树并

* 基金项目: 国家自然科学基金(61033015, 60831160525, 60933001, 61173023); 中央高校基本科研业务费专项资金(HIT.NSRIF.201180); 黑龙江省研究生创新科研基金(YJSCX2011-239HLJ); 黑龙江大学学生学术科技创新项目(2011386, 2011387)

收稿时间: 2011-05-02; 定稿时间: 2011-07-29

使用剪枝等方法快速发现顶点相互联系紧密且存在概率高的不重叠连通区域.实验结果表明,*K*-CLOSE 算法可以高效地发现无线传感器网络中的紧密连通区域.

关键词: 无线传感器网络;不确定图;数据挖掘;紧密区域

在无线传感器网络中,大量传感器节点被部署在广泛的物理区域内,执行环境监测及对象跟踪等任务.由于传感器节点的电源能量、计算能力和通信能力受限,无线传感器网络的设计需要充分考虑如何减少传感器节点的计算和通信开销问题,从而节省能量.若可以快速找出无线传感器网络中哪些顶点通信更密集,则可以修改通信协议,使紧密的连通区域内各传感器节点共享感知数据,从而在执行数据收集等任务时只需与该节点集中的部分节点通信即可获得该区域内的感知信息,从而降低区域内节点能耗;又如,将紧密节点集建簇,使用动态簇头将簇内共享的感知数据传回 sink 节点,即可提高感知数据传输的效率,进一步降低传感器节点的能耗,具有重要的实际意义.然而,在无线传感器网络中,传感器节点分布式工作,每个节点都有侦听、等待及睡眠等多个状态,同时,一些节点还存在因能量耗尽而失效的情况,这些问题都造成了无线传感器网络的复杂性与不确定性;除无线链路自身的不稳定性外,地理环境中障碍物影响及通信过程中其他信号的干扰和冲突等原因也都造成了网络拓扑结构的不确定性,由于上述原因,很多节点在地理位置上相互处于对方的通信半径内,节点之间实际的通信却经常无法有效完成,两节点以某概率连通,这使得整个无线传感器网络以一个不确定的形式存在,由于传感器网络中节点数量通常很多,网络拓扑结构复杂,存在大量不确定性因素,因此在网络中快速发现联系紧密的区域这一重要问题变得非常困难.

若不考虑不确定性,在无线传感器网络中发现最紧密连通区域**的问题抽象至图的角度分析,即在图 $G(V,E)$中发现子图 Sub,使 Sub 的稠密函数 $des(Sub)=|E(Sub)|/|V(Sub)|$(边的个数除以点的个数)最高,如图 1 所示,网络中的区域 A(节点 1,2,3,4,5)内共有 7 条边,其稠密函数值为 7/5,是整个拓扑结构中最稠密的区域.然而网络环境远比一个确定的拓扑图复杂,相互有边的节点有时由于很少通信或障碍物原因难以通信,使得节点间的边以很低的概率存在,如图 1 中的边(1,2)及边(2,3),虽然节点间可以连通,但边的存在概率仅为 0.1,其实际联系并不紧密,此时稠密函数无法表达子区域内各节点间的通信联系状况,只有顶点间的边以高概率存在(如图 1 中的顶点 5 和顶点 6 所示),区域内各边也以高概率连通(图 1 中区域 A'),该区域才是真正内部节点联系紧密的区域,也才更具有实际意义.在网络上附加不确定性后,发现紧密连通区域问题就转化为在一个不确定图中发现紧密子图的问题,本文使用期望紧密度对子区域节点间的紧密程度进行度量,这一概念将在第 3 节详述.在无线传感器网络中,不确定图中的顶点表示节点,边则表示节点之间是否可以进行通信,边上的权值即传感器节点间成功通信的概率,此时可以利用不确定图的挖掘技术完成传感器网络中发现紧密子区域的实际需求.至今为止,仍未见使用不确定图建模无线传感器网络并利用相关技术发现网络中紧密区域的有效解决方案.

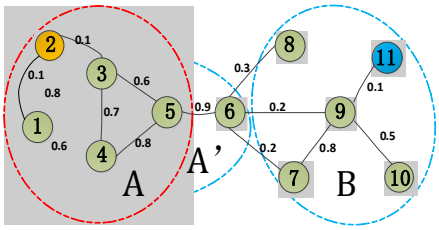


Fig.1 Discovery the close subgraph from uncertain graph

图 1 不确定图中发现紧密子图

然而,一方面由于无线传感器网络节点分布的复杂性和随机性,每一次网络的部署和每一段时间内节点相互通信的状态都可能有其自身的特点,根据不同网络不同时期的特点设计发现紧密区域的有效算法非常困难;

** 本文使用紧密区域(close region)及紧密子图(close subgraph)描述带有不确定性的网络与图中结构及关系密切的连通点集,不同于由单位顶点内边的个数度量的稠密子图(dense graph),紧密子图还需保证各边在可能世界中较高的存在概率.

另一方面,因传感器网络要求很高的实时性,只有快速获得结果才能让挖掘的信息符合时效性要求,这使得该问题对算法效率的要求非常高.

针对以上挑战,本文提出了适用于无线传感器网络的紧密连通区域的发现算法 *K*-CLOSE(finding the most ***K CLOSE*** regions in wireless sensor networks)以解决上述问题.

K-CLOSE 方法具有以下优点:首先,该方法依据传感器网络自身特点采用分布式的预处理方法将网络拓扑快速地构建为不确定图,图附加不确定性后可以有效地描述无线传感器网络中通信链路的结构特性与不确定性,分析并挖掘不确定图中的紧密子图即为传感器网络中的紧密区域;其次,*K*-CLOSE 充分利用了不确定图的特性,组织树型搜索空间,使得输出的结果具有不发生重叠的唯一性,搜索树中一系列有效的剪枝策略提高了算法的效率;最后,*K*-CLOSE 发现网络中紧密联系且存在概率最高的 *K* 个连通区域.

综上所述,本文的主要贡献如下:

(1) 使用不确定图准确刻画了无线传感器网络的拓扑特性和实际通信过程中的不确定性,提出在无线传感器网络中发现最紧密连通区域的问题并在理论上证明了其具有 NP-Hard 的复杂性;

(2) 提出分布式构建、集中式处理、自适应于网络环境且可以对解空间进行有效剪枝的快速紧密区域发现算法 *K*-CLOSE,证明了算法的近似比,并通过输入参数 *K* 设计了可调节的结果输出机制;

(3) 通过实验验证了文中理论分析的结果,并考察了不同网络环境及边的不确定性对算法性能的影响.

本文第 1 节综述相关研究工作.第 2 节定义在网络模型及问题中的相关概念.第 3 节提出自适应于网络环境的 *K*-CLOSE 算法.第 4 节给出实验结果及分析.第 5 节总结全文.

1 相关工作

自 Girman 和 Newman 提出在复杂网络中发现社区结构的问题[1],如何在网络中发现结构密集的区域或有特征的群体一直在通信网络、生物网络及社会网络等诸多领域受到人们的关注,相关的研究包括对网络进行划分、聚簇等.文献[2]提出一种无线传感器网络中有效进行边界划定的方法;文献[3]提出划分网络的聚簇方案,该方案可以保证簇中任意两节点间距离的跳数小于指定参数;最近,文献[4]则提出了一种在网络中发现社区结构的通用概率模型;针对无线传感器网络中数据因通信问题无法有效获得的问题,文献[5]提出了一种基于树结构的分布式数据收集算法,文献[6]提出数据收集的分布式监测方案.针对网络环境的复杂多变,文献[7]提出了一种在动态网络中自适应的社区发现算法.在算法学的最新研究中,文献[8,9]提出了在确定图上发现稠密子图的有效近似算法,然而以上所有的算法都没有考虑图或网络的不确定性.

对于不确定图的相关研究刚刚开始,但不确定图良好的建模能力,具有广泛的应用前景,日渐受到人们的关注,文献[10,11]提出挖掘不确定图中频繁模式及极大频繁模式的有效算法,文献[12]在社会网络上利用随机游走解决了不确定图上的 *k*NN 问题.最近,文献[13]还提出了在不确定图中挖掘极大团的有效方法,本文也是在不确定图中寻找紧密的集团,但在无线传感器网络中,算法必须考虑对不同网络环境的适应性及与网络中其他算法、协议的合作问题,同时还要符合传感器网络实时性要求高等特点.本文的早期工作[14]提出了不确定图的紧密子图概念,并对在大网络上挖掘紧密子图提出了相应的解决方法,然而该工作的解决方法由于对计算性能要求较高,并不适用于无线传感器网络,本文综合考虑传感器网络分布式运算,计算能力弱且单一处理的子图规模并不大等特性,从网络构建、通信开销及算法设计等多方面有效地解决传感器网络中的紧密子区域发现问题.

2 问题定义

无线传感器网络将节点随机部署于监测地区,节点自组织地进行通信,节点的能量和计算能力都非常有限.由于算法的目标是获得网络中节点在通信上的逻辑关系,不需要知道地理拓扑,只需对底层节点的数据收集算法作很小的调整即可快速将传感器网络建模抽象为对应的不确定图,具体方法将在下文算法 1 中详述.

本文所采用的不确定图数据模型是文献[10]提出的不确定图数据模型的特例.简单地讲,本文所采用的不确定数据模型只考虑边的不确定性,而认为顶点是确定的.为了论述得完整,给出如下不确定图的定义:

定义 1. 不确定图是一个三元组 $G=(V,E,p)$,其中,V 是不确定图的顶点集,$E\subseteq V\times V$ 是边集,$p:E\to(0,1)$为给边的赋予存在概率值的函数.

不确定网络中节点及边与不确定图一一对应,边的存在概率表示两个端点间边实际存在的可能性.1 表示边一定存在,确定图即为一个所有边的存在概率皆为 1 的特殊不确定图.一个不确定图 G 蕴含确定图 g,当且仅当图 $g=(V(g),E(g))$使 $E(g)\subseteq E(G)\cap(V(g)\times V(g))$,表示为 $G\Rightarrow g$,对于每个确定图 g 的存在概率为

$$P(G\Rightarrow g)=\prod_{e=(u,v)\in E'}p(e)\prod_{e=(u,v)\in E\cap(V'\times V')\setminus E'}(1-p(e)) \tag{1}$$

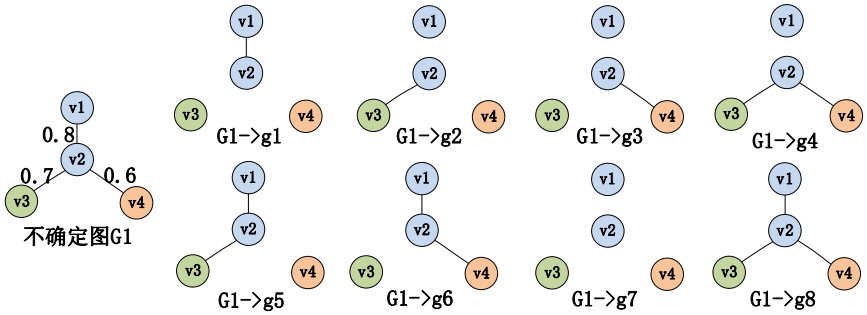


Fig.2 All probability worlds implicated by uncertain graph $G1$

图 2 不确定图 $G1$ 的可能世界实例

例:如图 2 所示,在不确定图 $G1$ 中,图中每条边上的权值表示相连两节点通信成功的概率;$G1$ 的边 $e_0(v1,v2)$ 的存在概率 $p(e_0)$为 0.8,则节点 $v1$ 与节点 $v2$ 通信成功的概率是 0.8.对于一个不确定图 G,若其边 $e_1=(m,n)$的存在概率 $p(e_1)=1$,则 e_1 的两端通信 100%成功(这在实际的链路中很难出现),$p(e_1)=0$(如图 1 中边$(v9,v11)$),则表示两节点无法通信.对一个有$|E|$条边的不确定图,其蕴含 $2^{|E|}$个确定子图.如图 2 所示,一个只有 4 个顶点,3 条边的小图,其蕴含图实例达 $2^3=8$ 之多,具有边的指数级个可能实例,而传感器网络通常由大量节点构成,所以,对于传感器网络,通过枚举每个可能世界,不能有效解决紧密区域发现问题.

本文将网络抽象为无向图,忽略边的方向,但可以很容易地将本文方法扩展至有向的不确定图.扩展时仅需为有向图的每条边赋以 0 或 1 的标识,分别表示顶点的出边和入边,在算法执行时对出边和入边分别进行计算,即可将算法扩展为适于考虑数据传送方向性的有向网络.

定义 2. 对于不确定图 G,不确定图 Gd 为 G 的直接子图当且仅当:

$$|G|=|Gd|+1 \tag{2}$$

$$Gd\subset G \tag{3}$$

因不确定性,在确定图中判定子图 Sub 稠密程度的函数 $des(Sub)$不再适用,只有子图内边个数的期望与顶点个数的比才能真正体现网络顶点间的紧密程度.

定义 3. 给定不确定图 G,g 为 G 的子图,g 期望紧密度(***Ex****pect* ***D****ensity*)为

$$ExD=\sum_{e=(u,v)\in g}p(e)/n,\ n=|V(g)| \tag{4}$$

其中,$p(e)$为边 $e=(u,v)$的存在概率,n 为子图 g 顶点的个数,子图 g 中所有边存在概率之和除以子图内顶点的个数即为每个顶点所连边个数的期望,在一个子区域内,该期望表达的是区域内部各顶点相互联系的紧密情况,即区域内各顶点间的紧密程度.

定理 1. 给定带有不确定性的网络 W,整数 $s>2$,在 W 中发现不小于 s 的最紧密连通区域是 NP-Hard 问题.

证明:不确定性网络 W 与不确定图 G 一一对应,在确定图 G'中发现不小于 k'的稠密子图(Dalk'S, the densest at-least-k'-subgraph problem)问题已证明具有 NP-Hard 的复杂性[15],对于在确定图上的 Dalk'S 问题的任意实例$\langle G',k'\rangle$,在多项式时间内可归约至不确定图的紧密子图发现问题$\langle G,s\rangle$.

Dalk'S 问题中计算稠密子图使用$|E(g)|/|V(g)|$衡量,获得该比值最大的子图,将$|E(g)|$与不确定图中边存在概率的期望与 $\sum_{e=(u,v)\in G}p(e)$ 对应,令$|G|=|G'|,k'=s$,不确定图 G 每条边 e 的存在概率 $p(e)=1$,不确定图 G 在多项式时

间内可转化为确定图 G',可知对于任意 Dalk'S 问题,实例$\langle G',k'\rangle$中稠密子图 sub'是最优解当且仅当其是转化前不确定图 G 的最紧密子图 sub,即不确定网络 W 中最紧密连通区域得证. □

由定理 1 可知,在不确定网络中获取大小至少为 s 的最紧密子区域是 NP-Hard 问题,即在 P≠NP 条件下,不存在多项式时间的有效算法,为了具有更好的实用性,本文研究在不确定网络中近似获得存在概率前 K 高的不重叠紧密子区域问题.具体定义为:

输入:不确定性网络 W,不小于 1 的正整数δ,s 和 K.

输出:W 中大于 s,期望紧密度不低于 W 最紧密区域紧密度 $1/\delta$,不重叠结果中存在概率最高的前 K 个.

3 *K*-CLOSE 紧密子图发现算法

本节首先介绍紧密区域存在概率的计算方法;然后介绍无线传感器网络中不确定图的分布式构建方法;提出适用于传感器网络的快速近似算法发现单个最紧密子区域问题,并分析图中顶点的连通情况,从而设定适应于网络环境的紧密阈值;最后介绍完整的 *K*-CLOSE 算法及优化机制.

3.1 紧密子图存在概率的计算

在无线传感器网络中发现内部联系最紧密的区域不仅要求该区域内各节点相互之间可能进行通信,还必须保证每对顶点相互间通信成功的概率较高,即必须考虑其不确定性.对于不确定图中的紧密子图发现问题,文献[14]已进行了相关分析,为保证文章完整性,仅列出简要证明过程.

定理 2. 给定不确定图 G,若子图 g'为 G 的紧密子图,则其存在概率为 $\prod_{e\in E(g')} p(e)$.

证明:设 P 是不确定图的概率分布,如式(5)所示,G 的 2^m 个可能世界中的第 i 个为 g_i,则可能世界 g_i 的存在概率为式(6).

$$P=\sum_{i=1}^{2} P(g_i) \tag{5}$$

$$P(g_i)=\prod_{e=(u,v)\in E(g)} P(e) \prod_{e=(u,v)\in E(G)\setminus E(g)} (1-P(e)) \tag{6}$$

$$P(g')=\sum\nolimits_{g'\in g_i,i=1}^{i=2^m} p(g_i)=\prod_{e\in E(g')} p(e)\cdot 1 \tag{7}$$

根据可能世界模型中不确定图蕴含子图的基本定义,一个最有价值的关键区域子图 g'在不确定图 G 中的存在概率为所有可能世界中存在概率之和,根据式(5)、式(6)可得式(7),得证[14]. □

推论. 在紧密子区域内边数确定的条件下,*ExD* 值越大,则该子区域的存在概率越大.

证明:由 *ExD* 的定义易得,若将边存在概率视为变量,*ExD* 随边数增加为单调递增.由定理 2 可知,子图 g'的存在概率为 $\prod_{e\in E(g')} p(e)$, 当$|E(g')|$确定时,图中任意一条边 e 的存在概率 $p(e)$越大,显然其存在概率越大. □

3.2 在无线传感器网络中分布式构建不确定图

无线传感器网络最重要的特性就是能量有限,计算能力差,为了降低构建不确定图对能量和计算资源的消耗,*K*-CLOSE 算法首先使用分布式的方法构建不确定图,基本过程为在运行的传感器网络上修改数据收集算法,要求各节点回传感知数据时附加自身节点编号及其与邻居节点通信的记录,在收集一段时间数据后,就可获得节点与节点相互通信成功与失败的次数,据此计算节点间边的概率,不确定图随即构建完成.

算法 1. *DISTRIBUTED_BUILD*.

输入:传感器网络的拓扑结构;

输出:由网络拓扑构建的不确定图 G.

步骤 1. 计时 t_0,网络中所有发送数据的节点开始记录发送数据情况,记录向所有邻居节点发送数据的次数 n 及成功收到校验回复的次数 sn,计时达到 t_0+t 时,下一步;

步骤 2. 将步骤 1 时间 t 内与所有邻居节点发送记录,包括发送和接收数据的节点 id,对应每个 id 的 n 和 sn

回传 *sink*,由 *sink* 上传用户;

步骤 3. 若在上传的数据中存在节点发送节点 *S* 与接收节点 *R* 的记录,在用户端计算两节点间通信的成功率 *sn*/*n*,将权值赋至 *S* 和 *R* 的边,若同时还存在 *R* 发送给 *S* 的记录,则与原赋值相加取平均;

步骤 4. 时间 *t* 内收集所有数据记录后,将网络中各部分已赋值的边及节点合并,不确定图构建完成,算法结束.

算法 1 中网络资源的消耗,对于网络中节点 *v*,其节点编号大小为|*id*|,设其邻居个数为 *N*,在 *t* 时间内,若其平均与每个邻居发生 *m* 次通信,则总传送的信息为 $N\times m\times(n+sn)+2\times|id|$.一般情况下,在一个有 500 个节点的网络中,式中平均邻居个数 $N<10$,设 *t* 为 1 分钟,平均每 6 秒发生一次通信,则 *m* 约为 10,*n*,*sn* 由 1 字节存储,|*id*|占 2 字节,则网络中总传送数据约为 $500\times(10\times10\times(1+1)+2)=101000$,约为 100*K*,这个数据量对于实际传感器网络的通信压力是非常小的.事实上在真正工作的网络中,节点的邻居个数及通信频率等达不到上例中所述,所以利用分布式的算法 1 从传感器网络中构建不确定图完全可以在实际中达到用户要求.

3.3 紧密子图衡量阈值*T*的确定

3.3.1 获得图中节点分布情况的 2-近似算法

本节设计一个 2-近似快速算法计算图中节点的分布情况,并由此确定发现紧密子图的期望最紧密阈值.考虑无大小限制的紧密子图发现问题,在不确定图中基于贪心思想有以下简单算法.

算法 2. *PRE_STUDY*.

输入:不确定图 *G*;

输出:图中最紧密子图 *C*.

步骤 1. 初始化大小为|*n*|的图队列 *R* 使 *R*[*n*]=*G*;

步骤 2. 在 *R* 中找到度的期望(所连边权和)最小的顶点 *v*,删除 *v* 及与其所连的各边,剩余的图存储至 *R*[*n*−1];

步骤 3. 迭代步骤 2 至 *R*[2]即子图只有两个顶点为止,则 *R* 队列存储具有 *n*,*n*−1,…,2 个顶点的近似最紧密子图;

步骤 4. 计算 *R* 中 *ExD* 最大值对应的图 *C* 并返回,算法结束.

由于算法 2 中步骤 4 输出 *ExD* 最大的结果并不需要单独进行排序,该结果仅需在构造 *R* 队列时记录并使用 *O*(1)时间输出,易得算法 2 的时间复杂度为遍历了图中所有顶点和边的时间,即 $O(|V|+|E|)$.接下来简要证明算法的近似比,类似于文献[9],但不确定图的属性信息因不确定性由期望刻画,设优化解 $ExD^*=d^*$,则最终图中每个顶点的期望度都大于 d^*,因为根据贪心思想,若在每一阶段有顶点度的期望小于 d^*,则该顶点一定已被删除,所以总边数期望大于等于 $d^*\times s/2$(*s* 为输出图的大小),依 *ExD* 的定义有 $ExD^*\geqslant d^*/2$,即近似算法输出的结果至少为最优解的 1/2,近似比为 2.

定理 3. 在不确定图中发现期望紧密度不小于最紧密子图 $1/\delta$的子图当且仅当所发现子图的期望紧密度 $T\geqslant ExD^*/2\delta$.

证明:由算法 2 分析,该算法具有 2 近似比,设算法 2 返回结果图的期望紧密度为 ExD^*,而在全图范围内最紧密子图的期望紧密度为 *W_ExD*,则其在最坏情况下,算法 1 的 2 近似比保证 $ExD^*\geqslant 2\times W\_ExD$,$ExD^*/2\geqslant W\_ExD$,易得$(ExD^*/2)/\delta\geqslant W_ExD/\delta$,则设子图期望紧密度阈值 $T^*=ExD^*/2\delta$,只需子图的紧密度 $T\geqslant T^*$即可,得证. □

3.3.2 计算紧密子图的基本方法

由定理 3 可知,使用算法 2 即可获得期望紧密度阈值 T^*,设 $T^*=ExD^*/2\delta$,很容易得到在不确定图中发现最紧密子图的基本算法 BASIC.

算法 3. *BASIC*.

输入:不确定图 *G*,参数 *s*,*K*;

输出:*G* 中大于 *s* 的子图中,不小于最紧密子图期望紧密度 $1/\delta$,存在概率最大的 *K* 个.

步骤 1. 运行算法 1 获得期望最紧密阈值 T^*;

步骤 2. 枚举 *G* 中所有不小于 *s* 的子图,计算其期望紧密度,当期望紧密度大于 T^*/δ时,计算每个子图在所有

可能世界的存在概率;

步骤 3. 排序所有结果后输出存在概率最高的前 K 个子图,算法结束.

在算法 3 中,容易发现,第 2 步需枚举所有大于 s 的顶点子集,虽然此步并不需要在图中特别发现连通子图,而只是大小超过 s 的顶点集(因为不连通的子图存在概率为 0),但算法所枚举子集的个数却通常是图中顶点数的指数,步骤 2 复杂性过高,当节点数量较大时,很难有效完成,并不适用于实时性要求很高的无线传感器网络.

3.4 基本的有效紧密子图发现算法*K*-CLOSE

文献[13]设计了一种分枝限界法用以在不确定图中挖掘极大团,本文基于此工作在自适应获得了期望最紧密阈值的条件下利用有效的剪枝策略设计 *K*-CLOSE 算法.

由于算法 1 已构建出了无线传感器网络上的不确定图 G,我们将 G 的顶点集组织成一棵搜索树,其中根节点为空集,每个非根节点都为其孩子节点的直接子图.此时 K 最紧密子图发现问题可以转化为在搜索树的第 s+1 以下层中找出存在概率最大的 K 个节点问题.由问题定义可知,所发现的 K 个节点所代表的子图间没有覆盖,因为在无线传感器网络中的多数应用(如建簇,制定睡眠机制等)需要将传感器节点划分为互不重叠或者重叠很小的组,本文算法采用按节点标号自左向右扩展的方式建树,已经扩展的节点将不在以后的算法过程中重复扩展,最终获得的 K 个结果将是相互没有覆盖的紧密子图.

算法 4. *K*-CLOSE.

输入:无线传感器网络 W,参数 s,K.

输出:不小于最紧密子图期望紧密度 $1/\delta$的不重叠子图中存在概率最大的 K 个.

步骤 1. 运行算法 1,快速构建不确定图 G;

步骤 2. 运行算法 2,获得期望紧密阈值 T^*;

步骤 3. 由空集开始每次增加一条边,构建所有可能世界实例的搜索树,自上而下搜索并计算搜索树中各节点所构建的可能世界实例,自第 s 层开始计算其期望紧密度,当期望紧密度大于 T^*/δ时,计算每个子图在所有可能世界的存在概率;

步骤 4. 排序所有结果后输出存在概率最高的前 K 个子图,算法结束.

算法 4 在每次迭代过程中维护一个大小为 K+1 的极大堆,堆中存储 K 个元素,这 K 个元素即为最终输出结果的候选集,空集作为堆的开始,使用堆可以使结构中有序的调整时间接近 $O(\lg n)$,同时堆内还能够记录上文所构造搜索树中每个中间结果的期望紧密度以及节点个数信息.当算法结束时,使堆中的每个元素都是最后输出的结果之一,输出堆内的 K 个结果即可.

3.5 *K*-CLOSE的优化剪枝策略

算法 4 构建了一棵可能世界实例的搜索树,在算法迭代至搜索树的第 s 层后,使用最大堆的数据结构记录达到紧密阈值的子图中存在概率最高的前 K 个,利用树结构采用每次增长一条边的方式不再枚举计算所有实例及实例的存在概率,有效降低复杂性.算法 4 较 BASIC 算法有很大提高,但仍在很多步骤上存在冗余和计算的浪费,以下将从几个方面对算法作进一步优化.

(1) 搜索过程中冗余遍历过程的合并,在基本算法 BASIC 中,步骤 2 中树的构建过程遍历了树中所有节点,而步骤 3 中又重新遍历计算了每个节点的期望紧密度和存在概率.事实上,容易发现步骤 2 和步骤 3 可以同步进行,在构建的过程中每增加树中一个节点就同时计算其期望紧密度和存在概率,几乎可以降低 50%的计算花费,同时对期望紧密度未达到紧密阈值的节点计算其存在概率也没有意义,所以,当发现期望紧密度小于阈值的节点时,可直接标识,无需计算其存在概率.

(2) 搜索树上的层次剪枝方法,算法要求在输出大小至少为 s 的子图中,达到紧密阈值且存在概率高的前 K 个.事实上,在搜索树的前 s–1 层中不可能有达到大小 s 的子图被发现,所以在该阶段无需计算其紧密度和存在概率,优化算法中可直接将搜索树构建至第 s–1 层再进行步骤 3 中对期望紧密度和存在概率的计算即可.

(3) 基于存在概率递减的剪枝方法,由定理 2 可知,在树自根向下搜索的过程中,随着边的增加,图实例的存

在概率减小,可以同时计算树中每个节点的期望紧密度和存在概率,搜索到期望紧密度阈值以上的节点,并剪枝其孩子.

使用以上剪枝策略优化后的*K*-CLOSE算法这里不再赘述,在第4节,我们将进一步考察优化后算法的效率.

4 实验结果

我们利用实验考察本文算法的执行效率,不同数据规模与不同参数对算法的影响以及算法所获结果的质量.所有算法都在BGL[***]库支持下用C++实现,G++编译器通过.用于实验的计算机具有Intel Core 2 Duo1.66GHz CPU和2G内存,运行Ubuntu10.10操作系统.

由于未见利用不确定图模型研究在无线传感器网络中进行 K 紧密区域发现问题的相关算法,且其他网络模型及确定性数据上的算法与本文解决的问题缺乏可比性,本文主要在不同网络环境及不同参数条件下验证理论分析的结果及算法执行的效率,并通过不同环境对比本文算法在无线传感器网络不同网络环境和节点分布条件下的适应性及基础预测能力.

4.1 实验数据

在本文算法 1 中已详细描述了在无线传感器网络中不确定图的构建和转化策略,但是由于真实实验中同一套实验节点系统难以实现多种不同网络环境,无法验证本文算法自适应于网络的重要特性,因此本文采用模拟数据.在不确定图的研究工作中,蛋白质交互网络的真实不确定信息具有很好的典型性,被不确定图的研究人员广泛使用,该图可以模拟复杂无线传感器网络环境,本文利用蛋白质交互网络的两组真实不确定数据及 4 组模拟的不确定数据实验,以使各网络具有较大的环境变化,从而验证本文所提算法的效率和结果的质量.

生物数据来自由BioGRID数据库获得的真实的不确定图,每个蛋白质交互网络都是一个不确定图,顶点代表蛋白质模拟传感器节点,边代表蛋白质间的交互模拟节点之间的通信,边上的概率由欧洲分子生物实验室的STRING数据库提供.4组模拟数据由计算机随机生成,并由程序对图上的边进行不确定性处理.

Table 1 Graphs used in performance evaluation

表 1 本文所使用的6组数据的相关参数

数据名称	顶点数	边数	边存在概率平均数
(Graph1)果蝇	4 038	5 564	0.469
(Graph2)蠕虫	827	961	0.392
(Graph3)模拟	150	100	(*aver*=0.7, *d*=0.2)0.748
(Graph4)模拟	200	150	(*aver*=0.3, *d*=0.2)0.306
(Graph5)模拟	500	1 500	(*aver*=0.4, *d*=0.2)0.349
(Graph6)模拟	3 500	3 800	(*aver*=0.6, *d*=0.2)0.539

不确定性处理的参数有*average*及*d*,在不确定性处理过程中,先扫描不确定图的所有边,使用随机数产生器以*aver*为均值,d^2为方差对图的每条边进行赋值,将不同的概率值存入图的各条边中.

将表1中的Graph2和Graph3可视化后的效果如图3所示,该图主要展示顶点间的情况,两点间概率越大,其颜色越深,本文算法的目标就是快速找到联系密集且存在概率高的K个子区域.

图4是6组网络数据中边存在概率的分布情况,易见其服从不同的分布,在实际的传感器网络环境中,节点相互间的通信状况也因各种原因呈现不同的分布,本文使用多组不同规模、不同分布的网络数据考察算法对不同网络的适用性.

图 5 所示为各数据中不考虑不确定性的顶点平均度与带有不确定性的顶点平均期望度的对比示意图,易见其分布在不同的网络环境中,因边存在概率的影响,其分布情况并不相同,所以,由于实际网络通信中通常带有不确定性,仅以顶点度为度量标准计算聚簇、划分等工作可能导致非优解甚至错误解,所以在实际网络进行相关算法设计时对不确定性的考虑非常必要.

*** Boost Graph Library(BGL).http://www.boost.org/doc/libs/1_42_0/libs/graph/

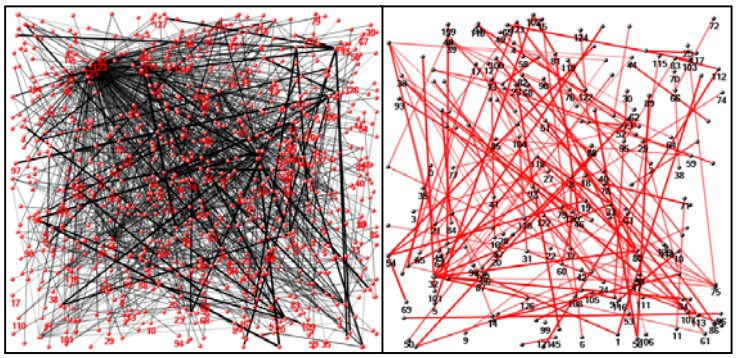


Fig.3 The topological structure of Graph2 and Graph3

图 3 Graph2 和 Graph3 构建的网络示意图

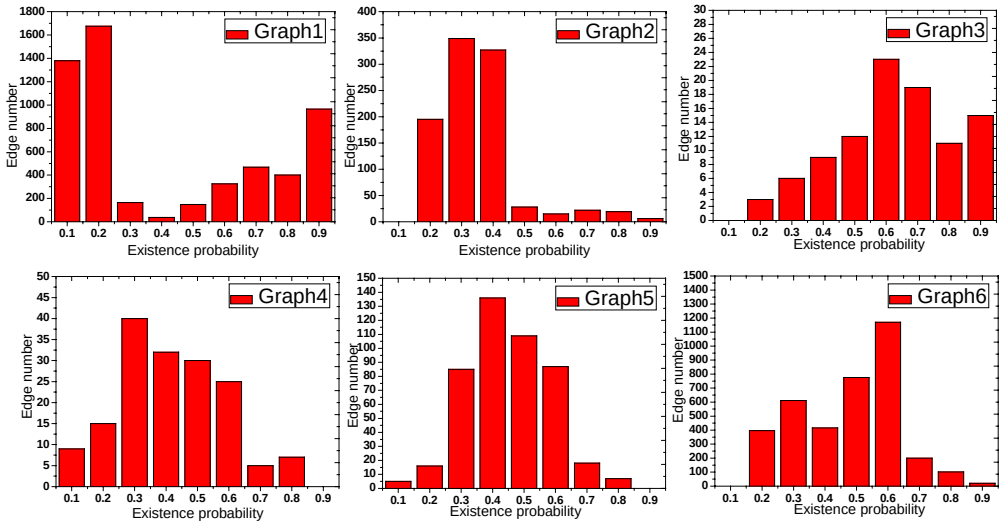


Fig.4 The existence probabilities of edges of every graph

图 4 各组数据中边的概率分布情况示意图

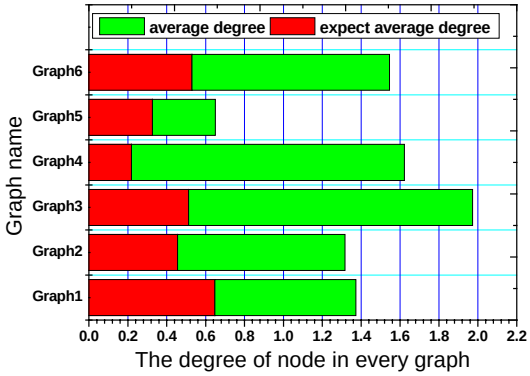


Fig.5 Comparison between expect degree and degree

图 5 各组数据中顶点度与期望度对比示意图

4.2 算法性能及结果分析

运行算法 2 对各图进行快速预处理,获得网络的近似最紧密子图,并计算其紧密度,如图 6 所示,不同网络中最紧密子图的近似度事实上因网络不同会有很大差别,由此可以发现,只有根据不同网络环境设定紧密阈值才能有效获得适合网络分布状况的紧密子区域.

考察算法运行的效率和不同参数对算法结果的影响,如图 7 所示,该图显示为参数δ(Delta)=5 时,不同的子图最低大小限制 s 运行时间情况,可以发现,当最小限制 s 增大时,运行时间显著增加,这是因为搜索树的构建要达到 s 层才开始剪枝,而随着搜索树深度的增加,每层元素的个数也快速增加.

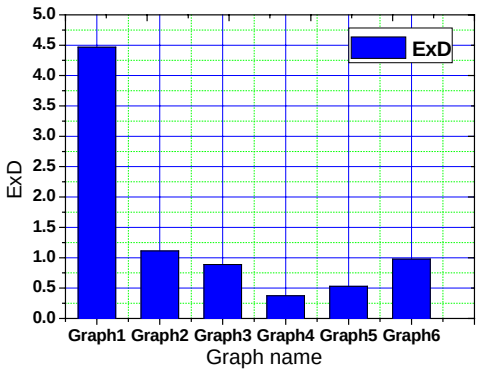


Fig.6 The *ExD* value of different graph

图 6 不同数据的期望紧密度 *ExD* 值

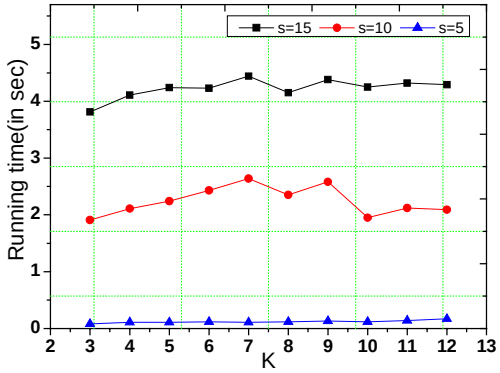


Fig.7 Impact of the variation of s on the execution time

图 7 不同参数 s 运行时间示意图

图 8 为同样当 s=10 时,δ参数对运行时间的影响,可见δ越大,则算法运行时间越长,这是因为δ的增大使我们判定紧密子图的条件变得更为苛刻,算法必须搜索更深的子树才能获得最优的 K 个结果.

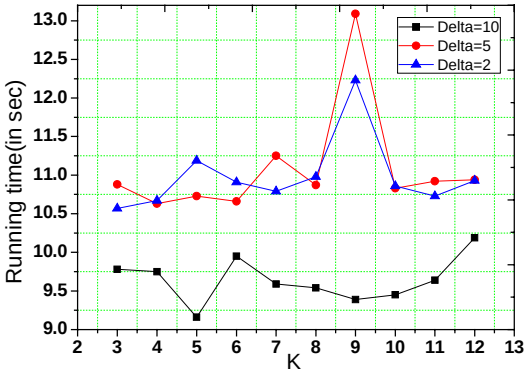


Fig.8 Impact of the variation of δ on the execution time

图 8 不同参数δ运行时间示意图

综上所述,本文实验验证了算法理论分析的结果,在 500 个节点的图上,由自适应近似算法计算的紧密阈值,在δ不小于 5 的条件下,算法的运行时间不超过 10s,完全可以适用于无线传感器网络中用户的使用.

5 结 论

本文研究了在带有不确定性的无线传感器网络上的发现 K 紧密子区域的问题,证明其具有 NP-Hard 的复杂性,提出了将网络构建为不确定图,利用不确定图的挖掘技术提出自适应于网络环境的 K-CLOSE 算法,K-CLOSE 可以快速获得 K 最紧密子图,从而可以有效解决在传感器网络上发现 K 紧密区域的问题.实验结果表明,本文提出的算法可以高效发现网络中的紧密子区域,具有重要的实际意义.

References:

[1] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99, 2002, 7821.

[2] Wang Y, Gao J, Mitchell JSB. Boundary recognition in sensor networks by topological methods. In: Proc. of the MobiCom 2006. New York: ACM Press, 2006. 122−133.

[3] Fernandess Y, Malkhi D. *K*-Clustering in wireless ad hoc networks. In: Proc. of the 2nd ACM Int'l Workshop on Principles of Mobile Computing 2002. New York: ACM Press, 2002. 31−37.

[4] Chang C-S, Hsu C-Y, Cheng J, Lee D-S. A general probabilistic framework for detecting community structure in network. In: Proc. of the IEEE INFOCOM 2011. Washington: IEEE Press, 2011.

[5] Li J, Khuller ADS. On computing compression trees for data collection in wireless sensor networks. In: Proc. of the IEEE INFOCOM 2010. Washington: IEEE Press, 2010. 2115−2123.

[6] Liu CL, Cao GH. Distributed monitoring and aggregation in wireless sensor networks. In: Proc. of the IEEE INFOCOM 2010. Washington: IEEE Press, 2010. 2097−2105.

[7] Nguyen NP, Dinh TN, Xuan Y, Thai MT. Adaptive algorithms for detecting community structure in dynamic social networks. In: Proc. of the IEEE INFOCOM 2011. Washington: IEEE Press, 2011.

[8] Andersen R, Chellapilla K. Finding dense subgraphs with size bounds. In: Proc. of the Workshop on Algorithms and Models for the Web-Graph, WAW 2009. Barcelona: Springer-Verlag, 2009. 25−36.

[9] Khuller S, Saha B. On finding dense subgraphs. In: Proc. of the Int'l Colloquium on Automata, Languages and Programming, ICALP 2009. Rhodes: Springer-Verlag, 2009,(1):597−608.

[10] Zou ZN, Li JZ, Gao H, Zhang S. Mining frequent subgraph patterns from uncertain graph data. IEEE Trans. on Knowledge and Data Engineering, 2010,22(9):1203−1218.

[11] Han M, Zhang W, Li JZ. RAKING: An efficient *K*-maximal frequent pattern mining algorithm on uncertain graph database. Chinese Journal of Computers, 2010,33(8):1387−1395 (in Chinese with English abstract).

[12] Potamias M, Bonchi F, Gionis A, Kollios G. *k*-Nearest neighbors in uncertain graphs. In: Proc. of the Vldb Endowment, PVLDB. 2010,3(1):997−1008.

[13] Zou ZN, Li JZ, Gao H, Zhang S. Finding top-*k* maximal cliques in an uncertain graph. In: Proc. of the Int'l Conf. on Data Engineering 2010. Washington: IEEE Press, 2010. 649−652.

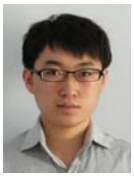
[14] Han M, Li JZ, Zou ZN. Finding the *K* close subgraphs in an uncertain graph. Journal of Frontiers of Computer Science & Technology, 2011,5(9):791−803.

[15] Feige U, Kortsarz G, Peleg D. The dense *k*-subgraph problem. ALGORITHMICA, 2001,29(3):410−421.

附中文参考文献:

[11] 韩蒙,张炜,李建中.RAKING:一种高效的不确定图 *K*-极大频繁模式挖掘算法.计算机学报,2010,33(8):1387−1395.

[14] 韩蒙,李建中,邹兆年.从不确定图中发现 *K* 紧密子图.计算机科学与探索,2011,5(9):791−803.



韩蒙(1987－),男,内蒙古通辽人,硕士生,主要研究领域为图数据挖掘,无线传感器网络.

邹兆年(1979－),男,博士,讲师,主要研究领域为图数据挖掘.

李建中(1950－),男,教授,博士生导师,主要研究领域为数据库系统,传感器网络.