# 一种基于随机游走算法的复杂网络生成

余思东 万荣泽* 黄 欣
(广西农业职业技术学院现代教育技术与网络信息中心 广西 南宁 530007)

**摘 要** 为了更好地研究复杂网络结构,采用随机游走算法实现复杂网络的生成。该算法采用扩散策略实现网络生成,根据活跃标签标注下一次抽取与前一次抽取的联系,除第一个节点之外的所有网络节点的加入均由网络节点属性决定,以最大转移概率为准则来实现下一个节点的选择,有效避免局部陷入与遍历不全的不足。实验证明,相比于传统的复杂网络生成方法,该算法能更好地反映复杂网络的原始网络结构,网络连通性好,度分布和聚类系数与原始网络更相似,网络层次性更强。

**关键词** 复杂网络 网络生成 随机游走算法 度分布 聚类系数

**中图分类号** TP399 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2015.02.030

# A COMPLEX NETWORK GENERATION METHOD BASED ON RANDOM WALK ALGORITHM

Yu Sidong Wan Rongze* Huang Xin
(*Modern Education Technology and Network Information Center*, *Guangxi Vocational and Technical College of Agriculture*, *Nanning 530007*, *Guangxi*, *China*)

**Abstract** In order to better study the complex network structure, we use random walk algorithm to realise the generation of complex network. The algorithm adopts diffusion strategy to achieve network generation, marks the connection between the next-time extraction and the previous one according to active labels. All the joining of the network nodes are determined by network node attributes except the first one, and the selection of next node is implemented by taking the maximum probability as criterion, thus effectively prevents from the insufficiencies of falling into local optimum and incomplete traversal. Experiment proves that compared with traditional complex network generation method, the algorithm proposed in this paper can better reflect the original network structure of complex networks. It has good network connectivity, the degree distribution and clustering coefficient are more similar to the original network, and the network hierarchy is higher as well.

**Keywords** Complex network Network generation Random walk algorithm Degree distribution Clustering coefficient

## 0 引 言

随着计算机网络应用与数据挖掘技术的发展,特别是社交网络平台的盛行,复杂网络的研究变得越来越重要。不论是对于社交平台用户关系网的数据分析[1],还是对于电子商务网站消费模式制定,抑或是电信营运商的资费方式调整,借助与复杂网络方法更能挖掘出关联用户的价值信息。比如腾讯公司根据用户的好友及所在群情况,建立了圈子好友,为用户推送更多具有间接联系的好友,这便是对复杂网络的一种应用。在利用复杂网络知识分析问题之前,首先必须对复杂网络进行可视化处理,这一步骤能够更好地反映整个网络的结构。文献[2]是运用复杂网络可视化对社会网的数据分析,文献[3,4]均是对网络社交平台的数据挖掘[2-4],可视化处理最大的优点是便于查看网络内部结构,从而挖掘网络中隐藏的有价值信息。这是数据处理比较关键的一步。本文正是基于此背景所做研究,目的在于找到一种性能更优的网络生成方法,为大型网络数据分析提供一定的帮助。

当前复杂网络的生成方式主要有三种[5]:随机点生成方式、随机边的生成方式和基于节点的度的生成方式。三种方式各有优劣,本文提出了基于随机游走复杂网络生成方法,网络抽样更加均匀,通过实验比较,能够更好地保持原网络结构,连通性优良。

## 1 基于随机游走的层次网络生成

对于复杂网络来说,研究其层次结构显得非常重要。复杂网络节点数目较多,节点之间的上行总括关系、下行隶属关系及并列关系错综复杂,而节点之间的这种关系对分析网络相邻节点属性以及整个网络的结构研究具有重要意义。这便需要采用相应的策略对复杂网络结构进行层次分析,将整个复杂网络按照一定的方式划分成不同层次的小网络,将三维空间网络结构划分成多张二维空间的网络进行探究分析,以降低复杂网络的分析难度。本文将采用随机游走算法来实现网络的层次生成。

### 1.1 常见层次网络生成方法

(1) 随机选点方法

随机选点方法,顾名思义是随机选择原始网络中的节点,并

收稿日期:2013-08-07。广西教育厅科研项目(201204LX350)。
**余思东**,副教授,主研领域:计算机技术,网络技术。**万荣泽**,副教授。**黄欣**,讲师。

参考原始网络中所选中节点的边关系对网络进行重新构建[6]。根据原始网络节点个数确定抽样比,选择一定量的随机节点,然后组建网络。一般而言,随机选点策略的精确性较差,因为网络中的节点都是随意抽取的,在样本量大的情况下,可能较好地恢复整个网络的原貌,但是在样本容量较小的情况下,误差较大,不适合复杂网络的结构生成。

(2) 随机选边方法

随机选边策略是对网络中的边进行随机选择,然后与选中边相连的节点选中,将选中的边与这些节点进行网络重建。在已经选中的边的相邻边中进行随机选择,该方法和随机选点方法的缺点一样,在样本容量不够大的情况下,误差较大,所以使用具有一定的局限性。

(3) 基于节点度方法

在随机点方法的基础上衍生了一种基于节点度的选点策略[7],节点的度在一定程度上反映了该节点与相邻节点的关联关系的强弱。该策略选择度较大的节点进行网络重建,因为度越大,与之相连的节点越多,表明该节点在整个网络中的重要性越高,选择这样的节点更能够保持原来网络的真实架构,该方法具有一定的应用范围,但是要计算出所有节点的度是比较耗时的,效率低下,特别是大型复杂网络,因此该方法不适合节点多的网络生成。

考虑到三种方法的适用范围及优缺点,提出了基于随机游走的网络生成方法,很好地弥补了这三种方法的不足。

## 1.2　基于随机游走的层次网络生成方法

随机游走是一种基于扩散策略的网络生成方法,基本实现原理是在整个原始网络中放置一个活跃标签[8],活跃标签的游走过程即为随机游走过程。活跃标签经过的节点即为网络重建的节点,随机游走策略在选取行走路由的时候,不是简单地根据当前节点与之相连的节点中随机选取下一个节点,而是根据节点网络属性。

随机游走策略强调下一次抽取与前一次抽取的联系,选中某一节点之后,从与之节点相连的节点中随机选取下一个节点,依次迭代重建网络。

在给定区域的随机游走算法数学模型如下所示:

考虑到网络节点个数为有限个,是具有一定的边界条件的,因此只需考虑在边界条件下活跃标签移动的概率求解即可。定义在给定边界条件 $\Gamma$ 下,随机游走活跃标签从非标记点出发第一次到达标记点的概率可参考下式:

$$\nabla^2 u = \frac{\partial^2 u}{\partial i^2} + \frac{\partial^2 u}{\partial j^2} = 0 \tag{1}$$

其中 $u(i,j)$ 在边界区域内具有二阶连续偏导数且满足拉普拉斯方程,因此该函数是典型的调和函数,该函数的边界条件为:

$$u(i,j)\mid_{\Gamma} = \begin{cases} 1 & (i,j) = s \\ 0 & \text{其他} \end{cases} \tag{2}$$

根据二维空间偏微分的数学关系,概率的求解过程转变为 $u(i,j)$ 的最小值求解过程。在映射图中定义联合拉普拉斯矩阵为:

$$L_{ij} = \begin{cases} d_i & i = j \\ -w_{ij} & v_i \text{与} v_j \text{为相邻节点} \\ 0 & \text{其他} \end{cases} \tag{3}$$

$L_{ij}$ 的值由节点 $v_i$ 与 $v_j$ 共同决定,$d_i$ 为节点 $v_i$ 的度。顶点间的关联矩阵,即图 G 的 $m \times n$ 条边的定义:

$$A_{e_{ij}v_k} = \begin{cases} +1 & i = k \\ -1 & j = k \\ 0 & \text{其他} \end{cases} \tag{4}$$

由式(4)得,关联矩阵由边 $e_{ij}$ 和节点 $v_k$ 共同决定,图中所有的 $e_{ij}$ 可以是任意一个指定的方向。一般称 $A$ 为联合梯度算子,$A^{\mathrm{T}}$ 为联合散度算子。

网络边的权值用对角阵 $C$ 来定义:

$$C_{e_{ij}e_{ks}} = \begin{cases} w(e_{ij}) & i = k \quad j = s \\ 0 & \text{其他} \end{cases} \tag{5}$$

由于正定阵 $L$ 可以分解为 $L = A^{\mathrm{T}}A$。这个构造矩阵 $C$ 可以理解为在向量上一个加权内积大小的度量,从这个意义上来说,通过 $L = A^{\mathrm{T}}CA$,即当 $C = I$ 时,$L = A^{\mathrm{T}}A$。

给定区域 $\Omega$ 上的积分为:

$$D[u] = \frac{1}{2}\int_{\Omega} |\nabla u|^2 \mathrm{d}\Omega \tag{6}$$

在已经固定标记点的值的前提下,求解非标记点到达标记点的概率值。式(6)可以改写为:

$$D[x] = \frac{1}{2}(Ax)^{\mathrm{T}}C(Ax) = \frac{1}{2}x^{\mathrm{T}}Lx = \frac{1}{2}\sum_{e_{ij}\in E} w_{ij}(x_i - x_j)^2 \tag{7}$$

根据线性代数相关知识,由式(7)可得,$L$ 是半正定矩阵,所以在唯一临界点处,$D[x]$ 可以取得最小值。将映射图 $G$ 的所有顶点划分为种子节点集 $V_M$(标记点集)和未标记点集 $V_U$ 两个子集,且使它们满足 $V_M \cup V_U = V$,$V_M \cap V_U = 0$,在不失一般性的前提下,我们假定 $L$ 和 $x$ 中的顶点按照优先排列种子点再排列非种子点的规则来完成,于是,式(7)可以分解为:

$$\begin{aligned} D[x_U] &= \frac{1}{2}\begin{bmatrix} x_M^{\mathrm{T}} & x_U^{\mathrm{T}} \end{bmatrix}\begin{bmatrix} L_M & B \\ B^{\mathrm{T}} & L_U \end{bmatrix}\begin{bmatrix} x_M \\ x_U \end{bmatrix} \\ &= \frac{1}{2}(x_M^{\mathrm{T}}L_M x_M + 2x_U^{\mathrm{T}}B^{\mathrm{T}}x_M + x_U^{\mathrm{T}}L_U x_U) \end{aligned} \tag{8}$$

其中,$x_B$ 和 $x_U$ 分别对应种子点和非种子点的电压值,此时,拉普拉斯矩阵也被分解为:

$$L = \begin{bmatrix} L_M & B \\ B_T & L_U \end{bmatrix} \tag{9}$$

对 $D[x_U]$ 求关于 $x_U$ 的微分可求得临界点,可得:

$$L_U x_U = -B_T x_M \tag{10}$$

这是一个含未知数 $|V_U|$ 的线性方程系统,只有整个图是连通的,或者每个连接区域都含有一个种子点,那么式(10)才是满秩的。

令 $x_i^s$ 表示非种子点 $v_i$ 第一次到达类别 $s$ 的种子点的概率值,并定义这样一个函数来表示所有标记点的集合:$Q(v_j) = s$,$\forall v_j \in V_M$,其中 $s \in Z, 0 < s \leqslant K$,再为所有在 $v_j \in V_M$ 的点定义一个 $|V_M|_{\times 1}$ 大小的矩阵:

$$m_j^s = \begin{cases} 1 & Q(v_j) = s \\ 0 & Q(v_j) \neq s \end{cases} \tag{11}$$

因此,通过求解式(12),就能得到单个标记点 $s$ 的联合狄利克雷问题的概率值大小。而对于所有的种子点的概率值大小,需通过求解式(13)得到[9]。

$$L_U x^s = -B^{\mathrm{T}}m^s \tag{12}$$

$$L_U X = -B^{\mathrm{T}}M \tag{13}$$

由 $K-1$ 个线性方程求解出非标记点到 $K$ 个标记点的概率值大小,由 $x^s$ 可以得出 $X$ 有 $K$ 列,且由 $m^s$ 可以得出 $M$ 的列数,由电路原理可知,每个节点的概率和为 1:

$$\sum_s x_i^s = 1 \quad \forall v_i \in V \tag{14}$$

在获得每个未标记点 $v_i$ 到标记点的 $K$ 个概率后，逐个比较它们的大小，以最大转移概率 $\max_s(x_i^s)$ 为准则来实现下一个节点的选择。

## 2 实例仿真

为了验证本文算法在复杂网络层次生成运算中的性能，采用某一社交网站服务器用户数据网络作为数据来源来进行实例仿真。社交网络平台需要发展，必须借助考虑用户之间的联系，以求为用户提供更优质便捷的推送服务，这就需要对用户之间的关联性进行深入研究，复杂网络方法则正好解决了该问题。

本文以某社交网站的用户数据进行提取，然后运用随机节点、随机边、基于节点的度方法和随机游走算法生成复杂网络，并对算法运行时间及层次网络属性进行分析，验证本文算法的优越性。

将每个注册的用户作为网络中的节点，用户发送消息对话作为网络中的边。首先抽取某一时间段的用户交互数据构建网络，网络节点共有 10 000 个，边 633 241 条。

由于要对节点度选点算法进行性能对比，有必要对网络中节点的度分布和聚类系数进行求解。其中聚类系数是对网络节点与之相邻节点紧密程度的衡量。

如图 1 所示为原始网络节点度分布结构图，其中横轴为节点的度，纵轴为网络中节点数量，从图中可见，网络中节点的度普遍集中在 1 ~ 6，度为 1 的节点最多，约有 3 200 个。

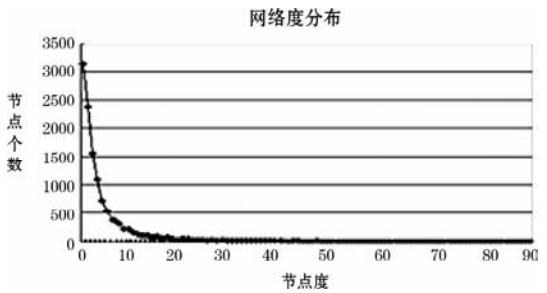


图 1 网络节点度分布图

如图 2 所示为原始网络节点聚类系数分布图，其中横轴为节点的度，纵轴为节点的平均聚类系数，从图中可见，度为 20 ~ 40 的节点的聚类系数最高，在 0.4 左右，表明度为 20 ~ 40 的节点与之相邻节点的紧密程度更高。

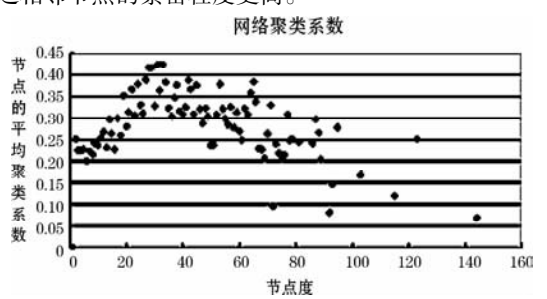


图 2 网络聚类系数分布图

在实验中，为了便于展示整个网络，在原始网络中按一定数量抽取节点，抽取节点占原始节点比例分别为：80%，60%，40%，20%，分别查看可得原始网络的拓扑结构图。考虑到篇幅原因，本文仅列举了抽样比例为 40% 时随机游走算法生成的网络结构图，如图 3 所示。

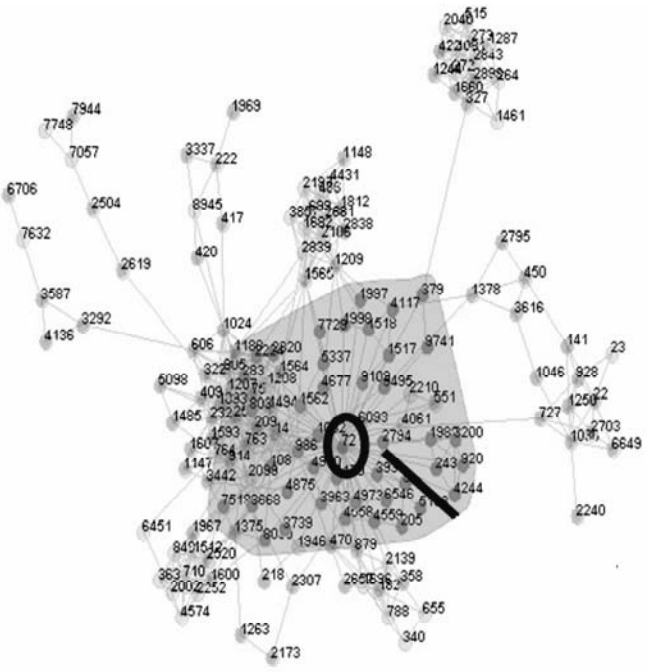

图 3 随机游走方法的层次生成结果图

在实际运算过程中，抽样比并不是越高越好，抽样比过高，系统运算量大，而且节点数目过多，对整个网络结构分析的贡献表现并不突出，相反，抽样比适中，系统运算效率高，而且能够很好地反映整个网络的层次架构，本文选择的抽样比例为 40%，满足本文分析需求。

考虑到层次生成的目的是用尽可能少的节点和边展示出网络特性，下面将对点边比例、网络连通分量个数、最大连通分量比例做仿真，因为这些参数在一定程度上反映了层次生成策略算法是否展示出原始网络特性。

从表 1 和表 2 数据可得，从网络连通程度上来看，随机节点，随机边，基于度的点策略均会因为节点抽样比下降导致网络连通性下降，随机游走则受节点抽样比的降低对网络连通性影响较小。

**表 1 网络连通分量个数**

| | Random Node | Random Edge | Random Degree | Random Walk |
|---|---|---|---|---|
| 100% | 1 | 1 | 1 | 1 |
| 80% | 731 | 1001 | 165 | 3 |
| 60% | 1633 | 1123 | 198 | 7 |
| 40% | 1812 | 996 | 173 | 11 |
| 20% | 1303 | 941 | 101 | 5 |

**表 2 网络最大连通分量比例**

| | Random Node | Random Edge | Random Degree | Random Walk |
|---|---|---|---|---|
| 100% | 1 | 1 | 1 | 1 |
| 80% | 0.881213 | 0.631681 | 0.983602 | 0.999921 |
| 60% | 0.703166 | 0.532468 | 0.960102 | 0.997012 |
| 40% | 0.536217 | 0.427713 | 0.943005 | 0.989019 |
| 20% | 0.170809 | 0.156839 | 0.929886 | 0.985818 |

下面将对不同策略的网络层次生成后的节点度分布和聚类系数与原图进行对比，比较网络生成的相似性与稳定性，考虑到篇幅原因，本文仅给出了随机游走策略生成的度分布和聚类系数图，如图 4 和图 5 所示。

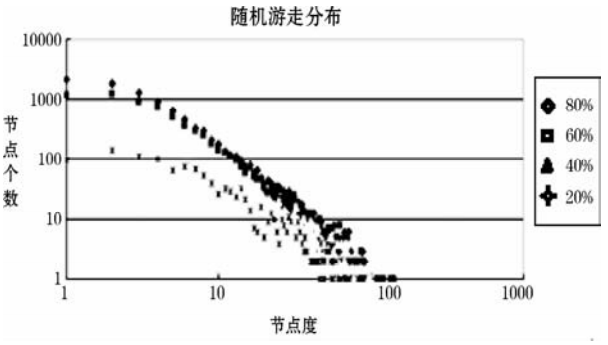


图 4　随机游走度分布

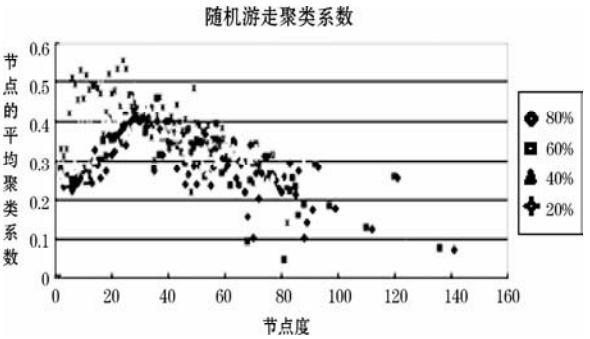


图 5　随机游走聚类系数

4 种层次生成策略在生成层次网络时,基本都能保持整个网络度分布的无标度特性,同时也能保证聚类系数的各层次之间也具有一定的相似关系,但是通过与图 1 和图 2 的原始网络度分布和聚类系数比较,策略算法性能有一定的差异。其中随机游走在度分布和聚类系数上更接近于原始网络。

从以上结果分析可得,随机游走策略运用于复杂网络层次生成,网络连通性好,度分布和聚类系数与原始网络更相似,是一种较好的复杂网络层次生成方法。

## 3　结　语

本文首先对常见层次生成方法进行了简要分析,接着提出了随机游走策略,并对其算法流程及数学模型进行详细分析,最后从某社交网站中取一部分用户交互信息进行实例仿真,对常见三种网络策略和随机游走策略的层次生成性能进行对比。实验对比结果表明,不论是在网络连通性,还是在度分布和聚类系数方面,随机游走策略更适合于复杂网络的层次生成,能更好地还原原始网络结构,失真度最小,具有一定的应用价值。

## 参　考　文　献

[1] 仇钧,刘功申. 基于关系的微博重要度算法研究[J]. 信息安全与通信保密,2013(1):51-53.

[2] 窦炳琳,李澍淞,张世永. 基于结构的社会网络分析[J]. 计算机学报,2012(4):741-753.

[3] 徐翔斌,涂欢,王佳强. 基于复杂网络的社交网站用户评价模型研究[J]. 华东交通大学学报,2012(5):38-43.

[4] 胡海波,徐玲,王科,等. 大型在线社会网络结构分析[J]. 上海交通大学学报,2009(4):587-591,595.

[5] 刘大有,杨建宁,杨博,等. 基于环路紧密度的复杂网络社区挖掘方法[J]. 吉林大学学报:工学版,2013(1):98-105.

[6] 周炎,刘亚冰,汪小帆. 一种基于层次化社团结构的复杂网络可视化平台[J]. 上海交通大学学报,2010(3): 332-335,339.

[7] 梁辰,徐健. 社会网络可视化的技术方法与工具研究[J]. 现代图书情报技术,2012(5): 7-15.

[8] 斜斐玲,胡延庆,黎勇,等. 空间网络上的随机游走[J]. 物理学报,2012(17): 571-577.

[9] 肖杰斌,张绍武. 基于随机游走和增量相关节点的动态网络社团挖掘算法[J]. 电子与信息学报,2013(4):977-981.

(上接第 115 页)

## 参　考　文　献

[1] Mitola J, Cognitive radio: An integrated agent architecture for software defined radio [D]. Doctor of Technology, Royal Inst. Technol. (KTH), Stockholm, Sweden,2000.

[2] Haykin S. Cognitive radio: brain-empowered wireless communications [J]. IEEE J. Sel. Areas Commu. ,2005, 23(2):201-220.

[3] 王钦辉,叶保留,田宇, 等. 认知无线电网络中的频谱分配算法[J]. 电子学报,2012, 40 (1): 147-154.

[4] 郭彩丽,张天魁,曾志民,等. 认知无线电关键技术及应用的研究现状[J]. 电信科学, 2006 (8): 50-55.

[5] 邝祝芳,罗孟宇, 陈志刚. 认知无线网络中基于遗传算法的异构信道频谱感知策略[J]. 计算机应用与软件,2011,28(11):102-104,135.

[6] 曲轶龙,张虎虎,齐冬莲. 认知无线网络多用户协同频谱感知协作机制研究[J]. 计算机应用与软件,2012,29(6):1-4,56.

[7] Liang Y C, Zeng Y H, Edward C Y, et al. Sensing-Throughput Tradeoff for Cognitive Radio Networks[J]. IEEE Trans. on Wireless Communications, 2009,7(4): 1326-1336.

[8] Pei Y Y, Hoang A T, Liang Y C. Sensing-throughput tradeoff in cognitive radio networks: How frequently should spectrum sensing be carried out? [C]//Proc. IEEE PIMRC, 2007:1-5.

[9] Zhou X W, Ma J, Li Geoffrey Ye. Probability based optimization of inter sensing duration and power control in cognitive radio[J]. IEEE Trans. on Wireless Communications, 2009, 8(10):4922-4927.

[10] Stotas S, Nallanathan A. Optimal sensing time and power allocation in multiband cognitive radio networks[J]. IEEE Trans. on Communications, 2011, 59(1): 226-235.

[11] Hoseini P, Beaulieu N. On the benefits of multichannel/wideband spectrum sensing with non-uniform channel sensing durations for cognitive radio networks[J]. IEEE Trans. on Communications, 2012, 60 (9):2434-2443.

[12] Kang X, Liang Y C, Garg h K, et al. Sensing-based spectrum sharing in cognitive radio networks[J]. IEEE Trans. on Vehicular Technology, 2009, 58( 8): 4649-4654.

[13] Khoshkholgh M G, Navaie K, Yanikomeroglu H. Access strategies for spectrum sharing in fading environment: overlay, underlay and mixed [J]. IEEE Trans. on Mobile Computing, 2010, 9 (12): 1780-1793.

[14] Lee W Y, Akyildiz Ian F. Optimal spectrum sensing framework for cognitive radio networks[C]. IEEE Trans. on Wireless Communications, 2008, 7(10):3845-3856.

[15] Tian Z, Giannakis G B. Compressed sensing for wideband cognitive radios[C]//Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)2007,4:1357-1360.