基于社交网络的犯罪团伙发现算法研究

作　　者：潘　　潇　　王　　斌　　君

来源：《软件导刊》2018 年第 12 期

摘要：如何快速、有效地发现犯罪团伙是公安机关侦查办案中的关键问题之一。针对通信网络特点，改进社区发现的 Louvain 算法，并根据电信诈骗犯罪团伙利用通信网络实施诈骗的特点，提出基于相似度的犯罪团伙发现算法，以及基于属性的犯罪团伙发现算法。初步实验结果表明，改进后的 Louvain 算法可以提高通信网络社区划分效率。然后在社区中利用结构特征进行相似度判断，并结合属性特征进行聚类分析，从而为公安机关发现可疑犯罪团伙提供有效的理论与技术支撑。

关键词：社交网络;Louvain 算法;犯罪团伙识别;通信网络社区;社区发现

Research on Criminal Gang Discovery Algorithm Based on Social Networks

PAN Xiao，WANG Bin;jun

（College of Information Technology and Cyberspace Security，

People′s Public Security University of China， Beijing 100038， China）

Abstract：How to quickly and effectively discover criminal gangs is one of the key issues in the investigation for the public security organs. According to the characteristics of communication network， the community detection Louvain algorithm is improved. According to the characteristics of telecommunication fraud gangs using communication network to implement fraud， a similarity;based criminal gang discovery algorithm and attribute;based criminal gang discovery algorithm are proposed. Preliminary experiments show that the improved Louvain algorithm can improve the efficiency of the communication network community， then it uses the structural features to judge the similarity in the community and combines the attribute characteristics for cluster analysis to provide effective theoretical and technical support for public security organs to discover suspicious criminal gangs.

Key Words：social network; Louvain algorithm; criminal gang identification; communication network community;;community detection

0;引言

社会成员通过在工作、学习、生活、娱乐等活动中的相互作用而逐渐形成了某种稳定关系，进而形成社交网络。社交网络分析主要通过对网络拓扑结构关系的分析，揭示网络中的社区结构。社区发现在生物学、物理学、计算机图形学与社会学中都有着广泛应用[1;2]。此外，

社交网络的社区发现还有着重要的现实意义，比如在商业或服务业领域，划分出对其产品或服务感兴趣的一类特定消费群体，进而可以有针对地进行产品推荐活动，或提供个性化服务，还可挖掘出更多潜在客户，进一步提高企业经济效益[3]。

在我国社会经济快速发展与转型的大背景下，犯罪行为的团伙化、组织化特征愈加明显。在社交网络中，犯罪团伙也表现为某种特定结构的社区。随着信息化的推进，积累了越来越多人员关系网络、通话记录等数据，即使简单的通信数据中也包含着丰富信息。充分利用相关信息，可以挖掘海量数据背后隐藏的犯罪团伙社区。

本文对社交网络中的犯罪团伙发现方法进行研究，在介绍目前常用的两类犯罪团伙分析方法后，提出改进的 Louvain 算法，并结合社交网络的拓扑结构与节点属性，可以有效发现社交网络中潜在的犯罪团伙，为公安机关侦查办案、打击犯罪提供支持。

1;相关研究与分析

在大数据分析中，有两种针对犯罪团伙的发现思路，一种是根据掌握的犯罪人员基本信息，采用社会网络分析方法，利用犯罪人员之间的联系发现可疑犯罪团伙;另一种是根据已掌握案件信息中的犯罪人员属性及其作案特征等，使用聚类方法发现具有共同属性特征的可疑犯罪团伙。

1.1;社会网络分析方法

各国安全部门高度重视收集与分析恐怖组织的有关数据，希望通过掌握相关恐怖分子的社会网络信息，了解其组织结构，以加强对恐怖分子的防范力度[4]。为了实现该目标，人们提出多种技术方法[5;8]应用于犯罪团伙发现与犯罪打击相关工作。一些研究人员也利用概念邮件系统与属性筛选支持向量机等方法对犯罪数据的社会网络进行分析，并提出多种分析方法[9;10]。

犯罪团伙早期分析研究大多采用基于社会网络的社区发现算法，通常分为两种：①图形分割。主要采用 Kernighan;Lin 算法[11]与谱平分算法（Spectral Bisection Method）[12];②数据挖掘中的层次聚类。主要采用 GN（Girvan;Newman）算法[13]，之后许多学者又提出一系列基于 GN 算法的改进算法[14;17]。如上海交通大学李亮[18]基于 Radicchi 等提出的改进型 GN 算法，设计了嫌疑人的社会网络分解与识别模块。

1.2;属性聚类方法

聚类分析根据不同属性识别对象，将相似事物聚类在一起，能够使聚类分析很好地解决不确定事物属性的分类问题。在公共安全应用领域，聚类分析可以从大量数据中发现一些特定集

合，以帮助公安部门专注于重点对象，从而极大减轻了公安部门工作量，并有可能发现潜在犯罪嫌疑人与犯罪团伙。

基于聚类技术的犯罪行为分析通常采用基于密度的聚类算法分析案件信息属性特征，由于同一类别犯罪行为更具有相似性，因此将具有高相似性的嫌疑人员聚为一类，可以帮助调查人员找出疑似犯罪团伙。例如：具有相同入室盗窃模式或逃离模式，以及犯罪地点相同的人员可能成为同一团伙。西南交通大学邓灵评[19]利用聚类方法对入室盗窃犯罪数据进行分析，将相似度较高的对象归为一类，进行串并案分析研判。在同一时间段内找出具有相似作案方式、相同受害者，或盗窃相似财物的案件，可推断同一类犯罪案件的实施者可能属于同一团伙。

2;犯罪团伙发现算法模型

犯罪团伙不仅是成员之间有通信联系，而且其在选择对象，以及选择时间、作案方式和作案工具等方面也呈现出高度的相似性。目前，尚未有在大量数据集上进行犯罪团伙发现的研究，本文采用社区发现与属性判定相结合的方法，改进社区发现中的 Louvain 算法[20]，使其能够更好地应用于日常生活的通信网络中。将对应通话号码的每个节点划分为不同社区，然后在社区中利用结构特征进行相似度判断，并结合属性特征进行聚类分析，从而发现具有相似特征的团伙。

2.1;Louvain 算法改进

Louvain 算法采用 Mark Newman 等研究发现的模块度作为衡量网络社区划分优劣的重要指标，通过比较现有网络与基准网络在相同社区划分情况下的连接密度差，以衡量网络社区划分的优劣。其中，基准网络是与原网络具有相同度序列的随机网络。假设 A 是复杂网络的邻接矩阵，k;v=∑wA;vw 表示节点 v 的度。一条边（v，w）在基准网络中存在的概率为 k;vk;w2m，其中 m 表示网络图 A 中的连边数目。模块度的完整数学表达如公式（1）所示。

其中，c;v 表示节点 v 所属社区。如果 u=v，δ（u，v）=1，反之，δ（u，v）;=0。该公式的数学意义为，网络中同一社区内部边的比例与在同样社区结构下基准网络内部边比例的期望值之差。

Louvain 算法首先为网络每个节点分配一个不同社区，因此在初始划分中的节点数量等于社区数量。对于每个节点;i 都考虑其邻居节点 j，计算将 i 从原社区删除并移动到 j 所在社区的模块度增益，将节点 i 移动到正增益最大的 j 所在社区。对所有节点重复应用该过程，在过程中一个节点可能被多次考虑，当模块度为局部最大值时，即没有单独节点可以移动并改善模块度时，则第一阶段停止。

算法第二阶段建立一个新网络，其节点是第一阶段发现的社区，新节点之间的连接权重为相应两社区中节点之间连接权重之和，同一社区节点之间的连接导致新网络中该社区的自我循环。一旦完成了第二阶段，即可将第一阶段算法重新应用到所得到的加权网络中进行迭代。

移动通信网络在日常生活中应用十分普遍，且数据量巨大，多数社区发现算法对其进行社区划分时，无法满足时间效率方面的要求。通信网络分析大多采用一段时间的通信数据，其存在大量两两节点之间的正常通信，而电信诈骗团伙的通话网络往往是许多节点的集聚与交叉。对于正常通信的两两节点，由于节点间的内部连接性与外部孤立性，必然会被划分到一个社区，如果直接使用传统 Louvain 算法对通信网络进行处理，则需要考量这些节点移动到相邻节点的模块度增益，从而浪费大量计算资源。

改进 Louvain 算法增加了预处理阶段，先计算每个节点的度，将仅有两两相连的节点对直接删除，而在两两相连节点对中如果存在一个度大于 1 的，则将两节点合并。通过预处理减少网络中的节点数量，可缩短 Louvain 算法第一阶段的模块度增益计算过程，提高算法效率。

算法 1：改进的 Louvain 社区发现算法

输入：网络图;G（V，E），包括节点 V 和连边 E 的信息;

输出：对 G（V，E）进行社区划分的社区集合;

1;将 G;中每个节点初始化为一个社团;

2;for （对所有节点;x）;//对所有节点执行预处理

3 计算节点 x 的度 d（x）;

4if （;d（x）==0） 删除节点 x;

5if （;d（x）==1）

6;for （对;x 所有邻接节点 x′）

7 计算节点 x′的度 d（x′）;

8if （;d（x′）==1） 将节点 x 和 x′;删除;

9else 将节点;x 和 x′;合并;

10;end for

11;end for;//结束预处理

12;计算此时模块度并存入;Q;1，Q;3=Q;1

13;Q;2=Q;3;

14;for i =;1 to n;//n 为网络中节点个数

15 将节点 v;i 从原来社区中取出;

16 将 v;i;加入到使模块度增益 Δ;Q 最大的社区中;

17;end for

18;计算此时模块度并存入;Q;1;;

19;将各个社区合并成一个点集合;

20;将不同集合中包含的点存入相应集合数组 communities;

21;if;Q;1>Q;2;， 转到步骤 12;

22;结束

2.2;基于相似度的电信诈骗团伙发现算法

多数电信诈骗团伙为了达到诈骗目的，在短时间内会对一批电话号码进行集中呼叫，并扮演不同身份进行团伙诈骗，在网络中相当于几个节点共享许多相邻的邻居节点。在使用改进 Louvain 算法对通信网络进行社区划分后，可能会发现其中的犯罪团伙。

如果网络中两个节点共享很多相同邻居节点，则两个节点是结构等价的，如图 2 中的节点 4、5。因此，利用通信进行犯罪的团伙在网络中反映为结构等价的，对于类似团伙可采用余弦相似性进行判别。

在几何学中，两个向量;x 和 y;的相似性可用余弦值公式（2）表示。

将社交网络邻接矩阵的第 i、j 行（或列）分别看成两个向量，然后将两个向量之间的夹角余弦值用于相似性度量。在无向网络中，对应邻接矩阵中两行点积为∑kA;ikA;kj，相似性测度如公式（3）所示。

因此，得到通信网络中诈骗团伙发现算法 2。

算法 2：通信网络中诈骗团伙发现算法

输入：;网络图 G（V，E），包括节点 V 和连边 E 的信息;

输出：可能的电信诈骗犯罪团伙;

1;调用算法 1;

2;for （对 communities 的每个社区;A）

3 计算 A 中每个节点 i 的度 d（i）;

4 提取 A 中前 x;%的节点;;//;x 是一个经验值

5 利用公式（3）计算这些节点两两之间的余弦相似性;

6 按照余弦相似性将 A;划分为不同的等价类，形成潜在的电信诈骗犯罪团伙;

7;end for

8;结束

2.3;基于属性的电信诈骗团伙发现算法

使用改进 Louvain 算法进行社区划分后，利用每个节点（号码）对应实体人属性信息，可在社区内采用基于密度的聚类方法进行犯罪团伙发现。节点（号码）除归属地等属性特征外，如果其对应实体人（号主）是公安机关已掌握的犯罪嫌疑人，可根据其作案手段、作案工具、选择对象、选择时间、选择处所等特征进行聚类分析，识别关于参数 ε 和 Minpts 的所有核心节点，核心节点及其邻域形成的簇则可能成为犯罪团伙，由此得到结合属性的犯罪团伙发现算法 3。

算法 3：结合属性的电信诈骗犯罪团伙发现算法

输入：;网络图 G（V，E），包括节点 V 和连边 E 的信息;

输出：基于密度簇的集合;

1;调用算法 1;

2;依次对每个社区中结合属性的节点进行如下操作

3;标记一个社区内所有包含属性的节点为 unvisited;

4;do

5 随机选择一个 unvisited 对象;p;

6 标记 p;为 visited;

7if;p 的 ε;邻域最少有 Minpts 个对象;

8;创建一个新簇;C，并将 p 添加到 C;

9;令 N 为 p 的 ε 邻域中的对象集合;

10;for;N 中每个点 p′;;

11;if;p;′是 unvisited

12;标记;p;′为 visited;

13;if;p′的 ε;邻域至少有 Minpts 个点，将这些点添加到;N;;

14;if;p′还不是任何簇的成员，将 p′添加到 C;

15;end if

16 输出 C;

17;end for

18;else 标记 p 为噪声;

19;until 没有标记为 unvisited 的对象

3;实验结果与分析

3.1;改进 Louvain 算法实验分析

实验平台处理器为 Intel（R）Core（TM）i5-6300HQ CPU@2.30GHz，内存为 8GB。为避免标准数据集的单一性，根据不同时间段与不同数据量特点，从通信话单网络中抽取 8 组实验数据集。每组数据集中的节点数量依次增长，并且包含不同数量的正常通信节点，根据通信的

主、被叫方建立节点间的网络关系连接。表 1 是两种算法在不同数据集上的时间效率与模块度值对比。

实验结果表明，对实际通信网络进行社区划分时，本文提出的改进 Louvain 算法在数据处理时间方面优于原始算法。在表 1 中，当数据量选取较少，即节点数分别为 522 与 3 043 时，改进算法与原始算法处理时间相差较小，且运行速度较快;当节点与边的数量增加后，改进 Louvain 算法在处理时间方面明显优于原始 Louvain 算法。同时计算改进算法与 Louvain 算法社区划分后的模块度值，发现两种算法模块度值完全相同。因此，改进算法与原始 Louvain 算法相比，不会降低社区划分结果的准确度。

3.2;基于相似性的电信诈骗犯罪团伙发现算法分析

在算法 2 实验中，选取不同日期、不同时间段的通信话单数据进行社区划分，发现某周五傍晚时的一个社区结构如图 3 所示，在图中存在出度较高的 8 个节点（号码），其对应电话号码经查询均来自同一地域。利用上文提出的相似性判别式进行计算可知，8 个节点中两两节点的相似度值远高于社区内其它节点的两两匹配相似度，推断其有较大概率为犯罪团伙，正集中对部分通信用户进行诈骗。

4;结语

本文根据通信网络结构特点，改进了现有 Louvain 算法，提高了通信网络社区划分效率。在此基础上，提出基于改进 Louvain 算法的犯罪团伙发现算法与结合属性的犯罪团伙发现算法。初步实验结果表明，将本文提出的改进 Louvain 算法应用于通信网络社区划分能有效提高社区划分效率，在划分后的社区使用两种犯罪团伙发现算法，可以帮助公安部门利用通信网络发现可疑犯罪团伙及其组织结构。然而，由于目前所掌握的社区中节点属性信息不足，无法充分结合属性对犯罪团伙发现算法进行实验验证，因此需要进一步结合公安业务，在完善与改进社区划分的基础上，针对属性聚类实验结果进行分析，从而更好地为公安机关利用社交网络发现犯罪团伙提供技术支持。

参考文献：

[1];PABLO M GLEISER， LEON DANON. Community structure in Jazz[J]. Advances in Complex Systems， 2011，6（4）： 565;573.

[2];HOLME P， HUSS M， JEONG H. Subnetwork hierarchies of biochemical pathways[J]. Bioinformatics， 2003，19（4）： 532.

[3];吴成钢，杨光，张翔，等.推荐系统的应用及其安全性研究[J].信息网络安全，2011（8）： 69;71.

[4];MCANDREW D. The structural analysis of criminal networks[J]. International Journal of Middle East Studies， 1999，8（2）： 272;281.

[5];谢秦川.非法交易犯罪团伙的社会网络分析研究[J].信息网络安全，2014（6）： 88;91.

[6];CHEN H， CHUNG W， XU J J， et al. Crime data mining： a general framework and some examples[J]. Computer， 2004， 37（4）： 50;56.

[7];XU J J， CHEN H. CrimeNet explorer： a framework for criminal network knowledge discovery[J]. ACM Transactions on Information Systems， 2005，23（2）： 201;226.

[8];VEL O D， ANDERSON A， CORNEY M， et al. Mining e;mail content for author identification forensics[J]. ACM Sigmod Record， 2001，30（4）： 55;64.

[9];温粉莲.基于犯罪数据挖掘系统的关键技术研究[D].成都：四川大学，2007.

[10];刘威，唐常杰，乔少杰.基于概念邮件系统的犯罪数据挖掘新方法[J]. 计算机科学，2007， 34（2）： 213;215.

[11];KERNIGHAN B W， LIN S. An efficient heuristic procedure for partitioning graphs[J]. Bell System Technical Journal， 2014， 49（2）： 291;307.

[12];POTHEN A， SIMON H D， LIOU K P. Partitioning sparse matrices with eigenvectors of graphs[J]. SIAM Suornal on Matrix Analysis and Aoolications，1990， 11（3）： 430;452.

[13];GIRVAN M， NEWMAN M E. Community structure in social and biological networks[J]. Proc Natl Acad Sci USA， 2002， 99（12）： 7821;7826.

[14];TYLER J R， WILKINSON D M， HUBERMAN B A. Email as spectroscopy： automated discovery of community structure within organizations [M].Netherlands Communities and Technologies.;2003：143;153.

[15];许为，林柏钢，林思娟，等. 一种基于用户交互行为和相似度的社交网络社区发现方法研究[J]. 信息网络安全， 2015（7）： 77;83.

[16];ZHOU H. Distance， dissimilarity index， and network community structure[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics， 2003， 67（1）： 061901.

[17];FORTUNATO S， LATORA V， MARCHIORI M. Method to find community structures based on information centrality[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics，2004， 70：056104.

[18];李亮.基于社会网络分析的犯罪团伙识别系统[D].上海：上海交通大学， 2008.

[19];邓灵评. 基于数据挖掘的犯罪行为分析及系统实现[D].成都：西南交通大学， 2014.

[20];BLONDEL V D， GUILLAUME J L， LAMBIOTTE R， et al. Fast unfolding of communities in large networks[J]. Journal of Statistical Mechanics， 2008（10）：155;168.

$$Q = \frac{1}{2m}\sum_{v,w}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(c_v, c_w) \qquad (1)$$

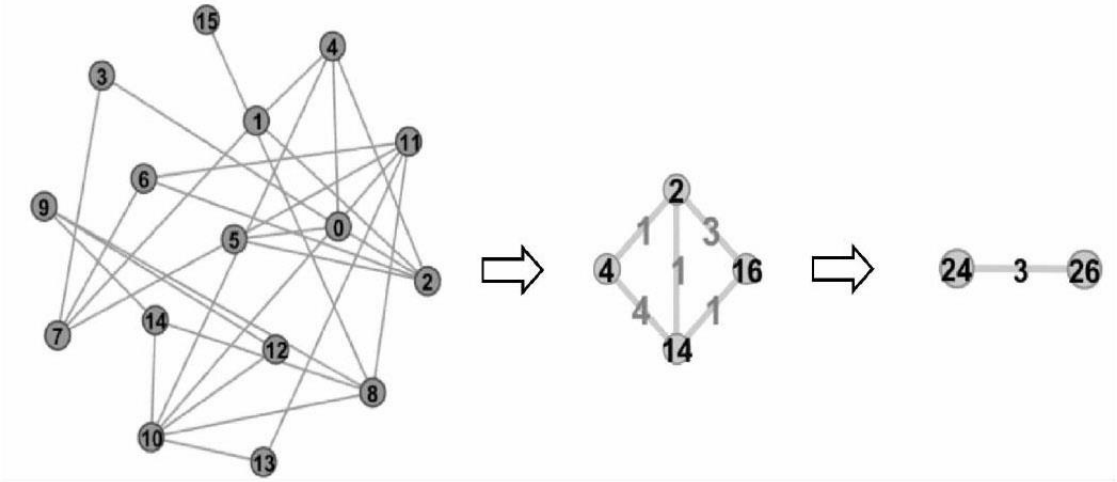

图 1　Louvain 算法过程

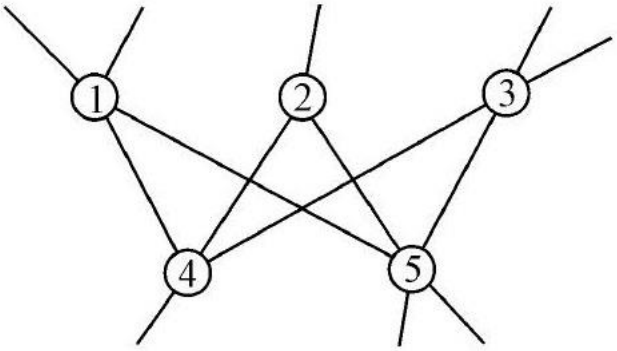

图 2　结构等价

$$cos\theta = \frac{x \cdot y}{|x||y|} \tag{2}$$

将社交网络邻接矩阵的第 i、j 行(或列)分别看成两个向量,然后将两个向量之间的夹角余弦值用于相似性度量。在无向网络中,对应邻接矩阵中两行点积为 $\sum_k A_{ik}A_{kj}$,

$$\sigma_{ij} = cos\theta = \frac{\sum_k A_{ik}A_{kj}}{\sqrt{\sum_k A_{ik}^2}\ \sqrt{\sum_k A_{jk}^2}} \tag{3}$$

表 1　改进算法效率与模块度比较

实验抽取数据集	算法处理时间(ms)		模块度(Q)	
(节点数/边数)	Louvain	改进算法	Louvain	改进算法
552/279	20	14	0.996	0.996
3 043/1 581	39	24	0.974	0.974
21 716/11 346	664	86	0.994	0.994
52 279/27 888	1 982	264	0.999	0.999
72 827/39 524	3 501	399	0.995	0.995
195 547/111 077	26 618	3 180	0.997	0.997
333 477/195 897	72 266	5 799	0.997	0.997
427 858/260 037	134 492	10 697	0.999	0.999

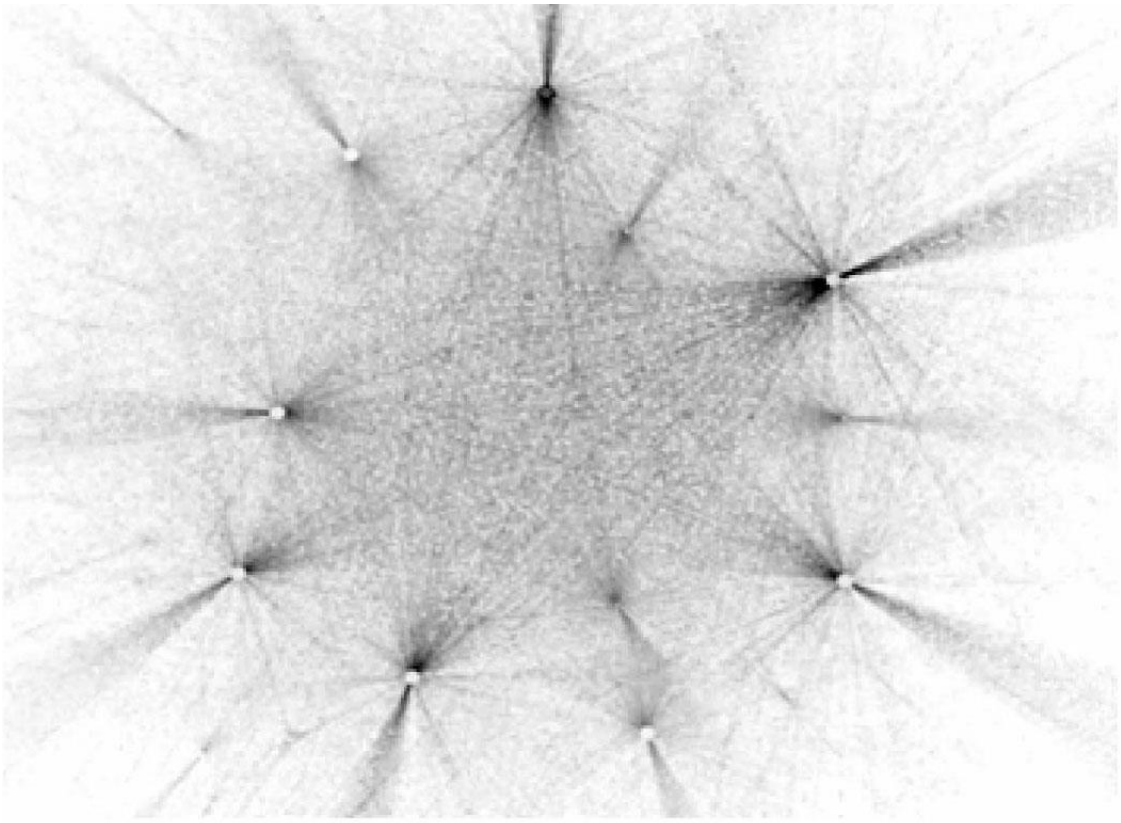

图 3　社区结构