软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2710−2720 [doi: 10.3724/SP.J.1001.2013.04477]

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

社群演化的稳健迁移估计及演化离群点检测*

胡 云[1,2], 王崇骏[1], 谢俊元[1], 吴 骏[1], 周作建[1]

[1](南京大学 计算机科学与技术系,江苏 南京 210093)

[2](淮海工学院 计算机工程学院,江苏 连云港 222005)

通讯作者: 王崇骏, E-mail: 15250998131@139.com

摘 要: 时序数据集中的社群演化模式是网络行为动力学研究与应用的重要领域.基于社群演化的离群点检测不仅能够发现新颖的异常行为模式,同时也有利于更准确地理解社群的演化趋势.运用成员关于社群隶属关系的变化,提出了社群演化迁移矩阵的概念,研究并揭示了迁移矩阵的若干性质及其与社群结构演化之间的关系.在采用稳健回归*M*-估计方法进一步优化迁移矩阵降低异常点干扰的同时,对社群演化离群点加以刻画和定义.鉴于复杂网络包含大量随机游走的边缘个体,所定义的离群点综合考虑其在社群中角色的变化和相对于社群总体迁移模式的差异.基于上述思想提出的演化离群点检测算法能够适应各类社群演化趋势,更有效地聚焦和发现大规模社会网络中重要成员的异常演化行为.实验结果表明,所提出的方法能够从大规模社会网络演化序列中发现重要的离群演化模式,并在现实中找到合理的解释.

关键词: 时序数据集;社群演化;迁移矩阵;稳健回归;离群点检测算法

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 胡云,王崇骏,谢俊元,吴骏,周作建.社群演化的稳健迁移估计及演化离群点检测.软件学报,2013,24(11):2710−2720. http://www.jos.org.cn/1000-9825/4477.htm

英文引用格式: Hu Y, Wang CJ, Xie JY, Wu J, Zhou ZJ. Robust transition estimation for community evolution and evolutionary outlier detection. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2710−2720 (in Chinese). http://www.jos.org.cn/1000-9825/4477.htm

Robust Transition Estimation for Community Evolution and Evolutionary Outlier Detection

HU Yun[1,2], WANG Chong-Jun[1], XIE Jun-Yuan[1], WU Jun[1], ZHOU Zuo-Jian[1]

[1](Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)

[2](School of Computer Engineering, Huaihai Institute of Technology, Lianyungang 222005, China)

Corresponding author: WANG Chong-Jun, E-mail: 15250998131@139.com

Abstract: Community evolutionary pattern analysis in temporal datasets is a key issue in social network dynamics research and applications. Identifying outlying objects against main community evolution trends is not only meaningful by itself for the purpose of finding novel evolution behaviors, but also helpful for better understanding the mainstream of community evolution. Upon giving the belonging matrix of community members, this study defines a type of transition matrix to characterize the pattern of the evolutionary dynamic between two consecutive belonging snapshots. A set of properties about the transition matrix is discussed, which reveals its close relation to the gradual community structural change in quantity. The transition matrix is further optimized using *M*-estimator Robust Regression methods by minimizing the disturbance incurred by the outliers, and the abnormality of the outlier objects can then be computed at the same time. Considering that large proportion of trivial but nomadic objects may exist in large datasets like those of complex social networks, focus is placed only on the community evolutionary outliers that show remarkable difference from the main bodies of their communities and sharp change of their membership role within the communities. A definition on such type of local and

* 基金项目: 国家自然科学基金(61375069, 61105069); 中国博士后科学基金(2011M500846); 江苏省科技支撑计划(BE2012181); 江苏省教育厅自然科学基金(11KJB520001); 江苏高校优势学科建设工程资助项目; 计算机软件新技术国家重点实验室自主课题(ZZKT2013B11)

收稿时间: 2013-04-29; 修改时间: 2013-07-17; 定稿时间: 2013-08-27

global outliers is given, and an algorithm on detection such kind of outliers is proposed in this paper. Experimental results on both synthetic and real datasets show that the proposed approach is highly effective in discovering interesting evolutionary community outliers.
Key words: temporal dataset; community evolution; transition matrix; robust regression; outlier detection algorithm

社群(community)是指网络中结点的聚簇,是社会网络的基本结构.自 Girvan 和 Newman[1]于 2002 年开创复杂网络社群结构研究以来,有关社群发现、社群结构分析和社群演化动力学领域的研究成果大量涌现.社群结构演化模式的研究主要关注生物或社会网络中的社群随时间发展变化的趋势、规律和内在动因.

所谓社群演化离群点是一类建立在社群演化趋势基础上的异常数据对象.它们混杂在包含多个相互交叠的社群随时间演化的序列数据中,但与数据中所有社群在演化的模式上又有显著的区别.例如,在 DBLP 数据库中,一位作者的研究兴趣一般涉及若干个不同领域,如果将每个研究领域的研究者视为一个社群,则该作者在不同领域研究成果的信息反映了其在若干个社群中的参与度.由于在特定时期内,某一领域(社群)中的研究者在研究兴趣上的迁移具有一定的同步性,这反映出了该社群随时间演化的特性.因此,如果该研究者在某一时刻研究兴趣突然转向与其他同行截然不同的研究领域,则可以认为其构成了相对于原社群的演化离群点.当然,在这种情景下,最引人注目的莫过于社群中核心成员的异常行为,其异常演化有可能标志着一个新的研究领域正在被开辟.

离群点检测始终是数据挖掘的重要研究主题.Knor[2],Hawkins[3],Breunig[4]等人提出了一系列面向各类应用场合的算法[5−8].近年来,面向时序数据集的演化离群点和社群离群点研究成为新的热门主题.Ge 等人[9]面向时空数据库开展基于路径(trajectory)检测的演化离群点检测研究,其研究对象为空间数据库中物体在时间维度上的异常变化特征.Gao 等人[10]在其有关社会网络结构的研究中首次提出了关于社群离群点的概念,通过在信息网络图中引入概率模型,以 k 个指标标识社群结点的正常行为模式,通过隐马尔可夫随机域模型将图中的连接信息用于表达结点的(隐)离群指标,其工作仍限于对数据集的静态分析,并没有考虑社群随时间的演化行为.Gupta 等人[11,12]首次将社群结构引入离群点的发现研究,其主要思想是:通过对演化中两个相邻数据视图中社群结构的匹配,分化出社群匹配中具有显著差异的个体,并给出了一种集成社群匹配与离群点发现为一体的算法.然而,Gupta 的工作并没有考虑成员在社群中的属性差异.大量的研究表明,复杂网络是由少数连接度高的核心成员和大量连接度非常低的一般成员组成的,如符合幂率分布,而大量的一般成员在行为上具有极大的不确定性.如不加区分地识别出所有的社群离群点,显然忽略了问题的关键.此外,Gupta 的社群匹配方法要求匹配矩阵的全部元素是非负的,这在应用中必然限制了匹配精度.事实上,从最小二乘回归角度看,其社群匹配过程等价于求解超定矩阵方程的最优解矩阵,而这一最优解矩阵并不能保证每个元素非负.

本文所做的创新性研究工作包括以下 3 个方面:(1) 基于社群演化的思想,从社群演化视图的最优拟合问题出发,提出了社群迁移矩阵的概念,并运用稳健回归(robust regression)中的 M-估计方法,通过不断发现并调整异常数据点的权重因子,求解社群演化的稳健迁移矩阵.该方法将传统的回归分析方法推广到矩阵分析的情形.(2) 研究并证明了迁移矩阵所具有的一系列特殊性质.这些性质有助于从社群演化的全局角度揭示其在迁移矩阵上的表征,说明了迁移矩阵在社群演化模式分析中的重要作用.(3) 借助稳健迁移矩阵对离群点的辨识能力,并结合其相对于社群的角色变化特征,提出了基于社群演化的具有显著意义的离群点的定义与检测算法,避免了大规模网络中随机游走个体对离群点识别的干扰.

本文第 1 节介绍社群演化,并证明其迁移矩阵的有关性质.第 2 节给出离群点的表示和检测算法,并分析算法的效率.第 3 节利用人工数据集与真实数据集开展实验研究,验证算法的优越性和有效性.第 4 节进行总结.

1 网络社群及其演化

1.1 社会网络社群分析

演化中的社会网络可以抽象为由结点及结点间连边组成的图序列 $\boldsymbol{G}_t=(\boldsymbol{V}_t,\boldsymbol{E}_t),t=0,1,\ldots,T$,其中,$\boldsymbol{V}_t$ 为时刻 t 时全体网络成员组成的结点集合;$\boldsymbol{E}_t$ 是时刻 t 时全体成员结点之间的关系集合,反映为结点之间的连边.根据研究问题的不同,$\boldsymbol{E}_t$ 的元素可以是有向弧、无向弧,并可以赋以权重,从而形成(加权)有向图、无向图、二分图等.本

文假设网络成员数是稳定的,即 $N=|\boldsymbol{V}_t|$保持不变,而对结点之间的连接性质不作要求.

网络的社群结构随时间的演化行为主要包括社群数量的增减和社群结构的改变,即社群分裂(split)、合并(merge)、结构扩张膨胀(expend)和收缩(contract)以及成员在社群间的迁移等[12].与社群结构发现研究相比,社群演化方面的研究仍处于初级阶段.近几年,在社群发现取得长足进步的基础上,社群演化研究得到了越来越多的关注[13,14].文献[15]基于完全子图渗流社群发现方法研究社群演化,得到一个有趣的结论:小社群的稳定性是保证其存在的前提,而大社群的动态性是其存在的基础.本文主要研究社群演化中的离群点检测问题,因此假设在任一时刻网络的社群结构是已知的.

1.2 社群演化及其迁移矩阵

给定由 N 个成员组成的社会网络图序列 $\boldsymbol{G}_t=(\boldsymbol{V}_t,\boldsymbol{E}_t),t=0,1,\ldots,T$,则在时刻 t 全体成员相对于 K_t 个社群的隶属矩阵 $\boldsymbol{P}(t)\in[0,1]^{N\times N}$ 可以定义为

$$\boldsymbol{P}(t)=\begin{bmatrix} p_{11}(t) & \ldots & p_{1K_t}(t) \\ \ldots & \ldots & \ldots \\ p_{N1}(t) & \ldots & p_{NK_t}(t) \end{bmatrix},\ t=0,1,\ldots,T \tag{1}$$

其中,K_t 为时刻 t 时的社群总数;$\boldsymbol{P}(t)$的第 i 个行向量表示社群成员 V_i 在时刻 t 关于 K_t 个社群的隶属度分布:$\boldsymbol{P}_i(t)=(p_{i1}(t),\ldots,p_{iK_t}(t))^T$,满足 $\sum_{j=1}^{K_t}p_{ij}(t)=1,i=1,\ldots,N$,且 $x_{ij}\geqslant 0$.

定义 1. 我们称成员 V_i 在时刻 t 隶属于社群 j,如果 $p_{ij}(t)=\max\{p_{i1}(t),\ldots,p_{iK_t}(t)\}$. 即,一个社群的成员相对于本社群的隶属度不小于其对其他社群的隶属度.

针对问题的不同,网络成员的社群隶属度可以采用不同的度量方法.最简单的情形是依据网络个体的连接关系.当然,研究者可以综合考虑各方面的因素,还可以采用手工标注或机器学习的方法加以定义.

定义 2(迁移矩阵). 设 $\boldsymbol{P}(t),\boldsymbol{P}(t+1)$为相邻两个时刻 t 和 $t+1$ 网络成员的社群隶属矩阵,称 $k_t\times k_{t+1}$ 阶矩阵 $\boldsymbol{S}(t)=[s_{ij}(t)]\in\boldsymbol{R}^{k_t\times k_{t+1}}$ 为 $\boldsymbol{P}(t)$到 $\boldsymbol{P}(t+1)$的迁移矩阵,如果:

$$\Phi(S)=\sum_{i=1}^{N}\sum_{j=1}^{k_t}\left(p_{ij}(t+1)-\sum_{l=1}^{k_{t+1}}p_{il}(t)s_{lj}(t)\right)^2=\text{Min!} \tag{2}$$

其中,k_t,k_{t+1} 分别表示时刻 t 和 $t+1$ 时的社群数.

容易证明,定义 2 中关于两个矩阵的最小二乘优化问题等价于求解矩阵方程:

$$\begin{bmatrix} \langle \boldsymbol{p}_{\cdot,1}(t+1),\boldsymbol{p}_{\cdot,1}(t)\rangle & \ldots & \langle \boldsymbol{p}_{\cdot,k_{t+1}}(t+1),\boldsymbol{p}_{\cdot,1}(t)\rangle \\ \ldots & \ldots & \ldots \\ \langle \boldsymbol{p}_{\cdot,1}(t+1),\boldsymbol{p}_{\cdot,k_t}(t)\rangle & \ldots & \langle \boldsymbol{p}_{\cdot,k_{t+1}}(t+1),\boldsymbol{p}_{\cdot,k_t}(t)\rangle \end{bmatrix}=\begin{bmatrix} \langle \boldsymbol{p}_{\cdot,1}(t),\boldsymbol{p}_{\cdot,1}(t)\rangle & \ldots & \langle \boldsymbol{p}_{\cdot,k_t}(t),\boldsymbol{p}_{\cdot,1}(t)\rangle \\ \ldots & \ldots & \ldots \\ \langle \boldsymbol{p}_{\cdot,1}(t),\boldsymbol{p}_{\cdot,kt}(t)\rangle & \ldots & \langle \boldsymbol{p}_{\cdot,k_t}(t),\boldsymbol{p}_{\cdot,k_t}(t)\rangle \end{bmatrix}\cdot\begin{bmatrix} s_{11} & \ldots & s_{1k+1} \\ \ldots & \ldots & \ldots \\ s_{k_t1} & \ldots & s_{kk+1} \end{bmatrix} \tag{3}$$

其中,$\boldsymbol{p}_{\cdot,i}(t+1),\boldsymbol{p}_{\cdot,j}(t)$分别为隶属矩阵 $\boldsymbol{P}(t),\boldsymbol{P}(t+1)$的第 i,j 个列向量,$i=1,\ldots,k_{t+1},j=1,\ldots,k_t$.

我们之所以将 $\boldsymbol{S}(t)$称为迁移矩阵,不仅是因为矩阵 $\boldsymbol{S}(t)$最好地满足了 $\boldsymbol{P}(t+1)\approx\boldsymbol{P}(t)\boldsymbol{S}(t)$这一条件,定量地刻画了社群从一个时刻到下一时刻的迁移性态,而且因为矩阵 $\boldsymbol{S}$ 各列的和值的变化能够反映出一个社群向另一个社群在数值上的迁移度.矩阵 $\boldsymbol{S}$ 具有若干特殊的类似于马尔可夫概率转移矩阵的性质,这些性质对于揭示社群演化规律和简化迁移矩阵的计算具有重要的作用.

性质 1. 迁移矩阵 $\boldsymbol{S}(t)$每行元素之和恒为 1,即 $\sum_{j=1}^{K_{t+1}}s_{ij}(t)=1,i=1,\ldots,K_t$.

证明:为简便起见,我们分别记两个相邻的隶属矩阵为 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 而略去时间变量,且仅以包含两个社团的情形加以证明.此时,$\boldsymbol{P}=(p_{ij})_{N\times 2}$ 和 $\boldsymbol{Q}=(q_{ij})_{N\times 2},\boldsymbol{S}=(s_{ij})$为 2 阶方阵.对公式(2)求偏导,得到:

$$\frac{\partial\Phi}{\partial s_{lj}}=2\sum_{i=1}^{N}p_{il}\left(q_{ij}-\sum_{l=1}^{2}p_{il}s_{lj}\right)=0,\ l,j=1,2.$$

即 $$\left(\sum_{i=1}^{N}p_{i1}p_{il}\right)\cdot s_{1j}+\left(\sum_{i=1}^{N}p_{i2}p_{il}\right)\cdot s_{2j}=\sum_{i=1}^{N}p_{il}q_{ij},\ j=1,2;\ l=1,2.$$

根据 Cramer 法则,s_{11},s_{12}的值分别为

$$s_{11}=\frac{\begin{vmatrix}\sum q_{i1}p_{i1} & \sum p_{i1}p_{i2}\\ \sum q_{i1}p_{12} & \sum p_{i2}^2\end{vmatrix}}{\begin{vmatrix}\sum p_{i1}p_{i1} & \sum p_{i1}p_{i2}\\ \sum p_{i1}p_{12} & \sum p_{i2}^2\end{vmatrix}},\ s_{12}=\frac{\begin{vmatrix}\sum q_{i2}p_{i1} & \sum p_{i1}p_{i2}\\ \sum q_{i2}p_{12} & \sum p_{i2}^2\end{vmatrix}}{\begin{vmatrix}\sum p_{i1}p_{i1} & \sum p_{i1}p_{i2}\\ \sum p_{i1}p_{12} & \sum p_{i2}^2\end{vmatrix}}.$$

我们记 $$s_{11}+s_{12}=\frac{\begin{vmatrix}\sum p_{i1} & \sum p_{i1}p_{i2}\\ \sum p_{12} & \sum p_{i2}^2\end{vmatrix}}{\begin{vmatrix}\sum p_{i1}^2 & \sum p_{i1}p_{i2}\\ \sum p_{i1}p_{12} & \sum p_{i2}^2\end{vmatrix}}=\frac{A}{B},$$

则有:

$$\begin{aligned}A&=\left(N-\sum p_{i2}\right)\cdot\sum p_{i2}^2-\sum p_{i2}\cdot\sum(1-p_{i2})p_{i2}=N\sum p_{i2}^2-\left(\sum p_{i2}\right)^2,\\ B&=\sum(1-p_{i2})^2\sum p_{i2}^2-\left(\sum(1-p_{i2})p_{i2}\right)^2\\ &=N\sum p_{i2}^2-2\sum p_{i2}\sum p_{i2}^2+\left(\sum p_{i2}^2\right)^2-\left(\sum p_{i2}\right)^2+2\sum p_{i2}\sum p_{i2}^2-\left(\sum p_{i2}^2\right)^2\\ &=N\sum p_{i2}^2-\left(\sum p_{i2}\right)^2\\ &=A.\end{aligned}$$

因此有 $s_{11}+s_{12}=1$.即行值保持为 1. □

性质 1 说明,迁移矩阵 $\boldsymbol{S}$ 具有与马尔可夫转移矩阵类似的性质,即行的和值始终保持为 1.但不同的是,$\boldsymbol{S}$ 的元素并不能保证全部非负,如性质 2 所述.

性质 2. 将隶属矩阵 $\boldsymbol{P}(t)$的第 i 列统一减去分量 h 并加到第 j 列上形成 $\boldsymbol{P}(t+1)$,则对应的迁移矩阵恰为单位矩阵经在 i 列统一减去 h 并加到 j 列后得到的矩阵.

证明:限于篇幅,仅就 $k=2$(即 $\boldsymbol{P}$ 和 $\boldsymbol{Q}$ 均只包含两列)的情形加以证明,其他情形的证明相类似.

此时,不妨假设 $q_{i1}=p_{i1}+h,i=1,\ldots,N$,则相应地,$q_{i2}=p_{i2}-h$.对 Φ 关于 s_{11},s_{21} 求偏导所得的方程组为

$$\begin{cases}\sum(p_{i1}+h)p_{i1}=\sum p_{i1}^2\cdot s_{11}+\sum p_{i2}p_{i1}\cdot s_{21}\\ \sum(p_{i1}+h)p_{i2}=\sum p_{i1}p_{i2}\cdot s_{11}+\sum p_{i2}^2\cdot s_{21}\end{cases}.$$

于是:

$$\begin{aligned}s_{11}&=\left[\left(\sum p_{i1}^2+h\sum p_{i1}\right)\sum p_{i2}^2-\left(\sum p_{i1}p_{i2}+h\sum p_{i2}\right)\sum p_{i1}p_{i2}\right]\Big/\left[\sum p_{i1}^2\sum p_{i2}^2-\left(\sum p_{i1}p_{i2}\right)^2\right]\\ &=1+h\cdot\left[\sum p_{i1}\sum p_{i2}^2-\sum p_{i2}\sum p_{i1}p_{i2}\right]\Big/\left[\sum p_{i1}^2\sum p_{i2}^2-\left[\sum p_{i1}p_{i2}\right]^2\right]=1+h,\\ s_{21}&=\left[\left(\sum p_{i1}p_{i2}+h\sum p_{i2}\right)\sum p_{i1}^2-\left(\sum p_{i1}^2+h\sum p_{i1}\right)\sum p_{i1}p_{i2}\right]\Big/\left[\sum p_{i1}^2\sum p_{i2}^2-\left(\sum p_{i1}p_{i2}\right)^2\right]\\ &=h\left[\sum p_{i2}\sum p_{i1}^2-\sum p_{i1}\sum p_{i1}p_{i2}\right]\Big/\left[\sum p_{i1}^2\sum p_{i2}^2-\left(\sum p_{i1}p_{i2}\right)^2\right]=h.\end{aligned}$$

类似地,可以证明:$s_{12}=-h;s_{22}=1-h$. □

显然,性质 2 可以直接推广到隶属矩阵多个列(即社群)间同步转化的情形.即,如果在一个周期的演化中,全体社群成员在社群间的迁移是同步的,则这种行为可通过迁移矩阵精确匹配而不存在误差.性质 2 还说明,迁移矩阵元素的非负性不能保证,它反映对应的社群在一轮演化中的网络成员参与度的增减量.

性质 3. 若隶属矩阵 $\boldsymbol{P}(t)$,$\boldsymbol{P}(t+1)$包含的社群数保持不变,即 $k_t=k_{t+1}$,且 $\boldsymbol{P}(t)$,$\boldsymbol{P}(t+1)$均不包含全 0 的列,则矩阵 $\boldsymbol{S}(t)$为非奇异矩阵,且$\sum_{i,j}s_{ij}(t)=k_t$(证明略).

性质 4. 若 $\boldsymbol{P}(t)$,$\boldsymbol{P}(t+1)$包含的社群数保持不同,即 $k_t \neq k_{t+1}$,则总可以在矩阵 $\boldsymbol{S}(t)$中对应的位置添加相应的 0 行或 0 列,使 $\boldsymbol{S}(t)$为方阵,且$\sum_{i,j} s_{ij}(t) = \min(k_t, k_{t+1})$(证明略).

由性质 3、性质 4 可知:在一步演化过程中,如果社群数不变,即 $k_t=k_{t+1}$,则矩阵 $\boldsymbol{S}(t)$保持为方阵;如果分化出更多的社群,即 $k_t<k_{t+1}$,则只需在 $\boldsymbol{P}(t)$中附加对应的全 0 列,并在 $\boldsymbol{S}(t)$中增加与该社群相应的全 0 行,使其保持为方阵;若某社群消失,即 $k_t>k_{t+1}$,则只需在 $\boldsymbol{P}(t+1)$中对应位置附加全 0 列,并在 $\boldsymbol{S}$ 中增加相应的全 0 列,使其保持为方阵.由于演化前后的社群数是已知的,因此总可以通过调整隶属矩阵,使 $\boldsymbol{S}$ 保持为方阵.

2 社群演化中的离群点

2.1 迁移矩阵的迭代优化

迁移矩阵的提出,为研究社群迁移的全局性质提供了有效的工具.由上述性质 2,当群体的演化行为完全同步时,其社群间的迁移量可以无误差地通过迁移矩阵反映出来,从而通过关系 $\boldsymbol{P}(t+1)\equiv\boldsymbol{P}(t)\boldsymbol{S}(t)$实现这种迁移模式的精确表达.但是,社会网络中的个体演化行为是复杂多样的,社群成员演化的同步也通常是渐进的,由于离群点的存在,通过迁移矩阵精确表达社群的演化模式无法实现.由于最小二乘估计对离群点的敏感性,由公式(2)定义的迁移矩阵拟合社群演化的精度将受到显著的影响.

为了降低离群个体对社群迁移模式的干扰,并显著地分化和识别社群演化离群点,采用稳健回归 *M*-估计方法对公式(2)中的目标函数加以优化.*M*-估计对异常数据引入的误差具有优越的稳健性,能够有效地排除异常值干扰.其基本思想是:采用迭代加权最小二乘估计回归系数,根据回归残差大小确定数据样本权重 w_i,从而抑制异常点对回归系数的贡献率,达到稳健估计的目的.稳健回归迭代优化采用的目标函数是:

$$\Phi(\tilde{S}) = \sum_{i=1}^{N} w_i \cdot \sum_{j=1}^{K(t)} \left[p_{ij}(t+1) - \sum_{l=1}^{K(t+1)} p_{il}(t)\tilde{s}_{lj}(t) \right]^2 \tag{4}$$

其中,w_i 为作用于数据行 i 的 Huber 权重因子,

$$w_i = \begin{cases} 1, & |u_i| \leqslant c_h \\ c_h / |u_i|, & |u_i| > c_h \end{cases} \tag{5}$$

其中,c_h的缺省值为 1.345;$u_i=e_i/\rho=0.6745e_i/med(|e_i-med(e_i)|)$为标准化的残差指标,$\rho$为残差尺度,$med(e_i)$为残差中位数.稳健回归后的迁移矩阵能够更精确地拟合社群的全局演化趋势,同时更凸显离群数据相对于社群总体趋势的差异.

M-估计通过迭代实现迁移矩阵的优化过程[16]:首先,建立最小二乘回归(公式(2))获得初始迁移矩阵,再据此确定各行残差按公式(5)赋予相应的权重,重新建立加权最小二乘回归计算新的迁移矩阵;通过反复迭代改进权重系数,直至残差和的改变小于给定阈值,或迭代达到指定次数.

容易发现,从迁移矩阵 $\boldsymbol{S}(t)$向稳健迁移矩阵 $\tilde{\boldsymbol{S}}(t)$ 的迭代优化过程是对隶属矩阵 $\boldsymbol{P}(t)$,$\boldsymbol{P}(t+1)$进行初等变换的过程,只需将 $\boldsymbol{P}(t)$,$\boldsymbol{P}(t+1)$左乘对角矩阵 $Diag(w_1,\ldots,w_N)$,并在加权隶属矩阵间求方程(3)的解,直到公式(4)中的 $\Phi(\tilde{S})$ 小于给定的阈值即可.我们称由公式(4)定义并经迭代优化获得的迁移矩阵 $\tilde{\boldsymbol{S}}(t)$ 为稳健迭代矩阵.容易证明,迭代优化后的加权迁移矩阵仍具备第 1.2 节所述的各项性质.

2.2 基于社群的演化离群点

从社群演化的角度检测离群点,为研究数据对象的动态行为开辟了新的视野.事实上,只有深入分析个体相对于群体的演化行为,才能深入理解社群演化规律和个体的离群模式.但是,恰恰由于个体的海量性和异构性,发现“散兵游勇”式的野点并无实际意义.例如前面所描述的始终处在网络社群边缘且随机游走的个体,其演化特征很少有规律可言,并大量存在于各类现实网络.因此,试图检测并解释这类离群点并无太多意义.

社群演化离群点研究的重点应着重于个体在社群中地位或影响力的显著变化.这类离群点可以归纳为以下 3 种:(1) 社群的(相对)核心个体演变为社群边缘个体;(2) 网络中的孤立个体演化为某个社群的(相对)核心;(3) 一个社群的核心演化为另一个社群的核心.

图 1 显示了这类个体与社群演化全局模式的差别.图 1 中,个体■逐渐从图 1(a)中的核心成员演化为图 1(d)中的边缘成员,个体●从一个社群的核心演化为另一个社群的核心,个体★从边缘状态演化为社群核心,而个体▲则始终在社群边缘随机游走.社群角色的显著变化是考察个体异常行为的关键.因此,前 3 类点可以视为有意义的离群点.而▲类个体是大量客观存在的随机游走个体,无须特别关注.本文的目标就是要刻画上述 3 类离群行为,并研究构造用于检测的方法.

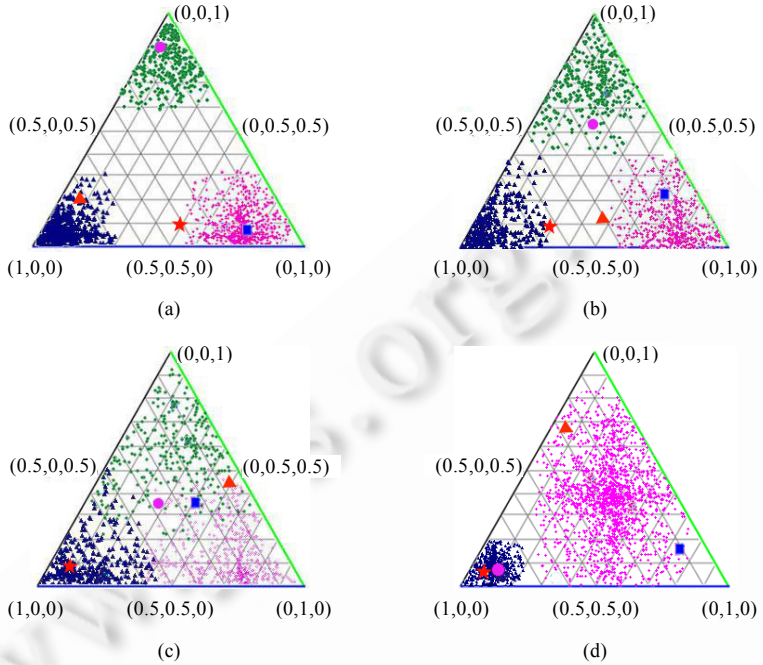


Fig.1 A illustration of the evolution difference between outlier data points and the members of communities

图 1 离群点相对于社群成员演化差异示意图

由隶属矩阵的定义,个体 V_i 在时刻 t 相对于社群的隶属度 $\boldsymbol{P}_i(t)=(p_{i1}(t),...,p_{iK_t}(t))$.显然,$V_i$ 对社群 j 的核心度越高,其隶属度向量的 j 分量越大.因此,我们可以定义 V_i 相对于社群 j 的核心度为

$$\mu_{V_i,j}(t)=(p_{ij}(t)-\min_{k=1:N}(p_{kj}(t)))/(\max_{k=1:N}(p_{kj}(t))-\min_{k=1:N}(p_{kj}(t))) \tag{6}$$

相应地,在时刻 t 到 t+1 的演化中,V_i 相对于社群 j 的核心度变化值可以表示为

$$\delta_{V_i,j}(t)=|\mu_{V_i,j}(t+1)-\mu_{V_i,j}(t)| \tag{7}$$

这里,$\delta_{V_i,j}(t)$ 表示 V_i 对社群 j 核心度的绝对变化量,即同时反映核心到边缘的双向变化.于是,个体在时刻 t 到 t+1 时相对于全体社群的核心度变化可以表示为

$$\delta_{V_i}(t)=\sum_{j\in k_t\cup k_{t+1}}\delta_{V_i,j}(t) \tag{8}$$

注意到,上式是对 t 和 t+1 时刻所有的社群上求和的,以涵盖同时存在社群新生和消失的情形.

定义 3. 给定$\varepsilon_1,\varepsilon_2>0$,称成员 V_i 为时刻 t 的 1-间隔离群点,如果

$$\Delta_{V_i}(t)=\delta_{v_i}(t)\cdot\sum_{j=1}^{K(t)}\left[p_{ij}(t+1)-\sum_{l=1}^{K(t+1)}p_{il}(t)\tilde{s}_{lj}(t)\right]^2>\varepsilon_1 \tag{9}$$

称成员 V_i 为全局演化离群点,如果

$$\Delta(V_i)=\sum_{t=1}^{T}\Delta_{V_i}(t)>\varepsilon_2 \tag{10}$$

定义 3 既考虑了成员在社群演化中与总体模式的差异,又兼顾了它在社群中的地位,可以有效抑制社群边

缘游走成员的干扰,有利于发现社群关键成员的异常行为.在实际应用中,可以定义任意两视图或任意时间区间内的离群点,还可以将参数化方式改变为发现离群点的个数,即在每个周期中或全局情况下发现离群点的个数,从而发现离群指标最大的前 k 个成员.

2.3 社群演化离群点检测的RoCTOD(robust community transition regression outlier detection)算法

根据以上讨论,社群演化离群点的发现可以归结为社群迁移矩阵的优化过程、成员离群度计算与离群点发现过程.该过程首先在相邻的两个数据视图中进行,从而求解出该时刻的稳健迁移矩阵和 1-步离群点.全局离群点发现则是 1-步离群点发现的叠加过程.当然,算法可以在任意的两个时间视图间或任意时间阶段上进行.本文方法与 Gupta 所提出的 1Sμ算法的显著区别在于以下两个方面:(1) 本文算法在通过稳健回归估计优化了对社群主体演化的拟合的同时,能够更为显著地体现离群个体的演化差异;(2) 本文重点关注社群核心成员演化过程中的离群行为,因此能够有效避免大量随机游走个体的干扰.

RoCTOD 算法.

输入:社群隶属度矩阵序列,迭代次数或误差限,离群度指标.

输出:社群不稳定成员列表,局部社群离群点,全局离群点.

(1) 求解矩阵方程(3),获得迁移矩阵 $\boldsymbol{S}$.

(2) while $\Phi(\boldsymbol{S})>\delta$ /* m-回归求解稳健迁移矩阵 */
依据公式(4)确定加权矩阵 $Diag(w_1,\ldots,w_N)$;
$\boldsymbol{P}(t)\leftarrow Diag(w_1,\ldots,w_N)\cdot\boldsymbol{P}(t)$;
$\boldsymbol{P}(t+1)\leftarrow Diag(w_1,\ldots,w_N)\cdot\boldsymbol{P}(t+1)$;
重新计算迁移矩阵 $\boldsymbol{S}$.

(3) 依据公式(7)、公式(8)计算成员的核心度变化值.

(4) 依据公式(9)输出 $\Delta_{V_i}(t)>\varepsilon$ 的成员作为 1-步离群点.

2.4 算法复杂度分析

RoCTOD 算法的复杂度与样本点个数 N 和社群数 k_t,k_{t+1} 有关,在社群数 k 保持不变的情形,其复杂度为 $O(m(N^2k^2+k^4)+Nk^2)$,其中,N^2k^2 为公式(2)中内积计算形成系数矩阵的复杂度,k^4 为求解迁移矩阵的复杂度,m 为求解稳健矩阵的迭代次数,Nk^2 为离群点检测的复杂度.由于网络中的社群数远少于网络成员数,因此,其总的时间复杂度为 $O(m(N^2k^2))$.当网络中的社群数较多时,$\boldsymbol{S}$ 将退化为稀疏矩阵,从而可以进一步优化稳健迁移矩阵的求解过程.

3 实验结果与讨论

第 1 组实验基于图 1 所示的人工数据集.实验首先在三维平面 $x+y+z=1$ 的第一卦限区域随机生成 3 个服从 Poisson 分布的类团,通过控制分布 Poisson 参数λ来控制类团中点的分布,从而形成模拟社群在演化中扩张、收缩、合并和分裂这 4 个演化模式的数据视图.根据数据集规模,按比例随机放置 4 类异常点(图 1 中的■,●,★,▲类异常点).由于点■和●的行为是对称的(即从边缘到核心或者相反),将其归为 Class-1 类;将从某视图中社群核心点转化为另一视图不同社群核心的点(点★)归为 Class-2 类;Class-3 类表示所有在社群边缘随机游走的数据点(点▲).对上述数据集,我们采用 AP 聚类算法对各视图进行聚类,从而生成各视图中数据点的隶属矩阵.

利用上述数据点的隶属矩阵关系,我们采用本文提出的算法对照 Gupta[12]提出的 1Sμ算法和 KNN 算法开展 3 种算法对 3 类离群点的发现能力比对实验.其中,1Sμ算法基于对隶属矩阵的迭代匹配和异常数据点的影响因子抑制,需要反复计算关于每个数据点的异常因子矩阵.由于 1Sμ算法没有考虑数据点在类团中的核心度变化因素,因此无法区分第 1 类、第 2 类和第 3 类离群点在社群重要度上的不同.对 1Sμ和 RoCTOD 算法采用统一迭代终止条件.

检测社群演化离群点的 KNN 算法的思想主要是:基于数据点在相邻视图间的演化量与其 K-近邻元素的平

均演化量之间的差异比较,从而发现全体数据中最显著区别于其 K-近邻元素的离群点,即计算数据点 p 自时刻 t 到 t+1 的演化离群指标:

$$\alpha_p(t)=\left\| p(t+1)-\frac{1}{K}\sum\nolimits_{q(t)\in KNN(p(t))} q(t+1)\right\| \tag{11}$$

和传统的基于最近邻关系的检测算法一样,KNN 对演化离群点的发现能力不仅受到计算效率的限制,而且 K-近邻均值化容易掩盖个体的演化差异度.

表 1 第 1 列表示各组实验的样本数据量,分别为 1 000,5 000 和 20 000.第 2 列为掺杂到原始数据中 3 类异常点相对于原始数据的比例.为了模拟真实应用场景,我们特别提高了 3 类随机游走数据的比重,并比较 3 种算法对该类异常点的检出能力.这组实验模拟了社群演化的扩张、收缩、合并和分裂这 4 种形态.我们采用 ROC (receiver operating characteristic)曲线下面积的 AUC 指标(area under ROC curve)[17]作为离群点检测精度的参考标准.表 1 为 NN,1Sμ与 RoCTOD 算法检测结果的 AUC 值,实验结果为 20 次同环境参数下计算结果的平均值.

Table 1 AUC values of the three algorithms to the community transition outlier detection

表 1 3 种算法检测社群迁移离群点的 AUC 值

样本点 N	异常点占比	社群扩展\收缩			社群合并			社群分裂		
		KNN	1Sμ	RoCTOD	KNN	1Sμ	RoCTOD	KNN	1Sμ	RoCTOD
1 000	Class1=2%	0.787	0.977	**0.986**	0.735	0.933	0.975	0.802	0.951	0.981
	Class2=1%	0.812	0.979	**0.989**	0.792	0.950	0.978	0.814	0.960	0.984
	Class3=5%	0.725	0.954	**0.013**	0.701	0.924	**0.095**	0.771	0.934	**0.051**
	Class1=8%	0.764	0.965	**0.976**	0.714	0.940	0.973	0.793	0.946	0.977
	Class2=4%	0.795	0.971	**0.985**	0.725	0.955	0.974	0.795	0.952	0.982
	Class3=10%	0.701	0.934	**0.025**	0.683	0.917	**0.101**	0.716	0.921	**0.063**
5 000	Class1=2%	0.756	0.964	**0.984**	0.736	0.941	0.977	0.786	0.952	0.983
	Class2=1%	0.803	0.971	**0.985**	0.803	0.953	0.985	0.801	0.959	0.984
	Class3=5%	0.695	0.945	**0.018**	0.694	0.926	**0.103**	0.715	0.921	**0.059**
	Class1=8%	0.743	0.959	**0.979**	0.724	0.953	0.976	0.782	0.950	0.980
	Class2=4%	0.789	0.959	**0.982**	0.758	0.960	0.981	0.788	0.954	0.983
	Class3=10%	0.657	0.931	**0.033**	0.637	0.932	**0.111**	0.697	0.918	**0.077**
20 000	Class1=2%	0.733	0.962	**0.981**	0.738	0.967	0.980	0.805	0.958	0.983
	Class2=1%	0.802	0.971	**0.984**	0.777	0.979	0.982	0.811	0.961	0.986
	Class3=5%	0.669	0.901	**0.027**	0.653	0.944	**0.093**	0.779	0.920	**0.078**
	Class1=8%	0.727	0.947	**0.974**	0.720	0.969	0.973	0.797	0.951	0.980
	Class2=4%	0.773	0.952	**0.980**	0.755	0.978	0.979	0.798	0.955	0.982
	Class3=10%	0.605	0.904	**0.048**	0.611	0.919	**0.115**	0.687	0.920	**0.089**

综合分析实验结果表明:本文算法对第 1 类、第 2 类异常点的检测能力优于 NN,1Sμ;同时,对第 3 类异常数据点(表中加框部分数据,即随机游走个体),算法能够有效抑制其对社群全局迁移模式的干扰,且绝大部分被检测算法视为噪声而忽略.在时间效率上,本文算法由于优化了社群成员间内积计算的过程,因此节约了大量的矩阵分量计算时间,其效率明显优于其他算法.而且算法收敛速度非常快.

第 2 个面向真实数据的实验对中国城市(地区)发展社会经济数据库[18]中 305 个主要城市 1996 年~2005 年份 GDP 中第一~第三产业占比数据集进行分析,其目标是提取各城市 3 类产业占比逐年演化中的共同模式,发现其中的显著差异.实验前,通过查阅资料对该数据集部分城市数据缺失、错误进行了修正.数据集每条记录包含各城市年度 GDP 三产比重分量.每个年度视图为 305 条三维数据,共 10 个视图.实验按产业门类对城市数据进行归一化并按产业占比手工进行分类标注,从而形成了按产业主导分类的 3 个社群.

实验首先求解每个年度视图向下一年度视图的稳健迁移矩阵,表 2 为 9 个迁移矩阵及经迭代 M-估计修正后的稳健迁移矩阵的累计残差平方和、正常数据最大残差值和离群点最大残差值.其中,稳健回归阈值统一设置为 0.001;MinΦ,MinΦ^*表示对全体 305 个记录项的残差绝对值之和.

表 2 说明,经 M-估计优化后的稳健迁移矩阵能够十分精确地拟合数据视图的总体演化趋势,并对异常数据项识别起到了放大作用.通过对各相邻年份演化离群点和全局离群点的分析发现,这些离群点的发生均符合对应城市相邻年份 GDP 构成的实际情况,即产业比重较之同类城市产生显著变化.例如:

(1) 桂林(1998-1999),离群指标为 0.388.该市因 1998 年地市合并,其三产也占比由 15.7%:42.1%:42.2%转化为 34.7%:30.5%:34.8%,三产比例显著改变.其原因在于桂林地区并入桂林市后,农业产业成分对产业结构的显著影响.

(2) 焦作(1997-2001),区间累计离群指标为 0.54.与全国其他城市工业强劲发展的趋势不同,焦作三产比重从 1997 年的 15%:62%:23%转变为 2001 年的 16%:51%:32%.其原因在于,作为衰老的矿区城市,该市资源开发利用呈逐年下降趋势,其矿业经济已失去其在城市经济中的主导地位.

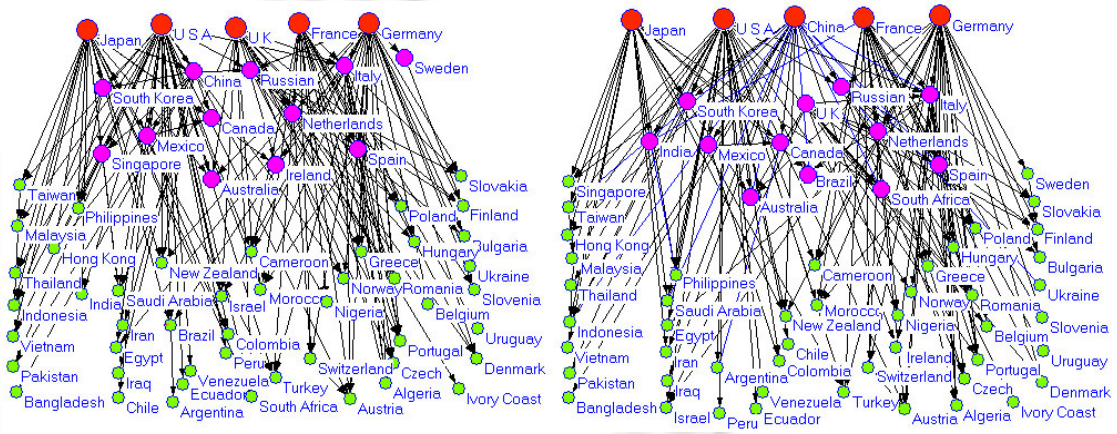
(3) 鄂尔多斯(1997-2005),10 年叠加离群指标最高,为 0.918.10 年间,该市从农业主导型转化为工业主导型城市.10 年中,多年 GDP 增长超过 20%,个别年份超过 40%.从最初以牧业为主 47%:22%:31%转变为 7%:53%:41%,创造了中国城市发展的奇迹.该实验结果表明,本文思想与算法能够在数据分析领域发挥有效作用.

Table 2 Outlier index and global outlier index of the top five cities of each year

表 2 各年度前 5 个城市的离群指数和全局离群指数

Year-to-year	1996-1997	1997-1998	1998-1999	1999-2000	2000-2001	2001-2002	2002-2003	2003-2004	2004-2005
Min Φ (for all 304 items)	2.997	2.624	3.166	2.365	1.88	2.19	3.16	3.26	5.29
Min Φ^* (for all 304 items)	0.458	0.396	0.544	0.280	0.269	0.360	0.375	0.401	0.439
Top most $\Delta_{V_i}(t)$	0.171 7	0.115 2	0.388 2	0.170 2	0.113 6	0.114 6	0.111 2	0.147 9	0.174 0
Outliers find	3	4	5	9	2	7	3	5	6

第 3 个实验采用 COW world Bilateral Trade 数据集[19].该数据集包含全球 197 个国家自 1870 年以来,对外进出口贸易的 791 491 条数据(其中,德国 2007-2009 年度数据缺失,自世界经合组织数据库[20]下载补充).选取其中 2000-2009 年的 11 万条数据,按国家统计其与他国的逐年进出口贸易额.先基于 Serrano 等人的研究[21]对该数据集进行社群关系分析,按贸易主导-依赖(dominance-dependence)关系形成社群隶属加权网络(如图 2 所示).



Year 2000 Year 2009

Fig.2 Dominance-Dependency change of foreign trade of the top 70 countries in the world

图 2 全球外贸排名前 70 名的国家对外贸易主导-依赖关系变化示意图

由图 2 可见,10 年间,国家间的主导-依赖关系发生了显著改变,中国已由 2000 年的半主导半依赖地位演化为 2009 年的主导地位,英国则正好相反.南非、印度和瑞士在国际贸易中的地位发生了转化.

依据这一社群关系,我们选取贸易总量排名前 70 名的国家,形成 10 个 70×70 的年度贸易关系隶属矩阵,求解相邻年度间的稳健迁移矩阵,并逐年检测离群点.表 3 列出了本文算法检出的各年度间基于稳健迁移矩阵的全局残差、最大离群度及对应的国家(地区).现结合原始数据,就离群度最高的国家举例说明如下:

(1) 多米尼加:与多个国家的贸易占当年度比例极不稳定.这个贸易严重依赖美国的南美国家在多个年份对美贸易出现剧烈震荡(2002-2009 年占比依次为 67.4%,73.6%,67%,63%,59.5%,60%,56%,61.5%).

其演化模式显然为外贸构成不稳定国家,原因在于中美洲自由贸易协定(CAFTA-DR)自始至终受多个国家抵制.

(2) 伊拉克:2007 年对美贸易占比 27.9%,而 2006 年占比 36.8%.此外,伊拉克对德国、台湾等主要贸易伙伴的贸易数据缺失(占比从大于 2%降为 0).同时,大幅度提高了与韩国、意大利等国家和地区的贸易比重,与多个国家新建立了贸易关系.这一演化模式成为该年度的最显著离群点.

(3) 中国:2003 年对外贸易出现显著增长,与多国贸易关系发生重大转变.如对美贸易由 2002 年占比 24.1%下降到 21%,对台贸易则由 2.3%增长到 6.5%(台湾对大陆贸易则由 9.2%跃升到 25.9%),而对原占比 0.5%以下的多个国家出现较大幅度的增长.该模式与其他国家显著不同.因此,台湾和中国大陆在 2002-2003 年成为最显著离群点.此外,中国在 10 年演进中的累计离群度达到 0.069,贸易地位指标发生显著变化,可以认为属全局离群点.

Table 3 Robust transition matrix fitting global residuals and yearly outliers of world trade relations

表 3 2000-2009 年度世界贸易关系稳健迁移矩阵拟合全局残差和各年度离群点

年度	2000-2001	2001-2002	2002-2003	2003-2004	2004-2005	2005-2006	2006-2007	2007-2008	2008-2009
全局残差	0.06	0.049	0.019	0.008	0.008	0.006	0.016	0.008	0.013
最大离群度	0.009	0.003	0.002 5	0.000 6	0.001 0	0.001	0.002	0.001	0.001 8
对应的国家	南非	加拿大	中国大陆	多米尼加	多米尼加	韩国	多米尼加	伊拉克	以色列

4 结 论

从社群演化观点分析研究各类关系网络参与者的演化模式,是揭示事物运动规律的有效工具,并成为社会网络分析的重要发展方向.网络社群演化中的离群点发现则着重于揭示网络个体变化的异常特征,能够发现新颖而富有实际意义的演化特征.社群演化分析方法较之传统基于静态数据集的离群点检测方法更符合多元、不断变化的社会网络分析的需求,但只有有效排除网络中大量随机游走边缘个体的干扰,才能聚焦重要的离群目标,把握演化离群点分析的正确方向.

本文首先采用稳健估计方法求解社群演化的稳健迁移矩阵,以更为精确地拟合社群的全局演化趋势.在此基础上,提出面向社群核心成员的演化离群点描述方法和检测算法.实验结果表明,该方法对重要的演化离群点具有敏锐的识别能力.

References:

[1] Girvan M, Newman MEJ. Community structure in social and biological networks. In: Shepp LA, ed. Proc. of the National Academy of Sciences. Washington: National Academies Press, 2002. 7821−7826. [doi: 10.1073/pnas.122653799]

[2] Knorr E, Ng R. Algorithms for mining distance-based outliers in large datasets. In: Ashish G, ed. Proc. of the 24th Conf. on VLDB. New York: Morgan Kaufmann Publishers, 1998. 392−403. [doi: 10.1073/pnas.122653799]

[3] Hawkins D. Identification of Outliers. London: Chapman and Hall, 1980. 1−45.

[4] Breunig MM, Kreigel HP, Ng RT, Sander J. LOF: Identifying density-based local outliers. In: Chen WD, Jeffrey F, Philip A, eds. Proc. of the 2000 ACM SIGMOD Conf. New York: ACM Press, 2000. 93−104. [doi: 10.1145/342009.335388]

[5] Breunig MM, Kreigel HP, Ng RT, Sander J. Optics-Of: Identifying local outliers. In: Jan M, Żytkow, Jan R, eds. Proc. of the PKDD'99. LNAI 1704, London: Springer-Verlag, 1999. 262−270. [doi: 10.1007/978-3-540-48247-5_28]

[6] Yu D, Sheikholeslami G, Zhang A. Findout: Finding outliers in very large datasets. Knowledge and Information Systems, 2002, 4(4):387−412. [doi: 10.1007/s101150200013]

[7] Bay SD, Schwabacher M. Mining distance based outliers in near linear time with randomization and a simple pruning rule. In: Mohammed JZ, Wang T, Hannu T, eds. Proc. of the SIGKDD 2003. New York: ACM Press, 2003. 29−38. [doi: 10.1007/s101150200013]

[8] Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Surveys, 2009,41(3):1−72. [doi: 10.1145/1541880.1541882]

[9] Ge Y, Xiong H, Zhou ZH, Ozdemir H, Yu J, Lee K. Top-Eye: Top-*K* evolving trajectory outlier detection. In: Huang J, ed. Proc. of the 19th ACM Conf. on Information and Knowledge Management. New York: ACM Press, 2010. 1733−1736. [doi: 10.1145/1871437.1871716]

[10] Gao J, Liang F, Fan W, Wang C, Sun Y, Han J. On community outliers and their efficient detection in information networks. In: Rao B, ed. Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2010. 813−822. [doi: 10.1145/1835804.1835907]

[11] Gupta M, Gao J, Sun Y, Jiawei H. Community trend outlier detection using soft temporal pattern mining. In: Flach PA, ed. Proc. of the 2012 European Conf. on Machine Learning and Principles and Knowledge Discovery in Databases. Berlin: Springer-Verlag, 2012. 692−708. [doi: 10.1007/978-3-642-33486-3_44]

[12] Gupta M, Gao J, Sun Y, Jiawei H. Integrating community matching and outlier detection for mining evolutionary community outliers. In: Yang Q, ed. Proc. of the 18th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, 2012. 859−867. [doi: 10.1145/2339530.2339667]

[13] Sun Y, Tang J, Han J, Gupta M, Zhao B. Community evolution detection in dynamic heterogeneous information networks. In: Brefeld U, ed. Proc. of the 8th Workshop on Mining and Learning with Graphs. New York: ACM Press, 2010. 137−146. [doi: 10.1145/1830252.1830270]

[14] Greene D, Doyle D, Cunningham P. Tracking the evolution of communities in dynamic social networks. In: Nasrullah M, Reda A, eds. Advances in Social Networks Analysis and Mining, 2010 Int'l Conf. on ASONAM. Washington: IEEE Computer Society, 2010. 176−183. [doi: 10.1109/ASONAM.2010.17]

[15] Cheng XQ, Shen HW. Community structure of complex networks. Complex Systems and Complexity Science,2011,8(1):57−70 (in Chinese with English abstract). [doi: 10.3969/j.issn.1672-3813.2011.01.007]

[16] Huber PJ, Ronchetti EM. Robust Statistics. 2nd ed., Hoboken: Wiley, 2009. 149−195. [doi: 10.1002/0471725250.refs]

[17] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006,27(8):861−874. [doi: 10.1016/j.patrec.2005.10.010]

[18] http://www.geodata.cn/Portal/mainbody/naturalRes/index.jsp

[19] http://www.correlatesofwar.org/

[20] http://stats.oecd.org/

[21] Serrano MÁ, Boguñá M, Vespignani A. Patterns of dominant flows in the world trade Web. Journal of Economic Interaction and Coordination, 2007,2(2):111−124. [doi: 10.1007/s11403-007-0026-y]

附中文参考文献:

[15] 程学旗,沈华伟.复杂网络的社区结构.复杂系统与复杂性科学,2011,8(1):57−70. [doi: 10.3969/j.issn.1672-3813.2011.01.007]

胡云(1978－),女,江苏连云港人,博士生,副教授,CCF 会员,主要研究领域为数据挖掘,社会网络分析.
E-mail: 15250998131@139.com

谢俊元(1961－),男,教授,博士生导师,主要研究领域为分布式人工智能.
E-mail: jyxie@nju.edu.cn

王崇骏(1975－),男,博士,教授,CCF 高级会员,主要研究领域为智能信息处理.
E-mail: chjwang@nju.edu.cn

周作建(1979－),男,博士生,主要研究领域为大数据分析及应用.
E-mail: lygzzj@163.com

吴骏(1981－),男,博士,副教授,主要研究领域为多 Agent 联盟.
E-mail: iip@nju.edu.cn