网络参与者见面概率的迭代估计方法

李永立[1], 樊宁远[1], 林亿民[1], 吴 冲[2]

(1. 东北大学工商管理学院, 辽宁 沈阳 110169;
2. 哈尔滨工业大学管理学院, 黑龙江 哈尔滨 150001)

摘要: 针对网络参与者见面概率的估计问题, 引入效用分析的方法, 并将效用函数的估计问题与logistic 回归分析相联系. 在深入分析网络参与者链接形成过程的基础上, 提出了迭代估计算法. 进一步给出了该方法的基本算例, 并在合著者网络中进行了方法的应用研究. 结果表明该方法能够实现对网络参与者见面概率进行估算的研究目标, 有助于优化网络的整体产出, 可行并实用.

关键词: 见面概率; 迭代估计方法; 合著网络; 社会网络分析
中图分类号: TP273 **文献标识码**: A **文章编号**: 1000−5781(2018)02−0167−08

doi: 10.13383/j.cnki.jse.2018.02.003

Iterative method of estimating the meeting probabilities of network individuals

Li Yongli[1], Fan Ningyuan[1], Lin Yimin[1], Wu Chong[2]

(1. School of Business Administration, Northeastern University, Shenyang 110169, China;
2. School of Management, Harbin Institute of Technology, Harbin 150001, China)

Abstract: This paper introduces connects the estimation of utility analysis and logistic regression analysis to estimate the meeting probability within network individuals. After a thorough analysis of the forming process of the linkage, this paper proposes an iterative algorithm to cope with the estimation problem. The method is also demonstrated with an example and an application for a coauthor network. The results show that the method can correctly estimate the meeting probabilities of network individuals and therefore uplift the monolithic network output, as proves our method is both feasible and practical.

Key words: meeting probability; iterative algorithm; coauthor network; social network analysis

1 引 言

社会网络构成了社会经济生活的骨架, 特别是在社会媒体平台上, 用户之间构成的人际网络成为了信息传播的主要渠道[1]; 无独有偶, 在科学计量学领域, 论文的作者构成合著网络, 成为完成一项科研工作重要的协作模式[2], 以上列举的两个网络都是社会网络在生产生活中的具体体现. 本文关注于社会网络参与者见面的概率问题, 该问题来源于对“有份无缘”现象的思考; 也即: 如果两个人成为朋友对于双方都有益, 但如果双方没有见面的机会, 朋友关系也不能建立. 事实上, 在社会媒体平台上(如: 微信、微博、Facebook 等)以

收稿日期: 2015−12−31; 修订日期: 2017−2−28.
基金项目: 国家自然科学基金资助项目(71501034; 71771041); 中国博士后基金资助项目(2016M590230; 2017T100183); 中央高校基本科研业务费资助项目(N160604001); 辽宁省教育科学“十三五”规划课题资助项目(JG16DB124).

及合作者的选择方面, “有份无缘”的现象比比皆是, 这不利于提高用户的交友体验和增加科研的产出. 本文旨在通过对网络参与者见面概率的估计, 找出“有份”但“ 无缘”的参与者, 提高推荐的针对性, 最终达到优化网络产出的目标[3]. 本文的研究意义主要体现在两个方面: 宏观上, 本文有助于社会媒体平台推荐合意的朋友, 提高社会媒体平台用户对于平台的满意度和忠诚度, 实现平台的盈利增长; 微观上, 个体依托于本文方法的推荐, 有助于找到心仪的朋友或理想的合作者, 实现个体利益的增加.

社会网络中链接的建立, 反映了参与个体的选择偏好. 而效用分析可以解释这种偏好. 在理性人假设下, 参与者倾向于选择使其效用水平最大的个体建立链接. 为了刻画个体间建立链接的福利水平, 本文将引入效用分析的方法, 通过考察建立链接后效用水平的变化对网络参与个体的福利水平进行定量分析. 这与既有的应用效用函数分析个体行为及文献评价的研究相一致, 比如: Wong 等[4], Poelmans 等[5], 以及俞立平等[6]等. 由于参与者的见面过程不易被观察, 既有的文献通常都将网络参与者见面的过程视为链路形成的一个中间环节, 仅仅视为一类“黑箱问题”加以分析[7−9]. 为了估计出各个参与者见面的概率, 找出“ 有份”但“无缘”的参与者, 本文将可观察到的网络链接视为“显状态”, 将不能观察到的见面过程视为“隐状态”, 拟通过逐步迭代的方法, 估计出“隐状态”中网络参与者见面的概率.

既有的社会网络领域的研究往往关注于网络的结构形成[10], 社团结构[11]和节点相似性评价[12]等问题的研究, 而关注网络参与者见面概率的研究较少. 由于网络参与者的见面过程不是可以直接观察的, 对其估计的研究具有理论难度, 是值得探索的研究方向. 由此, 本文将关注于网络参与者见面概率的估计问题, 将这一问题进行定量化和模型化. 本文从建立网络参与者的效用函数入手, 通过对节点建立链接前后效用变化的度量, 设计迭代算法用以推断网络参与者见面的概率. 在效用函数的设计中, 本文将网络的结构效应引入效用函数, 突出了研究的网络背景; 在对节点效用变化进行度量的分析中, 将 logistic 回归分析的思想引入对节点建立链接概率的估计, 适应因变量为 0-1 变量的情形; 并在迭代算法的提出中, 引入随机样本, 有效处理样本可能存在的不平衡问题. 基于一个数值算例的阐述和在真实数据集上的应用, 本文展示方法的合理性和实用性, 有助于深入解释网络链路形成的原因, 并给出有针对性地推荐.

2 网络参与者见面过程的效用分析模型

2.1 问题的提出

在社会网络中, 通常能够收集到节点的属性数据和节点间链接的情况, 但是很难直接观察到节点的见面过程, 这一情形表述在图 1 中.

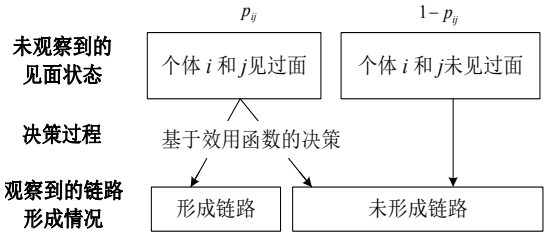

图 1 个体 i 和 j 链路形成的示意图

Fig. 1 Diagram of link formation between individual i and j

如图 1 所示, 如果观察到两个节点建立了链接, 那么他们之间必然见过面, 因此问题的重点是研究那些没有形成链接的“节点对”, 其中的两个节点可能见过面, 也可能没见过面. 以两个代表性个体 i 和 j 链路形成的过程为例, 如果两者没有见面的机会, 他们一定不能形成链路; 如果两者见过面, 则他们会基于各自的效用做出决策, 也即与对方形成链接是否有助于自身效用的增加: 如果是, 则形成链路; 如果不是, 则不形成链路. 在此过程中, 涉及到“见面的状态”, “ 决策的过程”和“ 可以观察到的链路形成情况”三个组成部分.

进一步, 记图 1 中个体 i 和 j 见面的概率为 p_{ij}, 即为本文求解的目标变量; 记 l_{ij} 为个体 i 和 j 链路

形成的情况, 则有 $l_{ij}=0$ 表示两者间没有形成链接, 若两者形成了链接, 则有 $l_{ij}=1$. 根据图 1 所示, 当 $l_{ij}=1$ 时, 必有 $p_{ij}=1$, 因为形成了链接, 必然两者见过面; 特别地, 对于 $l_{ij}=0$ 的"节点对", 在效用分析的框架下, 如果根据效用分析的结果, 两个个体应该以很大的概率值形成链接, 如果实际上两者并没有形成链接, 则推断两者没有见过面的概率很高, 这也是本文建模的基本出发点.

2.2　链接形成的效用分析模型

本节详细阐述图 1 中效用分析的过程. 以代表性个体 i 为例, 其效用函数为

$$U_i(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta}), \tag{1}$$

其中被观察到的网络的邻接矩阵记为 $\boldsymbol{L}$, 用以反映网络个体的链接情况; 网络个体的属性矩阵记为 $\boldsymbol{A}$, 用以反映可以获得的网络参与者的属性信息, 其中第 s 个个体的属性向量记为 $\boldsymbol{A}^s$; $\boldsymbol{\theta}$ 为待估计的参数向量.

对于图 1 中的链路形成过程, 倘若代表性个体 i 和 j 获得见面机会, 则个体 i 建立与个体 j 的链接时, 面临的效用变化为

$$\Delta U_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=U_i(\boldsymbol{L}_{+[l_{ij}=1]},\boldsymbol{A};\boldsymbol{\theta})-U_i(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta}), \tag{2}$$

其中 $\boldsymbol{L}_{+[l_{ij}=1]}$表示在原来网络链接的情形下, 增加个体 i 和个体 j 之间的链接.

根据式(2), 如果个体 i 的效用增加, 即 $\Delta U_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})>0$ 时, 个体 i 有向个体 j 建立链接的愿望. 同理, 对于个体 j 而言, 其面临的效用变化为

$$\Delta U_{j\to i}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=U_j(\boldsymbol{L}_{+[l_{ij}=1]},\boldsymbol{A};\boldsymbol{\theta})-U_j(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta}). \tag{3}$$

同样, 当建立链接有助于个体 j 的效用增加时, 即 $\Delta U_{j\to i}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})>0$ 时, 个体 j 有向个体 i 建立链接的愿望. 并且当两者都有建立链接的愿望时, 代表性个体 i 和 j 才能形成链接.

特别地, 作为有限理性的网络参与者, 其对于自己效用的感知可能存在偏差及其它不可观测的随机干扰因素, 为此令

$$\Delta\tilde{U}_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=\Delta U_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})+\varepsilon_i, \tag{4}$$

$$\Delta\tilde{U}_{j\to i}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=\Delta U_{j\to i}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})+\varepsilon_j, \tag{5}$$

其中 ε_i 和 ε_j 分别是代表性个体 i 和 j 对于效用感知的偏差.

根据 Coles 等[13]的研究, 这类来自于不完全理性人的感知偏差往往满足 Type I 的极值分布. 从建模的角度说, 引入随机干扰因素让模型更加接近决策的实际. 与此同时, 从方法的角度说, 引入随机干扰因素使得模型概率分析成为可能, 这类似于线性回归模型中对于随机干扰因素的引入. 事实上, 由于链路是否存在恰恰是 0-1 变量的形式, 这与 logistic 回归因变量为分类变量相一致, 这为在给出效用函数的具体形式下, 推断网络参与者见面概率提供了模型基础.

2.3　具体化的效用函数及回归方程

考虑到本文的模型将应用于合著网络中对合意的合作者进行挖掘, 以下根据这一应用背景将具体化式(1)的效用函数. 根据文献[14]的观点, "三度影响力是影响网络参与者行为的主要因素", 即来自朋友(一度影响力), 朋友的朋友(二度影响力), 乃至朋友的朋友的朋友(三度影响力)都会对行为产生影响, 并且发现这些影响力是逐渐递减的关系. 由此, 本文在构建具体的效用函数时, 考虑其中影响力较强的前两度影响力, 进而具体化的效用函数为

$$U_i(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=\mathrm{c}\sum_{k=1}^{n}l_{ik}+\theta_1\sum_{k=1}^{n}l_{ik}\|\boldsymbol{A}^i-\boldsymbol{A}^k\|+\theta_2\sum_{k=1}^{n}l_{ik}\sum_{s=1}^{n}l_{ks}\|\boldsymbol{A}^i-\boldsymbol{A}^s\|, \tag{6}$$

其中 c 是常数, n 是网络中个体的总数, $\|\cdot\|$ 用以度量属性之间的差异; 具体地, 作者的属性用作者文章的关键词及其频数来度量.

以式(6)给出的具体形式的效用函数为基础, 根据式(2)和式(3)的关系, 进一步推得个体 i 和个体 j 在建立链接的过程中, 感知到的效用变化分别为

$$\Delta\tilde{U}_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=\mathrm{c}+\theta_1\|\boldsymbol{A}^i-\boldsymbol{A}^j\|+\theta_2\sum_{s=1}^{n}l_{js}\|\boldsymbol{A}^i-\boldsymbol{A}^s\|+\varepsilon_i, \tag{7}$$

$$\Delta\tilde{U}_{j\to i}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=\mathrm{c}+\theta_1\|\boldsymbol{A}^i-\boldsymbol{A}^j\|+\theta_2\sum_{s=1}^{n}l_{is}\|\boldsymbol{A}^j-\boldsymbol{A}^s\|+\varepsilon_j. \tag{8}$$

事实上, 很难直接获得以上两式左端效用变化的具体数值; 由此, 对于式(7)或式(8)中的参数将不能通过以上两式直接进行估计. 但是, 通过对网络链接的观察, 可以获得 $\Delta\tilde{U}_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})$ 和 $\Delta\tilde{U}_{j\to i}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})$ 的符号; 这是因为若个体 i 和个体 j 形成了链接, 由 2.2 节可知, $\Delta\tilde{U}_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})>0$ 与 $\Delta\tilde{U}_{j\to i}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})>0$ 同时成立. 这一问题的结构与 logistic 回归问题的结构较为相似, 即因变量为符号变量, 进一步结合前面对随机变量 ε_i 的分布假设,有

$$\Pr(\Delta\tilde{U}_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})>0)=\frac{\exp(\Delta U_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta}))}{1+\exp(\Delta U_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta}))}. \tag{9}$$

式(9)等价于

$$\ln\left(\frac{\Pr(\Delta\tilde{U}_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})>0)}{1-\Pr(\Delta\tilde{U}_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})>0)}\right)=\Delta U_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta}). \tag{10}$$

式(10)左边即为 $\Delta\tilde{U}_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})$ 的 Logit 变换, 进而结合 $\Delta U_{i\to j}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})$的具体形式, 可得

$$\mathrm{logit}(\Delta\tilde{U}_{i\to j}>0)=\mathrm{c}+\theta_1\|\boldsymbol{A}^i-\boldsymbol{A}^j\|+\theta_2\sum_{s=1}^{n}l_{js}\|\boldsymbol{A}^i-\boldsymbol{A}^s\|+\varepsilon_i, \tag{11}$$

类似地, 针对个体 j 到个体 i 的链接形成过程, 得到

$$\mathrm{logit}(\Delta\tilde{U}_{j\to i}>0)=\mathrm{c}+\theta_1\|\boldsymbol{A}^i-\boldsymbol{A}^j\|+\theta_2\sum_{s=1}^{n}l_{is}\|\boldsymbol{A}^j-\boldsymbol{A}^s\|+\varepsilon_j. \tag{12}$$

式(11)和式(12)恰好为 logistic 回归的基本方程, 可以应用 logistic 回归的最大似然估计技术对效用函数中的参数进行估计. 考虑到本文所分析问题的结构特征, 关键在于如何合理的估计该参数向量, 并以此为基础推断网络参与者见面的概率.

3 算法实现及算例

3.1 算法步骤

将模型中的参数通过实际数据的校准是进一步进行见面概率估计的基础; 但是, 直接应用观察到的全部数据对式(11)和式(12)进行参数估计将产生有偏的估计结果, 这是因为 logistic 回归针对那些见过面的样本, 参数估计才有效. 具体地说, 等式(11)和式(12)左边对应的是可以观察到的链接情况, 表现为 0-1 变量, 但是 0-1 变量的值为 0 的原因可能是由于没有机会见面而造成的, 这个因素是不能被效用分析所解释的. 由此, 本文提出逐步迭代的方法解决该问题. 具体地, 该方法分为如下步骤:

步骤 1 对于被观察到的已经建立链接的样本, 同时添加原样本数一半数量的因变量为 0 的随机样本, 构成一个新的样本集(随机样本的含义是指自变量随机生成, 为取值范围内的均匀分布), 应用式(11) 和式(12) 进行 logistic 回归分析, 得到参数的初始估计值$(\hat{\mathrm{c}},\hat{\theta}_1,\hat{\theta}_2)$;

步骤 2 应用这一初始估计值, 对那些没有建立链接的样本进行估计, 得出其经过决策过程应该建立链接的概率, 记为 $\hat{p}_{ij}^{l}$;

步骤 3 将建立链接概率小于 50 % 的样本纳入估计样本, 这时全体样本由已建立链接的样本和建立链

接概率小于 50 % 的样本构成, 再应用式(11)和式(12)进行 logistic 回归分析, 得到更新后的参数估计值;

步骤 4 重复步骤 2, 步骤 3, 直到没有新的链接概率低于 50 % 的样本出现; 进行回归方程的参数估计;

步骤 5 对那些没有建立链接的“节点对”, 应用步骤 4 得到的参数计算 $\hat{p}_{ij}^{l}$, 用 $1-\hat{p}_{ij}^{l}$ 度量其见面的概率.

在步骤 1 中, 考虑到如果只用那些已经建立链接的样本进行分析, 因为这些样本的因变量都为 1, 会存在样本分布的不平衡问题, 导致第一步得到的结果存在严重偏差, 于是引入随机样本, 使得算法得到的初始估计值偏差小一些, 有助于减少迭代次数, 提高算法效率; 进一步, 在步骤 2 中试图增加估计的样本使得参数的估计值更接近于无偏估计, 于是用步骤 1 得到的参数估计结果, 计算那些没有建立链接的样本潜在的链接概率, 根据模型假设和逻辑分析, 如果经过决策过程其潜在建立链接的概率很低, 说明即便见面, 也不会形成链接. 步骤 3 中将概率低于 50 % 的“ 节点对”视为曾见过面的样本纳入参数估计; 步骤 4 重复以上的估计过程, 直到各类样本不再变化为止, 由此得到了一个较好的参数估计结果; 在步骤 5 中, 应用最终得到的参数估计结果, 计算那些未形成链接的“ 节点对”经历决策过程形成链接的概率, 注意到该值越大, 对应的“ 节点对”没有见面的概率就越大, 于是用 $1-\hat{p}_{ij}^{l}$ 度量两者见面的概率, 并进行排序.

3.2 数值算例

考虑有五个节点的网络如图 2 所示, 其中每个点的属性值标记在各个点的旁边.

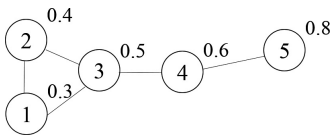

图 2　算例中包含 5 个节点的网络示意图及其属性值

Fig. 2　Network diagram and attribute values of the five nodes in the example

根据前面的算法步骤, 则每一步的结果如下.

在步骤 1 中, 根据图 2 的链接情况, 选择“节点对”(1– 2), (1– 3), (2– 3), (3– 4)和(4– 5)为样本, 同时添加 5 个因变量为 0 的随机样本. 利用式(11)和绝对值度量节点之间的属性差异进行 logistic 回归估计, 结果如表 1 所示. 根据表 1 的结果, 可以发现其参数的回归估计值在 0.10 置信度下通过统计检验.

表 1　参数的估计值及统计量

Table 1　Estimators and statistics of the parameters

参数	估计值	Wald统计量	显著性
c	8.466	5.220	0.022
θ_1	−20.399	4.072	0.044
θ_2	−8.410	5.480	0.019

根据步骤 2, 应用表 1 可以计算出在决策过程中图 2 中没有形成链接的“节点对”建立链接的概率, 结果如表 2 所示.

表 2　未形成链接的“节点对”应建立链接的概率

Table 2　The probability of forming a link between unlinked pairs

“节点对”$(i-j)$	$i\to j$ 概率值	$j\to i$ 概率值	概率值
1– 4	0.028 0	0.455 9	0.013 1
1– 5	0.014 0	0.009 2	0.000 1
2– 4	0.907 5	0.937 3	0.850 6
2– 5	0.201 7	0.044 9	0.009 1
3– 5	0.818 4	0.323 6	0.264 8

以“节点对”(1– 4)为例, 根据表 1 的参数估计值, 当建立从节点 1 到 4 的链接时, 节点 1 效用变化的方程为 $\Delta\hat{U}_{1\to4}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=8.466-20.399\|\boldsymbol{A}^1-\boldsymbol{A}^4\|-8.410\sum_{s=1}^{5}l_{4s}\|\boldsymbol{A}^1-\boldsymbol{A}^s\|$.

进一步, 算例中的节点差异$\|\cdot\|$定义为绝对值, 进而算得 $\Delta\hat{U}_{1\to4}(\boldsymbol{L},\boldsymbol{A};\boldsymbol{\theta})=-3.541$. 由此, 根据式(9)得到节点 1 到节点 4 建立有向链接的概率为 0.028 0, 类似地得到节点 4 到节点 1 建立有向链接的概率为 0.455 9, 其两者的乘积即为节点 1 和 4 之间建立链接的概率. 同理可得表 2 中其它节点间的见面概率.

结合表 2 的结果, 根据步骤 3 算出有四组“节点对”形成链接的概率小于 0.5, 将这四组“节点对”(1– 4), (1– 5), (2– 5)和(3– 5)纳入估计样本, 修正参数估计的结果, 得到 logistic 回归方程的参数估计结果见表 3.

表 3 参数的估计值及统计量

Table 3 Estimators and statistics of the parameters

参数	估计值	Wald 统计量	显著性
c	12.714	15.778	0.000
θ_1	−45.722	3.673	0.050
θ_2	−6.541	3.041	0.081

进一步进行迭代, 发现算法在该步收敛, 即用于训练的样本不再变化, 由表 3 的参数估计结果算得网络上未形成链接的参与者见面的概率值及其排序, 如表 4 所示. 由此, 根据表 4 的结果, 将推荐(2– 4) 这一“节点对”见面, 其见面将会建立相应的链接, 有助于彼此效用的增加和网络整体效用产出的增加.

表 4 未形成链接的“节点对”见面的概率值及其排序

Table 4 The meeting probability between the unlinked pairs and ranking

“节点对”$(i-j)$	见面的概率值	逆排序
1– 4	0.995 7	3
1– 5	1.000 0	4
2– 4	0.208 7	1
2– 5	1.000 0	4
3– 5	0.994 8	2

根据以上的步骤和算例可以发现, 本文提出的逐步迭代方法, 通过对效用函数的参数估计, 逐步推断不能直接被观察到的网络个体见面的概率. 而在逐步迭代的每一步, 应用的是 logistic 估计, 所以从参数估计的角度说, 该方法在估计的有效性上, 继承了 logistic 估计的特性, 这是本文参数估计在理论上的依据. 通过逐步迭代的过程, 本文将网络中可能见面的个体逐步加入到估计样本中, 逐步修正估计的参数, 进而修正“节点对”间的见面概率, 不断更新用于分析的样本, 直到样本不再变化, 迭代过程停止.

4 方法应用

本文选取发表在 www.arxiv.org 上从 1995–01–01～2003–06–30 之间“高分子”物理领域的文献, 该数据集属于一个开放的数据集, 已被 Newman[15]用于研究社会网络的统计特性、社团特征以及链路预测等方面. 获取该数据的网络地址为 www-personal.umich.edu/mejn/netdata/, 整个数据集包含了 31 163 名作者和他们发表的 120 029 篇论文, 他们合著网络的度分布如图 3 所示. 在本文计算出的合著网络中, 边表示相连的两位作者发生过合作关系, 而没有考虑合作的次数. 为了突出研究的主体, 本文关注节点度大于等于 100 的作者(意味着作者在统计区间内发表的论文中有不少于 100 名的合作者), 满足此条件的作者共计 50 名, 其度分布由图 3 中的小图所示, 以及链接情况如图 4 所示.

进一步, 根据 2.3 节的论述, 本应用中网络参与者个体的属性取其发表文章的关键词及关键词的频率. 具体地, 以代表性个体 i 和 j 为例, 两者的关键词及其频数和频率分别由表 5 和表 6 所示.

表 5 和表 6 中的粗体字是个体重复的关键词, 将这些重复的关键词构成一个集合并按顺序编号, 不失一般性地, 令该集合有 m 个元素; 进一步回顾式(8), 属性差异具体定义为

$$\left\|\boldsymbol{A}^i-\boldsymbol{A}^j\right\|=1-0.5\sum_{r=1}^{m}\left(f_i(r)+f_j(r)\right), \tag{13}$$

其中函数 $f_i(r)$ 表示集合中第 r 个重复的关键词对于作者 i 的频率, 类似地可得 $f_j(r)$的含义.

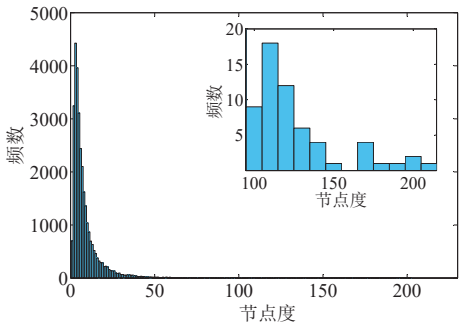

图 3　全部样本及遴选样本的节点度分布

Fig. 3　Node degree distribution of the whole and the selected samples

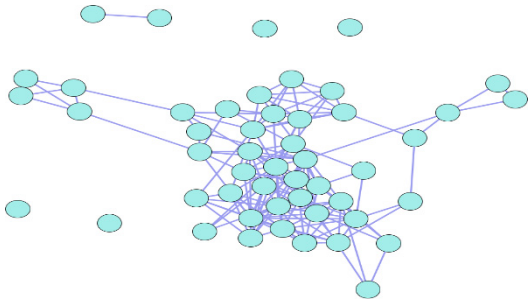

图 4　遴选样本合著者网络的图示

Fig. 4　Co-author network diagram of the selected samples

表 5　个体 i 发表文章的关键词及其频数和频率

Table 5　Keywords in the papers of of individual i as well as keywords' frequency and rate

关键词	频数	频率
Semiconductors	5	0.33
Gas diffusion	4	0.27
Scattering	3	0.20
Statistical Physics	3	0.20

表 6　个体 j 发表文章的关键词及其频数和频率

Table 6　Keywords in the papers of of individual j as well as keywords' frequency and rate

关键词	频数	频率
Susceptibility	4	0.31
Semiconductors	3	0.23
Gas diffusion	3	0.23
Scattering	2	0.15
Phase evolution	1	0.08

注: 表 5 和表 6 中粗体字是个体重复的关键词

定义式(13)的一个特性在于当两者重复的关键词多时, 两者的属性差异小. 以给出的数据为例, 经式(13)可以算得个体 i 和 j 的属性差异为 0.30.

进一步, 根据本文 3.1 节的步骤, 经四步后各个状态的样本不再变化, 求得最终参数的统计信息以及网络参与者见面概率的前 5 组"节点对", 分别由表 7 和表 8 所示. 本文的计算主要在 IBM SPSS Statistics 22 上进行.

表 7　参数的估计值及统计量

Table 7　Estimators and Statistics of the parameters

参数	估计值	Wald统计量	概率值
c	0.502	37.408	0.000
θ_1	−5.464	49.285	0.000
θ_2	−0.288	32.820	0.000

表 8　集中网络参与者见面概率的前 5 对

Table 8　The top 5 pairs according to meeting probability

作者"节点对"	见面概率值
MAREL, D.V.D.- AEPPLI, G.	0.078
FUJIMORI, A.- CHEONG, S.W.	0.085
UCHIDA, S.- TAKAGI, H.	0.139
KIM, H.J.- CAVA, R.J.	0.210
MAREL, D.V.D.- EISAKI, H.	0.353

表 8 中列出了最可能由于没有见面机会而没有形成链接的作者"节点对". 通过效用分析的结果, 这些"节点对"形成链接后, 将有助于彼此的效用增加, 从宏观上有助于整体的科研产出. 经过以上分析, 可以进一步考虑建立某种推送或推荐机制, 使得表 8 中列出的作者有见面和合作的机会.

5　结束语

不同于绝大多数的社会网络领域的既有研究, 本文将研究视角关注于网络参与者见面的概率问题, 注意到网络参与者见面是形成链接的前提. 为解决网络参与者见面的概率问题, 本文引入了效用分析的方法, 并在理论上将效用函数的估计问题与 logistic 回归分析联系起来, 在深入分析网络参与者链接形成过程的基础上, 提出了迭代估计算法, 并在一个算例上进行方法的展示, 在合著者网络中对方法进行应用研究. 结果

表明本文提出的方法可以解决网络参与者见面的概率估计问题. 虽然本文将模型在情报文献学领域做了一个简单的应用, 但是该模型依然有巨大的应用空间, 特别是在电子商务, 信息管理和市场营销领域. 期待进一步的研究在大量的真实数据集上对模型的实践效能进行全面地评估和验证.

参考文献:

[1] 李倩倩, 顾基发. 用户行为驱动的在线社交网络建模. 系统工程学报, 2015, 30(1): 9–15.
Li Q Q, Gu J F. Activity driven modeling of online social network. Journal of Systems Engineering, 2015, 30(1): 9–15. (in Chinese)

[2] Chan K C, Chang C H, Chang Y. The network effects of publishing in finance. The North American Journal of Economics and Finance, 2015, 33(6): 305–316.

[3] 陈 冀, 陈典发, 宋 敏. 复杂网络结构下异质性银行系统稳定性研究. 系统工程学报, 2014,29(2): 171–181.
Chen J, Chen D F, Song M. Heterogeneous bank system stability research under complex networks structure . Journal of Systems Engineering, 2014,29(2): 171–181. (in Chinese)

[4] Wong C Y, Goh K L. The sustainability of functionality development of science and technology: Papers and patents of emerging economies. Journal of Informetrics, 2012, 6(1): 55–65.

[5] Poelmans E, Rousseau S. Factors determining authors' willingness to wait for editorial decisions from economic history journals. Scientometrics, 2015, 102(2): 1347–1374.

[6] 俞立平, 张 全. 期刊评价中两类效用函数合成方法的本质研究. 情报学报, 2014, 33(10): 1077–1082.
Yu L P, Zhang Q. Two methods for utility function synthesis in the evaluation of journals. Journal of the China Society for Scientific and Technical Information, 2014, 33(10): 1077–1082. (in Chinese)

[7] Boucher V. Structural homophily. International Economic Review, 2015, 56(1): 235–264.

[8] Leung M P. Two-step estimation of network-formation models with incomplete information. Journal of Econometrics, 2015, 188(1): 182–195.

[9] Sauer N C, Kauffeld S. The ties of meeting leaders: A social network analysis. Psychology, 2015, 6(4): 415–416.

[10] 胡海波, 刘 璇. 在线社会网络增长中的优先连接. 系统工程学报, 2014, 29(3): 289–298.
Hu H B, Liu X. Preferential linking in the growth of online social networks. Journal of Systems Engineering, 2014, 29(3): 289–298. (in Chinese)

[11] Li Y, Zhang G, Feng Y, et al. An entropy-based social network community detecting method and its application to scientometrics. Scientometrics, 2015, 102(1): 1003–1017.

[12] 李永立, 吴 冲. 基于图模型和最优化的评价方法. 系统工程学报, 2013, 28(3): 403–409.
Li Y L, Wu C. Inventing an evaluation method based on graph model and optimization. Journal of Systems Engineering, 2013, 28(3): 403–409. (in Chinese)

[13] Coles S, Bawa J, Trenner L, et al. An Introduction to Statistical Modeling of Extreme Values. London: Springer, 2001.

[14] Christakis N A, Fowler J H. Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives. New York, NY: Little, Brown and Company, 2009.

[15] Newman M E J. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 2001, 98(2): 404–409.

作者简介:

李永立(1985—), 男, 辽宁沈阳人, 副教授, 硕士生导师, 研究方向: 社会网络分析, Email: ylli@mail.neu.edu.cn;

樊宁远(1994—), 男, 河南平顶山人, 博士生, 研究方向: 社会网络分析, Email: pdsfny@163.com;

林亿民(1994—), 男, 福建莆田人, 硕士生, 研究方向: 预测理论与方法, Email: ymlin@stumail.neu.edu.cn;

吴 冲 (1972—), 男, 黑龙江哈尔滨人, 教授, 博士生导师, 研究方向: 复杂网络与社会网络, Email: wuchong@hit.edu.cn.