考虑个体效用因素的社会网络演化分析模型①

李永立，陈 杨，樊宁远，高 馨
(东北大学工商管理学院，沈阳 110169)

摘要： 既有的社会网络演化分析模型往往通过统计分析的方法从宏观上描述网络演化规律，难以深入解释驱动网络演化的微观行为原因. 为了弥补以上不足，建立了网络参与者的效用函数，并引入效用分析的方法解释网络上链路形成或断开的现象，进而揭示驱动网络演化的微观行为原因. 与此同时，考虑到网络参与者见面过程是网络演化的内在过程，将其视为难以观测的潜变量引入模型，用以解释不能由效用分析所刻画的网络演化现象. 在以上理论模型基础上，为将其进一步量化和应用，基于对社会网络一个时期的观察和网络参与者个体属性的数据，发展了基于贝叶斯推断的参数估计方法，校准所建立效用函数中的参数并估计潜在的见面过程. 通过两组仿真分析验证了模型参数估计的准确性并讨论了模型的适用范围，并将模型应用于取自 Facebook 平台的真实数据集，实证了模型的解释力和预测力. 本文提出的模型将有助于解释社交媒体平台上社会网络形成的原因，并预测网络演化的趋势，为进一步优化社会网络结构和控制信息传播打下模型基础.

关键词： 网络演化；社会网络分析；效用分析；贝叶斯推断；社会媒体平台
中图分类号： TP391；N949 **文献标识码：** A **文章编号：** 1007 - 9807(2018)03 - 0041 - 13

0 引 言

随着信息技术的迅猛发展，社交媒体平台(如：微信、腾讯 QQ 和 Facebook 等)日益成为了主流的社交媒介，逐步取代以短信和电话为代表的传统交流方式. 在社交媒体平台上，人们通过建立朋友关系构成社会网络，并依托所形成的社会网络传递信息；由此，社交媒体平台上形成的社会网络构成了信息传播的骨架，并将进一步影响网络参与者的行为和决策[1]. 社交媒体平台上产生了海量的基于用户的信息，其不仅包含着用户的属性信息，还包含着用户间形成链接、以及组成社团的信息[2]等等. 本文旨在通过建立理论模型并结合对用户数据的分析，去揭示和探索社交媒体平台上，社会网络形成和演化的规律，以期解释影响网络形成的个体行为特征、并预测网络未来的发展趋势.

本文所建立的模型将有助于回答如下问题：具有相同职业或者相同年龄的用户是否更容易形成朋友链接？有更多共同朋友的用户间是否更容易形成链接？网络的局部结构会如何影响网络的演化趋势？等等. 因此，对于社会网络演化分析模型的研究将对于管理学领域，特别是信息管理领域，有着潜在和广泛的应用价值. 事实上，关于模型价值的论断，按照既有文献对于社交媒体平台研究所归纳的三个阶段“挖掘”，“理解”和“展示”[3]，本文的研究属于其中的第二阶段. 就这一阶段的研究而言，其引起了信息管理领域广大研究者的关注；比如：在 Twitter 数据基础上对于信息传播的研究[4]和在 Facebook 数据基础上对于用户行为的研究[5]. 虽然同为对于社会网络媒体

① 收稿日期：2016 - 06 - 14；修订日期：2017 - 02 - 21.
基金项目：国家自然科学基金资助项目(71501034)；中国博士后科学基金资助项目(2016M590230；2017T100183).
作者简介：李永立(1985—)，男，辽宁沈阳人，副教授，硕士生导师. Email：ylli@mail.neu.edu.cn

平台“理解”层面的研究，本文的研究不同于以上列举的文献：以上文献偏重于对理论的实证研究，是对网络形成后某些现象和规律的分析与总结；而本文偏重于建模的研究，考查的是社会网络媒体平台上用户间的社会网络因何形成、会怎样演化、未来将走向何方、以及如何控制和预测的问题，这是本文在理论出发点上不同于既有研究的地方，是本研究的动机之一.

既有的分析社会网络演化规律的主流模型主要有两类：一类是“指数随机图模型”[6,7]，另一类是“策略网络形成模型”[8]. “指数随机图模型”认为网络形成遵循着一种随机增长的机制，该机制在既有的文献中被称为 CHANCE；其强调了随机性是网络演化的核心因素，因此该类模型可以被视为纯粹的统计模型，其没有考虑网络参与者的效用函数、收益和由此形成的策略. 不同于“指数随机图模型”，“策略网络形成模型”强调了网络是基于网络参与者的决策形成的，认为形成网络的机制在于网络参与者的 CHOICE；由此，“策略网络形成模型”建立在网络参与者效用函数的基础上，从参与者建立链接的收益分析入手，从收益的角度考虑是否建立相应的链接. 对比这两类模型，指数随机图模型提出较早，是目前应用较广泛的一类模型，特别是在管理科学领域，近年来涌现了大量应用指数随机图模型进行网络演化分析的文献[9-12]. 但是，既有的指数随机图模型有两个显著的缺陷：一是参数估计的不一致问题[13]，二是估计算法复杂度高的问题[14]. 而“策略网络形成模型”由于从效用函数入手分析，相对于纯粹的统计模型，其结果更加稳健并且契合本文研究动机中能够预测网络演化的要求，因此本文建立的分析模型采用“策略网络形成模型”的思想体系，从网络参与者的效用函数刻画入手，并从收益的角度考察建立或断开链接的行为，这是本文建立模型的思想基础.

进一步，不同于既有的“策略网络形成模型”，本文强调了网络个体见面机制对于社会网络演化的影响，并将建立一种灵活的网络个体见面机制，这是本文对于既有模型的一个理论贡献. 具体地，以文献[15，16]为代表其分别给出了见面过程的两个极端状态. 文献[15]限定了每组网络个体对见面且仅见面一次，而文献[16]考查的是经过无限次见面过程后，网络达到稳定状态的情况，也即每组网络个体对可以有无限次的见面机会. 事实上，这两种状态都过于理想化：一方面，不一定每组个体对都有见面的机会，文献[15]的假定不符合这一现实情形；另一方面，现实中的网络演化可能会经历不同的时期，并不是所有的网络都必然达到稳态，因此文献[16]的结果仅仅适用于部分网络. 本文根据现实的情形，放宽了网络个体见面的次数假定，允许他们之间有的“个体对”没有见过面，有的“个体对”见过多次面，以期本文提出的模型将更加接近于“个体对”真实的见面过程，这将有助于补充既有的研究结论，是本文另一个主要的研究动机.

更进一步，灵活的见面机制事实上也对本文模型的算法提出了挑战. 作为考虑个体效用的社会网络演化分析模型，其求解算法建立在对效用函数中参数的估计和潜在的见面过程估计的基础上. 具体地，通过收集一个时间点上的网络结构数据以及个体属性的数据，推断效用函数中的参数和潜在的见面过程，并以估计值为基础，解释网络演化的主要驱动因素并预测网络可能的演化状态. 注意到，在网络的演化过程中，不易直接被观察到的见面过程是网络形成和演化重要的环节，因为如果一组节点对没有获得见面的机会，即便根据个体效用的分析，他们形成链接或断开链接能够给彼此带来巨大的效用增加，其链接情况也不会改变. 由此，本文的算法事实上可以划归为一个包含隐状态的参数估计方法. 事实上，关于这一领域的算法研究已经有了一定的积累，比如：隐马尔科夫模型；但是，其采用的“前向—后向算法”并不适用于这一问题，这是因为在传统的隐马尔可夫模型中，隐状态的数量是有限的，并且往往很少. 而本文由于考虑了灵活的见面机制，对于含有 n 个节点的网络，其见面的隐状态有 $2^{n(n-1)/2}$ 种情形，这是一个非常庞大的数字，直接套用隐马尔科夫模型是难以奏效的. 鉴于此，本文有发展既有算法的研究动机，需要提出新的算法在保证估计精度的同时，适用于本模型特定的结构.

综上所述，本文将建立一个考虑个体效用和网络个体见面过程的社会网络演化分析模型，通过对网络一个时间点上的数据收集，发展新的精度较高和复杂度较低的算法，以期估计潜在的节点对见面次序和效用函数中的参数，通过估计的参数值，

解释网络演化的规律并预测网络未来的发展趋势.

1　效用函数

1.1　效用函数的建立及一个示例

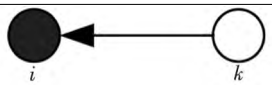
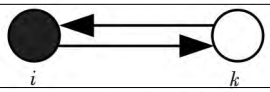
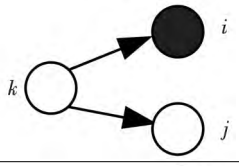
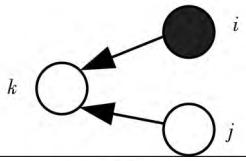
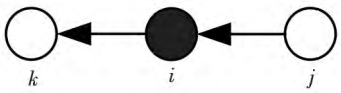
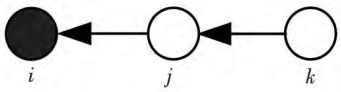
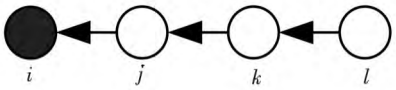
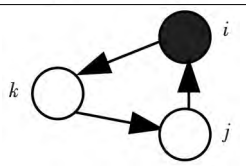
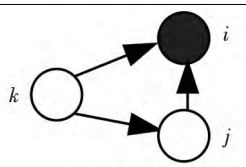
在考虑社会网络的背景下,网络个体的效用函数往往取决于网络的结构特征和个体的属性特征[17,18];由此,以社会网络上代表性个体 i 为例,在 t 时刻该个体效用函数的一般结构如下

$$U_i(g^t, C; \theta) \tag{1}$$

其中 g^t 表示 t 时刻的网络结构(这里用邻接矩阵表示),矩阵 C 表示全体网络个体的属性矩阵(其中代表性个体 i 的属性向量记为 C_i),向量 θ 为待估计的参数向量. 为了便于在具体研究背景下设计效用函数,表 1 列出了社会网络中常用的网络结构及其数学表达.

表 1　常用的网络结构效应的表示方法

Table 1 The commonly-used statistics of reflecting network structure effect

图形展示	网络结构效应的名称	数学表示
	直接影响 (direct effect)	$\sum_k g_{ik}$
	交互影响 (reciprocated effect)	$\sum_k g_{ik}g_{ki}$
	基于来源的共同朋友效应 (effect from common friends (origin))	$\sum_{j,k} g_{ik}g_{jk}$
	基于目标的共同朋友效应 (effect from common friends (target))	$\sum_{j,k} g_{ki}g_{kj}$
	混合效应 (popularity effect)	$\sum_{j,k} g_{ki}g_{ij}$
	来自距离为 2 的朋友的效应 (effect from 2-path actors)	$\sum_{j,k} g_{ij}g_{jk}$
	来自距离为 3 的朋友的效应 (effect from 3-path actors)	$\sum_{j,k,l} g_{ij}g_{jk}g_{kl}$
	环效应 (cyclic effect)	$\sum_{j,k} g_{ij}g_{jk}g_{ki}$
	传递效应 (transitive effect)	$\sum_{j,k} g_{ij}g_{ik}g_{jk}$

注: 表中的网络效应是以有向网络为基础,如果不强调链接的有向性,其中的"直接影响"和"交互影响"是等价的,两种不同来源的共同朋友效应是等价的,以及"环效应"和"传递效应"是等价的. 还有在效用函数设计方面的一个细节需要指出: 这里的 $g_{ik} = 1$ 表示从个体 i 到个体 k 建立了链接,而表 1 左边的图形表示的是影响产生的方向,并不是建立链接的方向. 具体地以表 1 中的第一行为例,图示中的箭头反映了当个体 i 向个体 k 建立链接时,个体 k 对个体 i 的影响作用,所以根据建立链接的方向,数学表示中给出的是 g_{ik}. 这一点在建模的时候很容易混淆,特别加以说明.

需要指出的是，本文模型的效用函数并不是唯一的，可以根据具体研究的背景和研究目标设计不同形式的效用函数. 以下提供效用函数设计的一个具体示例，其也成为本文后续部分所应用效用函数的具体形式. 该示例以研究 Facebook 社交平台上朋友网络(无向网络)的演化驱动因素为背景，以考查是否“朋友的属性差异”、“共同朋友数”和“来自于距离为 2 的朋友的属性差异”会影响网络的演化. 由此，在示例的研究背景和研究目标下，所设计的效用函数应包含直接影响、共同朋友的影响和距离为 2 的朋友的影响这三个模块，从表 1 中选择这三个模块，则代表性个体 i 的效用函数可以设计为

$$U_i(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta}) = \theta_0 + \theta_1 \sum_k g_{ik}^t \| \boldsymbol{C}_i - \boldsymbol{C}_k \| + \theta_2 \sum_{k,l} g_{il}^t g_{kl}^t + \theta_3 \sum_{k,l} g_{ik}^t g_{kl}^t \| \boldsymbol{C}_i - \boldsymbol{C}_l \| \quad (2)$$

这里 $\| \boldsymbol{C}_i - \boldsymbol{C}_k \| := \sum_m | c_{im} - c_{km} |$，其中 c_{im} 是个体 i 的属性向量 $\boldsymbol{C}_i$ 的第 m 个元素. 等式后面的第一项是“常数项”，第二项反映“直接效应”，第三项反映“共同朋友的影响”，以及第四项反映“距离为 2 的朋友效应”，注意到该效用函数的设计与示例中研究的目标和背景是相适应的. 诚然，效用函数不仅仅只有本文示例所给出的一种形式，本文的模型允许其他形式的效用函数. 另一方面，本文将在第 3 部分阐述的估计方法是一个普适意义的方法，其建立在一般意义的效用函数基础上，并不依赖于效用函数具体的特定结构，这也与以上允许多种形式的效用函数的思想相一致.

1.2 链接形成或断开的条件

根据效用分析的基本思想，本文假定每个网络个体都根据自身的效用情况做出是否建立或断开链接的决策，这里简称为“利己性假定”. 具体地，以个体 i 和 j 为代表性“个体对”，以下分链接形成和链接断开的情况分别进行分析.

1) 对于链接形成的情形

如果个体 i 和个体 j 之间尚未形成链接，则当个体 i 考虑与个体 j 建立链接时，个体 i 的效用变化为

$$\begin{aligned}\Delta U_{i\leftarrow j}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta}) &= U_i(\boldsymbol{g}^{t-1}|_{g_{ij}=1},\boldsymbol{C};\boldsymbol{\theta}) - U_i(\boldsymbol{g}^{t-1}|_{g_{ij}=0},\boldsymbol{C};\boldsymbol{\theta}) \\ &= \theta_1 \| \boldsymbol{C}_i - \boldsymbol{C}_j \| + \theta_2 \sum_{l\neq i,j} g_{il}^{t-1} g_{jl}^{t-1} + \theta_3 \sum_{l\neq i,j} g_{jl}^{t-1} \| \boldsymbol{C}_i - \boldsymbol{C}_l \| \end{aligned} \quad (3a)$$

其中 $\boldsymbol{g}^{t-1}|_{g_{ij}=1}$ 表示 t 时刻的网络除了限制 $g_{ij}=1$ 外，与 $t-1$ 时刻的网络相同；由此可推知 $\boldsymbol{g}^{t-1}|_{g_{ij}=0}$ 的含义，下同. 类似地，则当个体 j 考虑与 i 建立链接时，个体 j 的效用变化为

$$\begin{aligned}\Delta U_{j\leftarrow i}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta}) &= U_j(\boldsymbol{g}^{t-1}|_{g_{ij}=1},\boldsymbol{C};\boldsymbol{\theta}) - U_j(\boldsymbol{g}^{t-1}|_{g_{ij}=0},\boldsymbol{C};\boldsymbol{\theta}) \\ &= \theta_1 \| \boldsymbol{C}_i - \boldsymbol{C}_j \| + \theta_2 \sum_{l\neq i,j} g_{il}^{t-1} g_{jl}^{t-1} + \theta_3 \sum_{l\neq i,j} g_{il}^{t-1} \| \boldsymbol{C}_j - \boldsymbol{C}_l \| \end{aligned} \quad (3b)$$

由此，根据“利己性假定”，尚未建立链接的个体 i 和 j 在获得见面机会的前提下，建立连接的充要条件为

$$\Delta U_{i\leftarrow j}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta}) \geqslant 0 \text{ 且 } \Delta U_{j\leftarrow i}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta}) \geqslant 0 \quad (4)$$

2) 对于链接断开的情形

如果个体 i 和 j 之间已经形成链接，则当个体 i 考虑与个体 j 断开链接时，个体 i 的效用变化为

$$\begin{aligned}\Delta U_{i\leftarrow j}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta})' &= U_i(\boldsymbol{g}^{t-1}|_{g_{ij}=0},\boldsymbol{C};\boldsymbol{\theta}) - U_i(\boldsymbol{g}^{t-1}|_{g_{ij}=1},\boldsymbol{C};\boldsymbol{\theta}) \\ &= -\theta_1 \| \boldsymbol{C}_i - \boldsymbol{C}_j \| - \theta_2 \sum_{l\neq i,j} g_{il}^{t-1} g_{jl}^{t-1} - \theta_3 \sum_{l\neq i,j} g_{jl}^{t-1} \| \boldsymbol{C}_i - \boldsymbol{C}_l \| \end{aligned} \quad (5a)$$

类似地，则当个体 j 考虑与 i 断开链接时，个体 j 的效用变化为

$$\begin{aligned}\Delta U_{j\leftarrow i}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta})' &= U_j(\boldsymbol{g}^{t-1}|_{g_{ij}=0},\boldsymbol{C};\boldsymbol{\theta}) - U_j(\boldsymbol{g}^{t-1}|_{g_{ij}=1},C;\theta) \\ &= -\theta_1 \| \boldsymbol{C}_i - \boldsymbol{C}_j \| - \theta_2 \sum_{l\neq i,j} g_{il}^{t-1} g_{jl}^{t-1} - \theta_3 \sum_{l\neq i,j} g_{il}^{t-1} \| \boldsymbol{C}_j - \boldsymbol{C}_l \| \end{aligned} \quad (5b)$$

由此，根据“利己性假定”，已经建立链接的个体 i 和 j 在获得见面机会的前提下，断开连接的充要条件为

$$\Delta U_{i\leftarrow j}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta})' \geqslant 0 \text{ 或 } \Delta U_{j\leftarrow i}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta})' \geqslant 0 \quad (6)$$

事实上，以上两种情形的数学形式虽然有差别，但公式在本质上表达的含义是一致的. 为了说明这一点，观察式(3a)和式(3b)和式(5a)和式(5b)就是相反数的关系，进而式(4)和式(6)表达的条件互为逆否命题，因此以上两种情况可以合并为一种情况计算. 无论代表性节点对 i 和 j 是否已经形成链接，按照式(3a)和式(3b)分别计算 $\Delta U_{i\leftarrow j}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta})$ 和 $\Delta U_{j\leftarrow i}(\boldsymbol{g}^t,\boldsymbol{C};\boldsymbol{\theta})$，如果两者同时非负，则在 t 时刻建立或保持链接的状态，如果两者至少有一个非正，则在 t 时刻断开链接或保持未链接的状态. 由此，不必预先判断“个体对”当

前的链接状况并分别讨论,以免提高算法的复杂度和降低计算的效率.

2　演化分析模型

2.1　潜在的见面过程

网络个体对的见面过程是网络形成和演化的内在过程和必要前提,而由于见面过程往往不易直接被观察到,所以在分析网络演化的过程时常常被忽略掉. 事实上,忽略了见面过程的分析,往往会导致有偏的估计结果,这是因为网络演化的结果不仅仅是个体决策的结果,还受制于见面的几率,这也是社交媒体平台能够在社会网络建立过程中发挥关键作用的原因所在. 本节将建立一个灵活的网络个体见面过程的模型. 将网络个体的见面过程定义为一个随机序列,表示为 $\boldsymbol{M}=(m^t)_{t=1}^{T}$,其开始于空网络,结束于时刻 T. 该序列由一系列的节点对组成,也即该序列的组成部分 $m^t=\{i,j\}$ 意味着个体 i 和 j 在时刻 t 相遇,或记为 $m_{ij}^t=1$.

关于见面过程,一个直接的事实是: 如果个体对在观察到的网络 g 中形成了链接,说明这两个个体肯定见过面,必然会出现在见面序列 $\boldsymbol{M}$ 中; 如果这两个个体没有形成链接,则这两个个体可能见过面,可能没有见过面,是否其出现在 $\boldsymbol{M}$ 中,需要对其进行推断. 由此,根据贝叶斯定理,有以下的推论 1 成立.

推论 1　(未形成链接的"节点对"出现在见面序列中的后验概率). 在观察到的网络 g 中,那些满足 $g_{ij}=0$ 的个体对$\{i,j\}$,见过面的概率(也即出现在序列 $\boldsymbol{M}$ 中的概率) 为

$$p(m_{ij}=1\mid g_{ij}=0,C;\boldsymbol{\theta})=\frac{0.5\cdot\left(1-p(\Delta U_{i\leftarrow j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0,\Delta U_{j\leftarrow i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0)\right)}{1-0.5\cdot p(\Delta U_{i\leftarrow j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0,\Delta U_{j\leftarrow i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0)} \tag{7}$$

证明　首先在没有关于"节点对"是否见面的信息时,其先验概率为

$$p(m_{ij}=1)=p(m_{ij}=0)=0.5 \tag{8}$$

即: 两者见过面与未见过面具有相同的概率. 而后,根据贝叶斯公式,得到

$$p(m_{ij}=1\mid g_{ij}=0,\boldsymbol{C},\boldsymbol{\theta})=\frac{p(m_{ij}=1,g_{ij}=0\mid \boldsymbol{C},\boldsymbol{\theta})}{p(m_{ij}=0,g_{ij}=0\mid \boldsymbol{C},\boldsymbol{\theta})+p(m_{ij}=1,g_{ij}=0\mid \boldsymbol{C},\boldsymbol{\theta})} \tag{9}$$

为了计算分子和分母,进一步根据条件概率的等式关系,成立

$$p(m_{ij}=1,g_{ij}=0\mid \boldsymbol{C},\boldsymbol{\theta})=p(m_{ij}=1)\times p(g_{ij}=0\mid m_{ij}=1,\boldsymbol{C},\boldsymbol{\theta}) \tag{10}$$

其中式(8) 给出了 $p(m_{ij}=1)=0.5$,以及根据式(4),在两者见面的条件下,没有形成链接的概率为

$$p(g_{ij}=0\mid m_{ij}=1,C,\boldsymbol{\theta})=1-p(\Delta U_{i\leftarrow j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0,\Delta U_{j\leftarrow i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0) \tag{11a}$$

类似的

$$p(m_{ij}=0,g_{ij}=0\mid \boldsymbol{C},\boldsymbol{\theta})=p(m_{ij}=0)\times p(g_{ij}=0\mid m_{ij}=0,\boldsymbol{C},\boldsymbol{\theta}) \tag{11b}$$

其中式(8) 给出了 $p(m_{ij}=0)=0.5$; 以及在两者没有见面的条件下,必然不会形成链接,成立以下等式关系 $p(g_{ij}=0\mid m_{ij}=0,\boldsymbol{C},\boldsymbol{\theta})=1$. 将以上两个结果带入式(9),即可得到式(7) 的结论.

由此,根据推论 1,在给定网络 g 和参数向量 $\boldsymbol{\theta}$ 时,可以推断见面序列 $\boldsymbol{M}$ 的构成: 其包含那些已经形成链接的节点对和未形成链接、但基于后验概率推断认为见过面的节点对.

2.2　模型的似然函数

由以上的分析可以发现: 本文将社会网络链路形成的过程分为了见面过程和决策过程两个部分. 由此,本文中观察到网络 g 的似然函数也同时包含这两个过程,其表达式如下

$$l(\boldsymbol{\theta}\mid \boldsymbol{g},C):=p(\boldsymbol{g}\mid \boldsymbol{C},\boldsymbol{\theta})=p(\boldsymbol{M}\mid \boldsymbol{C},\boldsymbol{\theta})\times p(\boldsymbol{g}\mid \boldsymbol{M},\boldsymbol{C},\boldsymbol{\theta}) \tag{12}$$

其中 $p(\boldsymbol{M}\mid \boldsymbol{C},\boldsymbol{\theta})$ 表示在收集到的属性数据基础上和特定的参数取值下,出现见面过程 $\boldsymbol{M}$ 的概率,而 $p(\boldsymbol{g}\mid \boldsymbol{M},\boldsymbol{C},\boldsymbol{\theta})$ 表示在给定见面过程、个体属性和参数取值时,出现观察到的网络 g 的概率. 进一步根据最大似然估计的原理,可能估计出参数的最优取值,使得出现网络 g 的概率最大. 值得一提的是,这里见面过程 $\boldsymbol{M}$ 是潜变量,因此这个概率的获得事实上是对潜变量的推断. 以下分别对这两个组成部分进行分析,以期明确其计算的具体流程,为进一步估计参数做好铺垫.

1) 对 $p(\boldsymbol{M}\mid \boldsymbol{C},\boldsymbol{\theta})$ 的分析与计算

根据推论 1 和给定的网络 g,每得到一组参数向量 $\boldsymbol{\theta}$ 的估计值都可以推断出见面过程 $\boldsymbol{M}$ 中

所包含的全部节点对. 但考虑到节点对是有序的，本文将采用基于 MCMC 的随机抽样方法找到在某组参数条件下最优的见面序列，也即在给定参数向量 $\boldsymbol{\theta}$ 的条件下，使得式(12)取得最大值的序列 $\boldsymbol{M}$.

2) 对 $p(\boldsymbol{g}\mid \boldsymbol{M},\boldsymbol{C},\boldsymbol{\theta})$ 的分析与计算

在对这一概率值进行分析之前，首先回顾式(3a)和式(3b)，考虑到 Logistic 回归估计问题与这里效用函数中参数的估计问题有着比较类似的结构，因为效用的变化作为被解释变量，是不能直接观察和获得的，取而代之的是网络个体真实的链接情况，这时被解释变量为 0 - 1 的形式. 由此，根据 Logistic 回归得到如下概率表达式

$$p(\Delta U_{i\leftarrow j}(g^t,\boldsymbol{C};\boldsymbol{\theta})\geqslant 0)=\frac{\exp\left(\theta_1\|\boldsymbol{C}_i-\boldsymbol{C}_j\|+\theta_2\sum_{l\neq i,j}g_{il}^{t-1}g_{jl}^{t-1}+\theta_3\sum_{l\neq i,j}g_{jl}^{t-1}\|\boldsymbol{C}_i-\boldsymbol{C}_l\|\right)}{1+\exp\left(\theta_1\|\boldsymbol{C}_i-\boldsymbol{C}_j\|+\theta_2\sum_{l\neq i,j}g_{il}^{t-1}g_{jl}^{t-1}+\theta_3\sum_{l\neq i,j}g_{jl}^{t-1}\|\boldsymbol{C}_i-\boldsymbol{C}_l\|\right)} \tag{13a}$$

以及

$$p(\Delta U_{j\leftarrow i}(g^t,\boldsymbol{C};\boldsymbol{\theta})\geqslant 0)=\frac{\exp\left(\theta_1\|\boldsymbol{C}_i-\boldsymbol{C}_j\|+\theta_2\sum_{l\neq i,j}g_{il}^{t-1}g_{jl}^{t-1}+\theta_3\sum_{l\neq i,j}g_{il}^{t-1}\|\boldsymbol{C}_j-\boldsymbol{C}_l\|\right)}{1+\exp\left(\theta_1\|\boldsymbol{C}_i-\boldsymbol{C}_j\|+\theta_2\sum_{l\neq i,j}g_{il}^{t-1}g_{jl}^{t-1}+\theta_3\sum_{l\neq i,j}g_{il}^{t-1}\|\boldsymbol{C}_j-\boldsymbol{C}_l\|\right)} \tag{13b}$$

并且根据“利己性假定”，两个个体独立做出自己的行为决策，成立

$$p(\Delta U_{i\leftarrow j}(g^t,\boldsymbol{C};\boldsymbol{\theta})\geqslant 0,\Delta U_{j\leftarrow i}(g^t,\boldsymbol{C};\boldsymbol{\theta})\geqslant 0)=p(\Delta U_{i\leftarrow j}(g^t,\boldsymbol{C};\boldsymbol{\theta})\geqslant 0)\cdot p(\Delta U_{j\leftarrow i}(g^t,\boldsymbol{C};\boldsymbol{\theta})\geqslant 0) \tag{14}$$

进一步，记序列 $\boldsymbol{M}$ 中第 t 个节点对为 (i_t,j_t)，则 t 时刻对于形成链接或保持既有链接的情形，成立

$$p(\boldsymbol{g}^t\mid \boldsymbol{g}^{t-1},m^t,\boldsymbol{C},\boldsymbol{\theta})=p(\Delta U_{i_t\rightarrow j_t}(\boldsymbol{g}^{t-1},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0,\Delta U_{j_t\rightarrow i_t}(\boldsymbol{g}^{t-1},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0) \tag{15a}$$

否则，在 t 时刻对于断开链接或无链接的情形成立

$$p(\boldsymbol{g}^t\mid \boldsymbol{g}^{t-1},m^t,\boldsymbol{C},\boldsymbol{\theta})=1-p(\Delta U_{i_t\rightarrow j_t}(\boldsymbol{g}^{t-1},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0,\Delta U_{j_t\rightarrow i_t}(\boldsymbol{g}^{t-1},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0) \tag{15b}$$

基于以下条件概率的等式

$$p(\boldsymbol{g}\mid \boldsymbol{M},\boldsymbol{C},\boldsymbol{\theta})=\prod_{t=1}^{T}p(\boldsymbol{g}^t\mid \boldsymbol{g}^{t-1},m^t,\boldsymbol{C},\boldsymbol{\theta}) \tag{16}$$

并结合式(15a)和式(15b)，则 $p(\boldsymbol{g}\mid \boldsymbol{M},\boldsymbol{C},\boldsymbol{\theta})$ 可具体表示为

$$p(\boldsymbol{g}\mid \boldsymbol{M},\boldsymbol{C},\boldsymbol{\theta})=\prod_{t=1}^{T}p(\Delta U_{i_t\rightarrow j_t}(\boldsymbol{g}^{t-1},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0,\Delta U_{j_t\rightarrow i_t}(\boldsymbol{g}^{t-1},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0)^{g_{i_t,j_t}^t}\times$$
$$(1-p(\Delta U_{i_t\rightarrow j_t}(\boldsymbol{g}^{t-1},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0,\Delta U_{j_t\rightarrow i_t}(\boldsymbol{g}^{t-1},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0))^{1-g_{i_t,j_t}^t} \tag{17}$$

由此，基于对见面序列 $\boldsymbol{M}$ 和参数向量 $\boldsymbol{\theta}$ 的推断，可以计算出网络 g 的发生概率，这是式(12)所示的似然函数的一个重要组成部分.

3 估计方法

结合以上对网络演化过程的分析，本文估计方法的输出是参数向量 $\boldsymbol{\theta}$ 的估计值，输入是观察到的网络 g 和个体的属性矩阵 $\boldsymbol{C}$，所估计的原理是最大似然估计，也即使得式(12)所示的似然函数取得最大值. 考虑到见面过程 $\boldsymbol{M}$ 的存在，结合以上给出的推论和公式，其估计的流程图如图 1 所示.

根据图 1 所示的步骤，因为每一步的结果都使得似然函数的值变大，对于一个有上界的似然函数而言(如式(12)所示的似然函数必然不超过 1，因此是有上界的)，其算法必然会收敛. 表 2 中概述了本文算法的伪代码.

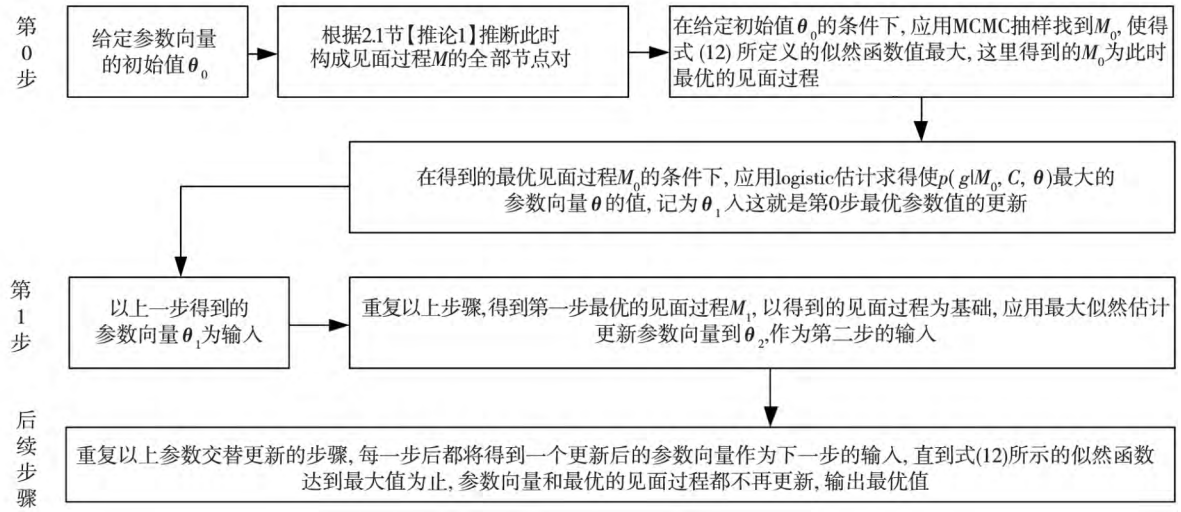


图 1　估计方法的基本流程

Fig. 1 Basic procedure of the proposed estimation method

表 2　估计算法的伪代码

Table 2 Algorithm of the proposed estimation method

输　入: 观察到的网络 g 和网络参与者的属性矩阵 C

输　出: 最优的参数向量估计值 θ^*

初始化: 参数向量的初始值 θ_0 和随机生成的初始见面序列 M_{-1} ,以及 $S_0 = \{(i,j) \mid g_{ij} = 0\}$,

$S_1 = \{(i,j) \mid g_{ij} = 1\}$,$S = \Phi$ **和** $k = 0$　　# Φ **表示空集**

过　程:

while $\eta < 10^{-5}$

　for　**所有属于集合** S_0 **的节点对** (i,j)

　　根据式(7) ,计算 $p(m_{ij} = 1 \mid g_{ij} = 0, \boldsymbol{C}; \boldsymbol{\theta}_k)$ **的值**

　　if *rand* ( ) $> p(m_{ij} = 1 \mid g_{ij} = 0, \boldsymbol{C}; \boldsymbol{\theta})$　　# *rand*( ) **用于生成** 0 – 1 **之间的随机数;**

　　将该节点加入集合 S **,**

　　end if

　end for

　则第 k **次循环的节点对集合** $SS_k = S_1 \cup S$.　　# **得到序列中全部的节点对;**

　#(1) **应用** MCMC **算法得到此时最优的见面序列** $\boldsymbol{M}_k$

　$q = 1$;

　while **超过** 25 **步见面序列没有更新**

　　为当前的序列赋初值: $\boldsymbol{MM}_q = M_{k-1}$,

　　随机安排集合 SS_k **中节点对的顺序,生成一个候选见面序列** $\boldsymbol{MM}'$,

　　根据标准的 MCMC **算法,计算选择概率,如下所示**

$$\beta(\boldsymbol{MM}', \boldsymbol{MM}_q \mid \boldsymbol{\theta}_k, \boldsymbol{g}, \boldsymbol{C}) = \min\left\{1, \frac{p(\boldsymbol{g} \mid \boldsymbol{MM}', \boldsymbol{C}, \boldsymbol{\theta}_k)}{p(\boldsymbol{g} \mid \boldsymbol{MM}_q, \boldsymbol{C}, \boldsymbol{\theta}_k)}\right\},$$

　　#**其中,** $p(\boldsymbol{g} \mid \boldsymbol{MM}', \boldsymbol{C}, \boldsymbol{\theta}_{k-1})$ **和** $p(\boldsymbol{g} \mid \boldsymbol{MM}_q, \boldsymbol{C}, \boldsymbol{\theta}_k)$ **的计算请回顾式(16);**

　　进一步,序列的更新规则为

$$\boldsymbol{MM}_{q+1} = \begin{cases} \boldsymbol{MM}' & \text{with probability } \beta(\boldsymbol{MM}', \boldsymbol{MM}_q \mid \boldsymbol{\theta}_k, \boldsymbol{g}, \boldsymbol{C}) \\ \boldsymbol{MM}_q & \text{with probability } 1 - \beta(\boldsymbol{MM}', \boldsymbol{MM}_q \mid \boldsymbol{\theta}_k, \boldsymbol{g}, \boldsymbol{C}) \end{cases}$$

　　$q = q + 1$,

　end while

　$\boldsymbol{M}_k = \boldsymbol{MM}_q$;

　#(2) **应用最大似然估计求得此时最优的参数向量** $\boldsymbol{\theta}_{k+1}$

　$\boldsymbol{\theta}_{k+1} = \arg\max_{\theta} p(\boldsymbol{g} \mid \boldsymbol{M}_k, \boldsymbol{C}, \boldsymbol{\theta})$,　　# $p(\boldsymbol{g} \mid \boldsymbol{M}_k, \boldsymbol{C}, \boldsymbol{\theta})$ **可根据式(16) 得到;**

$k = k + 1$;

#**计算更新前后最优值的变化率**

$\eta = |p(\boldsymbol{g} \mid \boldsymbol{M}_{k-1}, \boldsymbol{C}, \boldsymbol{\theta}_k) - p(\boldsymbol{g} \mid \boldsymbol{M}_{k-2}, \boldsymbol{C}, \boldsymbol{\theta}_{k-1})| / p(\boldsymbol{g} \mid \boldsymbol{M}_{k-2}, \boldsymbol{C}, \boldsymbol{\theta}_{k-1})$,

end while

注: #**号后为注释的部分**.

4 基于仿真分析的模型验证

本节将生成指定参数的随机网络，并通过参数估计值与真实值差异的比较，用以验证本文方法解释网络演化分析的能力，特别是对效用函数中参数值的估计能力. 为此，首先介绍随机网络的生成过程，总结在表 3 中. 基于表 3 所列出的随机网络相关参数的取值，可以得到不同演化时长或不同见面概率下的随机网络. 为了全面的验证模型，仿真分两种情况进行分析，以期获得参数估计误差的规律.

4.1 不同的见面概率

给定网络的演化时长为 10 000，注意到对于 200 个节点的网络而言，每组“节点对”见过一次面至少需要 19 900 个周期的演化时长，因此这里给定的演化时长不能保证节点对都至少见过一次面. 而后，对见面概率分别均匀取 0.20 至 1.00 中的五个值，考察在不同的见面概率下，本文估计算法的有效性，结果如表 4 所示.

通过表 4 的数据可以发现，本文提出的模型在参数估计方面的表现与见面概率的取值有关：当概率值越大时，也即网络个体获得更多的见面机会时，模型的参数估计值越准确. 虽然模型的估计结果随着概率的降低，表现的越来越偏离真实值，但是以上的结果也可以发现模型的稳健性，也即在概率为 0.20 时，该模型依然能够给出参数正确的符号估计.

表 3 随机网络的生成过程

Table 3 Generation process of random network

生成的过程或相关参数	注释
节点数	不失一般性，随机网络设定为 200 个节点
节点属性	节点的属性随机生成，取自区间 [0,1] 中的随机变量
见面概率	参照随机网络的生成特点，假定节点间见面是独立同分布的，则见面概率设为 p，在本文中是一个可以控制的变量
演化时长	演化时长设为 T，在本文中是一个可以控制的变量
参数真实值	采用本文式(2)中的效用函数形式，有效的三个参数值 θ_1，θ_2 和 θ_3 分别设为 −1.00，2.00 和 −0.01

表 4 不同见面概率条件下模型参数的估计结果

Table 4 Summary statistics of parameter estimates under different meeting probabilities

	θ_1		θ_2		θ_3	
p	平均值	标准差	平均值	标准差	平均值	标准差
0.20	−0.391	0.036 7	0.101	0.033 5	−0.001	0.013 3
0.40	−0.498	0.035 2	0.753	0.029 5	−0.005	0.006 8
0.60	−0.911	0.040 1	1.204	0.024 0	−0.006	0.004 7
0.80	−1.092	0.042 2	1.556	0.029 9	−0.007	0.004 0
1.00	−1.121	0.043 0	1.947	0.035 2	−0.010	0.003 9

注：在执行表 2 所示的算法步骤时，3 个参数的初始值都取为 0，下同.

4.2 不同的演化时长

类似于 4.1 节的分析，本节给定网络个体见面的概率为 0.50，而后考察在不同的演化时长条件下，验证算法的表现，结果如表 5 所示. 具体地，本文提出的模型在参数估计方面的表现与随机网络固有的演化时长有关，随着演化时长的变短，参数估计的效果越差. 特别是当网络的演化时长为 1 000时，这意味着网络是新兴的，网络个体之间见面的过程并不充分，这时模型不能给出参数正确的符号估计，而当网络的演化逐步进行到较为充分的时候，模型能够给出更加准确的估计结果.

表 5　不同演化时长下模型参数的估计结果

Table 5 Summary statistics of parameter estimates under different evolution periods

	θ_1		θ_2		θ_3	
T	平均值	标准差	平均值	标准差	平均值	标准差
1 000	-0.120	0.021 9	-0.863	0.022 4	0.016	0.047 6
5 000	-0.501	0.031 1	0.233	0.026 1	-0.001	0.010 7
10 000	-0.772	0.036 2	1.045	0.024 3	-0.006	0.005 6
15 000	-0.823	0.038 8	1.585	0.030 2	-0.004	0.004 4
19 900	-1.077	0.040 1	2.033	0.034 4	-0.009	0.003 7

事实上,以上两类不同情形的仿真分析,抽象了在现实生活中遇到的社交媒体平台的两大基本特征: 交流的便利性和平台运行时间的长短. 对于一个交流便利、乃至于有朋友推荐功能的社交媒体平台而言,将对应具有较大 p 值的情形,这是因为仿真中的 p 是网络个体间的见面概率,一个高效的社交媒体平台恰恰相当于提高了这个值. 另一方面,不同的社交媒体平台有着不同的运营时间,比如: 微信就比 QQ 上线的时间晚,由此也有了形成社交网络不同的演化时长,这对应于仿真分析中的 T. 由此,不同的 p 值和 T 值恰恰是对形成社会网络的社交媒体平台自身固有特性的抽象和描述,而这里的仿真结果阐明了本文模型的适用范围: 该模型适用于那些较为成熟和高效的社交媒体平台,有助于给出更加准确的参数估计结果,这一定程度也是本文模型对数据质量的客观要求.

5　模型的应用研究

本节应用 Facebook 社交媒体平台上的数据集,对"引言"中所提及的两个功能 "解释影响网络形成的行为特征"和"预测网络未来的发展趋势",进行应用研究与验证分析. 注意到,前一个模型功能指的是模型的解释力,也即能够揭示哪些用户的行为特征会影响网络的演化; 后一个模型功能指的是模型的预测力,也即揭示所建立的模型在多大程度上能够准确预测网络演化的趋势.

Facebook 是全球知名的社交媒体平台,根据 4.2 节的发现, 本文的模型适用于成熟和高效的社会媒体平台,而 Facebook 恰好满足了这一属性. 进一步,本节选取的数据来源于 http://snap.stanford.edu/data/egonets-facebook.html,该数据集包含了 Facebook 上的 10 个个体网络的数据,也即: 包含该个体的全部朋友,其朋友之间形成的链接关系和每个个体的属性信息[19],该数据中全部的节点都进行了匿名化处理. 注意到个体网络是一个好的研究单位,其提供了研究的边界,并且包含了以个体为中心的全部朋友及他们之间关系的信息,这比在 Facebook 上随机抽样节点获得的样本更具分析的价值,因为不会遗漏主要的网络结构信息. 以其中的包含 347 个节点和 5 038 条边的个体网络为例,其网络结构情况如图 2 所示.

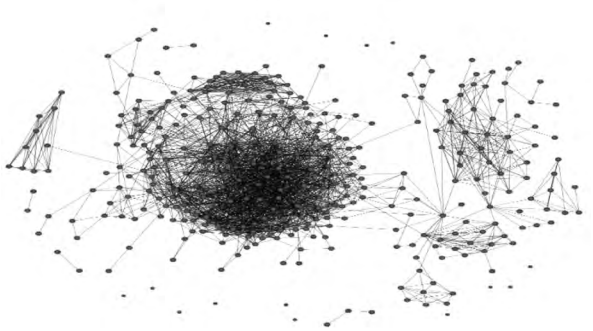


图 2　所选取数据集的网络图示

Fig. 2 Graphical representation of the selected dataset

5.1　模型在解释能力上的应用研究

以图 2 中所示的网络数据以及数据集中包含的个体属性数据为输入,采用式(2)所示的效用函数,运行本文提出的模型,经历 10 000 次迭代,参数更新的过程如下图 3 所示. 经由图 3 可以发现: 在运行到 2 400 次迭代的时候,模型估计的参数值发生了一次跳跃,而后呈现出平稳分布的形式,可以认为模型的参数达到了稳态分布,进一步截取 3 000 次迭代以后的数据进行参数的统计分析,结果在表 6 中列出.

表 6 的结果指出了在 Facebook 平台上驱动网络演化的因素: 参数 θ_1 表示当两个个体的属性差异越小时,他们越容易形成链接; 类似地,参数 θ_3 表示当目标个体的朋友与自身属性差异较小时,越容易和目标个体形成链接,但是这个作用的强度远远小于 θ_1 体现的直接作用; 参数 θ_2 表示当共同朋友数越多时,个体对越容易形成链接,并且这个效用的强度很显著. 这里所展示的结果是应用本模型进行解释网络演化的一个实例,一定程度上体现了模型的解释能力.

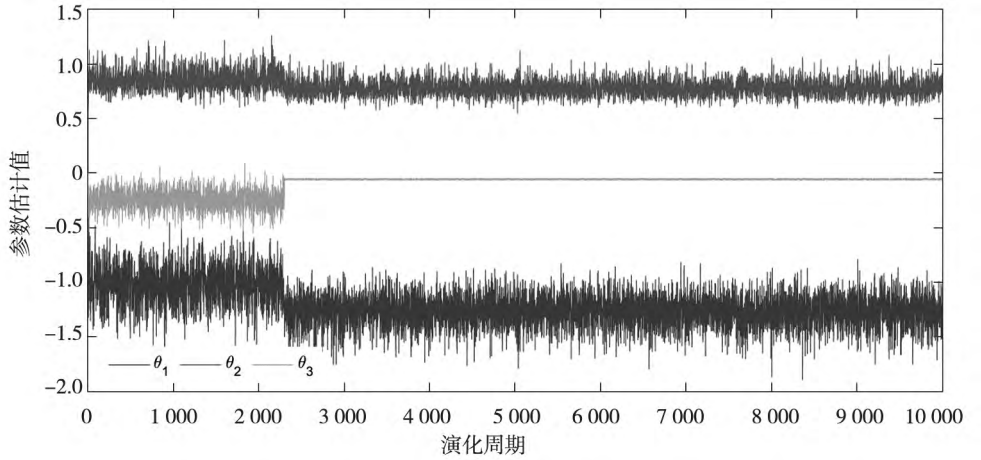


图 3 参数估计值的迭代更新过程

Fig. 3 The updating process of parameter estimates

表 6 基于迭代过程后 7 000 个参数估计值的统计分析

Table 6 Statistics of parameter estimates during the last 7 000 iterations

参数	含义	均值	标准差	95% 置信区间		中位数
				下界	上界	
θ_1	直接朋友的效应	−1.265 3	0.137 4	−1.546 8	−1.000 6	−1.260 4
θ_2	共同朋友数的效应	0.777 7	0.070 5	0.653 0	0.930 3	0.773 9
θ_3	距离为 2 的朋友的间接影响效应	−0.057 2	0.006 4	−0.067 6	−0.046 8	−0.057 3

5.2 模型在预测能力上的应用研究

在表 6 所列出的参数估计值的基础上,将其代入式(3a) 和式(3b) 中,对于任意的个体对 i 和 j ($i \neq j$),可以得到如下校准的效用函数

$$\Delta U_{i\leftarrow j}(\boldsymbol{g}^t, \boldsymbol{C}; \boldsymbol{\theta}) = -1.27 \times \| \boldsymbol{C}_i - \boldsymbol{C}_j \| + 0.78 \times \sum_{l \neq i,j} g_{il}^{t-1} g_{jl}^{t-1} - 0.06 \times \sum_{l \neq i,j} g_{jl}^{t-1} \| \boldsymbol{C}_i - \boldsymbol{C}_l \| \quad (18a)$$

$$\Delta U_{j\leftarrow i}(\boldsymbol{g}^t, \boldsymbol{C}; \boldsymbol{\theta}) = -1.27 \times \| \boldsymbol{C}_i - \boldsymbol{C}_j \| + 0.78 \times \sum_{l \neq i,j} g_{il}^{t-1} g_{jl}^{t-1} - 0.06 \times \sum_{l \neq i,j} g_{il}^{t-1} \| \boldsymbol{C}_j - \boldsymbol{C}_l \| \quad (18b)$$

由此,将这对具体的效用增益函数应用于数据集中标号为 2 的个体网络. 预测从空网络(记为 g^0) 开始,在第 t 轮迭代中,所有的节点对根据 g^{t-1} 的结构信息,结合属性值,代入式(18) 计算出效用增益,而后根据式(13) 和式(14) 计算节点对形成或断开链接的概率,进而得到 g^t 的网络结构,如此反复进行. 注意到第 t 轮迭代的结果仅与第 $t-1$ 轮相关,以上的迭代过程是一个马尔可夫过程; 根据马氏链抽样的基本方法,当迭代次数达到一定规模后(这里取 10 000) ,间隔地进行抽样(这里每隔 50 个网络进行一次抽样,总共抽取 1 000个) ,对以上抽样得到的网络进行平均,算得每个节点对形成链接的概率. 由此,根据真实的标号为 2 的个体网络对预测的结果进行评价,这里用 ROC 曲线及其 AUC 的值②进行预测结果的评价.

② 具体含义可参见 http://baike.so.com/doc/5499846-5737282.html

以上的预测过程事实上实现了动态网络背景下链路预测的功能,为了验证所提出模型预测功能的有效性,选取一些常用的链路预测方法作为基准进行对比分析. 为此,选取 Common Neighbour 法(简记为 CN),Jaccard 指数法(JI),Hub Promoted 指数法(HPI),Resource Allocation 指数法(RAI),Katz 指数法(KI),Adamic-Adar 指数法(AAI),SimRank 方法(SR)和 Preferential Attachment 指数法(PAI)等 8 种方法作为基准方法[20],同时将本文提出的方法简记为 UA,其 ROC 曲线及 AUC 值分别呈现在图 4 和表 7 中.

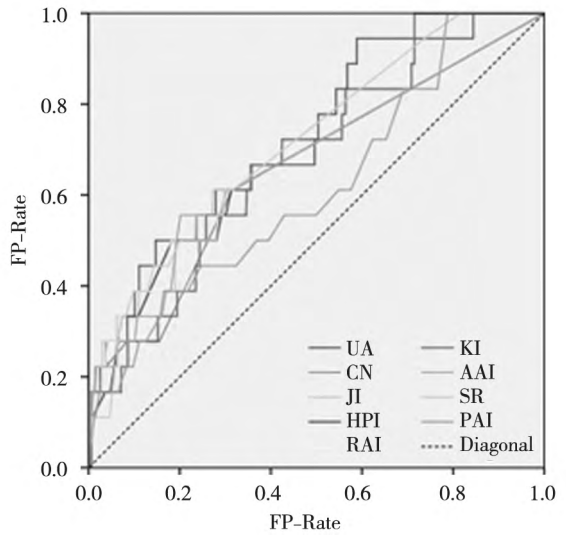


图 4　方法间的 ROC 曲线图

Fig. 4 ROC curves of these methods

表 7　方法间的 AUC 值

Table 7 AUCs of these methods

方法简写	AUC 值
UA	0.726
KI	0.687
CN	0.661
AAI	0.676
JI	0.682
SR	0.713
HPI	0.675
PAI	0.617
RAI	0.685

读图 4 和表 7 可见: 本文方法在预测的准确性上要优于所选取的基准方法,在一定程度上例证了模型在预测方面的有效性,特别是在所研究的具体实际问题上的实用性. 通过以上基于实际数据的分析,可以发现本文所提出的模型不仅具有解释的能力,还具有预测的能力,能够以较高的精度推断动态网络演化的状态,并在所列举的实例上,优于常用的基准方法.

6　结束语

本文提出了基于效用函数的社会网络演化分析模型,将决策科学中效用分析的方法引入社会网络演化问题的研究中,有助于解释驱动社会网络演化的个体行为原因并预测网络未来的演化趋势,进而为优化社会网络的结构打下模型基础,从属于信息管理、决策科学和社会网络分析领域的交叉研究. 概括一下,本文有如下工作和主要贡献: 1) 相比于既有的将社会网络演化视为黑箱的研究而言,本文提出了应用指向型的效用函数,可以根据研究目标和研究背景,设计灵活多样的效用函数,以期突出网络个体的决策能力; 2) 将网络参与者的见面过程纳入模型的分析,将其视为模型不易观察到的潜变量,并且建立了灵活的见面机制模型,这有助于完善网络演化分析模型,弥补了只用决策分析解释网络演化的不足; 3) 在证明所推断的后验概率正确性的基础上,提出了以马尔可夫链蒙特卡洛抽样方法和贝叶斯推断为基础的参数估计算法,该算法适用于含有潜变量的推断问题,可以较为准确地估计出效用函数中参数的具体值,使得模型不仅仅是一个理论模型,还具有了实用价值.

本文还设计了一系列的仿真分析实验用以验证模型估计参数的准确性. 对于社交媒体平台而言,其内在的两个机制是平台上交流的效率和平台运营的时间. 仿真分析通过两个参数: 见面概率和演化时长,与社交媒体平台这两个内在属性相对应,通过调控这两个参数的值,验证本文提出的模型在不同的社交媒体平台上的适用性. 结论表明: 1) 模型的表现受到这两个参数的影响,当提高见面概率(也即平台交流的效率)或增加演化时长(也即平台运行时间的长短)的情况下,模型能够给出更加准确的参数估计结果; 2) 本文提出

的模型具有一定的鲁棒性，当两组参数值较小时，也基本能够给出参数正确的符号估计. 同时，依托于 Facebook 平台上真实的数据集，进行了模型的应用研究，例证了模型在解释功能和预测功能两个方面的实用性.

限于本文的篇幅及研究重点，进一步的研究可能包含以下几个方面: 1) 估计算法的局部收敛性问题. 本文提出的估计算法可能会收敛于局部极小值，而非全局最小值，因此，估计算法虽然具有一定的准确性，但是依然有进一步发展和改进的空间; 2) 同类方法的对比分析问题. 本文模型重点介绍了新方法提出和验证的过程，在进一步的研究中，可以增加与同类方法的对比研究，具体分析各种方法的优缺点和适用范围，给出一个较为全面的模型使用建议; 3) 模型的衍生应用问题. 本文重点关注了效用函数中参数的估计值及其管理含义，事实上，基于估计出的效用函数值和参与者见面过程，可以在动态的模型框架下，对模型未来的演化结果做依托于参数调控的控制分析，以期深入地应用模型，获得更多的管理洞见.

参 考 文 献:

[1] 杨善林，周开乐. 大数据中的管理问题: 基于大数据的资源观 [J]. 管理科学学报，2015，18(5): 1 - 8.
Yang Shanlin，Zhou Kaile. Management issues in big data: The resource-based view of big data [J]. Journal of Management Sciences in China，2015，18(5): 1 - 8. (in Chinese)

[2] Handcock M S，Gile K J. Modeling social networks from sampled data [J]. The Annals of Applied Statistics，2010，4(1): 5 - 25.

[3] Fan W，Gordon M D. The power of social media analytics [J]. Communications of the ACM，2014，57(6): 74 - 81.

[4] Stieglitz S，Dang-Xuan L. Emotions and information diffusion in social media: Sentiment of microblogs and sharing behavior [J]. Journal of Management Information Systems，2013，29(4): 217 - 248.

[5] Lin H，Fan W，Chau P Y K. Determinants of users' continuance of social networking sites: A self-regulation perspective [J]. Information & Management，2014，51(5): 595 - 603.

[6] Currarini S，Jackson M O，Pin P. Identifying the roles of race-based choice and chance in high school friendship network formation [J]. Proceedings of the National Academy of Sciences，2010，107(11): 4857 - 4861.

[7] Currarini S，Jackson M O，Pin P. An economic model of friendship: Homophily，minorities，and segregation [J]. Econometrica，2009，77(4): 1003 - 1045.

[8] Jackson M O. Social and Economic Networks [M]. New Jersey: Princeton University Press，2010.

[9] Robins G，Bates L，Pattison P. Network governance and environmental management: Conflict and cooperation [J]. Public Administration，2011，89(4): 1293 - 1313.

[10] Quintane E，Conaldi G，Tonellato M，et al. Modeling relational events: A case study on an open source software project [J]. Organizational Research Methods，2014，17(1): 23 - 50.

[11] Yang D H，Yu G. Static analysis and exponential random graph modelling for micro-blog network [J]. Journal of Information Science，2015，40(1): 3 - 14.

[12] 王越乙，徐枞巍. 指数随机图 (p*) 模型不同描述的对比研究 [J]. 清华大学学报: 自然科学版，2015，55(4): 422 - 427.
Wang Yueyi，Xu Congwei. Comparative analysis of different descriptions in exponential random graph models (P*) [J]. Journal of Tsinghua University (Science and Technology)，2015，55(4): 422 - 427. (in Chinese)

[13] Shalizi C R，Rinaldo A. Consistency under sampling of exponential random graph models [J]. The Annals of Statistics，2013，41(2): 508 - 535.

[14] Chatterjee S，Diaconis P，Sly A. Random graphs with a given degree sequence [J]. The Annals of Applied Probability，2011，21(4): 1400 - 1435.

[15] Christakis N A，Fowler J H，Imbens G W，et al. An Empirical Model for Strategic Network Formation [R]. London: Na-

tional Bureau of Economic Research, Working Paper, 2010.

[16] Mele A. A Structural Model of Segregation in Social Networks [R]. London: Cemmap Working Paper, 2013.

[17] 李永立, 吴　冲, 张晓飞. 考虑网络交互影响效应的评价者权重分配方法 [J]. 管理科学学报, 2016, 19(4): 32 – 44.
Li Yongli, Wu Chong, Zhang Xiaofei. An evaluator's weight allocation considering network peer effect [J]. Journal of Management Sciences in China, 2016, 19(4): 32 – 44. (in Chinese)

[18] 李永立, 罗　鹏, 张书瑞. 基于决策分析的社交网络链路预测方法 [J]. 管理科学学报, 2017, 20(1): 64 – 76.
Li Yongli, Luo Peng, Zhang Shurui. Link prediction in social networks based on decision analysis [J]. Journal of Management Sciences in China, 2017, 20(1): 64 – 76. (in Chinese)

[19] Mcauley J, Leskovec J. Discovering social circles in ego networks [J]. ACM Transactions on Knowledge Discovery from Data, 2013, 8(1): 1 – 28.

[20] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks [J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019 – 1031.

Utility-based model for interpreting evolution patterns of social networks

LI Yong-li, CHEN Yang, FAN Ning-yuan, GAO Xin

School of Business Administration, Northeastern University, Shenyang 110169, China

Abstract: Existing models often uncovered the evolution patterns via statistical analysis, which would be unable to explain micro behavior reasons driving the social network evolution. To make up the deficiency, a utility function of network individuals is established and a utility analysis is introduced to model the social network evolution. Meanwhile, the meeting sequence, embedded in the social network evolution, is further modeled as a latent variable in order to explain the evolution phenomenon that the mentioned utility analysis cannot explain. Subsequently, taking one-period observation of social network structure and individual attributes as the input, a Bayes-inference-based method is developed for estimating the preference parameters and the latent meeting states. Through two groups of simulation analysis, the accuracy of parameter estimation and the applicable scope are verified, and the proposed model is also applied to validate its explanatory power and predictive force on the collected real data from Facebook platform. In all, the proposed model will be helpful to explain how social network forms on social media platforms and also to predict the tendency of social network evolution, so that it can lay a foundation for achieving the expected network structure and further controlling the information spreading within social networks.

Key words: network evolution; social network analysis; utility analysis; Bayesian inference; social media platform