基于决策分析的社交网络链路预测方法①

李永立[1]，罗　鹏[2]，张书瑞[3]
(1. 东北大学工商管理学院，沈阳 110169; 2. 哈尔滨工业大学管理学院，哈尔滨 150001;
3. 东北大学信息科学与工程学院，沈阳 110169)

摘要: 社交网络是社会媒体信息传播的骨架，对其进行链路预测的研究将有助于社会媒体平台上的信息管理和舆论控制. 在既有网络链路预测方法研究的基础上，以决策分析的思想为出发点，提出了引入效用函数分析的社交网络链路预测方法; 针对效用函数中参数的估计问题，进一步提出了允许一定误差度的马尔科夫链蒙特卡洛参数校准方法，并对方法的正确性进行理论上的论证. 在收集到的 5 个腾讯 QQ 群的数据集上，进行了新方法的验证研究，并与既有的链路预测方法在准确性方面进行了比较分析. 研究表明: 本文提出的预测方法考虑了链路形成的微观行为基础，具有较好的预测准确性，并且参数估计算法中“允许误差”的引入有助于模型应用者在模型效率和准确性方面的折衷中做出合理的决策.

关键词: 链路预测; 决策分析; 效用函数; 马尔科夫链蒙特卡洛方法; 社交网络; 数据分析
中图分类号: TP391; N949　**文献标识码**: A　**文章编号**: 1007 - 9807(2017)01 - 0064 - 11

0　引　言

网络构成了社会与经济生活的骨架，大到全球视角的贸易网络，小到一个单位的人际网络，都是“网络”这一组织形式在现实生活中的具体体现. 特别是随着社会媒体平台的高速发展，人们日常联系的方式变得更加多样，新兴的微信、微博、腾讯 QQ 等联系方式层出不穷，由此形成的社交网络的规模也呈现超速发展的态势，成为信息管理领域一个热点的研究方向[1-3]. 在这一研究背景下，本文将重点研究如下的问题“在分析社交网络既有链接形成规律的基础上，如何预测潜在的链接，进而为研究社交网络形成和演化的规律奠定微观基础”.

本文的研究问题属于网络链路预测的研究领域. 具体地说，网络链路预测是通过分析网络现有的链接规律，从而探究和预测网络未知或潜在的链接. 网络链路预测作为网络研究领域较为热门的研究方向之一，已经有了深厚的研究基础，较为早期的综述可以参见 Lü 等[4] 发表在 Physica A 上的文献. 本文进一步从研究方法的角度，对该领域既有的研究进行一个简单的概述: 一方面便于文章的整体性和可读性，另一方面为进一步的对比研究提供方法的索引. 本文将既有的网络链路预测方法分为了三类，分别为“基于节点相似性的链路预测”、“基于最大似然估计的链路预测”和“基于概率模型的链路预测”. 其中，(1)“基于节点相似性的链路预测”以两个节点之间相似性越大，它们之间存在链接的可能性就越大为研究前提，从度量节点相似性的角度出发进行网络链路预测的研究; 代表性的研究有“基于局部信息相似性指标的链路预测方法”[5-6]，“基于全局相似性指标的链路预测方法”[7-9]，以及“基于局部 - 全局相似性指标的链路预测方法”[10-12]. (2)“基

① 收稿日期: 2015 - 05 - 17; 修订日期: 2016 - 02 - 05.
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项目(2016M590230); 辽宁省财政科研基金项目(16C024).
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

于最大似然估计的链路预测”方法主要以“层次结构模型”[13]和“随机分块模型”[14-15]为典型代表. 具体地, “层次结构模型”将任意网络结构做树形网络结构的近似,应用最大似然估计的方法,得到最接近于目标网络的树形网络结构,进而进行预测和分析的研究; “随机分块模型”根据网络具有模块性的特点,应用最大似然方法将网络的节点分组,以节点所在组的链接形成规律为基础进行链路预测. (3) “基于概率模型的链路预测”的基本思路是建立一个含有一组可调参数的模型,然后使用优化策略寻找最优的参数值,使得所得到的模型能够更好地再现真实网络的结构和关系特征,进而以求出的最优参数估计值为基础得到潜在链路形成的概率; 这一方法的代表性研究有贝叶斯网络模型[16]、马尔科夫网络模型[17]、关系依赖网络模型[18,19]、有向无环概率实体关系模型[20]、随机关系模型[21]等等.

特别地,针对本文所具体研究的社交网络上的链路预测问题,考虑到社交网络的时间属性、节点的差异化以及结构的多样性,一些针对性较强的方法陆续被提出; 比如: Huang 等通过改进核投影机(Kernel Projection Machine, KPM) 的算法,提出了一种名为截断 KPM 社会网络链路预测的算法[22]; 田儒雅研究了领袖引导模型,根据该模型分析社交舆论网络的演化以及网络链路预测等一系列问题[23]; 李倩倩等研究社交网络中的用户行为,从而分析社交网络中节点的链路行为[24]; 以及基于网络链路预测建立的网络演化模型[25-27]等方面的研究.

纵观以上列举的研究, “基于节点相似性的链路预测”方法只涉及到网络的结构信息,虽然方法比较简单,计算复杂度相对较低,但是受制于“节点相似性影响链路形成概率”的研究假设,该方法并不通用,在不同网络中的预测能力差异很大. “基于最大似然估计方法的链路预测”方法针对的是网络整体结构的分析,该类方法需要计算似然函数,往往具有较高的计算复杂度,不适用于规模较大的网络. “基于概率模型的链路预测”方法基于机器学习的思想,从参数学习的角度建立模型,是近五年来研究链路预测的趋势之一. 本文的模型从属于概率模型的研究范畴,但在建模的思想上将不同于既有的模型: 本文以决策分析的理论方法为着眼点,将效用函数引入链路预测的分析,关注了网络宏观结构形成的微观基础,并以已知的网络结构为模型的输入对效用函数的参数进行校准,进行完成链路预测的目标.

针对以上的分析,本文的主要工作及贡献如下: (1) 将考虑效用函数的决策分析方法引入到社交网络链路预测的问题中,赋予了网络参与者具有智能属性的内涵; (2) 将效用函数中的参数估计问题视为一个机器学习问题,发展了既有的马尔科夫链蒙特卡洛方法、并加以创造应用到参数估计的问题解决中,论证了所提出算法的正确性; (3) 收集并利用“腾讯 QQ 群”上的社交网络数据,在 5 个随机选择的 QQ 群社交网络结构上应用和验证本文的模型,以期验证所提出模型的实用性.

1　模型建立

1.1　模型建立的基本思想

本文的模型立足于决策分析的思想,将社交网络参与者的效用函数引入对网络链路形成的分析,认为网络参与者是有限理性的个体,在形成社交网络的链接时,根据自身的效用情况来决定链接的形成; 也即: 如果建立链接有助于自身效用的提升,则建立相应的链接,否则,不建立链接. 具体地,以社交网络中代表性个体 i 为例,记效用函数为

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

其中矩阵 $\boldsymbol{C}$ 由全体网络参与者的属性向量组成(其形式为 $\boldsymbol{C}=[\boldsymbol{C}^1;\boldsymbol{C}^2;\cdots;\boldsymbol{C}^i;L;\boldsymbol{C}^{\Omega}]$,这里 $\boldsymbol{\Omega}$ 为网络参与者的数量) ; 以代表性个体 i 的属性向量 $\boldsymbol{C}^i$ 为例,可以包含诸如年龄、职业、学历、爱好等等方面的信息,其第 l 个分量记为 c_l^i; 矩阵 $\boldsymbol{g}$ 是代表性个体 i 做决策时网络的邻接矩阵,具体地,邻接矩阵 $\boldsymbol{g}$ 中的第 m 行和第 n 列($m\neq n$) 的元素记为 g_{mn},反映了网络中第 m 个个体和第 n 个个体建立链接的情况,如果已经建立了链接,则 $g_{mn}=1$,否则 $g_{mn}=0$; 特别地,当 $m=n$ 时(也即邻接矩阵对角线上的元素) ,根据邻接矩阵的定义,设定为 0. 此外,效用函数中的 $\boldsymbol{\theta}$ 是未知的参

数,是模型求解的对象. 接下来分析两种情形:

情形 1 如果个体 i 与个体 j 还未形成链接,也即在决策时,邻接矩阵 g 中的元素 $g_{ij}=0$.

这时涉及两个代表性个体是否要建立链接的问题,这一情形对应于新链接的建立. 注意到,在个体 i 的决策过程中,其会考察当与个体 j 建立连接时,自身效用的变化情况. 这里效用的变化记为 $\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})$,具有如下的量化表达

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

式中 $\boldsymbol{g}+[\boldsymbol{g}_{ij}=1]$ 表示在现有邻接矩阵的基础上,增加了个体 i 和 j 之间的链接;同理,对于代表性个体 j,表达如下

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

注意到两个公式中,效用函数下标是不同的. 如果两个决策个体都仅以自身的效用变化为建立链接的决策依据,则两者建立链接的充要条件为

$$\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0 \text{ 并且 } \Delta U_{j\to i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})\geqslant 0 \tag{4}$$

也即:当两者建立链接时,如果效用都增加了,链接(在社交网络背景中,也可称为“朋友关系”)才能形成;否则,链接无法形成. 注意到这里定义的朋友关系是相互的,不能单方面形成.

情形 2 如果个体 i 与个体 j 已经形成链接,也即在决策时,邻接矩阵 g 中的元素 $g_{ij}=1$.

这时涉及两个代表性个体是否要终止链接的问题,这一情形对应于旧链接的断裂,这一情形也是在现实生活中经常出现的,比如在微信中删除某朋友的行为. 这时,针对个体 i,其在决策过程中考虑的问题是:当终止与个体 j 的链接时,其效用是否会增加;为此,首先需要度量效用函数的变化,其量化表达为

$$\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) = U_i(\boldsymbol{g}-[\boldsymbol{g}_{ij}=1],\boldsymbol{C};\boldsymbol{\theta}) - U_i(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) \tag{5}$$

式中 $\boldsymbol{g}-[\boldsymbol{g}_{ij}=1]$ 表示在现有邻接矩阵的基础上,断开个体 i 和 j 之间的链接;同理,对于代表性个体 j,断开两者链接后,其效用的变化为

$$\Delta U_{j\to i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) = U_j(\boldsymbol{g}-[\boldsymbol{g}_{ij}=1],\boldsymbol{C};\boldsymbol{\theta}) - U_j(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) \tag{6}$$

注意到,这时如果两个个体仅以自身的效用变化作为决策的依据,则断开链接的充要条件为

$$\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})>0 \text{ 或 } \Delta U_{j\to i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})>0 \tag{7}$$

这意味着当断开链接时,若两个代表性个体中有一个的效用增加,则他们之间的链接(或朋友关系)将不再继续保持;否则,他们之间的链接依然继续存在.

1.2 考虑网络结构效应的效用函数

在上一节分析的基础上,本节考虑效用函数的具体形式,这是模型定量化的基础. 首先,在社交网络的背景下,Christakis 和 Fowler 所著的《Connected》一书提出如下的观点 “三度影响力原则是社会网络上决定网络参与者行为的一个规律,即:我们所做的事情会影响到我们的朋友(一度影响力),我们朋友的朋友(二度影响力),和我们朋友的朋友的朋友(三度影响力)”[28]. 根据这一观点,本文设计如下形式的效用函数

$$U_i(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) = \theta_1\cdot f_1(\boldsymbol{C}^i)+\theta_2\sum_{k\neq i}\boldsymbol{g}_{ik}f_2(\boldsymbol{C}^k) + \theta_3\sum_k \boldsymbol{g}_{ik}\sum_{s\neq i,k}\boldsymbol{g}_{ks}f_3(\boldsymbol{C}^s) + \theta_4\sum_k \boldsymbol{g}_{ik}\sum_{s\neq i,k}\boldsymbol{g}_{ks}\sum_{t\neq i,k,s}\boldsymbol{g}_{st}f_4(\boldsymbol{C}^t)+\varepsilon_i \tag{8}$$

其中 $\theta_1,\theta_2,\theta_3$ 和 θ_4 是参数;函数 f_1,f_2,f_3 和 f_4 分别对应网络中的常数,一度影响力,二度影响力和三度影响力,并受相应的网络参与者的属性影响;随机干扰项 ε_i 反映影响代表性个体 i 决策时没有被模型解释的随机因素,不失一般性,假定其满足均值和方差都未知的正态分布,也即 $\varepsilon_i\sim N(\mu,\sigma^2)$. 以下根据个体 i 与个体 j 是否已经建立链接的两种情形分别进行讨论:

情形 1 如果个体 i 与个体 j 还未形成链接,则结合式(8)和式(2),可知在“个体 i 是否与个体 j 建立链接”的决策过程中,个体 i 考虑如下的效用变化问题

$$\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) = \theta_2 f_2(\boldsymbol{C}^j)+\theta_3\sum_{s\neq i,j}\boldsymbol{g}_{js}f_3(\boldsymbol{C}^s) + \theta_4\sum_{s\neq i,j}\boldsymbol{g}_{js}\sum_{t\neq i,j,s}\boldsymbol{g}_{st}f_4(\boldsymbol{C}^t) \tag{9a}$$

情形 2 如果个体 i 与个体 j 已经形成链接,则

结合式(8)和式(2),可知在"个体 i 是否与个体 j 断开链接"的决策过程中,个体 i 考虑如下的效用变化问题

$$\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) = -\theta_2 f_2(\boldsymbol{C}^j) - \theta_3 \sum_{s\neq i,j} g_{js} f_3(\boldsymbol{C}^s) - \theta_4 \sum_{s\neq i,j} g_{js} \sum_{t\neq i,j,s} g_{st} f_4(\boldsymbol{C}^t) \tag{9b}$$

注意到式(9a)刻画的效用变化与式(9b)刻画的效用变化恰好互为相反数. 这一性质结合式(4)和式(7),并根据原命题和逆否命题相互等价的事实,推知: 情形一和情形二的充要条件是一致的. 类似地可以得到代表性个体 j 的效用变化情况.

进一步,令模型中的待定函数 f_2,f_3 和 f_4 取为 Bonacich 中心性,以 f_2 和代表性个体 j 为例,具体定义为

$$f_2(\boldsymbol{C}^j) = [\boldsymbol{I} - \gamma \boldsymbol{g}]^{-1} \cdot \boldsymbol{1}_j \tag{10}$$

其中 $\boldsymbol{I}$ 是单位矩阵,$\boldsymbol{1}_j$ 是第 j 个元素为 1、其余元素为 0 的列向量,β 是 Bonacich 中心性中设置的一个参数,反映网络中邻居的影响强度,其取值要求定义的矩阵求逆有意义,也即: β 的取值范围为 $\gamma \in (0, 1/\max(eigenvalue(\boldsymbol{g})))$,这里 $\max(eigenvalue(\boldsymbol{g}))$ 是矩阵 $\boldsymbol{g}$ 的最大特征根. 关于参数 γ 的性质和取值技巧,可以进一步参考 Li, et al. (2014) 的论述[19].

本文从网络参与者建立链接是否有利于其效用增加的角度出发,将建立链接的行为过程视为一个决策的过程,其不仅考虑了网络参与者的三度影响力,还考虑了网络参与者的位置属性,以及涉及了决策的随机干扰问题,考虑了决策问题的不确定性. 注意到,以上给出的效用函数是模型校准参数的形式基础,其将待确定的参数 θ_2,θ_3 和 θ_4,以及随机干扰项中的 μ 和 σ 视为随机变量,以下的模型校准过程将应用马尔科夫链蒙特卡洛方法(MCMC)进行参数的校准.

1.3　社交网络链路预测的步骤

基于以上设计的效用函数及已知的网络结构 $\boldsymbol{g}$,针对两个代表性个体 i 和 j,及其所在的网络的结构信息,预测他们之间是否建立链接的步骤依次为

步骤 1　通过已经获得的网络结构矩阵 $\boldsymbol{g}$,校准代表性个体效用变化中的参数 θ_2,θ_3 和 θ_4;

步骤 2　分别计算出两个代表性个体各自的 Bonacich 中心性值;

步骤 3　将以上数据分别获得两个代表性个体在决策过程中效用变化的情况;

步骤 4　若算出的 $\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) \geqslant 0$ 并且 $\Delta U_{j\to i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta}) \geqslant 0$,则建立链接或已建立的链接继续保持; 否则不建立链接或已建立的链接断开.

注意到以上步骤中,第一步的"参数校准"是后续步骤进行的前提,也是本文模型算法的核心部分,将在下面详述参数校准的过程.

2　参数校准

2.1　参数校准算法的流程

对本文的参数校准而言,基于线性回归假设的最小二乘法将不再适用,因为从现实中可以获得的数据来看,用于反映效用变化的被解释变量 $\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})$ 和 $\Delta U_{j\to i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})$ 的数据是无法直接获得的,而从网络链接的情况,只能推断出 $\Delta U_{i\to j}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})$ 和 $\Delta U_{j\to i}(\boldsymbol{g},\boldsymbol{C};\boldsymbol{\theta})$ 的符号. 不仅如此,现实中的网络往往具有较大的规模,模型校准计算的算法复杂度不能过高. 基于以上两点,本文拟将 MCMC 方法的思想应用于本文模型的参数校准,具体地算法流程如下:

根据以上的算法可以获得待估计参数取值的一个集合,通过对该集合的分析,可以进一步获得待估计参数的均值、中位数、方差等分布情况的统计信息. 在以上的算法流程中,可以发现算法的整体复杂度主要由第一步控制,其中计算各个节点的 Bonacich 中心性的复杂度为 $O(n^{2.373})$,而后面循环过程的复杂度为 $O(n^2)$,因此整个的算法复杂度为 $O(n^{2.373})$,其在目前计算机的计算能力下,可以适应于较大规模的样本数据集.

表 1 模型参数校准算法的流程表

Table 1 Procedure of parameter calibration algorithm

步骤	具体操作	解释及注释
第 1 步	给出每个参数的初始概率函数 π 和初始值,以及转移概率函数 q;根据式(10)计算每个网络参与个体的 Bonacich 中心性(本文中,γ 取值为 $1/\Omega$,这里 Ω 是网络的参与人数)	本文涉及的参数有 5 个,分别为 θ_2,θ_3 和 θ_4,以及 μ 和 σ;转移概率表示为 $q(\alpha\rightarrow\beta)$,其中 α 和 β 分别表示某参数转移前后的值
第 2 步	根据此刻参数的值 $\boldsymbol{\theta}=(\theta_2,\theta_3,\theta_4,\mu,\sigma)$ 和给定的每个参数的转移概率函数 q,生成参数的候选值 $\boldsymbol{\theta}'=(\theta_2',\theta_3',\theta_4',\mu',\sigma')$	这一步实现参数的更新过程,由参数此刻的值和各自的转移概率生成新的候选值,实现参数的抽样
第 3 步	将候选参数的值代入本文的式(9)中,并根据式(4)和式(7)给定的链路形成与断开的规则,可以判定每两个个体之间的链接情况,进而得到基于候选参数的一个新的社交网络,其邻接矩阵记为 g'	以代表性个体 i 和 j 为例,把候选参数代入式(9),其各自的效用函数的增量可以定量地获得,由此根据式(4)和式(7)给出的规则,可以确定任意节点对的链接情况;得到基于候选参数的社交网络
第 4 步	给出距离的度量 ρ,如果 $\rho(g,g')\leqslant\Delta$,则进行第五步,否则返回第二步重新抽样	距离的度量用来刻画基于候选参数的社交网络与真实的社交网络之间的差距.
第 5 步	计算: $p(\boldsymbol{\theta},\boldsymbol{\theta}')=\min\left(1,\frac{\pi(\boldsymbol{\theta}')\,q(\boldsymbol{\theta}'\rightarrow\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})\,q(\boldsymbol{\theta}\rightarrow\boldsymbol{\theta}')}\right)$	由于各个参数是独立取值的,成立 $\pi(\boldsymbol{\theta})=\pi(\theta_2)\,\pi(\theta_3)\,\pi(\theta_4)\,\pi(\mu)\,\pi(\sigma)$,类似地,推得 $\pi(\boldsymbol{\theta}')$、$q(\boldsymbol{\theta}\rightarrow\boldsymbol{\theta}')$ 和 $q(\boldsymbol{\theta}'\rightarrow\boldsymbol{\theta})$ 的表达
第 6 步	以 $p(\boldsymbol{\theta},\boldsymbol{\theta}')$ 的概率接受 $\boldsymbol{\theta}'$,否则参数向量仍保持为 $\boldsymbol{\theta}$,返回第二步重新抽样,直到达到抽样的规模为止	这里得到的参数集合被认为来自于参数真实分布的抽样,由此基于得到的参数集合对其真实分布进行推断

2.2 算法正确性的论证

以下论证根据表 1 的算法得到的参数集合是来自于参数真实分布函数的一个抽样. 为此,令参数向量真实的分布函数为 $f(\boldsymbol{\theta}\mid g)$,则根据贝叶斯公式,有

$$f(\boldsymbol{\theta}\mid g)=\frac{p(g\mid\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{p(g)} \tag{11}$$

式中 $p(g\mid\boldsymbol{\theta})$ 表示在参数向量为 $\boldsymbol{\theta}$ 的条件下,观察到的社交网络为 g 的概率;而 $p(g)$ 表示社交网络呈现为 g 的概率,是贝叶斯公式的一个正则化系数,形式上有 $p(g)=\int p(g\mid\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$. 本质上,如果能够直接得到参数向量的概率分布,也即得到了 $f(\boldsymbol{\theta}\mid g)$,则参数的校准和估计问题将变得非常直接. 但是,这个概率通常很难获得,而本文算法的思想在于获得一个符合这个真实概率函数的参数样本集合,应用集合的数据对参数进行校准和估计.

考虑到抽样的效率问题,根据候选的参数向量,想要得到与真实网络一模一样的网络往往是非常困难的,这里的算法允许得到的网络结构 g' 与真实的结构 g 存在不超过 Δ 的偏差,由此在对候选参数取舍的判断上,没有严格地要求 $g'=g$,而是采用了 $\rho(g,g')\leqslant\Delta$ 的标准,由此本文中参数的校准算法事实上获得的是以 $f(\boldsymbol{\theta}\mid\rho(g,g')\leqslant\Delta)$ 为概率分布的一个参数抽样集合. 虽然这样做会损失一定的估计精度,但是保证了抽样算法的效率,一定程度上与大数据分析的要求相切合. 基于以上分析,表 1 中算法正确性的论证归结为如下的一个命题:

命题 基于表 1 得到的参数样本集合是来自概率函数 $f(\boldsymbol{\theta}\mid\rho(g,g')\leqslant\Delta)$ 的一个抽样.

证明 表 1 的算法事实上实现了一个马尔科夫过程,具体地,令转移概率为 $tr(\boldsymbol{\theta}\rightarrow\boldsymbol{\theta}')$,根据表 1 的算法,有

$$\begin{aligned}tr(\boldsymbol{\theta}\rightarrow\boldsymbol{\theta}')&=q(\boldsymbol{\theta}\rightarrow\boldsymbol{\theta}')\,p(\rho(g,g')\\&\leqslant\Delta\mid\boldsymbol{\theta}')\,p(\boldsymbol{\theta},\boldsymbol{\theta}')\end{aligned} \tag{12}$$

式中等式右边三项的含义分别为:从参数向量 $\boldsymbol{\theta}$ 转移到候选参数向量 $\boldsymbol{\theta}'$ 的概率,以候选参数向量

θ' 为条件,满足 $\rho(g,g') \leqslant \Delta$ 的概率,和确定接受候选向量 θ' 的概率. 根据表 1 的算法,这三项的乘积是参数向量从 θ 转移到 θ' 的概率. 以下论证: 以上基于马尔科夫的参数抽样过程,其稳定概率恰为 $f(\theta \mid \rho(g,g') \leqslant \Delta)$. 注意到

$$\begin{aligned} & f(\theta \mid \rho(g,g') \leqslant \Delta)\, tr(\theta \to \theta') \\ &= f(\theta \mid \rho(g,g') \leqslant \Delta)\, q(\theta \to \theta')\, p(\rho(g,g') \leqslant \Delta \mid \theta')\, p(\theta,\theta') \\ &= \frac{p(\rho(g,g') \leqslant \Delta \mid \theta)\, \pi(\theta)}{p(\rho(g,g') \leqslant \Delta)} q(\theta \to \theta')\, p(\rho(g,g') \leqslant \Delta \mid \theta')\, \frac{\pi(\theta')\, q(\theta' \to \theta)}{\pi(\theta)\, q(\theta \to \theta')} \\ &= \frac{p(\rho(g,g') \leqslant \Delta \mid \theta')\, \pi(\theta')}{p(\rho(g,g') \leqslant \Delta)} q(\theta' \to \theta)\, p(\rho(g,g') \leqslant \Delta \mid \theta) \\ &= f(\theta' \mid \rho(g,g') \leqslant \Delta)\, q(\theta' \to \theta)\, p(\rho(g,g') \leqslant \Delta \mid \theta)\, p(\theta',\theta) \\ &= f(\theta' \mid \rho(g,g') \leqslant \Delta)\, tr(\theta' \to \theta) \end{aligned}$$

当 $\frac{\pi(\theta')\, q(\theta' \to \theta)}{\pi(\theta)\, q(\theta \to \theta')} > 1$ 时,$p(\theta,\theta') = 1$. 而由于 $p(\theta',\theta) = \frac{\pi(\theta)\, q(\theta \to \theta')}{\pi(\theta')\, q(\theta' \to \theta)}$,以上等式的推导过程倒过来书写即可,不改变等式成立的性质.

注意到,推导过程的最后一个等号意味着满足马尔科夫过程的细致平稳性条件,这意味着表 1 第六步得到的参数集合是来自如下条件概率函数 $f(\theta \mid \rho(g,g') \leqslant \Delta)$ 的一个抽样. 证毕.

2.3　算法的实现细节

对于表 1 中第一步涉及的初始概率和转移概率需要给定. 为此,根据系数的含义,考虑到来自朋友的三度影响力是逐步递减的关系,以及这种影响力具有正向传递的特点,本文给定 θ_2,θ_3 和 θ_4 的初始概率分别满足 $\chi^2(9)$,$\chi^2(5)$ 和 $\chi^2(3)$ 的分布函数. 对于随机干扰项分布中的参数 μ 和 σ,考虑到模型的意义,μ 应为负值,否则网络中所有的个体之间都将形成链接,这与大部分实际网络都不相符; 由此,令 μ 满足 $(-\infty, 0]$ 上的均匀分布,而 σ 满足 $\chi^2(2)$ 的分布函数.

针对转移概率,由于没有先验的知识判断转移概率具体的形式,本文取可等概率转移的概率函数,也即 $q(\theta' \to \theta) = q(\theta \to \theta')$,由此表 1 中第二步的参数抽样则变为参数可行域中的等概率抽样,而第五步的比值简化为初始概率的比值. 基于以上两类给定的概率分布,可以应用表 1 中的算法校准模型.

对于表 1 中涉及的距离度量 $\rho(g,g')$,在本文中的定义如下:

$$\rho(g,g') = \frac{\sum_{i=1}^{\Omega} \sum_{j=1}^{\Omega} \mid g_{ij} - g'_{ij} \mid}{\Omega(\Omega - 1)} \tag{13}$$

其中 Ω 是网络参与者的数量,g_{ij} 和 g'_{ij} 分别为网络 g 和 g 第 i 行第 j 列对应的链接情况; 该式反映了错误预测的链接数量占总链接数量的比例.

3　模型应用及验证

本文以“腾讯 QQ”的群为研究的对象和边界,收集一个群内参与者的链接信息和相关的个体属性信息. 注意到: 尽管人们可能根据自己的兴趣选择加入一个主题群,如“80 后群”、“驴友群”、“股票投资群”等等,但并不是群里所有的成员都会成为朋友(建立的 QQ 朋友关系),通常的情况是部分的成员会成为朋友,由此每个群内的成员又构成了社交网络,图 1 给出了其中一个群的社交网络结构. 之所以选取“QQ 群”为研究对象,是因为两个原因: (1) 研究的边界非常明确,便于数据收集; (2) 有助于模型准确性的验证: 相比于随机抽样的数据收集方法,本文获得的是以“群”为单位的完整的社交网络结构,而不是来自于完整网络的部分链接信息. 由此,本文收集 5 个 QQ 群的社交网络的信息,基于本文提出的方法进行模型的应用研究和交叉验证. 具体地,收集到的 5 个 QQ 群的基本统计信息如表 2 所示,可以发现: 这 5 个被分析的 QQ 群在成员数、链接数和

平均度方面有着较为显著的差别. 特别地,其中编号为 1 的群的社交网络结构如图 1 所示. 从图 1 中可以清楚地发现,图示的编号为 1 的 QQ 群有 5 个人没有与其他人建立朋友关系,并且有着显著的社团结构,形成了三个独立的社团,在不同的社团里,链接的密度也有着较大的差异.

表 2 5 个 QQ 群的基本统计信息

Table 2 Basic statistical information of five QQ groups

群编号	成员数	链接数	平均度	平均路径长度	孤立节点数	聚类系数
1	66	270	8.182	1.927	5	0.758
2	59	146	4.949	2.553	7	0.572
3	227	3 192	28.123	2.523	3	0.557
4	159	1 693	21.296	2.691	9	0.684
5	170	1 656	19.482	2.425	2	0.557

图 1 群编号为 1 的 QQ 群成员的社交网络结构示意图

Fig. 1 The social network structure of QQ members in group 1

根据 2.1 节表 1 中给出的算法步骤,在 Δ 取不同值时得到以上 5 个群各自效用函数的参数统计信息,如表 3 所示. 表 3 中模型准确率的度量采用的是“准确性指标(Accuracy)”,该指标定义为“准确性: = 预测正确的节点对数目 ÷ 总的节点对数目”. 进而,该指标是度量分类器性能的一个重要指标; 诚然,还有其他的度量指标,比如: AUC 指标、precision 指标、recall 指标和 F-measure 指标等等,各指标的关系可参见: http://alexkong.net/2013/06/introduction-to-auc-and-roc/的说明,本文不再赘述. 同时,表 3 中的 MCMC 拒绝率指的是在参数更新过程中,没能更新的数量占总的迭代次数的比例,其可以反映参数的收敛速度; 这里各个群参数的初始值都取为一致,具体为 $(\theta_2, \theta_3, \theta_4) = (3.00, 1.00, 0.50)$. 读表 3 可见: (1) 模型样本内预测的准确率基本都在 80% 以上,并且当允许误差 Δ 值取的较小时,模型的预测精度好于允许误差 Δ 值取较大值的情形; (2) 从三个参数数值的大小看,可以发现无论针对哪个群的样本,都有 $\theta_2 > \theta_3 > \theta_4$,这意味着一度影响力最强、二度影响力次之,而三度影响力最弱; (3) 针对各个群样本,还可以发现无论对于哪个参数而言,允许误差变大将导致参数的标准差变大; (4) 较小的误差允许值将导致较大的 MCMC 抽样拒绝率. 注意到,本模型中提出的基于 MCMC 的参数校准方法允许一定的误差存在,其类似于统计学中置信度的概念和作用,该参数可以实现调控的功能,使得在拒绝率(或称为“算法效率”)和准确度之间实现一个折衷的选择,以期方便管理者针对实际的问题选择不同的误差允许值.

表 3　参数估计结果

Table 3 Results of parameter estimation

参数	群号	群 1		群 2		群 3		群 4		群 5	
	Δ 值	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
θ_2	均值	10.28	10.40	10.38	10.43	11.45	10.90	11.19	11.34	10.90	11.03
	标准差	3.20	3.97	3.17	3.24	3.02	3.24	3.49	3.69	3.20	3.88
θ_3	均值	5.28	4.99	5.31	5.33	5.11	5.23	5.55	5.44	5.32	5.43
	标准差	2.19	2.30	2.42	2.61	2.22	2.38	2.58	2.70	2.40	2.67
θ_4	均值	3.36	3.22	3.34	3.26	3.45	3.54	3.20	3.21	3.67	3.54
	标准差	1.78	2.08	1.94	2.15	1.96	2.12	1.67	1.99	1.76	1.98
MCMC 拒绝率		24.3%	8.88%	26.1%	7.76%	27.1%	12.3%	27.5%	11.0%	26.6%	11.1%
模型的准确率		88.8%	87.4%	90.7%	88.2%	83.3%	81.0%	84.1%	83.1%	84.0%	82.1%

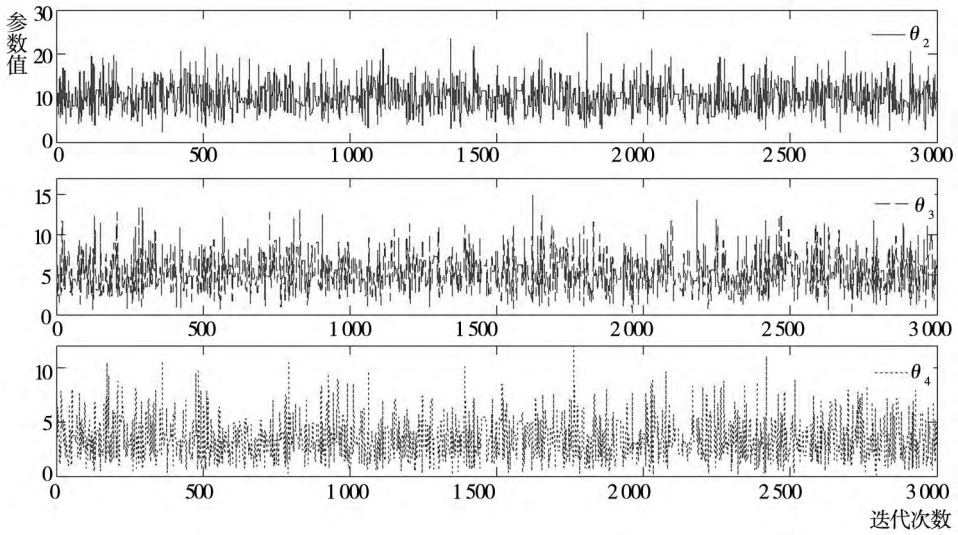

图 2　以群 1 为例的参数演化过程

Fig. 2 The updating parameter values (group 1 as the example)

进一步,以群编号为 1 的群为例,当 Δ 值取 0.05 时,三个参数的演化过程和统计分布情况分别由图 2 所示. 由图 2 可见: 三个参数在经历了较为短暂的演化过程后基本都达到了平稳的状态,这意味着表 3 中得到的均值和标准差能够在一定的统计意义下体现被估计参数的信息. 本文进一步进行了方法间的对比工作,针对收集到的 5 个群的信息,应用"引言"中提及的部分链路预测方法与本文的方法进行预测准确性的对比,具体的结果列在表 4 中. 读表可以发现: 在本文选取的数据集上,本文的方法在允许误差为 0.05 时,以选取的准确性指标为标准,较之比较的方法具有较好的预测精度,这从实践的角度说明了本文的方法可以成为既有链路预测方法的一个有力补充.

表 4　准确性的对比结果

Table 4 Comparisons of prediction accuracy

用于比较的方法	群 1	群 2	群 3	群 4	群 5
共同邻居法(CN) [3]	0.839 6	0.829 3	0.763 9	0.778 4	0.756 8
Jaccard 指数 [4]	0.864 3	0.893 0	0.774 5	0.810 0	0.812 1
Katz 指数 [7]	0.874 1	0.703 3	0.815 6	0.805 2	0.824 7
SimRank 方法 [8]	0.727 3	0.689 7	0.749 2	0.754 9	0.761 3
Liu 方法 [30]	0.837 3	0.829 3	0.720 7	0.783 4	0.691 1
Meng 方法 [31]	0.833 6	0.869 1	0.750 3	0.784 4	0.779 9
本文方法	**0.888 1**	**0.907 0**	**0.833 2**	**0.841 4**	**0.840 3**

4 结束语

本文立足于决策分析的理论，将网络参与者视为有智能的个体，将效用分析的方法引入社交网络链路预测问题的研究中，并应用 QQ 群这一社交媒体平台真实数据的例子对方法的适用性和准确性进行了验证. 在本文模型的求解过程中，需要对效用函数中的参数进行估计和校准，这是研究的一个难点和关键点. 本文基于模型的结构特征，采用了允许有一定误差的马尔科夫链蒙特卡洛方法(MCMC) 对参数进行校准，并论证了方法的正确性; 该估计方法具有算法复杂度较低，需要模型前提假设较少的优点，相比于普遍采用的最小二乘方法、Logistic 回归估计方法等，更适合于本模型的参数估计. 通过模型建立的过程和对比的结果，发现本文建立的模型具有以下特点: (1) 将效用函数引入社交网络链路预测的问题中，有助于解释网络链路形成的行为特征和微观基础，相比于纯粹的统计学习模型和机器学习模型而言，更多地考虑了参与者自身的效用因素、利益因素和属性特征; (2) 本文提出的参数估计方法能够完成效用函数中参数校准的目标，并且具有较好的预测准确度，可以适用于一定规模的社交网络的链路预测问题; (3) 模型的学习过程允许一定的误差存在，当误差设计的较小时，模型的参数学习过程有着较高的拒绝率，收敛会变慢; 而误差设计的较大时，模型的参数学习过程虽然会变快，但是也会有较大的预测误差; 这样的机制为模型的应用者在针对实际问题时提供了灵活选择的空间，综合权衡模型的效率和精度，以期符合实际背景的适用要求.

本文的模型解决的是静态网络上的链路预测问题，即没有考虑网络动态演化的过程和网络参与者见面的顺序与机制问题，这可以作为进一步的研究方向. 此外，本文将建立链接的过程视为一个双赢的过程，也即建立链接对双方都有利时才建立相应的链接; 事实上，在现实生活中，当建立链接两者整体的利益为正时，建立链接也是可能的，这涉及建立链接的机制设计问题，可以进一步作为本文研究的方向. 而且，本文中采用了全部样本对效用函数的参数进行校准，得到的参数估计值是在全体样本平均意义上的，这可能会忽略样本间的差异; 可以考虑将样本划分为若干个社团，在社团的层面上对参数进行校准，由此可以得到各个社团差异化的校准参数值，使得本文提出的效用函数针对不同社团的节点有差异化，这可能是进一步提高预测精度的手段之一，是进一步研究的方向. 另外，本文仅仅以 QQ 群这一社交媒体为例应用和验证了模型，进一步可以考虑在其他社交媒体平台上扩展本文模型的应用，特别是与既有的更多的链路预测方法的对比研究，这也是本文致力于进一步开展工作的地方.

参 考 文 献:

[1] 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)

[2] 赵正龙，陈 忠，孙武军，等. 具有差异化选择特征的复杂社会网络扩散研究 [J]. 管理科学学报，2010，13(3)：38－49.
Zhao Zhenglong，Chen Zhong，Sun Wujun，et al. Diffusion with property of differential choice complex social networks [J]. Journal of Management Sciences in China，2010，13(3)：38－49. (in Chinese)

[3] 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)

[4] Lü L，Zhou T. Link prediction in complex networks: A survey [J]. Physica A: Statistical Mechanics and its Applications，2011，390(6)：1150－1170.

[5] Li L, Qian L, Wang X, et al. Accurate similarity index based on activity and connectivity of node for link prediction [J]. International Journal of Modern Physics B, 2015, 29(17): 1550108.

[6] Liu Z, Dong W, Fu Y. Local degree blocking model for link prediction in complex networks [J]. Chaos: An Interdisciplinary Journal of Nonlinear Science, 2015, 25(1): 013115.

[7] Sherkat E, Rahgozar M, Asadpour M. Structural link prediction based on ant colony approach in social networks [J]. Physica A: Statistical Mechanics and Its Applications, 2015, 419(1): 80 - 94.

[8] Li D, Zhang Y, Xu Z, et al. Exploiting information diffusion feature for link prediction in sina weibo [J]. Scientific Reports, 2016, 6(1): 20058.

[9] Li Y, Luo P, Wu C. Information loss method to measure node similarity in networks [J]. Physica A: Statistical Mechanics and its Applications, 2014, 410: 439 - 449.

[10] Gong N Z, Talwalkar A, Mackey L, et al Joint link prediction and attribute inference using a social-attribute network [J]. ACM Transactions on Intelligent Systems and Technology, 2014, 5(2): 27.

[11] He Y, Liu J N K, Hu Y, et al. OWA operator based link prediction ensemble for social network [J]. Expert Systems with Applications, 2015, 42(1): 21 - 50.

[12] 俞　谈, 邱广华. 基于局部随机游走的在线社交网络朋友推荐算法 [J]. 系统工程, 2013, 31(2): 47 - 54.
Yu Yan, Qiu Guanghua. Algorithm of friend recommendation in online social networks based on local random walk [J]. Systems Engineering, 2013, 31(2): 47 - 54. (in Chinese)

[13] Clauset A, Moore C, Newman M E J. Hierarchical structure and the prediction of missing links in networks [J]. Nature, 2008, 453(7191): 98 - 101.

[14] White A, Murphy T B. Mixed-membership of experts stochastic blockmodel [J]. Network Science, 2015, 29(2): 1 - 33.

[15] Xu Z, Ke Y, Wang Y. A Fast Inference Algorithm for Stochastic Blockmodel [C]. Data Mining (ICDM), 2014 IEEE International Conference on. IEEE, 2014: 620 - 629.

[16] Johnson S, Abal E, Ahern K, et al. From science to management: Using Bayesian networks to learn about Lyngbya [J]. Statistical Science, 2014, 29(1): 36 - 41.

[17] Schlüter F, Bromberg F, Edera A. The IBMAP approach for Markov network structure learning [J]. Annals of Mathematics and Artificial Intelligence, 2014, 72(3 - 4): 197 - 223.

[18] Chen S H, Ip E H, Wang Y J. Gibbs ensembles for incompatible dependency networks [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2013, 5(6): 478 - 485.

[19] Natarajan S, Khot T, Kersting K, et al. Gradient-based boosting for statistical relational learning: The relational dependency network case [J]. Machine Learning, 2012, 86(1): 25 - 56.

[20] Yan L, Ma Z M. Incorporating fuzzy information into the formal mapping from web data model to extended entity-relationship model [J]. Integrated Computer-Aided Engineering, 2012, 19(4): 313 - 330.

[21] O'Rourke S. Gaussian fluctuations of eigenvalues in Wigner random matrices [J]. Journal of Statistical Physics, 2010, 138(6): 1045 - 1066.

[22] Huang L, Li R X, Wen K M, et al. A self training semi-supervised truncated kernel projection machine for link prediction [J]. Advanced Materials Research, 2012, 580: 369 - 373.

[23] 田儒雅, 刘怡君, 牛文元. 舆论超网络的领袖引导模型 [J]. 中国管理科学, 2014, 22(10): 136 - 141.
Tian Ruya, Liu Yijun, Niu Wenyuan. Leader-guiding model of online opinion supernetwork [J]. Chinese Journal of Management Science, 2014, 22(10): 136 - 141. (in Chinese)

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

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

[26] 刘怡君, 唐先一, 李倩倩, 等. 超链路预测 [J]. 管理评论, 2012, 24(12): 131 - 145.

Liu Yijun, Tang Xianyi, Li Qianqian, et al. Superlink prediction [J]. Management Review, 2011, 41(7): 816 – 823. (in Chinese)

[27] 张古鹏. 小世界创新网络动态演化及其效应研究 [J]. 管理科学学报, 2015, 18(6): 15 – 29.
Zhang Gupeng. Dynamic evolution of small world innovation network and its effects [J]. Journal of Management Sciences in China, 2015, 18(6): 15 – 29. (in Chinese)

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

[29] Li Y, Wu C, Wang X, et al. A network-based and multi-parameter model for finding influential authors [J]. Journal of Informetrics, 2014, 8(3): 791 – 799.

[30] Liu H, Hu Z, Haddadi H, et al. Hidden link prediction based on node centrality and weak ties [J]. Europhysics Letters, 2013, 101(1): 18004.

[31] Meng B, Ke H, Yi T. Link prediction based on a semi-local similarity index [J]. Chinese Physics B, 2011, 20(12): 128902.

Link prediction in social networks based on decision analysis

LI Yong-li[1], *LUO Peng*[2], *ZHANG Shu-rui*[3]

1. School of Business Administration, Northeastern University, Shenyang 110169, China;
2. School of Management, Harbin Institute of Technology, Harbin 150001, China;
3. College of Information Science and Engineering, Northeastern University, Shenyang 110169, China

Abstract: Social networks constitute the backbone of information transmission in social media platforms, where link prediction will contribute to managing information diffusion and controlling public opinion. Based on the existing studies in the field of link prediction, this paper starts from the theoretical foundation of decision analysis and presents a novel link prediction method by introducing the utility analysis. In order to solve the problem of parameter estimation, this paper further develops a Markov Chain Monte Carlo method with error degrees and demonstrates its correctness. Based on the selected information from five QQ groups, a comparison between the proposed method and the classic ones is carried out in terms of the prediction accuracy. The results indicate that the proposed method enjoys satisfactory prediction accuracy because the individual behavior is considered in this method, and particularly the introduced error degrees would benefit the potential model users in making reasonable decisions by weighing the model's efficiency and accuracy.

Key words: link prediction; decision analysis; utility function; Markov Chain Monte Carlo method; social network; data analysis