RUC-BK-071201-2013202003

中国人民大学本科毕业论文

基于样本网络数据的空间自回归成对极大似然估计

作　　者：陈韵然
学　　院：统计学院
专　　业：统计学
年　　级：2013 级
学　　号：2013202003
指导教师：
论文成绩：
日　　期：2017 年 4 月 19 日

摘要

空间自回归模型（Mixed Spatial Autoregressive Model，简称 SAR）经常用于网络结构数据的拟合，其参数估计常常采用极大似然方法。然而，大规模网络总体的全似然函数涉及到高维矩阵求逆和行列式求解问题，带来很大的计算量。现今的社交网络平台（如推特、脸谱网、人人网等）拥有的日活跃用户数均超过三千万，采用全网络极大似然估计计算量将难以想象。因而，实际估计中常常从全网络中随机抽取一部分节点组成样本网络，通过极大化样本似然函数来估计总体参数。但是，基于样本网络数据计算得到的空间自相关性的极大似然估计量有偏。针对大规模网络的巨大计算量问题，本文基于成对极大似然方法提出新的空间自回归模型参数估计方法。主要基于自相关系数很小的假定，利用泰勒展式将其似然函数近似为空间自相关系数的二次函数。同时，可以利用邻接矩阵的稀疏性有效地减少计算量。另一方面，针对样本网络极大似然估计有偏的问题，本文采用滚雪球抽样法极大的保留了网络的拓扑结构，并且利用社交网络中真实节点信息求解新估计量。我们通过数值模拟实验考察了所提估计方法的有限样本性质，并且采用滚雪球抽样收集微博用户数据进行案例分析。

关键词：空间自回归模型　成对极大似然估计　样本网络数据

Abstract

The Spatial Autoregressive Model (SAR) is often used to capture the characteristics of network data. Maximum likelihood method is widely used to estimate parameters in the model. However, the likelihood function of the large-scale network contains the inverse and determinant of the high-dimensional matrix, which bring a great deal of computation. Today's social networking platforms (such as Twitter, Facebook) have more than 30 million active users. It is hard to imagine taking the whole network into consideration. Practically, the actual estimation often based on a sampled network data, which is constructed by a set of nodes randomly extracted from the whole network. However, the estimation of the spatial autocorrelation based on sampled network data is biased. In order to settle large computational complexity of large-scale network, we propose a new parameter estimation of spatial autoregressive model based on pairwise maximum likelihood method. Based on the assumption that the autocorrelation is very small, we could apply a quadratic function to approximate the likelihood function with application of the first order Taylor expansion. At the same time, the calculation could be effectively reduced because of the sparseness of the adjacency matrix. To improve the biased estimator, we adopt snowball sampling method to preserve the topology of the network greatly, and use the real value of degree. We did several numerical simulations to confirm the good properties of estimation based on limited sample. In addition, we employed snowball sampling to collect a real-life data set crawling from Weibo to serve as a practical application of the estimation.

Key words: Mixed Spatial Autoregressive Model Pairwise Maximum Likelihood Estimation Sampled Network Data

内容目录

图目录

表目录

1 引言

人类作为社会性群体，其行为决策或多或少会受到他人的影响，是个人特质和他人行为共同作用下的产物。常言“物以类聚，人以群分”。一个人朋友的行为决策常常与这个人的行为决策存在很强的相关性。这一人际相关性特征可以在商业市场中得到有效利用。例如，信贷机构可以通过这个人周边朋友的信用状况来评估这个人的信誉和违约风险。对于产品销售部门而言，可以通过了解一个人朋友的偏好来估计这个人对产品的偏好进行有针对性地推销。特别地，人际关系的强度很大程度上影响了推断的准确性（Hartmann et al. 2008；Franzese & Hays 2007）。现有文献显示，人际相关性与消费者选择偏好（Yang & Allenby 2003）、广告投放的最优策略（Aravindakshan et al. 2012）等密切相关。

近年来，计算机与互联网的普及和高速发展催生了各类社交网络平台，如国外的推特（Twitter）、脸谱网（Facebook），和国内的微博、微信等。除了社交网络软件外，许多其他的网络平台都纳入社交功能以获取用户的人际关系信息。这些社交网络平台捕捉了大量的网络结构数据，使得人际关系这一对人们行为活动有重要影响的因素，能够被观测记录进而得到高效的利用。如何利用人际网络拓扑学信息以提升预测效果成为许多国内外学者关注的重点。选用合适的模型和参数估计方法是有效度量社会网络中人际相关性的关键问题（Doreian 1989；Franzese & Hays 2007；Chen et al. 2013；Zhou et al. 2015）。

Ord（1975）提出的空间自回归模型被广泛应用于拟合社会网络结构数据（Doreian 1989；Bradlow et al. 2005；Bronnenberg 2005；Lee et al. 2010）。Doreian（1989）引入空间自回归模型以解决社交网络中人际关系问题，并指出纳入了人际关系网络可以更好地估计行为特征，同时对空间自回归模型进行了多过程、节点类型不同情况下的拓展。Anselin（2002）指出，当研究关注点的在于人际关系是如何使人们产生相似的行为或呈现集群模式这类问题时，常常使用空间自回归模型来识别人际关系的强度。当自相关性 ρ 是作为一冗余参数，关键点在于去除自相关性以获得回归系数 β 的无偏估计时，因为回归系数 β 的估计必然涉及到冗余参数 ρ 的估计，所以不适合采用空间自回归模型。Brueckner（2002）提出了两个理论框架说明空间自回归模型在探究社交网络的人际效应中的适用性。

空间自回归模型如下：

$$Y = \rho WY + X\beta + \varepsilon$$

其中，W 为度量人际关系网络的矩阵，X 度量个人特征。ρ 表示自相关性，衡量一个人的人际关系对其的影响程度。而 β 表示这个人的个人特质对其行为决策造成的影响程度。

对于空间自回归模型参数估计，传统的估计方法如最小二乘法会导致估计量的有偏性和不一致性（Ord 1975），进而提出其它参数估计方法，如最大似然估计、工具变量法、两阶段最小二乘法、贝叶斯框架下 MCMC 模拟等（Ord 1975；Anselin 1988；Kelejian & Robinson 1993；Kenneth

& Glenn 1992；Yang & Allenby 2003）。其中，Ord（1975）提出的极大似然估计方法由于其解释的便利性、估计量的一致性、有效性和良好的渐近性质，得到了最广泛的使用（Anselin 1980；Lee et al. 2010）。但是这一估计方法面临着两方面的挑战。

一方面，求解极大似然函数极值点的计算量过大。

极大似然函数包含了矩阵行列式$|I-\rho W|$，求解似然函数极值的过程中涉及到此行列式的计算。这一非线性优化问题需要反复迭代求解，计算量很大。Ord（1975）提出了极大似然估计的简化计算方案。在行列式的计算中，把其拆解为矩阵特征根的乘积，即$|I-\rho W|=\prod_{i=1}^{N}(1-\rho\lambda_i)$。当$N$很大时，可以通过只关注最大的几个特征值，或者将矩阵W简化为分块对角阵的形式简化计算。但是一些学者随后指出，随着N增大，矩阵特征值的计算不稳定。另外，社交网络数据邻接矩阵的稀疏性这一特征并没有得到很好的利用。因而对于大规模网络数据，这一简化计算的方法并不高效（Roncek & Montgomery 1984；Land & Deane 1992）。另外，一些其他的简化计算的方法被提出。第一类是关注计算$I-\rho W$，通过对矩阵W增加假定以极大程度稀疏化矩阵或者寻求$I-\rho W$的简化形式。Pace & Berry（1997）假定几乎没有直接联系的关系网络。Pace & Zou（2000）假定只有最近的节点才能造成直接影响。LeSage & Pace（2007）假定$I-\rho W$可以被指数函数替代。Barry & Pace（1999）采用蒙特卡洛模拟的方法来近似计算雅各比行列式。第二类是寻求整个似然函数的简化形式。Besay（1975）提出了伪似然函数，在给定一些其他节点的情况下写出条件密度。Smirnov & Anselin（2000）采用特征多项式逼近似然函数。Zhou et al.（2015）研究不含外生变量的空间自回归模型（pure SAR），通过泰勒展式简化似然函数为二次函数形式，并在此基础上提出了成对似然方法进一步简化计算。

另一方面，对于样本网络数据采用最大似然方法会导致自相关系数的估计有偏。

Stephen（2012）指出现今网络和20世纪网络研究的关键区别在于网络规模。1984年Roncek & Montgomery为指出极大似然估计不高效，所举的反例仅涉及到了15000个节点的网络。而现今随着互联网的普及和迅速发展，网络社区如脸谱网（Facebook）、领英（linkedin）、推特（twitter）、微博（weibo），和以特有的兴趣聚集的专业化网络社区如知乎、果壳、谷歌学术等等都蓬勃发展起来。网络数据的大规模是计算机和互联网发展下的必然趋势。据 Statista 网站的数据显示，截止2017年1月，脸谱网（Facebook）的日活跃用户达到了1.87亿。而排名前十的社交网络网站均拥有3千万以上的日活跃用户。64%的互联网用户会使用网络社交平台。随着智能手机进一步普及和推广，这一数字还将上升。

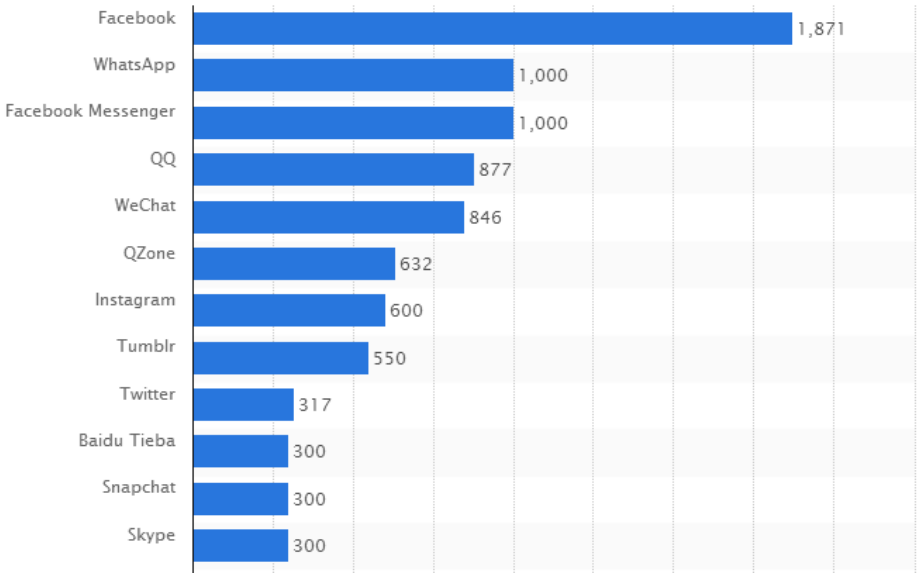

图 1：社交网络日活跃用户数

这一数量级意味着，即使我们可以获取整个网络的观测数据，也很难将全部的网络数据纳入模型以估计参数。因此，人们常常从中抽取样本，通过样本推断总体特征。特别是针对极大似然函数估计，抽取的样本量会受到更大的约束。

尽管极大似然估计量具有很多优良性质，但是对于网络结构的数据，基于样本数据的最大似然估计方法会导致自相关系数被低估。

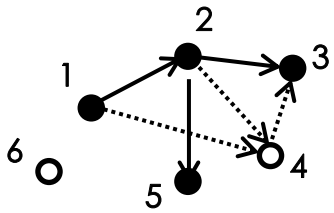

图 2：社会网络关系图

这一低估其实源自抽样的过程中，一些节点被忽略了，因而在样本中每个节点的度都被低估了。以实心点 1、2、3、5 代表被抽中的节点，以空心点 4、6 代表未被抽中节点。由于节点 4 与其余 4 个被抽中的节点相连，所以被抽中的节点的度是被低估的，样本网络结构被扭曲。样本网络中节点的度被低估，会造成自相关系数的低估（Chen et al. 2013）。Chen（2014）通过对几种网络的模拟也说明了极大似然估计有偏。

针对样本网络极大似然估计量的有偏性问题，目前有两种解决方案。

一是通过改进抽样方案以最大程度保留真实网络结构。不同的抽样方法会影响到样本网络对全网络重要性质的保留，因而采用合适的网络抽样方法可以得到更精确的网络推断（Leskovec & Faloutsos 2006；Choudhury et al. 2010；Chen et al. 2013；Chen et al. 2014； Ebbes et al. 2016）。其中，基于网络拓扑结构的高蔓延率森林火势蔓延抽样（Forest Fire Sampling with high burning

rate）和基于每一步已知信息构建抽样规则的适应性抽样法（Adaptive Sampling）得到广泛应用。Ebbes et al （2016）指出若研究重点在于局部的影响效应，采用较高蔓延率（Burning Rate）的森林火势蔓延抽样（Forest Fire Sampling）方法是最优的选择（Leskovec & Faloutsos, 2006； Ebbes et al. 2016）。Handcock & Gile（2010）通过模拟指出基于这一抽样方法得到的样本网络的极大似然估计只有中等程度的偏差，并且随着未观测节点的增加，偏差增加的速度较为缓慢。Chen（2013）通过一系列的模拟提出滚雪球抽样（蔓延率为 100%的森林火势蔓延抽样方法）为最能保留网络结构的抽样方式，可以极大程度改善极大似然估计的有偏性。Thompson（2006）和 Chen（2014）通过模拟提出序贯抽样（适应性抽样）可以加强极大似然方法的有效性。他们基于每一步抽取的样本构造条件似然函数，通过一定的准则筛选下一步抽取的样本。但是为了保证估计的连续型和无偏性，在每一步都要考虑到条件选择概率，计算较为复杂，计算量较大。

另一种解决方法则是利用已知信息。对于现今的社交网络，每个节点真实的度是已知的。以微博为例，每个用户的粉丝数和关注数均为已知。因此可以利用真实的度的数据修正估计量。Chen（2014）提出的 SEQ-MCLE 估计量和 Zhou（2015）提出的 AMLE、PMLE 估计量均利用了真实的度进行估计。但是，Chen（2014）提出的 SEQ-MCLE 基于每一步抽取都要构造似然函数进行估计，并且根据特定准则选取下一步抽取的样本，反复迭代，计算量较大。Zhou（2015）通过泰勒展式将似然函数转换为二次函数形式，并在此基础上提出成对似然减少计算量。但是，Zhou（2015）等人只考虑了无外生变量的空间自回归模型，即 $Y = \rho WY + \varepsilon$ 。并没有将个人特质 $X\beta$ 对个人决策行为的影响纳入考虑。而在实际中，个人特质才是对个人决策行为的决定性因素，因此本文将建立含有外生变量的空间自回归模型，并提出相应的参数估计方法。

为了解决极大似然估计的计算量大的问题，本文采用成对似然函数进行估计，同时利用一阶泰勒展式将似然函数化简为二次函数形式以简化计算。大规模、有方向性和稀疏性是社交网络邻接矩阵的三大特征。考虑到其稀疏性特征，本文引入成对最大似然估计极大地减小了计算量。同时，成对极大似然估计保留了极大似然估计的一致性和渐近性质（Shao 2003）。基于样本数据估计有偏问题，本文采用滚雪球抽样以最大程度保存网络的拓扑结构（Henry 2005；Salganik & Heckathorn 2004；Tepper 1994；Frenzen & Davis 1990）。并且利用了真实的度（在社交网络中通常为已知）来修正估计。通过模拟，可以说明本文提出的估计方法有不错的估计效果。

本文组织框架如下：第二章介绍基于社交网络数据建立空间自回归模型，及其背后的两个理论框架。另外，将详细介绍成对似然函数作为复合似然函数的一种最为简单的形式，其定义和相应的优良性质；第三章引入成对极大似然方法对空间自回归模型进行参数估计，提出成对似然估计量；第四章介绍模拟过程及结果展示；第五章基于微博数据进行实际案例分析；第六章为进一步的讨论与思考。

2 空间自回归模型及成对似然估计

2.1 空间自回归模型

本文以微博这一线上社交网络为例来进行建模说明。

对于任意两位用户 i 和 j，存在下图四种关系。其中，由 i 指向 j 的箭头，表征用户 i 关注了用户 j。由于 i 关注 j 与 j 关注 i 这两个事件并不等同，因此这一网络是有向性的。并且，用户不能自我关注，所以不会有指向自己的自循环箭头。

图 3：节点（用户）关系图

将上述图形用数学语言表示。则引入 a_{ij} 作为示性变量，表征用户 i 和 j 之间的关系。如果用户 i 关注了 j，则 a_{ij} 取值为 1，反之为 0。数学表达式如下：

$$a_{ij}=\begin{cases}1 & \text{如果用户 } i\text{关注 } j\\ 0 & \text{其它}\end{cases}$$

因此上述四种状态可以分别由下述的 4 个矩阵唯一识别，即：

$$\begin{pmatrix}0&1\\0&0\end{pmatrix}\qquad\begin{pmatrix}0&0\\1&0\end{pmatrix}\qquad\begin{pmatrix}0&1\\1&0\end{pmatrix}\qquad\begin{pmatrix}0&0\\0&0\end{pmatrix}$$

注意，可以发现矩阵具有两个特点：非对称和对角线为 0。对于有 N 个用户（节点）的有向性网络，可以被 $N\times N$ 个 a_{ij} 构成的矩阵唯一识别，即 $A=(a_{ij})\in R^{N\times N}$。我们称这一矩阵 A 为邻接矩阵。

对于热衷关注他人的用户来说，关注者们对此人的影响程度可能会被分散，因此常常对邻接矩阵 A 做一个“行正态化”处理，得到矩阵 W 来表征网络结构，即 $W=(\frac{a_{ij}}{d_i})\in R^{N\times N}$。其中 d_i 表示节点 i 的度，即用户 i 关注的人数总量。矩阵 W 的每个元素 w_{ij} 可以衡量用户 i 和 j 联系的紧密程度（Ord 1975）。当用户 i 没有关注用户 j 时，$w_{ij}=0$。而用户 i 关注的人数越多，即分母 d_i 越大，相应的与关注的人关系的紧密程度会被相应的分散，因而 w_{ij} 会越小。

考虑有 N 个节点的网络。网络结构由 $A=(a_{ij})\in R^{N\times N}$ 度量。运用空间自回归模型来拟合网络数据，有：

$$Y = \rho WY + X\beta + \varepsilon \qquad (1)$$

其中，$\rho \in R^1$ 表示空间自相关系数，度量人际效应。$Y \in R^N$ 表示响应变量，度量个人的行为表现，假定其为连续性变量。$X \in R^{N\times P}$ 表示外生变量，度量个人特征。$\beta \in R^{P\times 1}$ 表示回归系数，度量个人特征对决策行为的影响。$W = (\frac{a_{ij}}{d_i}) \in R^{N\times N}$ 为行正态化后的邻接矩阵，度量人际网络关系的强弱，是非随机化的外生变量。$\varepsilon \in R^N$ 度量随机效应，服从正态分布 $\varepsilon \sim N_N(0, \sigma^2 I)$。

这一空间自回归模型，可以从两种理论框架上来予以解释（Brueckner 2002）。

第一种称为溢出效应模型（spillover model），指一个用户的行为受到他人的直接影响。即用户 i 的行为决策 y_i 是由其他用户的行为决策 y_{-i}（这里，y_{-i} 中的 $-i$ 表示除 i 外的所有角标）共同作用。例如，微博用户对一则微博的转发行为在一定程度上受到其关注的用户是否转发了这一则微博的影响。因而，对于每一个用户 i，目标函数（objective function）可写作：

$$U(y_i, y_{-i}; x_i^{'})$$

其中，$x_i^{'}$ 代表了用户 i 的个人特征（外生变量）。对这一目标函数求解最大值则得到了如下决策函数（reaction function）：

$$y_i = R(y_{-i}, x_i^{'})$$

设 $R(\bullet)$ 为一个线性形式，且将人际关系以一个空间权重矩阵 W 表示，则可以得到如下形式：

$$Y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon$$

注意到这是式（1）的移项变换。其中，$(I - \rho W)^{-1}$ 称为列昂惕夫逆矩阵（Leontief inverse）。可以发现 y_i 的值并非只受到 $X\beta$ 和误差项 ε 的影响，而且有一个空间累积效应，故也称 $(I - \rho W)^{-1}$ 为空间效应乘子（spatial multiplier）。

另一种理论称为资源流通模型（resource flow model），即用户的决策并不受到其他用户的直接影响，而是间接影响。一个用户所做的决定依赖于个人特质和在其他用户的间接影响下可用“资源”。故用户 i 的目标函数可以如下表示：

$$U(y_i, s_i, x_i^{'})$$

其中，s_i 代表用户 i 的可用资源，在社会网络中可以表示这个人的影响力，可表示为个人特质和他人决策的函数，如下：

$$s_i = H(y_i, y_{-i}; x_i^{'})$$

将上式带入目标函数，可以得到和溢出模型同样的决策函数，Brueckner（2002）也建议采用空间自回归建模。

空间自回归模型可以解释以上所述的两种机制，适用于社交网络数据建模研究。

2.2 成对似然估计

似然函数在频率学派和贝叶斯学派的统计推断中具有至关重要的地位。但是随着现有数据规模的爆炸式增长和变量间复杂的相依关系，写出模型的全似然函数并且求解其极大值是不切实际的。所以复合似然（composite likelihood）近年来持续受到关注，可以有效的解决大规模数据的计算和高维复杂相依结构下的识别。

复合似然函数是一系列边际密度函数或条件密度函数的乘积，为全似然函数的近似形式。这些边际密度函数或条件密度函数为复合似然函数的“元素”，称成分似然函数（component likelihood）。考虑随机变量 X 服从分布 $f(x;\theta)$，其中 $x=(x_1,...,x_p)^T\in R^p$，$\theta\in R^d$。记每个成分似然函数（component likelihood）为 $L_k(\theta;x)$，其中 $k=1,2,...,K$。则复合似然函数可写作：

$$L_C(\theta;x)=\prod_{k=1}^{K}L_k(\theta;x)^{\lambda_k}$$

其中，λ_k 为非负权重。

成对似然函数为复合似然函数的一种特殊情况，即

$$L_{pairwise}(\theta;x)=\prod_{r=1}^{p-1}\prod_{s=r+1}^{p}L_k(x_r,x_s;\theta)$$

复合似然函数不仅有效的减小了计算量和高维复杂的相依问题，并且保留了全似然函数的很多优良性质。复合似然在一定的条件下具有一致性、有效性、稳健性以及渐近无偏性、渐近正态性（Lindsay 1988；Molenberghs & Verbeke 2005；Jin 2009；Xu & Reid 2011；Varin et al. 2014；Wang & Wu 2014），其渐近协方差阵为Godambe信息矩阵的逆（Godambe 1960）。Varin et al.（2011）对复合似然估计量性质做了很全面的阐述。

3 参数估计

下面我们将利用泰勒展式对似然函数化简并且引入成对似然函数来进一步简化计算。

对上述方程（1）进行移项，我们有

$$Y=(I-\rho W)^{-1}X\beta+(I-\rho W)^{-1}\varepsilon \tag{2}$$

由于 $\varepsilon\sim N_N(0,\sigma^2 I)$，所以有

$$Y\sim N_N((I-\rho W)^{-1}X\beta,(I-\rho W^T)^{-1}(I-\rho W)^{-1}\sigma^2) \tag{3}$$

写出似然函数估计参数。省略了常数项的似然函数如下：

$$\begin{aligned}Loglik\propto&-\log|(I-\rho W^T)^{-1}(I-\rho W)^{-1}|-N\log\sigma^2\\&-\sigma^{-2}(Y-(I-\rho W)^{-1}X\beta)^T(I-\rho W^T)(I-\rho W)(Y-(I-\rho W)^{-1}X\beta)\end{aligned} \tag{4}$$

分别对参数σ^2和β求导，我们有

$$\hat{\sigma}^2=\frac{1}{N}(Y-(I-\rho W)^{-1}X\beta)^T(I-\rho W^T)(I-\rho W)(Y-(I-\rho W)^{-1}X\beta) \quad (5)$$

$$\hat{\beta}=(X^TX)^{-1}X^T(I-\rho W)Y \quad (6)$$

将$\hat{\sigma}^2$和$\hat{\beta}$带入似然函数，我们有

$$\begin{aligned}Loglik \propto & -\log|(I-\rho W^T)^{-1}(I-\rho W)^{-1}| \\ & -N\log[N^{-1}(Y-(I-\rho W)^{-1}X\hat{\beta})^T(I-\rho W^T)(I-\rho W)(Y-(I-\rho W)^{-1}X\hat{\beta})\end{aligned} \quad (7)$$

似然函数形式较为复杂，我们先将$\hat{\beta}$代入，化简第二部分。

$$-Nlog\{N^{-1}Y^T[I-(I-\rho W)^{-1}H(I-\rho W)]^T[(I-\rho W^T)(I-\rho W)][I-(I-\rho W)^{-1}H(I-\rho W)]Y\} \quad (8)$$

其中，H为帽子矩阵，$H=X(X^TX)^{-1}X^T$。

由于

$$\begin{aligned}[I-(I-\rho W)^{-1}H(I-\rho W)] &= (I-\rho W)^{-1}[(I-\rho W)-H(I-\rho W)] \\ &= (I-\rho W)^{-1}(I-H)(I-\rho W)\end{aligned} \quad (9)$$

所以式（8）可以继续被化简为如下形式：

$$\begin{aligned}(8) &= -N\log\{N^{-1}Y^T[(I-\rho W)^{-1}(I-H)(I-\rho W)]^T[(I-\rho W^T)(I-\rho W)][(I-\rho W)^{-1}(I-H)(I-\rho W))]Y\} \\ &= -N\log\{N^{-1}Y^T(I-\rho W^T)(I-H)(I-\rho W)Y\}\end{aligned} \quad (10)$$

即似然函数可化简为如下形式。对下式求极值则可解出ρ的极大似然估计。

$$Loglik \propto -log|(I-\rho W^T)^{-1}(I-\rho W)^{-1}|-Nlog\{N^{-1}Y^T(I-\rho W^T)(I-H)(I-\rho W)Y\} \quad (11)$$

可以注意到化简后的似然函数仍涉及到$N\times N$矩阵的求逆运算和行列式计算。对于普通的社交平台而言，单就日活跃用户而言均在三千万以上。但对于普通的计算机而言，涉及到上万维的矩阵求逆和行列式运算就已难以运行。在 N 很大时，利用全似然函数求解的计算量是难以实现的。退一步，即使抽取样本网络数据，也会耗费大量的计算时间。因此，我们考虑对上式进行进一步的简化。

为简化上式，应用一阶泰勒展式$\log(1-t)\approx -t$，把式（10）做类似变换，得到式（11）。

$$-Nlog(N^{-1}Y^T(I-T)Y)\approx\sigma_Y^{-2}Y^TTY \quad (12)$$

其中，$T=H+\rho[(I-H)W+W^T(I-H)]-\rho^2W^T(I-H)W$，$\sigma_Y^{-2}=N^{-1}Y^TY$。注意，这一步近似计算的前提是$\frac{1}{N}\sigma_Y^{-2}Y^TTY$趋于 0， 在自相关系数很小的假设下，我们还需要在实际计算时对Y进行标准化处理。

因此似然函数可以被简化为下式：

$$Loglik \propto log|(I-\rho W^T)^{-1}(I-\rho W)^{-1}|+\sigma_Y^{-2}Y^T\{H+\rho[(I-H)W+W^T(I-H)]-\rho^2W^T(I-H)W\}Y \quad (13)$$

我们将式（13）拆分为两部分分别化简。

对于第一部分化简有：

$$\begin{aligned}-log|(I-\rho W^T)^{-1}(I-\rho W)^{-1}| &= -log|(I-\rho(W+W^T)+\rho^2W^TW)^{-1}| \\ &= log|I-\rho(W+W^T)+\rho^2W^TW|\end{aligned} \quad (14)$$

仅考虑节点 i 和节点 j。将上式化为成对的形式，我们有：

$$-log|(I-\rho W^T)^{-1}(I-\rho W)^{-1}| \quad =log[1-\rho^2(\frac{a_{ij}}{d_i}+\frac{a_{ji}}{d_j})^2+\rho^2(\frac{a_{ij}^2}{d_i^2}+\frac{a_{ji}^2}{d_j^2})]$$

$$\approx -\rho^2(\frac{a_{ij}}{d_i}+\frac{a_{ji}}{d_j})^2+\rho^2(\frac{a_{ij}^2}{d_i^2}+\frac{a_{ji}^2}{d_j^2}) \qquad (15)$$

上式化简中利用了一阶泰勒展式 $\log(1-t)\approx -t$ 。

下面，我们对式（13）的第二部分化简为成对形式。

首先，我们提取出帽子矩阵的第 i 和 j 行，第 i 和 j 列构建成对帽子矩阵。即

$$H=(X(X^TX)^{-1}X^T)_{i,j}=\begin{pmatrix} h_{ii} & h_{ij} \\ h_{ij} & h_{jj} \end{pmatrix} \qquad (16)$$

因为 $X(X^TX)^{-1}X$ 是一个对称幂等阵，因此成对帽子矩阵 H 也是对称和幂等的。这些性质会在稍后的计算中用到。

在定义了成对帽子矩阵后，可以把式（13）的各项转为成对形式。

$$(I-H)W=\begin{pmatrix} -\frac{a_{ji}}{dj}h_{ij} & \frac{a_{ij}}{di}(1-h_{ii}) \\ \frac{a_{ji}}{dj}(1-h_{jj}) & -\frac{a_{ij}}{di}h_{ij} \end{pmatrix} \qquad (17)$$

$$W^T(I-H)W=\begin{pmatrix} -\frac{a_{ji}}{d_j}h_{ij} & \frac{a_{ji}}{d_j}(1-h_{jj}) \\ \frac{a_{ij}}{d_i}(1-h_{ii}) & -\frac{a_{ij}}{d_i}h_{ij} \end{pmatrix}\times\begin{pmatrix} 0 & \frac{a_{ij}}{d_i} \\ \frac{a_{ji}}{d_j} & 0 \end{pmatrix}=\begin{pmatrix} \frac{a_{ji}^2}{d_j^2}(1-h_{jj}) & -\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}h_{ij} \\ -\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}h_{ij} & \frac{a_{ij}^2}{d_i^2}(1-h_{ii}) \end{pmatrix} \qquad (18)$$

以 T 表示记式（13）中弧形括号里的式子，即

$$\sigma_Y^{-2}Y^TTY=\sigma_Y^{-2}Y^T\{H+\rho[(I-H)W+W^T(I-H)]-\rho^2W^T(I-H)W]\}Y$$

记 $T=\begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}$，将（17）（18）代入式（13），可以得到 T 的具体形式如下：

$$T_{11}=h_{ii}+\rho\times(-2\frac{a_{ji}}{d_j}h_{ij})-\rho^2\frac{a_{ji}^2}{d_j^2}(1-h_{jj}) \qquad (19)$$

$$T_{12}=T_{21}=h_{ij}+\rho\times(\frac{a_{ij}}{d_i}(1-h_{ii})+\frac{a_{ji}}{d_j}(1-h_{jj}))+\rho^2\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}h_{ij} \qquad (20)$$

$$T_{22}=h_{jj}+\rho\times(-2\frac{a_{ij}}{d_i}h_{ij})-\rho^2\frac{a_{ij}^2}{d_i^2}(1-h_{ii}) \qquad (21)$$

将式（19）至式（21）代入下式

$$Y^TTY=T_{11}y_i^2+T_{22}y_j^2+2T_{12}y_iy_j \qquad (22)$$

同时，结合化简式（15），我们可以得到成对样本似然函数，是一个二次函数形式：

$$loglik \propto -(\frac{a_{ij}}{d_i}+\frac{a_{ji}}{d_j})^2\rho^2+\rho^2(\frac{a_{ij}^2}{d_i^2}+\frac{a_{ji}^2}{d_j^2})+\rho^2\sigma_Y^{-2}[-\frac{a_{ji}^2}{d_j^2}(1-h_{jj})y_i^2+2\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}h_{ij}y_iy_j-\frac{a_{ij}^2}{d_i^2}(1-h_{ii})y_j^2]$$
$$-2\rho\sigma_Y^{-2}[\frac{a_{ji}}{d_j}h_{ij}y_i^2-(\frac{a_{ij}}{d_i}(1-h_{ii})+\frac{a_{ji}}{d_j}(1-h_{jj}))y_iy_j+\frac{a_{ij}}{d_i}h_{ij}y_j^2]+c \quad (23)$$

其中，$c=\sigma_Y^{-2}(h_{ii}y_i^2+2h_{ij}y_iy_j+h_{jj}y_j^2)$与$\rho$的最优解无关，故省略。

对式（23*）成对形式的似然函数求和即为成对似然函数，由于是二次函数，可以很容易得到ρ的成对最大似然估计$\hat{\rho}$，即

$$\hat{\rho}=\frac{\sum_{i,j}\sigma_Y^{-2}[\frac{a_{ji}}{d_j}h_{ij}y_i^2-(\frac{a_{ij}}{d_i}(1-h_{ii})+\frac{a_{ji}}{d_j}(1-h_{jj}))y_iy_j+\frac{a_{ij}}{d_i}h_{ij}y_j^2]}{\sum_{i,j}-2\times\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}+\sigma_Y^{-2}[-\frac{a_{ji}^2}{d_j^2}(1-h_{jj})y_i^2+2\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}h_{ij}y_iy_j-\frac{a_{ij}^2}{d_i^2}(1-h_{ii})y_j^2]} \quad (24)$$

注意，对于那些没有任何联系的成对节点，$a_{ij}=a_{ji}=0$。由式（24）可知这些节点与参数ρ的估计无关。所以这些没有任何联系的成对节点可以在计算时省略，在式（24）的求和中可以不予考虑，化简为如下形式：

$$\hat{\rho}=\frac{\sum_{a_{ij}+a_{ji}>0}\sigma_Y^{-2}[\frac{a_{ji}}{d_j}h_{ij}y_i^2-(\frac{a_{ij}}{d_i}(1-h_{ii})+\frac{a_{ji}}{d_j}(1-h_{jj}))y_iy_j+\frac{a_{ij}}{d_i}h_{ij}y_j^2]}{\sum_{a_{ij}+a_{ji}>0}-2\times\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}+\sigma_Y^{-2}[-\frac{a_{ji}^2}{d_j^2}(1-h_{jj})y_i^2+2\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}h_{ij}y_iy_j-\frac{a_{ij}^2}{d_i^2}(1-h_{ii})y_j^2]} \quad (24^*)$$

由于社交网络邻接矩阵的稀疏性，省略这些节点可以极大的减少计算量。

将式（24*）带入式（6），可以得到β的成对似然估计$\hat{\beta}$。

值得注意的是，由于式（12）处用到了一阶泰勒展式，前提是 T 趋于 0。因而在实际计算时，要将 X 和 Y 进行标准化处理，即

$$X_j^*=\frac{X_j-\bar{X}_j}{sd(X_j)},\quad Y^*=\frac{Y-\bar{Y}}{sd(Y)}$$

标准化 X,Y 后，$\sigma_{Y^*}^{-2}$=1，因而式（24*）可被进一步简化为下式：

$$\hat{\rho}^*=\frac{\sum_{a_{ij}+a_{ji}>0}[\frac{a_{ji}}{d_j}h_{ij}y_i^2-(\frac{a_{ij}}{d_i}(1-h_{ii})+\frac{a_{ji}}{d_j}(1-h_{jj}))y_iy_j+\frac{a_{ij}}{d_i}h_{ij}y_j^2]}{\sum_{a_{ij}+a_{ji}>0}-2\times\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}+[-\frac{a_{ji}^2}{d_j^2}(1-h_{jj})y_i^2+2\frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j}h_{ij}y_iy_j-\frac{a_{ij}^2}{d_i^2}(1-h_{ii})y_j^2]} \quad (25)$$

$\hat{\beta}^*$要做相应的调整，即真实值$\hat{\beta}_j$为：

$$\hat{\beta}_j=\frac{sd(Y)}{sd(X_j)}\hat{\beta}_j^* \quad (26)$$

注意，当空间自回归模型中$\beta=0$时（pure SAR），H 元素趋于 0。式（25）退化为下式：

$$\hat{\rho}_C = \frac{\sum_{a_{ij}+a_{ji}>0} (\frac{a_{ij}}{d_i} + \frac{a_{ji}}{d_j}) y_i y_j}{\sum_{a_{ij}+a_{ji}>0} 2 \times \frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j} + \frac{a_{ji}^2}{d_j^2} y_i^2 + \frac{a_{ij}^2}{d_i^2} y_j^2} \qquad (25^*)$$

而 Zhou et al.（2015）提出的 PMLE 估计量如下：

$$\hat{\rho}_Z = \frac{\sum_{a_{ij}+a_{ji}>0} (\frac{a_{ij}}{d_i} + \frac{a_{ji}}{d_j}) y_i y_j}{\sum_{a_{ij}+a_{ji}>0} 2 \times \frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j} + \frac{a_{ji}^2}{d_j^2} + \frac{a_{ij}^2}{d_i^2}} \qquad (27)$$

由于 Zhou et al.（2015）在化简过程中并没有保留$O(\rho^2)$项，因而式（25*）与式（27）化简结果有差异。注意，由于$\sum_{a_{ij}+a_{ji}>0} 2 \times \frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j} + \frac{a_{ji}^2}{d_j^2} y_i^2 + \frac{a_{ij}^2}{d_i^2} y_j^2 \neq \sum_{a_{ij}+a_{ji}>0} 2 \times \frac{a_{ij}}{d_i}\frac{a_{ji}}{d_j} + \frac{a_{ji}^2}{d_j^2} + \frac{a_{ij}^2}{d_i^2}$。所以本文提出的估计量$\hat{\rho}_C$在 pure SAR 下的表现可能会优于 Zhou et al.（2015）提出的$\hat{\rho}_Z$的表现，这一点将会在数值模拟中予以验证。

4 数值模拟

4.1 模拟过程

为了评估本文提出的成对似然估计量的表现，将通过模拟产生全网络数据，并基于此抽样产生样本网络数据。本文采用了与 Zhou et.al.（2016）相同的模拟过程产生数据。模拟的过程大致分为如下三步。

（1） 产生全网络数据

依据空间自回归模型产生全网络数据：

$Y = \rho WY + X\beta + \varepsilon$

在给定全网络规模N下，我们需要产生W、X和ε。然后在给定参数ρ、β下，我们可以计算出Y值。那么我们可以得到全网络数据，再进行第二步抽样。

①模拟网络结构W。为了使模拟的网络结构接近于真实社交网络中人际网络结构，我们假定每个节点的度独立同分布于指数分布，记其均值为expm。在这里我们设置expm=10。我们采用与 Zhou et al.（2015）相同的记号。记每个节点的度为E_i，$1 \le i \le N$。出于实际意义的考虑，E_i表示用户 i 关注的人数，为整数形式。因此我们需要采用合适的规则将E_i取为整数。这里我们取大于或等于E_i的最小整数，记作$[E_i]$。对每一个节点 i，从$S = \{1,2,...,N\}$中不放回的随机抽取样本量为$[E_i]$的样本，记作S_i。S_i这一指标集表示用户 i 关注的其他用户。我们定义

$a_{ij}=\begin{cases}1 & j\in S_i\\ 0 & j\notin S_i\end{cases}$，则可以初步得到一个邻接矩阵 $A=(a_{ij})\in R^{N\times N}$。接着我们对这个矩阵进行角对称化。即令 $a_{ij}=a_{ji}$。由于现实的社交网络中，一个人常常会关注许多热门的微博用户，而被同时关注的情况较少。因此现实世界中的社交网络常常呈现稀疏化特征。我们需要对邻接矩阵 A 进行稀疏化处理，同时将矩阵转化为非对称矩阵。我们引入稀疏化因子 d_{ij}，独立同分布于二项分布，且 $P(d_{ij}=1)=0.5$。令邻接矩阵的每个元素都乘上稀疏化因子 $a_{ij}=d_{ij}a_{ij}$。由于社交网络自我关注的功能设置，我们令对角线元素全为 0，即 $a_{ii}=0$。从而得到非自循环的、有方向的、稀疏的邻接矩阵 A。对 A 进行行正态化处理，则得到最终的 W。

②其余参数和数据设定。ε 每个分量独立分布于均值为 0，标准差为 0.05 的正态分布。X 每个分量相互独立分布于均值为 1，标准差为 0.1 的正态分布。β 取 0 或 1，ρ 取 0 或 0.2。可以通过式（2）计算出 Y 的值。

（2）滚雪球抽样法进行抽样

滚雪球抽样法（Snowball Sampling，简称 SNOW）最初由 Goodman（1961）提出。由于其可以最大程度保存网络的拓扑结构，被广泛应用于许多社会学和市场学的研究（Henry 2005；Salganik & Heckathorn 2004；Tepper 1994；Frenzen & Davis 1990；Chen et al. 2013；Zhou et al. 2015）。在由 N 个用户构成的全网络中，我们随机的选取一个用户 i，记为 $U_1=\{i\}$。下一步抽取与这个用户相连的所有用户，即满足条件 $a_{ij}=1$ 的所有用户 $U_2=\{j:a_{ij}=1\}$。第 k 步的抽取的用户即为与 k-1 步抽取的每个用户相连的所有用户，即 $U_k=\{j:a_{ij}=1,i\in U_{k-1}\}$。重复以上迭代过程直至累积的用户量达到预定的样本规模 n，结束抽样。如果累积抽取的用户量超过预定样本规模 n，则在最后一步抽样中随机舍弃掉一些样本点使得累积抽取用户量恰为预定样本规模 n，结束抽样。这里我们设定样本规模为总体规模的 10%。

（3）计算成对最大似然估计量

根据式（25）可以计算出成对似然估计。并将 $\hat{\rho}$ 带入式（6）得到 $\hat{\beta}^*$，再通过式（26）变换得到 $\hat{\beta}$。

4.2 模拟结果

本文通过数值模拟主要验证三点问题。（1）全网络数据下，本文提出的方法 MPMLE 是否具有一些良好的性质。取 β 为 0 或 1，ρ 为 0 或 0.2，分别比较本文提出的估计量 MPMLE 与 Zhou et al.（2015）提出的估计量 PMLE、本文提出的估计量 MPMLE 与极大似然估计量 MLE 的有限样本性质。其中，前两者网络规模 N 为 5000，后两者的设网络规模 N 为 1000。（2）样本网络数据下，采用滚雪球抽样以及真实的度修正估计量后，本文提出的估计量是否仍具有一些良好性质。设网络规模 N 为 5000。取 β 为 0 或 1，ρ 为 0 或 0.2，分别基于全网络数据和滚雪球抽样产生的样本数据比较本文提出的估计量 MPMLE 和 Zhou et al.（2015）提出的估计量 PMLE。在模拟（1）（2）中，模拟均基于相同的全网络结构，每次模拟重复产生 X、ε 计算估计量，模拟 100 次比较两个估计量的均值（Mean）、偏差（Bias）、方差（Ese）和均方误差（Mse），结果见

表 1、表 2、表 3。（3）本文提出的方法是否有效的减小了计算量，与全似然函数对比，看耗时上是否有缩减。在模拟（3）中，设置 N=2000。固定网络结构，重复模拟 10 次，每次不进行抽样，计算每次模拟数据下的本文提出的估计量 MPMLE 和全网络极大似然估计 MLE 耗时。

值得注意的是，在模拟（1）中将本文提出的估计 MPMLE 与极大似然函数相比较时，设置全网络规模较小，为 1000。有两点原因，一是在这样的网络规模下，本文提出的估计量在数值模拟中已取得了不错的估计效果，和极大似然估计效果差异不大。二是在网络规模较大时，极大似然估计计算量较大，计算速度很慢且普通电脑难以进行模拟计算。并且计算复杂度随网络规模 N 增大并非是线性增大。举例而言，N=1000 时一次模拟耗时约 214 秒，而 N=2000 时一次模拟耗时约 1553 秒。当 N 翻倍增长时，相应耗时约为原来的 7 倍以上。

（1）基于全网络估计量效果比较

当基于全网络数据进行估计时，$\beta=0$代表真实模型为 $Y=\rho WY+\varepsilon$，不含外生变量 X 的空间自回归模型 pSAR（Pure Spatial Autoregression），$\beta=1$代表真实模型为$Y=\rho WY+X\beta+\varepsilon$，是混合空间自回归模型 mSAR（Mixed Spatial Autoregression）。理论上，当$\beta=0$时，MPMLE 估计量会退化到 PMLE，两者估计效果应类似。当$\beta=1$时，由于 PMLE 模型假设错误，忽略了外生变量 X 的效应，因而估计有偏。此时 MPMLE 的估计应优于 PMLE 估计。$\rho=0$代表不存在空间自相关性，真实模型为普通线性回归，MPMLE 和 PMLE 不仅对ρ应该有一个好的估计。MPMLE 对β的估计应退化到一般的线性回归的情况。

表 1：全网络数据下 MPMLE 和 PMLE 下估计效果比较

		pure SAR（beta=0）			mixed SAR（beta=1）		
		PMLE	MPMLE		PMLE	MPMLE	
		rho	rho	beta	rho	rho	beta
rho=0	Mean	-0.004520	-0.004540	0.000072	-0.004890	-0.004900	1.002510
	Bias	-0.004520	-0.004540	0.000072	-0.004890	-0.004900	0.002510
	Ese	0.034618	0.034769	0.011563	0.032818	0.032761	0.004742
	Mse	0.001207	0.001217	0.000132	0.001090	0.001087	0.000029
rho=0.2	Mean	0.201192	0.201946	-0.000518	0.184211	0.204681	0.991505
	Bias	0.001192	0.001946	-0.000518	-0.015790	0.004681	-0.008495
	Ese	0.035371	0.034626	0.012372	0.028399	0.031439	0.009005
	Mse	0.002480	0.002382	0.000304	0.001048	0.001000	0.000152

由上表可以发现，当实际模型是不含外生变量的空间自回归时（$\beta=0$），两种估计方法的估计效果相差不大，均为无偏估计。MPMLE 的偏差较大，但均方误差较小，总体言相差不大。当实际模型是含有外生变量的混合空间自回归时（$\beta\neq0$），本文提出的估计方法 MPMLE 是优于 PMLE 的。PMLE 此时估计有偏。而本文提出的对β的估计均有很好的无偏性和有效性。

由箱线图可以直观的呈现估计量的分布。左侧的图表示$\rho=0$，右侧图表示$\rho=0.2$。图的横轴分别为 pure SAR（pSAR）和 mixed SAR（mSAR）。红色表示本文估计量 MPMLE，蓝色表示 Zhou et al.（2015）提出的估计量 PMLE。

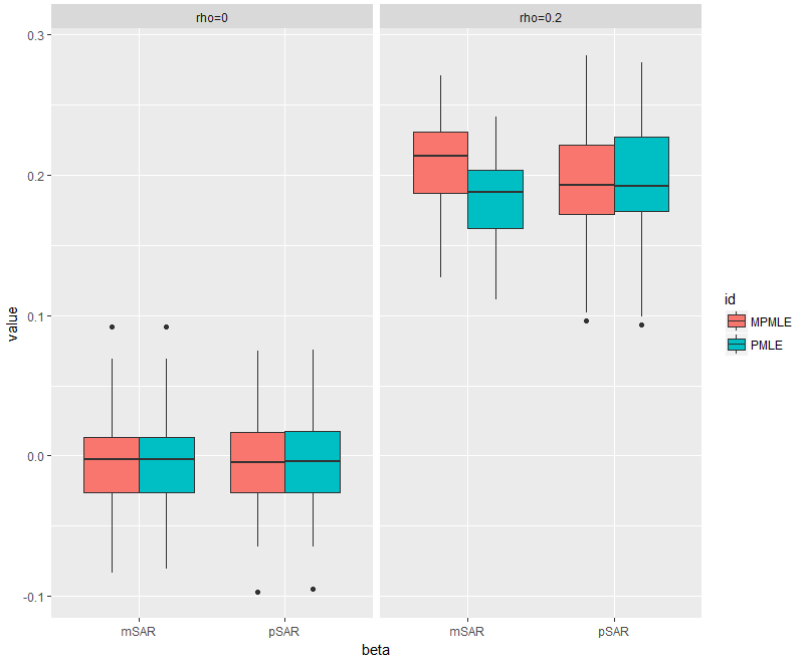

图 4：样本网络数据下 MPMLE 和 PMLE 下估计效果比较

取网络规模 N=1000，将本文的估计量 MPMLE 与极大似然估计量 MLE 进行比较，结果如下：

表 2：全网络数据下 MPMLE 和 MLE 下估计效果比较

		pure SAR(beta=0)				mixed SAR(beta=1)			
		MLE		MPMLE		MLE		MPMLE	
		rho	beta	rho	beta	rho	beta	rho	beta
rho=0	Mean	0.001279	-0.000087	0.001047	0.000463	0.000200	0.999721	0.003920	1.015716
	Bias	0.001279	-0.000087	0.001047	0.000463	0.000200	-0.000279	0.003920	0.015716
	Ese	0.048845	0.001575	0.048562	0.018515	0.007815	0.007352	0.045887	0.006447
	Mse	0.002364	0.000002	0.002336	0.000340	0.000061	0.000054	0.002100	0.000288
rho=0.2	Mean	0.200338	-0.000087	0.199555	0.000470	0.200173	0.999706	0.213694	1.003133
	Bias	0.000338	-0.000087	-0.000445	0.000470	0.000173	-0.000294	0.013694	0.003133
	Ese	0.048386	0.001577	0.048184	0.018489	0.006336	0.007435	0.043826	0.014216
	Mse	0.002318	0.000002	0.002299	0.000339	0.000040	0.000055	0.002089	0.000210

由上表可以发现，总体而言，尽管 MPMLE 普遍表现都不及 MLE，但是相差极小，仍保留了有限样本下的无偏性和有效性。特别地，当实际模型是不含外生变量的空间自回归时（$\beta=0$），MPMLE 对 ρ 的估计略优于 MLE 对 ρ 的估计，但对 β 的估计不及 MLE。结合模拟（3）得出，MPMLE 计算量远小于 MLE。因而，MPMLE 较 MLE 而言，牺牲一点点估计的精确性换取了计算量上的大幅削减。

（2）基于滚雪球抽样的样本网络下估计量的比较

全网络规模为 5000。取样本网络规模为总体的 10%。Leskovec & Faloutsos（2006）指出样本量在总体规模的 15%左右一般就够了。每次模拟通过滚雪球抽样方式得到样本网络数据，基

于真实的度计算估计量，模拟 100 次，得到模拟结果如下表：

表 3：样本网络数据下 MPMLE 和 PMLE 下估计效果比较

		pure SAR（beta=0）			mixed SAR（beta=1）		
		PMLE	MPMLE		PMLE	MPMLE	
		rho	rho	beta	rho	rho	beta
rho=0	Mean	-0.025599	-0.025093	0.002161	-0.001916	-0.000863	0.990343
	Bias	-0.025599	-0.025093	0.002161	-0.001916	-0.000863	-0.009657
	Ese	0.136139	0.131197	0.024085	0.114118	0.112730	0.010185
	Mse	0.019004	0.017670	0.000579	0.012896	0.012582	0.000196
rho=0.2	Mean	0.206017	0.203576	-0.002608	0.185817	0.186661	0.979614
	Bias	0.006017	0.003576	-0.002608	-0.014183	-0.013339	-0.020386
	Ese	0.142823	0.137968	0.024638	0.127739	0.130076	0.011261
	Mse	0.020231	0.018858	0.000608	0.016355	0.016929	0.000541

整体上看，两种基于样本网络数据的估计都表现良好，仅在 $\rho=0.2$，$\beta=1$ 的情形下出现了较大程度的低估，但偏差仍在可容忍范围以内。总体来说，MPMLE 的低估程度和方差均优于 PMLE。由图 5 可以直观的呈现估计量的分布。左侧的图表示 $\rho=0$，右侧图表示 $\rho=0.2$。每一图层的横轴分别为 pure SAR（pSAR）和 mixed SAR（mSAR）。红色表示本文估计量 MPMLE，蓝色表示 Zhou et al.（2015）提出的估计量 PMLE。

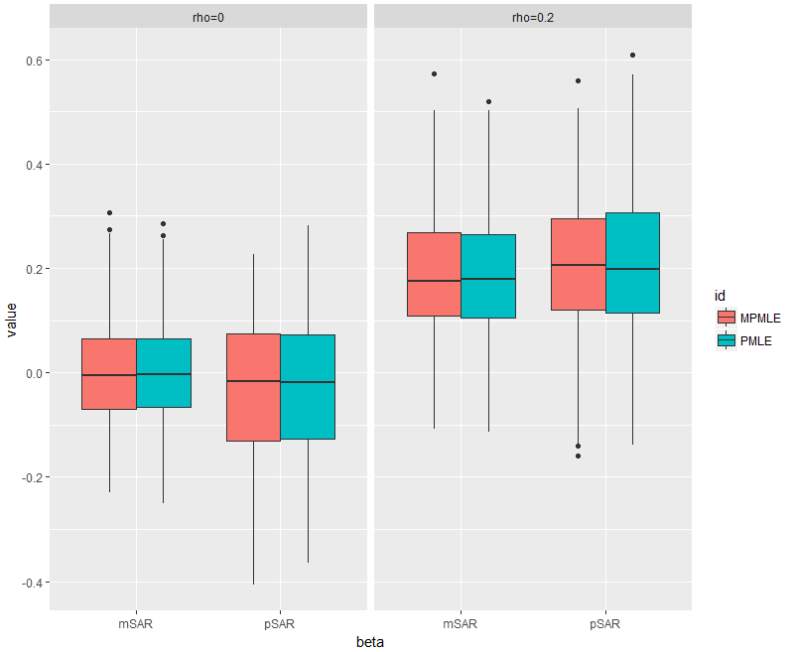

图 5：全网络数据下 MPMLE 和 PMLE 估计方法下ρ̂估计效果比较

但是与全网络相比，MPMLE 在混合空间自回归模型下的领先效应不再那么明显。与全网络估计相比，基于样本网络的估计偏差和方差都较大。在图 6 对比的箱线图中较为直观的呈现出来。左侧的图表示 $\beta=0$（pure SAR），右侧图表示 $\beta=1$（mixed SAR）。每一图层的横轴分别为 $\rho=0$ 和 $\rho=0.2$。红色表示基于全网络数据的估计量 $\hat{\beta}$，蓝色表示基于样本网络数据的估计量 $\hat{\beta}$。

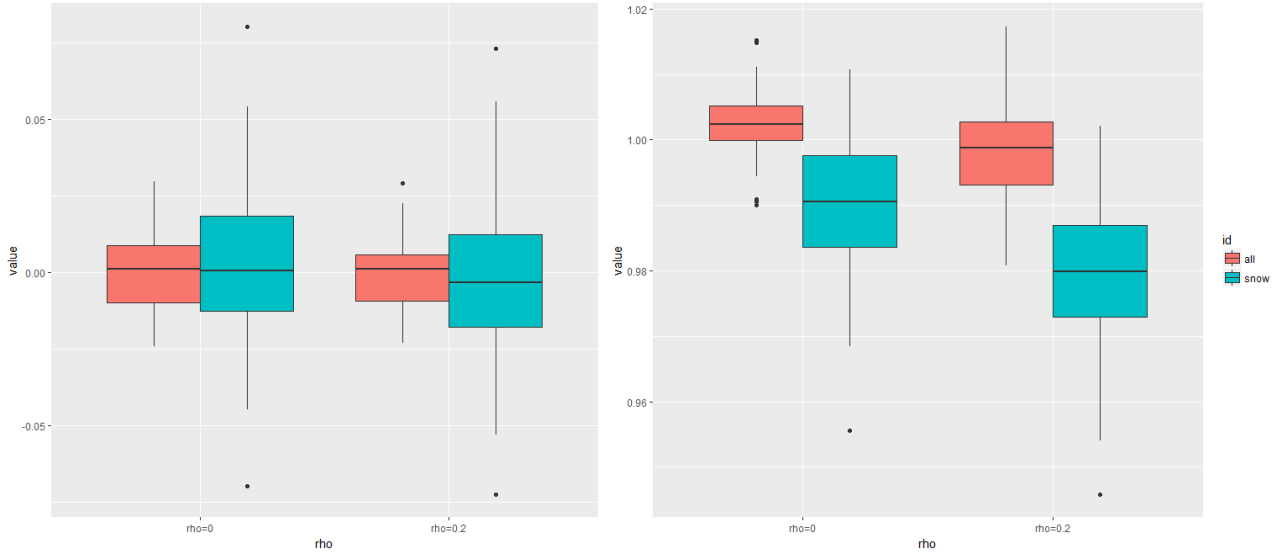

图 6：全网络数据和样本网络数据下 MPMLE 估计方法下β̂的估计效果

（3）计算耗时比较

除此外，我们比较了极大似然函数和成对似然函数的运行时间。设置 N=2000。固定网络结构，重复模拟 10 次，每次不进行抽样，计算每次模拟数据下的本文提出的估计量 MPMLE 和全网络极大似然估计 MLE 耗时。

表 4：MLE 和 MPMLE 耗时比较（单位：秒）

	用户	系统	流逝
MLE	1646.49	17.52	1669.89
MPMLE	83.08	1.22	91.74

可以发现全网络极大似然的耗时为本文估计方法近 20 倍。可见本文提出的估计方法在保证了估计量的良好性质的前提下，极大的减小了计算量。

通过以上模拟说明：（1）本文提出的估计方法在基于样本网络数据计算时，不仅极大减小了计算量，还保留了无偏性和有效性。（2）当实际网络数据源自空间自回归模型 pure SAR 时，本文提出的估计方法 MPMLE 的表现大部分优于 Zhou et al.（2015）提出的 PMLE 的表现。（3）当实际网络数据源自混合空间自回归模型 mixed SAR 时，本文的估计方法 MPMLE 明显优于 PMLE，并且 PMLE 此时为有偏估计。在现实生活中，一个人的决策并不仅仅依赖于他人的影响，很大部分是由个人特质决定，因而混合空间自回归应为更常见的形式。选用本文的估计量会更佳。

5 案例分析

本文利用滚雪球抽样法从新浪微博中爬取用户信息构成人际关系网络进行案例分析。微博作为社交网络平台，每个用户都可以在这里通过发言、点赞、转发、评论表达个人态度和意见。每个用户都可以关注他人发布的信息，同时也成为互联网上的信息而被关注，是以节点的延伸与扩张而互动新信息的生产，具有传播多向性和反馈的同步及时性等特征。同时，微博这类自

媒体可以通过共同话题形成人际信息交流的圈子，具有“圈子化”特征。本文主要关注于用户接收到的反馈（微博点赞数）和人际关系网络对用户的表达行为（发布微博数）的影响。以用户的发布的微博数为因变量 Y，以用户最近 5 条微博的点赞总数为外生变量 X，人际关系网络用 W 表示。建立含有外生变量的空间自回归模型估计参数。

$$Y = \rho WY + X\beta + \varepsilon$$

在实际数据收集过程中，采用的抽样方式在滚雪球抽样的基础上多增加了一个阈值设定。由于新浪微博的公众人物常有数以百万的粉丝，在样本量较小时，采用严格的滚雪球抽样会使得样本节点全来自于这一用户的粉丝，不具有代表性。因此，我们在滚雪球抽样中设置了一个阈值，最多抽取每个样本节点用户的 200 个粉丝用户纳入样本。具体的抽样过程如下：

以世界最快实现乒乓球大满贯得主张继科为起点，记为$U_1 = \{i\}$。从他的粉丝中随机抽取 200 个用户为样本节点，记为$U_2 = \{j : a_{ij} = 1\}$。接着，以这 200 个粉丝用户为起点，随机抽取每个用户下的粉丝为样本节点，并且保证每个用户下的粉丝纳入样本数目不超过 200。重复以上迭代过程直至累积的用户量达到预定的样本规模，结束抽样。

本文采用 Python 程序一共爬取了 1053 个用户数据，由于数据缺失，可用数据有 521 条。以黄色的点表示用户，以灰色线条表示两用户间存在联系，绘制如下的社会关系网络图。为了图形美观，省略了最外圈的节点和表征方向的箭头。因而两节点间相连的线条表示相互关注和单方面关注的三种状态。

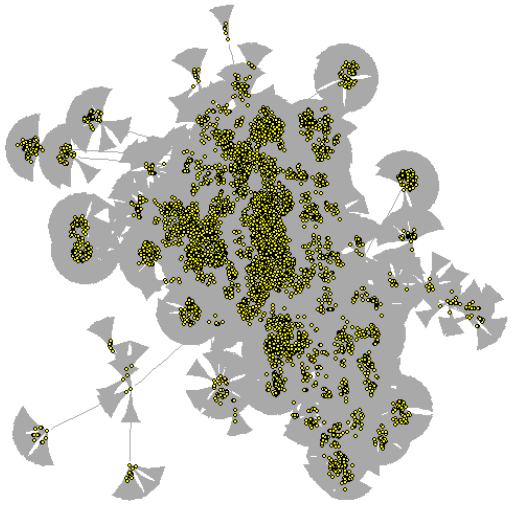

图 7：新浪微博人际网络关系图

采用本文的估计方法，计算得出 $\hat{\rho} = 0.064$， $\hat{\beta} = 4.885 \times 10^{-5}$。在以张继科为中心节点的人际网络中，回归系数 $\hat{\beta}$ 为正，表明他人的点赞对个人的发微博行为有正向促进作用。对某一微博点赞往往代表对这一微博反映的观点态度的认同，说明正向的反馈可以促进人们的个人表达，激励人们发布微博。自相关系数 $\hat{\rho}$ 为正，表明他人发微博行为对个人发微博行为有正向促进作用。发布微博这一行为常常代表一个人是否乐于自我表达，说明他人意见态度的表达会带动个

人表达。$\hat{\beta}$ 的值小于 $\hat{\rho}$，且两者相差 3 个数量级，说明 1 个点赞对个人表达的促进作用弱于 1 个关注者发布的一则微博的促进作用，这意味着正反馈对一个人的表达欲望的影响关注者发布的一则微博。这可能意味着个人表达行为更多受到个人对信息的接收行为（对一则微博中呈现态度的认同或反对）的影响，更少的受到他人反馈的影响。这也可能是由于点赞这一行为所蕴涵的个人表达的程度远弱于发布微博这一行为蕴涵的个人表达程度。这其中反映出的有意思的现象，需要基于更大的样本量和更严格的抽样设计予以验证。

6 讨论与思考

本文针对基于样本网络数据的极大似然估计有偏和计算量大的问题，提出了新的估计量有效的解决了这两方面问题。基于自相关系数很小的假定，本文构建成对极大似然函数并且利用一阶泰勒展式将其近似为空间自相关系数的二次函数求解。求解过程中同时利用了邻接矩阵的稀疏性有效地减少计算量。针对样本网络极大似然估计有偏的问题，本文利用社交网络中真实节点信息，并且采用滚雪球抽样法极大的保留了网络的拓扑结构，通过数值模拟证实了新的估计量具有无偏性和小计算量等优良性质，并且通过实际案例分析说明了本研究的重要实际意义。但是有几点值得注意之处：

（1）基于样本网络似然函数的有偏估计来自于每个样本节点真实的度被低估，对于微博等社交网络平台，度的真实数据（一个用户拥有的粉丝数和关注的用户数）极易获得。但是对于其他真实的度的数据不易获得的网络而言，这一修正方式是不适用的。

（2）若影响一个人决策的个人特质因素很多，即 $X \in R^{N\times P}$ 中 P 很大，则本文的估计方法会涉及到高维矩阵的求逆运算。成对似然函数难以避开帽子矩阵 H 的计算，而帽子矩阵的成对形式 $H=(X(X^TX)^{-1}X^T)_{i,j}=\begin{pmatrix} h_{ii} & h_{ij} \\ h_{ij} & h_{jj} \end{pmatrix}$ 涉及到了高维矩阵 X^TX 的求逆运算。因而当 X 为高维矩阵时，本文提出的估计量会面临着大计算量的问题。

除了上述两点不足，本文还有很多改进方向：

（1）本文提出的成对极大似然估计不仅保留的极大似然估计量的无偏性、有效性等优良性质，同时极大的减小了计算量。这一方法适用于现今的大规模网络数据。但是，本文只是采用了模拟证实成对极大似然估计量具有优良性质，这一块理论性质仍需要补充完善。

（2）滚雪球抽样法极大的保留了网络结构，使得基于样本的估计量呈现出较好的性质。但由于滚雪球抽样只抽取与上一样本点相连的所有节点，可能会造成自相关系数的高估。尽管 Chen et al.（2013）通过模拟证明，高估效应不明显。在本文模拟中也没有出现明显的高估效应。但是否采用高蔓延率的森林火势蔓延抽样（Forest Fire Sampling with high burning rate）会更利于保存网络结构需要进一步模拟探究。

作者签名：

参考文献

[1] Anselin L. Estimation methods for spatial autoregressive structures.[J]. Cornell University, 1980.
[2] Anselin L. Spatial Econometrics[M]// A Companion to Theoretical Econometrics. 2007:310-330.
[3] Anselin L. Under the hood Issues in the specification and interpretation of spatial regression models[J]. Agricultural Economics, 2002, 27（3）:247-267.
[4] Barry R P, Pace R K. Monte Carlo estimates of the log determinant of large sparse matrices[J]. Linear Algebra and its applications, 1999, 289（1-3）: 41-54.
[5] Besag J. Statistical analysis of non-lattice data[J]. The statistician, 1975: 179-195.
[6] Brueckner J K. Strategic interaction among governments: An overview of empirical studies[J]. International regional science review, 2003, 26（2）: 175-188.
[7] Chen Y, Chen X. The Impact of Sampling and Network Topology on the Estimation of Social Inter-Correlations[J]. Journal of Marketing Research, 2008, 50（1）: 95-110.
[8] Chen Y, Liu Q, Qian P Z G. Sequential Sampling Enhanced Composite Likelihood Approach to Estimation of Social Intercorrelations in Large-Scale Networks[J]. Social Science Electronic Publishing, 2014.
[9] Costenbader E, Valente T W. The stability of centrality measures when networks are sampled.[J]. Social Networks, 2003, 25（4）:283-307.
[10] Doreian P. Network autocorrelation models: Problems & prospects[C]// 1989.
[11] Ebbes P, Huang Z, Rangaswamy A. Sampling designs for recovering local and global characteristics of social networks ☆[J]. International Journal of Research in Marketing, 2016, 33（3）:578-599.
[12] Fienberg S E. A brief history of statistical models for network analysis and open challenges[J]. Journal of Computational and Graphical Statistics, 2012, 21（4）: 825-839.
[13] Franzese R J, Hays J C. Spatial Econometric Models of Cross-Sectional Interdependence in Political Science Panel and Time-Series-Cross-Section Data[J]. Political Analysis, 2007, 15（2）:140-164.
[14] Frenzen J K, Davis H L. Purchasing behavior in embedded markets[J]. Journal of Consumer Research, 1990, 17（1）: 1-12.
[15] Godambe V P. An optimum property of regular maximum likelihood estimation[J]. The Annals of Mathematical Statistics, 1960, 31（4）: 1208-1211.
[16] Goodman L A. Snowball Sampling[J]. Annals of Mathematical Statistics, 1961, 32（1）:148-170.
[17] Handcock M S, Gile K J. Modeling social networks from sampled data.[J]. Annals of Applied Statistics, 2010, 4（1）:5.
[18] Henry P C. Social class, market situation, & consumers' metaphors of （dis） empowerment[J]. Journal of Consumer Research, 2005, 31（4）: 766-778.
[19] Keith Ord. Estimation Methods for Models of Spatial Interaction[J]. Journal of the American Statistical Association, 1975, 70（349）:120-126.
[20] Kelejian H H, Robinson D P. A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model[J]. Papers in regional science, 1993, 72（3）: 297-312.
[21] Kelley R P, Barry R. Quick computation of regressions with a spatially autoregressive dependent variable[J]. Geographical Anal, 1997, 29: 232-247.
[22] Kossinets G. Effects of missing data in social networks[J]. Social Networks, 2003, 28（3）:247–268.
[23] Land K C, Deane G. On the Large-Sample Estimation of Regression Models with Spatial- Or

Network-Effects Terms: A Two-Stage Least Squares Approach[J]. Sociological Methodology, 1992, 22:221.
[24] Lee L F, Liu X, Lin X. Specification and estimation of social interaction models with network structures[J]. The Econometrics Journal, 2010, 13（2）:145-176.
[25] LeSage J P, Pace R K. A matrix exponential spatial specification[J]. Journal of Econometrics, 2007, 140（1）: 190-214.
[26] Leskovec J, Faloutsos C. Sampling from large graphs[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006:631-636.
[27] Lindsay B G. Composite likelihood methods[C]// Contemporary Math. 1988:221-239.
[28] Pace R K, Zou D. Closed-Form Maximum Likelihood Estimates of Nearest Neighbor Spatial Dependence[J]. Geographical Analysis, 2000, 32（2）: 154-172.
[29] Renard D, Molenberghs G, Geys H. A pairwise likelihood approach to estimation in multilevel probit models[J]. Computational Statistics and Data Analysis, 2004, 44（4）: 649-667.
[30] Roncek D W, Montomery A. Spatial autocorrelation: Diagnoses and remedies for large samples[C]//annual meeting of the Midwest Sociological Society, Des Moines, IA. 1984.
[31] Smirnov O, Anselin L. Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach[J]. Computational Statistics and Data Analysis, 2001, 35（3）: 301-319.
[32] Snijders T A B. Estimation on the basis of snowball samples: how to weight? [J]. Bulletin De Méthodologie Sociologique Bms, 1992, 36（1）:59-70.
[33] Tepper K. The role of labeling processes in elderly consumers' responses to age segmentation cues[J]. Journal of Consumer Research, 1994, 20（4）: 503-519.
[34] Thompson S K. Adaptive Web Sampling[J]. Biometrics, 2006, 62（4）:1224-1234.
[35] Varin C, Reid N, Firth D. An Overview of Composite Likelihood Methods[J]. Statistica Sinica, 2011, 21（1）:5-42.
[36] Wang X, Wu Y. Theoretical Properties of Composite Likelihoods[J]. Open Journal of Statistics, 2014, 04（3）:188-197.
[37] Xu X, Reid N. On the robustness of maximum composite likelihood estimate[J]. Fuel & Energy Abstracts, 2011, 141（9）:3047-3054.
[38] Zhou J, Tu Y, Chen Y, et al. Estimating Spatial Autocorrelation with Sampled Network Data[J]. Journal of Business & Economic Statistics, 2015.
[39] Most famous social network sites worldwide as of January 2017, ranked by number of active users （in millions） [EB/OL].https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ 2017 年 4 月 3 日访问