arXiv:2301.02871v1 [stat.ME] 7 Jan 2023

基于谱信息的网络数据模型选择

Jairo Ivan Pe˜na Hidalgo 和 Jonathan R. Stewart

佛罗里达州立大学

Abstract: 我们提出了一种用于网络数据建模的新方法，以进行模型选择。统计网络分析领域的文献已经发展出许多不同类别的网络数据模型，其中值得关注的模型类别包括随机块模型、潜在位置模型和随机图模型的指数族。统计网络分析领域的文献始终关注一个问题：如何比较不同模型，以实现模型选择并评估拟合优度，尤其是在模型具有不同数学基础的情况下。在这项研究中，我们针对网络数据场景开发了一种新颖的非参数模型选择方法，该方法利用图拉普拉斯矩阵谱中包含的信息，为一组既定的网络数据模型获取拟合优度度量。我们通过大量模拟研究，考察了所提出的方法在常见网络数据模型类别中的表现，并通过两个应用展示了该方法的实用价值。

Key words and phrases: 统计网络分析、网络数据、模型选择、社会网络分析。

1. 引言

近几十年来，网络数据在众多领域和学科中受到越来越多的关注，包括社会网络研究（Lusher 等人，2013）、网络流行病学（研究疾病如何通过接触网络传播）（Morris，2004）、犯罪活动和恐怖主义的隐蔽网络（Coutinho 等人，2020）、脑网络（Obando 和 de Vico Fallani，2017）、金融市场（Finger 和 Lux，2017）等。作为一种数据结构，网络数据通常表示为一个图（Kolaczyk，2009），由一组表示感兴趣总体中各元素的节点（例如，合作网络中的研究人员）以及一组节点间的成对观测或测量构成，并以节点之间的边表示（例如，在论文上的共同署名）。

人们提出并发展了许多类模型，用于研究和建模网络数据。以下简要回顾一些较为突出的例子，其中包括指数族

随机图模型（指数随机图模型）Lusher 等人，2013；Schweinberger 等人，2020）、随机块模型（随机块模型）（例如，Holland 等人，1983；Anderson 等人，1992；Wang 和 Bickel，2017）、潜在位置模型（例如，Hoff 等人，2002；Sewell 和 Chen，2015；Tang 等人，2013；Athreya 等人，2017）等。每一类模型都为基于观测网络数据构建网络模型提供了独特的数学框架，各有优缺点。指数族模型为构建具有相关边的网络模型提供了灵活的参数化框架。相比之下，随机块模型通过离散潜在空间捕捉网络结构和节点聚类，而潜在位置模型则通过节点在（例如）潜在欧几里得空间中的潜在位置，捕捉网络结构和边的依赖性。

统计网络分析应用中一个长期存在的挑战，是如何比较不同模型，并为特定网络数据集选择模型。目前，文献主要关注各类模型内部的模型选择问题，针对特定模型类别定制方法（随机块模型：Wang 和 Bickel (2017年); Latouche 等人(2014); 指数随机图模型：Hunter 等人(2008年); 潜在空间模型：Ryan 等人 (2017））。因此，文献中尚缺乏探讨如何比较不同数学体系下模型拟合优度或进行模型选择的方法，例如比较指数随机图模型、随机块模型和潜在位置模型。本文提出一种用于网络数据场景下模型选择的新型非参数方法。在较弱假设下，该方法可应用于广泛的模型类别，并能够促进具有不同数学基础的模型之间的比较。我们利用图拉普拉斯矩阵的谱信息，为观测网络选出拟合最佳的模型；该方法本质上只要求能够从候选模型中模拟生成邻接矩阵，并计算由这些邻接矩阵导出的图拉普拉斯矩阵的特征值。

本文其余部分安排如下。第2 节回顾网络的图拉普拉斯矩阵的谱性质，并阐述在网络数据模型选择问题中使用谱信息的动机。我们提出的方法将在第3节介绍。第4节将展示实验研究和模拟，并在第5节介绍该方法的两个应用。

2. 图拉普拉斯矩阵的谱性质

我们考虑定义在一组 N 节点上的简单无向网络，其对应的邻接矩阵为 $\boldsymbol{X} \in \{0,1\}^{N\times N}$，其中 $X_{ij} = 1$ 表示节点 i 和 j 之间存在一条边，而 $X_{ij} = 0$ 则表示不存在边。我们采用标准约定：$X_{ij} = X_{ji}$ 和$X_{ii} = 0$。我们在第 4 节讨论将本方法扩展到有向网络的情形。也可以将其扩展到边具有权值的网络，但这超出了本文的讨论范围。令 $\boldsymbol{d} = \deg(\boldsymbol{X}) = (\sum_{j=1}^{N} X_{i,j} : i = 1,\ldots,N) \in \mathbb{R}^N$ 表示该网络的节点度向量。拉普拉斯矩阵（也称为图拉普拉斯矩阵）定义为 $\boldsymbol{L}(\boldsymbol{X}) := \mathrm{diag}(\boldsymbol{d}) - \boldsymbol{X}$，其中 $\mathrm{diag}(\boldsymbol{d})$ 是一个 $N\times N$ 对角矩阵，其对角线为 $\boldsymbol{d}$。由于 $\boldsymbol{L}(\boldsymbol{X})$ 是对称半正定矩阵（Brouwer

和 Haemers，2012），$\boldsymbol{L}(\boldsymbol{X}$的特征值均为实数且非负。在全文中，我们用 $\boldsymbol{\lambda} \in \mathbb{R}^N$ 表示拉普拉斯矩阵 $\boldsymbol{L}(\boldsymbol{X}$按从小到大排序的特征值向量。向量 $\boldsymbol{\lambda}$ 取决于邻接矩阵 $\boldsymbol{X}$，具体通过 $\boldsymbol{L}(\boldsymbol{X}$体现。不过，为方便表述，我们不在符号上明确表示这种依赖关系，因为上下文会清楚地表明这一点。

拉普拉斯矩阵的特征值编码了网络的许多已知性质。例如，特征值 0 的重数对应于网络中连通分量的数量（Brouwer 和 Haemers，2012）。第二小的特征值（可能为 0）称为代数连通度（Fiedler，1973），用于衡量图的整体连通性（deAbreu，2007）。它还用于建立 Cheeger 不等式（Donetti 等人，2006），这些不等式在图像分割（Shi 和 Malik，2000）、图聚类（Kwok 等人，2013）和扩展图（Hoory 等人，2006）中都有应用。拉普拉斯矩阵的后续特征值与划分网络所需的最小割（删除加权边）有关（Bollob´as 和 Nikiforov，2004）。

具有邻接矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的两个无向图同构，当且仅当存在一个置换矩阵 $\boldsymbol{P}$使得 $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^t$，这要求邻接矩阵相似 $\boldsymbol{A} = \boldsymbol{P}\,\boldsymbol{B}\boldsymbol{P}^{-1}$。注意，置换矩阵 $\boldsymbol{P}$满足 $\boldsymbol{P}^t = \boldsymbol{P}^{-1}$。在这种情况下，相应的图拉普拉斯矩阵也相似：

$$L(\boldsymbol{A}) \;=\; \deg(\boldsymbol{P}\boldsymbol{B}\boldsymbol{P}^t) - \boldsymbol{P}\,\boldsymbol{B}\,\boldsymbol{P}^t \;=\; \boldsymbol{P}\deg(\boldsymbol{B})\boldsymbol{P}^t - \boldsymbol{P}\,\boldsymbol{B}\,\boldsymbol{P}^t \;=\; \boldsymbol{P}\,L(\boldsymbol{B})\,\boldsymbol{P}^t.$$

因此，由于 $L(\boldsymbol{B})$ 是厄米矩阵，存在特征分解 $L(\boldsymbol{B}) = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^t$。因此，$L(\boldsymbol{A}) = \boldsymbol{P}(\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^t)$ $\boldsymbol{P}^t = (\boldsymbol{P}\,\boldsymbol{U})\;\boldsymbol{D}(\boldsymbol{P}\,\boldsymbol{U})^t$。由此可知，如果 $\boldsymbol{\lambda}$是 $L(\boldsymbol{B})$ 的特征值向量，它也是 $L(\boldsymbol{A})$ 的特征值向量。在我们的语境中，这意味着只要两个非同构网络的特征值不同，就总能将它们区分开来。反过来的结论一般并不成立。有些图具有相同的特征值分解（称为余谱或等谱），但它们并不同构（Cvetkovi´c et al., 1980）。然而，数值证据表明，随着图中节点数量的增加，余谱图（非同构图）所占的比例趋于零（Brouwer 和 Haemers, 2012）。

网络分析文献中已提出拉普拉斯矩阵谱分解的若干应用。例如，谱聚类（Von Luxburg，2007）是一种广为人知的聚类算法，它基于相似度矩阵的拉普拉斯矩阵的主特征向量。Lei 和 Rinaldo (2015) 在温和条件下证明了随机块模型中谱聚类方法的一致性。另一个例子见于 Newman(2006)，其中提出了一类基于图拉普拉斯矩阵谱分解的网络社区检测算法。最后，Shore 和 Lubin(2015) 提出了一种用于评估网络模型拟合优度的统计量，其形式让人想起回归情形下的 R^2 统计量，

该统计量将模型拟合所生成的图拉普拉斯矩阵的特征值与预先指定的 $null$模型所对应的图拉普拉斯矩阵的特征值进行比较（通常采用伯努利随机图模型，称为仅密度模型）。

鉴于这些结果，自然可以将特征值向量 $\boldsymbol{\lambda}$ 视为网络的特征标识，其中包含重要的拓扑和结构信息，可用于模型选择。我们提出的方法将候选模型的图拉普拉斯矩阵谱的经验分布与观测网络的谱分布进行比较。我们提出的方法基于以下有关图拉普拉斯矩阵性质的考量。

首先，如果真实数据生成过程属于候选模型之列，那么从观测网络中得到的观测特征值预计会落在该数据生成过程的谱分布范围内。实际中，若所提出的模型均不是真实生成过程，仍可根据候选模型捕捉观测图拉普拉斯矩阵谱的能力对其进行评估，从而为模型选择方法的开发提供依据。其次，可以根据候选模型对观测图拉普拉斯矩阵谱的捕捉效果，对相互竞争的模型进行相对拟合度评估；这样不仅能够选出拟合最佳的模型，还能将该模型的拟合效果与未选中的备选模型进行比较。第三，我们的方法不需要对数据生成过程作任何参数假设，并且能够比较基于不同数学框架构建的模型，包括似然函数定义不明确的模型，或通过随机过程构建的模型；基于代理的模型（例如，Snijders 等人，2010；Jackson 和 Watts，2002）以及基于优先连接模型的生成算法（例如，Barabasi 和 Albert，1999；Zeng 等人，2013）便是其中的例子。

3. 方法

我们概述一种用于网络数据场景下模型选择的方法，该方法受到上一节相关考虑的启发，利用图拉普拉斯矩阵的谱性质。我们始终假设网络已被完全观测，记为 $\boldsymbol{X}_{\text{obs}}$。观测图拉普拉斯矩阵 $\boldsymbol{L}(\boldsymbol{X}_{\text{obs}})$ 对应的观测特征值向量记为 $\boldsymbol{\lambda}_{\text{obs}}$。我们的核心推断目标是从一组候选模型 $\boldsymbol{X}_{\text{obs}}$ 中，为观测网络 $\{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$ $(M \geq 2)$ 选择拟合效果最佳的模型。我们将这一问题表述为分类问题，并旨在构建分类器 $\mathcal{P}\colon \mathbb{R}^N \mapsto \{1, \ldots, M\}$，该分类器针对每个候选模型，以图拉普拉斯矩阵的谱作为训练数据，从而根据给定的特征值向量预测类别 $m^\star \in \{1, \ldots, M\}$，即 $\boldsymbol{\lambda}_{\text{obs}}$。我们的模型选择方法算法见表 1。

模型选择程序：

1. 模拟 K 网络 $\boldsymbol{X}^{(m,1)}, \ldots, \boldsymbol{X}^{(m,K)}$ 来自每个候选模型 model $\mathcal{M}_m \in \{\mathcal{M}_1, \ldots \mathcal{M}_M\}$。2. 对于每个 $\boldsymbol{X}^{(m,k)}$，计算拉普拉斯矩阵 $\boldsymbol{L}(\boldsymbol{X}^{(m,k)})$及相应的特征值向量 $\boldsymbol{\lambda}^{(m,k)} \in \mathbb{R}^N$。3. 将 KM个特征值向量 $\boldsymbol{\lambda}^{(m,k)}$ 堆叠起来作为 $\boldsymbol{D}$ 的各行，构建设计矩阵 $\boldsymbol{D} \in \mathbb{R}^{(KM)\times N}$。4. 训练分类器 $\mathcal{P}$：$\mathbb{R}^N \mapsto \{1, \ldots, M\}$，利用设计矩阵 $\boldsymbol{D}$中每个类别 $m \in \{1, \ldots, M\}$ 对应的 K 个模拟特征向量 $\boldsymbol{\lambda}^{(m,k)}$，预测模型 $m^\star \in \{1, \ldots, M\}$。此阶段建议进行特征工程。5. 计算观测网络 $\boldsymbol{X}_{\text{obs}}$ 的拉普拉斯矩阵 $\boldsymbol{L}(\boldsymbol{X}_{\text{obs}})$及相应的特征值向量 $\boldsymbol{\lambda}_{\text{obs}}$。6. 使用步骤 4 中训练的分类器为观测网络预测类别 $m^\star = \mathcal{P}(\boldsymbol{\lambda}_{\text{obs}})$，并设置 $\mathcal{M}^\star = \mathcal{M}_{m^\star}$。

表 1：模型选择算法说明。

3.1 分类器选择

现实网络可能包含数百、数千乃至数百万个节点。由于图拉普拉斯矩阵的特征值向量的维度等于网络中的节点数，基于拉普拉斯矩阵特征值的分类方法容易受到高维分类问题常见缺陷的影响。分类方法的相关文献相当丰富，因此，选择分类器是我们方法中的关键步骤。不过，我们将在第 4 节说明，在某些情况下，分类器选择对我们方法结果的影响可能并不显著。鉴于这些结果，我们将考虑选择分类器时的实际实施问题。

线性判别分析需要计算协方差矩阵的逆矩阵。实践中已发现，当变量数量增加而样本量固定时，其性能会下降（Bickel和Levina，2004）。其他方法包括支持向量机、神经网络、随机森林和提升算法，这些方法通常在高维场景下表现良好（Hastie等人，2011）。此类方法中包括极端梯度提升（XGBoost），该方法兼具可扩展性和先进的性能（Chen和Guestrin，2016）。本文其余部分均使用XGBoost，唯一的例外是第4节的模拟研究5。在该研究中，我们比较不同分类器的性能，以验证本节前文提出的论断。

3.2 拟合优度的相对度量

许多分类算法不仅会给出预测类别，还常常会返回一个倾向得分向量 $\boldsymbol{s} = (s_1, \ldots\ldots, s_M)$，其性质为 $\|\boldsymbol{s}\|_1 = 1$。如果考虑了多个模型，那么许多模型的倾向得分可能会仅仅因为所考虑的类别数量更多而缩小，这意味着对倾向得分 $s_1, \ldots\ldots, s_M$的解释可能取决于 M。为解决这一问题并便于比较不同模型的拟合情况，我们提出对倾向得分进行归一化，从而获得一种不受候选模型数量 M影响的拟合优度度量。为此，我们定义

$$\tilde{s}_i = \frac{s_i}{\|\boldsymbol{s}\|_\infty}, \qquad i = 1, \ldots, M,$$

即归一化得分，得分最高的模型对应的归一化得分为 1。对于所有其他模型，归一化得分衡量该模型相对于得分最高模型的拟合情况。以这种方式重新缩放所有倾向得分后，候选模型集中所考虑的模型数量 M 不会影响对（相对）倾向得分的解释。

4. 模拟研究

我们开展了若干模拟研究，以展示所提出方法的潜力。具体而言，我们旨在考察网络的特征在多大程度上包含于图拉普拉斯矩阵的谱中。模拟研究使我们能够知晓真实数据生成模型，从而开展实证研究，明确所提出的方法在何种条件下能够成功区分不同的网络模型及网络结构属性。

4.1 模拟研究1：曲指数族

我们研究了我们的方法在曲指数族上的表现。曲指数族在社会网络分析领域日益受到关注（例如，Snijders 等人，2006；Hunter 和 Hand- cock，2006），也应用于其他领域（例如，Obando 和 de Vico Fallani，2017；Schweinberger 等人，2020；Stivala 和 Lomi，2021）。随机图模型采用曲指数族参数化，源于解决早期且设定不当的模型设定所面临的退化及拟合难题的需要（Snijders 等人，2006）。此外，曲指数族参数化能够以简约的方式对复杂的图统计量序列进行建模，例如度序列和共享伙伴序列，同时不牺牲可解释

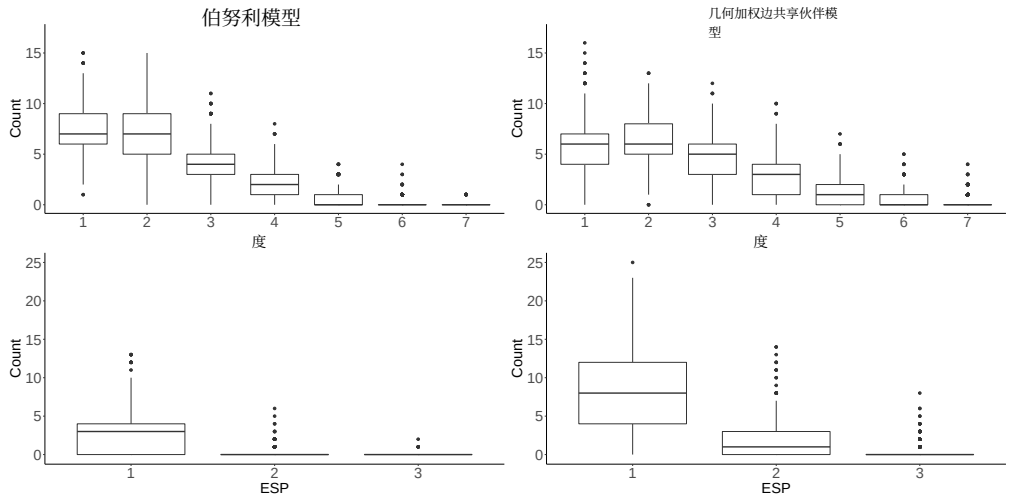

图1：我们通过从（1）模拟 1000 个网络，展示伯努利模型和几何加权边共享伙伴模型的度分布和边共享伙伴分布，其中数据生成参数向量为（θ_1，θ_2，θ_3）$=$（-2.5，0，1）（伯努利）和（θ_1，θ_2，θ_3）$=$（-2.5，$.3$，1）（几何加权边共享伙伴模型）。每一列对应一个模型；与伯努利模型相比，我们可以看到几何加权边共享伙伴模型的度分布和边共享伙伴分布向右偏移。

性（Hunter，2007；Stewart 等人，2019年）。社会网络分析文献中一个典型的例子是几何加权边共享伙伴模型，该模型通过共享伙伴序列对传递性进行建模（Snijders 等人，2006；Hunter，2007；Stewart 等人，2019年）。

我们根据以下模型模拟网络：

$$\mathbb{P}(\boldsymbol{X}=\boldsymbol{x}) \quad \propto \quad \exp\left(\theta_1 \sum_{i<j}^{N} x_{i,j} + \sum_{t=1}^{N-2}\sum_{i<j}^{N} \eta_t(\theta_2,\theta_3)\,\mathrm{SP}_t(\boldsymbol{x})\right), \tag{1}$$

其中，$\theta_1 \in \mathbb{R}$ 控制边形成的基线倾向，而

$$\eta_t(\theta_2,\theta_3) \quad = \quad \theta_2\,\exp(\theta_3)\,[1-(1-\exp(-\theta_3\,t)]\,, \qquad t \in \{1,\ldots,N-2\},$$

则用于参数化共享伙伴统计量序列

$$\mathrm{SP}_t(\boldsymbol{x}) \quad = \quad \sum_{i<j}^{N} x_{i,j}\,\mathbb{1}\left(\sum_{h\neq i,j}^{N} x_{i,h}\,x_{h,j} \;=\; t\right), \qquad t \in \{1,\ldots,N-2\}.$$

换言之，$\mathrm{SP}_t(\boldsymbol{x})$统计网络中恰有$t$ 个相互连接的节点之间的边数；在社会网络分析文献中，这些相互连接通常称为共享伙伴。尽管 $\theta_2 \in \mathbb{R}$，在典型应用中，$\theta_2 \geq 0$和 $\theta_3 \in (0,\infty)$作为 $\theta_3 < -$的取值时，log 2 对应

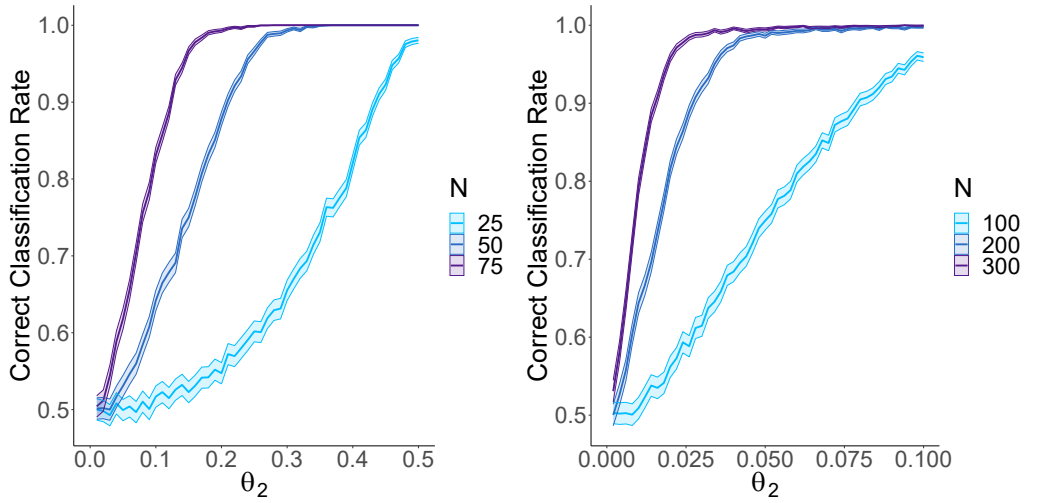

图2：模拟研究1的结果。（左）对于规模为 $N = 25$、50、75的网络，正确分类率的估计值及其95%置信带。（右）对于规模为 $N = 100$、200、300的网络，正确分类率的估计值及其95%置信带。

根据Schweinberger（2011）的定义，这些值对应的模型是不稳定的；经验性证据表明，在许多应用中，$\theta_3 \in$（0，∞）（Schweinberger，2011；Stewart 等人，2019年）。图1展示了式（1）所指定的 GWESP 模型对网络的度分布和共享伙伴分布的影响，其中 θ_2 正值会随机地促进网络形成更多传递性边，即至少有一个共享伙伴的节点之间的边；相较于 $\theta_2 = 0$的伯努利随机图模型而言，这一点尤为明显。与伯努利模型相比，GWESP 模型的 ESP 分布向右偏移，印证了这一点。

我们将真实数据生成模型 $\mathcal{M}^\star$ 设为由（1）指定的曲指数族，其参数向量为 $\boldsymbol{\theta}^\star = (-2.5, \theta_2, 1)$，其中 θ_2 取遍覆盖区间 $[0, 0.5]$的网格。注意，当 $\theta_2 = 0$时，该模型简化为边概率为$p = [1 + \exp(-2.5)]^{-1}$的伯努利随机图模型。我们考虑在两个模型 $\mathcal{M}_1$ 和 $\mathcal{M}_2$之间进行选择的问题，其中 $\mathcal{M}^\star = \mathcal{M}_1$，而 $\mathcal{M}_2$ 是边概率为 $p = [1 + \exp(2.5)]^{-1}$的伯努利随机图模型。通过改变 θ_2，我们能够研究效应大小阈值（θ_2），即能够正确检测网络中传递性是否存在的阈值；网络中的传递性由（1）中的几何加权边共享伙伴模型刻画。

我们改变网络规模 $N = 25$，50，75，100，200，300，并针对每种网络规模进行 5000 次重复试验。这项模拟研究的结果汇总于图2。当 θ_2 接近 0 时，正如上文所述，我们的方法倾向于以相同的概率选择 $\mathcal{M}_1 = \mathcal{M}_2$。然而，一旦 $\mathcal{M}_1$ 和 $\mathcal{M}_2$足够大（相对于网络规模 θ_2 而言），我们的方法几乎在每次重复试验中都能正确选择 N。改变 $\mathcal{M}_1$（从 25 到 300）可以观察到网络规模的影响。当网络规模较大（ N， $N = 100$*200*，*300*）时，我们能够

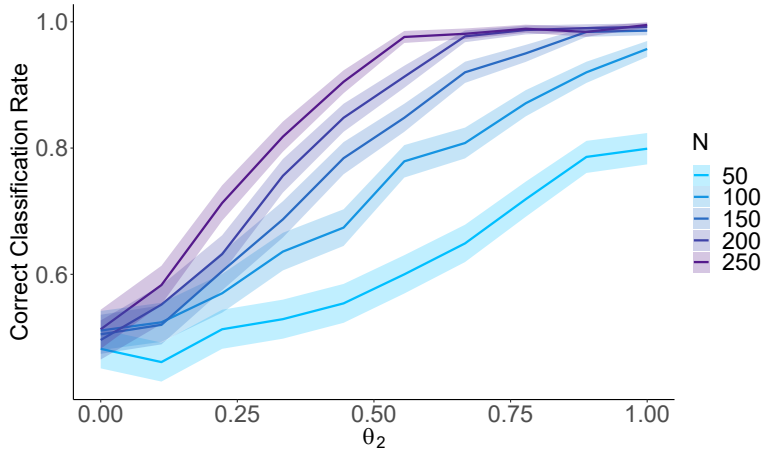

图3：模拟研究2的结果。针对不同网络规模 N，正确分类率估计值及其95%置信带。

以较高概率正确识别数据生成模型 $\mathcal{M}_1$，且 θ_2 的取值较小。相比之下，对于规模为 $N = 75$ 的网络，我们需要 $\theta_2 \geq .25$ 才能有较高把握正确选择数据生成模型；对于规模为 $N = 25$ 的网络，则需要 $\theta_2 \geq .5$。

4.2 模拟研究2：有向网络中的互惠性

当邻接矩阵 $\boldsymbol{X}$ 为无向时，相应的拉普拉斯矩阵 $\boldsymbol{L}(\boldsymbol{X})$ 将为半正定矩阵（Brouwer 和 Haemers, 2012），其特征值向量为实数$\boldsymbol{\lambda} \in \mathbb{R}^N$。然而，当 $\boldsymbol{X}$ 是有向网络的邻接矩阵时，按照无向网络的定义，图拉普拉斯矩阵可能不是半正定的，也可能具有复数特征值。文献中对有向网络的一种常见处理方式是考虑关联矩阵 $\boldsymbol{B} \in \{0, 1, -1\}^{N \times |E|}$，其中 $|E|$ 是网络中的边总数。关联矩阵的每一列恰有一个元素为 -1，表示一条边的起始节点；另有一个元素为1，表示该边的终止节点。其余元素均为零。这样，只需将所有现存的边列出为各列，标明相连的节点及其方向，即可完整描述一个有向网络。我们可以通过考虑如下定义的对称图拉普拉斯矩阵，将所提出的方法推广到有向网络：$\mathbf{L} := \mathbf{B}^t\mathbf{B}$ (Brouwer 和 Haemers, 2012）。

我们根据概率质量函数模拟有向网络

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) \quad \propto \quad \prod_{i<j}^{N} \exp\left(\theta_1 \left(x_{i,j} + x_{j,i}\right) + \frac{\theta_2}{2} x_{i,j}\, x_{j,i}\right), \tag{2}$$

我们应用该方法，并令 $\mathcal{M}_1$ 为式（ $\theta_2 = 0$ 2）中参数固定的仅密度模型。令 $\mathcal{M}^\star = \mathcal{M}_2$ 为通过式（2）指定且参数不受限制的一般模型。我们通过在 $\theta_1 = -2.5$ 和 $\mathcal{M}_1$ 中均取 $\mathcal{M}_2$，在 $\theta_2 = 0$中取 $\mathcal{M}_1$，并在 θ_2 上取 $[0, 1]$ 内的 100 个均匀网格值，对 $\mathcal{M}_2$开展模拟研究。图3 中的模拟结果基于每种情形下的 1000 次重复，进一步证实了先前模拟研究的发现：我们的方法检测真实数据生成模型的能力，取决于 θ_2与 0（即 $\mathcal{M}_1 = \mathcal{M}_2$所在的点）的距离，以及网络的规模。此外，本研究还独特地表明，我们的方法能够成功应用于有向网络。

4.3 模拟研究 3：潜在位置模型

自Hoff 等人（2002）的开创性研究以来，网络潜变量模型，尤其是潜在位置模型，日益受到关注并广泛应用。在这类模型中，为节点赋予潜在空间中的潜在位置 $\boldsymbol{z}_i \in \boldsymbol{Z}$（$i = 1, \ldots, N$），潜在空间通常取为欧几里得空间（即 $\boldsymbol{Z} = \mathbb{R}^k$），不过也有人提出了其他空间和几何结构，例如超度量空间（Schweinberger 和 Snijders，2003）、产生双线性形式的点积相似性（Hoff 等人，2002；Athreya 等人，2017）、双曲几何（Krioukov 等人，2010）以及椭圆几何（Smith 等人，2019年）。假设给定节点的潜在位置后，网络中的边条件独立。按照Hoff 等人（2002）的做法，我们在本研究中据此模拟网络：

$$\log \frac{\mathbb{P}(X_{i,j} = 1 \,|\, \boldsymbol{z}_i, \boldsymbol{z}_j)}{\mathbb{P}(X_{i,j} = 0 \,|\, \boldsymbol{z}_i, \boldsymbol{z}_j)} \quad = \quad \theta - \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2, \tag{3}$$

其中 $\theta \in \mathbb{R}$ 和 $\boldsymbol{z}_i$，$\boldsymbol{z}_j \in \mathbb{R}^k$。根据这一设定，两个节点形成边的几率会随着潜在度量空间中这两个节点的位置之间的欧几里得距离 $\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$ 增加而降低。

我们通过从上述潜在欧几里得模型生成网络，并改变潜在度量空间的维度 $k \in \{1, 2, 3, 4, 5\}$，来考察我们的方法检测潜在空间真实维度的能力。节点的潜在位置是在维度为 $k \in \{1, 2, 3, 4, 5\}$ 的空间中从均值向量为零、协方差矩阵为单位矩阵的多元正态分布中随机生成的。候选竞争模型以相同方式在维度 1、...、5 下生成。我们设定 $\theta = -2.5$，以确保边形成的基准概率较低，从而反映许多现实世界网络的稀疏性，并改变网络规模 $N \in \{50, 100, 150, 200, 250\}$。我们在每种情况下应用模型选择方法，并计算该方法选择各候选潜在空间模型的比例。

我们在图中总结了模拟研究的结果4，该图表明，我们的方法能够正确识别数据生成潜在空间的真实维度

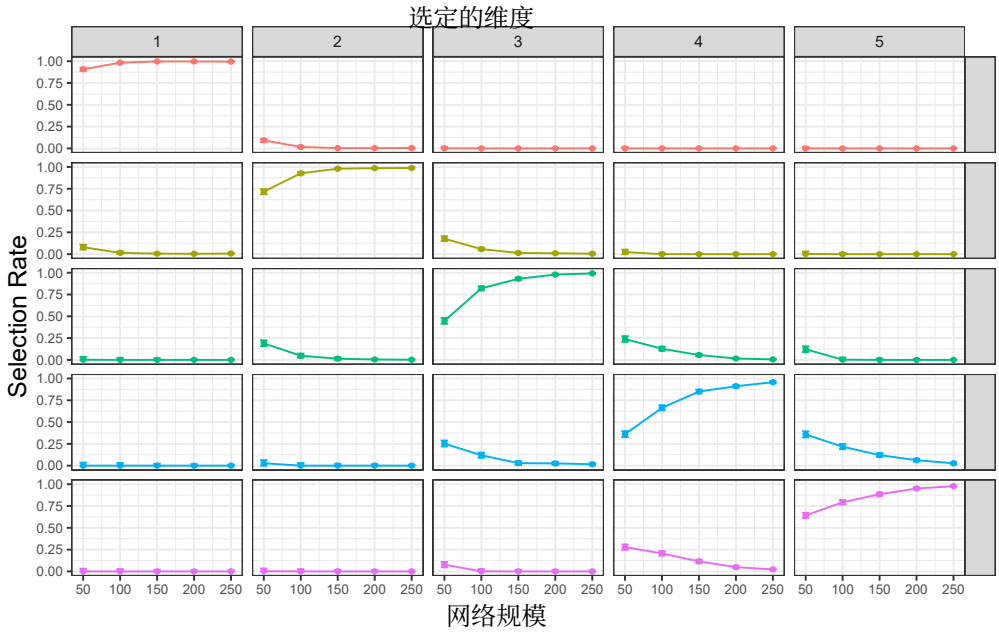

图4：模拟研究 3 的结果。对于选定的网络规模和不同的潜在空间维度，正确分类率估计值及其 95% 置信区间。对角线面板对应正确分类，此时选择率应达到最高。

前提是网络规模足够大，模型便能实现这一点。图4 中的对角线面板对应于对潜在空间维度的正确选择。值得特别注意的是，随着潜在空间维度增加，问题变得更具挑战性；但随着网络规模增大，这一影响会减弱。在本研究中，对于规模为$N = 250$的网络，正确选择率大多接近 1。

4.4 模拟研究 4：比较不同的潜在机制

接下来，我们研究所提出的方法是否能够区分潜在位置模型中边形成的不同潜在机制。第一个机制与式（3）中指定的潜在空间模型相同，而第二个机制则将欧几里得距离项 $-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$ 替换为点积 $\boldsymbol{z}_i^t \boldsymbol{z}_k$，这通常称为双线性形式。随机点积图（Athreya 等人，2017）是一类相关的潜在位置模型，它们利用潜在节点位置的双线性形式。与前一项模拟研究相同，节点的潜在位置随机取自均值向量为零的多元正态分布，但此次协方差矩阵为$\sigma^2 \boldsymbol{I}$，其中 $\boldsymbol{I}$为（适当维度的）单位矩阵，$\sigma^2 \in \{0.1, 0.2, \ldots, 1.0\}$ 为一个

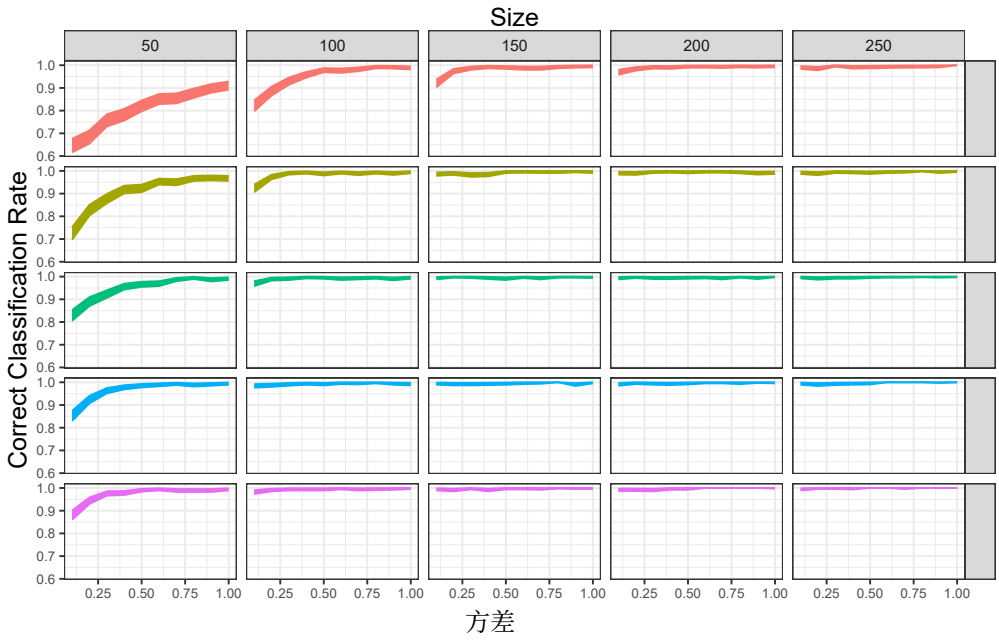

图 5：模拟研究 4 的结果，比较基于距离的模型（真实模型）与基于相似度的模型。展示了不同规模的不同网络在不同潜在空间维度下正确分类率的估计值及其 95% 置信带。

尺度因子。随着尺度因子趋近于零，两种模型都会收敛到仅密度模型，因此更难以检测真实生成过程。我们在图5中总结了模拟研究的结果。结果表明，与基于双线性形式（相似性）的模型相比，我们的方法能够正确识别真实模型（基于距离）。值得特别注意的是，与此前的研究结果一致，随着潜在空间维度和网络规模的增大，方法的表现也会有所提升。

4.5 模拟研究5：分类器选择的影响

在本研究中，我们使用三种不同的分类器重复模拟研究1：XGBoost（Chen和Guestrin，2016）、随机森林（Ho，1995；Liaw和Wiener，2002）和朴素贝叶斯（Hand和Yu，2001；Majka，2019年）。这样做使我们能够考察分类器的选择对本次模拟研究结果的影响，并探索各分类器在本次模拟研究中的相对有效性。图6 显示，在本次模拟研究中，所有分类器的表现相近，显著例外是网络规模为25时的朴素贝叶斯分类器。这表明，只要网络足够大，分类器的选择对我们所提出方法的性能影响较小。与先前模拟研究的结论一致，网络规模越大，模型特征越明显。鉴于

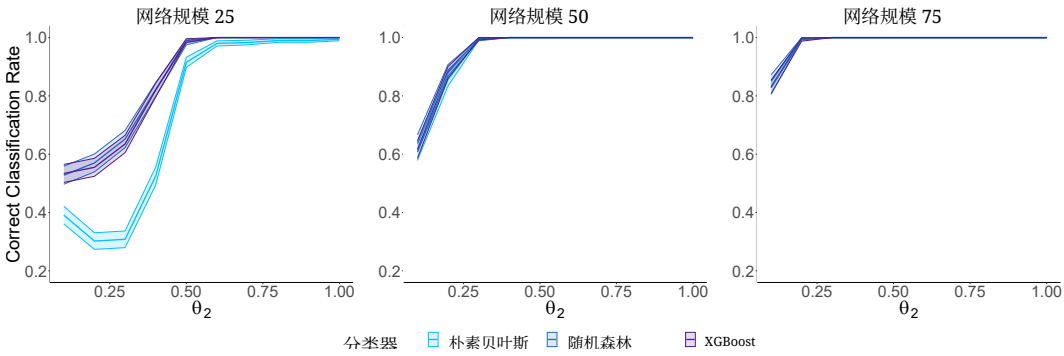

图6：模拟研究5的结果。不同分类器的正确分类率估计值及其95%置信带。

这些结果，如果模型信号足够强，分类器选择的影响似乎会减弱。

5. 应用

为了研究我们提出的模型选择方法在真实世界网络数据中的应用表现，我们研究了文献中已有研究的两个网络数据集，以此为基准，评估我们的方法是否能验证关于这些网络的既有结果与认识。

5.1 应用1：Sampson修道院网络

我们将模型选择方法应用于1968年新英格兰一座修道院中18名见习僧侣之间社会关系（亲近感）的Sampson修道院网络数据（Sampson，1968年）。根据研究该网络的现有文献，我们为该网络提出了不同的模型结构，这些结构经过精心设计，能够捕捉已知为该网络关键组成部分的社群结构。为对这种结构进行建模，研究者已将随机块模型应用于该网络（Airoldi 等人，2008年），也应用了在潜在位置上采用基于层次群组的先验分布结构的潜在位置模型（Handcock 等人，2007）。我们考虑以下模型：

- 随机块模型： $\mathcal{M}_1$–$\mathcal{M}_4$ 对应于具有 $K = 1$、2、3、4 个区块的随机块模型（$\mathcal{M}_1$ 等同于仅密度模型）。

- 潜在位置模型： $\mathcal{M}_5$–$\mathcal{M}_8$ 对应包含密度和互惠性模型项以及潜在空间维度 $K = 1$，2，3，4 的潜在位置模型。

模型	s_i	$\tilde{s}_i$	模型	s_i	$\tilde{s}_i$
$\mathcal{M}_1$ (随机块模型，$K=1$)	0.002	0.004	$\mathcal{M}_2$ (随机块模型，$K=2$)	0.003	0.007
$\mathcal{M}_3$ (随机块模型，$K=3$)	0.032	0.077	$\mathcal{M}_4$ (随机块模型，$K=4$)	**0.410**	**1**
$\mathcal{M}_5$ (潜在位置模型，$d=1$)	0.028	0.068	$\mathcal{M}_6$ (潜在位置模型，$d=2$)	0.028	0.069
$\mathcal{M}_7$ (潜在位置模型，$d=3$)	0.005	0.013	$\mathcal{M}_8$ (潜在位置模型，$d=4$)	0.003	0.008
$\mathcal{M}_9$ (广义潜在位置模型，$K=2$，$d=1$)	0.023	0.055	$\mathcal{M}_{10}$ (广义潜在位置模型，$K=3$，$d=1$)	0.044	0.108
$\mathcal{M}_{11}$ (广义潜在位置模型，$K=4, d=1$)	0.043	0.104	$\mathcal{M}_{12}$ (广义潜在位置模型，$K=2, d=2$)	0.060	0.147
$\mathcal{M}_{13}$ (广义潜在位置模型，$K=3, d=2$)	0.083	0.202	$\mathcal{M}_{14}$ (广义潜在位置模型，$K=4, d=2$)	0.036	0.089
$\mathcal{M}_{15}$ (广义潜在位置模型，$K=2, d=3$)	0.020	0.050	$\mathcal{M}_{16}$ (广义潜在位置模型，$K=3, d=3$)	0.041	0.101
$\mathcal{M}_{17}$ (广义潜在位置模型，$K=4, d=3$)	0.061	0.148	$\mathcal{M}_{18}$ (广义潜在位置模型，$K=2, d=4$)	0.012	0.029
$\mathcal{M}_{19}$ (广义潜在位置模型，$K=3, d=4$)	0.030	0.074	$\mathcal{M}_{20}$ (广义潜在位置模型，$K=4, d=4$)	0.035	0.085

表2：Sampson修道院网络中模型$\mathcal{M}_1$–$\mathcal{M}_{20}$ 的倾向得分 s_i 和归一化倾向得分 $\tilde{s}_i$ 。

- 广义潜在位置模型： $\mathcal{M}_9$–$\mathcal{M}_{20}$ 利用Handcock 等人(2007) 的层级分组先验分布结构，将前两种设定结合起来，并考虑群组数 $K=2$，3、4 以及潜在空间维度 $d=1$、2、3、4 的所有组合。

对每个模型进行拟合，并应用模型选择方法选出拟合效果最佳的模型。潜在空间模型采用Krivitsky和Handcock(2014)的方法进行拟合，随机块模型则采用Leger (2016)的方法进行拟合。表2展示了结果。倾向得分最高的模型是 $\mathcal{M}_4$，即具有 $K=4$ 个区块的随机块模型。

文献中早已确立，Sampson修道院网络具有显著的社群结构（Handcock 等人，2007；Airoldi 等人，2008年），呈现出三个标记群组。然而，统计分析揭示了可能存在第四个群组，这一点从采用混合成员随机区块模型的分析中可见（Airoldi 等人，2008年）；此外，采用潜在位置模型的研究也发现，某些节点可能与两个或更多标记群组有强连接（Handcock 等人，2007）。在本文所考察的模型框架下，选择一个含有 $K=4$ 个区块的随机块模型，似乎足以捕捉各社群的混合模式以及互惠性

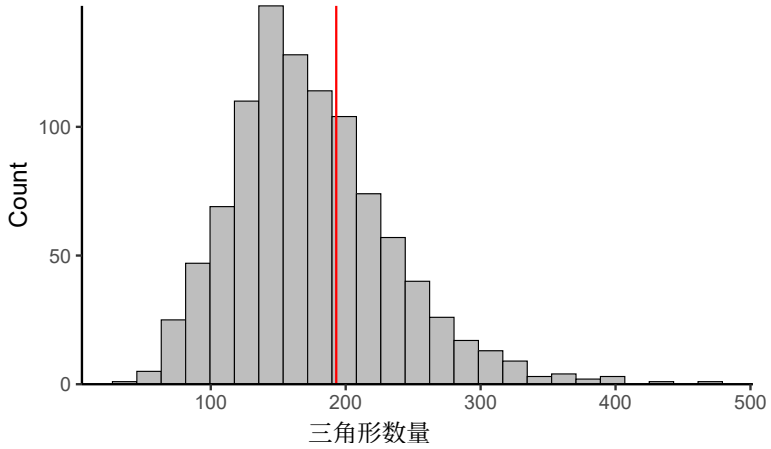

图 7：Sampson 网络中观测到的三角形数量与根据 $\mathcal{M}_4$ 模拟的网络所产生的三角形分布的拟合情况。观测到的三角形数量以红色标示。

（通过加入互惠项实现）。我们认为，该网络中的传递性并不十分显著；否则，含有 $K = 4$ 个群组的潜在位置模型可能会是更好的选择，因为潜在位置模型能够通过潜在度量空间捕捉网络传递性。图 7 通过模拟生成来自 $\mathcal{M}_4$ 的网络，并将这些模拟网络的经验三角形计数分布与该网络中观测到的三角形数量进行比较，支持了这一观点；就此而言，模型拟合良好。

5.2 应用 2：多层学校网络

本节最后，我们将研究应用于一个多层网络。该网络包含波兰 176 所小学、306 个班级中的 6,607 名三年级学生，数据采集于 2010/2011 学年（Maluchnik 和 Modzelewski，2014）。我们关注这个数据集，是因为 Stewart 等人 (2019年) 已对其进行了广泛研究，这为我们提供了最接近数据生成模型的参考。该网络包含 306 个班级，但有相当一部分学生未作答，导致网络中缺失边数据的比例很高。缺失数据问题需要仔细考虑，超出了本文的研究范围。因此，我们将本文的研究限制在该多层网络中没有任何缺失边数据的 44 个班级。所用数据集是一个有向网络，包含 906 个节点，分别对应这 44 个无缺失边数据班级中的学生；一条有向边$i \to j$ 表示个人 i 声称自己与个人 j是朋友。收集的数据中还包括每名学生的性别（记录为男性或女性）。这个多层网络数据集自然符合 Schweinberger 和 Handcock(2015) 的局部依赖框架；根据局部依赖假设，按班级抽样是合理的（命题 2 和定理 2,

模型项	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
边	✓	✓	✓	✓
互惠	✓	✓	✓	✓
出度（1–6）	✓	✓	✓	✓
女性出度	✓	✓	✓	✓
女性入度	✓	✓	✓	✓
性别匹配	✓	✓	✓	✓
几何加权共享伙伴（衰减参数固定为 0）		✓		
几何加权共享伙伴（衰减参数固定为 .25）			✓	
几何加权共享伙伴（估计衰减参数）				✓

表 3： Stewart 等人(2019年) 中模型 1–4 的描述。

Schweinberger 和 Stewart，2020）；有关该数据集的更多详情，见 Stewart 等人(2019年)。

在此应用中，我们通过研究 Stewart 等人 (2019年) 发表的模型 1–4，考察我们提出的模型选择方法是否与该网络已有的研究发现一致，相关内容总结于表 3。前三个模型项（边、互惠和出度项）控制网络中的结构效应，包括密度、互惠性以及度分布拟合。接下来的三个模型项用于调整不同的基于性别的边效应和同质性。最后三个模型项对应于式（1）中指定的几何加权共享伙伴（GWESP）项，该项曾在模拟研究1中进行研究。加入这一模型项旨在根据基参数（θ_2，见式（1））和衰减参数（θ_3，见式（1））的取值，捕捉网络传递性和三元组形成的随机倾向。模型 1 不包含 GWESP 项，而模型 2 和模型 3 将衰减参数固定为文献中给出的特定值，从而将曲指数族简化为典范指数族（参见 Hunter (2007) 和 Stewart 等人 (2019年) 中的讨论）。模型 4 对衰减参数进行估计。

我们对四个模型进行拟合 $\mathcal{M}_1$，$\mathcal{M}_2$，$\mathcal{M}_3$，$\mathcal{M}_4$ 并应用我们的模型选择方法，最终从所有其他候选模型中选出模型 $\mathcal{M}_4$ (倾向性 $= 0.997$)。这与Stewart 等人(2019年的研究结果一致。他们根据常用的启发式指标（Hunter 等人，2008年），以及样本外指标和贝叶斯信息准则（BIC），考察了各种模型对数据集的拟合情况。图8 展示了模型对重要网络特征的拟合情况。

参考文献

Airoldi, E., D. Blei, S. Fienberg 和 E. Xing（2008年）。混合成员随机块模型。
Journal of Machine Learning Research 9，1981–2014。

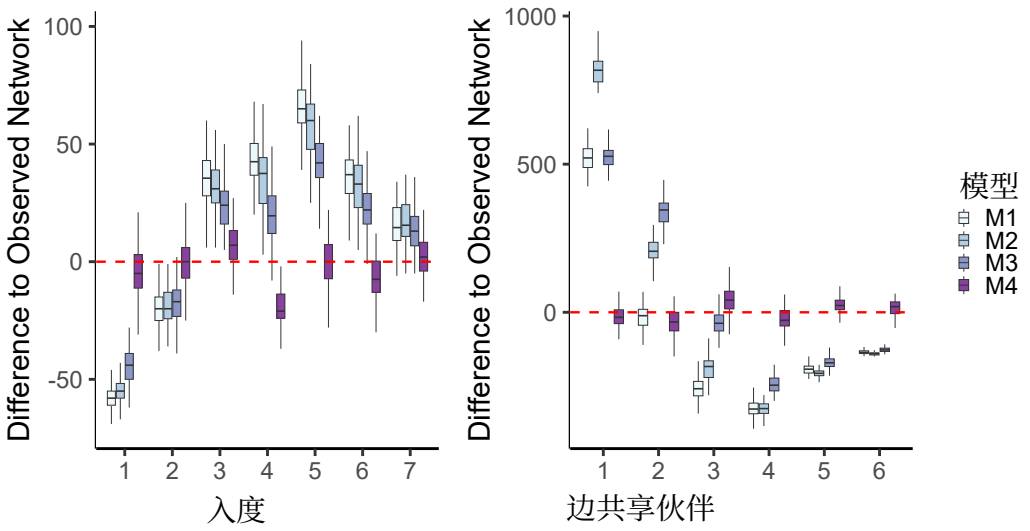

图8：拟合模型与波兰学校网络的选定统计量差异。

Anderson, C. J., S. Wasserman 和 K. Faust（1992）。构建随机块模型。 *SocialNetworks 14*, 137–161。

Athreya, A., D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin (2017). 随机点积图的统计推断：综述。 *The Journal of Machine Learning Research 18*(1), 8393–8484.

Barabasi, A.-L. and R. Albert (1999). 随机网络中标度的涌现。 *Science 286*(5439), 509–512.

Bickel, P. J. 和 E. Levina（2004）。当变量数量远多于观测数量时，关于费希尔线性判别函数、“朴素贝叶斯”及若干替代方法的一些理论。*Bernoulli 10*(6), 989–1010.

Bollob´as, B. 和 V. Nikiforov（2004）。图与厄米矩阵：特征值交错。*Discrete mathematics 289* (1-3), 119–127. Brouwer, A. E. 和 W. H. Haemers（2012）。 *Spectra of Graphs*. 学术出版社。Chen, T. 和 C. Guestrin（2016）。XGBoost：一种可扩展的树提升系统。载于 *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16，美国纽约州纽约市，第 785–794 页。ACM。

Coutinho, J. A., T. Divi´ak, D. Bright 和 J. Koskinen（2020）。有组织犯罪团伙之间合作的多层次决定因素。 *Social Networks 63*，56–69。

Cvetkovi´c, D.M., M. Doob 和 H. Sachs（1980）。 *Spectra of Graphs: Theory and Application*。纽约：学术出版社。

de Abreu, N. M. M. (2007). 图的代数连通度新旧研究成果。 *Linear Algebra and its Applications 423*(1), 53–73. 专刊收录了在阿威罗图谱研讨会上发表的论文。

Donetti, L., F. Neri, and M. A. Mu˜noz (2006, 8月). 最优网络拓扑：扩展图、笼图、Ramanujan图、纠缠网络及其他。 *Journal of Statistical Mechanics: Theoryand Experiment 2006*(08), P08007–P08007.

Fiedler, M. (1973). 图的代数连通度。 *Czechoslovak Mathematical Journal 23*(2), 298–305.

Finger, K. and T. Lux (2017). 银行间货币市场中的网络形成：行动者导向模型的应用。 *Social Networks 48*, 237–249.

Hand, D. J. and K. Yu (2001, 12月). 愚人贝叶斯——其实并没有那么愚蠢? *International Statistical Review 69*(3), 385–398.

Handcock, M. S., A. E. Raftery 和 J. M. Tantrum（2007）。社会网络的基于模型的聚类。 *Journal of the Royal Statistical Society, Series A (with discussion) 170*，301–354。

Hastie, T.、R. Tibshirani 和 J. Friedman（2011）。 *The Elements of Statistical Learning*（第2版）。纽约：施普林格出版社。

Ho, T. K.（1995）。随机决策森林。载于 *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*，ICDAR '95，M，第278–282页。电气电子工程师学会计算机学会。

Hoff, P. D.、A. E. Raftery 和 M. S. Handcock（2002）。社会网络分析的潜在空间方法。 *Journal of the American Statistical Association 97*（460），1090–1098。

Holland, P. W.、K. B. Laskey 和 S. Leinhardt（1983）。随机块模型：一些初步步骤。 *Social Networks 5*，109–137。

Hoory, S., N. Linial 和 A. Wigderson（2006）。扩展图及其应用。 *Bull.Amer. Math. Soc. 43*(04), 439–562。

Hunter, D. R.（2007）。社会网络的曲指数族模型。 *Social Networks 29*，216–230。

Hunter, D. R.、S. M. Goodreau 和 M. S. Handcock（2008）。社会网络模型的拟合优度。 *Journal of the American Statistical Association 103*，248–258。Hunter, D. R. 和 M. S. Handcock（2006）。网络曲指数族模型中的推断。 *Journal of Computational and Graphical Statistics 15*，565–583。Jackson, M. O. 和 A. Watts（2002）。社会与经济网络的演化。 *Journal ofEconomic Theory 106*（2），265–295。Kolaczyk, E. D.（2009）。 *Statistical Analysis of Network Data: Methods and Models*。纽约：施普林格出版社。Krioukov, D.、F. Papadopoulos、M. Kitsak、A. Vahdat 和 M. Bogu˜n´a（2010年9月）。复杂网络的双曲几何。 *Phys. Rev. E 82*，036106。Krivitsky, P. N. 和 M. S. Handcock（2014）。 *latentnet: Latent position and cluster models forstatistical networks*。R综合档案网。R软件包版本2.5.1。

Kwok, T. C., L. C. Lau, Y. T. Lee, S. Oveis Gharan 和 L. Trevisan（2013）。改进的切格不等式：通过高阶谱隙分析谱划分算法。载于 *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*，STOC '13，美国纽约州纽约市，第 11–20 页。计算机协会。

Latouche, P.、E. Birmel´e 和 C. Ambroise（2014）。重叠随机块模型中的模型选择。 *Electronic journal of statistics 8*（1），762–794。Leger, J.-B.（2016）。Blockmodels：用于潜在块模型和随机块模型估计的 R软件包，支持多种概率函数，并可选择是否使用协变量。Lei, J. 和 A. Rinaldo（2015）。随机块模型中谱聚类的一致性。 *TheAnnals of Statistics 43*（1），215–237。Liaw, A. 和 M. Wiener（2002）。基于随机森林的分类与回归。 *R News 2*（3），18–22。Lusher, D.、J. Koskinen 和 G. Robins（2013）。 *Exponential Random Graph Models for Social Networks*。英国剑桥：剑桥大学出版社。Majka, M.（2019）。 *naivebayes: High Performance Implementation of the Naive Bayes Algorithm inR*。R综合档案网。R软件包版本 0.9.7。

Maluchnik, M. 和 M. Modzelewski（2014）。研究过程与数据收集尝试。载于 R. Dolata（编），*Czy szkoła ma znaczenie? Zróżnicowanie wyników nauczania po pierwszym etapie edukacyjnym oraz jego pozaszkolne i szkolne uwarunkowania*，第 1 卷。华沙：教育研究院。

Morris, M.（2004）。 *Network epidemiology: A handbook for survey design and data collection*。牛津大学出版社按需出版。

Newman, M. E. J.（2006）。利用矩阵的特征向量发现网络中的社群结构。 *Physical review E 74* (3)，036104。

Obando, C. 和 F. de Vico Fallani（2017）。一种基于大规模电生理信号推断脑网络的统计模型。*Journal of The Royal Society Interface 14*(128)，20160940。

Ryan, C.、J. Wyse 和 N. Friel（2017）。社交网络潜在位置聚类模型的贝叶斯模型选择。*Network Science 5*(1)，70–91。

Sampson, S.（1968）。 *A novitiate in a period of change: An experimental and case study of relationships*。博士论文，社会学系，康奈尔大学。

Schweinberger, M.（2011）。离散指数族的不稳定性、敏感性和退化。*Journal of the American Statistical Association 106*(496)，1361–1370。

Schweinberger, M. 和 M. S. Handcock（2015）。随机图模型中的局部依赖：特征、性质与统计推断。 *Journal of the Royal Statistical Society, SeriesB 77*，647–676。

Schweinberger, M.、P. N. Krivitsky、C. T. Butts 和 J. Stewart（2020）。随机图的指数族模型：有限、超有限和无限总体情境下的推断。 *StatisticalScience 35*，627–662。

Schweinberger, M. 和 T. A. B. Snijders（2003）。社会网络中的情境：一种测量模型。载于 R. M. Stolzenberg（编）， *Sociological Methodology*，第 33 卷，第 10 章，第 307–341 页。波士顿和牛津：巴兹尔·布莱克威尔。

Schweinberger, M. 和 J. Stewart（2020）。随机图的典范指数族模型与曲指数族模型的集中性和一致性结果。 *The Annals of Statistics 48*，374–396。

Sewell, D. K. 和 Y. Chen (2015)。动态网络的潜在空间模型。 *Journal of the American Statistical Association 110*, 1646–1657。

Shi, J. 和 J. Malik (2000)。归一化割与图像分割。 *IEEE Trans. on Pattern Analysis and Machine Intelligence 22*(8)，888–905。

Shore, J. 和 B. Lubin (2015)。网络模型的谱拟合优度。 *Social Networks 43*，16–27。Smith, A. L., D. M. Asta 和 C. A. Calder (2019)。网络数据连续潜在空间模型的几何。 *Statistical Science 34* (3)，428–453。Snijders, T. A., G. G. van de Bunt 和 C. E. Steglich (2010)。网络动态随机行动者模型简介。 *Social Networks 32*(1)，44–60。

Snijders, T. A. B., P. E. Pattison, G. L. Robins, 和 M. S. Handcock（2006）。指数随机图模型的新设定。 *Sociological Methodology 36*, 99–153。

Stewart, J., M. Schweinberger, M. Bojanowski, 和 M. Morris（2019年）。多层网络数据有助于对含几何加权项的曲线指数随机图模型进行统计推断。 *SocialNetworks 59*, 98–119。

Stivala, A. 和 A. Lomi（2021）。使用指数随机图模型检验生物网络模体显著性。 *Applied Network Science 6*(1), 1–27。

Tang, M., D. L. Sussman, 和 C. E. Priebe（2013）。潜在位置图的普适一致顶点分类。 *The Annals of Statistics 41*, 1406–1430。

Von Luxburg, U.（2007）。谱聚类教程。 *Statistics and computing 17*(4), 395–416。Wang, Y. X. R. 和 P. J. Bickel（2017）。随机块模型的基于似然的模型选择。 *The Annals of Statistics 45* (2), 500–528。

Zeng, R., Q. Z. Sheng, L. Yao, T. Xu, and D. Xie（2013）。面向 Social Networks 的实用模拟方法。载于 *Proceedings of the First Australasian Web Conference-Volume 144*，第 27–34 页。