·专题一:双清论坛“大规模商务场景的统计管理论”·

网络数据建模、分析与应用研究综述*

任怡萌[1†] 陶春柏[1,2†] 朱雪宁[1,2**]

1. 复旦大学 大数据学院,上海 200433
2. 复旦大学 国家发展与智能治理综合实验室,上海 200433

[摘 要] 互联网、大数据、人工智能等信息技术催生海量数据,网络数据作为重要数据形式,具有极高的挖掘潜力与分析价值。本文首先回顾了经典网络数据建模方法及相关理论性质研究,其次综述了上述方法在金融风险、宏观经济、商业营销、社会民生方面的具体应用。在此基础上,本文针对大数据背景下复杂网络数据异质性、非线性、高维度、大规模的特征,以及众多场景中的具体应用需求,总结现有研究存在的不足,阐述了网络数据建模在理论方法与实证分析中面临的挑战。最后,基于网络数据的新特点与新需求,给出现实场景中的网络数据分析在理论建模与应用研究方面的建议。

[关键词] 复杂网络数据;网络自回归;异质性结构;非线性模型;高维数据分析;大规模网络

1 复杂网络数据建模的研究意义

党的十九大报告中明确指出,要“推动互联网、大数据、人工智能和实体经济深度融合”,数字中国和智慧社会已成为我国发展的重要目标。随着数字经济时代的到来,海量数据涌现,数据成为重要的生产要素,对各类应用场景中的管理范式产生深刻影响[1]。在多样化数据中,网络数据占有重要的地位,其商业价值受到广泛关注[2]。网络数据在不同领域中应用广泛,包括但不限于精准营销[3,4]、投资组合构建[5,6]、金融风险评估[7,8]、财税政策制定[9]等实际场景,这为建模与数据分析带来了新的机遇。

网络数据一般由网络节点、节点上的观测及节点间的连边构成。通过将结构关系抽象为简单的节点及连边,网络数据能够刻画复杂的关系模式,对其进行建模分析可揭示经济个体活动的相互作用、溢出效应和相关性,从而为实际场景中的管理决策提供参考[10]。在不同应用场景中,网络结构往往有不同的构造方式。例如,在社交媒体数据中,网络节点为平台用户,网络结构可基于节点之间的亲友、同事

朱雪宁 复旦大学大数据学院副教授,博士生导师。主要研究领域为网络数据分析、空间计量模型、高维数据建模等。入选2019年度上海市青年科技英才扬帆计划。研究成果发表于 *Journal of Econometrics*、*Journal of the American Statistical Association*、*Annals of Statistics*、《中国科学》等期刊,著有教材2本。主持国家自然科学基金优秀青年科学基金项目,参与国家自然科学基金重大项目1项。

任怡萌 复旦大学大数据学院在读博士生。主要研究方向为网络数据建模、空间计量模型、高维统计推断,研究成果发表于 *Journal of Econometrics*、*Journal of Business & Economic Statistics* 等期刊。

陶春柏 复旦大学大数据学院在读博士生。主要研究方向为分布式计算、高维数据分析、网络数据分析。

收稿日期:2023-12-25;修回日期:2024-01-28

† 共同第一作者。

* 本文根据国家自然科学基金委员会第344期“双清论坛”讨论的内容整理。

** 通信作者,Email: xueningzhu@fudan.edu.cn

等社交关系构建[11]。在金融风险研究中，网络节点为市场上的若干股票，网络结构可基于股票的共同股东或其他相关性而构建[12]。

复杂场景中的网络数据建模为传统的计量经济方法带来了新的挑战。在理论研究方面，需针对带有网络相关结构的数据提出合理的建模方法及高效、稳健的模型估计方案，建立估计方法的理论性质。在实证研究方面，需要将网络数据分析的新方法应用到具体场景中。一方面，验证新方法在复杂场景中的适用性；另一方面，提升网络数据的商业和社会应用价值，辅助管理决策。

2 复杂网络数据建模的研究现状

2.1 建模方法及理论性质研究

随着互联网技术与数字经济的发展，网络数据的体量增大、特征丰富化，衍生出大量针对复杂网络数据的建模方法及理论性质研究。目前对于网络数据的研究主要关注带网络结构的面板数据研究和网络形成与网络结构的建模方法。

2.1.1 带网络结构的面板数据建模方法研究

面板数据在计量经济、市场研究、环境科学等领域均有重要的应用。面板数据又可以分为静态与动态面板数据，其中，静态面板数据的网络结构可以刻画研究个体之间的社会相关性或交互影响，而动态面板数据的网络结构可以刻画个体之间的时序动态相关性，二者在实际问题上都具有重要的应用价值。在带网络结构的静态面板数据建模中，空间自回归模型[13]（Spatial Autoregression，SAR）是较为经典的方法，其将网络中 N 个节点之间的连接关系和强度抽象为一个 $N\times N$ 维的空间权重矩阵 W，将感兴趣的变量建模为其空间滞后项、协变量与随机误差之和，空间权重矩阵前的系数即为连接节点的平均网络效应。在此基础上，有诸多学者考虑了更加复杂的设定，丰富了其在空间计量经济学方面的应用。例如，Baltagi 等[14]在误差项中也引入空间权重矩阵，允许误差存在空间相关性。但此时，内生性问题导致传统的最小二乘估计方法不再适用，因此诸多研究讨论了空间模型的工具变量法[15, 16]，用于网络效应的估计。当工具变量难以获得时，基于最大似然法的估计方法被提出[17, 18]，Lee 等[19]建立了相关的理论性质。在互联网背景下，海量社交网络数据产生，研究静态面板数据模型的高效求解是重要课题。为此，Zhu 等[20]提出多变量空间自回归，面对传统的极大似然估计量中涉及到的高维矩阵求逆的复杂计算问题，通过建立每个节点的条件期望，构建出最小二乘类型的目标函数，极大降低了计算复杂度。基于现有研究进展，李龙飞等[10]对 SAR 模型进行了综述，给出 SAR 模型在纳什均衡模型下的经济学解释，并讨论了极大似然、伪极大似然、广义矩估计、广义经验似然等多种估计方法，为 SAR 模型的研究提供了全面而深入的探讨。

在静态面板数据建模的基础上，近年来一些研究在模型中加入回归变量的时间滞后项，形成包含时间滞后项、网络滞后项与同时段网络项的动态面板模型，用于刻画变量的动态影响，并研究了此类模型的估计方法与理论性质。例如，Yu 等[21]通过对最小二乘哑变量估计量进行纠偏，得到了带有固定效应的动态空间面板模型的伪极大似然估计量，并建立了 N 和 T（时期长度）不同的渐近关系下估计量的理论性质。Elhorst 等[22]在其基础上将模型推广至 T 较小的情况，并将伪极大似然估计量与矩估计量、极大似然估计量进行对比，证明了其在 T 较小情况下的优越表现。为了解决极大似然估计量在多个空间权重矩阵存在情况下计算效率低的问题，Lee 和 Yu[23]提出有效广义矩估计，并证明其可达到 $\sqrt{NT}$ 的收敛速度。此外，Zhu 等[24]将网络结构引入向量自回归模型，提出可应用于社交网络分析的网络向量自回归。在此基础上，考虑了节点分组异质性、稀疏性等特征的研究被提出[25-27]，增强了模型在互联网背景下的解释能力。

2.1.2 网络形成与网络结构建模的理论研究

对网络形成机制和网络结构进行分析和建模，可以辅助我们建模网络形成的内在规律。目前，常用的网络形成模型（Network Formation Model）可分为两大类：第一类是对概率进行建模的随机图模型，另一类则是将个体选择纳入考量的策略模型方法。在第一类研究方法中，早期随机图模型大部分考虑的是静态情况，即认为所有节点同时出现，依据概率规则绘制连边。较经典的模型为指数随机图模型[28]，其将节点的连边视为随机变量，通过一个指数函数，刻画网络连边的生成概率[29]。该模型在社交网络领域有丰富的应用[30, 31]。与这一类随机图模型研究思路建模较为相似的研究还包括使用统计工具对网络结构建模，其注重于对网络本身进行信息提取与刻画，以对不同网络特征结构进行学习。其中，基于潜在空间的表征方法（Latent Space Model，LSM）[32]是网络结构的经典建模手段，LSM 方法将网络中的节点映射到潜在向量空间，使用该

向量空间中的距离或相似度建模节点之间的连接信息。在其基础上，Handcock 等[33]提出潜在空间聚类模型，并同时考虑了网络数据的传递性、同质性等多种内在属性；Raftery 等[34]借用流行病学思想构造了一个近似对数似然函数，降低了计算复杂度，便于大规模网络的刻画和评估。不同于 LSM 方法，随机块模型（Stochastic Block Model，SBM）作为带有群组结构的传统图模型，聚焦于刻画网络结构中的群组特征，是社群发现的常用工具[35, 36]。其将网络节点分为若干群组，使得群组内部节点具有相似和相关属性，群组间节点具有不同属性，并可通过设置参数来调节群组内部和群组间节点的关系概率。为了解决传统 SBM 忽略顶点度的变化的问题，Karrer 和 Newman[37]引入节点度的异质性，提出度修正的 SBM 方法。后续研究在此基础上建立了度修正的 SBM 方法的理论性质，聚焦于检验群组检测的一致性，证明了模块化方法与似然类方法的一致性条件[38]。

不同于随机图模型的统计学建模思路，策略模型将网络生成时的成本和收益纳入考量，个体选择解释了网络形成的驱动因素[39]。策略模型主要通过效用函数来刻画网络中个体的总收益，若两个个体在建立连接后的效用均比建立连接前更大，则二者均会有建立连接的意向，并形成平衡状态。Jackson 和 Wolisnky[40]主要讨论了利己个体静态的策略模型，分析了不同网络的稳定性，并在网络整体层面分析不同网络的有效性（社会福利）。近年来，有较多文献对策略模型进行探索，如 Christakis 等[41]提出了网络形成的经验模型，该模型允许根据节点的特征以及网络的状态形成连接。Olaizola 和 Valenciano[42]将两种经典模型：Jackson 和 Wolinsky 的双边关系连接模型[40]及 Bala 和 Goyal 的双向流动模型[43]进行了结合，提高了模型的稳定性、有效性及稳健性。在策略模型中，个体选择会对结果产生影响，导致个体间形成博弈，在此基础上，另有一部分研究从博弈论角度进行了深入研究，考虑了多种个体策略对网络形成的影响。Dutta 和 Jackson[44]讨论了有向社交网络的有效性和稳定性，在有向网络的设定下，Bala 和 Goyal[43]将网络形成过程表述为非合作博弈，Saad 等[45]讨论了合作博弈中的联盟形成，合作的收益和成本决定了网络结构。传统的研究关注参与者之间是否产生网络连边，近年来，较多研究进一步针对个体博弈产生的连边权重进行研究[46-48]。Baumann 等[49]提出了一种加权网络形成博弈，其中每个参与者都通过有限的资源来与其他参与者形成不同强度的连接。此外，有部分研究关注个体博弈给网络带来的动态变化或结构演变[50, 51]。

2.2　实证分析研究

网络数据已成为大数据时代的一种重要数据形式，具有巨大的应用价值[52]。同时，我国大数据与人工智能技术快速兴起，实业领域内重大应用场景加速涌现，促进了对复杂场景中应用创新的大量需求。因此，从不同的复杂应用场景出发，基于网络数据进行具体应用的研究，实现场景与技术的循环促进与协同成长是至关重要的科学课题。目前针对网络数据模型的应用主要在金融风险管理、宏观经济分析、商业营销策略、社会民生分析等方面。

2.2.1　网络数据分析在金融风险管理的应用研究

在金融风险管理中，目前实证研究的主要思路是针对股票市场构建风险网络，同时结合分位数回归，考察金融风险的尾部影响及不对称影响。例如，Chen 等[53]提出了尾部事件驱动的分位数回归模型，使用时变邻接矩阵，量化系统重要性金融机构间的网络效应，捕获风险传播的动态模式。Feng 等[8]通过构建时变尾部风险网络，研究了 2008—2021 年“一带一路”股票市场的系统性风险溢出效应，反映了极端事件下贝加莱股票市场的总体风险水平和个体风险积累。欧阳资生和周学伟[54]采用分位数向量自回归模型，对金融机构的公共波动和异质波动进行建模，研究了 36 家国内金融机构股票的关联特征与风险传染效应，得到了公共波动关联的不变性及异质波动关联的非对称性结论。除股票风险外，另有一些研究考察了其他金融资产的风险。李欣珏等[55]关注我国省级城投债风险溢价网络，建立了考虑异质性溢出效应的自适应网络回归算法，为政策部门与投资者实时监控城投债风险提供重要参考。刘京军和苏楚林[56]则考察基金网络的传染性，以 2005—2014 年中国开放基金数据为样本，以基金持股关系构建网络，探讨基金网络对资金流量的溢出效应，证实基金网络间的资金传染现象及其带来的正向业绩影响。

2.2.2　网络数据分析在宏观经济分析的应用研究

除金融风险管理外，许多国内外研究也将网络数据分析应用于宏观经济分析中，为政府决策及宏观治理提供参考。其中，国内研究主要依据省份地理位置构建网络矩阵，并利用空间计量模型进行面板回归。例如，李欣先和李鲲鹏等[57]将动态双重空

间自回归模型应用于中国省际区域经济发展的研究中，考察了人力资本对国内生产总值（Gross Domestic Product，GDP）增长的贡献及空间溢出特征，拓展了原有静态分析研究的局限。骆永民和樊丽明[58]基于 1999—2009 年的省份面板数据，采用空间面板回归模型对中国农村基础设施增收效应的空间特征进行分析，得到农村基础设施建设投资对本省和邻省农民收入具有正向促进作用的结论。李婧等[59]以中国区域创新生产为背景，以省份地理位置和社会经济两种特征构建网络矩阵，考察了 1998—2007 年中国大陆 30 个省级区域创新的空间相关特性与聚集效应，为中国创新生产提供了宝贵建议。此外，一些国外研究也基于区域地理位置，分析了若干宏观因子的网络效应。Hauptmeier 等[60]研究了营业税率的选择以及地方政府的生产性公共投入，并发现邻国的税率以及公共投入的变化，都会导致本国的税收或公共投入策略发生变化，从而实现竞争力的均衡。

2.2.3 网络数据分析在商业营销策略的应用研究

近年来，数字商务与数字经济发展态势迅猛，海量电商数据催生网络数据分析在精准营销方面的应用。其中，诸多研究利用新浪微博上的海量数据进行网络分析，为用户的特征提取及个性化推荐提供重要参考。如 Zhu 等[25]将具有分组效应的网络自回归模型应用在社交平台新浪微博中，刻画用户的发帖行为模式。周静等[61]着眼于新浪微博中用户关系类型和发帖类型对于自身发帖行为的影响，并借助工具变量法解决内生性问题，从意见领袖和非意见领袖对社交平台的影响角度，洞悉了用户的发帖动机。蔡淑琴等[62]考虑了社交网络对消费者偏好的影响，基于社交网络构建用户相似度矩阵，提出基于社会网络修订的协同过滤推荐方法，更有效地实现个性化推荐。除社交平台研究外，网络数据也被应用于研究用户的点评模式及行为偏好上。Huang 等[63]提出的双模网络自回归模型将网络节点分为两种不同的类型，并将其应用在大众点评网顾客和商户的交互影响效应分析中。

2.2.4 网络数据分析在社会民生分析的应用研究

在社会民生分析中，网络数据同样起到至关重要的作用。面对 PM2.5 造成的空气污染问题，国内许多学者使用网络分析进行相关研究。邵帅等[64]基于 1998—2012 中国省域 PM2.5 数据，采用动态空间面板模型，以考虑雾霾污染的时空滞后效应，识别出影响雾霾污染的关键因素。此外，诸多国外研究也针对 PM2.5 污染、水污染问题进行了网络分析，给环境治理问题提供重要参考意见[65, 66]。在疾病传染研究中，许多研究使用网络数据分析考察了新冠疫情的变异及传播机制。Mollalo 等[67]采用包含空间滞后和空间误差项的模型研究美国县级新冠肺炎发病率的空间变异性。Sioofy 等[68]利用网络自回归模型同时考虑该地区邻国之间的疾病相互作用，基于 7 个地区/国家的新冠感染数据，预测伊朗直至 2021 年 12 月感染的总病例数。最后，也有学者在其他社会问题上使用了网络数据分析，如苏良军和孙便霞[69]聚焦空间相关性对高校学费的影响，依据两个学校是否属于同一省份构建网络矩阵，建立空间面板回归模型，证实除学校性质、当地经济水平、专业类别外，周边高校学费水平也是高校学费的重要影响因素。

3 复杂网络数据建模面临的挑战与关键科学问题

3.1 面临的挑战

由以上的研究进展梳理可知，目前对于网络数据建模的理论方法研究主要包括带网络结构的面板数据建模方法研究和网络形成与网络结构建模方法研究。网络数据实证分析应用研究中，研究者主要关注网络数据在金融风险管理、宏观经济分析、商业营销策略、社会民生分析等方面的研究。然而，在复杂场景下网络数据展现出新的特点。因此，需要设计更加灵活、高效的网络数据模型，同时适用于若干复杂场景的实际应用。

3.1.1 网络数据建模的理论方法挑战

针对复杂场景的网络数据建模，现有研究的理论方法主要面临异质性、非线性、高维度、大规模的挑战。

首先，复杂网络结构存在节点异质性和时变异质性的特征。其中，节点异质性体现在稀疏结构、分组结构以及不对称结构中。复杂网络中各个节点的度往往不同，连边数目会远小于所有可能的连边数目，形成稀疏网络，此时传统的网络结构建模方法可能失效[70]。因此，需要在已有工作的研究基础上考虑节点度的异质性，设计针对稀疏网络的建模方法。此外，网络中的节点可能会存在分组异质性。其中，同一个分组内的节点具有相似的特征，而不同分组内的节点具有不同特征，因此，需要设计考虑节点分组异质性的面板模型，讨论模型的估计方法。目前已有较多论文对网络的分组异质性进行了研

究[25-27]，但对于组数目的估计较多选用信息准则[71]的方法。如何设计数据驱动的组数估计方法，增加估计的灵活性，仍是值得研究的问题。最后，网络节点在网络中受到其他节点影响和影响其他节点的效应可能存在差异，从而产生网络结构不对称的特点，需要针对不对称的网络效应，设计对应的模型与估计方法。除节点异质性外，网络效应、分组结构、网络连边可能随时间变化，假设这些特征固定不变可能会疏于对于动态性的刻画。为刻画网络数据的实变性，Zhu 等[72]提出了函数型变系数网络模型，同时考虑了时变的节点效应系数和网络效应系数，并使用非参数最小二乘法得到求解高效的估计量。Guo 等[73]将此模型应用到了加密货币网络的分析中。但讨论时变异质性与分组异质性结合的文献相对较少，需要设计分组结构演变的面板模型以对两种异质性进行同时刻画。

其次，带有网络结构的面板模型建模需要考虑非线性的模型形式，例如，分位数回归模型、广义线性回归模型、点过程数据模型等，更大程度地丰富网络数据建模的适用场景。在分位数回归方面，Zhu 等[74]考虑了网络自回归模型的分位数回归形式，Xu 等[75]在其基础上，引入同时期网络效应，并利用工具变量法进行估计，对分位数网络自回归进行了拓展。虽然已有一些研究考虑了分位数的空间（网络）自回归模型，但其仍在建模、估计上存在多方面挑战。在建模中，分位数网络模型也可能存在节点异质性特征，需要在节点异质性的基础上考虑如何对因变量的条件分位数进行建模；在估计中，需要设计计算可行的目标函数，对条件分位数（尤其是极端条件分位数）实现高效估计。在点过程数据建模方面，Fang 等[76]在网络自回归中加入了霍克斯过程以进行点过程的建模，并同时考虑了节点异质性，对节点的潜在组结构进行了分析与估计。但点过程网络数据建模的研究依旧相对较少，需要考虑其他的离散过程对点过程进行建模，并刻画潜在的节点异质性。最后，如何将网络数据建模分析与机器学习、人工智能建模方法进行有效结合是另一大技术挑战。可考虑神经网络模型、决策树模型等高度非线性的建模方法，解决如何使用这些复杂非线性方法进行估计、纠偏与统计推断的问题。

此外，复杂网络数据建模需要考虑可能存在的数据高维度问题。一方面，当网络结构未知或部分结构不准确时，需通过对节点间的相关性进行估计，补全或纠正网络结构设定。该问题可以转化为一个高维矩阵估计的问题。已有对高维时间序列的自回归矩阵估计问题主要分为两个思路：其一，假设高维待估矩阵具有稀疏结构，并通过正则化的方法进行估计[77,78]；其二，假设存在潜在的因子结构进行降维[79-81]。然而，这些研究并没有充分利用网络结构这一重要信息。因此，需要针对网络数据中存在的相关问题设计对应的高维估计方法，这是网络数据模型的研究挑战之一。另一方面，数据采集技术的发展带来了形式多样的时间序列数据，例如，矩阵型、张量型时序数据。近年来，有研究者针对矩阵型高维时间序列模型的估计进行研究[82,83]，并应用在跨国交易数据上[84]；进一步地，张量型高维时间序列的降维方法研究也得到关注[85,86]。但是此类复杂时序的研究中未将网络结构考虑在内。因此，需要设计针对带有网络结构的复杂时序数据的建模方法，其面临的挑战包含两方面：其一，如何提出灵活高效的建模方案，利用数据结构减少待估参数的数量，处理高维问题；其二，如何建立稳健的模型估计方法，使得其能够处理数据存在缺失值、异常值的情况。

最后，复杂网络建模还面临大规模数据的计算挑战。随着电子商务与数据收集技术的发展，网络结构中的节点数量和连边数量大幅度增长，海量的数据为传统建模方法的实行带来较大困难。针对统计模型，目前已有较多工作考虑通过分布式框架来提升计算效率[87,88]，但是这些工作并没有针对网络数据模型进行考虑，无法解决大规模网络的计算问题。因此，需要针对大规模网络数据，设计计算高效的分布式估计与推断方法，降低模型计算复杂度，并建立对应的理论性质。另一方面，随着数据隐私保护的增强，大量基于联邦学习的方法研究被提出[89,90]，但这些研究对于数据的分布或结构具有较强的假设，并且未建立相关统计理论。因此，需要考虑网络数据的隐私保护，能够在较弱的假设条件下，实现网络模型去中心化的分布式估计与统计推断，并建立相关理论。

3.1.2　网络数据建模的实证分析研究挑战

网络数据模型的实证应用研究主要聚焦金融管理、宏观经济、商业营销、社会生活等诸多领域。不同领域的数据往往包含丰富的结构特征，需要使用更合适的方法对其中的潜在模式进行挖掘。例如，在金融风险研究中，具有潜在相似特征的股票往往在其持股人类型或各项指标的表现上具有同质性，而不同类型的股票则相反。如何通过构造股票的网

络结构，对股票收益率影响因素的潜在分组结构进行检测，从而更好地对其特征进行挖掘，此为一大实证挑战。在宏观经济政策研究中，不同经济发展水平的国家在税收、货币政策的制定中会存在不同的偏好。如何通过国家之间的地理邻接网络，对影响国家经济政策或宏观经济指标的关键因素分组结构进行研究，此为另一实证挑战。在用户画像研究中，大量用户在电子商务平台的直播互动中进行消费购买，如何结合此类非结构化点过程数据，进行网络关系的构造，更好地刻画用户购买行为与网络效应，是辅助精准营销的一大挑战。在社会民生研究中，分组结构与网络效应往往展现时变的特性，因此，通过构造更灵活的网络结构，研究疾病传播、污染评估等问题，具有较强的社会实践意义。针对上述现状，可以总结出网络数据建模实证分析的两大研究挑战：一方面，如何将不同领域数据的潜在复杂模式建模为统计问题，寻找最为合适的网络数据建模方法进行实证分析。另一方面，在人工智能与互联网技术快速发展的当下，针对复杂的网络数据，如何构造基于场景创新的网络数据分析方法，促使方法的落地实际应用与泛化。

3.2 关键的科学问题

针对复杂网络数据建模，需要从理论方法角度和实证应用角度解决以下两方面的科学问题，两个方面互为补充，相辅相成。

在理论方法方面，由于网络数据存在节点异质性特征，需要分别针对节点的稀疏、分组和不对称异质结构设计对应的网络数据面板模型，提出参数估计方法。对于时变异质性特征，需要考虑网络效应和分组的时变结构，并提出对应的统计建模方法。其次，考虑到带有网络结构的非线性面板数据模型，需要研究在分位数回归、广义线性模型、点过程数据模型等设定下的网络数据建模方案，并能够结合网络节点异质性提出新的理论方法。同时，需要将机器学习与深度学习方法中的非线性建模方法与网络数据建模有效结合。此外，针对网络模型中的高维待估参数提出估计方法，或利用矩阵型、张量型等数据结构减少参数量。在此基础上设计针对存在缺失值、异常值情况的估计方法，提高模型稳健性。最后，对于大规模网络数据的计算瓶颈，需要设计分布式网络模型的估计与推断方法，提高建模效率，并同时考虑隐私保护问题。

在实证分析方面，需要将复杂网络数据模型与方法应用在不同的实际场景中。例如，在金融研究中，需要构造不同股票、基金、投资组合之间的网络，研究网络效应与可能存在的异质性、动态性等，为金融风险管理提供建议。在宏观经济研究中，通过构造不同经济体之间的网络，研究空间溢出效应，辅助经济政策分析。此外，在国际贸易应用中，需要构造矩阵型时序模型，研究不同国家或地区的出口或进口网络效应。在商业营销策略实证研究中，通过构造第三方平台中用户与商户的网络结构，研究其群组行为模式，并对不同类型的网络个体进行画像。同时，对电商平台直播弹幕等点过程数据进行网络数据建模，实现对用户的精准定位，优化个性化推荐。在社会民生实证研究中，需要考虑网络结构自身时变特性，引入群组异质性结构，分析病毒传播、污染物传播的动态规律。此外，医学领域存在较多高维序列数据（如张量型脑科学数据），通过对此类张量型数据进行网络建模，可以研究人脑各区域的交互影响，实现与临床医学的良好结合。

4 网络数据建模、分析与应用的研究建议与技术路线

基于前文对于网络数据建模研究意义、研究现状及关键科学问题的分析，对网络数据建模、分析与应用研究进行如下总结：目前，网络数据建模的理论研究聚焦于带网络结构的静态面板、动态面板数据建模方法及网络结构建模方法，并已应用到金融风险控制、宏观经济政策、商业决策分析、环境污染治理等实证研究中。现存挑战主要在于大数据背景下网络数据异质性、非线性、高维度、大规模的特征，以及商业管理决策中的具体应用需求。基于前文内容，本章给出网络数据建模、分析与应用研究建议，相关技术路线如图 1 所示。

4.1 复杂网络数据的建模方法与理论研究

针对复杂的网络数据特征，本文给出以下四个方面的理论方法研究建议。

第一，针对网络数据的异质性，考虑对节点异质性和时变异质性进行建模。在节点异质性方面，首先，针对稀疏网络，聚焦稀疏网络的社群发现，建立相关统计推断理论。其次，考虑网络数据的分组结构，构建带有分组异质性的网络数据模型，提出估计模型参数和分组结构的估计方法，并建立渐近理论性质。对于分组数量的选择，考虑数据驱动的机制，增强模型灵活性。最后，针对不同网络节点的非对称性影响，可将节点划分为不同社群，刻画不同社群

之间的非对称效应，并设计相关估计方法。在时变异质性方面，可将传统常系数面板模型拓展至变系数面板模型，考虑网络稀疏性、分组结构、非对称效应随时间变化的情况，注重分析网络节点异质性的时变特征。

第二，考虑网络数据的非线性模型设置。例如，将带网络结构的面板回归与分位数回归结合，提出相应的估计方法，并建立统计理论。在此基础上，可以考虑更加复杂的网络数据设定，引入异质性结构，设计含有分组结构的条件分位数回归模型。对电子商务时代所产生的大量点过程数据，可考虑在传统泊松过程、霍克斯过程等相关模型中加入网络结构，对网络数据中的事件发生时间进行建模。在人工智能快速发展的当下，也可考虑使用机器学习、深度学习方法，刻画网络数据的非线性复杂特征，增加模型估计精度。

第三，需要针对复杂网络数据存在的高维度问题，给出高效灵活的建模方案。首先，当网络模型中网络结构设定有误或存在部分未知结构时，需要设定更为灵活的模型形式，从而对未知结构或错误结构的补全或纠正，并需要对可能产生的高维稀疏矩阵提出对应的估计方法。其次，需要针对更复杂的数据形式（如矩阵型、张量型时间序列），建立带有网络结构的动态面板模型，借助数据形式降低待估参数量。此外，可以引入分组异质性设定，提出分组参数与分组结构估计方法。同时，需要针对可能出现的缺失值、异常值情况设定更稳健、灵活的估计方案，使用深度学习方法进行处理，扩展此类模型在实际场景的应用范围。

第四，当网络数据体量很大时，需要设计网络模型的分布式估计与推断方法。考虑包含一个中心节点和多个子节点的分布式计算框架，设计传统空间面板模型的分配、局部计算和聚合方法，实现分布式估计；在推断过程中，需要在保证子节点之间无信息传输的前提下，设计计算高效的分布式推断方法。对该分布式架构，权衡整体的传输成本和估计准确性，建立相关的统计性质。此外，需要考虑分布式系统中的隐私保护问题，设计去中心化的分布式框架，借助联邦学习的思路，保证在数据隐私得到保护的同时，降低大规模网络模型的建模复杂度。

4.2　复杂网络数据的应用分析与管理决策

针对不同的应用场景，设计网络模型的实证分析，助力管理决策。以金融管理、商业营销和社交媒体领域的分析问题为例，提出如下应用分析建议。

在金融管理方面，已有研究主要对股票市场的波动率、收益率、系统性风险等指标建模，研究影响这些变量的网络效应[8, 54]，或针对其他金融资产的价格溢出效应进行建模[55, 56]。由于不同市值、账面市值比或其他指标水平的股票可能具有不同类型的收益表现，因此在金融资产相关指标分析中，需要对潜在异质性进行建模。尽管有部分研究对金融资产

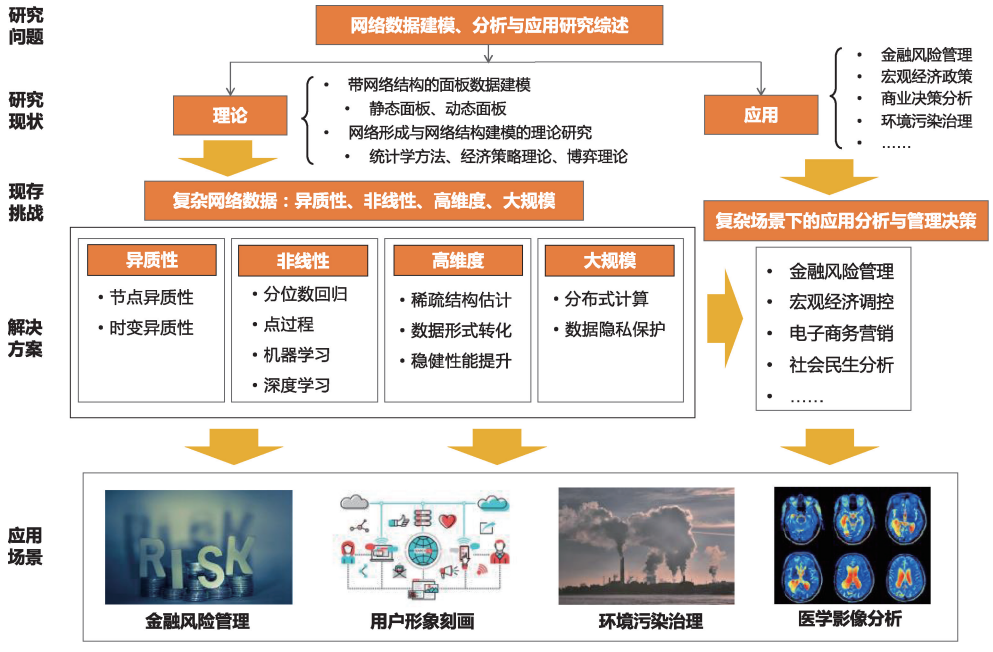

图 1　网络数据建模、分析与应用研究建议与路线图

的异质性进行了刻画[54]，但是对于网络效应的异质性研究还相对较少，或对于异质性的建模未能充分考虑网络结构特征。因此，需要针对该挑战建立更灵活的模型与估计方法，提升金融资产分析与预测的适用性。

在商业营销方面，以大众点评和 Yelp 等互联网第三方平台评分分析为例，目前已有研究通过用户与商户的交互关系、用户好友关系或用户地理位置关系构建网络，并利用网络自回归、图神经网络等模型对用户的行为偏好进行分析与预测。例如，Huang 等[63]利用的双模网络自回归模型以分析大众点评网中顾客和商户的交互影响效应，并提出最小二乘估计法以解决海量数据带来的计算负担。在此基础上，可考虑使用分布式计算框架，进一步提高大规模数据集的计算效率。此外，也可考虑分组异质性建模方法，以对不同用户群体的不同的消费习惯和品味偏好进行建模分析。最后，用户的动态浏览和评价过程可能形成高维的矩阵型时间序列数据，需要考虑针对此类数据的建模方法，以降低待估参数数量，降低计算负担。

在社交媒体领域，针对抖音、小红书等互联网平台上的网络直播或新浪微博等平台发帖行为，可利用网络建模对其中弹幕、点赞、评论等点过程数据进行建模，从而对用户进行行为画像。例如 Fang 等[76]基于组网络霍克斯网络过程，对新浪微博发帖进行分析，得到了四种人群的不同发帖模式及强度，对用户的网络异质性进行了刻画。后续研究可延续其思路，对网络直播中的点过程数据进行建模，捕捉用户对不同品类直播的偏好，以辅助优化平台的推荐系统及商户的直播选品。其次，可依据平台特征及数据特征，将组霍克斯网络过程推广至其他离散随机过程中，以更好刻画不同应用场景中的点过程数据特点；最后，由于抖音、小红书等社交媒体平台的用户数量非常庞大，会形成维度极高的关系网络，为高效计算和分析带来负担，故也需同时考虑大规模计算方法，进行分布式建模，提高网络分析的计算效率。

参 考 文 献

[1] 陈松蹊，毛晓军，王聪. 大数据情境下的数据完备化：挑战与对策. 管理世界，2022，38(1)：196—207.

[2] 洪永淼，汪寿阳. 大数据、机器学习与统计学：挑战与机遇. 计量经济学报，2021，1(1)：17—35.

[3] Carroni E，Pin P，Righi S. Bring a friend! privately or publicly? Management Science，2020，66(5)：2269—2290.

[4] Duan YR，Feng YX. Optimal pricing in social networks considering reference price effect. Journal of Retailing and Consumer Services，2021，61：102527.

[5] Caccioli F，Farmer JD，Foti N，et al. Overlapping portfolios，contagion，and financial stability. Journal of Economic Dynamics and Control，2015，51：50—63.

[6] Peralta G，Zareei A. A network approach to portfolio selection. Journal of Empirical Finance，2016，38：157—180.

[7] Zou T，Lan W，Wang H，et al. Covariance regression analysis. Journal of the American Statistical Association，2017，112(517)：266—281.

[8] Feng YS，Wang GJ，Zhu Y，et al. Systemic risk spillovers and the determinants in the stock markets of the Belt and Road countries. Emerging Markets Review，2023，55：101020.

[9] Yu JH，Zhou LA，Zhu GZ. Strategic interaction in political competition：evidence from spatial effects across Chinese cities. Regional Science and Urban Economics，2016，57：23—37.

[10] 李龙飞. 空间计量经济学中的空间自回归模型. 计量经济学报，2021，1(1)：36—65.

[11] Wasserman S，Faust K. Social network analysis：methods and applications. Cambridge：Cambridge University Press，1994.

[12] Jackson MO，Pernoud A. Systemic risk in financial networks：a survey. Annual Review of Economics，2021，13：171—202.

[13] Cliff AD，Ord JK. Spatial autocorrelation. London：Pion，1973.

[14] Baltagi BH，Song SH，Koh W. Testing panel data regression models with spatial error correlation. Journal of Econometrics，2003，117(1)：123—150.

[15] Anselin L. Spatial econometrics：methods and models. The Netherlands：Kluwer Acadmic Publishers，1988.

[16] Kelejian HH，Prucha IR. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. The Journal of Real Estate Finance and Economics，1998，17(1)：99—121.

[17] Ord K. Estimation methods for models of spatial interaction. Journal of the American Statistical Association，2012，70(349)：120—126.

[18] Smirnov O，Anselin L. Fast maximum likelihood estimation of very large spatial autoregression models：a characteristic polynomial approach. Computational Statistics & Data Analysis，2001，35(3)：301—319.

[19] Lee LF. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica, 2004, 72(6): 1899—1925.

[20] Zhu XN, Huang DY, Pan R, et al. Multivariate spatial autoregressive model for large scale social networks. Journal of Econometrics, 2019, 215(2): 591—606.

[21] Yu JH, de Jong R, Lee LF. Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. Journal of Econometrics, 2008, 146(1): 118—134.

[22] Elhorst JP. Dynamic panels with endogenous interaction effects when T is small. Regional Science and Urban Economics, 2010, 40(5): 272—282.

[23] Lee LF, Yu JH. Efficient GMM estimation of spatial dynamic panel data models with fixed effects. Journal of Econometrics, 2014, 180(2): 174—197.

[24] Zhu XN, Pan R, Li GD, et al. Network vector autoregression. The Annals of Statistics, 2017, 45(3): 1096—1123.

[25] Zhu XN, Pan R. Grouped network vector autoregression. Statistica Sinica, 2020, 30(3): 1437—1462.

[26] Ren YM, Zhu XN, Lu XL, et al. Graphical assistant grouped network autoregression model: a Bayesian nonparametric recourse. Journal of Business & Economic Statistics, 2024, 42(1): 49—63.

[27] Zhu XN, Xu GG, Fan JQ. Simultaneous estimation and group identification for network vector autoregressive model with heterogeneous nodes. Journal of Econometrics, 2023: 105564.

[28] Snijders TAB. Markov chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure, 2002, 3(2): 1—40.

[29] Robins G, Pattison P, Kalish Y, et al. An introduction to exponential random graph (p*) models for social networks. Social Networks, 2007, 29(2): 173—191.

[30] Robins G, Pattison P, Woolcock J. Missing data in networks: exponential random graph (p) models for networks with non-respondents. Social Networks, 2004, 26(3): 257—283.

[31] Robins GL, Pattison PE, Woolcock J. Social networks and small worlds. American Journal of Sociology, 2005, 110(6): 894—936.

[32] Hoff PD, Raftery AE, Handcock MS. Latent space approaches to social network analysis. Journal of the American Statistical Association, 2002, 97(460): 1090—1098.

[33] Handcock MS, Raftery AE, Tantrum JM. Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society), 2007, 170(2): 301—354.

[34] Raftery AE, Niu XY, Hoff PD, et al. Fast inference for the latent space network model using a case-control approximate likelihood. Journal of Computational and Graphical Statistics, 2012, 21(4): 901—919.

[35] Holland PW, Leinhardt S. An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association, 2012, 76(373): 33—50.

[36] Wang YJ, Wong GY. Stochastic blockmodels for directed graphs. Journal of the American Statistical Association, 1987, 82(397): 8—19.

[37] Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics, 2011, 83(1 Pt 2): 016107.

[38] Zhao Y, Levina E, Zhu J. Consistency of community detection in networks under degree-corrected stochastic block models. 2012, 40(4): 2266—2292.

[39] Jackson MO. Social and economic networks. Princeton: Princeton university press, 2008.

[40] Jackson MO, Wolinsky A. A strategic model of social and economic networks. Journal of Economic Theory, 1996, 71(1): 44—74.

[41] Christakis N, Fowler J, Imbens GW, et al. An empirical model for strategic network formation//The Econometric Analysis of Network Data. New York: Academic Press, 2020: 123—148.

[42] Olaizola N, Valenciano F. A unifying model of strategic network formation. International Journal of Game Theory, 2018, 47(4): 1033—1063.

[43] Bala V, Goyal S. A noncooperative model of network formation. Econometrica, 2000, 68(5): 1181—1229.

[44] Dutta B, Jackson MO. The stability and efficiency of directed communication networks. Review of Economic Design, 2000, 5(3): 251—272.

[45] Saad W, Han Z, Debbah M, et al. Coalitional game theory for communication networks. IEEE Signal Processing Magazine, 2009, 26(5): 77—97.

[46] Bloch F, Dutta B. Communication networks with endogenous link strength. Games and Economic Behavior, 2008, 66(1): 39—56.

[47] Cabrales A, Calvó-Armengol A, Zenou Y. Social interactions and spillovers. Games and Economic Behavior, 2011, 72(2): 339—360.

[48] Griffith A. A continuous model of strong and weak ties. Journal of Public Economic Theory, 2022, 24(6): 1519—1563.

[49] Baumann L. A model of weighted network formation. Theoretical Economics, 2021, 16(1): 1—23.

[50] Hernandez E, Menon A. Acquisitions, node collapse, and network revolution. Management Science, 2018, 64(4): 1652—1671.

[51] Joshi S, Mahmud A, Sarangi S. Network formation with multigraphs and strategic complementarities. Journal of Economic Theory, 2020, 188(6): 105033.

[52] Rathore AK, Kar AK, Ilavarasan PV. Social media analytics: literature review and directions for future research. Decision Analysis, 2017, 14(4): 229—249.

[53] Chen CYH, Härdle WK, Okhrin Y. Tail event driven networks of SIFIs. Journal of Econometrics, 2019, 208(1): 282—298.

[54] 欧阳资生, 周学伟. 金融机构时变关联的分位数特征研究——基于 QVAR 模型的实证分析. 计量经济学报, 2023(1): 213—237.

[55] 李欣珏, 夏红玉, 牛霖琳. 中国城投债风险溢价的及时性度量与预测——基于适应性网络自回归算法的分析. 计量经济学报, 2023(1): 259—285.

[56] 刘京军, 苏楚林. 传染的资金: 基于网络结构的基金资金流量及业绩影响研究. 管理世界, 2016(1): 54—65.

[57] 李欣先, 李鲲鹏, 李委明. 动态双重空间自回归模型与脉冲分析. 计量经济学报, 2021, 1(1): 66—83.

[58] 骆永民, 樊丽明. 中国农村基础设施增收效应的空间特征——基于空间相关性和空间异质性的实证研究. 管理世界, 2012(5): 71—87.

[59] 李婧, 谭清美, 白俊红. 中国区域创新生产的空间计量分析——基于静态与动态空间面板模型的实证研究. 管理世界, 2010(7): 43—55, 65.

[60] Hauptmeier S, Mittermaier F, Rincke J. Fiscal competition over taxes and public inputs. Regional science and urban economics, 2012, 42(3), 407—419.

[61] 周静, 沈俏蔚, 涂平, 等. 社交网络中用户关注类型与发帖类型对发帖行为的影响. 管理科学, 2019, 32(2): 67—76.

[62] 蔡淑琴, 袁乾, 周鹏. 基于社会网络关系的微博个性化推荐模型. 情报学报, 2014, 33(5): 520—529.

[63] Huang D, Wang F, Zhu X, et al. Two-mode network autoregressive model for large-scale networks. Journal of Econometrics, 2020, 216(1): 203—219.

[64] 邵帅, 李欣, 曹建华, 等. 中国雾霾污染治理的经济政策选择——基于空间溢出效应的视角. 经济研究, 2016, 51(9): 73—88.

[65] Zhou J, Liu J, Wang FF, et al. Autoregressive model with spatial dependence and missing data. Journal of Business & Economic Statistics, 2022, 40(1): 28—34.

[66] Ye WF, Ma ZY, Ha XZ. Spatial-temporal patterns of $PM_{2.5}$ concentrations for 338 Chinese cities. The Science of the Total Environment, 2018, 631/632: 524—533.

[67] Mollalo A, Vahedi B, Rivera KM. GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. The Science of the Total Environment, 2020, 728: 138884.

[68] Sioofy Khoojine A, Shadabfar M, Hosseini VR, et al. Network autoregressive model for the prediction of COVID-19 considering the disease interaction in neighboring countries. Entropy, 2021, 23(10): 1267.

[69] 苏良军, 孙便霞. 高校学费影响因素及空间相关性分析. 数理统计与管理, 2006, 25(4): 400—406.

[70] Krzakala F, Moore C, Mossel E, et al. Spectral redemption in clustering sparse networks. Proceedings of the National Academy of Sciences of the United States of America, 2013, 110(52): 20935—20940.

[71] Su L, Wang W, Xu X. Identifying latent group structures in spatial dynamic panels. Journal of Econometrics, 2023, 235(2): 1955—1980.

[72] Zhu X, Cai Z, Ma Y. Network functional varying coefficient model. Journal of the American Statistical Association, 2022, 117(540): 2074—2085.

[73] Guo L, Härdle WK, Tao Y. A time-varying network for cryptocurrencies. Journal of Business & Economic Statistics, 2022: 1—20.

[74] Zhu XN, Wang WN, Wang HS, et al. Network quantile autoregression. Journal of Econometrics, 2019, 212(1): 345—358.

[75] Xu X, Wang WN, Shin Y, et al. Dynamic network quantile regression model. Journal of Business & Economic Statistics, 2024, 42(2): 407—421.

[76] Fang G, Xu G, Xu H, et al. Group network Hawkes process. Journal of the American Statistical Association, 2023: 1—17.

[77] Davis RA, Zang P, Zheng T. Sparse vector autoregressive modeling. Journal of Computational and Graphical Statistics. 2016, 25(4):1077—1096.

[78] Basu S, Michailidis G. Regularized estimation in sparse high-dimensional time series models. The Annals of Statistics, 2015, 43(4): 1535—1567.

[79] Lam C, Yao QW, Bathia N. Estimation of latent factors for high-dimensional time series. Biometrika, 2011, 98(4): 901—918.

[80] Fan JQ, Liao Y, Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2013, 75(4):603—680.

[81] Wang D, Liu XL, Chen R. Factor models for matrix-valued high-dimensional time series. Journal of Econometrics, 2019, 208(1): 231—248.

[82] Chen EY, Tsay RS, Chen R. Constrained factor models for high-dimensional matrix-variate time series. Journal of the American Statistical Association, 2020, 115 (530): 775—793.

[83] Chen R, Xiao H, Yang D. Autoregressive models for matrix-valued time series. Journal of Econometrics, 2021, 222(1): 539—560.

[84] Chen EY, Chen R. Modeling dynamic transport network with matrix factor models: an application to international trade flow. Journal of Data Science, 2022, 21 (3): 490—507.

[85] Chen R, Yang D, Zhang CH. Factor models for high-dimensional tensor time series. Journal of the American Statistical Association, 2022, 117(537): 94—116.

[86] Wang D, Zheng Y, Lian H, Li G. High-dimensional vector autoregressive time series modeling via tensor decomposition. Journal of the American Statistical Association, 2022, 117(539): 1338—1356.

[87] Fan JQ, Wang D, Wang KZ, et al. Distributed estimation of principal eigenspaces. Annals of Statistics, 2019, 47(6): 3009—3031.

[88] Jordan MI, Lee JD, Yang Y. Communication-efficient distributed statistical inference. Journal of the American Statistical Association, 2019, 114(526): 668—681.

[89] Tang Hl, Lian Xr, Yan M, et al. Decentralized training over decentralized data// International Conference on Machine Learning. Proceedings of Machine Learning Research, 2018: 4848—4856.

[90] Richards D, Rebeschini P, Rosasco L. Decentralised learning with random features and distributed gradient descent// International Conference on Machine Learning. Proceedings of Machine Learning Research, 2020, 119: 8105—8115.

Network Data Modeling, Analysis and Application Review

Yimeng Ren[1†]　Chunbai Tao[1, 2†]　Xuening Zhu[1, 2*]

1. *School of Data Science, Fudan University, Shanghai* 200433
2. *MOE Laboratory for National Development and Intelligent Governance, Fudan University, Shanghai* 200433

Abstract　The advent of information technologies such as the Internet, big data, and artificial intelligence has generated massive data. Network data, as a crucial form of data, has high mining potential and analysis value. This paper begins by reviewing classical methods for network data modeling and related theoretical properties. Subsequently, it surveys the specific applications of these methods in financial risk, macroeconomics, business marketing, and societal well-being. Furthermore, considering the heterogeneity, nonlinearity, high dimensionality, and large-scale features of network data, this paper identifies the shortcomings of the current research and outlines challenges faced in theoretical methods and empirical analysis of network data modeling under the background of massive data. Finally, based on the novel characteristics and demands of network data, the paper provides recommendations for theoretical modeling and applied research in analyzing network data in real scenarios.

Keywords　complex network data; network autoregression; heterogeneous structures; nonlinear models; high-dimensional data analysis; large-scale network

（责任编辑　张　强）

† Contributed equally as co-first authors.

* Corresponding Author, Email: xueningzhu@fudan.edu.cn