rMultiNet：用于多层网络分析的 R 包

TingLi tingeric.li@polyu.edu.hk
应用数学系，香港理工大学**ZhongyuanLyu**
zlyuab@connect.ust.hk数学系，香港科技大学**ChenyuRen**
chenyu.ren@connect.polyu.hk 应用数学系，香港理工大学 **DongXia**
madxia@ust.hk 数学系，香港科技大学

arXiv:2302.04437v1 [stat.ML] 9 Feb 2023

编辑：

摘要

本文开发了一个 R 包**rMultiNet** ，用于分析多层网络数据。我们提供了近期文献中的两个通用框架，即混合多层随机区块模型（MMSBM）和混合多层潜在空间模型（MMLSM），用于生成多层网络。我们还提供了多种方法，用于揭示节点和层的嵌入，并进一步开展数据分析，例如聚类。该软件包包含三个真实数据示例。**rMultiNet** 的源代码可在 https://github.com/ChenyuzZZ73/rMultiNet 获取。

关键词：多层网络、张量分解、社区检测

1. 引言

近十年来，处理和分析复杂网络的需求迅速增长。尽管已有大量关于单层静态网络的研究（Amini 等（2013）；Gao 等（2017，2018）；Wang 等（2021）；Jing 等（2022）；Yu 等（2022）），研究人员对多层网络的研究兴趣日益浓厚（Paul 和 Chen（2020）；Le 等（2018）；Lei 等（2020）；Arroyo 等（2021）；Jing 等（2021）；Li 等（2021）；Chen 等（2022）），因为多层网络能够更有力地表示多关系数据。许多类型的现实世界数据都可以记录为多层网络，例如脑连接网络、基因间相互作用网络和世界贸易网络。

由于社区结构是静态网络分析中常见的现象，自然会引发这样一个问题：如何定义并检测多层网络中的社区结构。在多层网络中，节点代表我们感兴趣的个体，而不同层中节点之间的边代表不同的关系。多层网络中的这类复杂关系给识别和分析其社区结构带来了巨大挑战。具体而言，层间的异质性可通过个体链接、节点的群体归属，或社区内部及不同社区之间的连接模式来刻画。最近，Jing 等人（2021）提出了一种新颖的混合多层

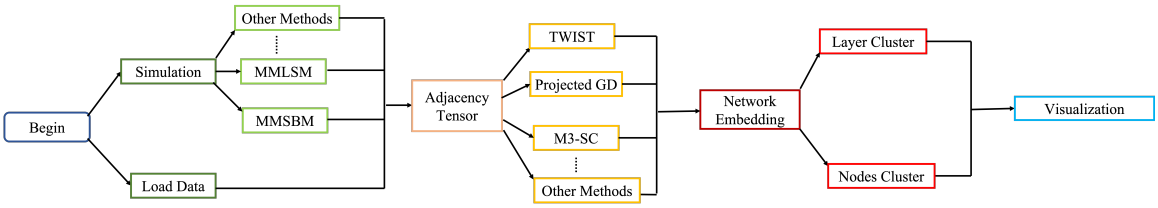

图 1：**rMultiNet** 包概览

随机块模型（MMSBM），并采用一种基于张量的新方法 TWIST，同时对网络进行聚类，并识别顶点的全局和局部群体归属。此外，Lyu等（2021）提出了一种新颖的混合多层潜空间模型（MMLSM），通过广义低秩张量分解估计节点的潜在位置。

本文提出了一个用于分析混合多层网络的 R 包 **rMultiNet** 。图 1 展示了我们的包 **rMultiNet** 的概览。我们提供了几种常用方法，供用户研究多层网络的潜在特征。首先，我们提供了三个真实数据集作为研究示例，并提供了两个生成模型 MMSBM 和 MMLSM，用于生成混合多层网络。接着，我们提供了 Jing 等人（2021）和 Lyu等（2021）提出的若干拟合方法，以揭示多层网络的嵌入结果；此外，还提供了几种常用的聚类方法来分析嵌入结果。最后，我们提供了若干数据可视化函数来展示分析结果。

2. 包概览

rMultiNet 包含用于生成多层网络的模型、用于学习复杂混合多层网络潜在结构的多种算法、用于进一步分析嵌入结果的多种聚类方法，以及若干用于展示的可视化函数。该包由以下模块组成：

- **生成**：**rMultiNet** 采用混合多层随机块模型（MMSBM）（Jing 等人（2021））和混合多层潜在空间模型（MMLSM）（Lyu等（2021））生成混合多层网络。

- **嵌入**：**rMultiNet** 包含两种用于混合多层网络嵌入的张量分解算法，例如 Jing 等人（2021）提出的 TWIST，以及 Lyu等（2021）介绍的 ProjectedGD。此外，Jing 等人（2021）提出的一种朴素方法——对所有层的邻接矩阵求和后进行谱聚类（Sum–Adj）以及模式3展平（M3-SC）——也被纳入作为比较基线。

- **聚类**：**rMultiNet** 提供多种聚类方法来分析嵌入结果，例如 K均值（Likas 等（2003））、谱聚类（Dong 等（2012））以及基于密度的含噪声应用空间聚类（DBSCAN）（Hahsler 等（2019））。

- **数据集**：**rMultiNet** 包含三个用于研究的数据集。

人类疟原虫基因网络：本研究所用数据为疟原虫基因序列的 9 个高变区。每个网络包含 212 个节点，这些节点都出现在全部 9 层中。有关背景和数据预处理的更多详情，请参阅 Larremore 等（2013）和 Jing 等人（2021）。

全球食品贸易网络：在这个多层网络中，层代表不同的产品，节点代表国家，每一层中的边代表各国之间特定食品产品的贸易关系。数据由 De Domenico 等（2015）收集，可在 http://www.fao.org 获取。经过数据预处理（Jing 等人（2021））后，我们得到一个包含 30 层、99 个节点的网络。

联合国商品贸易网络：该数据集包含联合国商品贸易数据库（https://comtrade.un.org）中各国 2019 年的年度贸易信息。我们选取按商品和服务出口额（美元）排名的前 48 个代表性国家。每一层代表按 97 个类别分类的一种商品类型（Lyu等（2021））。

3. 功能与示例

本节将详细介绍**rMultiNet** 软件包的用法。本部分所用符号的具体含义可参见 Jing 等人（2021）和 Lyu等（2021）。待分析的多层网络可以通过该软件包生成，也可以从外部文件加载。从外部文件加载的数据需要采用软件包**rTensor**中定义的“张量”类的形式。下面介绍**rMultiNet** 中生成混合多层网络邻接张量的两种方法，分别对应 MMSBM 和 MMLSM。

> library(rMultiNet)
>生成MMSBM(n, m, L, K, d = NULL, R = NULL)
>生成MMLSM(n, m, L, rank, U mean= 0.5, cmax =1, d, int type = ‘均匀分布’, ker- nel fun =‘logit’, scale par=1)
其中，n 表示顶点数，m 表示网络类型数，L表示层数，K 表示顶点组数，d 表示网络每层的平均度，r 表示每层的出入比。特别是在函数 $GenerateMMLSM$ 中，$rank$ 表示潜在位置矩阵 U 的秩，$U\,mean$ 表示 U 每个元素的正态分布均值，$cmax$ 表示核心张量 C 的逐元素上界，$int\ type$ 表示生成张量 C 的方式（‘均匀分布’或‘正态分布’），$kernel\ fun$ 表示生成邻接张量所用的连接函数（$‘logit’$ 或$‘probit’$），$scale\ par$表示参数张量的缩放因子。输出是一个列表，其中包含邻接张量和生成参数 Θ。

从上述内容加载或生成的多层网络将存储为邻接张量。**rMultiNet** 提供了用于学习邻接张量潜在结构的算法。

>InitializationMMSBM(tnsr, ranks=NULL)
>PowerIteration(tnsr, ranks=c(2,2,2), type=”TWIST”, U 0 list, delta1=1000, delta2=1000, max iter = 25, tol = 1e-05)
函数 $InitializationMMSBM$ 的输出是初始化 $U\ 0\ list$，可作为函数

*PowerIteration*的输入。这里，*tnsr* 是网络的邻接张量，*type* 指定要运行的迭代算法‘*TWIST*’或‘*Tucker*’；*delta1* 和*delta2* 是模态-1和模态-2正则化的调优参数，*max iter* 是最大迭代次数，*tol*是收敛容差。请注意，*ranks* 是根据公式 $m \times K -(m-1)$ 计算的核心张量秩（见 Jing 等人（2021））。输出是一个列表，其中包含核心张量 Z、网络嵌入和节点嵌入。

>*SpecClustering(tnsr, rank,* 嵌入类型 *= "Layer")*

在函数 *SpecClustering* 中，*tnsr* 是邻接张量，*rank* 是输出矩阵 U的列数，*embedding type* 指定嵌入类型（‘*Node*’对应邻接矩阵求和，‘*Layer*’对应 M3-SC）。输出矩阵 U 可用于 K 均值等聚类方法。>*InitializationLSM(gen list, n, rank, M, perturb = 0.1, int type)*

> *ProjectedGD(Ini* 列表*, Cmax,* 外层 *eta = 1e-04,* 外层 *tmax = 35, p* 类型 *='logit', rd='Non',* 显示 *= TRUE, sgma =1,* 样本量 *=5000)*

在函数 *InitializationLSM* 中，*gen list* 是一个列表，包含邻接张量以及混合多层网络的参数 Θ；n 是节点数，*rank* 是U的秩；M 是网络类型的数量，*perturb* 指定均匀分布的上界，*int type* 指定初始化 U 和 W 的方法（‘*spec*’、‘*rand*’或‘*warm*’）。函数 *InitializationLSM* 的输出是一个列表，包含邻接张量、U_0、W_0 和调优参数 $\{\delta_1, \delta_2, \delta_3\}$。在函数 *ProjectedGD* 中，*Ini list* 是函数*InitializationLSM*的输出，*Cmax* 是添加系数约束的上限，*eta outer*是梯度下降的学习率，*tmax outer* 是梯度下降的迭代次数，*p type* 指定连接函数类型（‘*logit*’、‘*probit*’或‘*poisson*’），*rd* 指定是否使用随机抽样（‘*rand*’或‘*Non*’），*sgma* 是连接函数参数 σ。输出为节点和层的嵌入结果。

rMultiNet 还实现了用于可视化的函数。

>网络嵌入*(network membership,L, paxis = 2)*

此函数分别用于绘制网络嵌入和节点嵌入。这里的*paxis* 指定绘图所使用的特征向量数量。如果特征向量数量超过两个，则会生成绘图表。默认情况下，绘制第二个和第三个特征向量。

>社区聚类 *km(embedding,type,cluster number)*

>社区聚类 *dbscan(embedding,type,eps value =.05,pts value=5)*

K均值和 DBSCAN 等聚类算法可应用于嵌入结果。这里，*type* 可以是节点嵌入‘*n*’或网络嵌入‘*N*’，*cluster number*是 K 均值的聚类数，*eps value* 和*pts value* 是 DBSCAN 的参数。

4. 总结

rMultiNet 推出了一个扩展型 R 包，其中包含多种传统及先进的张量分解方法，用于混合多层网络分析。该包采用模块化流水线模式开发，涵盖生成式建模、嵌入算法和可视化。面对不断增长的数据以及不同网络中的交互，**rMultiNet** 旨在帮助研究复杂网络，尤其是混合多层网络。该包也适用于动态网络；动态网络可视为多层网络的一种特殊情形，其中各层由时间索引。

致谢

参考文献

Arash A Amini、Aiyou Chen、Peter J Bickel 和 Elizaveta Levina。《大型稀疏网络社区检测的伪似然方法》。统计年刊，41 (4):2097–2122，2013。Jes´us Arroyo、Avanti Athreya、Joshua Cape、Guodong Chen、Carey E Priebe 和 Joshua T Vogelstein。《具有共同不变子空间的多个异质网络推断》。机器学习研究杂志，22(142):1–49，2021。Shuxiao Chen、Sifan Liu 和 Zongming Ma。《非齐次多层网络中的全局及个体化社区检测》。统计年刊，50(5):2664–2693，2022。Manlio De Domenico、Vincenzo Nicosia、Alexandre Arenas 和 Vito Latora。《多层网络的结构可约性》。自然通讯，6(1):1–9，2015。Xiaowen Dong、Pascal Frossard、Pierre Vandergheynst 和 Nikolai Nefedov。《多层图聚类：谱视角》。*IEEE*信号处理汇刊，60(11):5820–5831，2012。Chao Gao、Zongming Ma、Anderson Y Zhang 和 Harrison H Zhou。《随机块模型中实现最优误分类比例》。机器学习研究杂志，18(1):1980–2024，2017。Chao Gao、Zongming Ma、Anderson Y Zhang 和 Harrison H Zhou。《度修正块模型中的社区检测》。统计年刊，46(5):2153–2185，2018。Michael Hahsler、Matthew Piekenbrock 和 Derek Doran。《dbscan：使用 R 进行快速基于密度的聚类》。统计软件杂志，91:1–30，2019。Bing-Yi Jing、Ting Li、Zhongyuan Lyu 和 Dong Xia。《通过正则化张量分解进行混合多层网络社区检测》。统计年刊，49 (6):3181–3205，2021。Bingyi Jing、Ting Li、Ningchen Ying 和 Xianshi Yu。《使用对称化拉普拉斯逆矩阵（SLIM）进行稀疏网络社区检测》。华人统计学志，32(1): 1，2022。Daniel B Larremore、Aaron Clauset 和 Caroline O Buckee。《分析高度重组疟原虫基因的网络方法》。*PLoS*计算生物学，9(10): e1003268，2013。Can M Le、Keith Levin 和 Elizaveta Levina。《根据多个含噪实现估计网络》。电子统计学杂志，12(2):4697–4740，2018。

Jing Lei、Kehui Chen 和 Brian Lynch。《多层网络数据中的一致性社区检测》。*Biometrika*，107(1):61–73，2020。Ting Li、Jianchang Hu、Shiying Wang 和 Heping Zhang。《脑连接的超级变异体识别》。*Human Brain Mapping*，42(5):1304–1312，2021。Aristidis Likas、Nikos Vlassis 和 Jakob J Verbeek。《全局 K 均值聚类算法》。模式识别，36(2):451–461，2003。Zhongyuan Lyu、Dong Xia 和 Yuan Zhang。《高阶网络的潜在空间模型与广义张量分解》。*arXiv* 预印本 *arXiv:2106.16042*，2021。Subhadeep Paul 和 Yuguo Chen。《多层网络中一致性社区检测的谱方法与矩阵分解方法》。统计年刊，48(1):230–250，2020。

Jiangzhou Wang、Jingfei Zhang、Binghui Liu、Ji Zhu 和 Jianhua Guo。采用轮廓伪似然方法进行快速网络社区检测。美国统计学会杂志，第 1–14 页，2021 年。

Xianshi Yu、Ting Li、Ningchen Ying 和 Bing-Yi Jing。结合社交网络感知的协同过滤。《商业与经济统计期刊》，40(4):1629–1641，2022 年。