

参照零模型的双层网络结构相关性检测

崔丽艳¹, 许小可^{1,2}

1. 大连民族大学信息与通信工程学院, 大连 116600

2. 贵州大学, 贵州省公共大数据重点实验室, 贵阳 550025

摘要 多层网络是近年来提出的新型复杂网络模型,在级联故障分析、信息传播、链路预测和网络同步等诸多领域均有广泛应用。多层网络的任意两层网络结构间往往存在关联性和耦合性,如何检测两层网络之间是否具有结构相关性并对其进行定量刻画是一个非常重要且亟待解决的问题。本文首先分3个层次总结并提出度量双层网络结构相关性的方法与统计量,其中第1层次是检测双层网络整体上的连接相关性,第2层次检测两层网络所有节点之间整体上的度度相关特性,第3层次是检测双层网络富节点之间的连接相关性。由于这3种相关性的计算和度量都依赖于网络统计量进行,而统计量的绝对数值往往没有意义,因此提出了以多种双层网络的零模型作为参照物,通过假设检验方法来量化双层网络之间的结构相关性,分析了这种结构相关性存在的内在机理。最后,使用一个实证双层网络——全球语言多层网络验证了本文研究范式的有效性。本文研究可检测出实证多层网络中任意两层之间复杂的耦合作用。

关键词 多层网络;零模型;随机断边重连;结构相关性

复杂网络是指由一定数量的节点在一定时间尺度上相互关联共同构成的复杂系统,网络科学的研究重点关注复杂网络中存在的非平凡性质及这些特殊性质的产生机理。目前,网络科学的相关理论已经运用在生物学、神经科学、社会学、物理、信息科学等诸多领域,但是大多数复杂网络研究还主要针对单层网络,仅关注单层网络中存在的非平凡性质及其产生机理。然而在很多情况下,单层网络仅适用于研究复杂系统中的一个子集,因为它没有考虑不同层网络之间可能存在的相关性和因果性,这使得该模型对很多复杂系统的抽象过于简化。考虑到单层复杂网络的表达能力限制,人们提出了多种层内连接、层间相关的多层复杂网络模型,目前已成为很多复杂系统的标准研究模型和新的研究范式^[1-2]。

在多层网络中,不同层的网络既可以共用相同的节点集合,也可以使用不同的节点集合。最简单的多层网络就是双层网络,双层网络适用于很多实际问题的研究。如在社交网络中,一种数据记录了用户在不同时间点上和其他用户建立好友关系或解除好友关系,这种数据描述了社交网络上用户之间显性关系(谁和谁是朋友)的演化,构成了社交结构网络。另一种数据记录了用户在不同时间点上与其他用户共

同参与某项调查、发表微博、收发短消息、语音聊天、分享文件等行为,这种数据描述了社交网络上用户各种参与行为的演化,完成了用户的社交功能,构成了社交功能网络。很明显,社交结构网络和功能网络之间存在着一定的相关(耦合)关系,而且这种相关关系的计算和表达十分重要,这两层网络之间是否存在结构上的相关以及如何定量刻画相关强度是一个非常值得研究的科学问题。

在实证网络的研究中,研究者一般会使用多种统计量来刻画复杂网络的性质。具体的方式如图1(a)所示,首先得到一个实证网络,计算该实证网络的某个统计量,然后通过该统计特征的绝对数值对网络性质进行描述。由于网络统计量没有量纲而且网络规模千差万别,因此这种方式的结果往往不准确,不同网络之间结果也不具有可比性。而参照零模型的实证网络研究方法如图1(b)所示,不但要计算实证网络的统计量,也要计算它对应的多个零模型网络的统计量,然后将经典的假设检验过程引入到网络科学领域来说明结果的鲁棒性和可信性,通过统计分析来确定实证网络性质是否和零模型网络之间存在显著性差异,最终确定其性质是由于网络结构的哪种结构非平凡性导致的。在本文双层网络结

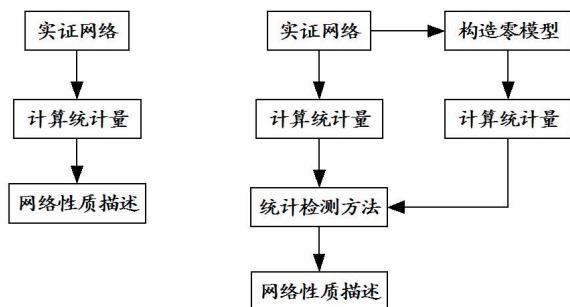
收稿日期:2017-03-13;修回日期:2017-06-30

基金项目:国家自然科学基金项目(61374170,61603073);大连市青年科技之星培育计划项目(2015R091);中央高校基本科研业务费专项资金项目(DC201502060201)

作者简介:崔丽艳,硕士研究生,研究方向为复杂网络,电子信箱:m18840910012@163.com;许小可(通信作者),教授,研究方向为复杂网络、社交网络大数据、计算传播学,电子信箱:xuxiaoke@foxmail.com

引用格式:崔丽艳,许小可.参照零模型的双层网络结构相关性检测[J].科技导报,2017,35(14):63-74;doi:10.3981/j.issn.1000-7857.2017.14.008

构相关性检测的研究中,使用了零模型作为参照的新研究框架,本研究主要涉及双层网络相关统计量、双层网络零模型和假设检验统计方法3部分。



(a) 传统实证网络研究过程 (b) 参照零模型的实证网络研究过程

图1 传统实证网络研究与参照零模型实证网络研究过程对比

Fig. 1 Comparison between the traditional study and the study based on null model for correlation detection

考虑到双层网络中层与层之间结构相关(耦合)特性的重要性,从3个层面(由强到弱)归纳和提出了多种双层网络结构耦合特性的量化指标。第一层次是检测双层网络中层与层之间所有连边整体上的连接相关性,第二层次是检测层与层之间所有对应节点之间的度度相关特性,第三层次是检测双层网络中两层中富节点之间的连接相关性。这3种相关性的计算和度量都依赖于网络统计量来进行,鉴于这些统计量的绝对数值往往无量纲,在研究中不仅关注实际网络统计量的绝对值,更关注实证网络和网络随机化零模型比较后的相对值^[3-6],通过假设检验的统计结果来说明实证网络的性质,系统性地提出了如何基于随机连边置乱和随机节点置乱构造出各种网络零模型作为实证双层网络的参照物。

通常,把与一个实际网络具有某些相同性质的随机网络称为该实际网络的随机化副本,这类随机化网络在统计学上被称为零模型^[7-9]。一个好的复杂网络零模型能为原始网络提供一个准确的参照,结合统计量指标就可以准确描述出实际复杂网络的非平凡特性。在本文中,以双层网络为研究对象,主要参照零模型来分析双层网络之间的相关性。双层网络的零模型构造方式主要有以下2种。一种是将双层网络中的一层进行断边重连,这种断边重连既可以进行随机断边重连,也可以使用有倾向性的断边重连。随机断边重连可以产生0、1、2、2.5阶网络零模型,用来检测实证网络 and 不同层次随机零模型网络之间的差别。有倾向性断边重连有2方面的应用:一是可用来检测网络是否具有富人俱乐部特性,二是可以用来检测网络的匹配特性。另一种构造方式是将双层网络中的一层进行随机节点置乱,即将该网络中的一些或全部节点的标号进行随机置乱,这种方式也可以理解为双层网络中每一层的拓扑结构均没有变化,但是层与层节点之间的

对应关系进行了随机置乱。

为检测双层网络的结构相关性,在计算出实证网络的某个统计量和对应的多个零模型网络统计量后,就需要采用假设检验来判断实证网络和零模型之间是否有显著性差异。其基本原理是将所有零模型视为总体,而将实证网络视为总体中的一个抽样,通过一些假设检验方法对它们的统计量进行统计推理,最后得出是否应该拒绝该假设的结论。

1 检测双层网络结构相关性的统计量

1.1 双层网络的形式化定义

尽管多层网络中不限定网络的层数,但是绝大多数研究中都以双层网络为研究对象,同时由于多层网络(多于2层)之间的相关关系也需要两两之间进行研究,因此本文中仅研究双层网络上层与层之间结构的相关性。本文中下层网络称为 G_α 网络,下层网络的节点连接关系用 A_{ij}^α 的形式表示;同时将上层网络称为 G_β 网络,上层网络的节点连接关系用 A_{ij}^β 的形式表示。如果每层网络中有 m 个节点,那么网络 G_α 的邻接矩阵 $A = (A_{ij}^\alpha)_{m \times m}$ 是一个 m 阶方阵,第 i 行第 j 列上的元素 A_{ij}^α 定义为

$$A_{ij}^\alpha = \begin{cases} 1, & \text{如果节点} i \text{和节点} j \text{之间有边} \\ 0, & \text{如果节点} i \text{和节点} j \text{之间没边} \end{cases} \quad (1)$$

一个双层无权无向网络,节点 i 在 G_α 网络的度 k_i^α 表示为

$$k_i^\alpha = \sum_j A_{ij}^\alpha \quad (2)$$

如果在网络 G_α 和 G_β 中各层共用一个节点 i ,则定义为:

$$A_i^{\alpha\beta} = \begin{cases} 1, & \text{如果各层共用一个节点} \\ 0, & \text{各层不共用一个节点} \end{cases} \quad (3)$$

1.2 检测双层网络链接整体相关性的统计量

在本节中,主要检测双层网络中层与层之间链接的整体相关性,能够检测双层网络中链接是否存在整体相关性的统计量有以下4种,分别为:长向量相关系数、长向量的互信息、整体重叠程度、局部重叠程度等。

1.2.1 网络对应长向量的Pearson相关系数

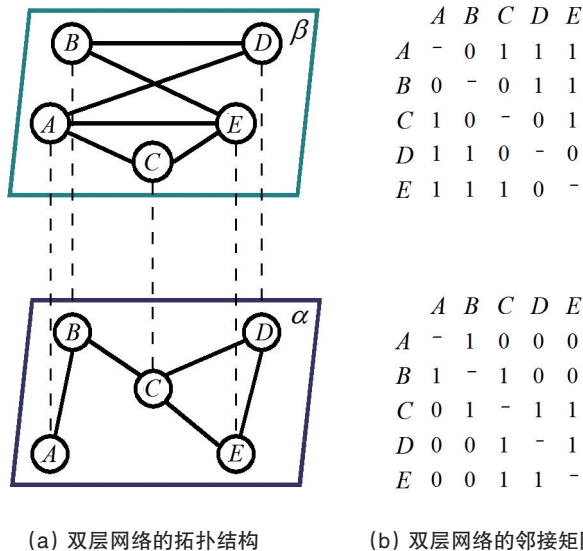
衡量两层网络链接之间的相关性,最常用的方式是计算两层网络的Pearson相关系数。假定 G_α 网络的邻接矩阵如图2(a)所示,它的拓扑结构如图2(b)所示。为比较两层网络之间的相关性,首先把每层网络对应邻接矩阵中的所有取值看成一个长向量,每个向量包含 $n(n-1)$ 个数字(对角线上的数字忽略不计)。就网络 G_β 而言,将它邻接矩阵中的数值先从左到右、再从上到下依次写下来,得到长向量是 $V^\alpha = [1, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1]$ 。对于另外一层网络 G_β ,也可以使用同样的方法得到它的长向量 V^β ,然后两个向量的Pearson相似系数就代表两个网络的相似性系数^[10]。

Pearson相关系数是最常用的计算节点之间度相关性的指标,同样可以拓展到双层网络中的度相关性测量中,具体公式为

$$p_{\alpha\beta} = \frac{\langle \nu_i^\alpha \nu_i^\beta \rangle - \langle \nu_i^\alpha \rangle \langle \nu_i^\beta \rangle}{\sigma(\nu_i^\alpha) \sigma(\nu_i^\beta)} \quad (4)$$

式中, ν_i^α 为 V^α 中的 i 个元素, ν_i^β 为 V^β 中的 i 个元素, $\sigma(\nu_i^\alpha) = \sqrt{\langle \nu_i^\alpha \nu_i^\alpha \rangle - \langle \nu_i^\alpha \rangle^2}$, 而 $\sigma(\nu_i^\beta) = \sqrt{\langle \nu_i^\beta \nu_i^\beta \rangle - \langle \nu_i^\beta \rangle^2}$ 。

采用公式计算,可以得到网络 G_α 和 G_β 之间的相关系数,



(a) 双层网络的拓扑结构

(b) 双层网络的邻接矩阵

图2 双层网络的拓扑结构和其对应的邻接矩阵

Fig. 2 Double-layer network topology and its corresponding adjacency matrix

该值为-1~1的数值。若该值大于0,可认为两个向量存在正相关;该值小于0,可认为两个向量存在负相关。但考虑到网络对应向量的特殊性,会发现 Pearson 相关系数的数值和网络的链接特性存在一定的关联性,即使是没有相关性的两个网络,它们之间的 Pearson 相关系数也可能不等于零,因此该统计特征的绝对数值和很多网络统计量一样,只有和它对应的零模型结果比较后才有实际意义。对于网络 G_β 而言,假定生成它的一个零模型网络 G'_β 如图2所示,该零模型的长向量是 [0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0], 也可以得到它和 G'_β 网络的 Pearson 相关性系数。从定性分析的角度,当实证双层网络的 G_α 和 G_β 之间的相关系数远远高于 G_α 和 G'_β 之间的相关系数,就认为 G_α 和 G_β 之间具有正相关性。相反,如果网络 G_α 和 G_β 之间的相关系数显著低于 G_α 和 G'_β 之间的相关系数,就认为 G_α 和 G_β 之间具有负相关性。

1.2.2 网络对应长向量的互信息

除了可以使用 Pearson 相关系数来计算两层网络对应长向量的线性相关系数,也可以使用互信息来计算两个长向量的非线性相关性。在信息论中,互信息主要是用来检测两个事件集合之间的非线性相关性来表示信息之间的关系,可以尝试用互信息这个统计量来检验两层网络间的结构相关性。

互信息的具体定义如下,设 X, Y 为两个离散时间序列,当 $X = \{x_1, x_2, \dots, x_n\}$, $Y = \{y_1, y_2, \dots, y_n\}$ 时的概率分别为 $p(x_i), p(y_i)$, ($i=1, 2, \dots, n; j=1, 2, \dots, m$)。其中,有 $\sum_i p(x_i) = 1$, $\sum_j p(y_j) = 1$ 。 X 和 Y 分别为信源时对应的信息熵为

$$H(X) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (5)$$

$$H(Y) = - \sum_{j=1}^m p(y_j) \log_2 p(y_j) \quad (6)$$

设由 X 与 Y 构成的联合信源 $XY = \{x_1, y_1, x_1 y_2, \dots, x_n y_m\}$ 时的概率分布为

$$XY: \begin{bmatrix} x_1 y_1, & x_1 y_2, & \dots, & x_n y_m \\ p(x_1 y_1), & p(x_1 y_2), & \dots, & p(x_n y_m) \end{bmatrix} \quad (7)$$

联合信源 XY 的信息熵为

$$H(XY) = - \sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) \log_2 p(x_i y_j) \quad (8)$$

则信源 X 与 Y 的互信息定义为(以 bit 为单位)

$$I(X, Y) = H(X) + H(Y) - H(XY) \quad (9)$$

互信息是对两个信源 X 与 Y 之间关联程度的度量指标,它满足对称性与非负性,即有

$$I(X, Y) = I(Y, X), \text{ 其中 } I(X, Y) \geq 0 \quad (10)$$

从绝对数值上看,当 X 与 Y 完全相同时,互信息最大;当 X 与 Y 完全独立时,互信息值为零^[11]。而在实际的应用中,互信息量要结合零模型网络才能发挥作用。由上述互信息的定义可知,如果采用 1.2.1 节中的方法将网络 G_α 映射成长向量 V^α , 同时将网络 G_β 映射成长向量 V^β 后,就可以计算 V^α 与 V^β 之间的互信息。

1.2.3 整体重叠程度

相对于 Pearson 相关系数、互信息等指标,还可以使用双层网络的整体重叠程度来刻画两层网络整体链接的相关性。网络的整体重叠程度定义为:双层网络中对应节点之间的连边重叠程度。如图2所示,在双层网络中,如果节点 i 和节点 j 在 G_α 网络中相连的同时在 G_β 网络中也相连,则称 $i-j$ 这条连边是重叠的。耦合网络中上下层网络连边的重叠程度,可通过链的重叠来公式化表示^[12]。经过归一化处理后,耦合网络的总体重叠程度可表示为

$$O^{\alpha\beta} = \frac{\sum_{i \neq j} A_{ij}^\alpha A_{ij}^\beta}{\min\{N^\alpha, N^\beta\}} \quad (11)$$

式中, N^α 为 G_α 中的连边数,而 N^β 为 G_β 网络中的连边数。如果在 G_α 网络中节点 i 与节点 j 之间相连,我们定义 $A_{ij}^\alpha = 1$, 若不相连则 $A_{ij}^\alpha = 0$ 。同理,在 G_β 网络中节点 i 与节点 j 之间相连,定义 $A_{ij}^\beta = 1$, 若不相连则 $A_{ij}^\beta = 0$ 。 $O^{\alpha\beta}$ 表示节点 i 与节点 j 在 G_α 网络中相连的同时也在 G_β 网络中相连的总链数^[13]。如果两个网络的链接特性完全相同,则 $O^{\alpha\beta}$ 的值取为 1。如果两个网络的链接特性完全不同,则 $O^{\alpha\beta}$ 的值为 0。在一般情况下, $O^{\alpha\beta}$ 应为 0~1 的某个值。

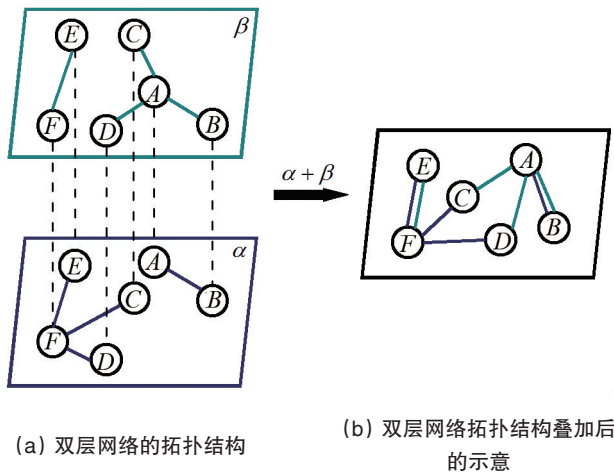


图3 双层网络整体重叠的示意

Fig. 3 Overall overlap of a double-layer network

1.2.4 局部重叠程度

除了定义全局重叠程度,针对每一个节点,也可以定义该节点的局部重叠程度。节点*i*的局部重叠程度定义为

$$O_i^{\alpha\beta} = \frac{\sum_{j=1}^{k_i^\alpha} A_{ij}^\alpha A_{ij}^\beta}{k_i^\alpha + k_i^\beta - \sum_{j=1}^{k_i^\alpha} A_{ij}^\alpha A_{ij}^\beta} = \frac{\sum_{j=1}^{k_i^\beta} A_{ij}^\alpha A_{ij}^\beta}{k_i^\alpha + k_i^\beta - \sum_{j=1}^{k_i^\beta} A_{ij}^\alpha A_{ij}^\beta} \quad (12)$$

式中, k_i^α 为节点*i*在 G_α 网络中的度, k_i^β 为节点*i*在 G_β 网络中的度。分子 $\sum_{j=1}^{k_i^\alpha} A_{ij}^\alpha A_{ij}^\beta$ 或 $\sum_{j=1}^{k_i^\beta} A_{ij}^\alpha A_{ij}^\beta$ 计算的是在双层网络 G_α 和 G_β 中节点*i*共用一个邻居节点*j*的数量^[14],而分母 $k_i^\alpha + k_i^\beta - \sum_{j=1}^{k_i^\alpha} A_{ij}^\alpha A_{ij}^\beta$ 或 $k_i^\alpha + k_i^\beta - \sum_{j=1}^{k_i^\beta} A_{ij}^\alpha A_{ij}^\beta$ 计算的是节点*i*在双层网络 G_α 和 G_β 中非重复链接的数量。分子和分母的比值为0~1。

局域重叠程度可以计算任意节点在双层网络上链接的重叠程度,但不能直接计算出两个网络整体上链接的相关程度。它的优点是可以组合出比较细致的特征,如将在网络 G_α 上度相同的节点集总起来计算它们的平均局域重叠程度,可以看出平均重叠程度这一指标是如何随着它们的度值而发生变化。可以将这一数值看成是检测双层网络链接整体相关性的一个统计量,而且展示出的信息比全局重叠程度更加丰富。

1.3 检测双层网络度度相关特性的统计量

在双层网络中,如果一层网络中的度大节点倾向于和另外一层的度大节点相连,这种倾向性即为双层网络中度匹配的正相关性,定义该双层网络为同配双层网络。如果一层网络中的度大节点倾向于和另外一层中的度小节点相连,这种倾向性即为度匹配的负相关性,定义该双层网络为异配双层网络。

1.3.1 双层网络度度相关的定义

将双层网络中节点的度度相关性定义为:如果同一节点在不同层上的度是相互关联的,即节点*i*在 G_α 网络中为度大

(小)节点的同时在 G_β 网络中也为度大(小)节点,称为正相关性双层网络^[14-15],如图4(a)所示。相反,如节点*i*在 G_α 网络中为度小(大)节点的同时在 G_β 网络中也为度大(小)节点,称为负相关性双层网络,如图4(c)所示。此外,如果两层网络的节点之间度值没有任何相关性,则该网络被称为无相关性双层网络,如图4(b)所示。

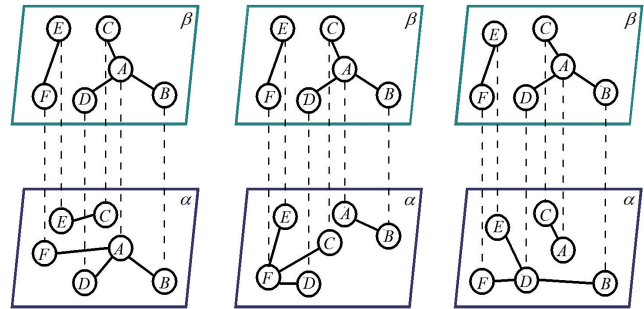


图4 不同度度相关性双层网络的示意

Fig. 4 Degree correlation of a double-layer network

基于上述双层网络中节点度度相关特性的定义,总结了4种计算方法去检测网络的度度相关特性:分别为度特征矩阵、余平均度、Pearson相关系数和Spearman相关系数^[13]。

1.3.2 度特征矩阵

度特征矩阵 $P(k^\alpha, k^\beta)$ 是一种对双层网络中节点度度相关性进行检测的方法^[9],它是相同节点在 G_α 网络的度值为 k^α 、在 G_β 网络的度值为 k^β 的数目。比如说,节点*i*在 G_α 网络中的度为 k^α ,在 G_β 网络中的度为 k^β ,则 $P(k^\alpha, k^\beta)$ 加1,直至统计完所有的节点。在得到 $P(k^\alpha, k^\beta)$ 值后,通过构建度度剖面来观察度特征矩阵。

1.3.3 余平均度

与单层网络类似,余平均度是另外一种刻画双层网络度相关性的方法。将 G_β 网络中度值为 k^β 的节点归类,归类后这些节点在 G_α 网络中的度值的平均值记为 $\bar{k}^\alpha(k^\beta)$,则有

$$\bar{k}^\alpha(k^\beta) = \frac{\sum_{k^\alpha} k^\alpha P(k^\alpha, k^\beta)}{\sum_{k^\alpha} P(k^\alpha, k^\beta)} \quad (13)$$

式中, $\bar{k}^\alpha(k^\beta)$ 可用于衡量 G_α 与 G_β 之间的相关性,如果函数 $\bar{k}^\alpha(k^\beta)$ 不随着度值 k^β 变化,则 G_α 与 G_β 之间是不相关的^[9]。如果函数 $\bar{k}^\alpha(k^\beta)$ 随着度值 k^β 的增加而增加,则 G_α 与 G_β 之间是正相关的。如果函数 $\bar{k}^\alpha(k^\beta)$ 随着度值 k^β 的增加而减小,则 G_α 与 G_β 之间是负相关的。

1.3.4 Pearson相关系数

Pearson相关系数除可以计算两层网络关联矩阵对应长向量的相关性,也是最常用的计算节点之间度相关性的指标,同样可以拓展到双层网络中的度相关性测量中,具体公

式为

$$r_{\alpha\beta} = \frac{\langle k_i^\alpha k_i^\beta \rangle - \langle k_i^\alpha \rangle \langle k_i^\beta \rangle}{\sigma(k_i^\alpha) \sigma(k_i^\beta)} \quad (14)$$

式中, k_i^α 为节点 i 在 G_α 网络中的度, k_i^β 为节点 i 在 G_β 网络中的度, $\langle k_i^\alpha k_i^\beta \rangle$ 为节点 i 在 G_α 网络中的度 k_i^α 与在 G_β 网络中的度 k_i^β 值乘积的平均数。 G_α 网络中 $\sigma(k_i^\alpha) = \sqrt{\langle k_i^\alpha k_i^\alpha \rangle - \langle k_i^\alpha \rangle^2}$, G_β 网络中 $\sigma(k_i^\beta) = \sqrt{\langle k_i^\beta k_i^\beta \rangle - \langle k_i^\beta \rangle^2}$ 。如果 $r_{\alpha\beta} > 0$, 那么网络是同配的; $r_{\alpha\beta} < 0$, 则网络是异配的。绝对值 $|r_{\alpha\beta}|$ 的大小, 反映了双层网络同配或异配的强弱程度。需要说明的是, Pearson 相关系数只有在 2 个输入变量均服从正态分布(更准确地说是二维正态分布)的情况下才能有效计算出两者之间的相关关系。在这里, 以 G_α 和 G_β 的度序列作为输入变量, 考虑到很多网络的度分布都是幂率的, 因此 Pearson 相关系数计算出的绝对数值有时是有偏差的。对于该问题, 一种解决方案是将每个节点的度值转化为它对应度值大小的升序排序, 然后再对其计算相关系数。

1.3.5 Spearman 相关系数

将双层网络中的节点按度值的顺序进行升序排列并生成相应的度序列 $\{x_i^\alpha\}$ 和 $\{x_i^\beta\}$, 节点 i 在 G_α 网络中的排列顺序记为 x_i^α , 节点 i 在 G_β 网络中的排列顺序记为 x_i^β , 则 Spearman 相关性系数 $\rho_{\alpha\beta}$ 可定义为

$$\rho_{\alpha\beta} = \frac{\langle x_i^\alpha x_i^\beta \rangle - \langle x_i^\alpha \rangle \langle x_i^\beta \rangle}{\sigma(x_i^\alpha) \sigma(x_i^\beta)} \quad (15)$$

式中, x_i^α 是节点 i 的度 k_i^α 在度序列 $\{k_i^\alpha\}$ 中的升序排名, x_i^β 是节点 i 的度 k_i^β 在度序列 $\{k_i^\beta\}$ 中的升序排名, $\sigma(x_i^\alpha) = \sqrt{\langle (x_i^\alpha)^2 \rangle - \langle x_i^\alpha \rangle^2}$, $\sigma(x_i^\beta) = \sqrt{\langle (x_i^\beta)^2 \rangle - \langle x_i^\beta \rangle^2}$ 。比较 Pearson 和 Spearman 两个相关系数的计算公式, 发现它们形式相同, 区别在于 Pearson 相关系数输入的是原始的度值, 而 Spearman 相关系数输入的是原始度值的排序值。Spearman 对变量的分布无要求, 它利用度的秩次进行计算, 可以用在度值任意分布的网络中。但考虑到网络中度相同的节点很多, 该方法根据网络中度序列来计算网络的相关性, 由于相同节点度的排序往往是随机的而不是唯一的, 因此通过该方法多次计算出来的 $\rho_{\alpha\beta}$ 值并不唯一, 存在一定的不稳定性。

1.4 检测双层网络中两层中富节点之间的连接相关性

在复杂网络中, 富节点(Hub 节点)指的是度大的节点, 由于度大节点拥有更多的连接关系, 因此在网络中相对于一般节点, 它们往往具有更高的地位、能起到更重要的作用^[16-17]。由于很多网络的度分布都具有幂率特性, 因此度大节点的度远远高于其他节点但是其数量却极少, 因此单独对富节点进行专门研究是必要的。本文中专门研究在双层网络中富节点之间的链接相关性, 即双层网络的富节点之间是否存在富

人俱乐部或非富人俱乐部现象。

1.4.1 双层网络的富人俱乐部系数定义

在单层网络中, 富人俱乐部现象是指富节点之间更倾向于相互连接这一特殊现象, 这一性质可以用富人俱乐部系数来表示^[18-19]。如果实证网络的富人俱乐部系数较高, 则说明实证网络存在富人俱乐部现象。相反, 如果实证网络的富人俱乐部系数较低, 则说明实证网络存在非富人俱乐部现象。

在双层网络中, 本研究不关注单层网络上富节点之间的链接关系, 而是关注于两层的富节点之间是否相关, 是否存在链接特性。因此, 双层网络中富节点之间的连接相关性指的是在一层网络中的富节点(度大节点)在另外一层网络中的对应节点是否也为富节点(度大节点)。如果在单层网络中某些节点为富节点, 而在另外一层网络中这些节点的对应节点也倾向于富节点, 则该双层网络具有富人俱乐部特性。相反, 如果在单层网络中某些节点为富节点, 而在另外一层网络中这些节点的对应节点倾向于都不是富节点, 则该双层网络具有非富人俱乐部特性。

1.4.2 双层网络富人俱乐部系数

考虑到双层网络中每层网络的节点度分布不同, 如果设定一个度值, 大于该度值为富节点, 小于该度值为非富节点, 那么双层网络中的富节点个数会不同, 尤其在上层和下层度分布相差较大的双层网络中, 这种不对称性会尤其明显。因此, 在这里不根据绝对的度值, 而是根据节点度大小的排序来确定富节点。在确定两层网络的富节点集合后, 双层网络的富人俱乐部系数可以定义为

$$\varphi_{\alpha\beta} = \frac{m(k^\alpha) \cap m(k^\beta)}{m(k^\alpha) \cup m(k^\beta)} \quad (16)$$

式中, $m(k^\alpha)$ 是 α 层网络中的富节点集合, 表示的是 α 层网络中度值大于 k^α 的节点, 而 $m(k^\beta)$ 是 β 层网络中的富节点集合, 表示的是 β 层网络中度值大于 k^β 的节点。考虑到两层网络的度分布可能具有差别, k^α 与 k^β 的值可以不同。因此, 上式表示双层网络的富人俱乐部系数为两层网络中富节点的交集除以富节点的并集。如果两层网络之间的富节点完全一致, 则富人俱乐部系数为 1。如果两层网络之间的富节点完全不同, 则富人俱乐部系数为 0。一般情况下, 两层网络之间的富节点有一定的重合性但是又不完全相同, 因此该值应该为 0~1。如果该值比较大, 表明实证网络有正向的富人俱乐部影响, 也就是说富节点之间更倾向于相互连接。相反, 如果该值比较小, 则表明原始网络有负向的加权富人俱乐部影响, 即富节点之间不倾向于相互连接。

1.4.3 富节点的度特征矩阵和余平均度

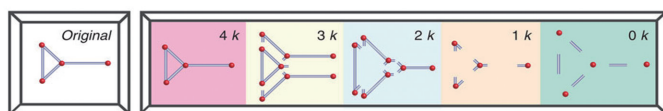
上文提到的度特征矩阵和余平均度是两种对双层网络中所有节点度度相关特性进行测量的方法, 富节点指的是网络中的度大节点, 它们自然也是在所有节点的集合中。因此, 如果分析度特征矩阵和余平均度中的度大节点(富节点)

体现出的特征,即仅研究富节点的度特征矩阵或余平均度,也可以描述双层网络的富人俱乐部现象。

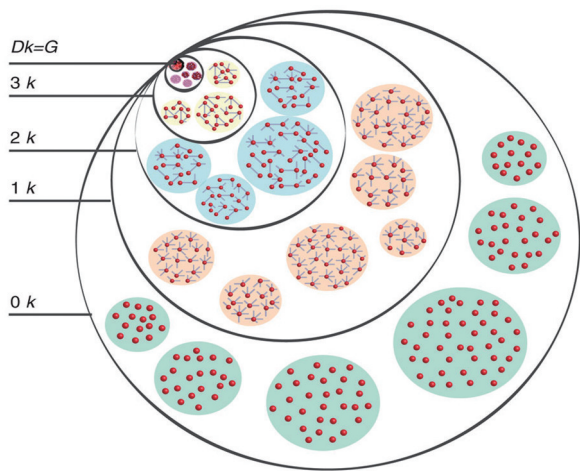
2 双层网络零模型的构造

2.1 基于层内断边重连的双层网络零模型

层内断边重连是构造双层网络零模型的一种方式,其主要思想是将双层网络中的其中1层进行连边置乱,但是层间节点的对应关系不会改变。在单层网络中,随机断边重连是最常用的零模型网络构造方式。使用网络零模型主要目的是检测实证网络的非平凡特性,这就需要从粗糙到精确逐步逼近原始网络的性质。Mahadevan 等^[20]为逐步逼近实证网络,根据原始网络的基本特性,可将网络中的基本组成单元分成不同的小模块如图5(a)^[21]所示,最左端为原始网络,右端从右到左依次介绍了产生不同阶数的零模型所需要原始网络的基本性质,从而引入了不同阶数网络零模型的概念,如图5(b)^[21]所示。不同阶数网络零模型之间并不是独立的,按照约束条件从少到多,存在一种包含关系,即 $0K \supset 1K \supset 2K \supset \dots (n-1)K \supset nK$, 任何一个 n 阶零模型都会包含 $(n-1)$ 阶零模型的性质。



(a) 原始网络以及 D_k 零模型的子图分布



(b) D_k 序列的收敛性与包含性分布

图5 不同阶数网络零模型之间的层次关系

Fig. 5 Hierarchical relationships between D_k null model network

0 阶零模型是最简单也是最随机化的网络零模型,只需要与原始网络具有相同的节点数和平均度。1 阶零模型与原始网络不但具有相同的节点数、相同的平均度,更重要的是具有相同的节点度分布 $p(k)$, 度分布是指原始网络中节点度

的概率或数目的分布。2 阶零模型与原始网络具有相同的联合度分布 $p(k_1, k_2)$, 联合度分布是指每条边两端连接节点的度值数目(概率)。2.25 阶零模型是根据原始网络中的联合度分布和平均聚类系数这两个物理属性构造的,要求与原始网络具有相同的联合度分布 $p(k_1, k_2)$ 和平均聚类系数 $\bar{C}(k)$ 。2.5 阶零模型与原始网络具有相同的联合度分布 $p(k_1, k_2, k_3)$ 和度相关的平均聚类系数 $\bar{C}(k)$, 3 阶零模型与原始网络具有相同的联合度分布,要求零模型网络和原始网络具有相同的开三角形和闭三角形分布。

随着生成零模型的约束条件逐步增加,构造的高阶零模型比低阶模型更接近于原始网络,因此精确的 3 阶零模型比 0~2.5 阶零模型更接近于原始网络。以此类推下去,理论上还可以定义更高阶的零模型^[7]。随着约束条件越多,可以被置乱的连边数也就越少,生成的网络零模型越接近于原始网络,但是这样进行连边置乱也就越困难。对于 3 阶网络零模型,该零模型受到的约束条件已经很多,多数情况下不能在一个有限的计算时间内生成相应的零模型,因此不具备实用性,对于大部分网络来说 2.5 阶零模型是现在能够实现的最高阶的实用零模型^[8,21]。

双层网络中同样采用随机断边重连的方式构造零模型。基于随机断边重连的双层网络零模型构造方式和单层网络零模型的构造方式大致相同,不同的是双层网络中各层共用一个节点集合,在进行断边重连过程中,保证双层网络中其中 1 层不变,而另 1 层网络根据不同的约束条件进行随机断边重连。

2.1.1 基于随机断边重连的 0 阶零模型

0 阶零模型网络中仅仅需要保证与原始网络的平均度不变,在对双层网络进行 0 阶零模型的构造过程中,保证 G_a 网络不做任何改变,根据生成 0 阶零模型的约束条件,将 G_b 网络进行连边的随机置乱。如图 6(a) 所示,每次随机选择原始网络中的 1 条边 CD , 先将 CD 这条边去掉,然后再随机选择网络中 2 个不相连的节点如 D 和 E , 并在他们之间添加 1 条连边,如图 6(b) 所示。为使网络充分随机化,一般应根据网络的规模将上述断边重连过程重复多次,直到原始网络足够随机化后才认为生成了对应的零模型网络。一般随机化的次数应该在 2 倍实证网络连边数以上。

2.1.2 基于随机断边重连的 1 阶零模型

1 阶零模型网络需要保证与原始网络的度分布不变,根据 1 阶零模型网络的约束条件,如图 7(a) 所示,每次随机的选择 G_b 网络中的两条边 BE 和 CD 。先将连边 BE 和 CD 边切断,再使节点 B 与 C 相连、 D 与 E 相连,从而得到两条新边 BC 和 DE ,如图 7(b) 所示。在随机断边重连的过程中,这 4 个节点的度值均保持不变,因此置乱前后保证了网络 G_b 的度序列不变。为了使网络充分随机化,一般应根据网络的规模将上述断边重连过程重复多次,直到原始网络足够随机化后才认为生成了对应的零模型。

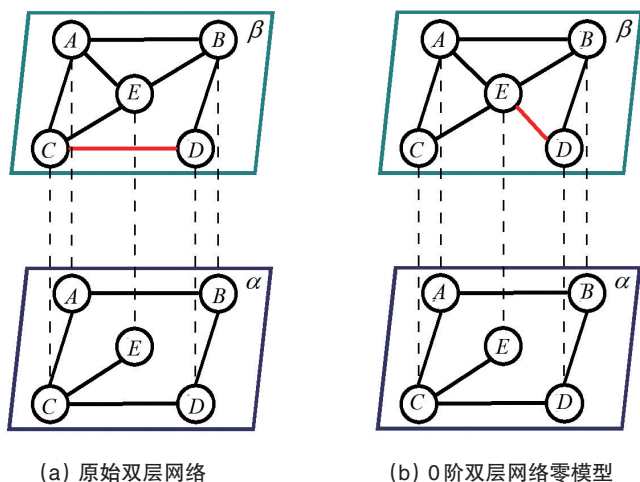


图6 基于随机断边重连生成0阶双层网络零模型的示意
Fig. 6 Random mixing process of 0k null model

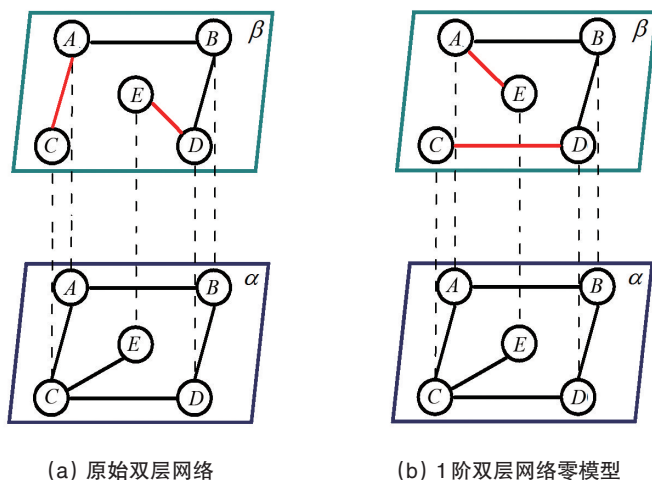


图7 基于随机断边重连生成1阶双层网络零模型的示意
Fig. 7 Random mixing process of 1k null model

2.1.3 基于随机断边重连的2阶零模型

2阶零模型需要保证与原始网络的联合度分布不变,每次随机选择 G_β 网络中的2条边如 AE 和 CD ,如图8(a)所示。节点 A 和 D 的度为2,节点 C 和 E 的度为1。为了保证断边重连前后 G_β 网络中2条边连接节点的度值不变,就将连边 AE 、 CD 断开,再使节点 A 与 C 相连, D 与 E 相连,从而得到2条新边 AC 、 DE ,如图8(b)所示。注意不能使 A 和 D 相连, C 和 E 相连,因为这样不满足联合度分布不变的约束条件。为使网络充分随机化,一般应根据网络的规模将上述断边重连过程重复多次,直到原始网络足够随机化后才认为生成了对应的零模型。

2.1.4 基于随机断边重连的2.5阶零模型

2.5阶零模型与原始网络具有相同的联合度分布和度相关的平均聚类系数。基于随机断边重连的2.5阶零模型构造的过程如图9所示。在原始网络 G_β 中随机选取2条边,首先

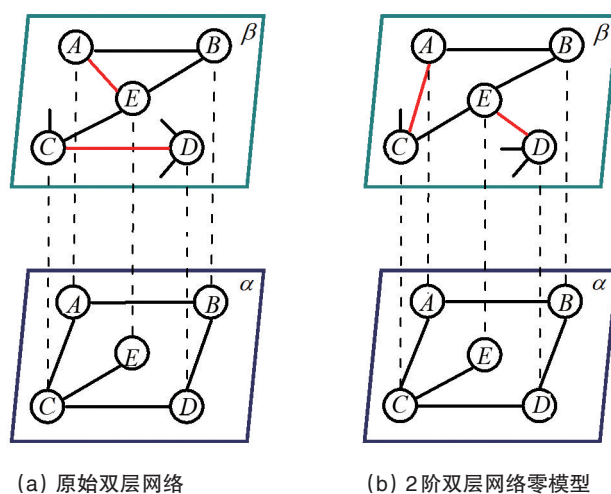


图8 随机断边重连生成2阶双层网络零模型的示意
Fig. 8 Random mixing process of 2k null model

要保证联合度分布不变即2阶零模型的基本性质不变。其次,断边重连之前计算2条连边两端每个节点聚类系数的值,以及邻居节点聚类系数的值。经过断边重连之后计算新生成的连边两端节点的聚类系数,以及邻居节点聚类系数的值。若断边重连前后所选边上的相同度的节点平均聚类系数的值不变,则证明断边过程是成功的,若聚类系数的值不一致,则证明所选的边不符合条件,需要重新再选。为了使网络充分随机化,一般应根据网络的规模将上述断边重连过程重复多次,直到原始网络足够随机化后才认为生成了对应的零模型。

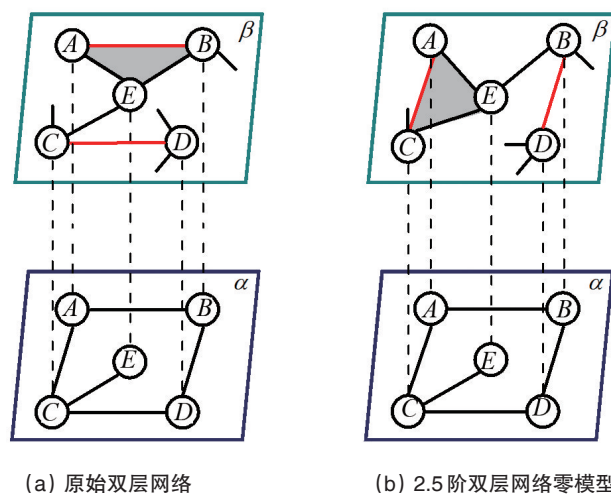


图9 基于随机断边重连生成2.5阶双层网络零模型的示意
Fig. 9 The random mixing process of 2.5k null model

2.2 基于节点置乱的双层网络零模型

前面介绍了不同阶次的随机断边重连零模型构造方法,其核心是不改变双层网络中每层节点的位置,仅置乱某一层节点之间的链接结构,因此零模型在拓扑结构上和原始网络

是不同的。下面介绍的基于节点置乱的方法,其核心是保持拓扑结构和原始网络相同的情况下随机置乱某一层节点的标号,使层间的链接关系发生变化,即在零模型的生成过程中每层网络的结构均保持不变,但是层间节点的对应关系发生改变。

2.2.1 基于节点随机置乱的零模型网络

节点随机置乱零模型的产生方法是将对应节点标号(位置)进行随机调换。节点随机置乱零模型的构造方式如图10所示。图10(a)为原始双层网络,网络中各层共用一个节点集合。图10(c)为随机置乱节点后的双层网络,在对双层网络进行随机置乱的过程中,保证 G_α 网络不变,仅置乱 G_β 网络。具体的置乱过程如图10(b)所示,经节点随机重排后的网络中,节点A置换到了节点D的位置,节点D置换到了节点A的位置。需要注意的是,仅仅将 G_β 网络中的节点标号进行随机重排,网络结构保持不变。除了像上述过程中那样一次置乱2个节点位置,也可以使用二次指派程序(quadratic assignment procedure, QAP)一次将所有节点的对应位置全部置乱^[10]。就其置乱的结果来看,2种方法的效果是一样的。

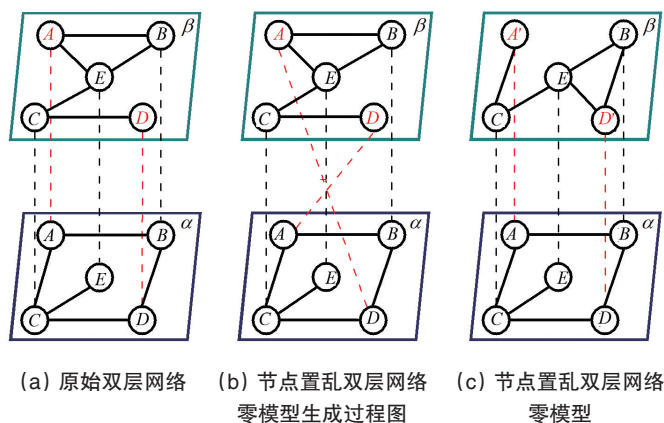


图10 基于节点随机置乱生成双层网络零模型的示意

Fig. 10 Process of node random rewiring for producing the null model

2.2.2 基于有倾向性节点置乱构造具有正匹配效应的零模型网络

双层网络匹配特性主要衡量的是相同节点在不同层网络中的相关性。有倾向性节点置乱算法构造的同配网络零模型过程如图11所示。在图11(a)中,原始网络中 G_α 网络的A节点为度大节点,E节点为度小节点;而原始 G_β 网络中A节点为度小节点,E节点为度大节点。为了构造同配网络,如图11(b)将 G_β 网络中A、E节点的标号互换位置,最终生成的网络如图11(c)所示, G_β 网络中的A'节点(交换前为E节点)与其对应的 G_α 网络的A节点都为度小节点, G_β 网络中的E'节点(交换前为A节点)与其对应的 G_α 网络的E节点都为度大节点。将这一过程反复进行,就可以生成强同配的网络。

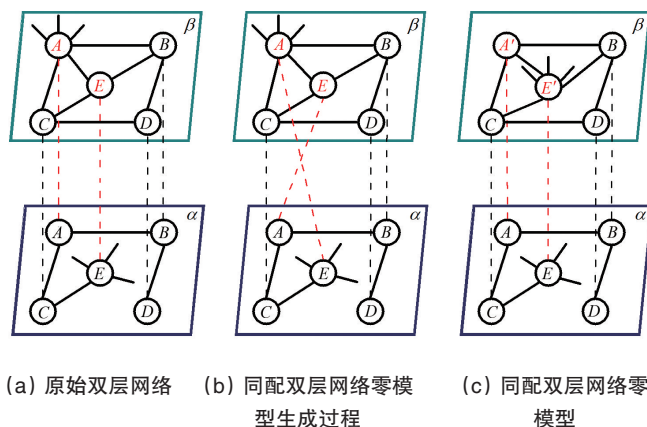


图11 基于有倾向性交换连边生成同配网络的示意

Fig. 11 Process of edge selective rewiring for producing the assortative null model

2.2.3 基于有倾向性节点置乱构造具有负匹配效应的零模型网络

有倾向性节点置乱算法构造负匹配网络零模型过程如图12所示。在图12(a)的双层网络中,无论在 G_α 网络或者 G_β 网络中,D节点均为度小节点,而E节点均为度大节点。为了构造异配网络,如图12(b)将 G_β 网络中D、E节点互换位置,最终生成的网络如图12(c)所示,此时 G_β 网络中的E节点(交换前为D节点)为度小节点与其对应的 G_α 网络的E节点为度大节点, G_β 网络中的D节点(交换前为E节点)为度大节点与其对应的 G_α 网络的D节点为度小节点。将这一过程反复进行,就可以生成强异配的网络。

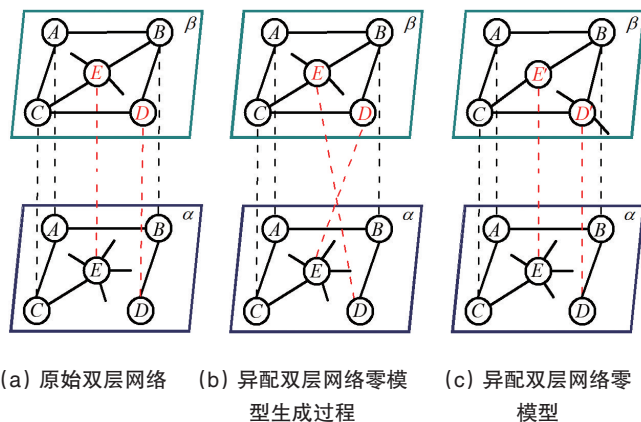


图12 基于有倾向性交换连边生成异配网络的示意

Fig. 12 Process of edge selective rewiring for producing the disassortative null model

2.2.4 基于有倾向性节点置乱生成具有富人俱乐部特性的零模型网络

在保证原始网络其他特性不变的情况下,尝试构造出具有或不具有富人俱乐部特性的双层网络零模型。这类零模

型网络有两方面的用处。一方面对于一个实证双层网络,在保持该网络某些特性不变的情况下,人们往往想知道该网络拥有富人俱乐部性质或不拥有富人俱乐部性质对于网络性质的影响。由于实证网络只有一个,那么由该网络来生成一系列可供比较的具有不同富人俱乐部特性的零模型网络就很有意义。另一方面,对于一个实证网络,即便是我们知道了该网络具有或不具有富人俱乐部特性,但是该网络具有该性质的强弱如果没有参考对象的话仍然是未知的。如果能构造出一系列富人俱乐部特性由弱到强的网络族作为参照,那么实证网络拥有的富人俱乐部特性有多强有多弱就一目了然了。

有倾向性随机节点置乱算法构造具有富人俱乐部特性的网络零模型的整个过程如图13所示。在图13(a)中,原始网络中 G_α 网络的 C 节点为非富节点, E 节点为富节点;而原始 G_β 网络中 C 节点为富节点, E 节点为非富节点。为了构造富人俱乐部网络,如图13(b)将 G_β 网络中 C 和 E 节点互换位置,最终生成的网络如图13(c)所示,此时 G_β 网络中的 E' 节点(交换前为 C 节点)与其对应的 G_α 网络的 E 节点都为富节点, G_β 网络中的 C' 节点(交换前为 E 节点)与其对应的 G_α 网络的 C 节点都为非富节点。将这一过程反复进行,就可以生成具有富人俱乐部特性的网络。

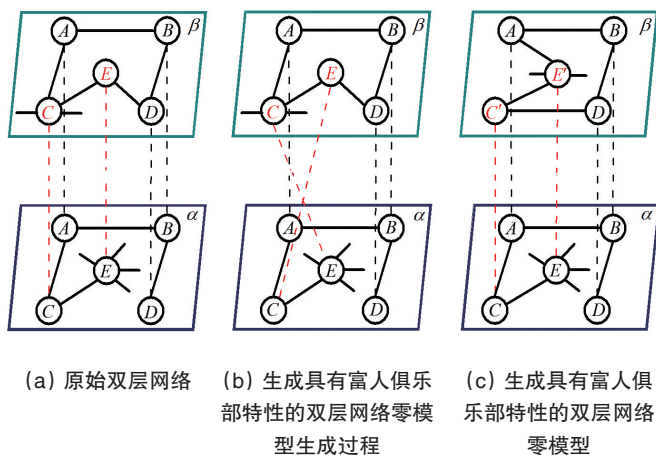


图13 根据节点置乱生成具有富人俱乐部特性的双层网络零模型示意

Fig. 13 Process of double network nodes rewiring producing the null model with rich club property

2.2.5 基于有倾向性节点置乱生成不具有富人俱乐部特性的零模型网络

有倾向性随机节点置乱算法生成不具有富人俱乐部特性的零模型网络如图14所示。在图14(a)中,原始网络中 G_α 网络的 C 节点为非富节点, E 节点为富节点;同时在 G_β 网络中 C 节点也为非富节点, E 节点也为富节点。为构造非富人俱乐部网络,如图14(b)将 G_β 网络中 C 和 E 节点互换位置,最终生成的网络如图14(c)所示,此时 G_β 网络中的 E' 节点(交换前

为 C 节点)为非富节点,与其对应的 G_α 网络的 E 节点为富节点, G_β 网络中的 C' 节点(交换前为 E 节点)为富节点,与其对应的 G_α 网络的 C 节点为非富节点。将这一过程反复进行,就可以生成不具有富人俱乐部特性的网络。

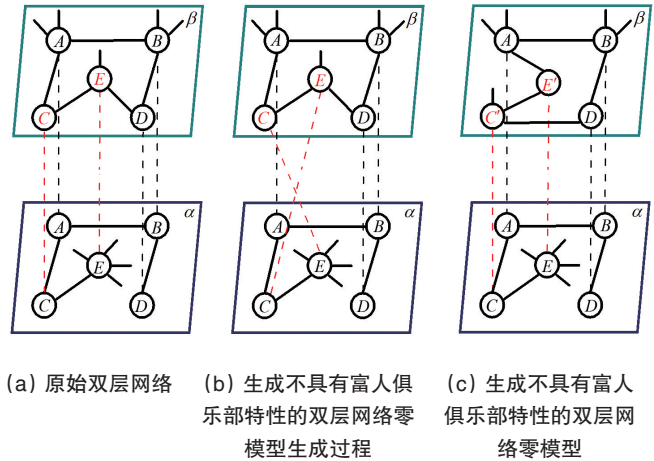


图14 根据节点置乱生成不具有富人俱乐部特性的双层网络零模型示意

Fig. 14 Process of double network node rewiring producing the null model without rich club property

3 检测双层网络链接相关性的假设检验方法

在计算出实证网络的某个统计量和对应的多个零模型网络统计量后,就需要采用假设检验来判断实证网络和零模型之间是否有显著性差异。其基本原理是假设将所有零模型看做是总体,而将实证网络看成是总体中的一个抽样,通过一些假设检验方法对得到的统计量进行统计推断,最后得出是否应该拒绝该假设的结论。

3.1 比值法

假设某种结构相关性统计量在实际网络中计算出来的值为 $N(G_\alpha, G_\beta)$,而在多个零模型中计算出来的平均值为 $\langle N_r(G_\alpha, G_\beta) \rangle$,则两者的比值为

$$R(G_\alpha, G_\beta) = \frac{N_r(G_\alpha, G_\beta)}{\langle N_r(G_\alpha, G_\beta) \rangle} \quad (17)$$

在实际应用中,如果 $R(G_\alpha, G_\beta) > 1$,那么意味着实际网络的设计或者演化过程促进了该连接相关性的出现;相反,如果 $R(G_\alpha, G_\beta) < 1$,则意味着实际网络的设计或者演化过程抑制了该连接相关性的出现。该方法适合于画图直观看出实证网络和零模型网络均值之间的差异,但是缺点是没有考虑到零模型之间的差异(方差),因此无法定量化推断出应该接受还是拒绝零假设。

3.2 Z检验方法

在统计学中Z检验方法(平均值差异检验方法)是一种参数统计方法,假定数据符合正态分布推断差异发生的概率,

从而比较两个平均数的差异是否显著,可以用Z检验方法检验结构相关性统计量在实际网络中出现的数值和在零模型中出现的数值差异的显著性。具体地说,连接相关性 $Z(G_\alpha, G_\beta)$ 的统计重要性可以用式(18)来刻画

$$Z(G_\alpha, G_\beta) = \frac{N(G_\alpha, G_\beta) - \langle N(G_\alpha, G_\beta) \rangle}{\sigma(G_\alpha, G_\beta)} \quad (18)$$

式中, $\sigma(G_\alpha, G_\beta)$ 是零模型中某个结构相关性统计量 $N(G_\alpha, G_\beta)$ 出现的平均数值 $\langle N(G_\alpha, G_\beta) \rangle$ 的标准差。Z值的绝对值越大就表示差异越显著,越能说明结构相关性 $N(G_\alpha, G_\beta)$ 产生于实证网络,而非随机造成的。在得到实证网络的 $N(G_\alpha, G_\beta)$ 值后,可以通过比较它和零模型所对应的均值 $\langle N(G_\alpha, G_\beta) \rangle$ 来分析实际网络的结构相关特性。在Z检验中,假设检验的统计量需要服从正态分布,对原始网络和零模型网络的统计量要使用高斯分布的参数检验方法进行检验,但很多情况下网络的统计量并不一定满足正态分布,此时使用参数检验方法就会有一定偏差。

3.3 秩检验方法

秩检验方法是一种非参数检验方法。Schreiber等^[22]提出的将统计量排序(Rank-Order)后再进行检验的秩方法。该方法首先产生 $1/\lambda - 1$ 组(单边检验)或 $2/\lambda - 1$ 组(双边检验)替代数据,然后计算原始数据和替代数据的检验统计量并进行排序,如果原序列的检验统计量为最小或最大,则拒绝零假设,检验的显著性水平为 $1 - \lambda$ 。该方法不需要对待假设的统计量分布做出限制,可以使用在总体方差未知或知道甚少的情况下,因此具有广泛的适用性。

4 参照零模型的实证双层网络结构相关性检测结果

4.1 所用实证网络的介绍

为度量双层网络之间的相关性,对本文中提出的度量方法和统计量基于实证网络并参照零模型进行了检验。本文中使用的网络为全球语言的多层网络,语言网络的实验数据取自文献[23],其中该多层网络有3层:一层为多种语言的Wikipedia网络,它包括134个语言节点;一层为多种语言的Twitter短信息网络,包含73个语言节点;还有一层是不同语言书籍之间的相互翻译关系网络,包括283个语言节点。

本文所构造的双层网络采用各层共用一个节点集合这一属性。为保证双层网络中 G_α 网络与 G_β 网络中的节点能够一一对应及数据的完整性,将Wikipedia和Twitter网络中的节点取并集,取并集后的语言节点有140个,并且将采用的这两个网络的原有节点都统一进行了编码。这样就构成了一个全球语言的Wikipedia和Twitter双层网络,且该双层网络共用一个语言节点集合。

4.2 基于互信息和比值法检测网络整体相关性

对于全球语言网络的Wikipedia和Twitter两层网络,它们

之间的互信息值如表1所示,计算出了不同阶数连边随机置乱和节点随机置乱零模型的互信息值的均值和标准差,每种零模型的样本数为50。由真实网络和零模型网络的互信息统计量比值可知,在连边置乱0阶零模型和节点置乱零模型中,真实网络中的互信息值远远大于零模型网络,说明与这2种零模型相比,真实网络特殊的拓扑结构促进了互信息值。相反,真实网络与连边置乱1阶和2阶零模型相比,其互信息比值略小于1,说明真实网络拓扑结构与这两种零模型网络结构相比,抑制了层间的互信息。最后,发现与连边置乱的2.5阶零模型相比,真实网络的互信息值又略高于该零模型。

表1 全球语言双层网络的互信息值

Table 1 Mutual information coefficient of the global language network

统计量	均值与标准差	原始网络与相应零模型的比值
原始实证网络	0.392	—
连边置乱0阶零模型	0.016±0.009	24.500
连边置乱1阶零模型	0.409±0.002	0.958
连边置乱2阶零模型	0.429±0.005	0.914
连边置乱2.5阶零模型	0.385±0.003	1.100
节点置乱零模型	0.040±0.009	9.800

4.3 基于Z检验方法和Pearson度相关系数检测度度相关性

对于全球语言网络的Wikipedia和Twitter两层网络,它们之间的Pearson度相关系数如表2所示,同时也计算出不同阶数连边随机置乱和节点随机置乱零模型的Pearson度相关系数的均值和标准差。每种零模型的样本数为50,样本数大于30可以认为是大样本。由于Pearson度相关系数的取值范围是在[-1,+1],需采用双边检测的方法。主要是要检测原始实证网络 and 不同零模型的Pearson度相关系数之间是否具有显著性差异,假定零模型的产生方法是将对应节点的标号(位置)进行随机调换。节点随机置乱零模型的Pearson度相关系数符合正态分布,那么95%的Z值是1.96,99%的Z值是2.58,原始实证网络对于连边随机置乱零模型的Z值为51.5远远大于2.58,因此两者之间存在显著性差异。同理,原始网络和节点置乱零模型之间也存在显著性差异。相反,连边随机置乱

表2 全球语言双层网络的Pearson度相关系数

Table 2 Pearson correlation coefficient of the global language network

统计量	均值与标准差	原始网络与相应零模型的比值
原始实证网络	0.515	—
连边置乱0阶零模型	0.309±0.004	51.5
节点置乱零模型	0.029±0.009	54.0
连边置乱1阶零模型	0.515±0.000	—

1阶零模型虽然仅保持了原始网络的度分布,但是该零模型和原始网络的Pearson度相关系数是完全相同的,两者之间没有显著性差异。同时,我们也认为该网络中度相关性的来源是由网络的度分布导致的。由于1阶以上的零模型均保持了网络的度相关特性,因此高阶零模型的Pearson度相关系数均为0.515,和原始网络的值是相同的。

4.4 基于秩检验方法和富人俱乐部系数检测富节点的链接相关性

对于全球语言网络的Wikipedia和Twitter2层网络,它们之间的富人俱乐部系数和对应的零模型中该统计量之间的秩检验如图15所示。秩检验方法不需要对待假设的统计量分布做出限制,可以使用在总体方差未知或知道甚少的情况下,因此具有广泛的适用性。对于每种类型的零模型取49组,由于双层网络的富人俱乐部系数范围是[0, 1],而真实网络的富人俱乐部系数为0.414,因此应采取双边检测方法。如图15所示,实证网络的富人俱乐部系数的排序值均比连边置乱0阶和1阶零模型、节点置乱零模型要高,因此其富人俱乐部特性和零模型相比具有显著性差异,显著性水平的衡量指标为 $1-\lambda$,其中 $\lambda=2/(49+1)=0.04$,实证网络和零模型已经具有很强的统计意义上的差异,具有较为明显的富人俱乐部性质。

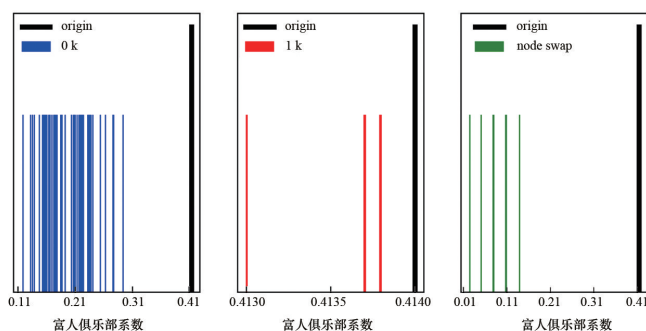


图15 全球语言双层网络的富人俱乐部系数

Fig. 15 Rich-club coefficient of the global language network

5 结论

综上所述,本文研究了双层网络的零模型构造及其在两层网络的结构相关性检测,重点总结了双层网络之间的连接相关性、度相关性和富节点之间的连接相关性检测算法,并应用于实证双层网络去检验本文算法的有效性。本文提出的参照网络零模型的研究方法,提供了一套统一的网络科学研究范式。一个好的网络零模型能为实证网络提供一个准确的基准,结合统计量指标就可以准确描述出实际复杂网络的非平凡特性以及这种非平凡特性的程度和来源。若实证网络与最简单的零模型在性质上没有差别,那么说明该网络没有非平凡性质。若已知实证网络与其对应零模型在性质

上具有统计意义上的差别,则说明该网络具有非平凡性质,并且对于不同网络可以通过比较同一个量化指标来说明它们拥有这种性质的程度。最终,参照零模型的假设检验方法遵循从简单到复杂、从低级到高级的顺序逐渐提高零模型和实证网络的吻合程度,可发现网络特殊性质是由哪些关键因素决定的。本文研究可检测出实证多层网络中任意两层层与层之间复杂的耦合作用,有助于学者们进一步深化对多层网络结构相关性、复杂性的理解和应用。

参考文献(References)

- [1] Lee K M, Min B, Goh K I. Towards real-world complexity: An introduction to multiplex networks[J]. The European Physical Journal B, 2015, 88(2): 1-20.
- [2] Boccaletti S, Bianconi G, Criado R, et al. The structure and dynamics of multilayer networks[J]. Physics Reports, 2014, 544(1): 1-122.
- [3] Milo R, Shenorr S, Itzkovitz S, et al. Network motifs: Simple building blocks of complex networks[J]. Science, 2002, 298(5594): 824-827.
- [4] Milo R. Superfamilies of evolved and designed networks[J]. Science, 2004, 303(5663): 1538-1542.
- [5] Foster J G, Foster D V, Grassberger P, et al. Edge direction and the structure of networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(24): 10815-10820.
- [6] Colizza V, Flammini A, Serrano M A, et al. Detecting rich-club ordering in complex networks[J]. Nature Physics, 2006, 2(3): 110-115.
- [7] Mahadevan P, Hubble C, Krioukov D, et al. Orbis: Rescaling degree correlations to generate annotated internet topologies[J]. ACM SIGCOMM Computer Communication Review, 2007, 37(4): 325-336.
- [8] Gjoka M, Kurant M, Markopoulou A. 2.5K-graphs: From sampling to generation[C]. IEEE International Conference on Computer Communications, Turin, Italy: 2012.
- [9] 汪小帆, 李翔, 陈关荣. 复杂网络引论——模型、结构与动力学[M]. 北京: 高等教育出版社, 2012.
- [10] 刘军. QAP: 测量“关系”之间关系的一种方法[J]. 社会, 2007, 27(4): 164-174.
- [11] Liu Jun. QAP: A unique method of measuring "Relationships" in relational data[J]. Society, 2007, 27(4): 164-174.
- [12] Fraser A M, Swinney H L. Independent coordinates for strange attractors from mutual information. Physical Review A, 1986, 33(2): 1134-1140.
- [13] Bianconi G. Statistical mechanics of multiplex networks: Entropy and overlap[J]. Physical Review E, 2013, 87(6): 62806.
- [14] Parshani R, Rozenblat C, Ietri D, et al. Inter-similarity between coupled networks[J]. Europhysics letters, 2010, 92(6): 2470-2484.
- [15] Szell M, Lambiotte R, Thurner S. Multirelational organization of large-scale social networks in an online world[J]. Proceedings of the National Academy of Sciences of the United States of America, 2010, 107(31): 13636-13641.
- [16] Min B, Yi S D, Lee K M, et al. Network robustness of multiplex networks with interlayer degree correlations[J]. Physical Review E, 2014, 89(4): 42811.
- [17] 姚尊强, 尚可, 许可. 加权网络的常用统计量[J]. 上海理工大学

- 学报, 2012, 34(1): 18–26.
- Yao Zunqiang, Shang Keke, Xu Xiaoke. Fundamental statistics of weighted networks[J]. Journal of University of Shanghai for Science and Technology, 2012, 34(1): 18–26.
- [17] Opsahl T, Colizza V, Panzarasa P. Prominence and control: The weighted rich–club effect[J]. Physical Retters Letters, 2008, 101(16): 168702.
- [18] Zhou S, Mondragon R J. The rich–club phenomenon in the internet topology[J]. IEEE Communications Letters, 2004, 8(3): 180–182.
- [19] Amaral L A N, Guimera R. Complex networks: Lies, damned lies and statistics[J]. Nature Physics, 2006, 2(2): 75–76.
- [20] Mahadevan P, Krioukov D, Fall K, et al. Systematic topology analysis and generation using degree correlations[J]. ACM Sigcomm Computer Communication Review, 2006, 36(4): 135–146.
- [21] Orsini C, Dankulov M M, Colomerdesimón P, et al. Quantifying randomness in real networks[J]. Nature Communications, 2015, 6: 8627.
- [22] Schreiber T, Schmitz A. Surrogate time series[J]. Physica D: Nonlinear Phenomena, 1999, 142(3/4): 346–382.
- [23] Ronen S, Goncalves B, Hu K Z, et al. Links that speak: The global language network and its association with global fame[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(52): E5616–E5622.

Correlation detection of double–layer network based on null models

CUI Liyan¹, XU Xiaoke^{1,2}

1. College of Information and Communication Engineering, Dalian Minzu University, Dalian 116600, China
2. Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China

Abstract Recently the framework of multi–layer networks was proposed as a new model of complex networks, and was used in widespread applications in many fields, such as the cascading failure, the information spreading, the link prediction and the network synchronization. In multi–layer networks, the correlation and the coupling between two layers of network structures might exist, so it is a very significant issue how to detect the structural correlation and quantify the correlation between the two layers. In this study we summarize and propose methods to measure the structural correlation of double–layer networks in three levels. The first level is to detect the overall connection relationship of the whole double–layer network. The second level is to test the degree correlation characteristics between all nodes at different layers. At last, the third level is to look for the connection relationship between the rich nodes at different layers. Although the three kinds of correlations are all dependent on the network statistics, these statistics are all without units. Furthermore, the sizes and the structures of different networks see a great difference. Therefore, absolute numerical values of some statistics often are not important and we put forward a variety of null models for double–layer networks as a reference. Through the hypothesis testing methods, we can quantify the structure correlation in double–layer networks, and try to analyze the intrinsic mechanism of inducing this kind of structure correlation. Finally, we use an empirical double–layer network (the global language multi–layer network) to verify the effectiveness of our methodology. This methodology can be used to detect the complex coupling between the layers in an empirical double–layer network, and for the better understanding of multi–layer networks and for new applications based on the structure complexity of multi–layer networks.

Keywords multi–layer network; null model; random mixing of edges; structure correlation

(责任编辑 田恬)