应用混合边模型在多层网络中进行社区检测

在单层网络中，已有前人提出混合边模型来刻画网络中的社区结构。 通过增加对层与层之间连边的考虑，本文将单层网络的混合边模型扩展到多层网络中。 同时，目用得比较多的用于刻画多层网络中社区结构的模型有：模块度(Modularity)，随机块模型(stochastic block model)和信息理论的方法(information-theoretic methods)这三种。通过对提出的混合边模型进行研究，证明可以由该模型等价地推导得到模块度和随机块模型(或者说，这两种模型是该模型的一种特例)。另外地，假设有一类刻画多层网络下社区结构的模型，该类模型目标函数由网络中每个社区的层内和层间连接关系的线性组合来表示。 本文亦证明了这类模型可由提出的混合边模型推导得到。

单层网络中刻画社区结构的混合边模型可以表示如下：

$$\begin{aligned}\mathcal{H}(g) = -\sum_{i\neq j} a_{ij} & \underbrace{A_{ij}\delta(g_i,g_j)}_{\text{Internal existing edges}} \\ +\sum_{i\neq j} b_{ij} & \underbrace{(1-A_{ij})\delta(g_i,g_j)}_{\text{Internal non-existing edges}} \\ +\sum_{i\neq j} c_{ij} & \underbrace{A_{ij}\big[1-\delta(g_i,g_j)\big]}_{\text{External existing edges}} \\ -\sum_{i\neq j} d_{ij} & \underbrace{(1-A_{ij})\big[1-\delta(g_i,g_j)\big]}_{\text{External non-existing edges}},\end{aligned} \tag{1}$$

where A_{ij} is the edge strength of nodes i and j, g_i indicates the label of the community that node i belongs to, and $\{a,b,c,d\}$ is the set of free hyper parameters. The delta function δ is the Kronecker delta. A lower $\mathcal{H}(g)$ value indicates a better partition. Despite great

在这个模型中，属于同一社区中的两点 i, j: 若它们间有边连接，则给予一个奖励项 a; 若它们之间无边连接，则给予一个惩罚项 b; 对不属于同一社区中的两点 i,j: 若它们之间有边连接，则给予一个惩罚项 c; 否则给予 d 的奖励。注: 该模型中奖励项前面是负号，惩罚项前面是正号，所以该模型的这个目标函数显然是值越低，表示社区划分得越好。另外，目标函数中的 delta 函数 delta(gi, gj)表示当 gi=gj 时，值为 1，否则为 0. 该模型的上述思路不适合于对二分图中的社区进行检测，因为二分图中同一社区的点倾向于无连边，连边更倾向于分布在不同的社区之间。

考虑层与层之间的连边(这种边叫 coupling，以区别于层内的连边 edge)，将该模型扩展到多层网络后，可表示如下:

$$
\begin{aligned}
\mathcal{M}(v) = & \sum_{i\neq j,s} \underbrace{\lambda(\{a\})a_{ijs}A_{ijs}P(v_{is},v_{js})}_{\text{Internal existing edges}} \\
& + \sum_{i\neq j,s} \underbrace{\lambda(\{b\})b_{ijs}(1-A_{ijs})P(v_{is},v_{js})}_{\text{Internal non-existing edges}} \\
& + \sum_{i\neq j,s} \underbrace{\lambda(\{c\})c_{ijs}A_{ijs}[1-P(v_{is},v_{js})]}_{\text{External existing edges}} \\
& + \sum_{i\neq j,s} \underbrace{\lambda(\{d\})d_{ijs}(1-A_{ijs})[1-P(v_{is},v_{js})]}_{\text{External non-existing edges}} \\
& + \sum_{s\neq r,i} \underbrace{\lambda(\{e\})e_{isr}C_{isr}P(v_{is},v_{ir})}_{\text{Internal existing couplings}} \\
& + \sum_{s\neq r,i} \underbrace{\lambda(\{f\})f_{isr}(1-C_{isr})P(v_{is},v_{ir})}_{\text{Internal non-existing couplings}} \\
& + \sum_{s\neq r,i} \underbrace{\lambda(\{g\})g_{isr}C_{isr}[1-P(v_{is},v_{ir})]}_{\text{External existing couplings}} \\
& + \sum_{s\neq r,i} \underbrace{\lambda(\{h\})h_{isr}(1-C_{isr})[1-P(v_{is},v_{ir})]}_{\text{External non-existing couplings}},
\end{aligned}
\tag{3}
$$

同单层网络中的混合边模型对比，由于考虑了层与层之间的连边，其相应地增加了 4 个参数: e, f, g, h(注: 单层网络里的点的奖惩项下标为 ijs，表示位于同一层 s 的不同点 i 和 j；而 e, f, g, h 则为了表示层间的关系，下标 isr 表示分别位于 s 层和 r 层的同一点 i)；另外地，之前表示惩罚/奖励项的前面分别用正、负号加以区别，现在则引入了 lambda 函数，可以更灵活地表示各种边对目标函数的影响。该函数的意义如下:

$$
\lambda(\{\omega\}) = \begin{cases} +1 & \text{if a larger value of } \omega \text{ leads} \\ & \text{to a greater contribution,} \\ -1 & \text{otherwise.} \end{cases}
\tag{4}
$$

亦即: 如果 w 的值越大表示其对目标函数的贡献越大的话，那么该函数就取 1，否则取-1。最后，该模型亦引入了一个矩阵 P(Vis, Vjr)来衡量位于 s 层的点 i 与位于 r 层的点 j，这两点属于同一社区的概率；这种方法有别于之前的 delta 函数，delta 函数是: 如果 i,j 所属的社区 gi 和 gj 相同，则为 1，否则为 0. 显然地，引入概率的表示方法更为灵活(delta 函数是其特例)。

总结一下这个扩展到多层网络的混合边模型，其实就是考虑了 8 种不同的边的线性表示，涉及到了 8 个参数{a, b, c, d, e, f, g, h}.

此外，值得注意的是，当 a, b, c, d, e, f, g, h 的取值相差很大时，这个模型有可能”退化”(比如 a 能取到的最大值是其他的 100 倍的话，那么这个模型就被和 a 相关的边所控制，其他参数涉及的边在这个模型就不起作用了)。为避免这种情况发生，模型对各参数作了如下限制:

$$\frac{\max_{x_1} \lambda(\{\omega_1\})\omega_1(x_1)}{\max_{x_2} \lambda(\{\omega_2\})\omega_2(x_2)} \to 1,$$
$$\forall \omega_1, \omega_2 \in \{a, b, c, d, e, f, g, h\}.$$

亦即：各参数对目标函数能起到的最大影响是相当的，谁也不会被忽略掉。

下面证明这个灵活的混合边模型可以等价地表示 3 类模型。

首先，拥有 8 个参数的这个混合边模型可以等价地表示**模块度模型**，亦即存在一组解使得模块度模型能由该混合边模型表示得到。证明过程如下，令:

$$\text{Here we take } \lambda(\{\omega\}) = +1 \text{ for } \omega \in \{a, d, e, h\},$$
$$\lambda(\{\omega\}) = -1 \text{ for other hyper parameters and}$$

$$\begin{cases} \text{existing edges} & a_{ijs} = c_{ijs} = 1 - \gamma_s p_{ijs} \\ \text{non-existing edges} & b_{ijs} = d_{ijs} = \gamma_s p_{ijs} \\ \text{existing couplings} & e_{isr} = g_{isr} = \varsigma \\ & f_{isr} = h_{isr} = 0, \end{cases} \tag{5}$$

其中 γ_s 为涉及社区检测的分辨率参数(通常取为 1，该参数设的越大，即：分辨率越高，则网络中的社区分得越细，此时原先一个大的社区可能被分为若干小社区)； P_{ijs} 表示零模型下位于同一层 s 中点 i 和点 j 的邻接情况； ς 表示层间的连接强度。
由此可将混合边模型(3)化为:

$$\begin{aligned} \mathcal{M}(v) &= \sum_{i \neq j, s} (A_{ijs} - \gamma_s p_{ijs})[2P(v_{is}, v_{js}) - 1] \\ &\quad + \sum_{s \neq r, i} \varsigma C_{isr}[2P(v_{is}, v_{ir}) - 1] \\ &= \underbrace{2 \sum_{ijsr} \left[(A_{ijs} - \gamma_s p_{ijs}) P(v_{is}, v_{js}) + \tilde{C}_{isr} P(v_{is}, v_{ir}) \right]}_{\text{Fuzzy multilayer modularity}} \\ &\quad \underbrace{- \sum_s (1 - \gamma_s) m_s - \sum_i m'_i}_{\text{Constant}}, \end{aligned} \tag{6}$$

where we utilize the fact that $\sum_{ij} A_{ijs} = \sum_{ij} p_{ijs}$ and define $\tilde{C}_{isr} = \varsigma C_{isr}$, $m_s = \sum_{ij} A_{ijs}$ and $m'_i = \sum_{sr} \tilde{C}_{isr}$ to keep the notations uncluttered.

最后，前面提到 delta 函数是矩阵 P(Vis，Vjr)是的特例，设 P 矩阵为对角阵，且其中的元素仅相应地取为 0 或 1，即可进一步地用 delta 函数来表示上式:

$$\mathcal{M}(v) = 2\sum_{ijsr}\left[(A_{ijs} - \gamma_s p_{ijs})\delta_{sr} + \tilde{C}_{isr}\delta_{ij})\right]\delta(v_{is}, v_{jr}) - \sum_s (1-\gamma_s)m_s - \sum_i m'_i, \tag{7}$$

where the δ function is the Kronecker delta.

优化(7)式亦即等价地优化模块度模型的目标函数，故得证。
注：模块度模型的目标函数如下:

$$Q_{\text{multislice}} = \frac{1}{2\mu}\sum_{ijsr}\left[\left(A_{ijs} - \gamma_s \frac{k_{is}k_{js}}{2m_s}\right)\delta_{sr} + \delta_{ij}C_{jsr}\right]\delta(g_{is}, g_{jr})$$

(其中的 2u，我的理解是网络中所有点的度数和，$\kappa_{is}\kappa_{js}/2m_s$ 为零模型，即：P_{ijs}，意为在单层网络中保持各点的度的前提下，随机情况里，i 和 j 这两点的期望连接强度。)

接着，证明**随机块模型**亦可由该混合边模型表示得到。
随机块模型是一种生成式模型，其目标函数可由如下似然函数表示:

$$\mathcal{L}(\mathbf{A}, \mathbf{C}|\mathbf{P}) = \prod_{i\neq j,s} P(v_{is}, v_{js})^{A_{ijs}}[1 - P(v_{is}, v_{js})]^{1-A_{ijs}} \prod_{s\neq r,i} P(v_{is}, v_{ir})^{C_{isr}}[1 - P(v_{is}, v_{ir})]^{1-C_{isr}}. \tag{8}$$

对社区进行划分，亦即最大化该似然函数。作 log 变换后，得到:

$$\begin{aligned}\log \mathcal{L}(\mathbf{A}, \mathbf{C}|\mathbf{P}) = &\sum_{i\neq j,s} A_{ijs}\log P(v_{is}, v_{js}) \\ &+ \sum_{i\neq j,s}(1 - A_{ijs})\log[1 - P(v_{is}, v_{js})] \\ &+ \sum_{s\neq r,i} C_{isr}\log P(v_{is}, v_{ir}) \\ &+ \sum_{s\neq r,i}(1 - C_{isr})\log[1 - P(v_{is}, v_{ir})].\end{aligned} \tag{9}$$

令:

$$\begin{cases} a_{ijs} = \log P(v_{is}, v_{js}) / P(v_{is}, v_{js}) \\ d_{ijs} = \log [1 - P(v_{is}, v_{js})] / [1 - P(v_{is}, v_{js})] \\ e_{isr} = \log P(v_{is}, vir) / P(v_{is}, vir) \\ h_{isr} = \log [1 - P(v_{is}, vir)] / [1 - P(v_{is}, vir)] \end{cases}$$

且$\forall \omega \in \{a,d,e,h\}, \lambda(\omega) = +1; b = c = f = g = 0$ 代入混合边模型(3)即可得到式(9); 得证。

最后对于**一类目标函数**可由网络中每个社区的层内和层间连接关系的线性组合来表示的模型，亦可证明其能由混合边模型得到。
该类模型的目标函数为式(11):

Here, we will concentrate on a specific kind of evaluators, the value of which is represented as the sum of a linear combination of within-layer and between-layer adjacency within every communities, which qualifies the detected communities based on the linking structures within them:

$$E = \sum_{ijsr} \Big\{ \big[k_1(\mathbf{x_1}) A_{ijs} + k_2(\mathbf{x_2})\big] \delta(v_{is}, v_{jr}) + \big[k_3(\mathbf{x_3}) C_{ijr} + k_4(\mathbf{x_4})\big] \delta(v_{is}, v_{jr}) \Big\}, \tag{11}$$

where k_x is an arbitrary function except for those taking the adjacency $\mathbf{A}$ and $\mathbf{C}$ directly as inputs. Here we consider a hard partition of the community assignment as is widely adopted in the literature.

该类模型的目标函数大括号内为 [k1 * Aijs + k2] + [k3 * Cijr + k4], 中括号的第一部分考虑了层内连接情况，第二部分考虑的是层间连接情况(Cijr 是否应为 Cisr?)，这两部分都由线性函数(线性组合)表示; 其中的 delta 函数表示该目标函数仅考虑点 i 和点 j 位于同一社区时的情况(当 i 和 j 位于不同社区时 delta 函数为 0).

如果将混合边模型式(3)换成另一种类似于式(11)的表示，则可表示成下式:

$$\begin{aligned}\mathcal{M}(v) = \sum_{ijs} \Big[& (a_{ijs} - b_{ijs} - c_{ijs} + d_{ijs}) A_{ijs} \delta(v_{is}, v_{js}) \\ & + (b_{ijs} - d_{ijs}) \delta(v_{is}, v_{js}) \Big] \\ + \sum_{isr} \Big[& (e_{isr} - f_{isr} - g_{isr} + h_{isr}) C_{isr} \delta(v_{is}, v_{ir}) \\ & + (f_{isr} - h_{isr}) \delta(v_{is}, v_{ir}) \Big] \\ + & constant, \end{aligned} \tag{12}$$

where *constant* refers to the terms that are independent of the community assignment. Here we omit the λ function to keep the notations uncluttered, so from here to

其中, constant 项表示和社区划分无关的常数项; 同上将 P 矩阵简化为 delta 函数; 此外, 8 个参数前的 lambda 函数也被省略了, 所以此时这 8 个参数是包含了相应的 lambda 函数信息的, 而不再只是单纯的值。

比较(11)和(12)式, 可得到一组方程:

$$\begin{cases} a_{ijs} - c_{ijs} = k_1 + k_2 \\ b_{ijs} - d_{ijs} = k_2 \\ e_{isr} - g_{isr} = k_3 + k_4 \\ f_{isr} - h_{isr} = k_4. \end{cases} \tag{13}$$

其中混合边模型有 8 个参数, 而上述这类模型只有 4 个参数, 可见上述方程的解是不唯一的。若再考虑前面推导模块度模型时用到的那组解, 即式(5), 将其代入式(13), 则可得:

$$\begin{cases} a_{ijs} - c_{ijs} = 2(1 - \gamma_s p_{ijs}) > 0 \\ b_{ijs} - d_{ijs} = -2\gamma_s p_{ijs} < 0 \\ e_{isr} - g_{isr} = 2\varsigma > 0 \\ f_{isr} - h_{isr} = 0, \end{cases} \tag{14}$$

(注: 如上提到的式(13)中的各参数是带 lambda 函数信息的, 所以代入时须将式(5)中各参数的 lambda 值也考虑进去。)

如何解读式(14)呢? 我们知道参数 a, c 考虑的是同一层 s 内的点 i 和 j 存在连边的情况, 而 a 是表示 i 和 j 位于同一社区, c 是 i 和 j 位于不同社区; 此时 k1 + k2 = a - c > 0, 表示我们的目标函数更希望位于同一层内的两点被划分到同一社区; 类似地, k3 + k4 = e - g > 0, 表示我们的

目标函数更希望层间连边上的两点被划分到同一社区；而 b - d < 0 则表示目标函数更希望不同的社区间不存在连边(社区间的连边是稀疏的)；最后，f, h 表示的是不同层间不存在连边的情况，f 表示同一社区不同层间不存在连边，h 表示不同社区间不同层不存在连边，f - h = 0 表示不同层不存在连边的整体情况对目标函数没有影响。上述结论正好和模块度模型的思路是一致的，只是阐述的角度不一样。

由于该混合模型涉及 8 个参数，表示灵活，故本文的实验部分主要探讨的是参数的不同赋值对模型的影响，如何对这 8 个参数进行合理赋值这类问题，此处未细读，略。

综上，混合边模型由于具有 8 个参数，表示灵活，可以由此等价地推导出模块度模型、随机块模型以及一类目标函数由网络中每个社区的层内和层间连接关系的线性组合来表示的模型。通过混合边模型，我们可以加深对模块度模型、随机块模型的理解；若对该模型中的 8 个不同参数赋予合理的不同权重，说不定我们还能得出更好的多层网络下的社区检测模型。（全文完）

林伟
20160802

论文:

［1］ Han Zhang et al. (2016)　Community Detection Using Multilayer Edge Mixture Model

链接: https://arxiv.org/abs/1605.07055

后记：本文原为我第二次的组会论文讲解内容，但感觉难度较大，讲公式比较 boring，所以作罢，用论文报告的形式补上，拖延时日，及至今日．秀隆也有关于这篇论文的报告在 bbs 里，感兴趣的可对照阅读．理解不当之处，还望指正．上述标红色字体部分，为我不了了之之处。