arXiv:2304.05894v1 [cs.LG] 12 Apr 2023

适用于加权标记网络的动态混合成员随机块模型

Gaël Poux-Médard 里昂大学第二校区，UR 3083，法国布龙
gael.poux-medard@univ-lyon2.fr

Julien Velcin 里昂大学第二校区，UR 3083，法国布龙
julien.velcin@univ-lyon2.fr

Sabine Loudcher 里昂大学第二校区，UR 3083，法国布龙
sabine.loudcher@univ-lyon2.fr

摘要

大多数真实世界网络会随时间演变。现有文献提出的动态网络模型要么不带标签，要么假设网络只有一种成员结构。另一方面，一类新型混合成员随机块模型（MMSBM）可在混合成员聚类假设下对静态标记网络建模。本文提出将这一类模型扩展到混合成员假设下的动态标记网络推断。我们的方法通过对模型参数设置时间先验来实现，其依据的唯一假设是动态变化并非突发性的。我们表明，该方法与现有方法有显著差异，并能够对更复杂的系统——动态标记网络——进行建模。我们通过在合成数据集和真实世界数据集上开展多项实验，证明了该方法的稳健性。我们的方法一项重要优势是只需极少的训练数据便能取得良好效果。在具有挑战性的条件下，性能提升拓宽了自动化学习工具的应用范围，例如社会科学领域。社会科学涵盖众多领域，而小型数据集是引入机器学习方法的一大障碍。

CCS 概念

• **信息系统** → **聚类**；《社会网络》；社会推荐；**数据挖掘**；协同过滤；网络数据模型；• **计算数学** → 概率推断问题；**贝叶斯网络**。

关键词

块模型、动态网络、聚类、推荐系- 统

ACM 参考格式：
Gaël Poux-Médard、Julien Velcin 和 Sabine Loudcher。2023年。面向加权标记网络的动态混合成员随机块模型。载于 第*46*届国际*ACM*信息检索研究与开发会议论文集（*SIGIR'23*），*2023*年*7*月*23*日至*27*日，中国台湾台北。ACM，美国纽约，纽约州，美国，9页。https://doi.org/10.1145/3539618。3591675

允许为个人或 乐彩购彩平台彩票平台彩票软件无码不卡
允许将本材料用于课堂教学，前提是不得为营利或商业利益制作或分发副本，并且副本首页须载有本声明及完整引文。对于本作品中由作者以外的其他人拥有版权的组成部分，必须予以尊重。注明出处后进行摘要是允许的。如需以其他方式复制、再版、发布到服务器或重新分发至列表，则必须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。*SIGIR '23*，*2023*年*7*月*23*日至*27*日，中国台湾台北。© 2023年版权所有归所有者/作者所有。出版权已授权给 ACM。ACMISBN 978-1-4503-9408-6/23/07...$ 15.00https://doi.org/10.1145/3539618.3591675

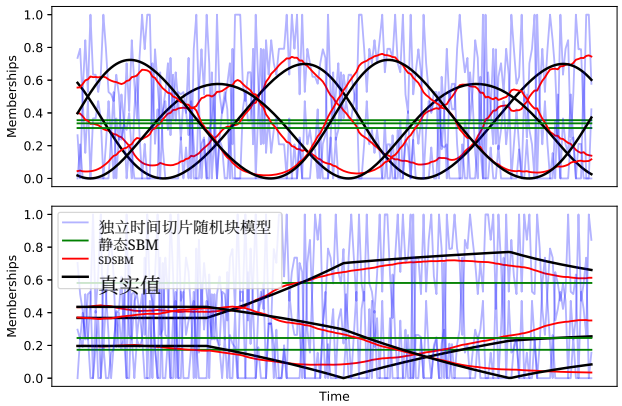


图1：用户对群组的偏好会随时间变化——一位音乐听众可能会循环地偏爱摇滚乐、爵士乐或流行音乐（上图），也可能毫无特定规律地听这些音乐（下图）。对于包含200个时期、每个时期仅有5次观测的数据，我们的方法（红色）能够推断出任意平滑的动态成员模式，且比静态模型（绿色 [9]）以及将每个时间切片独立考虑的模型（蓝色 [22]）推断得更准确。

1 引言

动态网络是可视化和建模不同实体之间随时间演化的交互的强大工具。网络的节点代表相互作用的实体，而这些节点之间的连接代表一种交互。在许多现实世界的情境中，连接的强度会随时间变化——例如，在音乐流媒体网站上，用户对各种音乐流派的偏好会随时间变化[14, 23] ——参见图 1 作为示例。该网络被称为动态网络。

如今，并非每次互动都具有相同的重要性。一位音乐听众可能既喜欢摇滚乐也喜欢爵士乐，但更偏爱其中一种。该用户与各音乐流派之间的连接强度并不相同；每条关系边都关联一个数值，用以表示互动的强度。这种网络被称为动态加权网络。

最后，简单的动态加权网络可能不足以捕捉特定情境的复杂性。音乐听众对不同音乐流派可能有不同的看法：可能喜欢、不喜欢或觉得厌烦，也可能只偏好在早晨、夜晚等时段听某些流派。每种关系都可以在网络中用一条独立的关系边表示，并为其赋予相应的强度。这样的网络称为动态加权带标签网络。

这类网络是高维对象，直接进行推断是一项艰巨的任务。已有研究提出了多种在静态情境下完成这一任务的方法；随机块模型（SBM）系列是最受欢迎的方法之一 [5, 10, 12]。其基本假设是，某些节点集合在与其他节点集合交互时表现相似。它们可以归入同一个簇；无需为网络中的每个节点逐一建模每条边，只需对簇之间的边进行建模，从而使推断任务更易处理。每个簇都与带标签的边相关联，每个节点则与一个簇相关联。随机块模型的一种变体具有更强的表达能力，称为混合成员随机块模型（MMSBM），其中每个节点可以按不同比例属于多个簇 [2, 9, 22, 24]。由此推断出的网络是加权且带标签的，但并非动态网络。随机块模型系列的一大优势是能够产生易于解释的结果（可解释的簇），这与现有大多数用于带标签网络的神经网络方法不同 [6]。

本文旨在利用混合成员随机块模型推断动态、加权且带标签的网络。我们的贡献在于通过使用时间狄利克雷先验，扩展了一大类混合成员随机块模型。我们首先仔细回顾了动态网络推断领域的文献，并指出此前没有研究涉及我们所关注的任务。尽管一些先前的研究探讨了类似问题，但它们并不适用于推断动态、加权且带标签的网络。此外，与文献中提出的方法相比，我们的方法在概念上要简单得多。最后但同样重要的是，我们提出的动态扩展可直接应用于大多数现有的、用于带标签和加权网络的混合成员随机块模型。

我们开发了一种 EM 优化算法，其计算规模随数据集大小线性增长；通过合成数据集和现实世界数据集展示了该方法的有效性，并进一步详细介绍了一种可能的应用场景。

2 背景

2.1 记号

我们考虑一个包含 I 个节点和 O 个标签的网络。本节讨论的所有聚类模型都可以用一组受限的参数表示：矩阵 $\theta(t) \in \mathbb{R}^{I\times K}$ 用于描述每个 I 个节点对每个 K 个可能簇的成员归属，块交互矩阵 $p(t) \in \mathbb{R}^{K\times K\times O}$ 则表示每个 K 个簇生成每个 O 个标签的概率。$\theta(t)$ 和 $p(t)$ 都可以随时间变化；下文中，除非另有说明，这种时间依赖是隐含的（$x(t) := x$）。当 $O = 1$ 时，该网络称为无标签网络；若一条边只能存在或不存在，则称为二元网络（相对于加权网络而言）。

2.2 动态无标签网络——单成员身份

单隶属关系随机块模型考虑如下成员矩阵$\theta \in \{0; 1\}^{I\times K}$：每个成员向量仅在一个簇上的取值为 1，在其他所有位置上的取值均为 0——这在文献中称为“硬”聚类。在 [28,29], 中，作者提出使用标签切换推断方法对二元无标签动态网络进行建模，并通过序贯蒙特卡洛算法 [13]进行优化。成员矩阵和交互矩阵都可以随时间变化，因此假设存在两个相互独立的底层马尔可夫过程 [13]。

在 [21, 30], 中，作者提出对二元动态无标签网络进行建模。簇交互矩阵 p 可以随时间变化，同时保持隶属关系 θ 静态不变。p 的各个条目均从狄利克雷分布中抽取，并表示为中国餐馆过程。该过程收敛于狄利克雷分布，并支持推断潜在的无限多个簇。因此，该模型是非参数的，并使用马尔可夫链蒙特卡洛算法进行推断。

文献 [16, 17] 提出了一种新颖的概念性方法，用于在单成员假设下对动态无标签网络建模。作者提出将簇交互矩阵和成员矩阵建模为泊松过程，以明确表示连续时间依赖性，而无需将数据集切分为多个时间片段。该方法可以推断动态二元网络或加权网络中随时间变化的成员矩阵和交互矩阵，但其结果表明，允许两者同时变化会导致可识别性和标签切换问题 [8]。鉴于这些随机块模型算法都无法达到似然函数的全局最优解，这一结论似乎合理。允许 θ 和 p 随时间变化的模型更有可能陷入局部最优解。

最后，我们提及随机块模型变体的存在，这些变体考虑了动态度修正 [25] 或强制实现无标度特性 [26]。

所有这些方法都考虑无标记网络，并采用硬聚类，其表达能力不及混合成员方法。

2.3 动态无标签的网络- 混合成员

混合成员随机块模型考虑诸如 $\theta \in \mathbb{R}^{I\times K}$ 这样的成员矩阵，其中每个成员向量 θ_i 的 L_1 范数为 1。文献中也将其称为“软”聚类。

与 [21, 30], 类似， [7] 提出了一种用于推断动态二值无标签网络的方法。根据某个节点此前与各个簇关联的次数，为每个信息片段的成员向量从中国餐馆过程（CRP）中抽取。该过程会在无限多个可用簇上产生一个分布。采用 CRP 表述自然而然，因为成员向量的先验通常是狄利克雷分布，而 CRP 会自然收敛到从狄利克雷分布中抽取的样本 [3]。块交互矩阵 p 不随时间变化。文章给出的复杂度分析表明，该方法的复杂度为 $O(N^2)$，因此不适用于大规模现实应用。

与我们最密切相关的工作是 [27]。这项开创性工作提出了 dMMSB，并将其作为一种使用变分算法对动态二元无标记网络建模的方法 [15]。为此，作者修改了原始的 MMSBM [2]，将逻辑正态分布作为成员向量的先验 θ_i。这一选择使得能够对成员向量演变之间的相关性建模 [1]。随后，成员向量被嵌入状态空间模型中，也就是说，在该空间中，可以为给定变量定义两个时间点之间的线性转移。作者将成员向量的这种轨迹定义为前一时间点的线性函数。该轨迹的估计采用了.....

并使用卡尔曼滤波器进行平滑。这种方法与我们的方法最为接近，因为它通过隶属关系上的先验分布来建模时间依赖性，记为 $P(\theta)$。 [27] 已扩展为将 $P(\theta)$ 视为逻辑斯蒂正态混合先验 [11], ，从而提升了其表达能力。

然而，该模型及其相关扩展并不适用于当前任务。首先，它考虑的是无标签的二值网络 [15],，若要扩展到带标签的加权网络，并非易事。所提出的优化算法在每次EM迭代中都需要循环直到收敛，因此无法处理大型数据集。此外，要使该方法奏效，簇交互矩阵 p 必须保持静态，而我们在此对这一限制进行了缓解。最重要的是，它假设状态空间中的时间步之间存在线性转移，而我们提出的方法不对任何核函数作此假设。

2.4 静态带标签的网络 - 混合成员

近年来，使用MMSBM变体 [9, 18–20, 22]推断静态加权带标签网络的贝叶斯方法日益增多。在这里，节点与一种类型 a相关联；每种类型的节点在多部网络中都有其自身的层，并通过成员矩阵 θ^a与该节点相关联的可用簇集合。来自某一层的节点按照混合比例属于与该层对应的每个簇。随后，所有层的簇之间的交互产生一个关于 O 可能输出标签的分布。由此形成了用于带标签网络的静态MMSBM。我们的贡献将 [9, 18–20, 22] 的所有内容扩展到了动态情形。这些工作将在第 3 节中进一步详述。

请注意，在 [22], 中，作者对数据进行时间切片，并将每个切片视为相互独立；每个时间切片都被视为三部网络中的一个节点。我们将在后文比较我们的方法与这种建模方式。

我们以这些工作作为基础模型，并将其与参数的动态先验分布相耦合。我们的工作重点是使 θ和 p的先验概率都具有时间依赖性，以便对这些参数的动态变化进行建模。我们为本节介绍的每项工作提供了可直接使用的时间插件。它适用于混合成员背景下的动态、加权和带标签的网络，并通过可扩展的变分EM算法进行推断。

3 动态带标签随机块模型

3.1 基础模型

为清晰起见，我们选择考虑带标签的标记混合成员随机块模型的 **最简单形式**：每个 I 节点都以混合比例与每个 K 簇相关联，而每个 K 簇又分别与 O 标签相关联（ [20]中的SIMSBM(1)）。如 [20], 我们对 SIMSBM(1) 的演示所示，其可平凡地扩展到 [9] （ SIMSBM(1,1) ）、[22]（SIMSBM(2,1)）和 [19] （SIMSBM(2)）——一般而言，扩展到任何表示为 SIMSBM(x,y,...) 的模型。

我们考虑一组 I节点，它们在离散时间间隔（或时期）内可以关联到 O 个可能的标签，该间隔记为 $t_\circ$。我们假设，在每个时间步，每个 I 节点都属于 K个可用簇的混合，每个簇又会产生一个关于 O 个标签的概率分布。每个 I 节点对每个 K 个可能簇的隶属关系在时间 t 由成员矩阵 $\theta_{(t)} \in \mathbb{R}^{I\times K}$编码。向量 $\theta_{(t)\,i}$表示i 在时间 t属于任意一个 K 簇的概率，并按如下方式归一化：

$$\sum_{k\in K} \theta_{i,k}^{(t)} = 1 \;\forall i, t \tag{1}$$

在时间 t，每个 K 簇与每个O标签相关联的概率由矩阵 $p_{(t)} \in \mathbb{R}^{K\times O}$编码。矩阵中的元素$p_{(t)\,k}(o)$表示簇 k 与标签 o 在时间 t 相关联的概率，因此其归一化方式如下：

$$\sum_{o\in O} p_k^{(t)}(o) = 1 \;\forall k, t \tag{2}$$

最后，节点 i在时间 t与标签 o相关联的概率（即在时间 t， i与 o之间存在一条边的概率）写作：

$$P(i \to o|t) = \sum_{k\in K} \theta_{i,k}^{(t)} p_k^{(t)}(o) \tag{3}$$

给定一组观测三元组 $R_\circ(i,\ o,\ t)$，模型的后验分布可写为 [9, 20]:

$$P(\theta, p|R^\circ) \propto P(R^\circ|\theta, p)\prod_t P(\theta^{(t)})P(p^{(t)}) \tag{4}$$

$$= \prod_{(i,o,t)\in R^\circ} \sum_{k\in K} \theta_{i,k}^{(t)} p_k^{(t)}(o) \prod_t \left(\prod_i P(\theta_i^{(t)}) \prod_k P(p_k^{(t)})\right)$$

现在，在描述优化过程之前，我们必须选择先验 $P(\theta_{(t)})$ 和 $P(p_{(t)})$。

3.2 SimpleDynamic SBM（SDSBM）先验

我们基于一个假设，为 $\theta_{(t)}$和 $p_{(t)}$定义先验分布：给定时间上的参数不会在短时间尺度内突然变化。这意味着，对于任意足够接近 t_1的 t_2，元素 $\theta_{(t_1)\,i,k}$与 $\theta_{(t_2)\,i,k}$的差异不应显著。与参考时间相近的元素称为**时间邻居**。

我们对每个条目 θ先验 $t_{(i)}$和 $p_{(t)\,k}$的先验知识是，它们与时间邻居的差异不应显著。这与 [27], 有着根本区别：后者使用卡尔曼滤波器估计下一时刻的参数值，而该滤波器只考虑前一个时间步。此外，作者假设线性转移函数，而我们不作此假设。图 2 展示了所提方法的示意图，其中成员关系的先验概率随时间的变化取决于其时间邻居。

3.2.1 狄利克雷分布。由于每个条目 $\theta_{(t)\,i}$和 $p_{(t)\,k}$都归一化为 1，我们将狄利克雷分布视为先验，它自然会生成归一化向量，使得 $\sum_n x_n = 1$。其形式为：

$$Dir(x|\alpha) = \frac{1}{B(\alpha)}\prod_n x_n^{\alpha_n - 1} \tag{5}$$

其中， $B(\cdot)$ 是多元贝塔函数。在公式5中，向量 α被称为浓度参数，必须提供给模型。该参数定义了狄利克雷分布的众数和方差。

我们将浓度参数设为 $\alpha = 1 + \beta\alpha_0$，因此当 $\beta = 0$ 时，我们可以得到单纯形上的均匀先验，且 $\beta \geq 0$

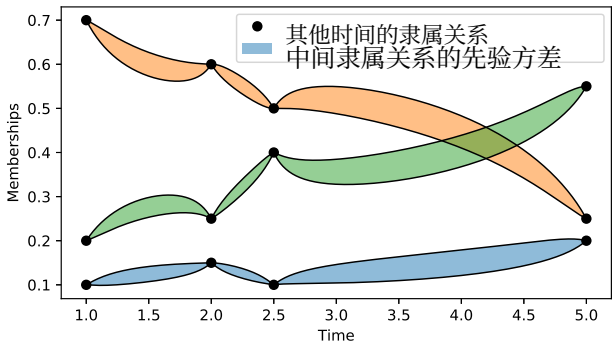


图 2：根据时间邻域，先验概率在所有时刻的隶属关系上的方差——给定一些时间邻居（黑点），隶属关系条目先验的方差（填充曲线；我们用不同颜色表示了 3 个此类条目以作说明）随时间的变化。该示例采用平均核 $\kappa(t, t\prime) = \frac{1}{|t-t'|}$。在时刻 t 推断参数 $x(t)$ 时，其先验概率 $P(x(t))$ 的方差取决于 t 相对于时间邻居的位置。例如，此处在 $t = 2$ 处方差为零，因为 $\kappa(t, t\prime)$ 发散，$\alpha(t)$ 也随之发散，因此方差收敛至 0。

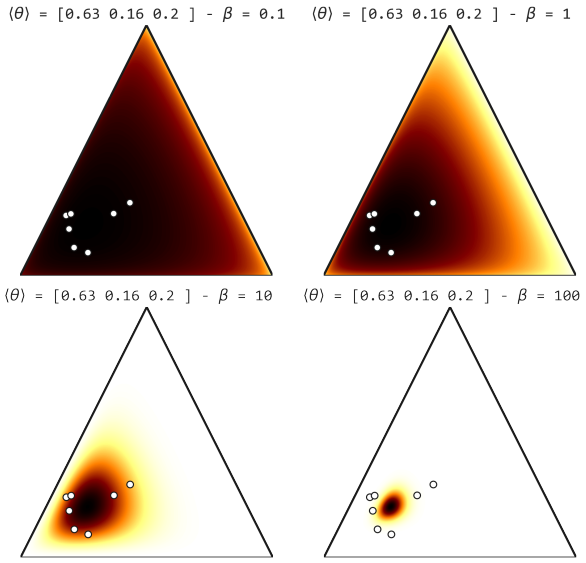


图 3：根据时间邻居，β 取不同值时成员向量上的先验概率——颜色越深，概率越高。投影到单纯形空间（3 个轴的范围均为 0 到 1）。白点表示所考虑三维向量的时间邻居。为便于说明，使用均匀权重函数 $\kappa(t, t\prime)$，其平均值表示为 $\langle\theta\rangle$。β 控制变量在其邻居周围的先验方差。

从而使先验具有唯一的众数。从公式5（或众数）中抽取的最频繁值为 $\mathbb{M}(x_n) = \frac{\alpha_n - 1}{\sum'_n (\alpha'_n - 1)} = \frac{\alpha_0 \sum_n \alpha_{0,n} - n}{\sum_n \alpha_{0,n} - n}$ 对于 $\beta = 0$，我们得到均匀先验；随着 $\beta \gg 1$ 趋于 $\frac{1}{\beta}$，方差逐渐消失。β 的不同取值对先验分布的影响如图3所示。

3.2.2 简单动态先验的众数。 我们的唯一约束是，$\theta(t)i$ 和 $p(t)k$ 不会在较小的时间尺度内发生突变。为此，我们根据相邻时间点定义它们先验概率的众数。超参数 β 控制先验的方差，即该约束对推断过程的影响程度。对于 $\theta(t)i$，我们将简单动态先验的参数表示为：

$$\alpha_{i,k}^{(t,\theta)} = 1 + \beta \underbrace{\left(\frac{\sum_{t' \neq t} \kappa(t,t') \theta_{i,k}^{(t')}}{\sum_{t' \neq t} \kappa(t,t')} \right)}_{\langle \theta_{i,k}^{(t)} \rangle} \tag{6}$$

其中，$\kappa(t, t\prime)$ 是一个权重函数，而 $\alpha(t,\theta)$ 对应于 θ 在时间 t 的浓度参数。在接下来的实验中，我们将权重函数定义为 $\kappa(t, t\prime) = \frac{N_{t'}}{|t-t'|}$，其中 $N_{t'}$ 是在时间 $t\prime$ 所获得的观测值数量。这样一来，时间邻居的影响会随时间距离的增加而减小，呈反比关系。我们在图 2 中展示了这一特定核函数对所有时刻成员归属先验概率的影响。具体而言，我们可以看到，采用这一表达式时，当所考察的时刻非常接近某个时间邻居时，先验概率的方差趋于 0。

先验的众数随后是其时间邻居在 $\kappa(t, t')$ 加权下的平均值，记作 $\theta(t)i,k$。注意，这成立是因为 $\sum_k \theta(t)i,k = 1 \; \forall i, t$。此外，先验方差是 β 的减函数；当 $\beta = 0$ 先验在单纯形上均匀分布时，而当 $\beta \to \infty$ 方差趋近于 0，如图3所示。同样的推理适用于 $p(t)k$，其先验参数 $\alpha(t,p)\,k,o = 1 + \beta p(t)k(o)$，以及 [9, 19, 20, 22] 中的任何其他参数。

3.2.3 先验表达式。 最后，我们给出关于 $\theta(t)i$ 和 $p(t)\,k$ 的最终对数先验：

$$P(\theta_i^{(t)} | \{\theta_{i,k}^{(t')}\}_{t' \neq t}) \propto \prod_k \theta_{i,k}^{(t)\,\beta \langle \theta_{i,k}^{(t)} \rangle} \tag{7}$$

$$P(p_k^{(t)}(o) | \{p_k^{(t')}(o)\}_{t' \neq t}) \propto \prod_o p_k^{(t)}(o)^{\beta \langle p_k^{(t)}(o) \rangle}$$

为清晰起见，我们省略了归一化因子——它不会影响推断过程。

3.3 推断

3.3.1 E 步。我们提出一种 EM 推断程序，用于最大化式 8 中定义的对数后验分布。期望步计算给定 R。的每个条目时，潜变量（此处为簇 k）被选中的期望概率。由于此类潜变量不会出现在先验表达式中，因此期望

步不受引入简单动态先验的影响；一般来说，先验分布不会影响期望步的计算 [4]。此类带标签网络的 E 步已在多处讨论过 [9, 19, 22]。它可以推导如下：

$$\begin{aligned}\log P(\theta,p|R^\circ) &\propto \log P(R^\circ|\theta,p)\prod_t\prod_i P(\theta_i^{(t)})\prod_k P(p_k^{(t)})\\ &= \sum_{(i,o,t)\in R^\circ}\log\sum_{k\in K}\theta_{i,k}^{(t)}p_k^{(t)}(o)\\ &\quad+\sum_t\sum_i\log P(\theta_i^{(t)})\sum_k\log P(p_k^{(t)})\\ &\geq \sum_{(i,o,t)\in R^\circ}\sum_{k\in K}\omega_{i,o}^{(t)}(k)\log\frac{\theta_{i,k}^{(t)}p_k^{(t)}(o)}{\omega_{i,o}^{(t)}(k)}\\ &\quad+\sum_t\sum_i\log P(\theta_i^{(t)})\sum_k\log P(p_k^{(t)})\end{aligned}\tag{8}$$

在公式 8 中，我们引入了一个提议分布 $\omega_{(t)\,i,o}(k)$，它表示给定观测值 (i,o,t) 时，某个簇分配 k 的概率。最后一行是在假设 $\sum_k\omega_{(t)\,i,o}(k)=1$ 的情况下，根据 Jensen 不等式得出的。我们注意到，当满足以下条件时，Jensen 不等式取等号：

$$\omega_{i,o}^{(t)}(k)=\frac{\theta_{i,k}^{(t)}p_k^{(t)}(o)}{\sum_{k'}\theta_{i,k'}^{(t)}p_{k'}^{(t)}(o)}\tag{9}$$

它为我们提供了在观测值（i,o,t）$\in R_\circ$下潜变量 k的期望，记作 $\omega_{(t)\,i,o}(k)$。

利用这一表达式，我们可以将对数似然 $\log P(R_\circ|\theta,p)$ 重写为 [4, 9, 20]：

$$\log P(R^\circ|\theta,p)=\sum_{(i,o,t)\in R^\circ}\sum_{k\in K}\omega_{i,o}^{(t)}(k)\log\frac{\theta_{i,k}^{(t)}p_k^{(t)}(o)}{\omega_{i,o}^{(t)}(k)}\tag{10}$$

3.3.2 M 步。 回到公式 8 的第一行，并代入公式 7 和公式 10，可得后验分布的无约束表达式。我们引入拉格朗日乘子，以考虑公式 1 ($\phi_{(t)\,i}$) 和公式 2 ($\psi_{(t)\,i}$) 的约束，最后计算关于模型参数的最大化方程。从成员矩阵的元素 $\theta_{(t)\,i,\ k}$ 开始：

$$\begin{aligned}&\frac{\partial\left(\log P(\theta,p|R^\circ)-\sum_{i',t'}\phi_{i'}^{(t')}(\sum_{k'}\theta_{i',k'}^{(t')}-1)\right)}{\partial\theta_{i,k}^{(t)}}=0\\ &\leftrightarrow\sum_{o\in\partial(i,t)}\frac{\omega_{i,o}^{(t)}(k)}{\theta_{i,k}^{(t)}}+\frac{\beta\langle\theta_{i,k}^{(t)}\rangle}{\theta_{i,k}^{(t)}}-\phi_i^{(t)}=0\\ &\leftrightarrow\sum_{o\in\partial(i,t)}\omega_{i,o}^{(t)}(k)+\beta\langle\theta_{i,k}^{(t)}\rangle=\phi_i^{(t)}\theta_{i,k}^{(t)}\\ &\leftrightarrow\sum_{o\in\partial(i,t)}\underbrace{\sum_k\omega_{i,o}^{(t)}(k)}_{=1\text{ (Eq.9)}}+\beta\underbrace{\sum_k\langle\theta_{i,k}^{(t)}\rangle}_{=1\text{ (Eq.1)}}=\phi_i^{(t)}\\ &\leftrightarrow\frac{\sum_{o\in\partial(i,t)}\omega_{i,o}^{(t)}(k)+\beta\langle\theta_{i,k}^{(t)}\rangle}{N_{i,t}+\beta}=\theta_{i,k}^{(t)}\end{aligned}\tag{11}$$

其中 $\partial(i,t)=\{o|(i,\cdot\ t)\in R_\circ\}$ 是与 i 和 t 都相关联的标签子集，而 $N_{i,t}=|\partial(i,t)|$ 是该集合的大小。注意，对于$\beta=0$，我们可以得到标准静态 MMSBM 模型 [9, 19, 20, 22] 的 M 步。

对于各项 $p_{(t)\,k}(o)$，M 步的推导完全相同，结果为：

$$p_k^{(t)}(o)=\frac{\sum_{(i,t)\in\partial o}\omega_{i,o}^{(t)}(k)+\beta\langle p_k^{(t)}(o)\rangle}{\sum_{(i,o,t)\in R^\circ}\omega_{i,o}^{(t)}(k)+\beta}\tag{12}$$

3.4 讨论

3.4.1 与现有模型的结合。 在现有 MMSBM 模型中使用 SDSBM（简单动态随机块模型）作为时间扩展，只需进行极少改动。在 [9, 19, 20, 22], 中，其引入归结为在最大化方程的分子中添加一项 βx，并在分母中添加相应的归一化项 β。这样，我们的方法便可直接用于构建动态网络模型。

3.4.2 灵活的动态建模。 该先验允许部分参数随时间变化，而其他参数保持不变。例如，当涉及多个成员矩阵时，如 [9, 20, 22])，将某些矩阵的 $\beta=0$ 设为相应值，就可以使其保持时间不变（或成为全局参数）。简单动态先验还允许选择块交互张量 p 是否随时间变化。此外，每个成员矩阵乃至其中的每一项 i，其 β 都不必相同。最后，β 本身也可以随时间变化。总而言之，β可以控制参数变化所涉及的时间尺度，从而联合建模全局参数（$\beta=0$）和动态参数（$\beta\neq0$）。

3.4.3 可调的时间依赖。 最后，选择平均核函数 $\kappa(t,t\prime)$ 也很重要。它允许选择推断某个变量时应依赖其时间邻居的范围。采用时间差倒数的形式似乎很合适：当时间 δt 越晚时，邻居的权重应按 $\delta t\to0$ 发散，从而确保连续性。此外，还可以通过将权重函数调节为 $\kappa(t,t\prime)=\frac{N_{t'}}{|t-t'|^a}$where 来控制曲线随时间变化的平滑度

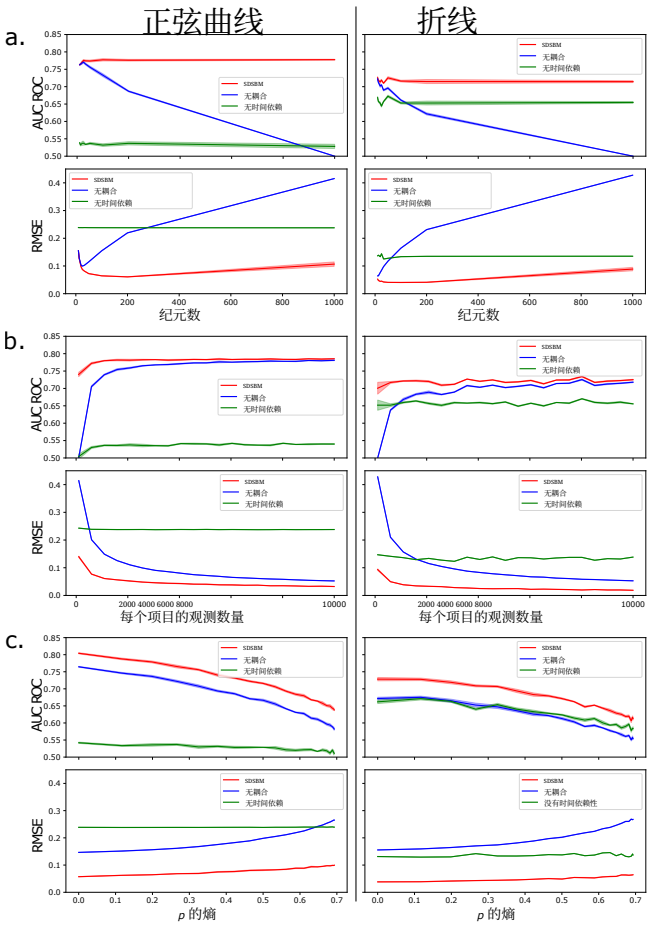


图 4：合成数据上的结果——(a.) SDSBM 找回了正确的动态成员关系，且几乎不受数据切片影响。(b.) SDSBM 在微型数据集上表现良好。(c.) SDSBM 在具有挑战性的情况下找回了动态成员关系。

$a = 1,2,\ldots$ 例如，其中 $N_{t'}$ 是时间切片 $t_{\prime}$中的观测值数量。总体而言，简单动态先验通过利用微观和中观时间尺度来推断变量。如果某个时间切片 t 的观测值较少，但其部分邻居拥有更多观测值，那么在 t 学习参数时，主要受其最近（$\frac{1}{|\Delta t|}$）且观测值最多（$N_{t'}$）的邻居群体帮助，受更远且观测值较少的时期影响较小。这正是图 2 所示的内容。

4 实验

4.1 合成数据

在本节中，我们介绍几种可能适用我们方法（简称 SDSBM，即简单动态 MMSBM）的情形[1]。实验针对 $I = 100$、$K = 3$ 和 $O = 3$进行，这些是动态网络文献中的标准测试参数。

[1]代码和数据可在 https://anonymous.4open.science/r/SDSBM-26DC 获取。

推断 [7, 16]。我们选择推断动态成员矩阵$\theta(t)$，并提供一个通用的块交互矩阵 $p\ \forall t$。请注意，当 p 也需要推断时，该模型依然表现良好；但由于 [16], 中提出的可识别性和标签切换问题，无法无偏地评估推断出的隶属关系值是否正确。额外实验表明，联合推断 p 和 θ 可获得相同的预测准确率。我们用于该矩阵 p 的表达式见式 13。我们系统地测试了 θ的两种变化模式：正弦模式（图 1 上）和折线模式（图 1 下）。每种模式针对不同项目采用不同的系数生成；隶属关系在所有时刻的总和仍为 1。

$$p = \begin{bmatrix} 1-s & s & 0 \\ 0 & 1-s & s \\ s & 0 & 1-s \end{bmatrix} \tag{13}$$

据我们所知，唯一一次尝试使用 MMSBM对带标签的加权动态网络中的动态参数进行建模的工作是 [22]。在该工作中，每个时期都独立于其他时期进行建模。我们将这一基线称为“无耦合”或“**NC**”基线。作为参考，我们还比较了一个不考虑时间维度、而是为每个变量推断单一通用值的基线（**SIMSBM(1)**）[20]。已有研究评估表明，在 [9] 和 [20]中，MMSBM 通常优于当前最先进的模型（NMF、TF、KNN 等）。由于不属于本文演示的范围，本文不再重复此类研究。我们重申，为了清晰起见，**我们特意选择了最简单形式的MMSBM**，因此报告的结果在绝对意义上可能并不理想，但凸显了我们的简单动态先验方法相较于同样简单的模型（NC 和 SIMSBM(1)）所带来的改进。

我们系统地进行了 5 折交叉验证。模型使用 80% 的数据进行训练，β使用 10% 的数据作为验证集进行调优，并使用剩余的 10% 数据评估模型。我们选择 ROC 曲线下面积和 θ 真实值（图 1 中的黑线）的均方根误差作为评估指标。该过程重复 5 次；实验结果中报告的误差棒表示这几折交叉验证的标准误差。

4.1.1 SDSBM 揭示复杂的时间模式。 在图 4a 中，我们针对每个项目 $i \in I$考察 1 000 个观测值，并将时期数从 10 改变到 1 000。在 p的表达式中，s设为 0.05。对于正弦型和折线型隶属关系，模型的预测表现（以 ROC 曲线下面积衡量）均优于所提出的基线方法。有趣的是，随着时期数增加，SDSBM 的表现仍保持稳定，不像 NC 基线方法那样；这意味着它缓解了 [22]中提出的时间建模所存在的偏差。相对于真实参数的均方根误差在所测试的整个时期数范围内都保持较低。随着时期数增加，均方根误差也随之增大，这是因为待估参数的数量也随之增加；这使推断更容易受到局部变化的影响，进而机械地增大均方根误差。总体而言，SDSBM 能以良好的表现恢复成员向量的动态变化；图 1 展示了推断得到的动态成员关系示例。

4.1.2 SDSBM 在数据稀少时仍能发挥作用。考虑时间数据时，一个主要问题是观测值稀缺，

因为切分会导致每个切片中的观测值数量减少。这一问题在社会科学中尤为突出，因为数据采集无法自动化，需要耗费大量繁琐的人力劳动。在此，我们展示了即使在数据稀少这一具有挑战性的条件下，我们的方法仍然有效。在图 4b 中，我们将每个项目可用的观测值数量从 100 变动到 10,000，并将其分布在 100 个时期内。因此，在最具挑战性的情况下，每个时期只有一个观测值用于确定 I 在 3 个簇上的动态成员关系。在表达式 p 中，s 设为 0.05。我们从图 4b 中看到，对于这两种模式，SDSBM 在这种条件下仍具有很强的预测能力。此外，在这种情况下，真实动态成员关系的均方根误差相当低，并且随着观测值数量增加而迅速下降。这是因为 SDSBM 在训练阶段将各个时间切片关联起来：平滑约束使每个时间切片都能间接受益于其相邻时间切片的训练数据。

当观测值数量较多时，“无耦合”基线 [22] 的表现可达到 SDSBM 的水平。这是因为随着每个切片中的观测值数量趋于无穷大，模型对时间邻居的依赖会降低。然而，即使每个项目有 10.000 个观测值（每个时期有 100 个观测值），SDSBM 仍能取得更好的结果。作为示例，图 1 中的结果仅使用每个时期 5 个观测值获得。

4.1.3 SDSBM 能处理高度随机的交互模式。最后，我们通过改变 s 来控制块交互矩阵 p 的确定性特征。我们将这一特征表示为 $\frac{1}{K}\sum_{k\in K}\sum_{o} p_{k\ o}\quad p_{k\ o}\qquad S(\ )$

相对于其可能输出的平均熵：$S(\ )(\ )\log(\ )$。对于所提出的 p 表达式，当 $s = 0.5$ 时，熵达到最大值。我们考虑分布在 100 个时期中的 1 000 个观测值。图 4c 展示了随着熵增加，所有三种方法的预测性能均有所下降。这是预料之中的，因为观测值是从方差更高的真实模型中生成的；随着 s 增大，每个观测值所包含的关于生成过程的信息都会减少。然而，使用 SDSBM 推断真实参数时的均方根误差在最大熵下仍然很低，这意味着该模型能够恢复正确的成员关系参数。

4.2 真实世界数据

4.2.1 实验设置。 最后，我们在真实世界数据上展示我们方法的有效性，以说明其效用和可扩展性。SDSBM 基于先前关于标记混合成员随机块模型的研究，并具有相同的线性复杂度 $O(|R_0|)$，其中 $|R_0|$ 表示数据集 [9]的规模。在实验中，我们使用了 [14]中的最新且有文档记录的数据集。**Reddit** 数据集（10.000 名用户、他们参与的 984 个子版块， ~670k 个观测值）、**LastFm** 数据集（980 名用户、用户听过的 1000 首歌曲， ~1.3M 个观测值）和维基百科（**Wiki**）数据集（8227 名用户、用户编辑过的 1000 个页面， ~157k 个观测值）。目标是预测用户随时间会参与哪个子版块、会听哪些歌曲，以及会编辑哪些维基百科页面。Reddit 和维基百科数据集包含 1 个月的数据；我们将其切分为每段 1 天的时间间隔。LastFm 数据集涵盖约 5 年；我们将其切分为每段 3 天的时间间隔。此外，我们还构建了一个额外的数据集（**Epi**），内容关于历史

表 1：真实世界数据集上的数值结果——指标缩写分别代表 ROC 曲线下面积（ROC）、平均精度（AP）和标准化覆盖误差（NCE）。模型缩写分别代表简单动态 SDM（SDSBM）、无耦合基线（NC）以及经典静态混合成员随机块模型（SIMSBM(1)）。总体而言，我们的方法具有更强的预测能力。

		ROC	AP	NCE
Epi	SDSBM	0。9025(11)	0.3700(17)	0。1151(11)
	NC	0.8420(22)	0.3435(36)	0.1582(19)
	SIMSBM(1)	0.8597(12)	0。2141（16）	0.1451(13)
Lastfm	SDSBM	0.8942(8)	0.0168(1)	0.1284(11)
	NC	0.8393(5)	0.0157(2)	0.1785(7)
	SIMSBM(1)	0.8647(5)	0.0115(2)	0.1493(4)
Wiki	SDSBM	0.9759(2)	0.0659(9)	0.0459(3)
	NC	0.9092(7)	0。0608(10)	0.1195(8)
	SIMSBM(1)	0.9576(7)	0。0622(4)	0。0565(8)
Reddit	SDSBM	0。9803（3）	0.4295(54)	0.0312(3)
	NC	0.8508(5)	0.3598(17)	0.1846(7)
	SIMSBM(1)	0.9798(2)	0.4269(40)	0.0322(3)

铭文学数据 [2]。该数据集由 117,000 条拉丁铭文组成，每条铭文包含 18 种社会地位中的一种或多种（奴隶、士兵、元老等），以及其所在地（共 62 个可能的地区），同时还包括从公元前100年到公元400年的估计年代。目标是根据时间推测某种社会地位出现的地区，并还原各类社会地位在罗马领土中的传播情况。我们将该数据集按每年一个时期进行切分。

评估仍采用 5 折交叉验证，每折使用 80% 的训练数据、10% 的验证数据和 10% 的测试数据。对于测试集中的每一对 $(i,\ o_{true})$，我们查询给定 i 时每个输出 o 的概率，并将其与 o_{true}进行比较，以构建混淆矩阵。在表 1 中，我们展示了所提方法与提出的各项基线方法在多种指标上的结果：ROC曲线下面积 **(ROC)**、平均精度 **(AP)** 和归一化覆盖误差 **(NCE)**。前两项指标评估模型为观测值分配概率的效果，最后一项则评估可能输出的排序顺序。

4.2.2 结果。 总体而言，我们的方法表现出更强的预测能力，但在 Reddit数据集 上，静态SBM的表现与SDSBM相当。我们将此归因于所考察时间段内缺乏显著的时间变化。这一结果并不意外，因为该数据集约有 80% 的行为是重复操作 [14], ，这意味着用户在一个月内并未显著探索新的社群。这表明SDSBM在静态情形下同样表现良好。在其他数据集上，SDSBM的表现通常更好，有时优势很大，尤其是在ROC-AUC 指标上，这意味着SDSBM能够有效地区分类别。我们还验证了关于参数随时间平滑变化的初始假设。对于每个真实世界数据集，相邻两个时间片之间成员关系变化的绝对平均值始终小于 0.005 ± 0.01。

[2]Clauss-Slaby存储库，http://www.manfredclauss.de/fr/index.html

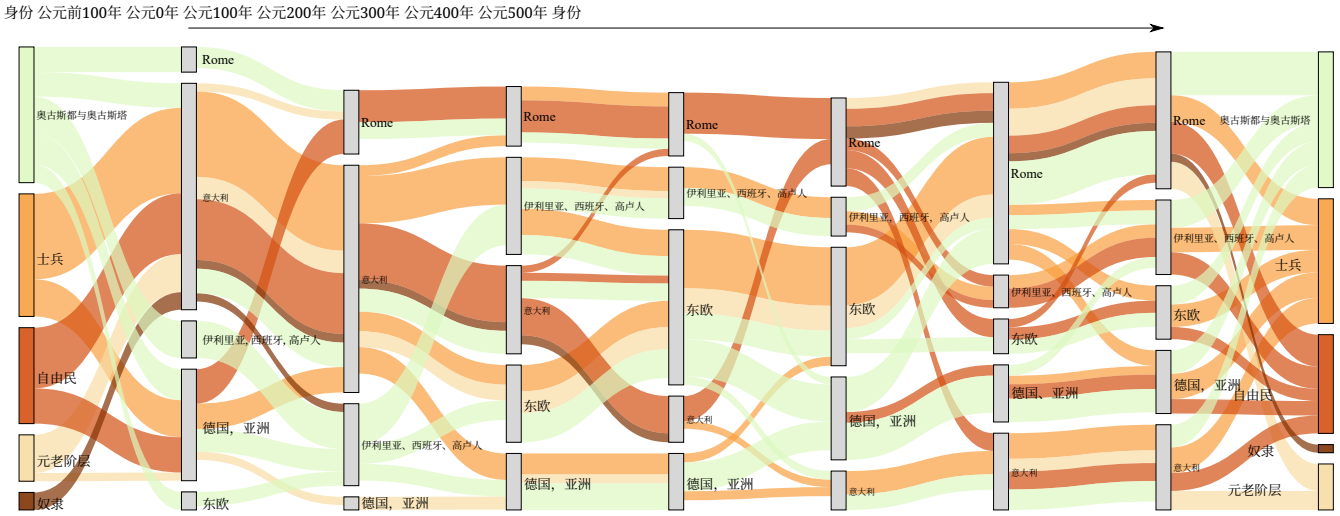


图5：拉丁墓葬中身份分布的地理演变（公元前100年—公元500年）——我们将 SDSBM 应用于 Epigraphy 数据集。回顾一下，我们的目标是根据身份（例如奴隶、元老等）和年份，预测一个罗马地区（例如伊利里亚、西班牙等）。我们绘制了身份归属五个人工标注的簇（灰色）的时间演变。为清晰起见，我们从数据中去除了少量成员转移，这解释了为何各簇的总人口可能随时间变化。该图有助于直观呈现罗马帝国演变的一些整体历史趋势（例如三世纪危机、军事存在扩展至欧洲、意大利非军事化等）。

经过 10 个时间片后的绝对变化小于 0.03 ± 0.04。因此，推断出的参数不会随时间突然变化。需要指出的是，此处使用的模型是为演示而有意简化的；指标较低并不意味着简单动态先验不起作用，而是说明它应与更复杂的模型（如任意 SIMSBM(x,y,...)）结合使用。

4.2.3 案例研究。 为了展示 SDSBM 所能提供的功能，我们在图 5 中绘制了铭文数据集的隶属关系随时间演变的一种可能可视化结果。在左侧和右侧，我们展示了可视化中所考虑的项目。时间从左侧（公元前100年）延伸至右侧（公元500年），流线表示各时期之间的隶属关系转移。灰色条带表示簇。我们通过观察其构成对这些簇进行了人工标注。从该图中，我们可以还原出若干历史事实：在所考察的大多数时期，罗马的军事存在都较为稀少；随着时间推移，意大利的军事存在逐渐减少（由于军事力量分散到如今疆域更为广阔的帝国各处），直到三世纪危机导致其重新军事化；在整个时期内，大多数获得铭文记录的奴隶都位于意大利；宗教职务（Augusti）始终均匀分布在各地；自由民（获释奴隶）的铭文基本上集中出现在罗马和意大利等地。显然，还需要开展专门研究来支持这些说明性论断，而我们相信，SDSBM 能够帮助人们更全面地理解这些过程。

5 结论

我们提出了一种简单的方法，通过假设动态混合成员随机块模型，对动态加权标记网络中的时间进行建模。我们的方法包括定义简单时间先验，该先验可直接应用于一整类现有的静态混合成员随机块模型。对时间的建模只基于一个假设：网络中的连接不会突然变化。

我们通过定义 SDSBM，并在合成数据集上的若干受控情形中进行测试，评估了所提方法的性能。具体而言，我们表明，我们的先验能够在数据集切片方面保持稳定的性能，并且在具有挑战性的条件下（数据量较少或块交互矩阵的熵较高）也能良好运行。我们在大规模真实世界数据集上测试了 SDSBM，结果表明其表现优于两种所提出的基线方法。最后，我们展示了该方法在拉丁铭文数据集上的应用价值；这些铭文间接叙述了罗马帝国的演变。

我们讨论了采用该方法的优势：对观测值进行不均匀的时间切片、异质动态时间尺度、时变块交互矩阵，以及先验的信息量；在真实世界数据上探索这些方向，或有助于在实际应用中发现有意义的簇。此外，我们认为该方法的一项关键优势是只需少量数据便能获得令人满意的性能。这一点对许多社会科学领域至关重要，我们相信该方法能够促进这些领域采用自动化学习方法。

参考文献

[1] Amr Ahmed 和 Eric P. Xing。2007。寻求真正相关的主题后验——关于逻辑斯蒂-正态混合模型的紧致近似推断。第 *11* 届人工智能与统计学会议论文集 2（2007），19–26。https://proceedings.mlr.press/v2/ ahmed07a.html[2] Edoardo M. Airoldi、David M. Blei、Stephen E. Fienberg 和 Eric P. Xing。2008。混合成员随机块模型。《机器学习研究杂志》9（2008 年 6 月），1981–2014。https://dl.acm.org/doi/10.5555/1390681.1442798

[3] Richard Arratia、A. D. Barbour 和 Simon Tavaré。1992。Ewens抽样公式的泊松过程近似。《应用概率年刊》2, 3（1992），519–535。[4] Christopher Bishop。2006。《模式识别与机器学习》。施普林格。450–455 页。[5] S. Cobo-López、A. Godoy-Lorite 和 J. Duch。2018。使用块模型预测社会困境中的决策并进行模型选择。*EPJ Data Sci* 7 48 2018 10 1140 13688 018 0175 3 ( ) ( )。https://doi.org/ 。/ [6] epjds/s - - - Feng-Lei Fan、Jinjun Xiong、Mengzhou Li 和 Ge Wang。2021。人工神经网络的可解释性：综述。*IEEE*辐射与等离子体医学科学汇刊 5, 6 2021 741 760 10 1109 2021 3066428 ( )，–。[7] https://doi.org/ 。/TRPMS. 。Xuhui Fan、Longbing Cao 和 Richard Yi Da Xu。2015。动态无限混合成员随机块模型。*IEEE*神经网络与学习系统汇刊26, 9 2015 2072 2085 10 1109 2014 2369374（），–。https://doi.org/ 。/TNNLS. 。[8] 2019 Thorben Funke 和 Till Becker。。随机块模型：变体与推断方法比较。*PLOS ONE*，(2019），1–40。https://doi.org/10.1371/ journal.pone.0215296[9] Antonia Godoy-Lorite、Roger Guimerà、Cristopher Moore 和 Marta Sales-Pardo。2016。使用混合成员随机块模型进行准确且可扩展的社会推荐。美国国家科学院院刊113, 50（2016），14207–14212。https: //doi.org/ 10 1073 1606316113 。/pnas. Roger Guimerà 和 Marta Sales-Pardo。2013。用于大规模无监督识别新型药物相互作用的网络推断方法。*PLOS*计算生物学 9, 12 12 2013 1 9 10 1371 1003374 ( )，–。[10] https://doi.org/ 。/journal.pcbi. Qirong Ho、Le Song 和 Eric P. Xing。2011。面向时变网络的演化簇混合成员块模型。*AISTATS*（2011）。https://proceedings. mlr.press/v15/ho11b.html [11] Paul W. Holland、Kathryn Blackmond Laskey 和 Samuel Leinhardt。1983。随机块模型：初步探索。《社会网络》5, 2（1983），109–137。https: //doi.org/ 10 1016 0378 8733 83 90021 7 。/ - ( ) - D. Jin、Z. Yu、P. Jiao、S. Pan、D. He、J. Wu、P. Yu 和 W. Zhang。2021。社区检测方法综述：从统计建模到深度学习。*IEEE*知识与数据工程汇刊01（2021），1–1。https://doi.org/10.1109/ TKDE.2021.3104155[13] Srijan Kumar、Xikun Zhang 和 Jure Leskovec。2019。预测时间交互网络中的动态嵌入轨迹。载于第 *25* 届 *ACM SIGKDD* 会议论文集（美国，阿拉斯加州，安克雷奇）*(KDD '19)*。计算机协会，纽约，美国，1269 1278 10 1145 3292500 3330895 – 。https://doi.org/ 。/ 。[14] Clement Lee 和 Darren J. Wilkinson。2019。用于图聚类的随机块模型及其扩展综述。《应用网络科学》4（2019），1–50。https://doi.org/ 10 1007 41109 019 0232 2 。/ s - - - Catherine Matias 和 Vincent Miele。2017。通过动态随机块模型对时间网络进行统计聚类。英国皇家统计学会杂志：*B*辑（统计方法） 79, 4 2017 1119 1141 10 1111 12200 ( )，–。[16] https://doi.org/ 。/rssb. C Matias、T Rebafka 和 F Villers。2018。纵向网络随机块模型的半参数扩展。《生物统计学》105, 3（2018 年 6 月），665 680 10 1093 016 – 。https://doi.org/ 。/biomet/asy Gaël Poux-Médard、Sergio Cobo-Lopez、Jordi Duch、Roger Guimerà 和 Marta Sales-Pardo。2021。将股市实验中的复杂决策策略解释为少数简单策略的组合。*EPJ*数据科学10（2021），1–14。第26 期。https://doi.org/10.1140/epjds/s13688-021-00280-z [19] Gaël Poux-Médard、JulienVelcin 和Sabine Loudcher。2021 年。结果预测中的信息交互：使用块模型进行量化与解释。第 *15*届 *ACM* 推荐系统会议（*RecSys*）论文集 （2021），199–208。https://doi.org/ 10 1145 3460231 3474254 。/ 。Gaël Poux-Médard、JulienVelcin[20]和 Sabine Loudcher。2022 年。序列化交互混合成员动态随机块模型。*2022 IEEE* 国际数据挖掘会议（*ICDM*） （2022）。https://doi.org/10.1109/ICDM54844。2022.00145[21] Xuning Tang和 ChristopherC. Yang。2014年。使用带时间狄利克雷过程的动态随机块模型检测社交媒体中的隐蔽社群。*ACM* 智能系统与技术汇刊 5，2（2014），21。https://doi.org/10.1145/ 2517085[22] Marc Tarrés-Deulofeu、Antonia Godoy-Lorite、RogerGuimerà 和Marta Sales-Pardo。2019年。用于分层网络和时间网络中链接预测的张量块模型与二部块模型。物理评论*E* 99（2019），032307。第 3 期。https: //doi.org/10.1103/PhysRevE.99.032307[23] Quentin Villermet、Jérémie Poiroux、Manuel Moussallam、Thomas Louail 和 CamilleRoth。2021 年。跟随指南：解构在线音乐消费中的人为策展与算法策展。计算机协会，美国纽约州纽约市， 380 389 10 1145 3460231 3474269 –。[24] https://doi.org/ 。/ 。Yuchung J. Wang 和 George Y. Wong。1987。有向图随机块模型。美国统计协会杂志 82，397（1987），8–19。https://doi.org/10.1080/ 01621459.1987.10478385[25] James D. Wilson、NathanielT. Stevens和 William H. Woodall。2019 年。通过度校正随机块模型对时间网络中的变化进行建模与检测。质量与可靠性工程国际期刊 35（2019），1363–1378。https://doi.org/10.1002/qre。[26] 2520 Xunxun Wu、Pengfei Jiao、Yaping Wang、TianpengLi、Wenjun Wang 和 Bo Wang。2019 年。具有无标度特性的时间复杂网络动态随机块模型。第 *24*届 *DASFAA* 论文集 （2019），502–518。https: //doi.org/ 10 1007 978 3 030 18579 4_30 。/ - - - - EricP.Xing、Wenjie Fu 和LeSong。2010 年。用于动态网络层析成像的状态空间混合成员块模型。应用统计年鉴 4（ 2010 535 566 10 1214 09 311），–。[27] https://doi.org/。/ -AOAS Kevin S.Xu和 Alfred O. Hero。2013。动态随机块模型：时变网络的统计模型。社会计算、行为—文化建模与预测 （ 2013 201 210 10 1109 2014 2310294 ），–。https://doi.org/ 。/ [28] JSTSP。。KevinS. Xu 和Alfred O. Hero。2014。时变社交网络的动态随机块模型。*IEEE*信号处理精选主题期刊 8（2014），552 562 10 1007 978 3 642 37210 0_22 – 。[29] https://doi.org/。/ - - - - Tianbao Yang、Yun Chi、Shenghuo Zhu、Yihong Gong和 RongJin。2010 年。检测动态社交网络中的社群及其演化——一种贝叶斯方法。机器学习 82（2010），157–189。https://doi.org/10.1007/s10994- 010-5214-7 [30]

收稿日期：2007年2月20日；修订日期：2009年3月12日；接受日期：2009年6月5日