arXiv:2301.04768v1 [physics.comp-ph] 12 Jan 2023

精确估计稀疏图模型的熵

Edward D.Lee*Complexity Science Hub Vienna*,
Josefstædter Strasse 39, Vienna, Austria

我们利用社会群体之间的桥，以及适用于稀疏样本的精确熵估计器，提出了一种估计成对概率图模型熵的算法。我们提议使用社会学文献中的一种投资度量——伯特结构洞约束——作为识别桥的启发式方法，从而将图划分为条件独立分量。我们将这一启发式方法与 Nemenman-Shafee-Bialek 熵估计器相结合，以获得更快速、更准确的估计器。我们在司法投票的成对最大熵模型（即伊辛模型）上展示了该方法，以改进朴素的熵估计。我们使用该算法近似估计配分函数，随后将其应用于模型选择问题，因为在该问题中估计似然十分困难。这相较于现有方法有所改进；现有方法依赖点关联函数来检验拟合效果，而通过对开源实现进行直接修改，该方法还可以扩展到其他图模型。

估计概率模型的熵是集体行为信息论分析中一项常见且必不可少、但又十分困难的任务。例如，广泛使用的神经元放电 [1–5], 政治投票 [6], 社会群体 [7, 8], 或集体运动 [9, 10] 的最大熵模型，都需要估计模型的熵，以评估其与多信息的拟合程度。此外，熵相对于独立模型的相对下降幅度，为集体性提供了一种通用度量，即衡量其偏离独立统计的程度；这有助于了解模型假设所施加的约束。熵与自由能密切相关，自由能描述能量约束与无序之间的平衡，因此可用于评估拟合优度。最后，自由能是配分函数的对数，而配分函数的导数能够揭示磁化强度、易感性以及距临界点的距离等物理性质 [11, 12]。

即使已知模型，熵估计这一普遍问题仍然很难解决。统计物理学方法提供了一种途径，包括矩展开（如贝特自由能近似） [13, 14] 和簇展开 [15] ，以及热力学积分 [16]。另一种方法是根据分布的样本推断熵，而使用蒙特卡罗马尔可夫链方法生成样本通常相对快速。然而，大多数有趣的系统规模都足够大，以至于其状态空间远超模拟所能产生的可能（或合理）样本数量；这一问题在离散结果空间中尤其难以处理。遗憾的是，我们很清楚，在样本不足的情况下，熵估计会产生偏置。更糟的是，在概率分布空间中从先验出发，以推断出良好的熵估计，可能适得其反，在数据量较少时引入其他无法消除的偏置。一种精度极高的估计器以作者 Nemenman、Shafee 和 Bialek 的名字命名为 NSB [17], ；它尝试通过在熵空间中采用平坦先验，使估计尽可能无偏，因此在样本极度不足时也能表现良好的情况下也能表现良好。然而，与其他估计器一样，它在大型系统中仍会遇到问题，因为先验不可避免地会占据主导地位，而且（就实际情况而言）计算速度较慢，在数值计算方面也颇具挑战。

在这里，我们关注这样一个问题：已知模型时，如何估计其熵，并利用文献中适用于小样本的熵估计器（例如 NSB），比如在刻画自然现象的统计特征时 [18]。简而言之，我们解决的是概率模型熵值数值估计中的一个技术问题，并利用许多问题具有稀疏结构这一事实。在物理模型中，这意味着各个组成部分可以划分为彼此连接稀疏的社群。许多政治互动图便是如此，例如司法法院或立法机关中的互动图；随着时间推移，绝大多数成员从未彼此见面或互动 [19]。一些社交图也是如此，它们往往分解为若干团，在小世界网络中，这些团之间只有间接连接 [20–22], 。但对于某些小型动物群体而言，情况并非如此；它们呈现格状、拓扑局部的相互作用，例如鸟群以及用于描述群集行为的局部相互作用模型 [23]。如下文所述，我们开发了一种启发式方法，利用现有估计器，并将其封装为 Python 库 TreeEnt，以便在离散状态空间中更快、更准确地估计稀疏连接图模型的熵。

熵及其估计概述

20世纪中期，克劳德·香农试图找到一种方法，用来描述通过 AT&T 电信网络传输的信号的统计结构 [24]。他将消息视为由离散字符组成的序列s ，其中每个字符出现的概率为 $p(s)$，由此确立了“信息熵”，用来衡量接收序列中每个新字符所带来的意外程度。

他还证明，除单位选择外，这是一种满足三个公理 [25] 的唯一度量：相对于 p的连续性、在最大不确定性下取最大值，以及将符号分层分解为集合时保持一致。形式上，离散状态空间 $\mathcal{S}$ 中构型 s 的概率分布 $p(s)$的信息熵 H 为 [26]

$$H[p] \equiv -\sum_{s\in\mathcal{S}} p(s)\log p(s). \tag{1}$$

当所有构型上的分布 s 均匀时，熵达到最大值（因此完全不可预测）；当只有一种构型以概率 1 出现时，熵为零（完全可预测）。因此，熵为分布中结构的多少提供了一种独特度量，并对其统计 可预测性设定了限度。

给定来自$p(s)$的随机样本来估计熵十分困难，因为即使对于规模适中的系统，要对状态空间进行全面采样，计算成本也很高，甚至不可能。例如，如果我们尝试用最¨朴素的方式估计熵，也就是依靠从大小为 K 的随机有限样本中观察到状态 s 的次数 k_s，我们会认为 $\hat{}p(s) = k_s/K$。随后，与真实值的偏差可表示为误差项 $p(s) = \hat{}p(s) + \epsilon(s)$。将熵按误差展开后，我们得到

$$\hat{H}[p] = H[p] - \frac{A}{K} - \mathcal{O}\left[\frac{1}{K^2}\right], \tag{2}$$

其中，我们将所有阶为 $(1/K)^2$ 及更高阶的项归入最后一项，且 A 是一个正数。因此，朴素估计器会给出一个有偏估计，低估该分布的真实熵。

缓解这一问题的一种方法可能是采用贝叶斯方法：将熵估计定义为一项推断任务，并以样本所源自的概率分布构成的先验分布为基础。严格来说，一种经典做法是写出一个狄利克雷族模型，其中指定每个唯一状态在可能的 S 个状态中出现的概率 q_i（例如，对于大小为 N 的二元自旋系统，我们有 $S = 2^N$）。考虑到概率集合 $\{q_\mathrm{i}\}$ 的归一化条件，我们有

$$P_\beta(\{q_\mathrm{i}\}) = \frac{1}{Z(\beta)}\delta\left(1-\sum_{\mathrm{i}=1}^{S} q_\mathrm{i}\right)\prod_{\mathrm{i}=1}^{S} q_\mathrm{i}^{\beta-1}. \tag{3}$$

这称为狄利克雷先验族，其中 β 决定了我们先验地认为较大概率值有多大可能；拉普拉斯计数法则对应于$\beta = 1$，而最大似然估计量为 $\beta = 0$。因子 Z 确保这是一个归一化的概率分布。给定公式 3 中的先验，关于

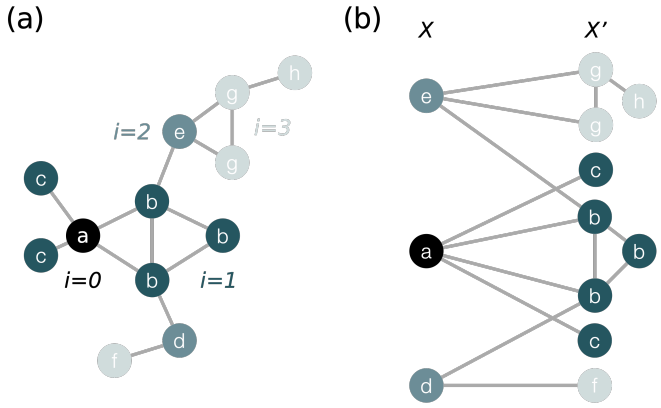

图 1. 图划分示例。(a) 划分前的图。边表示节点之间的相互作用。我们按照递增的索引 i 为树中远离根节点的每一代编号，根节点位于 $\mathrm{i} = 0$。不同代也以不同阴影区分。当每个分支都用字母表编号时，我们用代数和分支编号来指定节点集合，例如 s^a_i。(b) 通过我们的划分算法得到的二分图。固定上游节点的取值后，下游节点（例如，相对于 a 的 b）便条件独立。这使我们能够指数级缩减采样状态空间，因为最大的子图决定了要采样的最大状态空间。通过将最大簇大小设为 $K = 4$ 对图进行划分。

估计熵 H 的分布将由贝叶斯定理给出，

$$P_\beta(H) = \frac{P_\beta(H|\{q_\mathrm{i}\})P_\beta(\{q_\mathrm{i}\})}{P_\beta(\{q_\mathrm{i}\}|H)}. \tag{4}$$

那么，给定 β 时熵的平均估计为

$$\xi(\beta) \equiv \int_0^\infty P_\beta(H)H\,dH. \tag{5}$$

至关重要的是，β 的选择对确定熵的最终估计起着关键作用，因为先验分布的峰值很高，因此在低数据样本量极限 [17]下，熵基本上由先验决定。

相反，我们可以考虑一种元先验，它涵盖该族中所有可能先验的混合，或者以某种方式涵盖 β 的可能范围，从而弱化我们对熵的初始假设。这意味着，我们应在某个加权范围 $P(\beta)$ 内移动 β，使其对应于在 0 与 $\log H$ 之间移动 $\xi(\beta)$。于是，参考文献17提出的一种候选混合先验为

$$P(\{q_\mathrm{i}\};\beta) = \frac{1}{Z(\beta)}\delta\left(1-\sum_{\mathrm{i}=1}^{S} q_\mathrm{i}\right)\prod_{\mathrm{i}=1}^{S} q_\mathrm{i}^{\beta-1}\frac{d\xi(\beta)}{d\beta}P(\beta). \tag{6}$$

要计算这个积分，还必须了解估计熵如何随 ξ 变化，以便在估计范围内将先验展平；两者之间的关系见参考文献17。事实证明，公式6对于极小样本的熵估计效果出奇地好。

尽管如此，它仍无法克服一个根本问题：在任何欠采样系统中，熵估计都由先验主导。

解决这一难题的一种方法，是利用模型概率分布的结构来简化熵估计问题，实际上就是缩小估计器需要处理的状态空间。幸运的是，香农最后一条公理告诉我们，如果能将状态划分为若干独立子集，那么信息熵就等于簇标签的不确定性与给定标签后各子集内部不确定性的贡献之和。如果能巧妙地进行分解，我们就可以分别考察系统的各个子集。例如，若固定分量 k 后，s（其中索引为 i）的某些分量与索引为 j 的分量条件独立，就可以做到这一点。为简便起见，我们将集合 $\{s_\mathrm{i}\}$ 记作 s_i。换言之，这意味着我们可以将概率分布分解为

$$p(s) = p(s_\mathrm{i}|s_\mathrm{k})p(s_\mathrm{j}|s_\mathrm{k})p(s_\mathrm{k}). \tag{7}$$

若情况如此，则该集合的熵可分解为信息熵之和

$$H[p] = H[p(s_\mathrm{k})] - \sum_{s_\mathrm{k}} p(s_\mathrm{k}) \left(H[p(s_\mathrm{i}|s_\mathrm{k})] + H[p(s_\mathrm{j}|s_\mathrm{k})]\right). \tag{8}$$

我们可以将 s_k 的每种独特构型视为一个“标签”，给定标签后，我们必须计算集合 s_i 和 s_j 的不确定性，其相应权重由它们所基于的 s_k 的频率决定。

将这个基本例子推广到树时，我们再次用 s_i 表示一个集合，但现在用索引 $\mathrm{i}-1$ 表示沿树向上移动一代，用 $\mathrm{i}+1$ 表示向下移动一代。给每个分支标上一个不同的字母表字母作为上标，

$$p(s) = p(s_0) \prod_{i=1}^{N} \prod_{a,a'} p\left(s_\mathrm{i}^{a'} \middle| s_{\mathrm{i}-1}^{a}\right), \tag{9}$$

这里我们再次使用简写记法 s_i 来指代索引 i 对应的连通分量集合。于是，根节点是节点集合 s_0；第一个乘积遍历树中从索引 i 所指的层开始，直至第 N 代叶节点的各个连续层；第二个乘积则遍历第 $(\mathrm{i}-1)$ 层中每个分支 a 的后代节点 a'，这些节点在给定所有父分支的条件下相互独立。作为使用此记法进行因式分解的示例，我们在图 1 中展示了一幅图，其中标出了距根节点 i 代的树的连续各代；在给定父分支 $\mathrm{i}-1$ 的条件下，这些代彼此条件独立。

当概率分布可以按这种方式进行因式分解时，就可以由外向内计算整个系统的熵：计算每个叶节点的熵，而每个叶节点的熵都会对最终熵产生加和贡献。我们将其表示为递归形式

$$H[p] = H[p(s_0)] + \sum_{s_0} \sum_{a} p(s_0) \Big(H[p(s_1^a|s_0)] + \sum_{s_1^a} \sum_{a'} p(s_1^a|s_0) \Big(H[p(s_2^{a'}|s_1^a)] + \cdots + \sum_{s_i^{a(n-1)}} \sum_{a^{(n)}} p\left(s_i^{a^{(n)}}|s_{i-1}^{a^{(n-1)}}\right) \left(H\left[p\left(s_{i+1}^{a^{(n+1)}}|s_i^{a^{(n)}}\right)\right] + \cdots \right) \Big) \Big). \tag{10}$$

简而言之，我们可以从叶节点向上到根节点，实现式（10）中的计算。我们从叶节点的条件熵开始，对该叶节点进行剪枝，此时通向该叶节点的分支随之变成叶节点。以这种递归方式继续下去，最终便会到达图的根节点。

这种层次分解提供了一种缓解熵估计问题的方法：如果我们能够将连通分量分组成若干子集，且这些子集远小于整个图，那么状态空间就会呈指数级缩小 [27]。如果这些组确实远小于整个图，那么从原理上说，熵估计将更易于处理所需的蒙特卡洛马尔可夫链（MCMC）样本也会呈指数级减少。我们提出了一种算法，用于估计概率图模型的熵。该模型通过一种基于结构洞的启发式方法对图进行因子分解，并对条件概率分布进行快速 MCMC 采样。

算法

我们在表 I 中概述了算法，并在下文提供更多细节。

第一步是对图进行因式分解

从而将问题拆分为更易处理的熵估计问题。为此，我们借鉴社会学文献中的结构洞概念 [20, 28]。结构洞的概念源自社交网络中受约束行动的问题：其假设是，任何个体 u 受到邻居 v 的约束程度，直接取决于两个因素： u 在与 v 的关系上投入的相对资源，以及 w 的另一位邻居 u 投入到 v上的相对资源。后一个步骤背后的直觉是，当投入相互重叠时，u 在发挥影响力方面受到最大程度的制约，因为u 对 v的局部邻域产生了重复影响（它们在同一范围内争夺影响力）；而重叠程度较低时， v 较少构成约束，更多地充当通往网络其他部分的桥梁，而 u 无法触及这些部分。换言之，相对于 u 而言，结构约束较低的邻居v 周围存在“结构洞”，例如它属于另一个团。

就我们的目的而言，直观地说，低约束的节点位于连接稠密的社群之间；它们应构成一个较小的集合，我们可以固定这些节点，从而使相邻社群彼此独立。具体来说，这些节点周围存在空隙，因此其结构约束较小 c_u，结构约束定义为节点 u 的局部约束之和 l ，

$$c_u := \sum_{v \in \mathcal{N}(u)} l(u,v), \tag{11}$$

也就是对 u 的所有邻居 v 求和，或 $\mathcal{N}(u)$。边权重均匀时，局部约束定义为

$$l(u,v) := \left(\frac{1}{|\mathcal{N}(u)|} + \sum_{w \in \mathcal{N}(u) \cap \mathcal{N}(v)} \frac{1}{|\mathcal{N}(u)|} \frac{1}{|\mathcal{N}(w)|} \right)^2, \tag{12}$$

其中，节点 u 的邻居数量为 $|\mathcal{N}(u)|$。括号中的第一项表示 u 对这个邻居 v 的投资，这一投资在所有邻居之间均匀分配。第二项表示 u 和一个共同邻居 w 对邻居 v 的联合投资。请注意，公式 12 是针对我们这一情形的简化形式；节点对之间的“投资”，即边权重，通常不必均匀 [29]。

我们从图中移除节点，首先移除约束最小的节点，并在每次移除后重新计算局部约束。每个被移除的节点都放入集合 X' 中。这往往会沿着连接这些连通分量的桥，将剩余节点集合 X 中的连通分量分割开来。如果我们无限期地继续这一过程，那么位于已移除节点集合 X' 中的子图最终将会连通

表 I. 熵估计算法。如果起始图由多个连通分量组成，则分别处理它们。

1. 将图因式分解，生成“收缩”树。
   a. 将所有节点放入 X。 b. 计算 X 中每个节点的结构约束。 c. 从集合 X 中移除约束最小的节点，并将其放入 X'。 d. 如果 X 或 X' 中的最大连通分量现在大于步骤 c 之前的最大连通分量，或者 X 中的最大连通分量小于或等于阈值 M，则跳转到步骤 f。 e. 返回步骤 b。 f. 创建该图的粗粒化表示：如果两个连通分量之间至少存在一次交互，则将 X 和 X' 中的连通分量彼此连接。

2. 估计条件熵。
   a. 在粗粒化图中识别叶节点。 b. 对每个叶节点，生成一个大小为 K 的蒙特卡罗样本。 c. 对于大小为 K 的样本中的每个唯一状态，从该叶节点生成一个大小为 K' 的蒙特卡罗样本。 d. 使用 NSB 估计该叶节点和条件分支的熵。 e. 移除所有叶节点，对粗粒化图进行剪枝。 f. 返回第一步。

3. 汇总熵并计算误差。
   a. 将每个叶节点及其分支的估计熵相加。

它们构成了图的大部分，而且可能形成大型连通分量，完全没有简化问题。因此，只要被移除节点集合中的最大连通分量以及剪枝后的图仍在缩小，我们就会继续移除约束最少的节点；或者直到它们的规模都降至小于指定阈值M。

经过这一过程，我们得到两组节点X 和 X'。每组中都有互不连通的连通分量，分别标记为 x 和 x'；这些连通分量连接着另一组中的节点，但彼此之间并不直接相连，形成了我们在图1b中展示的二分结构。因此，该图呈现出这样一种结构：固定 x中的连通分量后， x' 中的连通分量便彼此独立，反之亦然。

最后一步，我们计算所得因子化树的熵。具体而言，我们识别出叶节点，或在保持互补集合中的单个连通分量固定时，在 X 或 X' 中条件独立的连通分量。

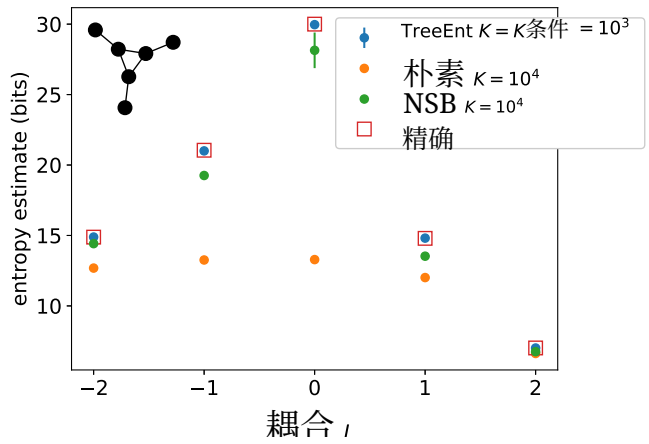

图 2. 在“尖角”三角形的五个副本上，对比 TreeEnt 与朴素和 NSB 熵估计器（左上角插图）。通过识别模型的相互作用结构，我们的算法以更快的速度、更少的样本取得了更好的效果。

虽然无法保证在结构约束移除过程中最后一定能找到叶节点，但移除约束最少的节点往往会将图分解为一条由条件独立部分组成的链。此外，始终存在一种图的划分方式，使得可以指定一个叶节点（在平凡情况下，可将单个节点放入 X'）。

最后，我们从条件独立的“叶节点”集合以及通向这些叶节点的唯一节点集合（即它们的“分支”）中进行采样。首先，我们从整个系统的样本中生成第 a 个分支 s^a_{i-1}分布的 MCMC 样本。然后，我们遍历唯一状态，并在分支的每种可能状态下，对第 a' 个叶节点 $s^{a'}$ i进行采样。样本集的大小也使我们能够估计采样熵的标准差。这便得到了公式 10 中的各项。随后，我们从图中剪去该叶节点，并递归地重复这一过程，直到只剩下树的根节点。每一步中，我们都使用 NSB 估计量计算熵。树的总熵为所计算熵的总和。

误差估计

对于该算法，我们必须考虑两个误差来源：条件集合的有限样本量和 NSB 估计量。

对于有限样本的贡献，式 10 中单个项（一个叶节点的熵）对熵的贡献是以下各项的组合

$$\left\langle H[p(s^{a''}_{i+1})]\right\rangle = \sum_{s^{a'}_i} p(s^{a'}_i) H[p(s^{a''}_{i+1}|s^{a'}_i)] \qquad (13)$$

其中，求和遍历分支 $s^{a'}_i$ 和特定叶节点 a'' 中出现的所有唯一状态。在此计算中，我们不考虑对分支 $s^{a'}_i$ 的所有父节点求和，因为我们可以直接从该子集中抽取马尔可夫链蒙特卡洛样本。这意味着我们也可以计算这个叶节点熵的方差，

$$\sigma^2_{s^{a''}_{i+1}} = \sum_{s^{a'}_i} p(s^{a'}_i)\Big(H[p(s^{a''}_{i+1}|s^{a'}_i)] - H[p(s^{a''}_{i+1}|s^{a'}_i)]\Big)^2. \qquad (14)$$

按照通常的做法，我们将方差除以样本数 K，以估计均值标准误。由于对公式 13 和 14 的求和中每一项的 NSB 误差 $\sigma_{\mathrm{NSB}}(s^{a'}_i)^2$ 都是独立于有限样本误差计算的，因此我们将两种误差的范数相加，以确定熵估计的总误差。对于单个叶节点，我们得到方差之和，即误差的范数平方

$$\Sigma^2_{s^{a''}_{i+1}} = \sigma^2_{s^{a''}_{i+1}} \Big/ K + \sum_{s^{a'}_i} p(s^{a'}_i)\sigma_{\mathrm{NSB}}(s^{a'}_i)^2. \qquad (15)$$

评估性能

我们将算法与朴素估计器和 NSB 估计量进行比较，所用示例中的熵可以通过枚举所有构型精确计算。我们发现，利用图的因子化结构，可以恢复出几乎完美的精确熵估计。作为示例，我们展示了由多个互不相连的图组成的系综的结果；每个图都由一个三角形以及从每个顶点伸出的一个额外节点组成（因此形似“尖头”三角形），总共有六个自旋。自旋以统一强度 J 耦合，该强度由我们改变。当我们同时从图的五个副本中进行采样时，在台式计算机上很难对$K = 2^{30} \sim 10^9$ 的状态空间进行充分采样。如图2所示，即使样本数量相对较少，我们的估计器在不同耦合尺度 J 下也表现良好。由于我们计算估计器所用的子集规模更小，它也明显快于一次性将 NSB 估计量应用于整个系综。

最大熵投票模型的应用

我们使用该算法估计美国巡回上诉法院法官投票的稀疏投票模型的熵。我们所考虑的概率模型源自最大熵原理，这是一种用于确定数据的最简统计模型的算法[30]。重要的是，在这一框架中，我们关注的量是模型的熵；在

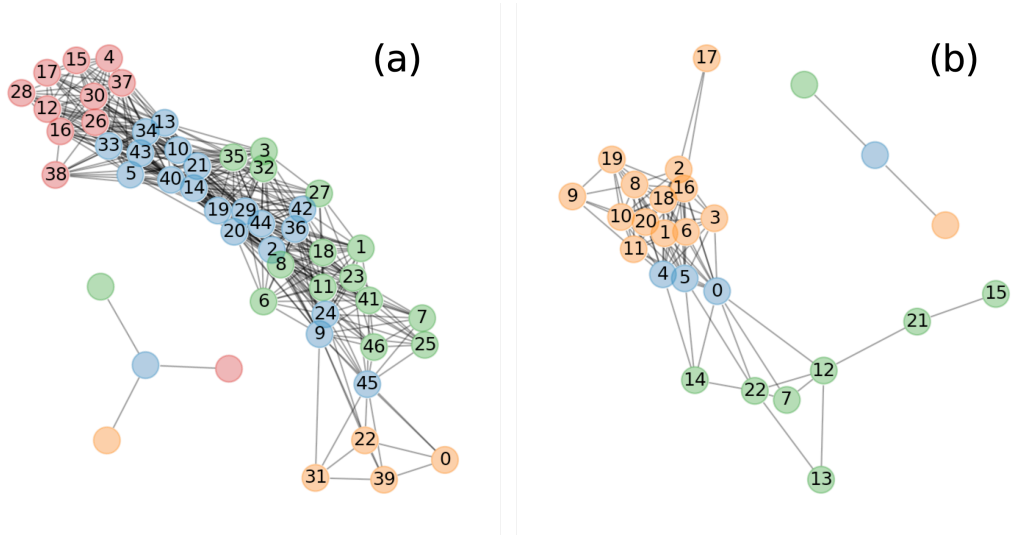

图 3. 将最小约束启发式应用于参考文献 中稀疏的美国上诉法院交互图所得图划分示例。(a) 当最大簇大小设为 $M = 14$ 时，哥伦比亚特区巡回上诉法院被分为四个连通分量。最大连通分量的大小为 $N = 19$，而整个图的大小为 $N = 47$。每种颜色代表一个不同的子图。插图展示了粗粒化图上的相互作用结构，该图是一棵树；也就是说，固定蓝色簇后，绿色、红色和橙色簇彼此独立。(b) 第一巡回上诉法院被划分为若干组，这些组构成一条直线。最大连通分量的大小为 $N = 13$，而整个图的大小为 $N = 23$。

对系统的少数几项关键属性（例如每位法官的平均投票及其成对相关性）施加约束的同时，模型的熵达到最大值 [31]。施加这两组约束后，我们得到成对最大熵模型；研究表明，该模型能够高精度地刻画美国最高法院的投票模式 [6, 19]。

根据该模型，概率分布采用玻尔兹曼形式

$$p(s) = \frac{1}{Z} e^{-E(s)} \tag{16}$$

其中归一化项 Z 在统计物理学中称为“划分函数”。对于每个构型 s（定义为由 -1 和 1 组成的向量，表示法官的投票），都会赋予一个“能量” $E(s)$；能量越低，该构型出现的可能性就越高。在成对最大熵模型中，它的形式为

$$E(s) = -\sum_{i=1}^{N} h_i s_i - \sum_{i<j}^{N} J_{ij} s_i s_j. \tag{17}$$

场 h_i 描述了每个连通分量的偏置 i：具有负偏置 $h_i < 0$ 的连通分量倾向于取值 -1，而具有正偏置 $h_i > 0$ 的连通分量倾向于取值 1。耦合 J_{ij}描述了各连通分量彼此趋于一致的倾向。与偏置类似，正耦合会在两个连通分量一致时降低能量，而负耦合会在两个连通分量不一致时降低能量。因此，成对最大熵模型同时捕捉了独立地偏向某一方向的倾向，或

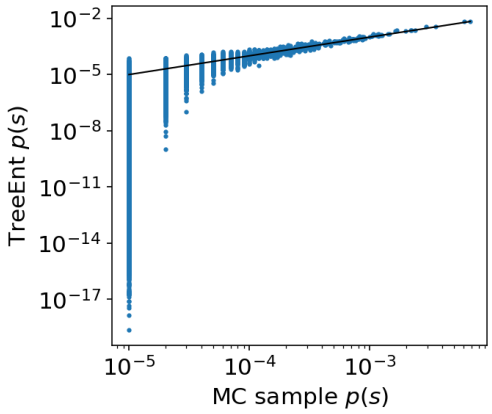

图 4. 对于哥伦比亚特区巡回上诉法院，使用我们的配分函数熵估计得到的状态 s 概率，与通过蒙特卡洛马尔可夫链抽样得到的概率。我们展示了在样本中至少出现两次的点。与公式 18 一样，我们还必须估计平均能量 $\langle E \rangle$；除非能量分布呈现重尾（例如接近临界点），否则使用 MCMC样本可以高精度地直接获得该估计——不过，这种情况同样会给熵估计带来问题。MCMC样本量为 $K = 10^6$。

另一连通分量，以及描述某个连通分量与其他所有连通分量之间如何相互作用的成对相互作用倾向。

重要的是，耦合描述了一个统计交互网络，类似于图 1 中显示的边。这一点可以从式 16 和 17 所描述的概率分布的因式分解中看出。显然，由于能量函数具有可加性，

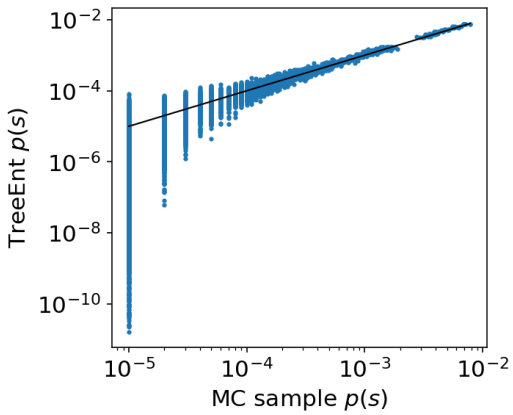

图5。使用我们对配分函数的熵估计得到的状态 s 的概率 p（s），与通过第一巡回上诉法院的 MCMC 采样得到的结果进行比较。更多详情请见图4。MCMC样本量为 $K = 10^5$。

对于每一对由 $J_{ij} \neq 0$ 关联的选民 i 和 j，乘积中都会出现一个单独的项 $p(s)$。因此，我们只需利用耦合矩阵的连通性，将图划分为条件独立的分量。

在美国上诉法院中，相应的成对最大熵模型是稀疏的，因为任何给定时间内在法院任职的法官人数都有上限。这意味着过去有许多法官与未来的法官没有互动，因此他们之间不存在耦合（更多详情请参阅参考文献19中的补充信息）。如图3a所示，我们的算法将哥伦比亚特区（DC）巡回上诉法院的相互作用图分解为四个部分：一棵简单树，其中给定蓝色簇后，绿色、红色和橙色集合条件独立。朴素方法需要在对应于 $N = 47$ 个选民的状态空间中进行估计，状态空间大小为 $2^{47} \sim 10^{15}$；而划分后最大的簇包含 $N = 19$ 个选民，对应约为 10^6，的状态空间，这在数值上是可行的，可以进行充分采样；我们预计，NSB 估计量只需远少于完整状态空间所需的样本量便能取得良好效果。图3b展示了第一巡回上诉法院的相互作用图，其中 $N = 23$；我们再次将其划分为三个簇，其中绿色和橙色的两个簇在固定中心蓝色簇后彼此条件独立。因此，TreeEnt 为估计这些投票模型的熵提供了一种更快速、更准确的方法。

对于拟合效果最佳的模型，我们的估计器在设定分支样本量为 $K = 10^4$ 、叶节点样本量为$K' = 10^3$后，估计哥伦比亚特区巡回上诉法院的熵为 $S_{\mathrm{DC}} = 17.84 \pm 0.05$ 比特，第一巡回上诉法院的熵为$S_1 = 11.34 \pm 0.03$比特。由于这些熵估计值相较于独立模型的熵有显著下降 $S_{\mathrm{DC}} = 47$比特和 $S_1 = 23$比特，这表明法官的投票行为存在强相关性。相比之下，1994—2005年美国最高法院的成对最大熵模型约为5比特，而可能的最大值为9比特。从每位投票者的熵来看，上诉法院受到的约束比美国最高法院更多；这是合理的，因为上诉法院通常有义务遵循后者确立的先例。

熵与配分函数密切相关，因此我们可以反过来从熵估计配分函数，这与通常的做法相反。从公式 16 的对数开始，

$$\begin{aligned} \log p(s) &= -E(s) - \log Z, \\ -\log Z &= \langle E \rangle - H[p], \end{aligned} \tag{18}$$

其中，在最后一步中，我们对等式两边关于分布 $p(s)$取了平均值。这便得到公式 18 中的亥姆霍兹自由能关系。因此，我们可以利用熵的估计值计算配分函数。配分函数是一个重要的量，除了决定归一化外，还直接参与模型物理性质的计算。如图 4 和图 5 所示，使用该配分函数估计值对概率进行归一化后，我们得到的状态 $p(s)$概率与通过 MCMC采样估计的概率高度吻合。

对于上诉法院的解空间，由于用于推断缺失投票的那组一致性条件 [19]，解空间存在退化。因此，我们需要一种方法来从恢复出的多个解中进行选择。每个解都具有相同的相互作用结构，这一结构由我们观察到的共同投票或未共同投票的法官配对决定，但场 h_i 或耦合 J_{ij} 的具体取值会有所不同。作为衡量拟合优度的标准信息论度量，我们采用数据分布与模型之间的 KL 散度，

$$D_{\mathrm{KL}} = \sum_s p_{\mathrm{data}}(s) \log \left(\frac{p_{\mathrm{data}}(s)}{p(s)} \right). \tag{19}$$

忽略数据熵后，这就简化为“伪似然”；数据熵是一个常数，对最小化散度没有影响，或者

$$\tilde{D}_{\mathrm{KL}} \sim \langle E \rangle_{p_{\mathrm{data}}} + \log Z. \tag{20}$$

我们将公式20称为伪似然，因为它在最大化似然时会得出相同的相对结果。请注意，我们已代入 $p(s)$ 的最大熵模型形式。因此，第一项是按数据分布加权的平均能量（对任何未观测到的选民进行边缘化后），第二项则是公式 18 给出的自由能。作为衡量拟合优度的另一种方法，

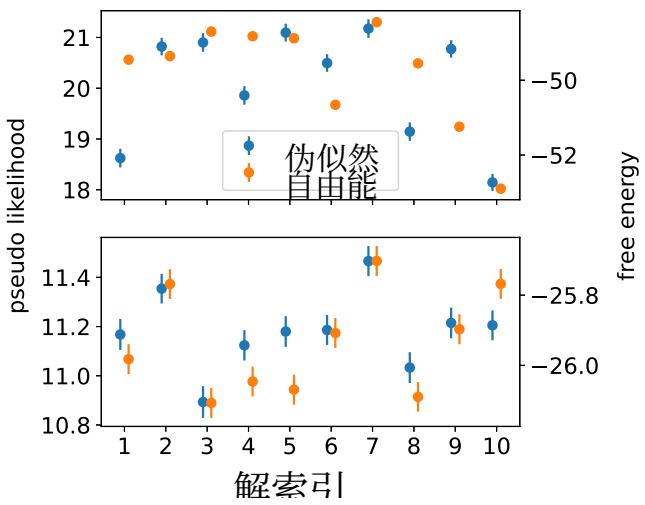

图6. 对 DC 和第一巡回上诉法院的十个不同简并解，拟合优度由伪似然（蓝色，定义见公式20）和自由能（橙色）衡量。综合指标显示，这些解在这两项指标上各有优劣。误差由正文所述的熵估计误差和能量估计的均值标准误差按范数合并（平方误差相加后再开平方）。

我们可以只考虑自由能，因为它平衡了能量（约束更多的分布会使其最小化）和熵（随机性更强、因而重数更高的分布会使其最大化）。如图 6 所示，这些度量可以区分某些解，其中拟合最佳的解与根据插补相关矩阵 [32] 对法官排序的定性检验结果一致。

值得注意的是，似然难以计算，因此评估最大熵模型拟合效果的一种常见方法，是使用未明确输入模型的相关性。例如，伊辛模型应能精确再现平均投票和成对相关性，但无法保证其能够再现高阶相关性。这些统计量常用于评估神经统计学模型和投票模型的拟合质量 [1, 6]。在投票数据集中，若投票数据缺失，依赖此类检验就会带来严重问题，例如只有部分个体共同投票时（如上诉法院中标准审判庭仅由三名法官组成），便无法检验相关性。本文将展示，使用我们的启发式方法，可以切实可行地对模型进行准确且更优的比较。

讨论

我们提出一种启发式算法，用于计算社会行为图模型的一项重要集体属性——信息熵。与通常通过预测低阶统计量来估计集体属性和拟合优度的做法相比[1，6]，，我们表明，利用稀疏图结构以及优良的熵估计器，可以计算出隐含纳入各阶相关性的集体统计量。作为该算法的示例，我们提出了两个美国上诉法院的司法投票模型。我们发现，可以对熵进行精确估计，并据此从一组简并解中识别出最优解。通过计算，我们还发现，每位法官的相对熵高于美国最高法院的相应数值；这一观察可为今后基于行为数据比较制度属性和约束的研究提供参考。

对于我们研究的问题，用于检测桥的社会学测度——结构约束——效果良好，但我们可以通过考虑其他图分解算法来扩展这项工作。实验中我们发现，其他常用测度（如中介中心性或 Louvain 社区聚类）在为递归熵计算分解图时效果不如结构约束。一项可能颇有价值的研究方向是考察更高阶的结构洞。此外，将有效电阻作为结构约束的边权重并未改善结果，不过，算法的效果无疑取决于图的属性。作为未来研究的铺垫，我们目前的工作为评估社会互动图模型中的集体属性和模型拟合度提供了有益的一步。

架构

我们的算法 TreeEnt 的开源软件包将在 https://github.com/eltrompetero/treeEnt 上发布。TreeEnt 是 Tree Entropy（树熵）的简称（同时也致意《指环王》：其中会移动的树被称为树人）。

该软件包的核心由“measures.py”、“test measures.py”和“NSB toolbox.py”中的两个 Python 模块组成。正如名称所示，正文中介绍的算法作为 TreeEntropy 类的一部分在 measures 模块中实现。测试模块提供了一些可通过 pytest 运行的自动化测试，以及用于检查算法有效性的例程；该算法的实现见随附的 Jupyter 笔记本。最后一个模块实现了 NSB 估计量，大量借鉴了 Bryan Daniels 编写的现有代码库：https://github.com/bcdaniels/toolbox.git。

TreeEnt 类接收一个用于描述概率图模型的模型实例。该实例必须提供从目标概率模型中进行采样的例程

。在当前实现中，这个类基于 ConIII 编码项目中实现的伊辛模型类。使用 ConIII 运行计算所需的依赖项，远多于 TreeEnt 中明确列出的依赖项。

致谢

EDL 感谢奥地利科学基金通过项目编号 ESP 127-N 提供资助。我们感谢 Cris Moore 提供的有益讨论，以及 Bryan Daniel 对其开源代码的帮助。

我们声明不存在利益冲突。

[1] Elad Schneidman、Michael J. Berry、Ronen Segev 和 William Bialek。成对相关性较弱，意味着神经元群体中的网络状态具有强相关性。*Nature*，440(7087):1007–1012，2006年4月。[2] Gasper Tkacik、Elad Schneidman、Michael J. Berry II 和 William Bialek。真实神经元网络的伊辛模型。*arXiv:q-bio/0611072*，2006年11月。[3] John Barton 和 Simona Cocco。通过选择性簇展开推断神经活动的伊辛模型：结构与编码特性。*J. Stat. Mech.*，2013 03 03002，2013 ( ):P March 。[4] Xiaowen Chen、Francesco Randi、Andrew M. Leifer 和 William Bialek。在小型脑中寻找集体行为。*arXiv:1810.07623 [cond-mat,* ，10月。*physics:physics, q-bio]* 2018 [5] Edward D. Lee、Xiaowen Chen 和 Bryan C. Daniels。发现神经活动中的稀疏控制策略。*PLoS Comput Biol*，18(5):e1010072，2022年5月。[6] Edward D. Lee、Chase P. Broedersz 和 William Bialek。美国最高法院的统计力学。*J. Stat.* , ( ): – *Phys.* 160 2 275 301, 2015 7月。[7] Gavin Hall 和 William Bialek。Twitter 社群的统计力学。 *J. Stat. Mech.*，2019(9):093406，2019年9月。[8] Edward D Lee、Daniel M Katz、Michael J Bommar- ito II 和 Paul H Ginsparg。集体结果的敏感性揭示了关键组成部分。*J. R. Soc. In-* , ( ), . *terface* 17 20190873 2020 [9] Yair Shemesh、Yehezkel Sztainberg、Oren Forkosh、Tamar Shlapobersky、Alon Chen 和 Elad Schneidman。鼠群中的高阶社会互动。 *eLife*，2 2013 :e00759，9月。[10] W. Bialek、A. Cavagna、I. Giardina、T. Mora、E. Silvestri、M. Viale 和 A. M. Walczak。自然鸟群的统计力学。*Proc. Natl. Acad. Sci. U.S.A.*，109 13 4786 4791，2012 ( ): – 3 月。[11] Frederick Reif。 *Fundamentals of Physics*。1965。[12] Bryan C. Daniels、David C. Krakauer 和 Jessica C. Flack。在小规模社会系统中控制有限临界行为。 *Nat. Comms.*，8:14301，2017年2月。[13] Jonathan S. Yedidia、William T. Freeman 和 Yair Weiss。贝特自由能、菊池近似与置信传播算法。技术报告 TR2001- 16，三菱电机研究实验室，2001年5月。[14] Marc Mezard 和 Andrea Montanari。*Information, Physics and Computation* 。《牛津研究生教材》。牛津大学出版社，牛津；纽约，2009 年。[15] S. Cocco 和 R. Monasson。利用含噪数据推断玻尔兹曼机的自适应簇展开。*Phys. Rev. Lett.*，106(9):090601，2011 年 3 月。[16] Daan Frenkel 和 Berend Smit。《*Understanding Molecular Simulation: From Algorithms to Applications* 1》。计算科学丛书，第版。学术出版社，圣迭戈，第 2 版，2002 年。[17] Ilya Nemenman、F. Shafee 和 William Bialek。《熵与推断：再探》。载于 T. G. Dietterich、S. Becker 和 Z. Ghahramani 编，《*Advances in Neural In- formation Processing Systems 14* 471 478 》，页码 –。麻省理工学院出版社，2002 年。[18] S. P. Strong、Roland Koberle、Rob R. de Ruyter van Steveninck 和 William Bialek。神经脉冲序列中的熵与信息。 *Phys. Rev. Lett.*，80(1):197–200，1998 年 1 月。[19] Edward D. Lee。党派直觉掩盖了关于美国最高法院投票集团的强大机构共识与宽广齐夫定律。*J. Stat. Phys.*，173(6):1722–1733，2018 年 12 月。[20] Ronald S. Burt。《*Structural Holes: The Structure of Social Competition* 》。哈佛大学出版社，马萨诸塞州剑桥，1992 年。[21] M. E. J. Newman 和 Juyong Park。社交网络为何不同于其他类型的网络。*Phys. Rev. E*，68 3 036122，2003 ( ): 9 月。[22] Duncan J Watts 和 Steven H Strogatz。“小世界”网络的集体动力学。393:3，1998 年。[23] M Ballerini、N Cabibbo、R Candelier、A Cavagna、E Cisbani、I Giardina、V Lecomte、A Orlandi、G Parisi、A Procaccini、M Viale 和 V Zdravkovic。支配动物集体行为的相互作用取决于拓扑距离而非度量距离：一项实地研究。*Proc Natl Acad Sci USA* 105 4 1233 1237, 2008 , ( ): – 1 月。[24] James Gleick。《 *The Information: A History, a Theory, a Flood* 》。Knopf Doubleday出版集团，纽约，2011 年。[25] C E Shannon。通信的数学理论。 *Bell Syst. Tech. J.*，27:379–423，623–656，1948 年。[26] Thomas M. Cover 和 Joy A. Thomas。《*Elements of In- formation Theory* 》。约翰·威立父子公司，霍博肯，第 2 版，2006 年。[27] 类似地，还可以证明，正方晶格上伊辛模型的配分函数复杂度并非按 2^N 增长，而是按 2^L增长，其中 L 是晶格一边的长度，因为可以对将问题一分为二的一条自旋线进行条件化，并递归地执行这一过程。[28] Ronald S. Burt。《结构洞与好想法》。 *American Journal of Sociology*，110(2):349–399，2004 年 9 月。[29] Burt 提出了另一种结构约束度量：其中，方程 12 中最后一项不再使用 w 赋予 v 的权重，而是使用该权重除以 w的最大边权重。由于这里的权重是均匀的，这意味着该项变为 1 [20]。[30] E. T. Jaynes。信息论与统计力学。 *Phys. Rev.*，106(4):620–630，1957 年 5 月。[31] Edward D. Lee 和 Bryan C. Daniels。逆伊辛（ConIII）的便捷接口：用于求解伊辛型最大熵模型的 Python 3 软件包。 *JORS*

,

7(1):3，2019年3月。[32] 我们使用插补相关矩阵的主维度来定义法官相对排名，并且该排序与一位法律学者在保守—自由维度上进行的定性检验结果一致。