arXiv:2305.04192v1 [math.PR] 7 May 2023

两种物种树模型下，与基因树和物种树相匹配的祖先构型总数的分布

Filippo Disanto*, Mic hael Fuchs†, Chun-Yen Huang‡, Ariel R. Paningbatan§, Noah A. Rosenberg¶

2023年5月9日

摘要

给定一个基因树标记拓扑 G 和一棵物种树 S，在 S 中考虑 G 的所有可能实现时， S 的内部节点k 处的祖先构型代表了在 k 处可能存在的基因谱系的不同组合集合。祖先构型已被引入作为一种数据结构，用于计算给定物种树时基因树标记拓扑的条件概率；枚举这些构型有助于描述这一计算的复杂性。当基因树标记拓扑 $G = t$与物种树 S的拓扑相匹配时，我们运用解析组合学技术，研究在均匀概率模型和尤尔概率模型下选取的随机带标签拓扑t 中，各节点的总祖先构型数的分布性质。在这两种概率情形下，我们证明，含有 n 个分类单元的随机带标签拓扑的祖先构型总数 T_n 渐近服从对数正态分布。对于均匀分布的带标号拓扑， T_n 的均值和方差的渐近增长分别满足 $\mathbb{E}_{\mathrm{U}}[T_n] \sim 2.449{\cdot}1.333^n$ 和 $\mathbb{V}_{\mathrm{U}}[T_n] \sim 5.050{\cdot}1.822^n$。在尤尔模型下，该模型赋予较为平衡的带标号拓扑更高的概率，我们得到均值 $\mathbb{E}_{\mathrm{Y}}[T_n] \sim 1.425^n$ 和方差 $\mathbb{V}_{\mathrm{Y}}[T_n] \sim 2.045^n$。

关键词： 解析组合学、祖先构型、基因树、系统发育学、物种树。

数学主题分类（2010年）： 05A15 · 05A16 · 05C05 · 92D15

1 引言

祖先构型是一些列表，用于描述给定基因树拓扑 G 和物种树 S 时，在 S 的给定节点处可能存在的基因谱系集合（图 1）。Wu [35] 引入了祖先构型，将其作为一种数据结构，用于在多物种溯祖模型下计算以物种树为条件的基因树拓扑概率。具体而言，对于给定的物种树 S，Wu 的算法“STELLS”通过递归计算G 在 S不同节点处祖先构型的概率，从叶端向 S [35]根推进，来评估基因树拓扑 G 的概率。STELLS 的运行时间取决于 G 在 S中的祖先构型总数 $c(G, S)$，也就是说，取决于 S不同节点处 G 祖先构型数量之和。

*比萨大学数学系，意大利比萨 56126。电子邮箱：filippo.disanto@unipi.it。

†国立政治大学数学科学系，台湾台北 116。电子邮箱：mfuchs@nctu.edu.tw。

‡国立交通大学应用数学系，台湾新竹 300。§菲律宾大学迪利曼分校数学研究所，菲律宾奎松市 1101。电子邮箱：arpaningbatan@math.upd.edu.ph。

¶美国加利福尼亚州斯坦福市斯坦福大学生物学系，邮编 94305。电子邮件：noahr@stanford.edu。

如果基因树的拓扑 $G = t$ 与物种树 S的拓扑相匹配，那么 G 在 S 中的构型总数就成为关于 t的函数 $c(G, S) = c(t)$，其在规模不断增大的树族上的行为可以用枚举组合学和解析组合学工具进行分析 [18]。在早期研究中 [7, 11], ，我们通过考察叶节点数为 n、随机选取的拓扑 t 在根处的祖先构型数量，推导出了确定总构型数均值 $\mathbb{E}[c]$ 和方差 $\mathbb{V}[c]$ 指数增长率的定理；此处用符号"$\bowtie$"表示指数增长率。具体而言，我们发现，对于规模为n的拓扑，均匀分布和 Yule 分布下的均值分别以 $\mathbb{E}_{\mathrm{U}}[c] \bowtie 1.333^n$ 和 $\mathbb{E}_{\mathrm{Y}}[c] \bowtie 1.425^n$ 的速度呈指数增长。在相同分布下，方差的指数增长率满足 $\mathbb{V}_{\mathrm{U}}[c] \bowtie 1.822^n$和 $\mathbb{V}_{\mathrm{Y}}[c] \bowtie 2.045^n$。

然而，这些结果并未充分刻画随机变量 c 的均值和方差的次指数增长，而描述构型总数渐近分布的问题也一直悬而未决。在此，我们利用生成函数对随机树拓扑中的祖先构型总数进行计数，从而解决了这些问题。令人惊讶的是，除一个我们精确计算出的常数因子外， $\mathbb{E}_{\mathrm{U}}[c]$、 $\mathbb{E}_{\mathrm{Y}}[c]$、 $\mathbb{V}_{\mathrm{U}}[c]$和 $\mathbb{V}_{\mathrm{Y}}[c]$的指数增长决定了相关量的完整渐近行为。更具体地说，对于在均匀分布和 Yule 分布下选取、规模不断增大的随机拓扑 n ，我们证明了 $\mathbb{E}_{\mathrm{U}}[c] \sim 2.449 \cdot 1.333^n$、 $\mathbb{E}_{\mathrm{Y}}[c] \sim 1.425^n$、 $\mathbb{V}_{\mathrm{U}}[c] \sim 5.050 \cdot 1.822^n$ 和 $\mathbb{V}_{\mathrm{Y}}[c] \sim 2.045^n$。此外，在均匀模型和 Yule 模型下，我们都得到了祖先构型总数的渐近对数正态分布。我们研究了构型总数与密切相关的根祖先构型数之间的相关性。

我们的方法采用解析组合学的标准技术来推导整数序列的渐近增长，并结合一项关键观察：通过等价地考察均匀分布的平面树类（通常称为卡特兰树 和递增二叉树），可以研究在均匀模型和尤尔模型下选取的随机树拓扑中祖先构型数量的分布性质。这些结果有助于对基因树与物种树之间关系中的组合结构进行枚举研究 [2, 6, 8, 9, 10, 12, 13, 20, 25, 26, 27, 28], ，并有助于将基于祖先构型计算基因树概率的算法复杂度，与采用基于其他结构的评估方法的算法复杂度联系起来 [6, 33, 36]。

2 预备知识

我们首先介绍一些定义、预备结果和计数组合学的基本原理。在第 2.1 节中，我们引入带标号拓扑及其均匀概率分布和 Yule 概率分布。在第 2.2 节中，我们介绍用于分析整数序列渐近增长的生成函数技巧。

2.1 带标号拓扑与无标号拓扑

大小为 $|t| = n$ 的带标签拓扑 t，也称为系统发育树，是一棵二叉有根树，其 n 外部节点（即叶节点）各自带有不同的标号，通常按字母顺序取前 n 个 n 字母（图 1A）。带标号拓扑是非平面或无序的，也就是说，每对子节点都没有左右方向之分；交换某个内部节点下的两棵子树，得到的仍是同一个带标号拓扑。

将带标签拓扑 t 的内部节点用不同于叶节点所用的字母表示会比较方便（图1B）。我们用与边相邻、距离叶节点更近的直接后代节点（的标号）来标识每条边 t 。我们用符号 $\preceq$表示定义在 t 节点集合上的后代—祖先序关系。更确切地说，对于不同的节点 x 和 y，若节点 y 属于连接节点 x 与 t的根之间的路径，我们就记作 $x \prec y$在 t 中。以节点 k为根的 t 子树，包含 t 中满足 $x \preceq k$的节点 x ，记作 t^k。因此，特别地，若 k 是 t 的根，则 $t^k = t$ ；若 k 是 t的叶节点，则 $t^k = \bullet_k$ ，其中 $\bullet_k$ 是仅包含节点 k的子树。

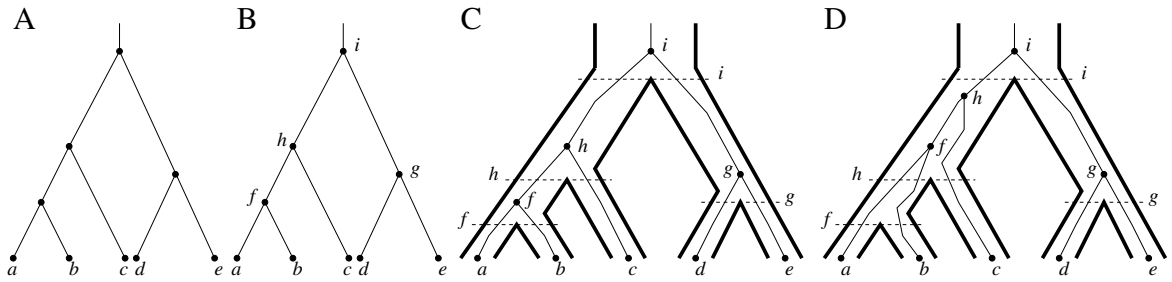

图1：带标签拓扑、基因树和物种树。**(A)** 一个大小为5的带标签拓扑。**(B)** 对（A）中的带标签拓扑的内部节点进行标记后的结果。我们通过树的直接后代节点来标识每条边；例如，谱系 h 由谱系 c 和 f合并而形成。**(C)** 在具有匹配带标签拓扑的物种树（粗线）中，（A）的基因树标记拓扑的一种可能实现（细线）。物种树节点 i处的祖先构型为 $\{g, h\}$。节点 h处的构型为 $\{c, f\}$。**(D)** 在匹配物种树中，（A）的基因树标记拓扑的另一种不同实现。物种树节点 i和 h处的祖先构型分别为 $\{g, h\}$和 $\{a, b, c\}$。

从带标签拓扑 t中移除标号后，我们便得到其底层树形 或无标签拓扑 $t_\circ$。无标签拓扑也称为*Otter* 树 [23]；随着叶节点数 $n \geqslant 1$增加，它们的数量由 Wedderburn-Etherington 数列给出：1，1，1，2，3，6，11，23，46，... [17, 34]。

不同的带标签拓扑 t_1 和 t_2 可能共享相同的无标签拓扑。例如，在 *Newick* 格式中，带标签拓扑 $t_1 = ((a, b)\ c)$ 和 $t_2 = (a, (b, c))$ 共享无标签拓扑 $((\bullet, \bullet)\ \bullet) = (\bullet, (\bullet, \bullet))$。形状为 t 的带标签拓扑数量 lab(t) 可通过 [7], 的式（22）递归求得。

$$\mathrm{lab}(t) = \mathrm{lab}(t_L)\,\mathrm{lab}(t_R)\binom{|t|}{|t_L|}\frac{1}{1+\delta_{t_L=t_R}}. \tag{1}$$

这里， t_L 和 t_R 是从 t 的根部延伸出的两棵子树（简称根子树），而 $\delta_{t_L=t_R}$ 是克罗内克 delta：当 t_L 和 t_R 是相同的无标签拓扑时，其值为 1。若 $|t| = 1$，我们令 lab(t) $= 1$ 。

令 L_n 表示大小为 n的带标号拓扑集合。对于 $n \geqslant 2$， L_n 的基数为 $|L_n| = (2n-3)!! = 1 \times 3 \times 5 \times \ldots \times (2n-3)$ [16, 17],，可写为

$$|L_n| = \frac{(2n)!}{2^n(2n-1)n!}. \tag{2}$$

在固定大小为 n [1]的带标号拓扑集合 L_n 上，可以考虑不同的概率模型。在均匀模型下，每个带标号拓扑 $t \in L_n$ 具有相同的概率

$$\mathbb{P}_U[t] = \frac{1}{|L_n|} = \frac{2^n(2n-1)n!}{(2n)!}.$$

Yule 模型是一种生成模型，其中每个谱系在时间向前推进时都有相同的概率成为下一个分叉的谱系；等价地，每对谱系在时间回溯时都有相同的概率成为下一对合并的谱系 [19, 31, 37]。该模型的许多组合特征已得到研究 [5, 14, 15, 22, 24]；具有 n 个叶节点的带标号拓扑 t 的概率为

$$\mathbb{P}_Y[t] = \frac{2^{n-1}}{n!\prod_{r=3}^{n}(r-1)^{d_r(t)}} \tag{3}$$

在尤尔模态下 l，其中 $d_r(t)$是 t 中具有 r 个后代叶 [4, 21, 32]的节点数。

由于分母中出现了乘积，在 Yule 分布下，越平衡的带标号拓扑往往具有越大的概率 [19]。例如，在大小为5的带标号拓扑中，图 1A 所示的拓扑具有最大的尤尔概率 $\mathbb{P}[(((a, b)\ c), (d, e))] = \frac{1}{60}$；将 $((((a, b)\ c)\ d)\ e)$ 和 $(((a, b), (c, d))\ e)$ 分别作为其未标记形状的代表性标记拓扑，我们有 $\mathbb{P}[((((a, b)\ c)\ \overline{d})\ e)] =$1 180 和 $\mathbb{P}[(((a, b), (c, d))\ \underline{e})] = \frac{1}{90^\circ}$ —

2.2 渐近增长与生成函数

本文研究非负整数序列的增长。我们使用以下记号。对于两个序列 $(a_n)_{n\geqslant 0}$ 和 $(b_n)_{n\geqslant 0}$，如果比值 b_n/a_n 在 $n\to\infty$ 时收敛于 1，我们记作 $a_n\sim b_n$。如果 $a_n\sim b_n$，那么我们称序列 a_n 和 b_n 渐近地具有相同的增长。若 $a_n\sim k^n s(n),$，则称序列 a_n 具有指数增长 k^n，或等价地，具有指数阶 k，其中 k 是一个常数，$s(n)$ 是一个次指数因子。如果 a_n 和 b_n 具有相同的指数增长，我们记作 $a_n\bowtie b_n$。

序列 $(a_n)_{n\geqslant 0}$ 的生成函数是幂级数 $A(z)=\sum_{n=0}^{\infty}a_n z^n$。将 $A(z)$ 与生成函数 $B\ z\ =\sum_{n=0}^{\infty}b_n z^n$ $A\ z\ B\ z\ =\sum_{n=0}^{\infty}\sum_{j=0}^{n}a_j z^j\cdot b_{n-j}z^{n-j}=$ ( ) 相乘，得到生成函数 $\sum_{n=0}^{\infty}\sum_{j=0}^{n}a_j b_{n-j}\ z^n$ $\sum_{j=0}^{n}a_j b_{n-j}$ n k ( ) ( ) ( )，其第 th 个系数为卷积。此外，如果 是常数，那么 $A(z)+B(z)=\sum_{n=0}^{\infty}(a_n+b_n)z^n$，且 $kA(z)=\sum_{n=0}^{\infty}(ka_n)z^n$。

如果将 $A(z)=\sum_{n=0}^{\infty}a_n z^n$ 视为复变量 z 的函数，那么研究 $A(z)$ 在其奇点附近的性质——即复平面中 $A(z)$ 不解析的点——有助于研究系数 $a_n=[z^n]A(z)$ 的渐近增长。最简单的情形是 $A(z)$ 具有唯一的主导奇点 $\alpha>0$，也就是说，$z=\alpha>0$ 是 $A(z)$ 中模最小的唯一奇点。在这种情况下，根据 [18]，中定理 IV.7 和 VI.4 所述的相当一般的条件，生成函数 $A(z)$ 在 $z=\alpha$ 处的奇异展开 $A_\alpha(z)$ 决定了系数 a_n 的渐近增长，其形式为

$$a_n\sim[z^n]A_\alpha(z)\bowtie\alpha^{-n}. \tag{4}$$

换言之，随着 n 增大，n 阶系数的 $A(z)$与展开式 $A_\alpha(z)$的 n 阶系数具有相同的增长速度，其中，特别地，$\frac{1}{\alpha}$ 是序列 a_n 的指数阶。例如，如 [18] 中的例 II.19 所示，$L(z)=1-\sqrt{1-2z}$ 是与序列 $|L_n|/n!$ 相关联的生成函数——其中，L_n 是大小为 n 的带标号拓扑数量（式 2）。$L(z)$ 的主导奇点为 $\alpha=\frac{1}{2}$，事实上，与式 (4) 一致，我们有 $|L_n|/n!=\binom{2n}{n}/[2^n(2n-1)]\bowtie 2^n=\alpha^{-n}$。

在以下各节中，我们将式 (4) 中的渐近关系应用于生成函数 $A(z)$。该函数具有唯一的主导奇点 $\alpha>0$，其奇异展开为 $A_\alpha(z)=1/(1-\frac{z}{\alpha})$ 或 $A_\alpha(z)=k_1-k_2\sqrt{1-z/\alpha}$，其中 k_1 和 $k_2>0$ 为常数。利用式 (4) 中与 $[z^n]\big(1/(1-z)\big)=1$ 和 $[z^n]\big(-\sqrt{1-z}\big)\sim 1/(2\sqrt{\pi n^3})$ 的等价关系（见 [18] 第 388 页），这些展开得到的结果为

$$A_\alpha(z)\ =\ \frac{1}{1-\frac{z}{\alpha}}\quad\Rightarrow\quad a_n\sim\alpha^{-n} \tag{5}$$

$$A_\alpha(z)\ =\ k_1-k_2\sqrt{1-\frac{z}{\alpha}}\quad\Rightarrow\quad a_n\sim k_2\,\frac{\alpha^{-n}}{2\sqrt{\pi n^3}}. \tag{6}$$

如果 $A'(z)=\sum_{n=0}^{\infty}na_n z^{n-1}$ 和 $\int_0^z A(t)\,dt=\sum_{n=0}^{\infty}[a_n/(n+1)]z^{n+1}$ 是通过对生成函数 $A(z)$求导和积分得到的生成函数，那么我们使用 [18] 的定理 VI.8 和 VI.9(ii)，计算 $A'(z)$ 和 $\int_0^z A(t)\,dt$ 在其主导奇点 α 处的奇异展开，如下所示

$$A'(z)\ \overset{z\to\alpha}{\sim}\ \Big(A_\alpha(z)\Big)' \tag{7}$$

$$\int_0^z A(t)\,dt\ \overset{z\to\alpha}{\sim}\ \int_0^z A_\alpha(t)\,dt. \tag{8}$$

特别是，$A'(z)$ 和 $\int_0^z A(t)\,dt$ 具有与 $A(z)$ 相同的主导奇点 α，它们的奇异展开分别通过对 $A(z)$ 的奇异展开 $A_\alpha(z)$ 求导和积分得到。请注意，我们将公式 eq. (8) 应用于生成函数 $A(z)$，其奇异展开为 $A_\alpha(z)=1/(1-\frac{z}{\alpha})^2$,，符合 [18]的定理 VI.9 情形 (i) 的假设。

3 匹配基因树与物种树的祖先构型

在本节中，我们定义了匹配的基因树和物种树的祖先构型（第3.1节），并解释了如何通过带标号拓扑和其他树族，对祖先构型数量的分布性质进行等价研究（第3.2节）。第3.3节回顾了 [7]和 [11] 关于随机选取的带标号拓扑的根处祖先构型的研究结果。

3.1 定义与示例

我们首先介绍基因树、物种树，以及基因树在物种树中的实现。接下来， [35], 我们为具有相同带标签拓扑的基因树与物种树定义祖先构型。

3.1.1 基因树与物种树

带物种树 标签的拓扑表示一组种群或物种之间的演化关系，这些种群或物种以树的叶节点标识。带基因树 标签的拓扑描述从一组种群或物种中采样的个体（基因组区域）的演化。

如果采样个体来自物种树叶节点所代表的种群，那么从生物学角度看，基因树可视为一组基因谱系（图 1C，细线），它们在演化力量的作用下，沿物种树的分支结构扩散（图 1C，粗线）。我们假设物种树每个叶节点所代表的种群中恰好采样一个个体。我们只考察匹配的 物种树与基因树对，即具有相同带标签拓扑的树对（图 1C 和 D）。

对于固定的物种树，同一基因树可能由进化过程的不同实例（即实现）产生。图 1 的 C 和 D 面板将 A 面板中的基因树标记拓扑（内部节点如 B 面板所示）描绘在物种树中，展示了两种不同的实现；这两种实现的区别在于，基因树的合并事件（内部节点）发生于物种树的哪些分支（边）。具体而言，从 C 面板切换到 D 面板时，我们发现基因树合并事件 f 发生在两个不同的物种树分支中。用数学术语来说，在物种树 S 中，具有匹配带标签拓扑 $S = G = t$ 的基因树 G 的一种实现是一个函数 R，它将 t 的内部节点集合映射到其自身，并满足以下两个条件：(i) 对所有内部节点 k、 $k \preceq R(k)$，以及 (ii) 对所有内部节点 k_1 和 k_2， $k_1 \preceq k_2 \Rightarrow R(k_1) \preceq R(k_2)$。通过用物种树分支的直接后代节点来标识该分支，基因树 G 的内部节点 k 所对应的合并事件由实现 R 指定，使其发生在物种树的分支 $R(k)$ 中。例如，图 1C 中所示进化情景对应的实现为：对于所有 $k \in \{f, g, h, i\}$， $R(k) = k$；而在图 1D 中，实现则为 $R(f) = R(h) = h$、 $R(g) = g$，以及 $R(i) = i$。

3.1.2 祖先构型

当物种树配备了用于衡量相邻节点对之间时间间隔的分支长度时，可以在多物种溯祖过程模型 [6] 下计算给定物种树的基因树标记拓扑的条件概率。Wu [35] 引入了祖先构型，将其作为递归计算该条件概率的一种数据结构，其中物种树的每个节点都关联着一组给定的祖先构型，这些构型取决于基因树可能的实现。在每一步中，Wu 的算法都会计算在共祖模型下，给定物种树节点处的一个祖先构型由其子节点处的祖先构型“演化”而来的概率，并按照从叶节点到根的顺序进行递归计算。Wu 算法的计算成本受到物种树所有节点上祖先构型总数的影响。

在我们的设定中，基因树带标签拓扑与物种树的带标签拓扑相匹配，祖先构型定义如下。给定一个基因树带标签拓扑 $G = t$ 和一棵匹配物种树 S，设 R 为 G 在 S中的一个实现。对于 S中的给定节点 k，考虑在k处、k、 R时的基因谱系集合 $C(k) = C(k, R)$（即 G的边）

这些基因谱系存在于 S 中，时间点紧邻节点 k之前，且时间从叶节点流向物种树根。集合 $C(k)$称为在实现 R 下、物种树节点 k 处基因树的祖先构型 。例如，在图 1C 的实现中，物种树内部节点处的祖先构型为$C(f) = \{a,b\}$、 $C(g) = \{d,e\}$、 $C(h) = \{c,f\}$和 $C(i) = \{g,h\}$，其中每个基因谱系均由其紧邻的后代节点标识。在图 1D 的实现中， S的内部节点处、记作 h 的祖先构型为 $C(h) = \{a,b,c\}$；其他节点处的祖先构型与前一种情况相同。

令 $\mathbf{R}(t)$ 表示在匹配物种树 S 中实现基因树标记拓扑 $G = t$ 的所有可能实现的集合。对于 S 中给定的节点 k，考虑所有可能的实现 $R \in \mathbf{R}(t)$，我们定义集合

$$C_k = C_k(t) = \{C(k,R) : R \in \mathbf{R}(t)\}, \tag{9}$$

其基数为

$$c_k = c_k(t) = |C_k|. \tag{10}$$

因此， c_k 统计了在考虑 S 中基因树标记拓扑 $G = t$ 的所有实现时， G 的基因谱系到达 S 中节点 k 之前这一时间点的方式数。例如，取图 1A 中的 t，则有 $C_f = \{\{a,b\}\}$、 $C_g = \{\{d,e\}\}$、 $C_h = \{\{a,b,c\},\{c,f\}\}$ 和 $C_i = \{\{a,b,c,d,e\},\{c,d,e,f\},\{d,e,h\},\{a,b,c,g\},\{c,f,g\},\{g,h\}\}$，共计 10 种祖先构型。注意，根据祖先构型的定义， $\{k\} \notin C_k$。事实上，只有在物种树中节点 k 以上的部分，基因谱系才能合并形成 G 的节点 k。为与这一观察保持一致，若节点 k 是叶节点，则令 $c_k = 0$。

集合 $C_k \cup \{\{k\}\}$ 可以视为以节点k为根的 t 的子树 t^k 的极大反链集合。特别是，如果 r 是 t的根，那么 $C_r \cup \{\{r\}\}$ 对应于 t的极大反链集合。子树 t^k 的反链确实是其节点（可能包括叶节点）的一个子集，其中的元素在 t中定义的后代—祖先序关系 $\preceq$ 下两两不可比。 C_k 的祖先构型是 t^k 的一个“极大”反链，也就是说，它不真包含于 t^k的任何其他反链中。

将带标签拓扑 t 的节点集合 $N(t)$ 中的 k 取值对应的数量 c_k 求和，我们便得到 t 的祖先构型总数，记作

$$c = c(t) = \sum_{k \in N(t)} c_k. \tag{11}$$

等价地， $c + 2|t| - 1$ 是 t的各子树中极大反链的总数，包括 $N(t)$中的 $|t|$ 叶节点，以及 $|t| - 1$ 个内部节点各自的计数，其中节点本身也作为一个极大反链。

对于基因树标记拓扑 $G = t$ 和匹配物种树 S， t 的祖先构型总数 $c(t)$ 可通过递归分解 t 的左、右根子树 t_L 和 t_R 来计算（固定将 t 嵌入平面的一种方式后）。若 $c_r(t)$ 表示 t根处的祖先构型数，则

$$\begin{aligned} c(t) &= c(t_L) + c(t_R) + c_r(t) && (12)\\ c_r(t) &= [c_r(t_L)+1][c_r(t_R)+1], && (13) \end{aligned}$$

其中，对于 $|t| = 1$ [11]， $c(t) = c_r(t) = 0$。

例如，假设 t 是图1A中的带标签拓扑，其内部节点如图1B所示。回顾一下， t^k 指的是以节点 k为根的 t 子树，通过应用方程（12）和（13），我们得到

$$\begin{aligned} c(t) &= c(t^h) + c(t^g) + c_r(t)\\ &= [c(t^f) + \cancel{c(t^c)} + c_r(t^h)] + [\cancel{c(t^d)} + \cancel{c(t^e)} + c_r(t^g)] + [c_r(t^h)+1][c_r(t^g)+1]\\ &= c(t^f) + c_r(t^h) + c_r(t^g) + ([c_r(t^f)+1]\cancel{[c_r(t^c)}+1] + 1)(\cancel{[c_r(t^d)}+1]\cancel{[c_r(t^e)}+1] + 1)\\ &= \cancel{[c(t^a)} + \cancel{c(t^b)} + c_r(t^f)] + [c_r(t^f)+1]\cancel{[c_r(t^c)}+1] + \cancel{[c_r(t^d)}+1]\cancel{[c_r(t^e)}+1] + [c_r(t^f)+2]2\\ &= 4c_r(t^f) + 6\\ &= 4\cancel{[c_r(t^a)}+1]\cancel{[c_r(t^b)}+1] + 6\\ &= 10. \end{aligned}$$

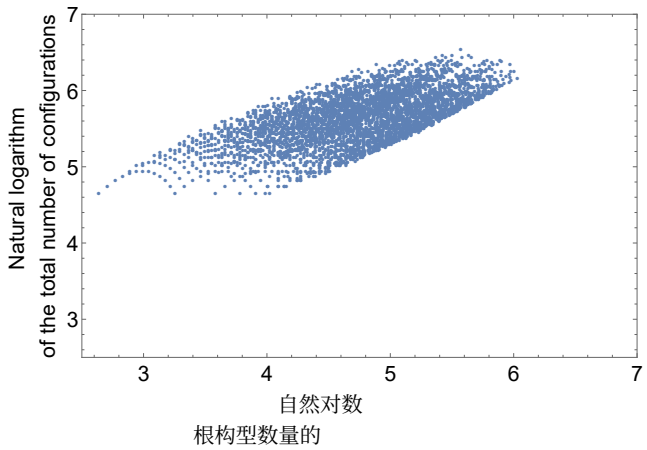

图2：对于规模为 $n = 15$ 个叶节点的4850种无标号拓扑中的每一种，取其代表性标号，给出了祖先构型总数和根祖先构型数量的自然对数。

当带标签拓扑 t 的大小为 n时， c 祖先构型的总数可以根据 c_r 根祖先构型的数量界定为

$$c_r \leqslant c \leqslant (2n-1)c_r. \tag{14}$$

事实上， t 中有 $|N(t)| = 2n-1$ 个节点；对于 t中的每个节点 k，都有 $c_k \leqslant c_r$。

由于 c 和 c_r 至多相差树的规模 n的多项式因子，因此在规模不断增大的树族中考察时，它们具有相同的指数阶。基于这一观察，关于随机树中根祖先构型数量渐近增长的 [7] 和 [11]研究，得出了在均匀分布和尤尔分布下从规模为 n的带标号拓扑中随机选取的树，其祖先构型总数的均值和方差的指数阶。在第4节中，我们进一步细化这些结果，得出祖先构型总数的完整渐近分布。我们还研究了规模不断增大的随机有标号拓扑中，祖先构型总数与根祖先构型数量之间的相关性。

图 2 在对数尺度上展示了大小为 $n = 15$ 的每种无标签拓扑中，代表性带标签树的祖先构型总数和根祖先构型数。该图表明，祖先构型总数超过根祖先构型数。图中还显示，这两个量在不同树之间呈正相关。

3.2 有序树族 及祖先构型的等价概率模型 s

根据带标签拓扑 t 中某个节点处祖先构型集合的方程（9）定义，该集合仅取决于 t 的形状。对于多种类型的二叉有根树 t （例如带标签、无标签、有序、无序），也可以采用相同的方式定义祖先构型以及方程（10）和（11）中的量。本节说明，对于随机有标号拓扑所考察的祖先构型数量的概率性质，可以通过在不同树族上进行等价分析来研究。在第3.2.1节和第3.2.2节中，我们介绍有序无标号拓扑族和有序无标号历史族。接下来，在第3.2.3节中，我们回顾关于祖先构型数量分布的若干等价关系结果 [7] 。特别地，引理1指出，当在均匀分布的带标签拓扑上考察祖先构型数量时，其分布与在相同大小的均匀分布的有序无标号拓扑上考察时相同。这种等价关系还扩展到在 Yule 分布的带标记拓扑和均匀分布的有序无标记历史上考察的祖先构型数量的分布。最后，我们以引理2结束本节，该引理是第4节计算的一个预备结果。

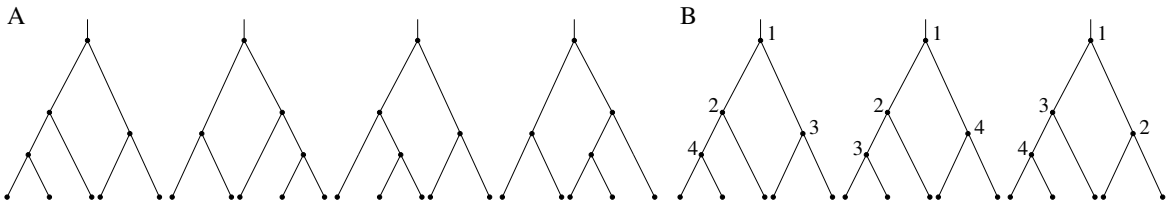

图3：有序无标号拓扑和有序无标号历史。（**A**）底层无序无标签拓扑为 **(((•, •) •), (•, •))** 的四种可能的有序无标号拓扑。（**B**）底层有序无标号拓扑与（A）中最左侧形状相同的三种可能的有序无标号历史。

3.2.1 有序无标号拓扑

一种有序无标号拓扑是二叉有根平面树，也就是说，是一种无标签拓扑 t ，其内部节点向下延伸的子树带有左右定向。有序无标号拓扑的树形状是其底层无序无标签拓扑。每个有序无标号拓扑都是其形状在平面上的一种嵌入。

图3A展示了形状为 **(((•, •) •), (•, •))** 的四种有序无标号拓扑。记 out(t) 为形状为 t 的有序无标号拓扑数，则文献 [7] 中的式（23）给出

$$\mathrm{out}(t) = 2\,\mathrm{out}(t_L)\,\mathrm{out}(t_R)\,\frac{1}{1+\delta_{t_L=t_R}}, \tag{15}$$

其中，当 $|t|=1$ 时，out(t) $=1$ 。

有序无标号拓扑也称为卡特兰树，因为按大小n计数时，其数量由第（$n-1$）个卡特兰数 C_{n-1} ([30], 习题 6.19d)给出，其中

$$C_n = \frac{1}{n+1}\binom{2n}{n} \sim \frac{4^n}{\sqrt{\pi n^3}}. \tag{16}$$

序列 C_n 的生成函数为

$$C(z) = \sum_{n=0}^{\infty} C_n z^n = \frac{1-\sqrt{1-4z}}{2z}. \tag{17}$$

$C(z)$在其主导奇点 $\alpha = \frac{1}{4}$ 处具有奇异展开 $C(z) \overset{z\to\alpha}{\sim} 2-2\sqrt{1-4z}$，这一点可通过将$z=\frac{1}{4}$代入分母看出；根据式（6），我们得到式（16）中的渐近表达式。

分解为均匀选取的具有 n 个叶节点的有序无标号拓扑的左、右根子树分别具有 j 和 $n-j$ 个节点的概率提供了一个有用的公式（$1 \leqslant j \leqslant n$）。每个具有 $n \geqslant 2$ 个叶节点的有序无标号拓扑 t 都通过以下递归构造生成：(a) 分别取大小为 j 和 $n-j$ 的两个有序无标号拓扑 t_L 和 t_R，以及 (b) 将 t_L 和 t_R 分别附加到一个公共根节点的左侧和右侧。例如，图 3A 最左侧的有序无标号拓扑是通过将 t_L = **((•, •) •)** 附加到共享根的左侧，并将 t_R = **(•, •)** 附加到其右侧而得到的。对于图 3A 中第三个有序无标号拓扑，我们则取 t_L = **(•, (•, •))** 和 t_R = **(•, •)**。由于存在 C_{j-1} 个可能的选择用于 t_L，且存在 C_{n-1-j} 个选择用于 t_R，因此，一个大小为 n 的均匀分布的有序无标号拓扑 t 具有大小为 j 的左根子树 t_L 和大小为 $n-j$ 的右根子树 t_R 的概率为

$$\mathbb{P}[|t_L| = j \,\&\, |t_R| = n-j] = \frac{C_{j-1}\,C_{n-1-j}}{C_{n-1}}. \tag{18}$$

3.2.2 有序无标号历史

大小为 n 的有序无标签 历史 是一个具有 n 个叶节点的无标签拓扑的平面嵌入，其中内部节点与区间 $[1, n-1]$ 中的整数一一对应地标记，并且每个非根内部节点的标号都大于其父节点的标号（图 3B ）。从生物学角度来看，内部节点上的标号定义了历史中合并事件的时间顺序。

用计算机科学的术语来说，有序无标号历史——去掉叶节点及其相连的边后——对应于所谓的递增二叉树（ [18] 中的例 II.17）。“递增”一词指的是节点的标号沿树中从根到叶节点的任意路径递增。为指定内部节点的线性顺序，我们在有序无标号历史的 Newick 格式中，将相应内部节点的标号以下标形式写在右括号旁边。例如，$(((\bullet,\ \bullet)_4,\ \bullet)_2,(\bullet,\ \bullet)_3)_1$ 表示图 3B 中描绘的第一个有序无标号历史。

有序无标号历史的树形是去除内部节点的标号并忽略左右定向后得到的底层无序无标签拓扑。采用方程 (1) 和 (15) 中的相同记法，树形为 t 的有序无标号历史数量 $\mathrm{ouh}(t)$ 递归计算如下

$$\mathrm{ouh}(t) = 2\,\mathrm{ouh}(t_L)\,\mathrm{ouh}(t_R)\binom{|t|-2}{|t_L|-1}\frac{1}{1+\delta_{t_L=t_R}}, \tag{19}$$

其中，若 $|t| = 1$，则 $\mathrm{ouh}(t) = 1$。每个形状为 t 的有序无标号历史，都是通过将两个形状分别为 t_L 和 t_R 的有序无标号历史 h_1 和 h_2 接到共同根节点的左侧和右侧构造而成，同时从 $\binom{|t|-2}{|t_L|-1}$ 种可能性中选择一种，将 h_1 的内部节点的线性排序与 h_2 的内部节点的线性排序合并。因子 $1/(1+\delta_{t_L=t_R})$ 用于计入这一过程中的可能对称性。

大小为 n 的所有可能有序无标号历史的集合由 F_{n-1} ([31], , 第 47 页)枚举，其中

$$F_n = n!. \tag{20}$$

通过扩展有序无标号拓扑的递归构造，可以得到一个类似于式（18）的有序无标号历史公式。要构造一个大小为 n 的有序无标号历史 t，我们执行以下步骤：(a) 取两个大小分别为 j 和 $n-j$ 的有序无标号历史 t_L 和 t_R；(b) 将 t_L 和 t_R 分别接到共同根节点的左侧和右侧；(c) 将 t_L 的内部节点的线性排序与 t_R 的内部节点的线性排序合并，以确定 t 的内部节点的线性排序。

例如，图 3B 最左侧的有序无标记历史，是通过在同一个根的左侧附加 $t_L = ((\bullet,\ \bullet)_2,\ \bullet)_1$，在右侧附加 $t_R = (\bullet,\ \bullet)_1$，然后将 t_L 和 t_R 的内部节点的排序进行合并，并将 t_R 的根节点置于 t_L 的内部节点之间而得到的。由于 t_L 有 F_{j-1} 种选择，有 F_{n-1-j} t_R $j-1$ t_L 种选择，并且将 的内部节点排序与 t_R 的 $n-1-j$ 个内部节点的排序进行合并有 $\binom{n-2}{j-1}$ 种方式，因此，一个大小为 n 的均匀分布的有序无标记历史 t，其左根子树 t_L 的大小为 j、右根子树 t_R 的大小为 $n-j$ 的概率为

$$\mathbb{P}[|t_L| = j\,\&\,|t_R| = n-j] = \frac{F_{j-1}\,F_{n-1-j}\,\binom{n-2}{j-1}}{F_{n-1}} = \frac{1}{n-1}. \tag{21}$$

3.2.3 祖先构型的等价模型

我们此前注意到， [7] 有序无标号拓扑和有序无标号历史分别可用于研究均匀分布和 Yule 分布的带标号拓扑的祖先构型数量。我们观察到，给定树结构的祖先构型数量仅取决于底层无标记拓扑。其次，如 [7], 引理 1、2 和 3 的证明所示，大小为 n 的有序无标号拓扑集合上的均匀分布，与大小为n 的带标号拓扑集合上的均匀分布，在大小为 n的底层无标记拓扑集合上诱导出相同的分布；均匀分布

表1：大小为5的无标号拓扑的诱导概率。

无标号拓扑 t	诱导概率 按均匀有序方式 无标号拓扑 $\frac{\text{out}(t)}{C_4}$	诱导概率 按均匀方式 带标号拓扑 $\frac{\text{lab}(t)}{\lvert L_5\rvert}$	诱导概率 按均匀有序方式 无标号历史 $\frac{\text{ouh}(t)}{F_4}$	诱导概率 按 Yule 模型 带标号拓扑 $\text{lab}(t)\times\mathbb{P}_Y[t]$
	$\frac{8}{14}=\frac{4}{7}$	$\frac{60}{105}=\frac{4}{7}$	$\frac{8}{24}=\frac{1}{3}$	$60\times\frac{1}{180}=\frac{1}{3}$
	$\frac{2}{14}=\frac{1}{7}$	$\frac{15}{105}=\frac{1}{7}$	$\frac{4}{24}=\frac{1}{6}$	$15\times\frac{1}{90}=\frac{1}{6}$
	$\frac{4}{14}=\frac{2}{7}$	$\frac{30}{105}=\frac{2}{7}$	$\frac{12}{24}=\frac{1}{2}$	$30\times\frac{1}{60}=\frac{1}{2}$

对于每个大小为 5 的无标签拓扑 t，有序无标号拓扑上的均匀分布所诱导的 t 概率，是 out(t) 与有序无标号拓扑总数 $C_4=14$ 之比。均匀有标号拓扑（$|L_5|=105$）、有序无标号历史（$F_4=24$）和 Yule 带标签拓扑也有类似的计算。量 out(t)、lab(t) 和 ouh(t) 分别根据方程（15）、（1）和（19）递归计算。Yule 分布在带标号拓扑上诱导的 t 概率，可通过将 lab(t) 乘以每个以 t 为底层无标记拓扑的带标号拓扑在方程（3）中的 Yule 概率得到。第二列与第三列一致，第四列与第五列也一致。

大小为 n 的有序无标号历史上的分布，与大小为 n 的带标号拓扑上的尤尔分布，在大小为 n的无标号拓扑上诱导出相同的分布。换言之，对于每个无标号拓扑 t，具有t 形状的均匀分布的有序无标号拓扑（或历史）的概率之和，等于树形为t的均匀分布（或尤尔分布）带标号拓扑的概率之和。这两个事实引出了下面的引理。表1展示了 $n=5$的情况。

引理1 在大小为 n 的均匀分布（或尤尔分布）带标号拓扑上，祖先构型数量的分布，与在大小为 n的均匀分布有序无标号拓扑（或历史）上祖先构型数量的分布相同。

因此，对于均匀分布和尤尔分布的带标号拓扑，其祖先构型数量的概率性质，分别可以通过均匀分布的有序无标号拓扑和均匀分布的有序无标号历史来等价研究。

为使用这些等价关系，我们需要以下引理。

引理2 令 R_n 和 T_n 分别为在均匀分布下选取的、大小为n 的随机有序无标号拓扑（分别为历史）中根祖先构型数和祖先构型总数的随机变量。等价地，根据引理*1*， R_n 和 T_n 分别表示在均匀分布（分别为 ***Yule*** 分布）下选取的、大小为 n的随机带标签拓扑中的根祖先构型数和祖先构型总数。那么我们有 $R_1=T_1=0$，并且对于 $n\geqslant 2$,

$$T_n \stackrel{d}{=} T_{I_n}+T^*_{n-I_n}+R_n, \tag{22}$$

$$R_n \stackrel{d}{=} R_{I_n}R^*_{n-I_n}+R_{I_n}+R^*_{n-I_n}+1, \tag{23}$$

其中， I_n 在区间 $[1,n-1]$ 上服从概率分布，概率为 $\mathbb{P}[I_n=j]=C_{j-1}C_{n-1-j}/C_{n-1}$ （分别为 $\mathbb{P}[I_n=j]=\frac{1}{n-1}$）；对于每个 $j\in[1,n-1]$， R^*_j和 T^*_j分别是 R_j和 T_j的独立副本；并且当 $j\in[1,n-1]$时， R_j与R^*_j、 T_j与 T^*_j均独立于 I_j。

证明。 分布递推关系可直接由方程 (12) 和 (13) 得出。 $\mathbb{P}[I_n=j]$ 遵循方程 (18) 和 (21)，给出了从 n 个分类单元中均匀随机选取的有序无标号拓扑或历史的左根子树大小为 $I_n=j$ 的概率。 □

3.3 关于祖先构型分布的已知结果 s

对于随机变量 R_n 和 T_n， $\mathbb{E}[R_n]$、 $\mathbb{E}[R_n^2]$和 $\mathbb{E}[T_n]$的矩以及 $\mathbb{V}[R_n]$ 和 $\mathbb{V}[T_n]$ 的方差在带标签拓扑的均匀模型下的渐近行为由 [11] 研究（命题5和命题6），在 Yule 模型下则由 [7] 研究（命题5.4和命题5.5）：

$$\mathbb{E}[T_n] \bowtie \mathbb{E}[R_n] \sim \begin{cases} \sqrt{\frac{3}{2}} \left(\frac{4}{3}\right)^n, & \text{Uniform model,} \\ \left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n, & \text{Yule model,} \end{cases} \tag{24}$$

$$\mathbb{V}[T_n] \bowtie \mathbb{V}[R_n] \sim \mathbb{E}[R_n^2] \sim \begin{cases} \sqrt{\frac{7(11-\sqrt{2})}{34}} \left[\frac{4}{7(8\sqrt{2}-11)}\right]^n, & \text{Uniform model,} \\ (2.0449954971\ldots)^n, & \text{Yule model.} \end{cases} \tag{25}$$

Yule模型下 $\mathbb{V}[T_n]$ 的指数阶2.0449954971 . . . 是通过一种数值程序近似得到的，该程序见 [7]的附录。

研究表明，在均匀模型和Yule模型下，随机选取的大小为 n 的带标签拓扑的根构型数目的对数渐近服从正态分布。 [7] 中的命题4.1和命题5.2指出，重新缩放后的随机变量

$$\frac{\log R_n - \mathbb{E}[\log R_n]}{\sqrt{\mathbb{V}[\log R_n]}}$$

收敛 到标准正态分布，其中 $\mathbb{E}[\log R_n] \sim \mu n$、 $\mathbb{V}[\log R_n] \sim \sigma^2 n$，且

$$(\mu, \sigma^2) \approx \begin{cases} (0.272, 0.034), & \text{Uniform model,} \\ (0.351, 0.008), & \text{Yule model.} \end{cases} \tag{26}$$

4 祖先构型总数的分布性质

以均匀分布和 Yule 分布选取带标签拓扑时，关于祖先构型的既有研究主要关注根构型的数量。现在，我们研究同样两种概率模型下随机祖先构型总数的情况。具体而言，我们确定其均值和方差的渐近增长。与方程（24）和（25）一致，我们发现，构型总数的均值和方差与根构型数量的均值和方差仅在次指数项上有所不同，而这些次指数项最终为常数。此外，我们发现，与根构型数量的情形相同，构型总数渐近服从对数正态分布。

4.1 有序无标号拓扑的均匀分布与带标号拓扑的均匀分布

根据引理1，在大小为n 的随机带标号拓扑中均匀随机选取拓扑时，祖先构型数量的分布与大小为 n的有序无标号拓扑均匀分布时祖先构型数量的分布相同。我们利用这一等价关系推导本节的结果，并分别用 R_n 和 T_n表示在均匀分布下选取的、大小为 n 的随机有序无标记拓扑中的根祖先构型数量和祖先构型总数。

我们的第一个命题利用引理2中的分布递推关系，确定了 T_n均值的渐近行为。

命题3 在大小为 n、均匀随机选取的有序无标号拓扑中，祖先构型总数的均值满足渐近关系 $\mathbb{E}[T_n] \sim 2\mathbb{E}[R_n] \sim \sqrt{6}(4/3)^n$。

证明。结合引理2中的公式（23）与 $\mathbb{E}[R_{I_n}\, R^*_{n-I_n}] = \sum_{j=1}^{n-1}\mathbb{P}[I_n = j]\,\mathbb{E}[R_j\, R^*_{n-j}] = \sum_{j=1}^{n-1}\mathbb{P}[I_n = j]\,\mathbb{E}[R_j]\,\mathbb{E}[R^*_{n-j}]$，我们得到，对于 $n \geqslant 1$， R_n 的期望满足

$$C_{n-1}\mathbb{E}[R_n] = \sum_{j=1}^{n-1} C_{j-1}\, C_{n-1-j}\,\Big(\mathbb{E}[R_j]\,\mathbb{E}[R_{n-j}] + \mathbb{E}[R_j] + \mathbb{E}[R_{n-j}] + 1\Big), \tag{27}$$

这对 $n = 1$ 作为 $\mathbb{E}[R_1] = 0$也成立。类似地，对于 $\mathbb{E}[T_1] = 0$，对于 $n \geqslant 1$，引理2中的等式 (22) 给出

$$C_{n-1}\mathbb{E}[T_n] = 2\left(\sum_{j=1}^{n-1} C_{j-1}C_{n-1-j}\mathbb{E}[T_j]\right) + C_{n-1}\mathbb{E}[R_n]. \tag{28}$$

定义生成函数

$$R(z) \equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n \tag{29}$$

$$T(z) \equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n]z^n, \tag{30}$$

其系数 $C_{n-1}\mathbb{E}[R_n] = C_{n-1}\sum_{i=0}^{\infty}|\{t:\, c_r(t) = i\}|i/C_{n-1}$ 和 $C_{n-1}\mathbb{E}[T_n] = C_{n-1}\sum_{i=0}^{\infty}|\{t:\, c(t) = i\}|i/C_{n-1}$分别给出了所有 n 个分类单元的有序无标号拓扑 t 中根构型数量之和以及构型总数之和。

方程 (27) 和 (28) 中的递推关系转化为关于 $R(z)$ 和 $T(z)$ 的方程组：

$$S_1 \equiv \begin{cases} R(z) = R(z)^2 + 2zC(z)\,R(z) + z^2C(z)^2 \\ T(z) = 2zC(z)\,T(z) + R(z), \end{cases}$$

其中， $C(z)$ 是 C atalan 生成函数（公式 17）。事实上，将公式（27）乘以 z^n，我们得到 e

$$\begin{aligned} C_{n-1}\mathbb{E}[R_n]z^n &= \sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[R_j]z^j \cdot C_{n-1-j}\mathbb{E}[R_{n-j}]z^{n-j} + z\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[R_j]z^j \cdot C_{n-1-j}z^{n-1-j} \\ &\quad + z\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}\mathbb{E}[R_{n-j}]z^{n-j} + z^2\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}z^{n-1-j}. \end{aligned}$$

将 S_1 对 $n \geqslant 1$ 求和并化简，即可得到第一个方程：

$$\begin{aligned} R(z) &= \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n = \sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[R_j]z^j \cdot C_{n-1-j}\mathbb{E}[R_{n-j}]z^{n-j} + z\sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[R_j]z^j \cdot C_{n-1-j}z^{n-1-j} \\ &\quad + z\sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}\mathbb{E}[R_{n-j}]z^{n-j} + z^2\sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}z^{j-1} \cdot C_{n-1-j}z^{n-1-j}. \end{aligned}$$

同样，由方程（28）可得 S_1 的第二个方程：

$$T(z) = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n]z^n = 2z\sum_{n=1}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[T_j]z^j \cdot C_{n-1-j}z^{n-1-j} + \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n.$$

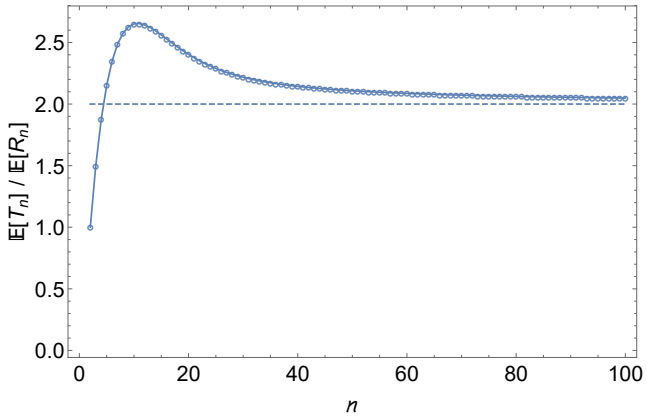

图4：对于大小为 $2 \leqslant n \leqslant 100$ 的均匀分布的有序无标号拓扑（或均匀分布的带标号拓扑），祖先构型总数的均值 $\mathbb{E}[T_n]$ 与根构型的平均数量 $\mathbb{E}[R_n]$ 之比。$\mathbb{E}[R_n]$ 和 $\mathbb{E}[T_n]$ 的值分别根据方程（27）和（28）中的递推关系计算得出。

解方程组 S_1，求得 $T(z)$:

$$T(z) = \frac{R(z)}{1-2zC(z)} = \frac{\sqrt{1-4z} - \sqrt{2\sqrt{1-4z}-1}}{2\sqrt{1-4z}}, \tag{31}$$

其具有主导奇点 $\alpha \equiv \frac{3}{16}$—— $2\sqrt{1-4z}-1$ 的根。奇异展开为

$$T(z) \overset{z\to\alpha}{\sim} k_1 - \sqrt{\frac{3}{2}} \cdot \sqrt{1-\frac{16z}{3}},$$

其中 k_1 是某个常数。因此，由方程（6）可得

$$[z^n]T(z) \sim \sqrt{\frac{3}{2}} \frac{(16/3)^n}{2\sqrt{\pi n^3}}.$$

利用 $C_{n-1} \sim 4^{n-1}/\sqrt{\pi n^3}$ （方程 16）这一事实，我们得到

$$\mathbb{E}[T_n] = \frac{[z^n]T(z)}{C_{n-1}} \sim \frac{\sqrt{\frac{3}{2}}\frac{(16/3)^n}{2\sqrt{\pi n^3}}}{\frac{4^{n-1}}{\sqrt{\pi n^3}}} = \sqrt{6}\left(\frac{4}{3}\right)^n,$$

这是方程（24）中给出的均匀情形下 $\mathbb{E}[R_n]$ 的渐近值的两倍。 □

图4绘制了 $\mathbb{E}[T_n]/\mathbb{E}[R_n]$ 随 n增加而变化的精确比值。与命题3一致，随着 n 增加，比值$\mathbb{E}[T_n]/\mathbb{E}[R_n]$ 趋近于2。

现在，我们考察固定大小为 n的均匀分布的有序无标号拓扑中，祖先构型总数的方差 $\mathbb{V}[T_n]$ 及其与根构型数 $\rho[T_n, R_n]$ 之间的相关系数。下一个引理给出了一系列分布递推关系。

引理4 考虑随机变量 $R_n \equiv R_n + 1$ 和 T_n。我们有 $R_1 = 1$、$T_1 = 0$，且对于 $n \geqslant 2$，

$$\tilde{R}_n \overset{d}{=} \tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + 1, \tag{32}$$

$$(\tilde{R}_n)^2 \overset{d}{=} (\tilde{R}_{I_n})^2(\tilde{R}^*_{n-I_n})^2 + 2\tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + 1, \tag{33}$$

$$T_n\tilde{R}_n \overset{d}{=} T_{I_n}\tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + T^*_{n-I_n}\tilde{R}^*_{n-I_n}\tilde{R}_{I_n} + T_{I_n} + T^*_{n-I_n} + (\tilde{R}_n)^2 - \tilde{R}_n, \tag{34}$$

$$(T_n)^2 \overset{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + 2T_{I_n}T^*_{n-I_n} + 2T_nR_n - (R_n)^2, \tag{35}$$

˜其中 I_n 以概率 $\mathbb{P}[I_n = j] = C_{j-1}C_{n-1-j}/C_{n-1}$、 R_j^*、 R_j^*分布于区间 $[1, n-1]$ 上，并且对于每个 $j \in [1, n-1]$，T_j^*是 R_j、 R_j和 T_j的独立副本，分别如此；此外，对于 $j \in [1, n-1]$， R_j和 R_j^*以及˜ ˜R_j、 R_j^*、 T_j和 T_j^*均与 I_j独立。

证明。由引理2中的式（23）可直接得到式（32）。将式（32）平方即可得到式（33）。对于式（34），将引理2中的式（22）与式（32）结合可得

$$T_n\tilde{R}_n = (T_{I_n} + T^*_{n-I_n} + R_n)\,\tilde{R}_n = (T_{I_n} + T^*_{n-I_n})\tilde{R}_n + R_n\tilde{R}_n = (T_{I_n} + T^*_{n-I_n})(\tilde{R}_{I_n}\tilde{R}^*_{n-I_n} + 1) + (\tilde{R}_n - 1)\tilde{R}_n.$$

最后，将引理2中的式（22）平方，可得

$$\begin{aligned}
(T_n)^2 &\stackrel{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + (R_n)^2 + 2T_{I_n}T^*_{n-I_n} + 2T_{I_n}R_n + 2T^*_{n-I_n}R_n \\
&\stackrel{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + 2T_{I_n}T^*_{n-I_n} + 2(R_n)^2 + 2T_{I_n}R_n + 2T^*_{n-I_n}R_n - (R_n)^2 \\
&\stackrel{d}{=} (T_{I_n})^2 + (T^*_{n-I_n})^2 + 2T_{I_n}T^*_{n-I_n} + 2R_n(R_n + T_{I_n} + T^*_{n-I_n}) - (R_n)^2,
\end{aligned}$$

由此得到式（35），因为 $R_n + T_{I_n} + T^*_{n-I_n} = T_n$ 再次由式（22）可得。 □

为对方差 $\mathbb{V}[T_n]$ 和相关性 $\rho[T_n, R_n]$进行渐近分析，我们现在确定期望 $\mathbb{E}[T_n^2]$和 $\mathbb{E}[T_nR_n]$的渐近行为。定义如下生成函数:

$$\tilde{R}(z) \equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n + \sum_{n=1}^{\infty} C_{n-1}z^n = R(z) + zC(z), \tag{36}$$

$$\tilde{S}(z) \equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2]z^n, \tag{37}$$

$$\begin{aligned}
S(z) &\equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n^2]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2 - 2\tilde{R}_n + 1]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2 - 2(R_n + 1) + 1]z^n \\
&= \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2]z^n - 2\sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[R_n]z^n - \sum_{n=1}^{\infty} C_{n-1}z^n = \tilde{S}(z) - 2R(z) - zC(z),
\end{aligned} \tag{38}$$

$$\tilde{V}(z) \equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n\tilde{R}_n]z^n, \tag{39}$$

$$\begin{aligned}
V(z) &\equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_nR_n]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n(\tilde{R}_n - 1)]z^n = \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n\tilde{R}_n]z^n - \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n]z^n \\
&= \tilde{V}(z) - T(z),
\end{aligned} \tag{40}$$

$$U(z) \equiv \sum_{n=1}^{\infty} C_{n-1}\mathbb{E}[T_n^2]z^n. \tag{41}$$

Here， $R(z)$ 来自式（29），$C(z)$ 由式（17）给出， $T(z)$ 由式（31）给出。

引理4中的分布递推关系确定了期望 $\mathbb{E}[\tilde{R}_n^2]$、 $\mathbb{E}[T_n\tilde{R}_n]$ 和$\mathbb{E}[T_n^2]$的递推关系，进而转化为一个函数方程组:

$$S_2 \equiv \begin{cases} \tilde{S}(z) - z = \tilde{S}(z)^2 + 2\tilde{R}(z)^2 + z^2C(z)^2 \\ \tilde{V}(z) = 2\tilde{V}(z)\,\tilde{R}(z) + 2zC(z)\,T(z) + \tilde{S}(z) - \tilde{R}(z) \\ U(z) = 2zC(z)\,U(z) + 2T(z)^2 + 2V(z) - S(z). \end{cases}$$

聚焦于 S_2 中的第一个方程，我们注意到方程（33）给出了 $n \geqslant 2$

$$C_{n-1}\mathbb{E}[\tilde{R}_n^2] = \sum_{j=1}^{n-1} C_{j-1}C_{n-1-j}\Big(\mathbb{E}[\tilde{R}_j^2]\,\mathbb{E}[\tilde{R}_{n-j}^2] + 2\mathbb{E}[\tilde{R}_j]\,\mathbb{E}[\tilde{R}_{n-j}] + 1\Big),$$

将其乘以 z^n 后可改写为

$$\begin{aligned} C_{n-1}\mathbb{E}[\tilde{R}_n^2]z^n &= \sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[\tilde{R}_j^2]z^j \cdot C_{n-1-j}\mathbb{E}[\tilde{R}_{n-j}^2]z^{n-j} + 2\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[\tilde{R}_j]z^j \cdot C_{n-1-j}\mathbb{E}[\tilde{R}_{n-j}]z^{n-j} \\ &\quad + z^2\sum_{j=1}^{n-1} C_{j-1}z^{j-1}\cdot C_{n-1-j}z^{n-1-j}. \end{aligned}$$

对 $n \geqslant 2$ 求和，得到

$$\begin{aligned} \tilde{S}(z) &= z + \sum_{n=2}^{\infty} C_{n-1}\mathbb{E}[\tilde{R}_n^2]z^n = z + \sum_{n=2}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[\tilde{R}_j^2]z^j \cdot C_{n-1-j}\mathbb{E}[\tilde{R}_{n-j}^2]z^{n-j} \\ &\quad + 2\sum_{n=2}^{\infty}\sum_{j=1}^{n-1} C_{j-1}\mathbb{E}[\tilde{R}_j]z^j \cdot C_{n-1-j}\mathbb{E}[\tilde{R}_{n-j}]z^{n-j} + z^2\sum_{n=2}^{\infty}\sum_{j=1}^{n-1} C_{j-1}z^{j-1}\cdot C_{n-1-j}z^{n-1-j} \\ &= z + \tilde{S}(z)^2 + 2\tilde{R}(z)^2 + z^2C(z)^2. \end{aligned}$$

同样， S_2 的第二个方程由方程（34）得出，而第三个方程由方程（35）得出。

为求解系统 S_2，我们首先根据命题3证明中的系统 S_1 的第一个方程求出 $\tilde{R}(z)$。然后利用方程（36）求出 $\tilde{R}(z)$。由系统 S_2 的第一个方程，我们可以得到 $\tilde{S}(z)$，并且还可由方程（38）得到 $\tilde{S}(z)$。随后，将 $\tilde{S}(z)$ 与 $T(z)$（方程31）结合，用于根据 S_2 的第二个方程计算 $V(z)$。得到 $V(z)$ 的公式后，我们根据方程（40）得到 $V(z)$，最后根据 S_2 的第三个方程计算 $U(z)$。令 $r \equiv \sqrt{1-4z}$，我们得到

$$V(z) = \frac{-\sqrt{2r-1} + r\left(-r + \sqrt{2r-1} - \sqrt{-2r+4\sqrt{2r-1}-1} + 3\right) - 1}{2r\sqrt{2r-1}}, \tag{42}$$

$$U(z) = \frac{1}{2}\left(-\frac{1}{r^3} + \frac{-\frac{2\sqrt{-2r+4\sqrt{2r-1}-1}}{\sqrt{2r-1}} + \sqrt{-2r+4\sqrt{2r-1}-1} + \frac{4}{\sqrt{2r-1}} + 3}{r} - \frac{6}{\sqrt{2r-1}} + 1\right). \tag{43}$$

生成函数 $V(z)$ 和 $U(z)$ 的主导奇点位于 $\alpha \equiv 7(8\sqrt{2}-11)/16$，这是方程 (42) 和 (43) 中出现的平方根 $\sqrt{-2r+4\sqrt{2r-1}-1}$ 的主导奇点。通过将展开式代入，我们得到 $V(z)$ 和 $U(z)$ 在其主导奇点 α 处的展开式

$$\sqrt{-2r+4\sqrt{2r-1}-1} \overset{z\to\alpha}{\sim} \sqrt{\frac{7(11-\sqrt{2})}{34}}\cdot\sqrt{1-\frac{16z}{7(8\sqrt{2}-11)}}$$

在方程（42）和（43）中，同时在其他位置令 $z=\alpha$。随后通过代数运算可得

$$V(z) \overset{z\to\alpha}{\sim} k_1 - \frac{1}{2}\left(1+\frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}}\cdot\sqrt{1-\frac{16z}{7(8\sqrt{2}-11)}},$$

$$U(z) \overset{z\to\alpha}{\sim} k_1' - \frac{1}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}\cdot\sqrt{1-\frac{16z}{7(8\sqrt{2}-11)}},$$

对于某些常数 k_1 和 k_1'。式（6）与式（16）中的渐近展开最终给出

$$\mathbb{E}[T_nR_n] = \frac{[z^n]V(z)}{C_{n-1}} \sim \left(1+\frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n, \tag{44}$$

$$\mathbb{E}[T_n^2] = \frac{[z^n]U(z)}{C_{n-1}} \sim \frac{2}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n. \tag{45}$$

通过这些计算，我们得到以下结果。

命题5 从大小为 n 的有序无标签拓扑中均匀随机选取一个拓扑时，其中祖先构型总数 T_n 的方差满足渐近关系 $\mathbb{V}[T_n] \sim \mathbb{E}[T_n^2]$，其中 $\mathbb{E}[T_n^2]$ 按公式（*45*）中的方式增长。随着 n 增大，祖先构型总数 T_n 与根构型数 R_n 之间的相关系数 $\rho[T_n, R_n]$ 收敛于常数 $\rho[T_n, R_n] \to 0.9004\ldots$.

证明。首先，对于方差，我们有 $\mathbb{V}[T_n] = \mathbb{E}[T_n^2] - \mathbb{E}[T_n]^2 \sim \mathbb{E}[T_n^2]$。事实上，根据命题3，$\mathbb{E}[T_n]^2 \bowtie [(4/3)^2]^n = (16/9)^n$；而根据公式（45），$\mathbb{E}[T_n^2] \bowtie \left[4/[7(8\sqrt{2}-11)]\right]^n$，且 $4/[7(8\sqrt{2}-11)] > \frac{16}{9}$。

其次，协方差 $\mathrm{Cov}[T_n, R_n]$ 的增长速度与 $\mathrm{Cov}[T_n, R_n] = \mathbb{E}[T_nR_n] - \mathbb{E}[T_n]\,\mathbb{E}[R_n] \sim \mathbb{E}[T_nR_n]$ 相同。事实上，根据命题3以及均匀模型下的方程（24），我们有 $\mathbb{E}[T_n]\,\mathbb{E}[R_n] \bowtie (4/3)^n \cdot (4/3)^n = (16/9)^n$；根据方程（44），我们有 $\mathbb{E}[T_nR_n] \bowtie \left[4/[7(8\sqrt{2}-11)]\right]^n$，其中 $4/[7(8\sqrt{2}-11)] > \frac{16}{9}$。因此，对于相关系数，

$$\rho[T_n, R_n] = \frac{\mathrm{Cov}[T_n, R_n]}{\sqrt{\mathbb{V}[T_n]}\sqrt{\mathbb{V}[R_n]}} \sim \frac{1+\frac{\sqrt{2}}{2}}{\sqrt{\frac{2}{17}(15+11\sqrt{2})}} \approx 0.9004,$$

其中，我们使用了公式（25）中给出的均匀模型的 $\mathbb{V}[R_n]$ 渐近公式，以及方程（44）和（45）中分别关于协方差 $[T_n, R_n]$ 和 $\mathbb{V}[T_n]$的渐近结果。 □

总之，本节将证明：从大小为 n 的有序无标号拓扑中均匀随机选取一个拓扑，其祖先构型总数 T_n 渐近服从对数正态分布。由引理1和第3.3节可知，从大小为 n 的有序无标号拓扑中均匀随机选取一个拓扑，其根构型数量 R_n 的对数渐近收敛于标准正态分布，即

$$\frac{\log R_n - \mathbb{E}[\log R_n]}{\sqrt{\mathbb{V}[\log R_n]}} \xrightarrow{d} N(0,1), \tag{46}$$

其中，$\mathbb{E}[\log R_n] \sim 0.272 \cdot n$ 和 $\mathbb{V}[\log R_n] \sim 0.034 \cdot n$（常数为近似值）。

根据式（14），在同一个大小为n的随机有序无标记拓扑上测得的变量 R_n 和 T_n 满足 $R_n \leqslant T_n \leqslant (2n-1)R_n$。该不等式给出 $\log T_n = \log R_n + \epsilon_n$，其中随机变量 ϵ_n 的取值为 $[0, \log(2n-1)]$。因此，我们有

$$\mathbb{E}[\log T_n] = \mathbb{E}[\log R_n] + \mathbb{E}[\epsilon_n] = \mathbb{E}[\log R_n] + \mathcal{O}(\log n) \sim \mathbb{E}[\log R_n] \tag{47}$$

$$\mathbb{V}[\log T_n] = \mathbb{V}[\log R_n] + \mathbb{V}[\epsilon_n] + 2\mathrm{Cov}[\log R_n, \epsilon_n] \sim \mathbb{V}[\log R_n], \tag{48}$$

其中，我们利用 $\mathbb{V}[\epsilon_n] \leqslant [\log(2n-1)]^2/4$ 由 Popoviciu不等式得出的有界随机变量的最大方差 [3]，，因此，与线性增长的 $\mathbb{V}[\log R_n]$ 相比，得到 $\lim_{n\to\infty} \mathbb{V}[\epsilon_n]/\mathbb{V}[\log R_n] = 0$；我们还利用 $\mathrm{Cov}[\log R_n, \epsilon_n] \leqslant \sqrt{\mathbb{V}[\log R_n]}\sqrt{\mathbb{V}[\epsilon_n]}$，这由柯西 - 施瓦茨不等式得出。

接下来，我们写出

$$\frac{\log T_n - \mathbb{E}[\log T_n]}{\sqrt{\mathbb{V}[\log T_n]}} = \frac{\log R_n - \mathbb{E}[\log R_n]}{\sqrt{\mathbb{V}[\log R_n]}} \cdot \frac{\sqrt{\mathbb{V}[\log R_n]}}{\sqrt{\mathbb{V}[\log T_n]}} + \frac{\epsilon_n - \mathbb{E}[\epsilon_n]}{\sqrt{\mathbb{V}[\log T_n]}}. \tag{49}$$

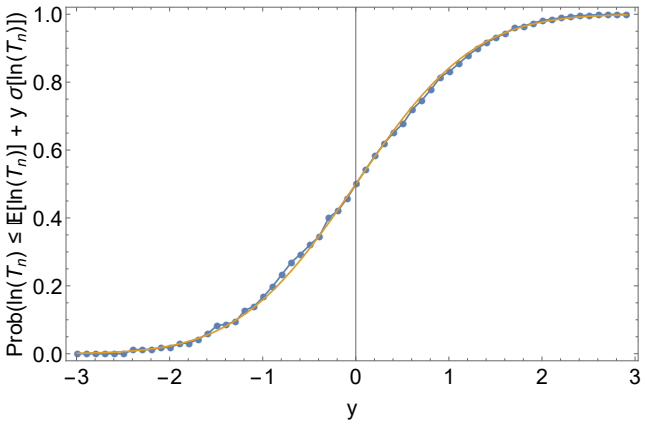

图5：对于大小为 $n = 15$ 的均匀分布的有序无标号拓扑（或均匀分布的带标号拓扑），总构型数的自然对数的累积分布（点线）。对于每个以0.1为步长递增的 $y \in [-3, 3]$，所绘数量是指：将 $n = 15$均匀随机选取时，一个有序无标号拓扑（或带标号拓扑）的总构型数小于或等于$\exp(\mathbb{E}[\log T_n] + y\sigma[\log T_n])$的概率，其中$\mathbb{E}[\log T_n]$ 和 $\sigma[\log T_n] = \sqrt{\mathbb{V}[\log T_n]}$ 分别是具有 $n = 15$ 个分类单元的均匀分布的有序无标号拓扑（或带标号拓扑）的总构型数的对数的均值和标准差。实线是均值为0、方差为1的高斯随机变量的累积分布。

表达式 $(\log R_n - \mathbb{E}[\log R_n])/\sqrt{\mathbb{V}[\log R_n]}$ 依分布收敛于均值为 0、方差为 1 的正态随机变量（公式 46）。比值 $\sqrt{\mathbb{V}[\log R_n]}/\sqrt{\mathbb{V}[\log T_n]}$ 是一个数列，根据公式（48）收敛于有限常数 1。表达式 $(\epsilon_n - \mathbb{E}[\epsilon_n])/\sqrt{\mathbb{V}[\log T_n]}$ 依均方收敛于 0，因为

$$\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{\epsilon_n - \mathbb{E}[\epsilon_n]}{\sqrt{\mathbb{V}[\log T_n]}} - 0\right)^2\right] = \lim_{n\to\infty} \frac{\mathbb{V}[\epsilon_n]}{\mathbb{V}[\log T_n]};$$

分母 $\mathbb{V}[\log T_n]$ 随 n 线性增加（第3.3节），同样根据Popoviciu不等式，分子被 $[\log(2n-1)]^2/4$从上方界定，因此 $\lim_{n\to\infty} \mathbb{V}[\epsilon_n]/\mathbb{V}[\log T_n] = 0$。

由于均方收敛蕴含依概率收敛（见 [29] 第 10 页），我们可以将斯卢茨基定理应用于式 (49)：该定理说明，依分布收敛的随机变量受到依概率收敛的随机变量扰动后仍依分布收敛（见 [29] 第 19 页）。具体而言，$(\log R_n - \mathbb{E}[\log R_n]) \times (\sqrt{\mathbb{V}[\log R_n]})^{-1}$ 依分布收敛，$\sqrt{\mathbb{V}[\log R_n]}/\sqrt{\mathbb{V}[\log T_n]}$ 平凡地依概率收敛，且 $(\epsilon_n - \mathbb{E}[\epsilon_n])/\sqrt{\mathbb{V}[\log T_n]}$ 依概率收敛，据此我们可以得出结论：$(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathbb{V}[\log T_n]}$ 依分布收敛于均值为 0、方差为 1 的正态随机变量。

图 5 显示了当以均匀随机方式选取大小为 15 的有序无标号拓扑时，累积分布 $\mathbb{P}[\log T_n \leqslant \mathbb{E}[\log T_n] + y\sqrt{\mathbb{V}[\log T_n]}]$ 随 y 的变化。为得到该分布，我们对每个具有 15 个叶节点的无标号拓扑统计其构型总数 t，然后统计具有 t 形状的有序无标号拓扑数量（公式 15）。该图展示了祖先构型的精确累积分布与标准正态分布之间的一致性。

根据引理1中所述的有序无标号拓扑与带标号拓扑之间的等价关系，我们可以将本节的主要结果表述如下。

定理 6 对于在均匀分布下随机选取的大小为 n 的带标签拓扑，总数为 T_n 的祖先构型的均值和方差渐近增长为

$$\mathbb{E}[T_n] \sim \sqrt{6}\left(\frac{4}{3}\right)^n, \tag{50}$$

$$\mathbb{V}[T_n] \sim \frac{2}{17}(15 + 11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n. \tag{51}$$

此外，大小为 n 且均匀随机选取的带标签拓扑中祖先构型总数的对数，经缩放后为 (对数 $T_n - \mathbb{E}[$对数 $T_n]$)/$\sqrt{}$ 方差[对数T_n]，收敛于标准正态分布，其中 $\mathbb{E}[$对数 $T_n] \sim \mu n$ 和 $\mathbb{V}[$对数 $T_n] \sim \sigma^2 n$，$(\mu, \sigma^2) \approx (0.272, 0.034)$。

4.2 均匀分布的有序无标记历史与尤尔带标签拓扑

根据引理1，在尤尔概率模型下，从大小为 n 的随机有标号拓扑中选取的祖先构型总数的分布，与大小为 n 的均匀分布的有序无标记历史中构型总数的分布相同。本节将利用这一等价关系，并分别用 R_n 和 T_n 表示均匀分布下大小为 n 的随机有序无标记历史中的根祖先构型数和祖先构型总数。

引理7 考虑随机变量 R_n 和 T_n。我们有 $R_1 = T_1 = 0$，且对于 $n \geqslant 2$，

$$
\begin{aligned}
R_n &\stackrel{d}{=} R_{I_n} R^*_{n-I_n} + R_{I_n} + R^*_{n-I_n} + 1, && (52)\\
T_n &\stackrel{d}{=} T_{I_n} + T^*_{n-I_n} + R_n, && (53)\\
T_n R_n &\stackrel{d}{=} T_{I_n} R_{I_n} R^*_{n-I_n} + T_{I_n} R_{I_n} + T_{I_n} R^*_{n-I_n} + T_{I_n} \\
&\quad + T^*_{n-I_n} R_{I_n} R^*_{n-I_n} + T^*_{n-I_n} R_{I_n} + T^*_{n-I_n} R^*_{n-I_n} + T^*_{n-I_n} + (R_n)^2, && (54)\\
(T_n)^2 &\stackrel{d}{=} (T_{I_n})^2 + (T^*_{I_{n-I_n}})^2 + 2T_{I_n} T^*_{n-I_n} + 2T_n R_n - (R_n)^2, && (55)
\end{aligned}
$$

其中，I_n 是在区间 $[1, n-1]$ 上均匀分布的变量，R^*_j 和 T^*_j 分别是 R_j 和 T_j 的独立副本，对任意 $j \in [1, n-1]$ 均成立；R_j 和 R^*_j 以及 T_j 与 T^*_j 均彼此独立，并且当 $j \in [1, n-1]$ 时，它们与 I_j 独立。

证明。方程（52）和（53）来自引理2。展开 $T_n R_n \stackrel{d}{=} (T_{I_n} + T^*_{n-I_n})(R_{I_n} + 1)(R^*_{n-I_n} + 1) + (R_n)^2$,后可得方程（54）。最后，对方程（53）两边平方可得方程（55）；它也复用了引理4中的方程（35）。 □

引理7中的分布递推关系可用于确定期望 $\mathbb{E}[R_n]$、$\mathbb{E}[T_n]$、$\mathbb{E}[T_n R_n]$ 和 $\mathbb{E}[T_n^2]$ 的递推关系。对于 $n \geqslant 1$，我们可以写出

$$
\begin{aligned}
(n-1)\mathbb{E}[R_n] &= \left(\sum_{j=1}^{n-1} \mathbb{E}[R_j]\,\mathbb{E}[R_{n-j}]\right) + 2\left(\sum_{j=1}^{n-1} \mathbb{E}[R_j]\right) + (n-1), && (56)\\
(n-1)\mathbb{E}[T_n] &= 2\left(\sum_{j=1}^{n-1} \mathbb{E}[T_j]\right) + (n-1)\mathbb{E}[R_n], && (57)\\
(n-1)\mathbb{E}[T_n R_n] &= 2\left(\sum_{j=1}^{n-1} \mathbb{E}[T_j R_j]\,\mathbb{E}[R_{n-j}]\right) + 2\left(\sum_{j=1}^{n-1} \mathbb{E}[T_j R_j]\right) + 2\left(\sum_{j=1}^{n-1} \mathbb{E}[T_j]\,\mathbb{E}[R_{n-j}]\right) \\
&\quad + 2\left(\sum_{j=1}^{n-1} \mathbb{E}[T_j]\right) + (n-1)\mathbb{E}[R_n^2], && (58)\\
(n-1)\mathbb{E}[T_n^2] &= 2\left(\sum_{j=1}^{n-1} \mathbb{E}[T_j^2]\right) + 2\left(\sum_{j=1}^{n-1} \mathbb{E}[T_j]\,\mathbb{E}[T_{n-j}]\right) + 2(n-1)\mathbb{E}[T_n R_n] - (n-1)\mathbb{E}[R_n^2]. && (59)
\end{aligned}
$$

定义以下生成函数

$$R(z) \equiv \sum_{n=1}^{\infty}\mathbb{E}[R_n]z^n, \tag{60}$$

$$T(z) \equiv \sum_{n=1}^{\infty}\mathbb{E}[T_n]z^n, \tag{61}$$

$$S(z) \equiv \sum_{n=1}^{\infty}\mathbb{E}[R_n^2]z^n, \tag{62}$$

$$V(z) \equiv \sum_{n=1}^{\infty}\mathbb{E}[T_nR_n]z^n, \tag{63}$$

$$U(z) \equiv \sum_{n=1}^{\infty}\mathbb{E}[T_n^2]z^n. \tag{64}$$

方程 (57)、(58) 和 (59) 中的递推关系可转化为一个微分方程组：

$$S_3 \equiv \begin{cases} T'(z)-\frac{z+1}{z-z^2}T(z)=R'(z)-\frac{R(z)}{z} \\ V'(z)-\left(\frac{2R(z)}{z}+\frac{z+1}{z-z^2}\right)V(z)=\frac{2T(z)\,R(z)+\frac{2z}{1-z}T(z)+zS'(z)-S(z)}{z} \\ U'(z)-\frac{z+1}{z-z^2}U(z)=\frac{2T(z)^2+2zV'(z)-2V(z)-zS'(z)+S(z)}{z}, \end{cases}$$

由于方程 (57)、(58) 和 (59) 中的因子 $n-1$，导数 $R'(z)$、$T'(z)$、$S'(z)$、$V'(z)$ 和 $U'(z)$ 会出现在 S_3 中。下面以 S_3 中的第三个方程为例进行推导。首先，将方程 (59) 两边乘以 z^n，得到

$$\begin{aligned} zn\mathbb{E}[T_n^2]z^{n-1}-\mathbb{E}[T_n^2]z^n &= 2\sum_{j=1}^{n-1}\mathbb{E}[T_j^2]z^j\cdot z^{n-j}+2\sum_{j=1}^{n-1}\mathbb{E}[T_j]z^j\cdot\mathbb{E}[T_{n-j}]z^{n-j}+2zn\mathbb{E}[T_nR_n]z^{n-1} \\ &\quad -2\mathbb{E}[T_nR_n]z^n-zn\mathbb{E}[R_n^2]z^{n-1}+\mathbb{E}[R_n^2]z^n. \end{aligned}$$

对 $n\geqslant 1$ 求和，得到

$$\begin{aligned} z\sum_{n=1}^{\infty}n\mathbb{E}[T_n^2]z^{n-1}-\sum_{n=1}^{\infty}\mathbb{E}[T_n^2]z^n &= 2\sum_{n=1}^{\infty}\sum_{j=1}^{n-1}\mathbb{E}[T_j^2]z^j\cdot z^{n-j}+2\sum_{n=1}^{\infty}\sum_{j=1}^{n-1}\mathbb{E}[T_j]z^j\cdot\mathbb{E}[T_{n-j}]z^{n-j} \\ &\quad +2z\sum_{n=1}^{\infty}n\mathbb{E}[T_nR_n]z^{n-1}-2\sum_{n=1}^{\infty}\mathbb{E}[T_nR_n]z^n-z\sum_{n=1}^{\infty}n\mathbb{E}[R_n^2]z^{n-1}+\sum_{n=1}^{\infty}\mathbb{E}[R_n^2]z^n. \end{aligned}$$

为完成推导，我们注意到该方程可以改写为：

$$zU'(z)-U(z)=2U(z)\left(\frac{z}{1-z}\right)+2T(z)^2+2zV'(z)-2V(z)-zS'(z)+S(z),$$

因为 $U'(z)=\left(\sum_{n=1}^{\infty}\mathbb{E}[T_n^2]z^n\right)'=\sum_{n=1}^{\infty}n\mathbb{E}[T_n^2]z^{n-1}$、$V'(z)=\left(\sum_{n=1}^{\infty}\mathbb{E}[T_nR_n]z^n\right)'=\sum_{n=1}^{\infty}n\mathbb{E}[T_nR_n]z^{n-1}$、$S'(z)=\left(\sum_{n=1}^{\infty}\mathbb{E}[R_n^2]z^n\right)'=\sum_{n=1}^{\infty}n\mathbb{E}[R_n^2]z^{n-1}$，以及 $\frac{z}{1-z}=\sum_{n=1}^{\infty}z^n$。

我们还注意到，在 [7] 的第 5 节中，对带标号拓扑的 Yule 模型下根构型的分析研究了生成函数 $R(z)$ 和 $S(z)$（方程 60 和 62）。特别是，[7] 中命题 5.3 的证明里的方程 (39) 发现，$R(z)$——在该文中记作 $E(z)$——具有显式形式

$$R(z)=\frac{2z\sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right)}{(z-1)\left[\sqrt{3}\cos\left(\frac{\sqrt{3}}{2}\log(1-z)\right)+\sin\left(\frac{\sqrt{3}}{2}\log(1-z)\right)\right]}. \tag{65}$$

主导奇点为 $\alpha_1 \equiv 1 - e^{-2\pi\sqrt{3/9}}$，在主导奇点处的奇异展开为

$$R(z) \overset{z\to\alpha_1}{\sim} \frac{1}{1-\frac{z}{\alpha_1}}. \tag{66}$$

生成函数 $S(z)$ 在 [7] 第 5.3 节中被证明具有如下奇异展开：

$$S(z) \overset{z\to\alpha_2}{\sim} \frac{1}{1-\frac{z}{\alpha_2}}, \tag{67}$$

其中，主导奇点 $\alpha_2 \equiv 0.4889986317\ldots$ 已在附录中进行了近似。根据奇点分析（公式 5），方程（66）和（67）中的展开给出方程（24）和（25）中的渐近关系：

$$\mathbb{E}[R_n] \sim \alpha_1^{-n} \quad \text{and} \quad \mathbb{E}[R_n^2] \sim \alpha_2^{-n}. \tag{68}$$

请注意，公式（25）中出现的渐近常数 2.0449954971 ... 是通过 α_2^{-1} 得到的。

现在我们注意到，方程（66）和 S_3 的第一个方程给出了在大小为 n 的有序无标记历史中均匀随机选取时，祖先构型平均数量 $\mathbb{E}[T_n]$ 的渐近增长。

命题 8 在大小为 n 的有序无标记历史中均匀随机选取时，祖先构型总数的均值满足渐近关系 $\mathbb{E}[T_n] \sim \mathbb{E}[R_n] \sim \alpha_1^{-n} = [1/(1-e^{-2\pi\sqrt{3/9}})]^n$.

证明。我们首先将 S_3 的第一个方程改写为

$$T'(z)\,M(z) - \frac{z+1}{z-z^2}M(z)\,T(z) = \left[R'(z) - \frac{R(z)}{z}\right]M(z), \tag{69}$$

其中，$M(z) \equiv (z-1)^2/z$ 是积分因子。

由于 $M'(z) = -\frac{z+1}{z-z^2}M(z)$，方程 (69) 的左侧可改写为 $[T(z)\ M(z)]'$，从而得到

$$\left[T(z)\frac{(z-1)^2}{z}\right]' = \left[R'(z) - \frac{R(z)}{z}\right]\frac{(z-1)^2}{z}. \tag{70}$$

因为 $T_1 = 0$ ，$T(z)$ 的展开以一个非零二次项开头。因此，我们有

$$\left[T(z)\frac{(z-1)^2}{z}\right]_{z=0} = 0,$$

因此，式（70）中的微分方程给出 $T(z)^{\frac{(z-1)^2}{z}} = \int_0^z [R'(t) - R(t)/t]^{\frac{(t-1)^2}{t}}\,dt$，即，

$$T(z) = \frac{z}{(z-1)^2}\int_0^z \left[R'(t) - \frac{R(t)}{t}\right]\frac{(t-1)^2}{t}\,dt.$$

为了得到 $T(z)$ 的奇异展开，我们必须分析函数 $R'(t)$、$[R'(t)-R(t)/t]^{\frac{(t-1)^2}{t}}$ 和 $\int_0^z [R'(t)-R(t)/t]^{\frac{(t-1)^2}{t}}\,dt$ 在其主导奇点处的情况。由于 $\alpha_1 = 1-e^{-2\pi\sqrt{3/9}}$ 是 $R(t)$ 和 $R(t) \overset{t\to\alpha_1}{\sim} 1/(1-\frac{t}{\alpha_1})$（式 66）的主导奇点，根据式（7），$R'(t)$ 的主导奇点为 $t = \alpha_1$。其奇异展开为

$$R'(t) \overset{t\to\alpha_1}{\sim} \frac{1}{\alpha_1(1-\frac{t}{\alpha_1})^2},$$

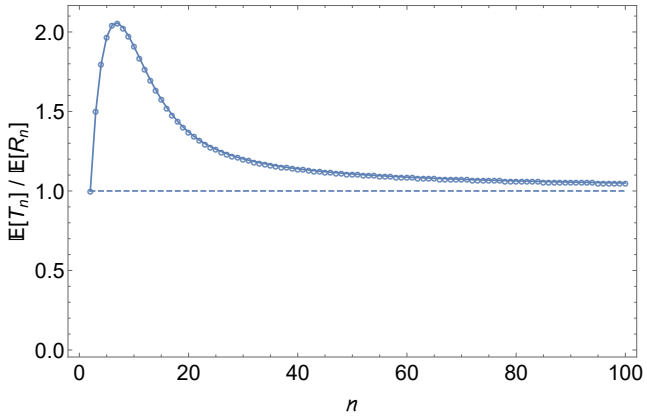

图6：对于大小为 $2 \leqslant n \leqslant 100$的均匀分布的有序无标记历史（或 Yule 分布的带标记拓扑），祖先构型总数的均值 $\mathbb{E}[T_n]$ 与根构型的平均数量 $\mathbb{E}[R_n]$ 之比。$\mathbb{E}[R_n]$ 和 $\mathbb{E}[T_n]$ 的值分别根据方程（56）和（57）中的递推关系计算得出。

通过对 $R(t)$ 的展开求导即可得到。由此可知， α_1 也是函数 $[R'(t) - R(t)/t]^{\frac{(t-1)^2}{t}}$ 的主导奇点，其奇异展开如下

$$\left[R'(t) - \frac{R(t)}{t}\right] \frac{(t-1)^2}{t} \overset{t\to\alpha_1}{\sim} \left[\frac{1}{\alpha_1(1-\frac{t}{\alpha_1})^2}\right] \frac{(\alpha_1-1)^2}{\alpha_1}. \tag{71}$$

最后，根据方程（8），通过对被积函数的奇异展开（方程 71）进行积分，可以在主导奇点 α_1 处展开 $\int_0^z [R'(t) - R(t)/t]^{\frac{(t-1)^2}{t}} dt$。因此， $T(z)$ 在其主导奇点 α_1 处的展开满足

$$T(z) \overset{z\to\alpha_1}{\sim} \frac{\alpha_1}{(\alpha_1-1)^2} \int_0^z \frac{1}{\alpha_1(1-\frac{t}{\alpha_1})^2} \frac{(\alpha_1-1)^2}{\alpha_1} dt = \frac{1}{1-\frac{z}{\alpha_1}} - 1 \overset{z\to\alpha_1}{\sim} \frac{1}{1-\frac{z}{\alpha_1}}.$$

根据方程（5）和（68），我们得出

$$\mathbb{E}[T_n] = [z^n]T(z) \sim \alpha_1^{-n} \sim \mathbb{E}[R_n]. \quad \square$$

图6展示了比值 $\mathbb{E}[T_n]/\mathbb{E}[R_n]$ 随 n变化的数值图。根据该命题，随着 n 增加，数值比值趋近于1。

现在我们研究固定大小的均匀分布的有序无标号历史中，祖先构型总数的方差 $\mathbb{V}[T_n]$ ，以及总构型数与根构型数之间的相关系数 $\rho[T_n, R_n]$ 。再次利用奇异展开在微分与积分下的性质，我们确定期望 $\mathbb{E}[T_n R_n]$ 和 $\mathbb{E}[T_n^2]$ 的渐近增长。我们从 $\mathbb{E}[T_n R_n]$ 开始。我们将 S_3 的第二个方程中的一项简记为 $P(z)$:

$$P(z) \equiv \frac{2T(z)R(z) + \frac{2z}{1-z}T(z) + zS'(z) - S(z)}{z}.$$

S_3的第二个方程变为 $V'(z) - V(z)\left[2R(z) + \frac{z+1}{z-z^2}\right] = P(z)$。我们引入积分因子 $M(z)$,

$$M(z) \equiv \left[e^{\int_0^z -\frac{2R(t)}{t} dt}\right] \cdot \frac{(z-1)^2}{z},$$

使得 $M'(z) = -[\frac{2R}{(z}\frac{z}{} + \frac{z+1}{z-z^2}]M)(z)$。我们求得 $[V(z)\ M(z)]' = P(z)\ M(z)$，因此

$$V(z) = \frac{1}{M(z)} \int_0^z P(t)\, M(t)\, dt.$$

为确定 $V(z)$ 在其主导奇点处的奇异展开，我们注意到 $P(t)\ M(t)$ 是 $T(t)$、$R(t)$、$S(t)$ 和 $S'(t)$ 的函数。如命题 8 所示，$T(t)$ 和 $R(t)$ 具有相同的主导奇点 $\alpha_1 \approx 0.702$ （式 66），其值大于 $S(t)$ 和 $S'(t)$ 的主导奇点 $\alpha_2 \approx 0.489$ （式 67）。因此，$R(t)$ 和 $T(t)$ 在包含 α_2 的 0 的某个邻域（记为 $|t| \leqslant \frac{1}{2}$）内是解析函数。由此，我们可以通过将 $P(t)\ M(t)$ 中的 $S(t)$ 和 $S'(t)$ 分别替换为其展开式 $S(t) \overset{t\to\alpha_2}{\sim} 1/(1 - \frac{t}{\alpha_2})$（式 67）和 $S'(t) \overset{t\to\alpha_2}{\sim} [1/(1 - \frac{t}{\alpha_2})]' = 1/[\alpha_2(1 - \frac{t}{\alpha_2})^2]$，并在其他位置代入 $t = \alpha_2$，得到其在主导奇点 α_2 处的奇异展开。我们得到

$$P(t)\, M(t) \overset{t\to\alpha_2}{\sim} \left[\frac{2T(\alpha_2)\, R(\alpha_2) + \frac{2\alpha_2}{1-\alpha_2} T(\alpha_2)}{\alpha_2} + \frac{1}{\alpha_2(1 - \frac{t}{\alpha_2})^2}\right] M(\alpha_2) \overset{t\to\alpha_2}{\sim} \frac{M(\alpha_2)}{\alpha_2(1 - \frac{t}{\alpha_2})^2}.$$

因此，对积分后的奇异展开（方程 8）可得 $\int_0^z P(t)\ M(t)\ dt \overset{z\to\alpha_2}{\sim} \int_0^z M(\alpha_2)/[\alpha_2(1 - \frac{t}{\alpha_2})^2]\, dt \overset{z\to\alpha_2}{\sim} M(\alpha_2)/(1 - \frac{z}{\alpha_2})$，由此得到 $V(z)$ 在其主导奇点 α_2 处的奇异展开

$$V(z) \overset{z\to\alpha_2}{\sim} \frac{1}{M(\alpha_2)} \frac{M(\alpha_2)}{1 - \frac{z}{\alpha_2}} = \frac{1}{1 - \frac{z}{\alpha_2}}.$$

因此，结合方程 (5) 和方程 (68)，可得

$$\mathbb{E}[T_n R_n] = [z^n]V(z) \sim \alpha_2^{-n} \sim \mathbb{E}[R_n^2]. \tag{72}$$

我们采用相同的方法来确定 $\mathbb{E}[T_n^2]$的渐近增长。将 S_3 的第三个方程两边乘以证明命题 8 时所用的积分因子 $M(z) \equiv (z-1)^2/z$，可得

$$U(z) = \frac{z}{(z-1)^2} \int_0^z \left[\frac{2T(t)^2}{t} + 2\left(V'(t) - \frac{V(t)}{t}\right) - \left(S'(t) - \frac{S(t)}{t}\right)\right] \frac{(t-1)^2}{t}\, dt.$$

我们简记

$$G(t) \equiv \frac{2T(t)^2}{t} + 2\left(V'(t) - \frac{V(t)}{t}\right) - \left(S'(t) - \frac{S(t)}{t}\right).$$

奇异展开 $V(t) \overset{t\to\alpha_2}{\sim} 1/(1 - \frac{t}{\alpha_2})$、$V'(t) \overset{t\to\alpha_2}{\sim} 1/[\alpha_2(1 - \frac{t}{\alpha_2})^2]$、$S(t) \overset{t\to\alpha_2}{\sim} 1/(1 - \frac{t}{\alpha_2})$ 和 $S'(t) \overset{t\to\alpha_2}{\sim} 1/[\alpha_2(1 - \frac{t}{\alpha_2})^2]$ 给出如下展开

$$G(t) \overset{t\to\alpha_2}{\sim} \frac{1}{\alpha_2(1 - \frac{t}{\alpha_2})^2}.$$

因此，在其主导奇点处，α_2, $U(z)$ 满足

$$U(z) = \frac{z}{(z-1)^2} \int_0^z G(t) \frac{(t-1)^2}{t}\, dt \overset{z\to\alpha_2}{\sim} \frac{\alpha_2}{(\alpha_2 - 1)^2} \int_0^z \frac{1}{\alpha_2(1 - \frac{t}{\alpha_2})^2} \frac{(\alpha_2 - 1)^2}{\alpha_2}\, dt \overset{z\to\alpha_2}{\sim} \frac{1}{1 - \frac{z}{\alpha_2}}.$$

结合式（5）和式（72），最终可得

$$\mathbb{E}[T_n^2] = [z^n]U(z) \sim \alpha_2^{-n} \sim \mathbb{E}[T_n R_n] \sim \mathbb{E}[R_n^2]. \tag{73}$$

根据这些计算，我们得到以下结果。

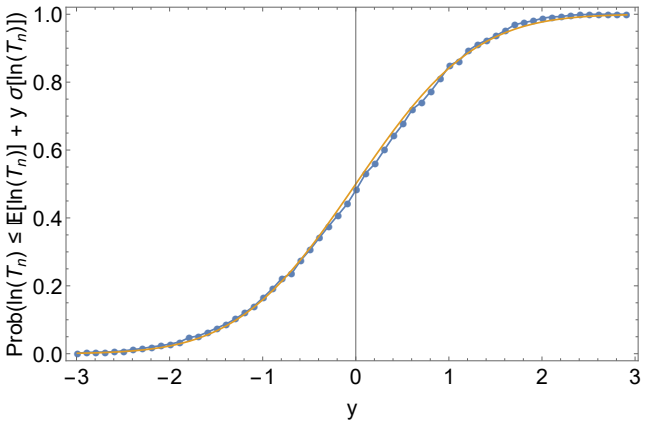

图 7：对于大小为 $n = 15$ 的均匀分布的有序无标记历史（或 Yule 分布的带标号拓扑），其构型总数的自然对数的累积分布（虚线）。对于每个以 0.1 为步长的 $y \in [-3, 3]$，所绘制的数量是随机选取的、具有 $n = 15$ 的有序无标记历史（或 Yule 分布的带标号拓扑）的构型总数小于或等于 $\exp(\mathbb{E}[\log T_n] + y\sigma[\log T_n])$ 的概率，其中，$\mathbb{E}[\log T_n]$ 和 $\sigma[\log T_n] = \sqrt{\mathbb{V}[\log T_n]}$ 分别是具有 $n = 15$ 个分类单元的均匀分布的有序无标记历史（或 Yule 分布的带标号拓扑）的构型总数的对数的均值和标准差。实线表示均值为 0、方差为 1 的高斯随机变量的累积分布。

命题 9 从均匀分布中选取大小为 n 的有序无标记历史时，其中祖先构型总数 T_n 的方差满足渐近关系 $\mathbb{V}[T_n] \sim \mathbb{E}[T_n^2] \sim \mathbb{V}[R_n] \sim \alpha_2^{-n} = (2.0449954971\ldots)^n$。随着 n 增大，祖先构型总数 T_n 与根构型数 R_n 之间的相关系数 $\rho[T_n, R_n]$ 收敛于 1，$\rho[T_n, R_n] \to 1$。

证明。对于方差，我们观察到 $\mathbb{V}[T_n] = \mathbb{E}[T_n^2] - \mathbb{E}[T_n]^2 \sim \mathbb{E}[T_n^2]$。事实上，根据命题 8，我们得到 $\mathbb{E}[T_n]^2 \bowtie (\alpha_1^{-2})^n$，其中 $\alpha_1^{-2} < \alpha_2^{-1}$，其中 α_2^{-1}是 $\mathbb{E}[T_n^2]$的指数阶，如方程 (73) 所示。此外，$\mathbb{E}[T_n^2] \sim \mathbb{V}[R_n]$，因为 $\mathbb{E}[T_n^2] \sim \mathbb{E}[R_n^2] \sim \mathbb{V}[R_n]$ 由方程 (73) 和 (25) 得出。

同样，对于 T_n 和 R_n 之间的协方差，我们得到

$$\mathrm{Cov}[T_n, R_n] = \mathbb{E}[T_n R_n] - \mathbb{E}[T_n]\,\mathbb{E}[R_n] \sim \mathbb{E}[T_n R_n] \sim \mathbb{E}[T_n^2].$$

事实上，由命题 8 可得 $\mathbb{E}[T_n]\,\mathbb{E}[R_n] \bowtie (\alpha_1^{-2})^n$，而由公式 (73) 可得 $\mathbb{E}[T_n R_n] \sim \mathbb{E}[T_n^2] \bowtie \alpha_2^{-n}$，其中$\alpha_2^{-1} > \alpha_1^{-2}$。因此，$T_n$ 与 R_n 之间的相关系数为

$$\rho[T_n, R_n] = \frac{\mathrm{Cov}[T_n, R_n]}{\sqrt{\mathbb{V}[T_n]}\sqrt{\mathbb{V}[R_n]}} \sim \frac{\mathbb{E}[T_n^2]}{\mathbb{E}[T_n^2]} = 1. \quad \square$$

根据方程（47）、（48）和（49）中的同一论证，对于均匀分布的有序无标号历史，总数 T_n 祖先构型可证明渐近服从对数正态分布。特别是，对于大小为 n 的随机有序无标号历史，变量 $(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathbb{V}[\log T_n]}$ 和 $(\log R_n - \mathbb{E}[\log R_n])/\sqrt{\mathbb{V}[\log R_n]}$ 渐近收敛于正态分布，其中$\mathbb{E}[\log T_n] \sim \mathbb{E}[\log R_n] \sim 0.351n$ 和 $\mathbb{V}[\log T_n] \sim \mathbb{V}[\log R_n] \sim 0.008n$（方程26）。

图7中，我们绘出了累积分布 $\mathbb{P}[\log T_n \leqslant \mathbb{E}[\log T_n] + y\sqrt{\mathbb{V}[\log T_n]}]$ 关于 y的函数关系，其中大小为15的有序无标号历史是均匀随机选取的。为得到该分布，我们可以统计每种含有15个叶节点的无标签拓扑 t 对应的构型总数，然后统计树形为 t 的有序无标号历史的数量（方程19）。该图展示了 $(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathbb{V}[\log T_n]}$ 与正态分布之间的一致性。

根据引理1中均匀分布的有序无标记历史与 Yule 分布的带标号拓扑之间的等价关系，我们总结本节的结果。

定理10 对于在尤尔分布下随机选取的大小为 n 的带标签拓扑，祖先构型总数 T_n 的均值和方差渐近增长为

$$\mathbb{E}[T_n] \sim \left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n, \tag{74}$$

$$\mathbb{V}[T_n] \sim (2.0449954971\ldots)^n. \tag{75}$$

此外，随机选取大小为 n、服从 *Yule* 分布的带标号拓扑，其祖先构型总数的对数经标准化为 $(\log T_n - \mathbb{E}[\log T_n])/\sqrt{\mathrm{Var}[\log T_n]}$后，收敛于标准正态分布，其中 $\mathbb{E}[\log T_n] \sim \mu n$ 且 $\mathbb{V}[\log T_n] \sim \sigma^2 n$，$(\mu,\ \sigma^2) \approx (0.351,\ 0.008)$。

5 结论

对于随机选取的、大小为 n 且带标号拓扑匹配的基因树和物种树 t，在均匀概率模型和 Yule 概率模型下，我们研究了 t的祖先构型总数 T_n 的渐近分布。借助解析组合学的方法，我们扩展了 [7] 和 [11], 的结果；他们关注的是 t的根祖先构型数量 R_n。

我们发现，在均匀模型和 Yule 模型下，构型总数均服从渐近对数正态分布； [7]也证明了根构型数量具有这一性质。此外，在定理6和定理10中，我们证明，对于均匀分布的带标号拓扑，祖先构型总数的均值和方差分别以$\mathbb{E}[T_n] \sim \sqrt{6}(4/3)^n$ 和 $\mathbb{V}[T_n] \sim 5.050 \cdot 1.822^n$的速度增长；而在 Yule 概率模型下选取大小为 n 的带标号拓扑时，则分别以 $\mathbb{E}[T_n] \sim 1.425^n$ 和 $\mathbb{V}[T_n] \sim 2.045^n$的速度增长。具体而言，我们观察到，对于服从均匀分布的带标号拓扑，构型总数的均值是根构型平均数量的两倍；当 T_n 和 R_n 之间的相关系数接近0.9且 n 较大时，更是如此。对于服从 Yule 分布的带标号拓扑，构型总数的均值渐近地与根构型的均值相当；随着 n增大， T_n 和 R_n 之间的相关系数趋近于1。总结见表2。

T_n 和 R_n 在带标签拓扑的尤尔分布下具有相同的渐近增长，且相关性趋近于1，这多少令人惊讶。 R_n 仅统计根处的祖先构型，而 T_n 则对所有 $n-1$ 内部节点的构型求和，包括根节点。相关性结果表明，在尤尔分布下，非根节点处的构型对总数的贡献可以忽略不计。

均匀模型与尤尔模型的结果差异表明，树平衡性与构型总数之间可能存在相关性。事实上， [11] 曾提出根构型存在这种关系。构型总数也可能存在类似关系；我们发现，在尤尔模型下（该模型赋予平衡拓扑更高的权重），构型的平均总数比均匀模型下增长得更快（图8）。

我们的研究自然引出了若干后续方向。例如，我们尚未刻画给定大小下祖先构型总数最多的带标签拓扑。 [11] 的第4节描述了根祖先构型数量达到最大值的带标签拓扑的递归结构。然而，如图9所示，根构型数与构型总数未必会在同一个带标签拓扑上达到最大值。我们也未研究不匹配的基因树和物种树。对于不匹配的情形，即基因树和物种树具有不同的带标签拓扑，值得进一步分析，因为不匹配的基因树标记拓扑可能比与物种树 [11] 相匹配的拓扑具有更多的构型总数。值得探究的是，本文采用的技术能否推广，以推导基因树和物种树拓扑不同时祖先构型数量的分布性质。

致谢 本项目源于在班夫国际研究站举办的一次研讨会上的讨论。美国国立卫生研究院向 NAR 提供了 R01 GM117590 资助；台湾科技部（MOST）向 MF、CYH 和 ARP 提供了 MOST-107-2115-M-009-010-MY2 和 MOST-109- 2115-M-004-003-MY2 资助；意大利教育、大学与研究部向 FD 提供了丽塔·莱维-蒙塔尔奇尼资助。

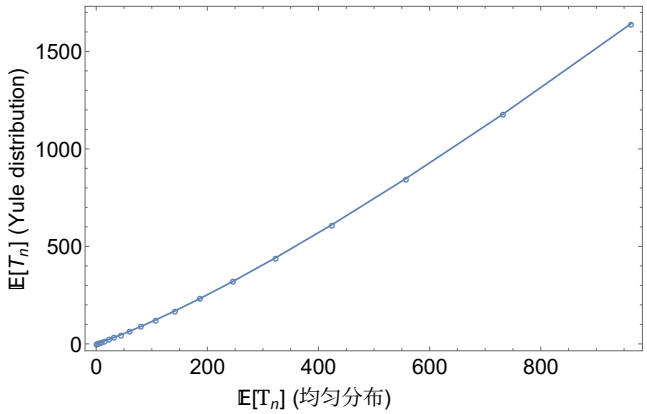

图 8：在尤尔分布和均匀分布下，大小为 n 的带标号拓扑的构型总数的均值，其中$2 \leqslant n \leqslant 20$。均匀分布下的数值通过结合方程 (27) 和 (28) 中的递推关系计算得出；尤尔分布下的数值通过结合方程 (56) 和 (57) 中的递推关系计算得出。

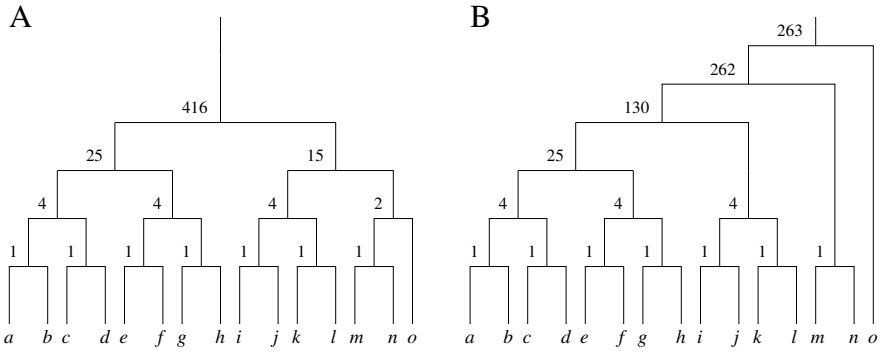

图 9：大小为 $n = 15$ 且（忽略标号的排列）具有最大构型数的带标号拓扑。(A) 根构型（416）。(B) 构型总数（690）。每个内部节点旁标出构型数。图 2 中最右侧的点坐标为 (log 416，log 477)；最上方的点坐标为 (log 263，log 690)。

表2：带标号拓扑上均匀模型和 Yule 模型下根构型数量及构型总数的渐近等价关系汇总。

	均匀模型		Yule 模型	
数量	结果	参考文献	结果	参考文献
$\mathbb{E}[R_n]$	$\sqrt{\frac{3}{2}}\left(\frac{4}{3}\right)^n$	公式 (24)	$\left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n$	公式 (24)
$\mathbb{E}[R_n^2]$	$\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	公式（25）	$(2.0449954971\ldots)^n$	公式（25）
$\mathbb{V}[R_n]$	$\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	公式（25）	$(2.0449954971\ldots)^n$	公式（25）
$\mathbb{E}[T_n]$	$\sqrt{6}\left(\frac{4}{3}\right)^n$	命题 3	$\left(\frac{1}{1-e^{-2\pi\sqrt{3}/9}}\right)^n$	命题 8
$\mathbb{E}[T_n^2]$	$\frac{2}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	公式（45）	$(2.0449954971\ldots)^n$	命题 9
$\mathbb{V}[T_n]$	$\frac{2}{17}(15+11\sqrt{2})\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	命题5	$(2.0449954971\ldots)^n$	命题 9
$\mathbb{E}[T_nR_n]$	$\left(1+\frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	方程（44）	$(2.0449954971\ldots)^n$	命题 9
协方差$[T_n,R_n]$	$\left(1+\frac{\sqrt{2}}{2}\right)\sqrt{\frac{7(11-\sqrt{2})}{34}}\left[\frac{4}{7(8\sqrt{2}-11)}\right]^n$	命题5	$(2.0449954971\ldots)^n$	命题 9
$\rho[T_n,R_n]$	$\frac{1+\frac{\sqrt{2}}{2}}{\sqrt{\frac{2}{17}(15+11\sqrt{2})}}$	命题5	1	命题 9

我们注意到反复出现的常数的数值：$1/(1-e^{-2\pi\sqrt{3}/9}) \approx 1.4253868277$, $1-e^{-2\pi\sqrt{3}/9} \approx 0.70156394081$, $4/[7(8\sqrt{2}-11)] \approx 1.8215272244$, $7(8\sqrt{2}-11)/4 \approx 0.5489898732$。常数 2.0449954971 在 [7],的附录中进行了求值，其倒数为 0.4889986317。

参考文献

[1] D. J. Aldous. 从 Yule 到今天，系统发育树的随机模型与描述性统计。统计科学，16:23–34，2001年。

[2] E. Alimpiev 和 N. A. Rosenberg. 毛毛虫状物种树与 p-伪毛毛虫状基因树的共祖历史计数。应用数学进展，131:102265，2021年。

[3] R. Bhatia 和 C. Dav is. 关于方差的一个更优界。美国数学月刊，107:353–357，200 0.

[4] J. K. M. Brown。进化树的概率。*Syst. Biol.*，43:78–91，199 4.

[5] H. Chang 和 M. Fuchs。系统发育树中模式的极限定理。*J. Math. Biol.*，60:481–512，2010年。

[6] J. H. Degnan 和 L. A. Salter。共祖过程中基因树的分布。*Evolution*，59:24–37，2005年。

[7] F. Disanto、M. Fuchs、A. R. Paningbatan 和 N. A. Rosenberg。在两种物种树模型下，匹配基因树和物种树的根祖先构型数量的分布。*Ann. Appl.Prob.*，32:4426–4458，2022。

[8] F. Disanto 和 E. Munarini。随机游走的加权 Dyck 模型中的局部高度，以及毛毛虫形基因树和物种树的共祖历史数量的变异性。*SN* 应用科学，1:578，2019年。

[9] F. Disanto 和 N. A. Rosenberg。扭叶松物种树的共祖历史。计算生物学杂志，22:918–929，2015年。[10] F. Disanto 和 N. A. Rosenberg。毛毛虫状物种树族中匹配的共祖历史数量的渐近性质。IEEE/ACM计算生物学与生物信息学汇刊，13:913–925，2016年。[11] F. Disanto 和 N. A. Rosenberg。匹配基因树和物种树的祖先构型枚举。计算生物学杂志，24:831–850，2017年。[12] F. Disanto 和 N. A. Rosenberg。匹配基因树和物种树的紧致共祖历史枚举。数学生物学杂志，78:155–188，2019年。[13] F. Disanto 和 N. A. Rosenberg。关于匹配基因树和物种树的非等价祖先构型数量。数学生物学通报，81:384–407，2019年。[14] F. Disanto、A. Schlizio 和 T. Wiehe。受节点不平衡约束的尤尔生成树。数学与生物科学，246:139–147，2013年。[15] F. Disanto 和 T. Wiehe。共祖模型下排序树中樱桃和叉形结构的精确枚举。数学与生物科学，242:195–200，2013年。[16] J. Felsenstein。进化树的数量。系统动物学，27:27–33，1978年。[17] J. Felsenstein。《系统发育推断》。Sinauer，马萨诸塞州桑德兰，2004年。[18] P. Flajolet 和 R. Sedgewick。《解析组合学》。剑桥大学出版社，剑桥，2009年。[19] E. F. Harding。随机二分生成的有根树形的概率。应用概率进展，3:44–77，1971年。[20] Z. M. Himwich 和 N. A. Rosenberg。受阻单调路径与非匹配毛毛虫状基因树和物种树的共祖历史枚举。应用数学进展，113:101939，2020年。[21] M. C. King 和 N. A. Rosenberg。单败淘汰制体育锦标赛与进化树之间的数学联系。数学杂志，96:即将出版，2023。[22] A. McKenzie 和 M. Steel。两种树模型的樱桃分布。数学与生物科学，164:81–92，2000年。[23] R. Otter。树的数量。数学年刊，49:583–599，1948年。[24] N. A. Rosenberg。尤尔生成的谱系树中 r叉节点和 r毛毛虫的均值与方差。组合学年刊，10:129–146，2006年。[25] N. A. Rosenberg。共祖历史计数。计算生物学杂志，14:360–377，2007年。[26] N. A. Rosenberg。毛毛虫状族的共祖历史。IEEE/ACM计算生物学与生物信息学汇刊，10:1253–1262，2013年。[27] N. A. Rosenberg。利用对跖樱桃枚举基因树和物种树的孤立对。应用数学进展，102:1–17，2019年。[28] N. A. Rosenberg 和 J. H. Degnan。不一致基因树和物种树的共祖历史。理论种群生物学，77:145–151，2010年。[29] R. J. Serfling。《数理统计近似定理》。Wiley，纽约，1980年。

[30] R. P. Stanley。《组合计数学（第2卷）》。剑桥大学出版社，纽约，1999年。[31] M. Steel。《系统发育：进化中的离散与随机过程》。工业与应用数学学会，费城，2016年。[32] M. Steel 和 A. McKenzie。尤尔型物种形成模型所生成的系统发育树的性质。《数学与生物科学》，170：91–112，2001年。[33] J. Truszkowski、C. Scornavacca 和 F. Pardi。在多物种溯祖过程中，计算与物种树一致的基因树的概率。《理论种群生物学》，137：22–31，2021年。[34] J. H. M. Wedderburn。函数方程 $g(x^2) = 2\alpha\, x + [g(x)]^2$。《数学年刊》，24：121–140，1922年。[35] Y. Wu。在不完全谱系排序条件下，基于最大似然从基因树拓扑结构推断物种树。《进化》，66：763–775，2012年。[36] Y. Wu。一种用于计算多物种溯祖过程下基因树概率的算法及其在种群树推断中的应用。《生物信息学》，32：i225–i233，2016年。[37] G. U. Yule。一种基于 J. C. Willis 博士、F. R. S. 研究结论的进化数学理论。《伦敦皇家学会哲学汇刊B辑》，213：21–87，1925年。