Yule-Simon 过程及其应用

修格致

IRSGIS of PKU

October 15, 2018

Outline

- Yule-Simon Process 的前世今生
- 生成模型
- 数学工具 *

Remark

随机过程与生成模型的方法是可以互相借鉴的。试验方法上，随机过程利用的工具主要是概率论中的矩估计、组合方法、和生成函数；而生成模型的试验方法则经常用到 MCMC。二者可以相互验证。进一步的，概率论中的结论有助于我们理解生成模型预测出的复杂网络的大尺度特征。

Yule 模型

研究对象：

- 大尺度超文本（天然的 sf 网络）
  - 研究对象是网页的 popularity 时，需要考虑的就是网络的 connectivity
- connectivity 的极限分布会是幂律的。

Yule 模型之前：Yule 过程

- 纯生过程 (pure birth process)
  - 状态序列为 $\{0, 1, 2, 3, 4, 5 \dots\}$.
  - 转移概率为 $P_t(n, n+1) = \lambda_n \Delta t$.
- Yule 过程：纯生过程的实例
  - $\lambda_n = n\lambda$

Yule 模型的描述

- Yule 模型是连续时间随机过程。
- Yule 模型是由不同的参数的 Yule 过程组成的：
  - 第一个尤尔过程记录结点个数：$\{N_\beta(T)\}_{T\geq 0}, \beta > 0$
  - 每加入一个新的结点，都开始一个新的 Yule 过程 $\{N_\lambda(T)\}_{T\geq 0}, \lambda > 0$. 这个 Yule 过程用来描述到这个结点的边的个数。
- 解释：生物的属中，物种数量的分布。这种分布有幂律分布的特征。

Yule 的构造

- 定义两个独立的线性纯生过程：
  - 一个是属出现的过程
  - 一个是种出现的速度
- 物种规模的极限分布：

$$\lim_{T\to\infty} \mathcal{P}(\mathcal{N}=k) = \rho\frac{\Gamma(k)\Gamma(1+\rho)}{\Gamma(k+1+\rho)} = \rho B(k+1+\rho)$$
$$\sim \rho k^{-(1+\rho)}$$

Simon 模型

- 每个时刻加入一个单词。对于 $\alpha \in (0,1)$,
  - 那么 $t+1$ 时刻加入的词是新的的概率为 α,
  - 加入的词出现 k 次的概率为

$$(1-\alpha)\frac{k\vec{N}_{k,t}}{t} \tag{1}$$

  其中 $\vec{N}_{k,t}$ 代表出现过 k 次的词的个数。

- 设词数为 V_t, 那么

$$\frac{\vec{N}_{k,t}^{Simon}}{V_t} \to \P \frac{1}{1-\alpha}\frac{\Gamma(k)\Gamma(1+\frac{1}{1-\alpha})}{\Gamma(k+1+\frac{1}{1-\alpha})} \sim \frac{1}{1-\alpha}k^{-1-\frac{1}{\alpha}} \tag{2}$$

Yule 模型和 Simon 模型的极限，都是幂律分布

- 我们实际上可以将两个过程对应起来：
  - Simon 过程中的第一个过程相当于 Yule 过程中，$\beta = (1 - \alpha)$
  - Simon 过程中的第二个过程相当于 Yule 过程中，$\lambda \to 1$

为什么 Yule 过程很重要？

- Yule 过程给出了线性偏好依附模型的一个物理背景
- 该模型不依赖于连边的生成（这是一个可有可无的条件。）所以减少了假设的数量
- 自然地给出了两个量的数学期望：
  - 总的 generation 数（相当于层级数）
  - 总生存时间（相当于网络问题中的路径长度）
- Yule 过程可以适当推广：将种类的纯生过程改成生灭过程。

改进：生灭过程

Generalization: Death

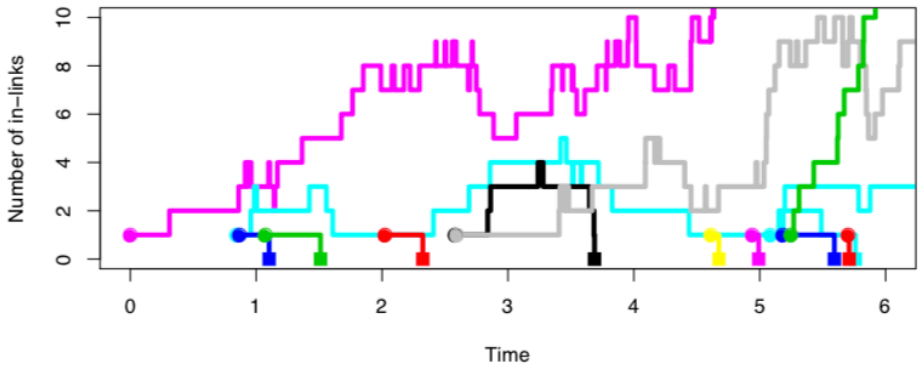

Figure : An example of a realization of the generalized Yule model with $\beta = 0.1$, $\lambda = 1.1$, $\mu = 1$ (supercritical case). The evolution of the number of in-links in different webpages is highlighted with different colors. The instants when new webpages are introduced are indicated with colored dots while the moments when webpages disappear due to the removal of the last in-link are denoted by colored squares.

federico.polito@unito.it

Figure 1: Yule Process with a generalization

改进：生灭过程

分布：

$$\mathcal{P}(\mathcal{N}=0) = \mathit{Hypergeo}(1,0,\beta/\lambda) \tag{3}$$

$$\mathcal{P}(\mathcal{N}=n) = (\beta/\lambda)\Gamma(n)\mathit{Hypergeo}(n,0,\beta/\lambda) \tag{4}$$

衰减速度比幂律分布要快。

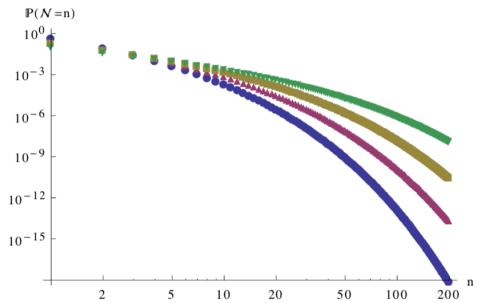

Figure : Distribution of the number of in-links for different values of the webpage rate constant, $\beta = \{1, 0.25, 0.5, 0.1\}$. Critical case, $\lambda = \mu = 1/2$; we can see that the tails decay faster than a power-law.

Figure 2: Caption

生长网络

- 网络的类型决定于网络的功能
- 最简单的网络就是随机网络[1]
- 对于随着时间流逝，网络节点和连接关系随着人口变化改变的场景[2]，我们就需要用生长随机网络理论来解释

[1]以 ER 随机图为代表
[2]电网、运输网络

生长网络的描述

1. 文字描述
2. rate equations

$$\frac{dN_k}{dt} = \frac{1}{A}[A_{k-1}N_{k-1} - A_kN_k] + \delta_{k,1}$$
$$A = \sum_k A_kN_k$$

其中，N_k 是度分布函数；A_k 是连接核，A 是归一化常数。最后一项 $\delta_{k,1}$ 表示在 $k = 1$ 时，该项为 1，其余时候为 0。

这个方程的意义是什么？

工具 1：度分布的低阶矩

n 阶矩定义为：

$$M_n(t) = \sum_j j^n \cdot N_j \tag{5}$$

- 可以得到的结论：
  - 线性核下的度分布函数

$$N_k(t) = \frac{4t}{k(k+1)(k+2)}$$

$$\frac{N_k(t)}{t} \sim 4k^{-3}.$$

  - 一般的核 A_k

$$n_k := \frac{N_k(t)}{t} = \frac{\mu}{A_k} \prod_{j=1}^{k} (1 + \frac{\mu}{A_j})^{-1}$$

$$A(t) = \mu t$$

GRN 的组织

- 每一个时刻，网络加入一个结点。并于之前存在的结点建立 link
- 与度为 k 的结点建立联系的概率为 $A_k(t) \sim k^{-\alpha}$
  - 这个概率决定了偏好依附的强度。

Spatially Distributed Social Complex Networks

社会联系的无无标度网网络，使其看起来像夜空图像里的城市的蔓延; 导出一一个城市规模排名与城市大大小小之间关系的幂律分布; 并反映出高高度联系的个体更倾向于生活在人口密度更高的区域。第一个结点放在方形区域的中心。

Scaling Behaviours in the growth of networked systems and their geometric origins

1. 给定一个有界 d 维欧式空间 $S = |x_1|, \cdots, |x_d| \leq \frac{L}{2}$. $t = 0$ 时在原点插入一个结点.
2. 结点生成：每个 t 时刻，以均匀分布在 S 中放置一个新结点 P_t. 记它的坐标为 x_t. 如果存在一个结点 $P_q, q \in \{1, 2, \cdots, t-1\}$, 使得 $||x_p - x_q|| < r$, 则结点 P_t 存活。否则 P_t 死亡。
3. 连边：将新加入结点与其 $r-$ 临域的所有结点相连。
4. 重复这个过程，直到存活的结点达到 N 个。

Scaling Behaviours in the growth of networked systems and their geometric origins

结论：

- 半径 $R(t) \sim t$（定义为 $\max\{||P_0 - P_i||, i = 1, 2, \cdots t\}$）
- 距中心距离为 ρ 时，密度 $\mu(\rho, \Theta, t) \sim \frac{R(t)-\rho}{L^d}$，$\mu(\rho, t) \sim \frac{(R(t)-\rho)\rho^{d-1}}{L^d}$. 到中心距离在 ρ 以内的结点个数 $\sim \rho^d$.
- 网络中结点的总数

$$N(t) = \int_0^{R(t)} \mu(\rho, t) d\rho \tag{6}$$

$$\sim R(t)^{d-1} \tag{7}$$

$$\sim t^{d-1} \tag{8}$$

- 总边数：$E(t) \sim N(t)^{\frac{d+2}{d+1}}$.

Scaling Behaviours in the growth of networked systems and their geometric origins

标度律

- 边数与结点数 $\gamma = \frac{d+2}{d+1}$.
- 体积与结点数 $\gamma = \frac{d}{d+1}$.
- 加速增长效应 $t \sim N(t)^{\frac{1}{d+1}}$.

修正

- 一个新加入的结点的生存概率是插入点的点密度的一个负指数倍：$P(\text{survive}) = \mu(\rho, \Theta, t)^{-\alpha}$.
  - 边：$R(t) \sim N(t)^{1+\frac{1}{1+(1+\alpha)d}}$
  - 体积 $V(t) \sim N(t)^{1+\frac{1+\alpha}{1+(1+\alpha)d}}$
  - $\alpha \to \infty$ 时，上述两个幂律指数都会趋近于 1. 也就是随着 α 的增加，超/亚线性性会降低至线性。

矩的计算

此处需手写。

取热力学极限

** 热力学极限 **: $L, t \to \infty$.
在这种情况下，网络会渐近形成一个 d 维球。

生成模型与 YS 模型的对应

从种类的产生上：

- Yule model：作为连续时间模型，用速率来控制新种类出现的时间；
- Simon model：作为离散时间模型，用转移概率来控制新单词出现的时间；
- KR model：作为拓扑模型，通过连接关系来决定种类的产生；
- 张江 model：作为空间分布模型，用空间扩张速度来控制密度的变化。

生成模型与 YS 模型的对应

从幂律的产生上

- Yule 模型：产生速率导致的长期稳定性
- Simon 模型：线性偏好依附 kernal
- KR 模型：线性偏好依附 kernal
- 张江模型：度分布与局部密度的关联而局部密度与空间维数的关系导致了幂律分布