数据挖掘课程报告(2019 年春)

Learning and Interpreting Complex Distributions in Empirical Data

Chengxi Zang (Tsinghua University); Peng Cui (Tsinghua University); Wenwu Zhu (Tsinghua University)

作者，题目，发表会议或期刊，年份，卷期号(只有期刊有)，页码。

专　　业 软件工程

学　　生 李洋

班　　号 1611101

学　　号 161110117

日　　期 2019-05-28

内容概要

问题定义：

学习和解释经验数据中的复杂分布。

内容简介：

现实生活中，有各种数据分布模型。比如高斯分布（正态分布），它在数学、物理及工程等领域都非常重要的应用，在统计学的许多方面有着重大的影响力；幂律分布，它在计算机文件大小的分布、国家姓氏的分布、每类生物中物种数的分布的统计方面都有重要作用；韦布尔分布，它被广泛应用于各种寿命试验的数据处理。诸如此类，每一种分布都适用于特定的领域，因此可以想到，是否可以将这些复杂的分布抽象为一种模型？这一篇论文就是通过一系列微分方程，构建了一个四参数的模型，可以简约地描述这些分布。而且研究人员收集了来自不同学科的 16 个具有代表性的真实数据集，验证了这个模型的准确性。

学习和解释经验数据中的复杂分布

Chengxi Zang*
中国北京清华大学
计算机科学与技术系
chengxi.zang@gmail.com

Peng Cui
中国北京清华大学
计算机科学与技术系
cuip@tsinghua.edu.cn

Wenwu Zhu
中国北京清华大学
计算机科学与技术系
wwzhu@tsinghua.edu.cn

摘要

拟合经验数据分布，然后解释它们生成的方式，是理解不同学科中数据背后的结构和动态的常见研究范式。但是，以前的工作主要是尝试以个案的方式拟合或解释经验数据分布。面对现实世界中复杂的数据分布，我们可以通过统一但简约的参数化模型来拟合和解释它们吗？

本文将复杂的经验数据看作是以均匀随机性为输入的动态系统生成的。通过对数据生成动力学的建模，我们展示了一个包含推理和仿真算法的四参数动态模型，该模型能够拟合和生成一系列分布，从高斯分布、指数分布、幂律分布、拉伸指数分布(Weibull)到具有多尺度复杂性的复变函数。我们的模型可以用一个统一的微分方程来解释，而不是一个黑匣子，这个微分方程可以展现这些分布的生成。我们的框架可以有原则地构建更强大的模型。我们通过各种合成数据集来验证我们的模型。然后我们应用我们的模型来自不同学科的 16 个真实数据集。我们用最广泛使用的方法对这些数据集进行了系统的拟合，并证明了我们的模型的优越性。简而言之，我们的模型可能提供一个框架，以便在经验数据中拟合复杂的分布，更重要的是，了解它们的生成机制。

CCS 概念

计算数学→分布函数；随机过程；
网络→网络动态；社交媒体网络；

关键词

复杂分布；重尾分布；生存分析；动态模型；解释性。

ACM 参考格式：

Chengxi Zang, Peng Cui, and Wenwu Zhu. 2018. Learning and Interpreting Complex Distributions in Empirical Data. In *Proceedings of The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18).* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3220073

1 引言

通过参数模型拟合经验数据分布，然后解释它们生成的方式，是理解数据背后的结构和生成方式的主要科学范式，其广泛用于各种领域，包括生物学[9]，物理学[1,5]，社会科学[12,15]，计算机科学[8,32]等。例如，通过研究网络的幂律度分布[2]，物理学家发现了网络演化随机网络中的动态。通过检查达尔文和爱因斯坦[18]的通信模式的响应时间分布，社会科学家试图揭示人类行为的决策动态。通过拟合高斯混合模型[19]，贝叶斯方法[14]，甚至深度生成模型[10]的数据分布，计算机科学家试图找到观察数据集的聚类结构和生成动力学。简而言之，这种科学范式适用于广泛的数据科学任务。

然而，以往的研究主要是对复杂的实证数据进行个案拟合或解释。例如，高斯分布最广泛用于拟合窄尾数据分布。大量的文献试图通过幂律分布[5]，威布尔分布（或拉伸的指数分布）[11]等方法对重尾数据进行建模。特定混合模型也用于拟合复杂的多尺度分布[23,30,32]。像 GAN 这样的深层生成网络在拟合一维参数分布[24]时显现出有限的能力。因此，我们是否可以拥有一个统一的模型来拟合和解释现实世界中各种复杂的数据分布？回答这个问题至关重要。

在本文中，我们试图通过调查其生成动

态来拟合经验数据中的复杂分布。我们的模型的直观显示如下：我们将具有复杂分布的经验数据视为从动态系统生成，其采用均匀随机性的方式输入。我们不是直接以个案的方式对各种复杂的分布进行建模，而是尝试对其统一的、可能是简约的生成动态进行建模，从而生成所有这些复杂的分布。表 1 显示了一个例子：我们可以通过一个四参数动态模型，而不是逐个地拟合高斯函数、指数函数、幂律函数、伸缩指数函数及其在多尺度系统中具有复杂性的复变函数。我们的框架可以以原则的方式构建更复杂的动态模型，提供了有效的推理方法和模拟算法。此外，我们不是通过黑盒模型，而是通过统一的动态微分方程来解释这些复杂分布的生成动力学。至于实验，我们通过各种合成数据集分析模型的属性，并通过 16 个来自不同学科的经验数据集进一步验证了模型的有效性。我们的模型准确地拟合了所有这些复杂的经验数据（图 5）。我们的模型可能提供一个框架，以适应在现实世界中观察到的复杂分布，更重要的是，了解它们的生成机制。我们总结了我们的贡献如下：

- 统一：我们提出了一个通用模型来拟合经验数据中的各种复杂分布，以及推理和模拟算法。
- 简约：我们的模型只有四个参数来捕捉经验分布中的多尺度复杂性。
- 可解释性：我们的模型由一个统一的生成动态方程解释。所有参数都有明确的物理意义。
- 实用性：我们的模型可以准确地拟合各种经验数据集，并且可以有原则地推广到更复杂的情况。。

论文的大纲是：综述、模型、机制、实验、讨论和结论。复用性：软件和数据库在 www.calvinzang.com 上是开源的。

2 相关工作

我们主要回顾以下两个方面的相关工作：

从简单到复杂的经验数据分布。窄尾分布，如指数分布和高斯分布，可以很好地捕捉到它们的均值和方差，这些分布的潜在结构和动力学特征都得到了很好的研究。相比之下，重尾分布，如幂律分布，拉伸指数分布，对数正态分布等，表现出更大的均匀无穷大方差，这意味着复杂的基础结构和数据动态。在重尾分布中，幂律分布以其尺度特性[22]和生成机制[2]最为著名。关于幂律分布的广泛证据和讨论可以在[13,16]中找到。最近，越来越多的文献发现经验数据的分布比纯粹的幂律更复杂，从人类行为数据[26,32]，网络数据[3]到各种数据集，如图 5 所示。

拟合复杂的分布。最大似然估计(可能带有先验或正则化因子)用于拟合窄尾分布[14]。另一方面，像 GAN 这样的深度生成网络通过拟合一维参数分布得到验证，但存在较大的偏差[24]。相比之下，对于复杂分布的拟合理论，比如偏斜或重尾分析[17]，还没有很好的建立。以最典型的幂律分布为例，首先采用目视检验和最小二乘拟合来拟合幂律分布。随后，著名著作[5]展示了最小二乘拟合方法的偏倚，提出了基于最大似然原理拟合幂律分布的参数法（$f(x)=\frac{\alpha_{PL}-1}{x_{\min}}\left(\frac{x}{x_{\min}}\right)^{\alpha_{PL}}$）根据最大似然原则来拟合幂律分布。PL 方法已被广泛用于通过大量科学论文拟合合理的幂律分布。然而，然而，我们发现 PL 方法在检测真实数据中的幂律信号时存在较大的偏差，如图 5 所示。失败的根源在于 PL 方法忽略了真实数据的复杂性[23,26,32]。如何通过统一模型拟合和解释经验数据集中的各种复杂分布在很大程度上是未知的。

3 提出的方法

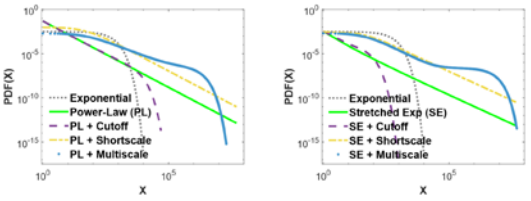
3.1 直观模型

我们的直观模型如下：我们将具有复杂分布的经验数据视为从（非线性）动态系统生成，该系统将均匀随机性作为输入。我们没有对这个动态系统的复杂输出(即各种数据分布)逐个建模，而是试图捕捉它们的统一生成动态。简而言之，我们试图对产生复杂现象的简单生成动力学进行建模。

我们的模型基于生存分析[32]，点过程[26]和动态系统[25,27,31]。数据 $X=$

$(x_1,\dots,x_{n-1},x_n)$ 可以通过危险率函数 $\lambda(x)=\frac{f_X(x)}{S_X(x)}$来建模，其描述了以$X\geq x$为条件的随机变量$X=x$的出现率，其中$S_X(x)=1-\int_{-\infty}^{x}f_X(s)ds.\Lambda(x)=\int_{-\infty}^{x}\lambda(s)ds$表示累积危险率。通过对危险率的建模，我们可以根据关系$f_X(x)=\lambda(x)e^{-\int_{-\infty}^{x}\lambda(s)ds}$得到复杂的概率密度函数。在下一节中，我们将进一步建立λ(x)与其对应的动态系统之间的联系，以解释数据分布的生成机制。

3.2 模型



(a) $\theta=1$ (b) $\theta\neq1$

图 1：基本模型功能的图示。我们的模型生成一系列分布，包括幂律（PL），具有截止的 PL，具有短尺度复杂度的PL，具有多尺度复杂度的PL，指数，拉伸指数（SE），具有短尺度复杂度的 SE，SE 具有多尺度复杂性等。

在这里，我们提出了基本模型，简单但通用，可以得到各种分布，如表 1 和图 1 所示。指定模型的危险率函数是：

$$\lambda(x)=\beta+\alpha(x+\Delta)^{-\theta} \quad (1)$$

3.2.1 关键案例：基于幂律的分布。当$\theta=1$时，$\lambda(x|\theta=1)$基于幂律分布生成一系列分布。

引理 3.1 $\lambda(x|\theta=1)$产生一系列分布，从指数分布，幂律分布，指数截止的幂律分布到复杂的多尺度分布。

证明：随机变量x的概率密度函数为：

$$\begin{aligned}f(x|\theta=1)&=\lambda(x|\theta=1)e^{-\int_0^x\lambda(s|\theta=1)ds}\\&=\left(\beta+\frac{\alpha}{x+\Lambda}\right)e^{-\beta x-\alpha\ln\left(\frac{x}{\Delta}+1\right)}\\&=\beta e^{-\beta x}\left(\frac{x}{\Delta}+1\right)^{-\alpha}\\&\quad+\frac{\alpha}{\Delta}\left(\frac{x}{\Delta}+1\right)^{-(\alpha+1)}e^{-\beta x}\end{aligned}\quad (2)$$

指数分布。当$\alpha=0$时，无论其他三个参数如何，$\lambda(x|\alpha=0)$都会产生指数分布，概率密度函数$f(x|\alpha=0)=\beta e^{-\beta x}$，如图 1a 灰度曲线所示。

幂律分布。当$\beta=0$且$\Delta\ll x$时，$f(x|\theta=1,\beta=0)=\alpha\Delta^{\alpha}(x+\Delta)^{-(\alpha+1)}\propto x^{-(\alpha+1)}$。Δ的另一个含义是最小值，称为$x_0$，$x$可以取：$f(x|\theta=1,\Delta=0)=\frac{\alpha}{x}e^{\int_{x_0}^{x}\frac{\alpha}{s}ds}=\alpha x_0^{\alpha}x^{-(\alpha+1)}$。

具有截止的幂律分布。当$\beta\gg0$且$\Delta\ll x\ll\frac{\alpha}{\beta}-\Delta$，$f(x|\theta=1)=\left(\beta+\frac{\alpha}{x+\Delta}\right)\left(\frac{x}{\Delta}+1\right)^{-\alpha}e^{-\beta x}\approx\alpha\Delta^{\alpha}x^{-(\alpha+1)}e^{-\beta x}$。

复杂的多尺度分布。当$\beta\to0$时，复杂的多尺度分布具有恒定的短尺度、幂律的中尺度和指数的长尺度特征。当$x\to0$时，$f(x|\theta=1)\to\frac{\alpha}{\Delta}$。短尺度方程 $x\in(0,\Delta],f(x|\theta=1)\approx\frac{\alpha}{\Delta}\left(\frac{x}{\Delta}+1\right)^{-(\alpha+1)}$，它慢慢衰减到幂律中尺度范围。当$\beta e^{-\beta x}\left(\frac{x}{\Delta}+1\right)^{-\alpha}\gg\frac{\alpha}{\Delta}\left(\frac{x}{\Delta}+1\right)^{-(\alpha+1)}e^{-\beta x}$，即$x\gg\frac{\alpha}{\beta}-\Delta,f(x|\theta=1)=\beta e^{-\beta x}\left(\frac{x}{\Delta}+1\right)^{-\alpha}+\frac{\alpha}{\Delta}\left(\frac{x}{\Delta}+1\right)^{-(\alpha+1)}e^{-\beta x}\approx\beta e^{-\beta x}\left(\frac{x}{\Delta}+1\right)^{-\alpha}$，这是指数长尺度范围。当$\Delta\ll x\ll\frac{\alpha}{\beta}-\Delta,f(x|\theta=1)\approx\alpha\Delta^{\alpha}(x+\Delta)^{-(\alpha+1)}\propto x^{-(\alpha+1)}$，这是幂律中等尺度范围。

3.2.2 一般情况：基于拉伸指数的分布。当$\theta\neq1$，$\lambda(x|\theta\neq1)$时，生成了一系列基于拉伸指数分布的分布。

引理 3.2。$\lambda(x|\theta\neq1)$生成了一系列分布，从指数分布，拉伸指数（Weibull）分布，带指数截止的拉伸指数分布到复杂的多尺度分布。

证明：与上述理由类似，随机变量 x 的概率密度函数为：

Table 1: Capability table. Our basic model encompasses all the following distributions. Illustrations are shown in Fig. 1.

Capability	Exponential	Power law	Power law + cutoff *	Power law + Shortscale	Power law + Multiscale
PDF ($\theta = 1$)	$\beta e^{-\beta x}$	$\alpha\Lambda^{\alpha}x^{-(\alpha+1)}$	$\alpha\Lambda^{\alpha}x^{-(\alpha+1)}e^{-\beta x}$	$\alpha\Lambda^{\alpha}(x+\Lambda)^{-(\alpha+1)}$	$(\beta+\frac{\alpha}{x+\Lambda})(\frac{x}{\Lambda}+1)^{-\alpha}e^{-\beta x}$
Hazard rate	β	$\frac{\alpha}{x}$	$\beta+\frac{\alpha}{x}$	$\frac{\alpha}{x+\Lambda}$	$\beta+\frac{\alpha}{x+\Lambda}$
Our model	✓	✓	✓	✓	✓
Capability	**Exponential**	**Stretched exponential** **	**Stretched exponential + Cutoff ***	**Stretched exponential + Shortscale**	**Stretched exponential +Multiscale**
PDF ($\theta \neq 1$)	$\alpha e^{-\alpha x}$	$\alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$	$\alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}-\beta x}$	$\alpha(x+\Lambda)^{-\theta}e^{-\frac{\alpha}{1-\theta}[(x+\Lambda)^{1-\theta}-\Lambda^{1-\theta}]}$	$[\beta+\alpha(x+\Lambda)^{-\theta}]e^{-\beta x-\frac{\alpha}{1-\theta}[(x+\Lambda)^{1-\theta}-\Lambda^{1-\theta}]}$
Hazard rate	α	$\frac{\alpha}{x^{\theta}}$	$\beta+\frac{\alpha}{x^{\theta}}$	$\frac{\alpha}{(x+\Lambda)^{\theta}}$	$\beta+\frac{\alpha}{(x+\Lambda)^{\theta}}$
Our model	✓	✓	✓	✓	✓

* For the Power law distribuion with cutoff case and Stretched exponential distribution with cutoff case, the probability density functions of which are derived approximately by the hazzard rates. Refer to the Model Section.
** Special case is approximately Normal Distribution when $\theta = -1$. Refer to the Model Section.

$$f(x|\theta \neq 1) = \lambda(x|\theta \neq 1)e^{-\int_0^x \lambda(s|\theta \neq 1)ds}$$

$$= \left[\beta + \alpha(x+\Delta)^{-\theta}\right]e^{-\beta x-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}$$

$$= \beta e^{-\beta x}e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]} \quad (3)$$

$$+ \alpha(x+\Delta)^{-\theta}e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}e^{-\beta x}$$

指数分布。当$\beta = \theta = 0$, $f(x|\beta = 0, \theta = 0) = \alpha e^{-\alpha x}$。

拉伸指数（Weibull）分布。当$\beta = 0$且$\Delta = 0$，$\lambda(\mathrm{x}|\theta \neq 1)$会有

$$f(x|\theta \neq 1) = \alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}} \quad (4)$$

累积密度函数是$F(x|\theta \neq 1) = 1 - e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$。这是拉伸的指数分布。一些特殊情况：当$\theta = 0$时为指数分布$\alpha e^{-\alpha x}$，当$\theta = -1$近似正态分布$\frac{\alpha}{x^2}e^{-\frac{\alpha x^2}{2}}$。

带指数截止的拉伸指数分布。当$\beta \gg 0$且$\Delta \ll x \ll \left(\frac{\alpha}{\beta}\right)^{\frac{1}{\theta}} - \Delta$，$f(x|\theta \neq 1) = \left[\beta + \alpha(x+\Delta)^{-\theta}\right]e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta}-\Delta^{1-\theta}]}e^{-\beta x} \approx \alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}e^{-\beta x}$。

复杂的多尺度分布。当$\theta \neq 1$时，复多尺度分布基于拉伸指数分布。当$\beta \to 0$时，复杂多尺度分布具有不变的短尺度、指数中尺度和指数长尺度特征，当$x = 0$时，$f(x|\theta \neq 1) \approx \frac{\alpha}{\Delta^{\theta}}$。在短尺度方案中$x \in (0, \Delta]$, $f(x|\theta \neq 1) \approx \alpha(x+\Delta)^{-\theta}e^{-\frac{\alpha}{1-\theta}}\left[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}\right]$，它缓慢地衰减到拉伸的指数中尺度范围。当$\beta \gg \alpha(x+\Delta)^{-\theta}$，即$x \gg \left(\frac{\alpha}{\beta}\right)^{\frac{1}{\theta}} - \Delta$，$f(x|\theta \neq 1) \approx \beta e^{-\frac{\alpha}{1-\theta}}\left[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}\right]_e - \beta x$。它是指数长尺度范围。当$\Delta \ll x \ll \left(\frac{\alpha}{\beta}\right)^{\frac{1}{\theta}} - \Delta$，$f(x|\theta \neq 1) \approx \alpha(x+\Delta)^{-\theta}e^{-\frac{\alpha}{1-\theta}}\left[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}\right]$, 这是指数中等规模尺度。

我们在图 1 中说明了上述理由。我们发现基于幂律分布的复杂分布（图 1a）和拉伸指数分布（图 1b）是由我们的简单危险率函数生成的，可能在多尺度机制中具有复杂性。

3.3 参数推理

我们的模型参数可以通过最大似然估计（MLE）框架来学习。观察一组数据$\{\mathrm{x}_1, \ldots, \mathrm{x}_{n-1}, \mathrm{x}_n\}$的对数似然函数由下式给出：

$$\ln L(x_1, \ldots, x_n) = \ln \prod_{i=1}^{n} \lambda(x_i)e^{-\Lambda(x_i)} \quad (5)$$

$$= \sum_{i=1}^{n} \ln \lambda(x_i) - \sum_{i=1}^{n} \Lambda(x_i)$$

根据θ的值，$\Lambda(\mathrm{x})$采用不同的形式。当$\theta \neq 1$时，对数似然函数是

$$\ln L\,(x_1,\dots,x_n|\theta\neq 1)=\sum_{i=1}^{n}\ln\left[\beta+\alpha(x_i+\Delta)^{-\theta}\right]-\beta\sum_{i=1}^{n}x_i-\frac{\alpha}{1-\theta}\sum_{i=1}^{n}\left[(x_i+\Delta)^{1-\theta}-\Delta^{1-\theta}\right] \quad (6)$$

然而当$\theta=1$时，对数似然函数是：

$$\ln L\,(x_1,\dots,x_n|\theta=1)=\sum_{i=1}^{n}\ln[\beta+\alpha(x_i+\Delta)^{-1}]-\beta\sum_{i=1}^{n}x_i-\alpha\sum_{i=1}^{n}\ln\left(\frac{x_i}{\Delta}+1\right) \quad (7)$$

关于$\{\beta,\alpha,\Delta,\theta\}$的最大化等式 5 或 6，受$\{\beta,\alpha,\theta\geq 0;\Delta>0\}$的约束导致估计的建模参数。但是，由于参数的物理意义明确，可以将先验知识应用于初始化。我们稍后会说明这一点。

该模型的另一个优点是所有参数都具有闭合形式的梯度。$\theta\neq 1$情况的梯度，即基于拉伸指数的模型，是：

$$\frac{\partial \ln L}{\partial\beta}=\sum_{i=1}^{n}\frac{1}{A(i)}-\sum_{i=1}^{n}x_i \quad (8)$$

$$\frac{\partial \ln L}{\partial\alpha}=\sum_{i=1}^{n}\frac{(x_i+\Delta)^{-\theta}}{A(i)}-\frac{1}{1-\theta}\sum_{i=1}^{n}\left[(x_i+\Delta)^{1-\theta}-\Delta^{1-\theta}\right] \quad (9)$$

$$\frac{\partial \ln L}{\partial\Delta}=-\alpha\theta\sum_{i=1}^{n}\frac{(x_i+\Delta)^{-\theta-1}}{A(i)}-\alpha\sum_{i=1}^{n}\left[(x_i+\Delta)^{-\theta}-\Delta^{-\theta}\right] \quad (10)$$

$$\frac{\partial \ln L}{\partial\theta}=-\alpha\sum_{i=1}^{n}\frac{(x_i+\Delta)^{-\theta}\ln(x_i+\Delta)}{A(i)}$$

约束条件：$A(\mathrm{i})=\beta+\alpha(\mathrm{x_i}+\Delta)^{-\theta}$

当$\theta=1$时，基于幂律的模型的梯度是：

$$\frac{\partial \ln \mathrm{L}}{\partial\beta}=\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{1}{\mathrm{B(i)}}-\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{x_i} \quad (12)$$

$$\frac{\partial \ln \mathrm{L}}{\partial\alpha}=\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{(\mathrm{x_i}+\Delta)^{-1}}{\mathrm{B(i)}}-\sum_{\mathrm{i}=1}^{\mathrm{n}}\ln\left(\frac{\mathrm{x_i}}{\Delta}+1\right)$$

$$\frac{\partial \ln \mathrm{L}}{\partial\Delta}=-\alpha\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{(\mathrm{x_i}+\Delta)^{-2}}{\mathrm{B(i)}}+\alpha\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{\mathrm{x_i}}{\mathrm{x_i}\Delta+\Delta^2}$$

（13，14）约束条件：$B(i)=\beta+\frac{\alpha}{x_i+\Delta}$

我们可以用许多基于梯度的优化算法来解决优化问题。例如，我们采用内点算法[4]，为了重现性，可以查看代码代码，参见第 7 节。

3.4 生成器

从累积分布函数$F(\mathrm{x})$生成随机数x的最简单和最优雅的方法是逆变换方法[7]。首先我们从标准均匀分布$U(0,1]$生成一个随机数 u。通过求解方程$F(\mathrm{x})=u$表示x，x是跟随分布$F(\mathrm{x})$的数。我们将这个逆变换方法扩展到由$F(\mathrm{x})=1-\mathrm{e}^{-\Lambda(\mathrm{x})}$的事实导致的危险率函数，其中$\Lambda(\mathrm{x})=\int_{\mathrm{x_0}}^{\mathrm{x}}\lambda(\mathrm{s})\mathrm{d}\,s$。因此，$F(\mathrm{x})=u=1-\mathrm{e}^{-\Lambda(\mathrm{x})}$，我们可以通过求解$\Lambda(\mathrm{x})=-\ln(1-\mathrm{u})$得到所需的随机数，其中$u$和从$U(0,1]$采样时$1-u$没有区别。由于$\Lambda(\mathrm{x})$是单调递增函数，因此具有反函数$\Lambda^{-1}$，我们可以得到

$$x=\Lambda^{-1}(-\ln \mathrm{u}) \quad (15)$$

即使没有闭合形式的反函数Λ^{-1}，我们也可以通过求解方程$\ln \mathrm{u}+\Lambda(\mathrm{x})=0$来得到数值，

其中u是从均匀分布$U(0,1)$生成的。

Algorithm 1: Generating random samples specified by the hazard rate Equation 1

Input : Hazard function of model $\lambda(x) = \beta + \alpha(x+\Delta)^{-\theta}$, total number N
Output: $\{x_1, \ldots, x_N\}$
Set current number of events $n = 1$;
while $n \leq N$ **do**
  Sample $u \sim Uniform([0,1])$;
  Solve $\ln u + \Lambda(x) = 0$ for x by Algorithm 2.;
  $x_n = x$;
end

Algorithm 2: Newton's iterative method

Input : Equation $\Phi(x) = \log u + \Lambda(x)$.
Output: x
Set $\epsilon = 10^{-8}, x = 0$;
while $|\Phi(x)| \leq \epsilon$ **do**
  if $\theta == 1$ **then**
    $\Phi(x) = \ln u + \beta x + \alpha \ln(\frac{x}{\Delta} + 1)$;
  else
    $\Phi(x) = \ln u + \beta x + \frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]$;
  end
  $\Phi'(x) = \beta + \alpha(x+\Delta)^{-\theta}$;
  $x = x - \frac{\Phi(x)}{\Phi'(x)}$;
end

4 物理机制

在本节中，我们给出了模型的基本生成动力学，即等式 1 和各种分布，如表 1 所示。我们将复杂数据分布视为由（非线性）动态系统生成，采用均匀随机性作为输入：

4.1 统一输入和增长

为了给出数据生成过程的动态视图，我们的第一步是通过连接点过程和生存分析来计算动态系统的输入。我们从标准均匀分布$U(0,1]$中采样n个数的过程可以看作是一个随机点过程。给定泊松过程$N(t) = \{t_i | i = 1, \ldots, N(t) = n; t_i \leq t_2 \leq \cdots \leq t_n\}$，然后$t_i$是均匀分布在区间$(0,t]$上。如果我们将$t_i$标准化为$t$，则$u = \frac{t_i}{t}$遵循标准均匀分布$U(0,1]$。

我们将等式 15 中的u替换为$\frac{t_i}{t}$，导致代理i的增长动态，其中$(0,t]$的均匀到达时间t_i：

$$x_i(t) = \Lambda^{-1}(-\ln u) = \Lambda^{-1}\left(-\ln\left(\frac{t_i}{t}\right)\right) = \Lambda^{-1}\left(\ln\left(\frac{t}{t_i}\right)\right) \quad (16)$$

例如，让$\lambda(x) = \alpha x^{-\theta}$，当$\theta = 1$时它产生幂律分布$f(x) = \alpha\Delta^{\alpha}x^{-(\alpha+1)}$。并且当$\theta \neq 1$时，拉伸指数分布为$f(x) = \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$（详见第 3 节）。因此，$\Lambda(x|\theta = 1) = \alpha\int_{\Delta}^{x}\frac{1}{s}ds = \alpha\ln\left(\frac{x}{\Delta}\right)$，并且$\Lambda(x|\theta \neq 1) = \alpha\int_{\Delta}^{x}\frac{1}{s^{\theta}}ds = \frac{\alpha(x^{1-\theta} - \Delta^{1-\theta})}{1-\theta}$。我们可以得到它们的反函数$\Lambda^{-1}(y|\theta = 1) = \Delta e^{\frac{y}{\alpha}}$，和$\Lambda^{-1}(y|\theta \neq 1) = \left(\frac{1-\theta}{\alpha}y + \Delta^{1-\theta}\right)^{\frac{1}{1-\theta}}$。

通过在等式 16 中使用$\Lambda^{-1}(y|\theta = 1) = \Delta e^{\frac{y}{\alpha}}$，我们获得了超过时间的增长曲线$t$：

$$x_i(t) = \Lambda^{-1}\left(\ln\left(\frac{t}{t_i}\right)\middle|\theta = 1\right) = \Delta e^{\frac{\ln\left(\frac{t}{t_i}\right)}{\alpha}} = \Delta\left(\frac{t}{t_i}\right)^{\frac{1}{\alpha}} \quad (17)$$

类似地，当$\theta < 1$时，通过将$\Lambda^{-1}(y|\theta \neq 1) = \left(\frac{1-\theta}{\alpha}y + \Delta^{1-\theta}\right)^{\frac{1}{1-\theta}}$应用于等式 16，，我们得到增长曲线：

$$x_i(t) = \Lambda^{-1}\left(\ln\left(\frac{t}{t_i}\right)\middle|\theta \neq 1\right) = \left(\frac{1-\theta}{\alpha}\ln\left(\frac{t}{t_i}\right) + \Delta^{1-\theta}\right)^{\frac{1}{1-\theta}} \quad (18)$$

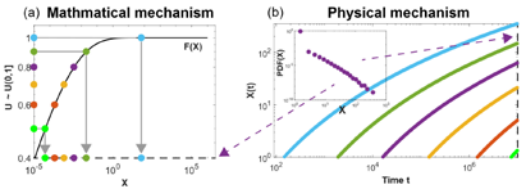


图 2:(a)生成器和(b)生成动态图，用于生成拉伸指数数据，如图(b)中的插图所示。由生成器生成的样本沿 x 轴分布，而物理过程生成的样本沿 x(t)轴作为横断面快照分布。

4.2 动态生成：基础模型

我们的第二步是通过连接生存分析和动态系统来逆向工程生成动力学。我们将等式 17 和 Eq.18 对时间 t 求导，我们得到幂律分布数据和拉伸指数分布数据的生成动态

关系如下：

$$\frac{\mathrm{dx_i(t)}}{\mathrm{dt}}=\frac{\mathrm{d\Delta}\left(\frac{\mathrm{t}}{\mathrm{t_i}}\right)^{\frac{1}{\alpha}}}{\mathrm{dt}}\frac{\Delta}{\mathrm{t_i^{\frac{1}{\alpha}}}}\frac{1}{\alpha}t^{\frac{1}{\alpha}-1}=\frac{x_i(t)}{\alpha t} \quad (19)$$

$$\frac{dx_i(t)}{dt}=\frac{d\left(\frac{1-\theta}{\alpha}\ln\left(\frac{t}{t_i}\right)+\Delta^{1-\theta}\right)^{\frac{1}{1-\theta}}}{dt}$$

$$=\frac{\left(\frac{1-\theta}{\alpha}\ln\left(\frac{t}{t_i}\right)+\Delta^{1-\theta}\right)^{\frac{\theta}{1-\theta}}}{\alpha t}$$

$$=\frac{x_i(t)^{\theta}}{\alpha t} \quad (20)$$

我们用 Eq.19 得到线性优先附件生成幂律分布，用等式 20 得到延伸型指数函数分布，与随机网络中无标度观测的文献一致[2]。

4.3 动态生成：一般模型

在这里，我们给出了模型的动态生成。当 $\lambda(\mathrm{x})=\beta+\alpha(\mathrm{x}+\Delta)^{-\theta}$ 时，$\Lambda(\mathrm{x})=\int_0^{\mathrm{x}}\lambda(\mathrm{s})\mathrm{d}s=\beta x+\frac{\alpha}{1-\theta}\left[(\mathrm{x}+\Delta)^{1-\theta}-\Delta^{1-\theta}\right]$。

通过我们的构造，我们得到：

$$\Lambda\left(x_i(t)\right)=\beta x_i(t)+\frac{\alpha}{1-\theta}\left[(x_i(t)+\Delta)^{1-\theta}-\Delta^{1-\theta}\right]=\ln\frac{t}{t_i} \quad (21)$$

通过将上述方程对t求导，我们得到：

$$\frac{dx_i(t)}{dt}=\frac{(x_i(t)+\Delta)^{\theta}}{\beta(x_i(t)+\Delta)^{\theta}t+\alpha t} \quad (22)$$

因此，从动态的角度来看，表现出复杂多尺度分布的复杂数据（由我们的模型等式 1 捕获）是由微分方程 22 的动态系统生成的，包括物理机制：非线性优先附着$(\mathrm{x_i(t)}+\Delta)^{\theta}$，增长系统，以及短期复杂度$\Delta$和长期复杂度$\beta(\mathrm{x_i(t)}+\Delta)^{\theta}t$。我们的动态包括特殊情况下的等式 19 和等式 20。

因此，我们在随机网络场景中描述数据生成过程如下：

- 新节点i在$0<\mathrm{t_i}<t$的泊松过程后进入网络，其中t是最大观测时间，
- 并且，节点i的程度，表示为$\mathrm{x_i(t)}$，

根据微分方程 22 随时间增长。

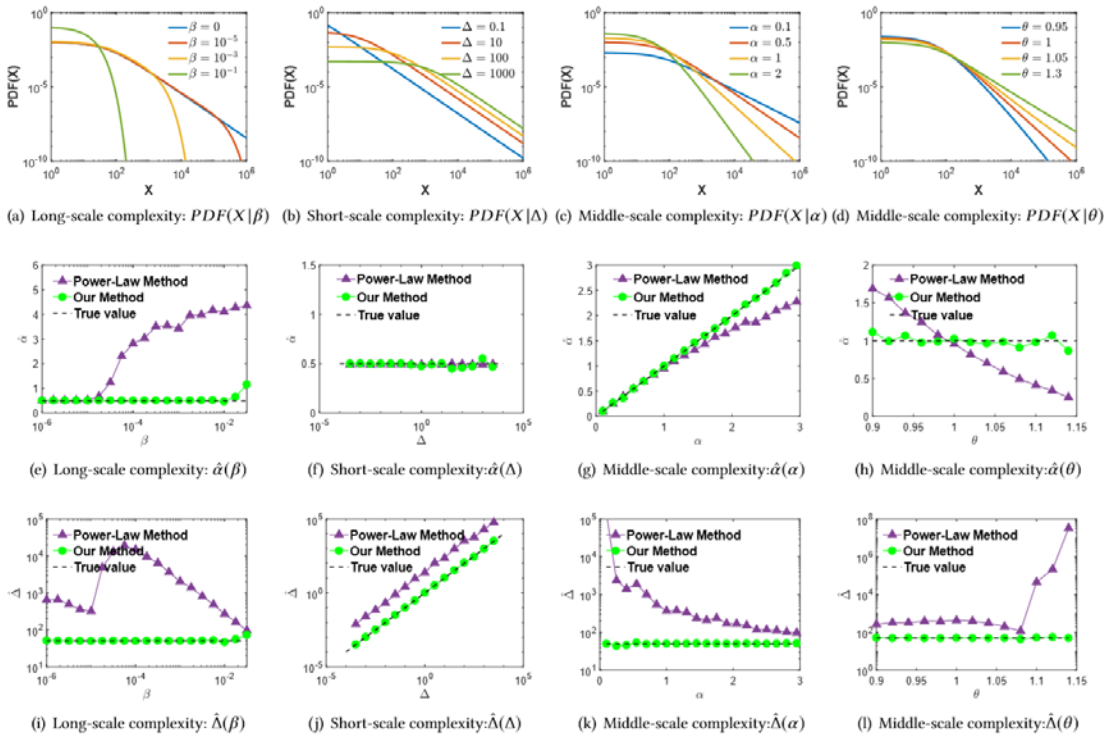
然后，该网络在时间t的横截面度分布为$f_X(x)=\lambda(x)e^{-\int_{-\infty}^{x}\lambda(s)ds}$，其中$\lambda(\mathrm{x})=\beta+\alpha(\mathrm{x}+\Delta)^{-\theta}$如等式 1 所示。

5.1 综合数据分析

*5.1.1 忽略完整性会导致系统偏差。*实际上，衡量现实世界数据量的分布要比纯幂律分布复杂得多。我们将在下一节中展示来自真实世界数据集的证据。在这里，我们研究了通过将着名的幂律拟合方法（表示为 PL 方法）[5]应用于复杂分布而引入的可能偏差。

长期复杂性。我们首先研究长期复杂性。参数β作为建模长尺度复杂性的最简单形式。通过改变$\lambda(\mathrm{x}|\beta,\alpha=0.5,\Delta=50,\theta=1)$中的$\beta$，我们得到了图 3a 中所示的一系列分布。当$\beta=0$时的情况，如图 3a 中蓝色曲线的直线部分（幂律指数$1+\alpha=1.5$）所示，表明缺少长尺度复杂性。随着β的增加，长尺度的复杂性将向短距离移动，直到两个部分重叠。实际上，长期制度的特征尺度是$\frac{\alpha}{\beta}$，短期制度的特征尺度是Δ（参见第 3 节）。我们通过限制$\beta<10^{-2}$来避免这种重叠。我们生成 104（一个相对较大的数据集，以获得合理的拟合结果，同时在基线方法的可扩展性限制内。我们将在稍后显示模型和基线的可扩展性。）采用每个特定β的$\lambda(\mathrm{x}|\beta,\alpha=0.5,\Delta=50,\theta=1)$样本，并通过 PL 方法和我们的方法拟合$\alpha$和$\Delta$。图 3e 和$i$绘制了$\alpha$（缩放指数）和$\Delta$作为$\beta$的函数的平均估计值。我们发现 PL 方法估计的幂律缩放参数α与虚线标记的真值之间的差异越来越大，β增加，如图 3e 所示。相比之下，我们的模型很好地获得了真正的缩放值。对于图 3i 所示的短尺度Δ的估计，PL 方法严重高估了真实值，当$\beta\approx3*10^{-5}$时，真实值高达 450 倍。随着β的增长，长期制度挤入短期制度，因此通过 PL 方法估计的Δ降低到真实值。

短期复杂性。然后我们考虑短期复杂性的影响。参数Δ是捕获短尺度方案特征尺度



(a) Long-scale complexity: $PDF(X|\beta)$ (b) Short-scale complexity: $PDF(X|\Lambda)$ (c) Middle-scale complexity: $PDF(X|\alpha)$ (d) Middle-scale complexity: $PDF(X|\theta)$

(e) Long-scale complexity: $\hat{\alpha}(\beta)$ (f) Short-scale complexity:$\hat{\alpha}(\Lambda)$ (g) Middle-scale complexity:$\hat{\alpha}(\alpha)$ (h) Middle-scale complexity:$\hat{\alpha}(\theta)$

(i) Long-scale complexity: $\hat{\Lambda}(\beta)$ (j) Short-scale complexity:$\hat{\Lambda}(\Lambda)$ (k) Middle-scale complexity:$\hat{\Lambda}(\alpha)$ (l) Middle-scale complexity:$\hat{\Lambda}(\theta)$

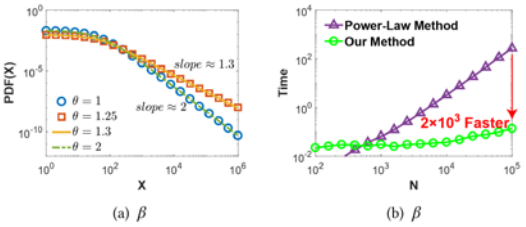
图 3:在短期、长期和中等规模的情况下，复杂性会对前面的方法引入系统偏差。我们的方法很符合实际。第一行显示了不同体系中具有不同复杂性的分布。最后两行显示了建模参数的平均估计值，这些估计值作为变化参数的函数被绘制出来，这些变化参数捕获了相应体系中的复杂性。每一列中的数字都具有相同的设置。在所有情况下，统计误差都小于数据点。真正的参数值用虚线表示。PL 方法中描述的 pdf 是$f(x) = \frac{\alpha_{PL}-1}{x_{min}}\left(\frac{x}{x_{min}}\right)^{\alpha_{PL}}$**其中**$\alpha = \alpha_{PL} - 1$。

的最简单形式。图 3b 绘制了$\lambda(x|\beta = 0, \alpha = 0.5, \Delta, \theta = 1)$随着Δ变化，其由短尺度方案和随后具有相同斜率的幂律方案组成。Δ越大，短期规模越大，因此范围越广。类似地，我们通过$\lambda(x|\beta = 0, \alpha = 0.5, \Delta, \theta = 1)$生成 104 个样本，每个特定的Δ，并通过 PL 方法和我们的方法拟合α和Δ。我们的模型很好地拟合了两个参数的实际数据，如图 3f 和图 3j 所示。对于这个特殊的实验设置，PL 方法在拟合比例指数α方面做得很好，但是α的良好结果是以严重高估Δ为代价，表明 PL 方法丢弃了短程方案中的数据样本，占整个数据集的 280%。

中等规模的复杂性。最后但同样重要的是，我们研究了中等规模的复杂性。当$\theta = 1$时，中等规模的制度遵循幂律，其中缩放指数为$1 + \alpha$。通过在控制其他参数时改变$\lambda(x|\beta = 0, \alpha, \Delta, \theta = 1)$中的$\alpha$，我们得到中等和规模体系下幂律分布的不同比例指数。图 3c 绘制了$\lambda(x|\beta = 0, \alpha, \Delta = 50, \theta = 1)$作为变化，$\alpha$越大，曲线越陡峭。然而，我们发现，随着生长，PL 方法低估了缩放参数α，并且差异变得越来越大，如图 3g 所示。此外，PL 方法同时严重高估了短程参数δ，最高达 3 个数量级，如图 3k 所示。相比之下，我们的模型始终如一地达到了真正的价值。

当$\theta \neq 1$时，中等规模的尺度遵循拉伸指数定律。图 3d 绘制$\lambda(x|\beta = 0, \alpha = 1, \Delta = 50, \theta)$，随着$\theta$的变化图像。$\theta = 1$的红色曲线是幂律分布（在中等和长期状态下），pdf 曲线的斜率$\alpha + 1 = 2$，而其他曲线是拉伸的指数分布。我们无法通过视觉检查来区分这些幂律或拉伸指数曲线之间的差异。幂律工具和泰勒展开可以用于在稍后的渐近分析部分中分析该方案中的复杂性。在$(-\infty, 1 + \epsilon]$且$\epsilon \to 0$（参考渐近分析部分）的范围内，θ越大，曲线的尾部越粗，如图 3d 所示。我们发现 PL 方法在$\theta < 1$时过高估计了缩放参数α，并且当图 3h 中

所示的$\theta > 1$时低估了α。同时，PL 方法一直高估Δ，如图 31 所示。相反，我们的模型再次给出了更好的估计。



(a) β (b) β

图 4:(a)拉伸指数分布的渐近行为。(b)可扩展性。我们的模型可以应用于大型数据集，而幂律方法不能。

5.1.2 渐近行为。我们分析了模型的渐近行为。以拉伸指数为例，其渐近行为导致与幂律分布的混淆和分布的复杂尺度定律。pdf 图片中，拉伸指数分布的函数是对于$x > 0$，$f(x|\theta \neq 1) = \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$。通过在$x \to 0$和$\theta < 1$时应用泰勒展开式，我们得到：

$$f(x \to 0|\theta < 1) = \frac{\alpha}{x^{\theta}}\left[1 - \frac{\alpha}{1-\theta}x^{1-\theta} + O\left(x^{2-2\theta}\right)\right] \quad (23)$$

因此，$f(x \to 0|\theta < 1) \approx \frac{\alpha}{x^{\theta}}$，然而，当Δ很大时，它可以通过短程方案容易地丢失。当 $x \to \infty$ ，$f(x \to \infty|\theta < 1) = \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$，其衰减速度快于幂律衰减$\frac{\alpha}{x^{\theta}}$但比指数衰减$e^{-\frac{\alpha}{1-\theta}x^{1}}$慢。

当$\theta > 1$且$x \to \infty$时，我们得到泰勒展开：

$$f(x \to \infty|\theta > 1) = \frac{\alpha}{x^{\theta}}\left[1 + \frac{\alpha}{(\theta-1)x^{\theta-1}} + O\left(\frac{1}{x^{2\theta-2}}\right)\right] \approx \alpha\frac{1}{x^{\theta}} + \frac{\alpha^2}{\theta-1}\frac{1}{x^{2\theta-1}}$$

近似于幂律分布，其中缩放指数由缩放参数θ和$2\theta - 1$交替控制。当$\theta = 1 + \epsilon$其中$\epsilon \to 0^+$时，看系数$\frac{\alpha^2}{\theta-1} \gg \alpha$，因此$f(x \to \infty|\theta = 1 + \epsilon, \epsilon \to 0^+) \approx \frac{\alpha^2}{\theta-1}\frac{1}{x^{2\theta-1}} = \frac{\alpha^2}{\theta-1}\frac{1}{x^{1+2\epsilon}}$，表示缩放指数$1 + 2\epsilon$,其中$\epsilon \to 0^+$。相反，当$\theta \gg 1$时，$f(x \to \infty|\theta \gg 1) \approx \frac{\alpha}{x^{\theta}}$号，表示缩放指数$\theta$，其中$\theta \gg 1$。因此，在数学上，我们得出结论经验幂律观察可以来自拉伸指数分布的渐近行为。此外，一个有趣的现象是，当θ从 1 增加时，分布曲线首先变得更胖然后回到先前的状态并且更陡峭和更陡峭。上述渐近分析可以通过具有不同θ值的分布的崩溃来验证，如图 4a 所示。

5.1.3 可扩展性。我们在数值上比较了模型的可扩展性和幂律（PL）方法。我们从幂律分布配置 1 生成 N 个样本$\lambda(x|\beta = 0, \alpha = 1, \Delta = 50, \theta = 1)$，并且通过改变$N$，我们绘制了图 4b 中两种方法消耗的平均时间。幂律方法与复杂度$\approx O\left(N^2\right)$成比例。幂律方法的样本大小的经验法则上限是$10^5$。PL 方法的可扩展性更差是由于Δ的网格搜索[5]。但是，我们的方法可以以更快的速度应用于更大的数据集。例如，当$N = 10^5$时，我们得到快约$2 * 10^3$倍。

5.2 现实世界的数据分析

5.2.1 数据集。我们通过来自各种不同人类努力的 16 个真实世界数据集来验证我们的方法。根据数据集的时间性质，我们将它们分类为横截面数据和动态数据。前八个数据集来自横截面上的变化：

（a）赫尔曼梅尔维尔在小说“白鲸记”中出现的词数[16]。

（b）1968 年 2 月至 2006 年 6 月全球恐怖主义袭击事件造成的死亡人数[6]。

（c）每个分类群体在地球上的哺乳动物数量[20]。

（d)1984 年至 2002 年期间受美国停电影响的客户数量[16]。

（e）2000 年美国人口普查中的美国城市人口[5]。

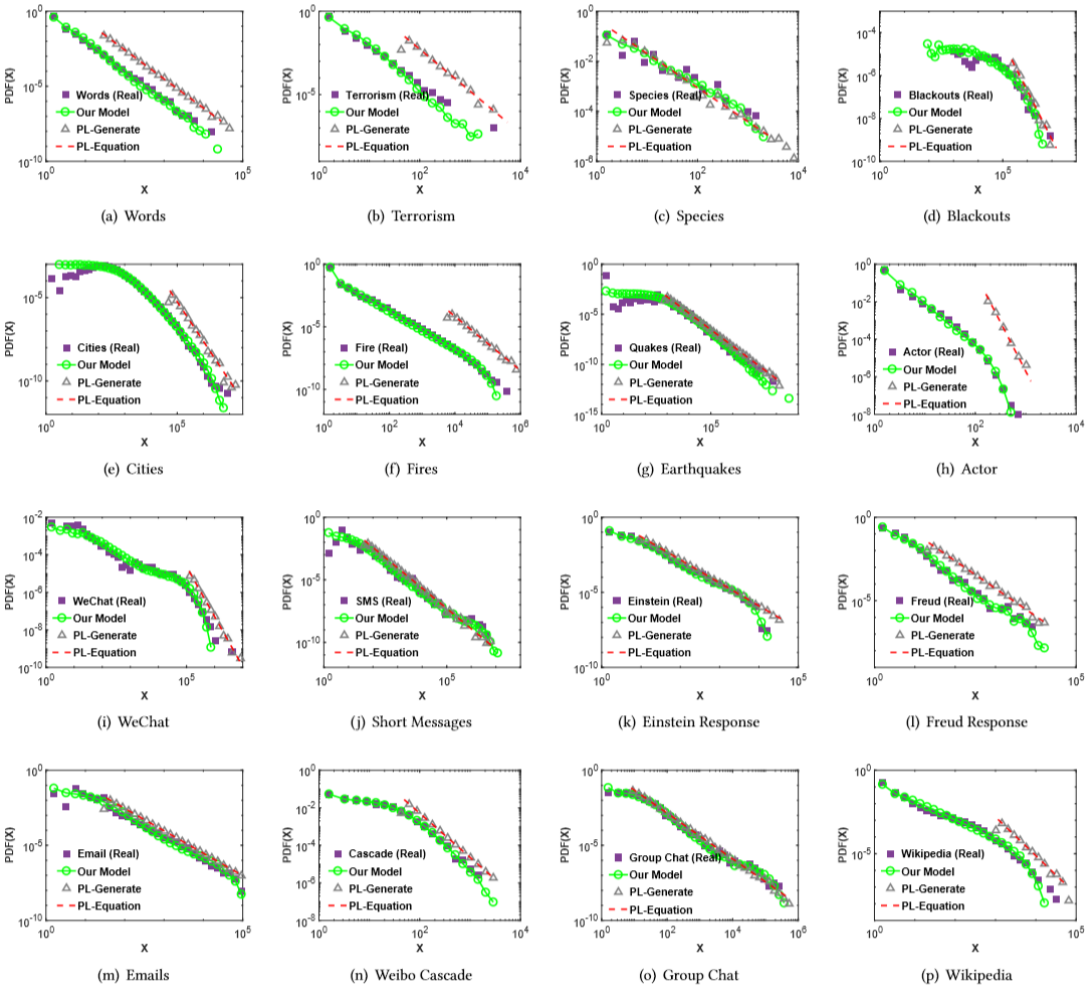


图 5: PDF(X)及其通过我们的方法和基线[5]对来自广泛不同学科的 16 个实际数据集的拟合结果。实际分布比较复杂。我们的方法符合实际。前八个数据集是交叉概念数据集:

(a) 赫尔曼·梅尔维尔的小说《白鲸记》中出现的词语数量。。

(b) 从 1968 年 2 月到 2006 年 6 月，全世界死于恐怖袭击的人数。。

(c) 地球上每个分类群的哺乳动物数量。

(d) 1984 年至 2002 年间，美国受停电影响的消费者数量。

(e) 2000 年美国人口普查中美国城市的人口。

(f) 1986 年至 1996 年间美国发生的森林大火的面积。

(g) 1910 年至 1992 年发生在加利福尼亚的强烈地震。

(h) 电影-演员双向网络中演员的程度。

后八个数据集是动态数据集:

(i) 活动用户在微信中添加连续好友的事件间时间。

(j) 手机用户发送的短信息。

(k)和(l) 爱因斯坦和弗洛伊德一生书信的回复时间。

(m) 一个人在大学里连续三个月发送两封电子邮件的间隔时间。

(n) 腾讯微博信息级联中两次转发的时间间隔。

(o) 腾讯 QQ 群聊天行为的时间间隔。

(p) 一个维基百科条目连续修改的时间间隔。

我们的模型(绿色圆圈)非常适合所有这些数据集(紫色方块)，而最先进的方法(方程曲线的虚线，方程生成的样本的三角形)显示出很大的偏差。

Table 2: Results of real-world datasets. Basic statistics of the 16 datasets, results of the PL method and the results of our method. With respect to KS-Dist error, our model captures the real datasets with much smaller error than the PL Method. The pdf described in PL method is $f(x) = \frac{\alpha_{PL}-1}{x_{min}}(\frac{x}{x_{min}})^{\alpha_{PL}}$.

	Real-World Data Statistics					PL Method			Our Method				
Dataset	N	$Min(X)$	$Max(X)$	$E[X]$	$Std(X)$	$\hat{x}_{min}$	$\hat{\alpha}_{PL}-1$	KS-Dist	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\Lambda}$	$\hat{\theta}$	KS-Dist
(a) Words	18855	1	14086	11.14	148.33	26.00	0.93	0.960	$3.63e-04$	5.00	6.31	1.34	0.319
(b) Terrorism	9101	1	2749	4.35	31.58	50.00	1.52	0.992	$1.18e-10$	5.00	7.20	1.21	0.348
(c) Species	29	1	1425	148.41	324.35	2.00	0.36	0.208	$1.00e-03$	0.33	3.93	0.95	0.138
(d) Blackouts	211	1000	7500000	253868.68	610308.58	230000.00	1.27	0.725	$1.58e-06$	0.03	10000.00	0.80	0.108
(e) Cities	19447	1	8008654	9002.05	77825.05	52457.00	1.37	0.970	$7.99e-07$	0.40	723.77	0.92	0.033
(f) Fire	203785	0	412050	89.56	2098.73	6324.00	1.16	0.997	$3.53e-05$	0.53	0.14	1.00	0.249
(g) Quakes	19302	1.00	63095734.45	24537.21	563830.70	794.33	0.64	0.439	$1.96e-15$	0.37	521.91	0.92	0.095
(h) Actor	383640	1	646	3.83	10.42	162.00	4.21	0.999	$2.05e-02$	1000.00	14.54	2.83	0.388
(i) WeChat	973	0	4073278	57644.40	159193.93	122841.00	1.66	0.887	$1.22e-05$	0.11	30.00	1.12	0.076
(j) SMS	1692	0	4932276	16502.89	201848.27	45.00	0.62	0.556	$4.76e-07$	1.75	26.17	1.17	0.134
(k) Einstein	5943	0	18496	197.32	819.46	9.00	0.53	0.483	$5.09e-04$	10.00	18.55	1.62	0.076
(l) Freud	1190	0	7760	44.38	369.65	22.00	0.66	0.911	$3.06e-04$	10.00	12.83	1.51	0.157
(m) Email	9856	1	228965	711.70	5086.52	34.00	0.49	0.661	$6.57e-05$	7.37	38.05	1.43	0.121
(n) Cascade	3087	0	1586	52.42	102.59	49.00	1.36	0.720	$6.82e-04$	10.00	96.22	1.26	0.021
(o) Group Chat	1055	0	266831	2200.11	16078.36	8.00	0.52	0.245	$1.57e-05$	10.00	40.81	1.47	0.082
(p) Wikipedia	4660	1	29594.	311.58	1143.88	1153.00	1.32	0.940	$4.77e-04$	0.22	1.52	0.92	0.111

（f）1986 年至 1996 年期间在美国发生的野火的大小[16]。

（g）1910 年至 1992 年期间在加利福尼亚发生的地震强度[16]。

（h）电影演员双方网络中的演员程度

最后八个来自人类或社会动态的动态记录：

（i）活跃用户在微信中添加连续朋友的事件间事件[25,26]。

（j）来自移动电话用户的短消息的 IET [23]。

（k），（1）爱因斯坦和弗洛伊德一生中信件通信的响应时间[18]。

（m）在大学 3 个月期间发送两封连续电子邮件的个人之间的时间间隔[1,21]。

（n）腾讯微博信息级联中两次转发的时间间隔[28,29]。

（o）腾讯 QQ 在线群聊聊行为的时间间隔[32]。

（p）连续修订一个维基百科项目的时间间隔[30]。

我们编译并公开所有数据集（参见第 7 节）的可重复性，其中最后八个数据集作为人类和社会动态的第一个数据集。

5.2.2 结果。我们通过回答我们的模型是否可以捕获所有经验数据集来验证我们的方法。我们将我们的方法与[5]中开发的最先进方法进行了比较，表示为 PL 模型，广泛用于拟合可能遵循幂律的胖尾分布。

图 5 绘制了真实数据集，通过 PL 模型和我们的模型拟合结果。我们发现现实世界数据集的分布比纯幂律分布复杂得多。对于不同的数据，其分布表现出不同的多尺度复杂性。然而，我们的模型（绿色圆圈）和所有图中的真实数据（紫色方块）的重叠表明模型的良好性能，即使通过视觉检查。根据合成数据分析，分布中的多尺度复杂性导致严重高估 PL 模型的 xmin，并且我们还在真实数据集中观察到这些偏差，如图 5 所示。因此，$f(x)$由 PL 模型学习的系统偏差，如表示 PL 结果的灰色三角形和表示实际数据的紫色方块的差异所示。

然后我们进行定量分析。给定实数$X = \{x_1, \dots, x_n\}$，我们学习了由 PL 方法和我们的方法建模的$f(x|\Theta)$的参数Θ。然后，我们从我们的方法生成模拟数据样本。从 PL 方法表示为 $X_{\text{our}} = \{x_1, \dots, x_{n'}\}$ 和 $X_{PL} = \{x_1, \dots, x_{n'}\}$。我们通过双样本 Kolmogorov-Smirnov 距离（KS-Dist）评估拟合精度，即 $\text{KS} - \text{Dist} = \max_x |\widehat{F}_i(x) - F(x)|$，越低越好。$F(x)$是从实际数据学习的非参数累积分布，而$\widehat{F}(x)$是通过方法$i$从模拟数据集学习的非参数累积分布。双样本 Kolmogorov-Smirnov 距离广泛用于此标准统计测试任务。为了消除由于生成的样本数量较少而导致的错误，我们设置$n' = 10 * n$。我们总结了表 2 中的数据和结果。我们发现对于所有 16 个数据集，我们的方法比 PI 方法得到更低的误差，

即 KS-Dist，表明我们的方法的优越性。

6 讨论

我们的展示模型的设计原则是：保持简单，捕捉复杂。只有一个参数Δ用于捕捉短尺度方案的复杂性，一个参数β用于捕捉长尺度方案的复杂性，一个参数θ用于包含中等尺度的幂律分布和拉伸指数分布政权。但是，可以通过我们的框架做出更多努力。例如，可以进一步利用混合物重尾模型，对数正态分布。我们可以预期经验数据和拟合结果之间的误差要小得多。但是，无论模型有多复杂，建模参数都应该是可解释的。此外，可以通过贝叶斯框架捕获关于参数的先验知识。应该检查更多真实世界的数据集。应重新检查以前基于在复杂分布上应用 PL 方法的结论。

7 结论

在本文中，我们发现各种经验数据的分布，从艺术，生物学，物理学，地质学，社会科学到计算机科学，从横断面观察到动态记录，具有多尺度的复杂性。我们开发了一个动态框架，以适应现实世界中复杂的分布。通过对数据的生成动力学建模，我们极大地简化了模型的数学形式，但同时生成了大量复杂的分布。提供了有效的推理方法和数据生成算法。我们用一个统一的微分方程来解释这些复杂分布的生成机制，而不是黑盒模型。我们通过各种合成数据集分析模型的属性，并通过各种实际数据集验证我们的模型。我们的模型很好地捕捉了所有这些数据的复杂性。我们的模型可能提供一个框架，以便在经验数据中拟合复杂的分布，并了解它们的生成机制。简而言之，我们总结了我们的贡献如下：

- 统一能力：我们提出了一个通用模型，以适应经验数据中的各种复杂分布，以及推理和模拟算法。
- 简约：有了四个参数，我们的模型有一个简单的形式来捕捉经验分布中的多尺度复杂性。
- 可解释性：我们的模型由统一的生成动力学方程解释。所有参数在随机网络场景中都具有明确的物理意义。
- 实用性：我们的模型准确地拟合了各种学科中复杂的经验数据集分布，并且可以以原则的方式推广到更复杂的案例。

我们在 www.calvinzang.com 开源代码和数据集。

致谢

作者感谢匿名审稿人提供了许多有用的讨论和富有洞察力的建议。这项工作部分得到国家重点基础研究项目 No.2015CB352300，国家自然科学基金重大项目 No.U1611461 的支持;国家自然科学基 No.61772304,61521002,61531006,61702296。感谢清华-腾讯互联网创新技术联合实验室的研究基金，以及 CAST 的青年精英科学家赞助项目。

参考文献

[1] Albert-Laszlo Barabasi.2005. The origin of bursts and heavy tails in human dynamics. Nature 435,7039(2005),207-211.

[2] Albert-Laszl6 Barabasi and Reka Albert.1999. Emergence of scaling in random networks. science 286,5439(1999),509-512.

[3] Anna D Broido and Aaron Clauset.2018. Scale-free networks are rare. arXiv preprint arXiv:1801.03400(2018).

[4] Richard H Byrd, Jean Charles Gilbert, and Jorge Nocedal.2000.A trust region method based on interior point techniques for nonlinear programming. Mathematical Programming 89,1(2000),149-185.

[5] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman.2009. Power-law distributions in empirical data. SIAM review 51,4(2009),661-703.

[6] Aaron Clauset, Maxwell Young, and Kristian Skrede Gleditsch.2007. On the frequency of severe terrorist events. Journal of Conflict Resolution 51,1(2007), 58-87.

[7] Luc Devroye.1986. Sample-based non-uniform random variate generation. In Proceedings of the 18th conference on Winter simulation. ACM,260-265.

[8] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos.1999. On power-law relationships of the internet topology. In ACM SIGCOMM computer communication review, Vol.29. ACM,251-262.

[9] Benjamin H Good, MichaelJ McDonald, Jeffrey E Barrick,

Richard E Lenski, and Michael M Desai.2017. The dynamics of molecular evolution over 60,000 generations. Nature 551,7678(2017),45.

[10] lan Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.2014. Generative adversarial nets. In Advances in neural information processing systems.2672-2680.

[11] Lei Guo, Enhua Tan, Songqing Chen, Zhen Xiao, and Xiaodong Zhang.2008. The stretched exponential distribution of internet media access patterns. In Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing. ACM,283-294.

[12] R Dean Malmgren, Daniel B Stouffer, Adilson E Motter, and Luis AN Amaral.2008. A Poissonian explanation for heavy tails in e-mail communication. Proceedings of the National Academy of Sciences 105,47(2008),18153-18158.

[13] Michael Mitzenmacher.2004.A brief history of generative models for power law and lognormal distributions. Internet mathematics 1,2(2004),226-251.

[14] Kevin P. Murphy.2014. Machine learning,a probabilistic perspective.(2014).

[15] Mitchell G Newberry, Christopher A Ahern, Robin Clark, and Joshua B Plotkin. 2017. Detecting evolutionary forces in language change. Nature 551,7679(2017), 223.

[16] Mark EJ Newman.2005. Power laws, Pareto distributions and Zipf's law. Contemporary physics 46,5(2005),323-351.

[17] John Nolan.2003. Stable distributions: models for heavy-tailed data. Birkhauser New York.

[18] Joao Gama Oliveira and Albert-Laszlo Barabasi.2005. Human dynamics: Darwin and Einstein correspondence patterns. Nature 437,7063(2005),1251-1251.

[19] Douglas Reynolds.2015. Gaussian mixture models. Encyclopedia of biometrics (2015),827-832.

[20] Felisa A Smith,S Kathleen Lyons, SK Ernest, Kate E Jones, Dawn M Kaufman, Tamar Dayan, Pablo A Marquet, James H Brown, and John P Haskell.2003. Body mass of late Quaternary mammals. Ecology 84,12(2003),3403-3403.

[21] Alexei Vazquez, Joao Gama Oliveira, Zoltan Dezso, Kwang-III Goh, Imre Kondor, and Albert-Laszlo Barabasi.2006. Modeling bursts and heavy tails in human dynamics. Physical Review E 73,3(2006),036127.

[22] GWest.2017. Scale: The universal laws of growth, innovation, sustainability and the pace of life in organisms and companies.(2017).

[23] Ye Wu, Changsong Zhou, Jinghua Xiao, Jirgen Kurths, and Hans Joachim Schellnhuber.2010. Evidence for a bimodal distribution in human communication. PNAS 107,44(2010),18803-18808.

[24] Manzil Zaheer, Chun-Liang Li, Barnabas Poczos, and Ruslan Salakhutdinov.2017. GAN Connoisseur: Can GANs Learn Simple 1D Parametric Distributions?(2017).

[25] Chengxi Zang, Peng Cui, and Christos Faloutsos.2016. Beyond Sigmoids: The NetTide Model for Social Network Growth, and Its Applications. In Proceedings of the 22Nd ACM SIGKDD(KDD '16). ACM,2015-2024.

[26] Chengxi Zang, Peng Cui, Christos Faloutsos, and Wenwu Zhu.2017. Long Short Memory Process: Modeling Growth Dynamics of Microscopic Social Connectivity. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,565-574.

[27] Chengxi Zang, Peng Cui, Christos Faloutsos, and Wenwu Zhu.2018. On Power Law Growth of Social Networks. IEEE Transactions on Knowledge and Data Engineering(2018).

[28] Chengxi Zang, Peng Cui, Chaoming Song, Christos Faloutsos, and Wenwu Zhu.2017. Quantifying Structural Patterns of Information Cascades. In Proceedings of the 26th International Conference on WWW Companion.867-868.

[29] Chengxi Zang, Peng Cui, Chaoming Song, Christos Faloutsos, and Wenwu Zhu.2017. Structural patterns of information cascades and their implications for dynamics and semantics. arXiv preprint arXiv:1708.02377(2017).

[30] Yilong Zha, Tao Zhou, and Changsong Zhou.2016. Unfolding large-scale onlinecollaborative human dynamics. Proceedings of the National Academy of Sciences 113,51(2016),14627-14632.

[31] Tianyang Zhang, Peng Cui, Christos Faloutsos, Yunfei Lu, Hao Ye, Wenwu Zhu, and Shiqiang Yang.2016. Come-and-go patterns of group evolution:A dynamic model. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,1355-1364.

[32] Tianyang Zhang, Peng Cui, Chaoming Song, Wenwu Zhu, and Shiqiang Yang.2016.A multiscale survival process for modeling human activity patterns. PloS one 11,3(2016),e0151473.

本文所提出方法存在的问题

对于本文所提出方法的改进，或者对本文提出的问题的全新的解决方法