复杂社交系统的数据驱动动力学建模研究

（申请清华大学工学博士学位论文）

培 养 单 位：计 算 机 科 学 与 技 术 系
学　　科：计 算 机 科 学 与 技 术
研　究　生：臧　承　熙
指 导 教 师：朱　文　武　教　授

二〇一八年十二月

Data-Driven Dynamical Modeling of Complex Social Systems

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the professional degree of

Doctor of Engineering

by

Chengxi Zang

(Computer Science and Technology)

Dissertation Supervisor : Professor Wenwu Zhu

December, 2018

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》，向国家图书馆报送可以公开的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

复杂社交系统，如微信、微博、脸书，推特，抖音等，接受用户的输入信号，产生输出信号，满足用户在系统中的各种信息需求，随时间不断演化，并产生社交大数据。解释复杂社交系统演化机制是本文的核心课题，其对理解自然界中复杂系统如何运行等科学问题，和在社交系统中如何提供可解释的推荐、广告投放等计算服务都具有重要意义。然而，研究复杂社交系统的演化机制极具挑战，其体现在：一，复杂社交系统由亿万链接的个体组成，其输出数据也多呈现链接的网络结构，呈现结构复杂性；二，复杂社交系统中微观个体动态地相互影响，导致其宏观整体输出与微观个体输入之和差别极大，即非线性；或在短时间尺度大规模涌现，即爆发性；两者合称为动态复杂性；三，复杂社交系统在微观呈现随机无序状态，但在宏观呈现确定有序状态，其演化呈现多尺度复杂性。传统分析方法基于物理动力学模型，试图刻画复杂社交系统动态变化的现象，并揭示其变化的动力学机制。然而，真实复杂社交系统如微信、微博等有着数十亿节点和数百亿条边，其动态变化现象体现在宏观网络演化，微观个体社交行为，及信息在网络上的动态传播等不同尺度及不同场景。通过数据驱动的方式，我们发现了上述复杂社交系统在不同尺度不同场景的很多全新的复杂现象，而传统动力学模型不能刻画。

本文创新地将计算机科学和物理学理论融合，通过数据驱动的方式，对复杂社交系统演化机制进行了动力学建模。具体而言，本文研究了复杂社交系统演化的三个核心子课题，包括：一，社交网络多尺度演化规律发现和建模，其旨在回答复杂社交系统在不同尺度如何增长的问题；二，信息流在网络中传播的复杂模式生成，其旨在回答信息流在复杂网络上如何传播的问题；三，宏观分布函数的微观动力学起源定理，其旨在回答如何连接微观行为和宏观现象的问题。我们所提出的研究方法在微信（首次）和腾讯微博等大规模社交数据上进行了实验验证。对于社交网络多尺度演化规律发现和建模子课题，我们发现宏观网络节点和边随时间的幂律增长，提出网潮模型，准确刻画了网络演化规律，提升了对网络演化长期预测的性能；并进一步发现了微观网络演化和个体社交行为的长期非线性和短期爆发性增长，提出长短记忆随机过程，准确刻画了个体动态随机行为，提升了微观行为的预测性能和可解释性。对于信息流在网络中传播的复杂模式生成子课题，我们发现了真实社交系统中信息流的复杂结构，通过量化信息流结构发现其复杂几何模式，并提出数据驱动的异构分支随机过程，解释了真实社交系统中信息流复杂模式生成的机制，极大提升了拟合信息流复杂几何结构的准确性。对

于分布函数的动力学起源子课题，我们提出了一种构造化的解释，认为宏观确定分布函数由随机到达的微观个体通过确定的动力学过程得到，给出了连接微观随机性和宏观确定性的定理，并极大提升了宏观分布拟合准确率和可解释性。

我们通过数据驱动的动力学建模，很好地刻画了观测到的全新复杂现象，并进一步从复杂系统的角度解释其动力学演化机制，有效地结合了数据科学的可计算性和统计物理的可解释性。我们提出的数据驱动的动力学建模研究方法，试图为建模，理解，预测真实世界中的大规模复杂社交系统奠定一定的理论基础。

关键词：复杂社交系统；数据驱动的动力学建模；微信；网络多尺度演化；网潮模型；长短记忆随机过程；信息流模式生成；异构分支过程；分布函数动力学起源；复杂多尺度分布拟合

Abstract

Complex social systems, such as WeChat, Weibo, Facebook, Twitter, Tik Tok, etc., accept users' input signals, generate output signals, meet users' various information and social media needs, constantly evolve over time, and generate social big data. Explaining the evolutionary mechanisms of complex social systems is the core topic of this paper. It is of great significance to understand the scientific problems of how complex systems evolve in nature, and how to provide interpretable recommendation, advertising and other computing-based services in social systems. However, to study the evolutionary mechanisms of complex social systems is extremely challenging. First, the complex social system is composed of hundreds of millions of linked individuals, and its output signals are always in the form of networked data, exhibiting structural complexity. Second, micro-individuals in complex social systems interact with each other dynamically, resulting in a large difference between the macroscopic output and the sum of micro-individual inputs, that is, nonlinearity; or resulting in large-scale emergence on a short time scale, that is, burst; these two are collectively called dynamic complexity. The complex social system presents a random disorder state at the microscopic level, but the macroscopic phenomena determine the ordered state, and thus its evolutionary process exhibits multi-scale complexity. Traditional analysis methods are based on physical dynamical models, trying to characterize the dynamics of complex social systems and reveal the mechanisms of their changing. However, complex social systems in the real world such as WeChat, Weibo, etc. have billions of nodes and tens of billions of edges, and their dynamic phenomena are reflected in different scales and different scenarios such as macro network evolution, micro-individual social behavior, dynamic propagation of information on the network and so on. Through a data-driven approach, we have discovered many new complex phenomena of the above-mentioned complex social systems in different scales and different scenes, and the traditional dynamical models failed in capturing the observed phenomena.

This paper tries to combine the theories in both computer science and physics to model the evolutionary mechanisms of complex social systems through a data-driven dynamical modeling approach. Specifically, this paper studies three core sub-topics of the evolution of complex social systems. First, the discovery and modeling of multi-scale

evolution of social networks, which aims to answer the question of how complex social systems grow at different scales. Second, Complex pattern formation of information flow, which is designed to answer the question of how information flows over complex social networks. Third, the theorem of the dynamic origin of distribution functions, which is designed to answer questions about how to connect micro-behavior and macro-phenomena. Our proposed methods have been experimentally verified on large-scale social network datasets such as WeChat (first time), Tencent Weibo and so on. For the discovery and modeling of multi-scale evolution of social networks, we find that the nodes and links of social networks follow the power-law growth over time, and we propose the NetTide model which accurately captures the evolutionary power law, and improves the performance of long-term network evolutionary prediction. Furthermore, we find long-term nonlinear growth and short-term bursty growth of micro-network evolution. We propose the long and short memory stochastic process, which accurately describes the dynamic random behavior of individuals and improves the predictive performance and interpretability of micro-social behaviors. For the complex pattern formation of information flow in the network, we find the complex structure of information flow in the real social systems, and then quantify their complex geometric patterns, and propose a data-driven heterogeneous branching processes to explain the mechanisms of complex pattern formation of information flow, which greatly improves the accuracy of fitting complex geometric structures. For the dynamic origins of distribution functions, we propose to explain that the distribution function is obtained from the randomly arriving microscopic individuals through a determined dynamic process, which connects the micro-randomness and the macro-determinism. The theorem greatly improves the accuracy and interpretability of fitting complex multi-scale distributions.

Through a data-driven dynamical modeling approach, we have well characterized the new observed phenomena in the complex social systems, and further explained their dynamic evolutionary mechanisms from the perspective of complex systems, which effectively combine the computability of data science and interpretability of statistical physics. Our proposed data-driven dynamical modeling approach attempts to lay a theoretical foundation for modeling, understanding, and predicting large-scale complex social systems in the real world.

Key words: Complex Social Systems; Data-Driven Dynamical Modeling; WeChat; Multiscale Network Evolution; NetTide Model; Long Short Memory Process;

Pattern Formation of Information Flow; Heterogeneous Branching Process;
Dynamic Origins of Distributions; Fitting Complex Multiscale Distributions

目 录

第 1 章　引言

如果只能用一句话概括本文，那就是：**本文以大数据驱动的动力学建模方法，试图解释复杂社交系统的演化机制。**

1.1　复杂社交系统

图 1.1　**复杂社交系统**

什么是复杂社交系统（Complex Social Systems）？复杂社交系统由连接的人组成。人和人之间相互连接，相互影响，导致其宏观整体特性与微观个体特性之和差别极大。如图1.2所示，我们把复杂社交系统当做一个黑盒系统，有输入信号，有输出信号，并满足用户在复杂社交系统中的各种信息需求。通过观测复杂社交系统的结构，复杂社交系统的输入和输出随时间的变换，我们收集了大量的数据，即复杂社交系统大数据。简单的输入信号，有可能产生极其复杂且意外的输出信号。而通过分析描述系统和输入输出的数据，我们试图打开复杂社交系统的面纱，分析其演化运行的规律和机制。本文的主题，正是尝试回答复杂社交系统的演化运行机制是什么的问题。

在线社交网络，如微信、微博、脸书，推特，抖音等，第一次提供了支持我们了解复杂社交系统的详细数据。我们以中国最大的（在线）社交网络，也是支持

本文研究的社交网络数据之一--微信（WeChat）[①]--举例说明：在 2018 年第三季度，微信每月活跃用户达到 10.8 亿，是一个覆盖中国几乎所有智能手机用户的社交系统。通过微信，人们可以通过短信，语音，群聊，视频通话等多种方式进行通信，构建出了上千亿规模的社交连接。此外，微信已经远远超出一款简单的通讯工具，它成为许多人的生活方式，记录着人行为的点点滴滴。例如，人们可以在“朋友圈”（Moments，一个社交信息流媒体功能，允许用户发布图像，文本，视频，音乐等等，好友所发表信息接收评论和喜欢等反馈）中分享他们美好的生活。此外，拥有 6 亿活跃的移动支付用户，提供中国最大的支付服务之一，涵盖数字钱包，余额账户，红包，幸运钱，现金转账，金融投资等功能。人们通过微信在餐厅付款，超市甚至供应商通过扫描 QR 码而不是使用现金或信用卡。此外，微信提供官方账号，城市服务，迷你程序，微信索引，新闻源，搜索等功能。通过微信提供的便捷个性化服务，人们现在可以在交朋友或开展业务时交换微信账号而不是名片。它是一个适用于中国及世界的连接“一切”的超级应用程序，是一个复杂社交系统。而每个用户在这个复杂社交系统中的每一个行为，都被记录产生海量的微观用户行为数据，而且反作用于微信系统使其随时间不断变化。

复杂社交系统，就像许多自然现象一样，时刻都在变化着，是动态的。以微信社交系统为例，当我们互相发短信，添加新朋友，发布照片，付款，加入群聊，进行对话，扫描 QR 码等等时，所有这些行为都会随着时间的推移动态变化，而每个用户动态变化的行为并致使整个系统无时无刻不在发生状态改变。例如，微信于 2011 年首次发布，只有少数用户，然后发展成为拥有数 10 亿用户和数百亿社交链接的超大型复杂社交网络。注册微信账号后，新用户建立与现有用户的社交链接，并向尚未注册的朋友推荐微信应用，即社交网络在不断演化。形成社会群体是一种内在的人性。通过网络，人们可以构建具有特定主题的各种群组。新的小组成员加入小组并且现有成员退出小组，导致组群在动态演化。社交媒体现在是我们获取信息的主要平台。信息在社交网络上产生，传播和消费，产生信息级联现象等等。复杂社交系统，在时间维度展开，产生了大量的动态变化的数据，覆盖不同尺度，不同场景。

传统统计物理等学科，例如对热力学系统的研究，无法跟踪和记录每一个原子在系统中的运动信息。类似，传统社会学研究同样没有处理过如此复杂且有海量数据的复杂社交系统场景。但是，随着在线社交网络的普及，复杂社交系统不同尺度不同场景的演化数据被详细记录下来了，包括网络演化数据，微观个体用户行为数据，信息流在网络中传播数据等等。正是这些随时间演化的，动态的社

① weixin.qq.com

交大数据的产生，使得我们以大数据驱动的方式探究复杂社交系统演化运行规律和机制成为可能。

1.2　研究意义

许多自然界的法则有着动力学的起源，我们也希望探究动态变化的复杂社交系统的动力学运行机制。通过真实世界数据的驱动，比如微信社交系统等，第一次为我们提供了对人类行为，网络演化，组群演化和信息级联现象等等真实世界复杂社交系统的大规模实证研究，支持我们对复杂社交系统运行规律的探索。通过真实世界大数据驱动的研究，我们对包括社会科学，经济，统计物理学和计算机科学等等复杂性研究领域的经典结论再次审视，并进一步探究未解决的问题。确实，通过大数据驱动的方式，我们在微信等复杂社交系统中观测到的很多全新的现象，但是他们和经典的复杂性研究中的理论并不一致。举例而言，微信传播即用户增长过程，是否遵循指数增长并长期呈现 Bass 模型所预测的 S-型曲线[1]？还是用户匀速到来呈现线性增长[2]？面对微信如此复杂且庞大的真实社交系统，很多现象尚不为人知。例如，复杂社交系统中社交连接如何增长呢？而在计算机领域对社交网络的研究，侧重于数据处理技术和面向性能的推荐，预测等等应用，缺乏通过物理动力学的视角来理解复杂社交系统的运行机制。所以，数据驱动的动力学研究，试图通过融合数据科学的可计算性和物理动力学的可解释性，对发现真实世界复杂社交系统的运行模式，理解其动力学运行机制，提出处理复杂动态数据的建模方法论等等，都有着重大科学研究意义。

另一方面，许多问题对工业应用同样至关重要。我们可以预测微信中每个用户的下次在线购买行为吗？我们能否通过组群的动态行为数据发现异常组群，而该组群有可能正在招募恐怖分子成员，或者传播非法信息？新发布的应用版本在发布时有多大的潜在受欢迎程度？发布下一版微信或投放广告的正确时机是什么时候？我们可以预测一篇文章是否会受欢迎吗？微信（或 Facebook，Twitter 等）下个月会有多少新用户（公司估值）？我们可以预测用户什么时候离开社交应用嘛？我们可以将我们在微信中开发的知识应用到其他社交系统吗？我们可以构建一个模拟的复杂社交系统，来进行各种在现实世界不能进行的实验嘛？其他应用包括社交产品的推广传播，流失预测，产品供应，政策制定等等。所有这些问题都以理解和建模复杂社交系统动态运行规律为核心。

1.3　研究挑战

然而，理解和建模复杂社交系统的演化机制是非常具有挑战性的，其总结为如下四个方面：

- **复杂社交大数据**。[①] 过去由于缺乏记录人类动态行为，网络演化过程，群体行为和信息级联传播的详细的数据，我们无法对复杂社交系统进行细致的微观层面的分析，更无法了解其复杂系统运行的内在机制。而微信，微博等在线社交网络，为我们提供有关现实世界复杂社交网络不同场景不同尺度的动态数据记录。但是，机遇与挑战并存，该数据极其复杂，数据量极大。例如，我们在网络宏观演化工作中，使用了微信从上线及之后两年的网络演化数据，就含有超 3 亿用户，47.5 亿条带创建时间戳的社交连接。而腾讯微博信息传播的一个数据集包括 7 天内超过 1 亿用户的 4.32 亿条信息级联传播，超过 5.6 亿条带时间戳和超过 1 亿节点的网络数据等等。如何处理如此规模的网络及动态数据带了极大的挑战。
- **复杂社交系统的结构复杂性**。复杂社交系统由亿万个异构个体组成，他们并不是独立的，而是链接在一起，呈现复杂网络结构。其次，系统输出信号也多是复杂的网络链接数据，如信息传播数据等等。如何分析和建模复杂社交系统的结构复杂性极具挑战。
- **复杂社交系统的动态复杂性**。复杂社交系统具有动态复杂性，其表现在非线性，爆发无处不在。复杂社交系统中微观个体动态地相互影响，导致其宏观整体输出与微观个体输入之和差别极大，输出信号不正比于输入信号，即非线性性。爆发是指短时间尺度发生的大规模涌现现象，例如比特币价格在短时间内的爆发增长等等。动态复杂性广泛存在于不同尺度场景下，例如社交系统宏观人口增长，微观个体行为，网络上信息传播等等，都随时间呈现动态复杂性。如何分析和建模复杂社交系统的动态复杂性极具挑战。
- **复杂社交系统的多尺度复杂性**。复杂社交系统在时间和空间都呈现多尺度复杂性。复杂社交系统每天都产生海量数据，包括个体行为，局部现象等。这些数据在微观呈现随机无序状态，但在宏观呈现确定有序状态，如集群行为，全局现象。如何在结构和时间维度连接微观和宏观，如何从底层微观数据推测上层宏观规律，如何从多尺度层面理解复杂社交系统极具挑战。

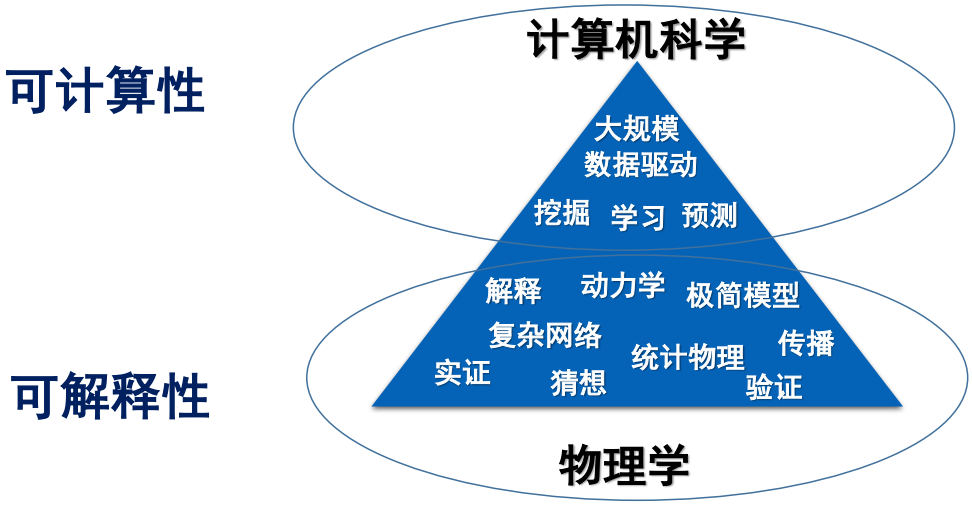


图 1.2　**研究思路：融合物理学的可解释性和计算机科学的可计算性**

1.4　解决思路

为了解决以上挑战，解释复杂社交系统演化机制，我们需要创新地将物理和计算机科学融合：面对社交大数据，我们需要通过计算机科学提供可计算性，在大规模数据上，以数据驱动方式，挖掘，学习，和预测；另一方面，我们需要通过物理融入可解释性，引入动力学建模来刻画复杂系统的内在演化过程，通过统计物理，复杂网络的理论，处理网络结构复杂性，动态复杂性和多尺度复杂性的社交大数据。简言之，解释复杂社交系统演化机制需要计算机科学与物理学的交叉融合。

1.5　论文贡献及结构

本文尝试将计算机科学和物理学理论融合，通过数据驱动动力学方法，对复杂社交系统演化机制进行了解释和建模，并通过微信（首次）和腾讯微博等大规模社交数据上对所提出的研究方法进行了实证。具体而言，本文研究了复杂社交系统演化的三个核心子课题：一，社交网络多尺度演化规律发现和建模，其旨在回答复杂社交系统在不同尺度如何增长的问题；二，信息流在网络中传播的复杂模式生成，其旨在回答信息流在复杂网络上如何传播的问题；三，宏观分布函数的微观动力学起源定理，其旨在回答如何连接微观行为和宏观现象的问题。

① 我们可以访问的所有社交网络数据都是匿名的行为数据，无权限访问内容数据。我们遵循了严格的隐私政策。

- **社交网络多尺度演化规律发现和建模**。我们从宏观和微观多尺度对复杂社交系统演化进行了分析，其分别为：
  - 宏观社交网络演化规律发现和建模： 我们研究了多个复杂社交系统的宏观演化过程，包括了中国最大社交网络微信自上线两年内网络演化的详细过程，覆盖 3 亿用户，47.5 亿条带创建时间的社交连接。我们发现多个社交系统用户数的增长不是指数增长，也不是线性增长，而是幂律增长。进一步，我们首先提出了社交连接的增长也是幂律增长的现象。我们给出了产生复杂社交系统节点和连接幂律增长的机制，并通过动力学方程网潮模型和对应的微观随机过程建模。我们的动力学方程，能产生广泛的复杂动力学增长现象，并准确地拟合和预测了真实社交系统的增长规律。（第二章）
  - 微观 社交网络演化及个体社交行为规律发现和建模： 我们研究了驱动复杂社交系统宏观演化的微观个体社交行为，发现微观个体加好友行为呈现极大的随机性和异构性。我们发现微观个体社交行为在长期遵循非线性幂律随机增长，在短期呈现爆发随机增长。我们提出了三个机制，即平均效应，多尺度效应和相关效应，来控制不同尺度下的个体随机行为模式，并给出一个长短记忆随机过程建模。通过模型分析大规模加好友行为，我们进一步发现了微观用户加好友的规律和典型行为类型，并用于用户画像聚类和异常检测等应用。（第三章）
- **信息流在网络中传播的复杂模式生成**。 我们研究了信息流在网络中传播产生有规律的复杂几何模式的过程。尽管越来越多的研究旨在了解信息流的传播机制，但对于这些传播模式的几何形状以及它们在传播过程中是如何形成的却知之甚少。通过探索了大规模在线社交媒体数据集中提取的 4.32 亿个信息流模式，我们在一个三维度量空间中发现了信息流传播结构的复杂几何模式。相比之下，对信息流传播结构的现有理解仅限于扇形展开或狭窄的树状的几何形状。我们发现了控制信息流复杂几何模式生成的三个关键因素：异质性，集体性，和记忆性。之后，我们提出了一个包含这些因素的随机过程模型，证明它可以成功复现真实信息流传播模式中发现的复杂几何模式。我们的发现为信息流的微观机制提供了理论基础，其可能的应用包括对信息的预测，控制和政策决策等等。（第四章）
- **分布函数的动力学起源定理**。我们总结了许多科学研究都遵循如下模式：从一个系统的截面状态数据来推断其动力学生成/演化机制。但是，正式且系统地学习它们之间关系的研究却少之又少。我们将复杂的截面状态数据视为

通过确定的动态系统生成，该系统以均匀的随机信号作为输入。我们构造了（截面状态）概率分布函数与其动力学生成系统之间的一个等价关系，然后开发了一个框架来从截面状态数据，或数据分布函数，来推断其动力学生成过程。通过这样的框架，我们能够从各种分布中发现新的动力学生成机制，而且可以从各种动力学生成机制中发现新的概率分布函数。我们通过合成数据和真实数据验证了我们的框架。实验结果表明，我们的框架能够准确地发现和拟合各种数据分布函数的动力学生成过程。我们的研究有助于发现现实世界中复杂截面数据的未知动力学生成机制（第五章）。进一步，我们给出一个统计模型，来拟合和解释真实世界中的复杂分布函数，而最常用的统计模型面对现实复杂数据表现出了系统误差。我们展示了刻画动力学机制的优势--通过刻画相对简单的动力学产生过程，来简化对复杂生成现象之间刻画的模型复杂度。（第六章）

最后在第七章，我们给出论文总结和对未来工作的展望。我们指出将数据科学（数据挖掘，机器学习）的可计算性和物理动力学（统计物理，网络科学）的可解释性相结合的研究范式，来分析复杂社交系统大数据，并进一步探究其演化规律。我们提出的数据驱动的动力学建模研究方法，试图为建模，理解，预测真实大规模复杂社交系统奠定一定的理论基础。

第 2 章　宏观社交网络演化规律发现和建模

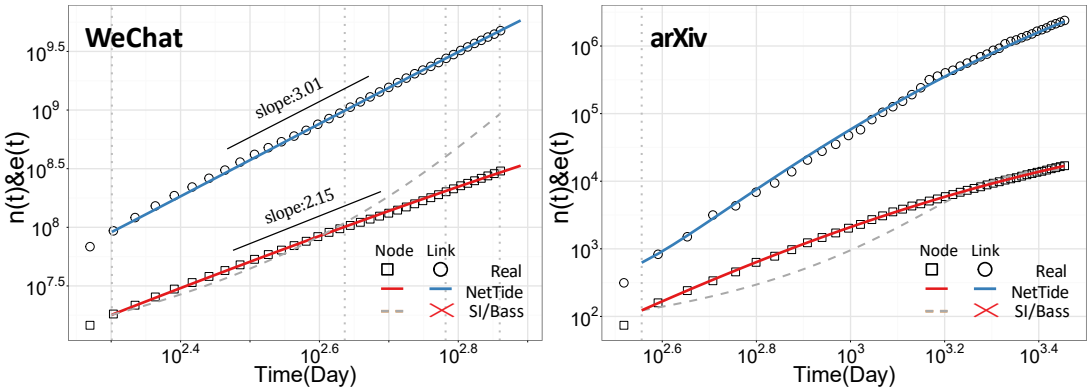
真实在线社交网络，例如微信（WeChat）[①]和脸书（Facebook）[②]，其用户数随时间如何动态地增长？他们早期是否呈现出指数增长模式，如著名的 Bass 模型所描述的那样？此外，社交网络的社交链接（边）如何增长？很少有模型给出社交网络边的增长规律。第一次，我们验证了中国最大的在线社交网络 – 微信，和其他多个真实的社交网络的演化过程。我们发现了社交节点和社交链接数随时间都呈现幂律增长（Power-law Growth）的规律。社交节点的增长规律完全打破了经典教科书所表述的 Simoid 曲线（即 S-型增长曲线）规律；社交链接的幂律增长规律第一次被发现提出。我们给出幂律增长的动力学机制，并提出了网潮（NetTide）模型，包含刻画节点和边增长的两个动力学方程。我们的模型很好地刻画了真实社交网络的动力学增长过程。我们地网潮节点模型有着简洁地参数，并且包含过去很多经典的增长模型为其特例。社交链接的动力学方程是第一次被提出，它不但很准确地刻画了真实社交网络边的增长过程，而且刻画了网络逐渐变稠密（Densification）的现象。进一步，我们给出两个微观随机过程，其一通过生存分析模型刻画了节点和边的随机增长过程，其二通过微观随机交互产生网络。我们提出的微观随机过程产生了逼真的网络增长动力学曲线。我们将网潮模型用于对微信未来的增长预测，其准确预测了两年后（730 天）的用户数，误差仅为三个百分点。

2.1　引言

社交系统的增长现象，包括社交网络的增长[3,4]，社交群组的增长[5]，和信息传播覆盖人数的增长[6–8] 等等，是社交系统演化机制研究的核心问题之一。但是，由于缺乏记录真实社交系统演化的详细数据，我们尚不清楚社交系统的增长规律及可能的机制。所以，我们很难回答如下涉及社交系统长期演化的问题，例如：推特（Twitter）下个月有多少会员？微信（或 Facebook 或 Google+）明年会有多少社交链接呢？社交网络（或信仰，宗教或流行病）用户的数量至关重要（社会产品的增长，供应，社会政策变化等等），对该问题也进行了广泛的研究（参见相关工作小节）。形成鲜明对比的是，虽然社交链接数量也很重要（例如，FaceBook 中连接良好的节点不太倾向于流失；大脑中连接良好的神经元可能表明对阿尔茨海默病的抵抗力等），但是对社交链接数量增长的研究却少之又少。

① www.wechat.com/en/

② www.facebook.com



(a) 微信　　(b) arXiv

图 2.1　我们发现（a）微信 WeChat 和（b）arXiv 的节点初期（如图方块标记）随时间都遵循幂律增长。我们提出的网潮-节点模型 NetTide-Node（实心红线）很好地拟合了真实数据，但是 SI 或 Bass 模型（灰色虚线）的 Sigmoid 曲线远远偏离实际。此外，社交链接数随着时间也遵循幂律增长（圆圈和我们的网潮-链接模型 NetTide-Link 如纯蓝色实线所示）。请注意，刻画链接增长的基准动力学方模型程是不存在的（SI 或 Bass 被划掉）。图 a-b 均为双对数坐标。

来自跨学科的研究人员在近二十年研究了网络增长的现象[2,9–13]，他们在理解无标度（scale-free）网络的生成，网络链接的稠密化现象，网络直径收缩现象等方面取得了重大进展。网络增长模型包括著名的 Barabási-Albert 模型和它的各种变种[2,14,15]。但是所有上述模型都假设网络节点是均匀线性到来的。著名的 Bass 模型[1] 和传染病模型 Susceptible-Infected (SI)[16]，可以产生 S-型 Sigmoid 曲线，其有着初期指数增长模式。此外，所有上述模型都没有研究社交链接，也就是网络边的增长规律。简言之，本文的重点是回答关于复杂社交系统宏观动态增长的如下三个问题：

- 复杂社交系统节点数量 $n(t)$ 随时间如何增长？
- 复杂社交系统链接的数量 $e(t)$ 随时间如何增长？
- 我们可以产生 $n(t)$ 和 $e(t)$ 的逼真的（随机）增长过程么？

读者可能会认为，至少第一个问题已经有了答案：Sigmoid 增长（这是 SI 和 Bass 模型的解决方案）。然而，现实与理论相违背，如图 2.1 a-b 所示，真实社交网络节点和边都表现出幂律增长（Power-law Growth）。具体而言，我们研究了四个真实社交网络的演化过程，包括微信，arXiv[17]，Enron[18] 和微博[19]，分别代表在线社交网络，科学合作社交网络，企业社交网络和信息级联网络。以微信为例，我们研究了其从零到三亿（300 million）节点和四十七点五亿（4.75 billion）条链接的详细演化过程。我们发现虽然四个社交网络的增长曲线具有不同的形状，但它们都遵循幂律增长模式。更具体来说，我们发现微信的动态增长遵循幂律增长，节点

数随时间增长的幂律指数为 2.15，边数随时间增长的幂律指数为 3.01（参阅图 2.1 a），arXiv 的动态增长在达到饱和前同样遵循幂律增长（参阅图 2.1 b）。上述观测到的现象，超出了我们对社会网络节点数或是创新传播（Diffusion of Innovations）数的指数增长或均匀增长的预期。

由于被广泛采用的 Sigmoid 模型和现实人口增长模型相矛盾，我们需要一个更加真实好用的动力学增长模型。我们希望一个好的模型应满足如下特征：

- **简约性**：模型应该只具有最少的必要参数，并且可以产生幂律增长。
- **通用性**：模型应该是通用的，最好可以包含传统增长模型，如 Bass，SI 和 Log-Logistic 等增长作为特殊情况。
- **链接增长**：模型应该能够刻画社交链接的动态增长规律。
- **符合直觉**：模型应该易于解释。
- **微观生成器**：宏观人口动态模型应该有着微观层面的随机过程，从微观层面的模拟可以产生逼真的宏观动态增长模式。

我们提出了全新的网络动力学模型，名为网潮模型（NetTide），用于刻画宏观网络动力学增长过程。网潮 NetTide 模型由两个组件组成，分别是节点动力学方程 NetTide-Node 和链接的动力学方程 NetTide-Link。正如我们稍后展示的那样，我们的 NetTide 模型实现了上述特征，并且很好的刻画了多个全然不同的真正社交网络增长演化规律。

我们的模型可对动态增长过程进行预测，聚类和异常值检测[20] 等等。例如，我们发现网潮 NetTide 模型能够预测微信未来两年（730 天）的增长情况，在大约三个百分点的误差范围内。这是令人印象深刻的结果，因为社会系统是复杂和非线性的，对于长期预测极其困难。作为参考，大多数预测工作通常对未来一步进行预测[21,22]，而不是试图解释并预测复杂系统的长期动态增长规律。

网潮模型的直观理解启发我们构造了两个随机过程生成器，即 NetTide-Survival 模型和 NetTide-Process 模型，它们都能产生逼真的节点和链接的随机动态增长过程。NetTide-Survival 通过对节点和链接的危险速率（Hazard Rate）建模来连接宏观网潮 NetTide 模型和生存分析框架。进一步，NetTide-Process 将网潮模型的宏观确定性规律解释为源自网络内微观个体的随机交互。通过广泛的数值模拟，两个随机过程生成器都可以产生真实的动力学增长曲线。总之，我们工作的贡献可以归纳为：

- **全新的动力学模型**: 网潮 NetTide 模型，它是简约的，包含许多增长模型为其特例，提供了第一个网络链接增长的动力学方程，而且直观容易解释。
- **准确性**: 网潮模型可以准确地拟合多个不同的真实社交网络的动态增长过

程。

- **实用性**: 网潮模型提供了出色的长期预测性能，可以预测微信未来两年后的用户数量，并且保持错误率在三个百分点内。
- **随机过程生成器**: 网潮模型对应两个随机过程生成器，即 NetTide-Survival 和 NetTide-Process，它们可以生成逼真的社交网络的随机动力学增长曲线。

本章的大纲如下：我们在本章第 2.2节给出了相关工作调查，本章第 2.3节介绍模型方法，本章第 2.4节介绍实验结果，最后在本章第 2.5节总结讨论。此外，本章所涉及部分数据是公开的，参见[17,18]，我们的代码是开源的，参见 https://github.com/calvin-zcx/NetTide。

2.2 相关工作

我们所研究的问题与网络演化，增长模型和人类行为动力学等课题密切相关，所以我们主要回顾了这三个领域的相关工作。

2.2.1 网络演化模型

研究网络演化规律的先驱性工作表明，真实网络的动力学增长过程在塑造其结构方面起着至关重要的作用，例如网络的幂律度分布[2], 逐渐缩短的网络直径[9], 网络边的逐渐稠密[9] 等等。然而，所有这些网络演化模型都假设节点在演化过程中是匀速到来的，例如 Barabási-Albert model (BA) 模型[2] 和其各种变体[15]。其他一些工作将真实网络节点的动态增长过程当做研究其他性质的输入，所以并没有刻意研究节点的动力学增长模式[9,11,23–26]。

另一方面，现有文献很大程度上忽略了社交链接（或称作网络的边）的动态增长。一些工作表明内部链接的双重优先连接，亦或者是随机连接，都可能使网络更加同质，但是其中每条链接的增长率也被认为是匀速的[14,27,28]。最近，信息传播扩散对链接创建的影响在[29,30] 中进行了研究，此外,[31] 提出了多变量 Hawkes 过程模型（Coevolve）来捕捉链接和信息扩散的微观随机演化过程。但是，它有两个严重的问题：其时间计算复杂度为 $O(N^2)$，其中 N 表示[31] 中提到的事件数量; 此外，基于 Hawkes 随机过程的模型并**不能** 产生我们观察到的幂律增长（指数 2.11，3.01，如图2.1所示），而主要产生线性或指数增长（请参阅相关工作增长模型中有关 Hawkes 过程的讨论）。社会网络增长的时间和空间相关性在[32,33] 中进行了研究。

表 2.1　模型能力对比表。只有我们的模型有着所有特性。

	网络增长模型			**增长动力学模型**					**网潮模型**
能力	BA	FF	Coevolve	SI	BASS	CS	SpikeM	PhoenixR	**NetTide**
指数增长（Exponential growth）			✓	✓	✓	✓	✓	✓	✓
幂律增长（Power-law growth with arbitrary exponent）									✓
节点动力学方程（Differential equation for $n(t)$ ）				✓	✓	✓	✓	✓	✓
节点显示表达式（Closed form $n(t)$）				✓	✓				✓
边动力学方程（Differential equation for $e(t)$）			✓						✓
微观生成器（Microscopic Generators）	✓	✓		✓					✓

2.2.2　增长模型

增长模型[34] 在很多领域都有所讨论。关于增长现象的最经典的模型是流行病学中的传染病模型 Susceptible -Infected（SI）模型[16] 和关于创新扩散的 Bass 模型[1]。它们都产生 S 形曲线，即 Sigmoid 曲线，其在早期呈指数生长，用于刻画“感染”节点的动态增长过程。它们对微观感染过程以平均场形式给出了直观的描述。由恒定感染率引起的指数增长违背了人类行为的遗忘天性[35,36]，和社交网络中普遍存在的衰退模式[37]。像 PhoenixR[38] 这样的模型试图通过 Susceptible Infected Recovered（SIR）模型框架[16] 引入动力学衰退机制。SIR 模型的进一步变种在[5,39] 等工作中被采用。然而，基于 SIR 的恒定恢复率不能使指数增长减缓到幂律增长。总而言之，所有上述模型都无法产生所观测到的真实社交网络的幂律增长模式。

最近，一种刻画自激励（Self-excited）机制的点过程（Point Process）模型，即 Hawkes 过程（HP)[40]，被用来刻画爆发式增长和扩散现象。我们可以将 HP 视为内生分支过程（Branching Process，BP）与外生移民过程（Immigration Process)[41] 的结合，如 Crane-Sornette（CS）模型[42]，SpikeM[43] 等等工作。上述 HP 模型及其变体可以产生三种典型的增长模式：1）超临界状态下的指数增长，如 SpikeM 等模型；2）临近状态下的幂指数 < 1 的幂律速率增长和幂指数 < 2 幂律累积量增长模式；3）和在亚临界状态下迅速消亡的增长模式。因此，所有上述模型都无法产生具有任意幂指数的幂律增长。具体来讲，我们在微信社交网络中观察到的幂指数为 2.15。此外，上述模型都没有描述网络边的动态增长模式。

2.2.3　人类行为动力学

人类行为的动力学过程通常呈现爆发式行为和长尾（Fat-tailed）时间间隔时间（Inter-Event Time, IET）分布[35]。幂律（Power-law）IET 分布被发现普遍存在，并通过基于优先级的排队论[44] 和调制泊松过程[45] 等行为建模解释。最近，IET 的多尺度复杂分布被广泛发现[4,46]，包括在长时间尺度的二项分布[47]，在短时间尺度的密集行为[46] 等。总而言之，人类行为的遗忘或衰退现象普遍存在，是构建人类行为动力学过程的基石之一。

我们在表 2.1 中总结了上述所有模型的相对优缺点。只有我们的网潮 NetTide 模型具备所有优势。

2.3　网潮模型

2.3.1　初步分析

传统模型如 SI 模型和 Bass 模型，都无法产生如图2.1所示的幂律增模式。其中，SI 模型功能强大，直观，且被许多领域广泛运用。基于 SI 模型框架有许多的变种模型。例如 Bass 模型，就是增加了一个非零噪音项。但是，他们都无法定量地产生如真实数据图2.1a 和2.1 b 所示的增长模式。它们只能随着时间的推移产生 Sigmoid 的增长，这导致指数早期增长，而不是幂律增长。

一些合理的尝试。我们或许应该改变固定的传染性因子 β。比如，随着时间推移该传染因子可以衰减（可能是因为新奇感消失了）。一种被广泛采用的兴趣减弱方式是指数衰减，即 $\beta = \beta_0 * \exp(-\xi t)$。其中 ξ 是放射性衰变的半衰期：

- 尝试一，放射性衰减，如下式：

$$\frac{dn(t)}{dt} = \beta_0 \exp(-\xi t) n(t) \, (N - n(t)) \ 。 \tag{2-1}$$

但是以上组合还是没法产生如图2.1所示的幂律增长模式。

增长率应同时取决于 $n(t)$ 个受感染者以及 $N - n(t)$ 个易受感染者影响，但可能不是线性的关系。也许并非所有易受影响的人都可以被感染（例如，其中一些人有免疫抗体等等）和/或并非所有受感染的实际上都是活跃的（例如，其中一些呆在家里）。通过刻画上述不完全参与的情况，等式变为：

- 尝试二，部分参与。我们尝试了 $n(t)^{\zeta}$ 和 $(N - n(t))^{\psi}$。其中 $\zeta < 1$，$\psi < 1$，用来建模不完全参与的情形，如下式：

$$\frac{dn(t)}{dt} = \beta n(t)^{\zeta} \, (N - n(t))^{\psi} \ 。 \tag{2-2}$$

这个模型被广泛的用来刻画两种物质的化学反应，但是 $n(t)^{\zeta}$ 和 $(N - n(t))^{\psi}$ 个参与者在社交情形中难通过微观过程解释非整数情形。之后我们会发现，不用通过引入额外的参数，我们的模型就可以产生幂律增长，如果只额外引入一个参数，我们就可以产生一系列复杂的动力学增长过程。

2.3.2　网潮-节点模型

事实证明即使全部用户参与，如果通过刻画传染率 β 的衰退，我们也可以产生现实中的幂律动态增长。一个好的模型应该是直观容易解释的 - 为什么人类的兴趣会衰减？衰减，尤其是幂律衰减（Power-law Decay），或称作按比例放缩（Scaling），

表 2.2　符号和定义

符号	定义
N	总人口数，Number of the total population
$n(t)$	至 t 时刻累计用户数，Cumulative number of users by time t
$dn(t)/dt$	t 时刻新增用户数，Number of new users at time t
$e(t)$	至 t 时刻累计社交链接数，Cumulative number of links by time t
$de(t)/dt$	t 时刻新增社交链接数，Number of new links at time t
β	最大节点增长速率，Maximum growth rate of nodes
θ	时序衰减指数，Temporal fizzling exponent
β'	最大建边速率，Maximum linking rate
γ	网络幂律稀疏指数，The scaling sparsity exponent
α	网络线性稀疏系数，The linear sparsity coefficient

在社交交互（电子邮件回复时间等，正如我们在相关工作小节中提到的那样），和随机游走理论中（零交叉时间遵循指数为-1.5 的幂律）都被广泛发现。幂律衰减是合理和直观的模型来刻画人类行为的决策过程，和社交系统的群体衰减效应。

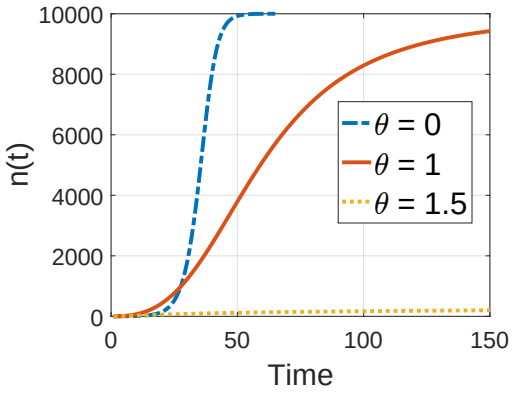
正是衰减效应导致了幂律增长，正如实际数据图2.1所示。此外，衰减效应也可以产生其他更普遍的增长，如图2.2所示。接下来，我们给出网潮-节点模型的详细描述和证明，来刻画节点的幂律增长动力学机制和过程。之后下一个小节，我们将致力于研究社交链接的动力学增长过程，并给出网潮-链接模型。我们将所使用的符号总结在表2.2中。

我们的网潮-节点模型，通过如下动力学方程描述：

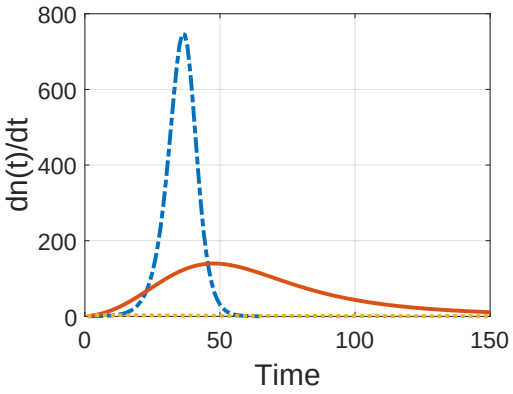
$$\frac{dn(t)}{dt} = \frac{\beta}{t^{\theta}} n(t)(N - n(t)) \text{ 。} \tag{2-3}$$

一个社交网络，如果其现有用户 $n(t)$ 越多，其在早期阶段越倾向于吸引更多的用户。但是，由于所有加入社交网络的潜在用户是有限的，其增长将受到不断减少的潜在用户 $(N - n(t))$ 的限制，特别是在趋近于人口饱和的阶段。这是一种普遍存在的自然现象，并且已经在许多学科中被观察到，从化学中的质量作用定律到模拟化学反应的速率，到在流行病中易感染者和感染者之间疾病的传播等等。创新的项 $\frac{\beta}{t^{\theta}}$ $(t > 0)$ 是自社交网络启动以来逐渐衰减的感染率，即人们逐渐开始厌倦感染他们的朋友加入社交网络。正是这个衰减的缩放指数 θ，产生了各种动态增长过程，包括幂律动力学增长作为特例（当方程2-3中 $\theta = 1$ 的时候）。我们把 θ 称作时间维度的衰减指数（temporal fizzling exponent）。

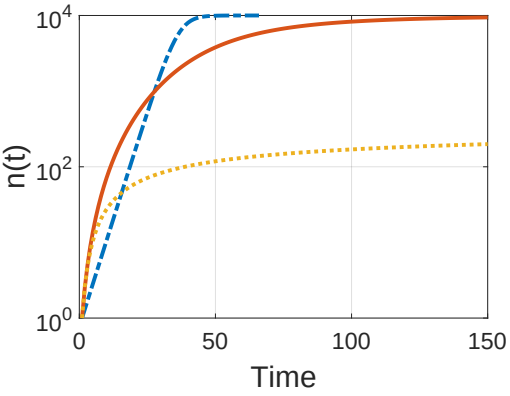
接下来，我们将给出动力学公式网潮-节点可以产生幂律增长的证明，及给出



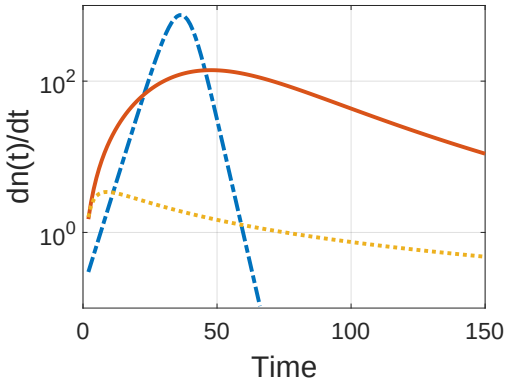
(a) 累积增长，线性-线性坐标



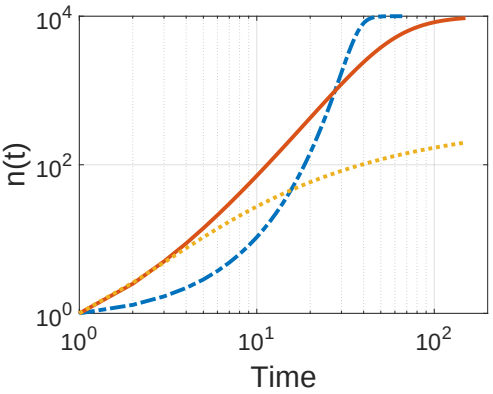
(b) 增长速率，线性-线性坐标



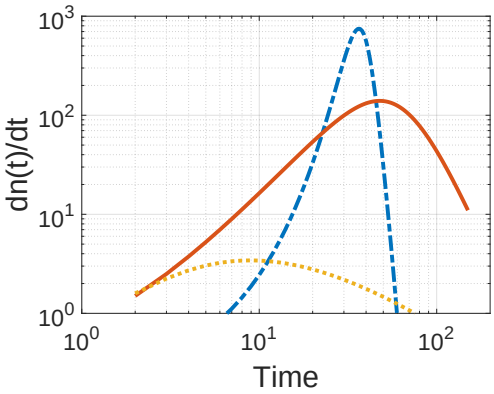
(c) 累积增长，对数-线性坐标



(d) 增长速率，对数-线性坐标



(e) 累积增长，对数-对数坐标



(f) 增长速率，对数-对数坐标

图 2.2　网潮-节点模型产生广泛的增长动力学曲线。三种典型的有着不同参数的累计增长曲线，在三种不同的坐标系统 a，c 和 e 下，其对应的增长速率曲线展示在 b, d 和 f 中：Sigmoid 增长 有着**指数初期增长** ($\theta = 0, N = 10^4, \beta = 3 \times 10^{-5}$)， Log-Logistic 增长 有着**幂律初期增长** ($\theta = 1, N = 10^4, \beta = 3 \times 10^{-4}$)，和 Stretched-Logistic 增长 有着**拉伸的指数初期增长** ($\theta = 1.5, N = 10^4, \beta = 4.5 \times 10^{-4}$)，分别用蓝色虚线，红色实线，还有黄色点线呈现。

动力学幂律增长的机制，并且进一步证明他将 Sigmoid 增长，即 SI 模型，Bass 模

型等当做特例，它还能产生更多全新的动力学增长过程!

引理 2.1：　当 $\theta = 1$ 时, 网潮-节点模型 NetTide-Node 将产生 **Log-Logistic** 动力学增长, 如公式 (2-5) 所示；当 $n(t) \ll N$ 时，产生近似的幂律 **Power-Law** 动力学增长，如公式 (2-8) 所示，其幂律指数是 βN.

证明 当 $\theta = 1$ 时, 网潮节点模型 NetTide-Node 是

$$\frac{dn(t)}{dt} = \frac{\beta}{t} n(t)(N - n(t)) \text{ 。} \tag{2-4}$$

由于这是个可分离的微分方程，我们可以将 $n(t)$ 项和时间项 t 分离然后分别积分，得到如下所示的增长曲线：

$$n(t) = N\frac{\lambda_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu} d\mu\}}{1 + \lambda_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu} d\mu\}} = N\frac{\lambda_0 (\frac{t}{t_0})^{\beta N}}{1 + \lambda_0 (\frac{t}{t_0})^{\beta N}} \tag{2-5}$$

且

$$\lambda_0 = \frac{n_0}{N - n_0} \text{ ,} \tag{2-6}$$

其中 n_0 是系统在初始时刻 t_0 时的总节点数。

如果 $n(t) \ll N$，即在系统增长初期，我们可以得到产生幂律增长的动力学机制：

$$\frac{dn(t)}{dt} \approx \frac{\beta N}{t} n(t) \text{ 。} \tag{2-7}$$

且该幂律增长指数是 βN，如下式所示：

$$n(t) = n_0 (\frac{t}{t_0})^{\beta N} \tag{2-8}$$

□

我们通过图2.2中的红色曲线来进一步说明什么是Log-Logistic 动力学增长。我们发现一个被拉伸的 S 型累计增长曲线，如图2.2a 所示，而 S 型 Sigmoid 曲线如蓝色虚线所示。幂律增长曲线（红色曲线）在对数-线性坐标下呈现次线性增长，如图2.2c 所示，并在对数-对数坐标系下呈现线性直线的增长，如图 Fig. 2.2e 所示，表

明了在拐点前后的幂律的动力学速率增长和衰退模式，而 Sigmoid 的动力学增长和衰退呈现更快的指数模式，如图2.2 b,d,f 所示。

引理 2.2：　当 $\theta \neq 1$ 时, 网潮-节点模型 NetTide-Node 按照如公式 2-9所示的动力学增长模式。当 $n(t) \ll N$ 时，节点在系统初期的增长遵循如公式2-10所描述的动力学增长模式。

我们将新发现的动力学增长公式2-9 称作 **Stretched-Logistic growth**，即拉伸的逻辑增长，并且在系统初期的动力学增长公式2-10称作 **Stretched-Exponential growth**，即拉伸指数增长。

证明 当 $\theta \neq 1$，公式 (2-9) 和其初期增长的公式 (2-10) 的推导过程都类似引理2.1的证明。我们得到：

$$\begin{aligned} n(t) &= N\frac{\lambda_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu^\theta} d\mu\}}{1+\lambda_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu^\theta} d\mu\}} \\ &= N\frac{\lambda_0 \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta}-t_0^{1-\theta})\}}{1+\lambda_0 \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta}-t_0^{1-\theta})\}} \ , \end{aligned} \tag{2-9}$$

其中 λ_0 在引理2.1中定义过了。特别的，当 $n(t) \ll N$，其初期增长曲线遵从：

$$\begin{aligned} n(t) &= n_0 \exp\{\int_{t_0}^{t} \frac{\beta N}{\mu^\theta} d\mu\} \\ &= n_0 \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta}-t_0^{1-\theta})\} \ 。 \end{aligned} \tag{2-10}$$

现在，我们说明它为什么称作拉伸的逻辑增长，即 Stretched-Logistic growth。现在值得我们回忆一下的是，如果随机变量 (random variable，r.v.）遵循对数逻辑分布 Log-Logistic distribution，那么它的对数变化后的值遵循逻辑 logistic 分布。类似的遵循 Log-Logistic 分布的命名规则，我们将说明如果一个随机变量 T 遵循如下 Stretched-Logistic 分布

$$P_T\{T \leq t\} = \frac{1}{Z_T}\frac{\lambda \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta}-t_0^{1-\theta})\}}{1+\lambda \exp\{\frac{\beta N}{1-\theta}(t^{1-\theta}-t_0^{1-\theta})\}} \ , \tag{2-11}$$

其中 Z_T 是归一化因子，λ 是一个常数，那么通过积分衰减因子 $X = \int_{t_0}^{T} t^{-\theta} dt =$

$\frac{T^{1-\theta}}{1-\theta} - \frac{t_0^{1-\theta}}{1-\theta}$ 遵循 Logistic 分布。当 $\theta < 1$，对于任意 $x \geq \frac{t_0^{1-\theta}}{\theta-1}$，

$$\begin{aligned} P_X\{X \leq x\} &= P_T\{\int_{t_0}^{T} t^{-\theta} dt \leq x\} \\ &= P_T\{\frac{T^{1-\theta}}{1-\theta} - \frac{T_0^{1-\theta}}{1-\theta} \leq x\} \\ &= P_T\{T \leq [(x + \frac{t_0^{1-\theta}}{1-\theta})(1-\theta)]^{\frac{1}{1-\theta}}\} \\ &= \frac{1}{Z_T} \frac{\lambda \exp\{\beta N x\}}{1 + \lambda \exp\{\beta N x\}} \ , \end{aligned} \tag{2-12}$$

上式说明 X 遵循 Logistic 分布。当 $\theta > 1$ 时，对于任意 $x < \frac{t_0^{1-\theta}}{\theta-1}$，相似的过程可以证明 X 遵循 Logistic 分布。 □

引理 2.3：　当 $\theta = 0$ 时，网潮-节点模型 NetTide-Node 遵循 **Sigmoid (或 Logistic)** 动力学增长模式，例如 SI 模型，其时引理 2.2 的特殊情况。当 $n(t) \ll N$ 时，逻辑增长曲线初期产生近似指数增长，**Exponential growth**，如下式所示：

$$n(t) = n_0 \exp\{\beta N (t - t_0)\} \ 。 \tag{2-13}$$

证明 将公式 (2-9) 和 (2-10) 中的 θ 替换为 0，即可以得到。此处说明，Sigmoid 函数一般用来描述 Logitic 的一种标准形式。此处我们将 Sigmoid 和 Logistic 混用来刻画有着指数初期增长的 S 型增长曲线。 □

网潮-节点模型 NetTide-Node 的进一步说明：

- 时间维度的衰减。此处我们并没有刻画每个人的衰减效应 $\frac{\beta}{(t-t_i)^\theta}$，其中 t_i 是用户 i 加入系统的时间，而是我们刻画了系统宏观的衰减效应 $\frac{\beta}{t^\theta}$，其中 t 是系统从初始时候到现在的时刻。因为像通过积分微观用户衰减效应的模型，例如 Hawkes 模型 $\frac{dn(t)}{dt} = n(t_0) + \sum_{t_i \leq t} \mu_i \frac{1}{(t-t_i)^\theta}$ 只能产生指数增长，或是幂律指数小于 2 的临界态幂律增长，如相关工作章节所述。该模型并不能产生如我们观测到的增长模式，即幂律指数大于 2 的幂律增长。相反，我们的网潮-节点模型很好的刻画了真实数据的动力学增长过程（实验章节进一步说明），而且它能产生很多动力学增长模式，包括有着任意幂律指数的幂律增长，拉伸指数增长，指数增长，还有包含上述增长的更长期的增长，正如如图2.2所展示的。

2.3.3　网潮-链接模型

社交网络的增长并不是仅限于节点的增长。之前的研究工作，都没有一个动力学方程来描述链接的动态增长过程。在这里，我们将介绍网潮-链接模型，来建模网络链接的动力学增长过程。我们假设存在一个潜在的组织结构作为社会网络形成和增长的潜在上下文结构。例如，共同作者社交网络的形成和发展受到导师 - 学生和研究者 - 合作者结构等组织结构的制约。因此，在对网络增长进行建模时，我们需要考虑底层组织结构的特征。我们将基础组织结构定义为图形 G_0，链接过程描述如下：对于每个现有节点 i，i 尝试链接到 G_0 中已存在的邻居 j。如果已经存在链接，则没有任何反应。如果尚未建立从 i 到 j 的链接，则 i 会尝试以感染率 β' 与 j 连边，且其感染率以 $\frac{1}{t^\theta}$ 随时间衰减。此外，新节点的到来将带来恒定数量的外部链接。网潮-链接模型 NetTide-Link 刻画了上述的链接过程如下：

$$\frac{de(t)}{dt} = \frac{\beta'}{t^\theta} n(t)(\alpha(n(t)-1)^\gamma - \frac{e(t)}{n(t)}) + 2\frac{dn(t)}{dt} \text{ 。} \tag{2-14}$$

网潮-链接模型的进一步说明：

- 外部链接：$2\frac{dn(t)}{dt}$ 刻画了新到达节点带来两个新链接的过程。因为我们将社交网络的边视为双向链接，并假设我们将每个新到达节点的第一条边视为外部链接。我们也可以进一步复杂化建模外部边的形成，比如我们可以将一个新到节点的头 m 条社交链接当做新形成的外部链接。但是无论如何，外部链接和下述内部链接比起来，对边的动力学增长过程的贡献都是相对较小的。
- 内部链接：内部链接是由已经存在的节点之间建立社交关系而形成的，并产生网络中边的稠密化现象。对于每个已经存在的节点，他/她尝试连接潜在组织结构 G_0 中尚未链接的现有邻居。由于组织结构的限制，他/她只能访问到局部平均而言 $\alpha(n(t)-1)^\gamma$ 个现有节点，即平均而言潜在可访问的现有邻居数量。而 $\frac{e(t)}{n(t)}$ 项是刻画了要排除的已连接的平均邻居数。同节点方程类似，$\frac{\beta'}{t^\theta}$ 项刻画了人链接行为随时间衰退的现象。
- 稠密化：通过实验部分的实证分析，链接方程通过幂律稀疏指数 γ 来建模边幂律稠密的规律。链接和节点之间的幂律稠密指数是 $1+\gamma$。例如，如果我们通过 Kronecker 图模型[48] 对 G_0 建模，我们得到 $\gamma = \log E / \log N$。

2.3.4　微观随机过程生成模型

在这里，我们构建了两个微观随机过程生成模型，分别是网潮-生存模型 NetTide-Survival 和网潮-过程模型 NetTide-Process，希望从微观随机模拟角度产生各

种观测到的宏观动力学增长模式。我们将说明他们确实可以生成逼真的社交网络随机动态增长曲线。

网潮-生存模型 NetTide-Survival。我们首先通过生存分析（Survival Analysis）框架中的危险率（Hazard Rate）来刻画节点和边的随机动态增长过程。

- *节点增长的危险率*。一个在 t 时刻还不是网络节点的用户，他在此时此刻想要加入网络（例如注册微信）的瞬时概率密度定义为：

$$\lambda_n(t) = \frac{\beta}{t^\theta} n(t) \text{ 。} \tag{2-15}$$

我们定义 $F(t)$ 为 t 时刻所有潜在用户中现在是注册用户的比例，而 $f(t)$ 表示 $F(t)$ 的导数。然后，$\lambda_n(t) = \frac{f(t)}{1-F(t)} = \frac{\beta}{t^\theta} n(t)$。请注意 $F(t) = \frac{n(t)}{N}$ 和 $f(t) = \frac{dn(t)}{N \times dt}$。我们通过以下公式连接了网潮-节点模型 (2-3) 和危险率公式 (2-15)：

$$\begin{aligned} \frac{dn(t)}{dt} &= f(t) \times N \\ &= \lambda_n(t) \times (1 - F(t)) \times N \\ &= \frac{\beta}{t^\theta} n(t)(N - n(t)) \text{ 。} \end{aligned} \tag{2-16}$$

确实，生存分析框架中的危险率函数和（Log / Fizzle-）Logistic 框架通过等式 (2-16) 连接，且基于潜在人口 N 是常数的假设。

- *链接增长的危险率*：建立外部链路的瞬时速率是新节点瞬时速率的两倍。对于内部链接而言，在 t 时刻还没有构建链接的潜在候选边在此时刻构建的瞬时速率是：

$$\lambda_e(t) = \frac{\beta'}{t^\theta} \text{ 。} \tag{2-17}$$

按照类似的分析过程，定义 $F_e(t)$ 是在 t 时刻构建的内部链接的比例，$f_e(t)$ 是 $F_e(t)$ 的导数。因此：

$$\begin{aligned} \lambda_e(t) &= \frac{f_e(t)}{1 - F_e(t)} \\ &= \frac{de(t)}{(\alpha n(t)(n(t) - 1)^\gamma - e(t))dt} \\ &= \frac{\beta'}{t^\theta}, \end{aligned} \tag{2-18}$$

其中 $\alpha n(t)(n(t) - 1)^\gamma - e(t)$ 是 t 时刻所有可能构建的链接数量减去已经构建的

链接。因此，我们通过公式 (2-18) 连接了网潮-链接模型 NetTide-Link (2-14) 和生存分析中的危险率公式 (2-17) 。

在此，我们给出完整的的网潮-生存模型 NetTide-Survival。给定输入参数 $(\beta, \theta, N, \beta', \alpha, \gamma)$，NetTide-Survival 模型通过刻画危险率方程产生了随机增长动力学过程 $n(t)$ 和 $e(t)$：

- 节点增长：在时刻 t 时，对于 $N - n(t)$ 个潜在的生存节点中的每一个点，在时间间隔 $(t, t+h)$（$h \to 0$）以 $\lambda_n(t)h + o(h)$ 概率选点，如果成功选中，我们将 $n(t)$ 增加 1，将 $e(t)$ 增加 2。
- 链接增长：对于 t 时刻所有 $\alpha n(t)(n(t)-1)^\gamma - e(t)$ 个潜在的边, 每一条边在时间间隔 $(t, t+h)$（$h \to 0$）以 $\lambda_e(t)h + o(h)$ 概率选边，如果选边成功，我们将将 $e(t)$ 增加 2。

网潮-生存模型 NetTide-Survival 的进一步扩展。我们可以扩展现有模型来刻画更复杂的现象。例如，节点的危险函数可以包含泊松项 λ_0，成为 $\lambda_n(t) = \lambda_0 + \frac{\beta}{t^\theta} n(t)$，来刻画用户加入网络，比如注册微信，的内在的线性随机速率。此外，我们可以引入时间滞后项来刻画初期的密集加入过程，例如：$\lambda_n(t) = \frac{\beta}{(t+\Delta)^\theta} n(t) = \frac{\beta/\Delta^\theta}{(1+t/\Delta)^\theta} n(t)$。通过引入具有特定参数的正弦函数，我们可以将周期性结合到网潮-生存模型中。为简洁起见，我们并没有进一步详细介绍这些扩展。

网潮-过程模型 NetTide-Process 。我们进一步从网络中的微观随机交互的角度来解释节点和链接的动态增长过程。在平均场假设下，每个人有相当的易感邻居，我们可以将总体危险率分解为微观成对的感染率：

$$\begin{aligned} \frac{dn(t)}{dt} &= \lambda_n(t) \times (N - n(t)) \\ &= \frac{\lambda_n(t)}{\langle k \rangle} \langle k \rangle N(1 - F(t)) \\ &= \frac{\beta N}{\langle k \rangle t^\theta} n(t) \langle k \rangle (1 - F(t)) \ , \end{aligned} \tag{2-19}$$

其中 $\langle k \rangle (1 - F(t))$ 是已经被感染的用户其周围可被感染的平均邻居数量，其中 $\langle k \rangle$ 是平均度数。因此，对于每个感染用户而言，他/她尝试在时间间隔 $(t, t+h)$ $(h \to 0)$ 内以 $p_n(t) = \int_t^{t+h} \frac{\beta N}{\langle k \rangle t^\theta} dt = \frac{\beta N}{\langle k \rangle t^\theta} * h + o(h)$ 概率将其每个邻居感染。对于链接的过程，在均匀混合的假设下，即 $\gamma = 1$ 并且 α 是随机图的线性稀疏下，我们将总体危

险率分解为微观成对的感染率：

$$\begin{aligned}\frac{de(t)}{dt} &= \lambda_e(t) \times (\alpha n(t)(n(t)-1) - e(t)) \\ &= \frac{\beta'}{t^\theta} n(t)(\alpha(n(t)-1) - \frac{e(t)}{n(t)}) \text{ 。}\end{aligned} \tag{2-20}$$

因此，对于每个感染用户而言，他/她尝试和其已经被感染的但是还没有建边的邻居在时间间隔 $(t, t+h)$（$h \to 0$）内以 $p_e(t) = \int_t^{t+h} \frac{\beta'}{t^\theta} dt = \frac{\beta'}{t^\theta} * h + o(h)$ 概率建边。

虽然在随机图中分析了宏观层面危险率与基于网络的微交互之间的关系，但是模型 - 过程可以应用于任意网络中。首先，我们需要基础组织结构 G_0，节点 β 的最大增长率，时间维度的衰减指数 θ 和最大链接率 β'。考虑 $G_1(t) = (Node(t), Edge(t))$ 是 G_0 上的不断发展的网络。$Node(t)$ 和 $Edge(t)$ 是 t 时刻 G_1 系统中现有的节点集合和链接集合。我们可以通过随机初始化 $Node(t_0)$，或者只是将初始状态作为输入来描述系统的初始阶段。对于 $Edge(t_0)$ 的情况也是如此。因此，网潮-过程模型 NetTide-Process 可以表述如下：

- *节点增长*：在 t 时刻，对于在 $Node(t)$ 中已经存在的任意节点 i，i 尝试感染它的邻居，比如 G_0 中的节点 j。如果 j 还没有被感染，即 j 在 G_1 中，那么在时间间隔 $(t, t+h)$（$h \to 0$）内，i 以 $p_n(t)$ 的概率感染 j。如果成功感染，我们将 (j, t) 加入到节点集合 $Node(t)$ 中，将 $(i, j, t), (j, i, t)$ 加入到边集合 $Edge(t)$ 中。
- *链接增长*：如果 j 已经在 G_1 但是还没有和 G_1 中的 i 相连接，那么 i 在时间间隔 $(t, t+h)$（$h \to 0$）内尝试和 j 以概率 $p_e(t)$ 连边。如果成功连接边，我们将 $(i, j, t), (j, i, t)$ 加入到边集合 $Edge(t)$ 中。
- *用户行为*：如果 j 已经在 G_1 中并且和 G_1 中 i 也连了边，那么之后 i 可以和 j 进行交互，比如发信息，但是对我们关心的网络 $G_1(t)$ 没有影响。随着时间的推移，网络 G_1 随时间增长。

这两个随机生成器旨在描述节点和边的随机动态增长。为了结果的可复现性，我们开源了代码，参见章节2.5。

2.3.5　模型参数学习

网潮模型 NetTide，包括节点和链接两个子部分，有着简明的参数集，即 $\Theta = \{\beta, \theta, \beta', \alpha, \gamma, N\}$。我们的参数学习过程分为两步，第一步先学习节点方程，之后

再学习边方程。给定真实的节点增长序列 $n(t)$，我们试图最小化如下平方和误差：

$$\min_{\beta,\theta,N} J(n(t), n^*(t)) = \sum_{t=t_0}^{T} (n(t) - n^*(t))^2 \text{ 。} \tag{2-21}$$

至于链接方程，给定真实链接和节点增长序列 $e(t)$ 和 $n(t)$，以及从节点步骤学习的时间衰减指数 θ，我们遵循与节点学习相同的过程步骤，以尽量减少平方和误差：

$$\min_{\beta',\alpha,\gamma} J(e(t), e^*(t)) = \sum_{t=t_0}^{T} (e(t) - e^*(t))^2 \text{ 。} \tag{2-22}$$

我们采用 *Levenberg-Marquardt* 算法 (LM)[49] 来求解如上非线性优化问题。

2.4 实验结果

在本节中，我们将通过一系列真实演化的社交网络来评估网潮模型 NetTide 的有效性。在这里，我们实验试图回答以下问题：

Q1. **准确性**。 网潮模型是否可以精确刻画真实社交网络节点和边的动力学增长过程？

Q1. **实用性**。网潮模型可以精确预测未来的节点数 $n(t)$ 和链接数 $e(t)$ 么？可以预测多远的未来？

Q1. **生成器**。两个微观随机过程生成器，即网潮-生成模型 NetTide-Survival 和网潮-过程模型 NetTide-Process 可以产生逼真的随机增长过程嘛？

2.4.1 数据集

表 2.3　社交网络数据集统计表

网络	时间	N	E	T	时间粒度
WeChat	01/2011-01/2013	300M	4.75B	726	1 天
arXiv	03/1992-03/2002	16,959	2,388,880	95	1 月
Enron	01/1998-07/2002	86,458	594,998	55	1 月
Weibo	21-26/06/2012	165,147	331,607	1409	5 分钟

微信在线社交网络。 微信是中国最大的在线社交网络，截至 2018 年 6 月 30 日，每月活跃用户超过 10.58 亿。我们收集了微信的历史数据，包括从 2011 年 1 月

21 日（WeChat 发布之日起）到 2013 年 1 月 16 日节点和链接增长的完整记录，总共有三亿个节点（注册用户而不是每月活跃用户）和超过 47.5 亿条社交链接。数据记录了每个用户的添加时间和每个社交链接的建立时间，精确至秒。我们将用户之间的双向关系视为两条链接。据我们所知，这是社交网络演化的最大的数据集之一。此外，我们通过 2015 年 12 月 17 日至 2016 年 1 月 14 日的微信社交网络的五个最新快照验证了我们模型的预测能力。我们可以访问的所有微信数据都是匿名的，且遵循了严格的隐私政策。

ArXiv **学者合作网络**。这是一个学者合作网络，自成立以来已有近十年[17]。如果任何两个人在一篇论文的作者名单中，那么他们形成了双向链接，其时间戳是其发布日期。人员的加入日期由他在此数据集中首次发表论文的日期表示。该数据集涵盖从 1992 年 3 月（在 arXiv 开始附近）到 2002 年 3 月。通过过滤没有明确日期的链接，总共有 16,959 个节点和 2,388,880 条边。

Enron **公司社交网络**。 通过安然 Enron 公司[18] 的电子邮件记录，我们恢复了安然员工构建的企业社交网络。该数据集涵盖了从 1998 年 1 月至 2002 年 7 月的网络演化，其间安然于 2001 年 12 月 2 日破产，导致 $n(t)$ 的急剧下降。总共有 86,458 个节点和 594,998 条边。

微博信息传播网络。我们在腾讯微博中选择一个大型信息级联社交网络[8]，这是由一个关于流行游戏的信息传播形成的微博子网络，有 165,147 节点和 331,607 条社交链接，揭示了对这个游戏感兴趣用户的社交网络的构建过程。

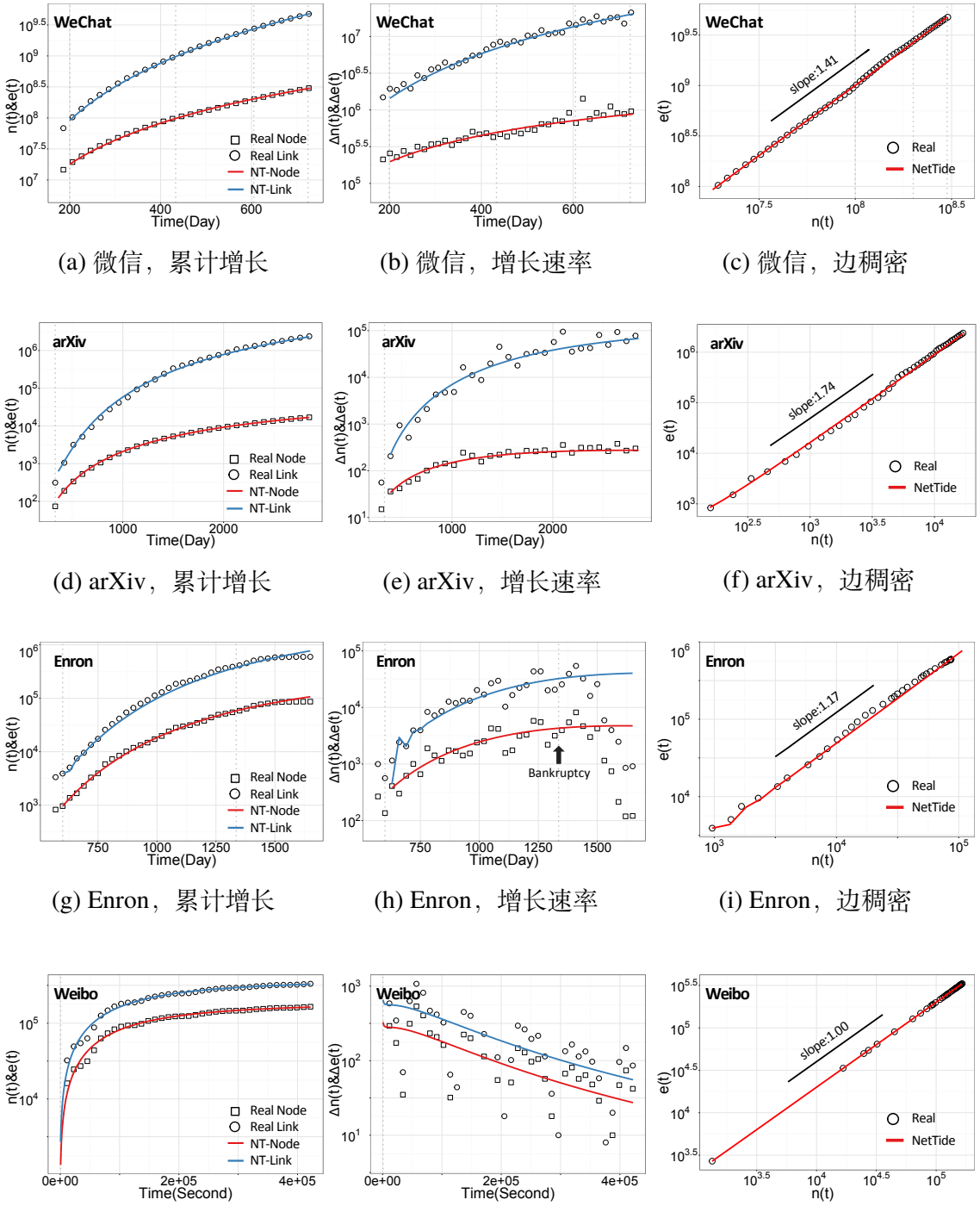
2.4.2　正确性

我们通过回答 Q1 来验证网潮模型的正确性，以确定我们的模型是否能够捕获真实社交网络中节点和链接的动态增长模式。

2.4.2.1　评价方法

我们在四个不同的真实社交网络中进行实验，并设置了五个评价指标，来验证我们模型的正确性和可推广性。这五个评价指标分别是节点动态累积数量 $n(t)$，节点动态增长速率 $\frac{dn(t)}{dt}$，链接动态累积数量 $e(t)$，链接动态增长速率 $\frac{de(t)}{dt}$，以及链接相对于节点的稠密化关系 $e(n(t))$。我们还将在2.2部分中讨论与其他四种基线方法的比较，他们分别是：传染病模型（SI），Bass 模型，SpikeM 模型和 Phoenix-R（PHR）模型。所有这些方法都是为节点设计的，据我们所知，**并没有适用于链接的动力学模型**。

我们通过归一化均方根误差（*NRMSE*）来评估模型的整体拟合准确性。具体



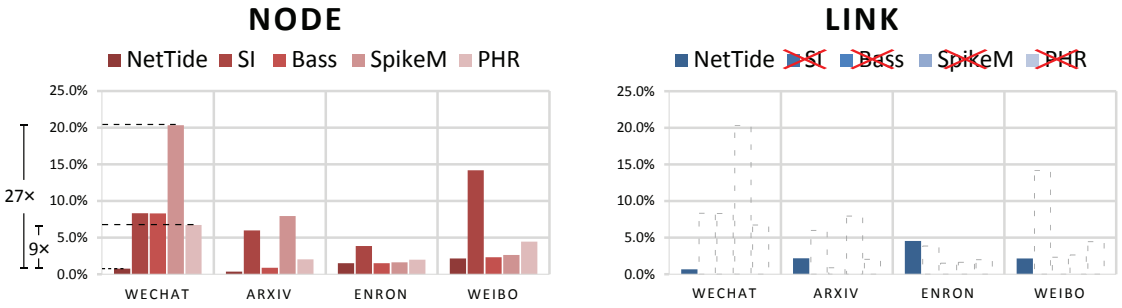
(a) 微信，累计增长　(b) 微信，增长速率　(c) 微信，边稠密

(d) arXiv，累计增长　(e) arXiv，增长速率　(f) arXiv，边稠密

(g) Enron，累计增长　(h) Enron，增长速率　(i) Enron，边稠密

(j) 微博，累计增长　(k) 微博，增长速率　(l) 微博，边稠密

图 2.3　**网潮模型准确地拟合了真实数据**。我们的模型准确地拟合了四个真实社交网络的动态增长过程。四行分别对应于微信（a-c），arXiv（d-f），安然 Enron（g-i）和腾讯微博（j-l）。在每一行中，有五个评价指标：$n(t)$ 和 $e(t)$ 在第一列图中，$\frac{dn(t)}{dt}$ 和 $\frac{de(t)}{dt}$ 在第二列图中，$e(n(t))$ 在第三列图中。



(a) 节点 $n(t)$ 的误差　　(b) 链接 $e(t)$ 的误差

图 2.4　**网潮模型 NetTide 打败了所有基线模型**。网潮节点模型 NetTide-Node 在 *NRMSE* 指标上都打败了其他模型。网潮链接模型 NetTide-Link 以较低的误差拟合了所有数据。更重要的是，现有的模型并没有刻画边增长的动力学模型。

来讲，给定两个增长序列，例如真实节点增长序列 $n(t)$ 和我们的模型给出的相应序列 $n^*(t)$，$NRMSE = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(n(t)-n^*(t))^2}}{max(n(t))-min(n(t))}$。作为 $T=1$ 时的特殊情况，*NRMSE* 退化为绝对百分比误差 $APE(x,x^*) = \frac{|x-x^*|}{x}$。*NRMSE* 与 *LM* 算法 *L2* 范数意义下的目标函数一致，并且其还可以在不同数据集之间进行比较。我们还通过其他标准度量指标来评价我们模型的性能，例如平均绝对百分比误差（*MAPE*）。我们得出了一致的结论，因此出于简明性考虑我们不做额外的报告。表2.5显示了数据集的最佳拟合参数。

2.4.2.2　数量准确性和曲线性状准确性

我们的网潮模型准确地刻画了微信节点和边的动态增长过程，包括微信自发布以来 726 天的时间。从如图 2.3 a-c 所示的五个检查点的拟合结果来看，我们模型生成的增长曲线几乎与所有实际数据点完全重叠。我们拟合的部分覆盖了微信获得其大部分用户的关键时期。对于累积增长的数值，网潮模型和真实数据之间的整体错误都小于 1%。具体来说，对于 $n(t)$ 误差为 0.76%，对于 $e(t)$ 误差为 0.66%，如表 2.4所示。尽管与累积增长数相比，速率表现出更大的波动，但我们的模型模型仍能很好地拟合速率（如图2.3 b 所示），且与基线模型相比具有最低的误差（如表 2.4所示）。$n(t)$ 和 $e(t)$ 之间的稠密化关系被网潮模型完美地描述，总体错误为 1.08%。此外，只有我们的网潮-链接模型可以刻画边的动力学增长（如图 2.4 b 所示）。

随后，我们验证了网潮模型对 arXiv 和 Enron 数据的刻画。尽管两个数据时间跨度较长（分别为 5 年和 10 年），增长逐渐趋于饱和，以及其他意外因素（如安然的破产），但是我们的模型再次准确地拟合了他们的动态增长过程。拟合范围涵盖了两个社交网络获得其 99% 人口的时期。我们对这两个数据的动态增长的时间单

位设置成月，是对于科研合作或企业上下文适当的粒度。我们网潮模型产生的红色和蓝色曲线几乎与 arXiv（如图 2.3 d-f 所示）和 Enron（如图 2.3 g-i 所示）的所有实际数据点重叠。具体来说，与基线模型相比，网潮节点模型在 arXiv 和安然案例中都得到的最低的误差，分别为 0.35% 和 1.51%（如图 2.4 a 所示）。此外，网潮链接模型分别为 arXiv 和 Enron 准确地捕获链接增长，2.18% 和 4.54%。所有基线模型都无法描述链接的动态增长过程（如图 2.4 b 所示）。

最后，我们验证了网潮模型对腾讯微博数据的刻画，这是一个易变的网络并且表现出很大的波动。尽管如此，我们的模型再次很好地捕捉了微博的动态增长过程。由于其不稳定的性质，我们将增长的最小时间单位设置为 5 分钟。虽然每日波动（对应于午夜和办公时间的退潮和峰值分别如图 2.3 k 所示）引入了相对较大的误差（如表 2.4所示），我们仍然可以看到图 2.3 j 和 l 中模型所给出的趋势，而且 $n(t)$ 和 $e(t)$ 的拟合结果仍然很好。具体来说，网潮-节点模型和网潮-链接模型分别对 $n(t)$ 和 $e(t)$ 获得了 2.15% 和 2.15% 的误差。我们的模型相对于其他模型而言，对 $n(t)$ 依旧是最低的误差，没有模型刻画 $e(t)$，如图 2.4 b 所示。

只有我们的网潮模型在刻画动力学增长的绝对数值和增长形状方面获得了准确的结果。到目前为止，我们的模型已经通过刻画真实动态增长的曲线形状和最低的整体拟合误差证明其能力。更重要的是，我们的网潮链接模型在捕捉链接增长动态方面是独一无二的。进一步，我们的反问是：我们的网潮节点模型在刻画节点的动力学增长方面是不是也是独一无二的?

所有最先进的基线模型都无法准确刻画动态增长过程的绝对数值或增长形状这两个测量方面，具体来说：早期 SI 和 Bass 的指数增长性质严重偏离了真实数据，完全无法刻画微信的幂律增长。在微信数据中，SI 和 Bass 具有非常相似的性能，误差比我们的网潮-节点的结果高出 10.0 倍。在其他数据集中，SI 也严重偏离现实，如图 2.4 a 所示，对 arXiv，安然和微博的结果相比，误差相对于我们的模型分别高出 16.1，1.5，和 5.6 倍。虽然 Bass 模型与 SI 模型相比，引入了对市场自然增长的刻画，相对而言减小了误差，但 Bass 模型早期阶段的指数形状与我们的幂律观测是完全相违背的。SpikeM 在不同数据集中的表现差异很大。SpikeM 在微信和微博案例中的最佳拟合结果是 Hawkes 过程的次-临界态中。但是，SpikeM 在微信数据中报告了最大的误差（比模型节点高 25.8 倍），而在微博中则达到了相对较低的错误（比模型节点高 21.9%）。而 SpikeM 模型在拟合 arXiv 和安然时达到 Hawkes 过程超临界状态，在早期阶段产生指数增长，误差相对于我们的网潮-节点模型而言，分别高出 21.7 倍，和 8.0%。Phoenix-R 也出现了错误形状和大幅波动的问题：它报告了微信基线中的最低误差，仍比我们的网潮节点模型大 7.9 倍。对

表 2.4　模型在五个评价方面的准确率。我们的网潮模型对于四个真实社交网络，5 个评价方面，都好于最先进的基线模型。所有基线模型都不适用于边的增长（---）。

WeChat	**n(t)**	**e(t)**	**dn(t)/dt**	**de(t)/dt**	**e(n)**
NetTide	**0.76%**	**0.66%**	**6.29%**	**5.07%**	**1.08%**
SI	8.32%	—	23.39%	—	—
BASS	8.31%	—	23.64%	—	—
SPIKEM	20.33%	—	48.19%	—	—
PHR	6.73%	—	8.59%	—	—
arXiv	**n(t)**	**e(t)**	**dn(t)/dt**	**de(t)/dt**	**e(n)**
NetTide	**0.35%**	**2.18%**	**9.91%**	**11.27%**	**3.32%**
SI	5.97%	—	33.83%	—	—
BASS	0.88%	—	11.18%	—	—
SPIKEM	7.95%	—	24.63%	—	—
PHR	2.03%	—	15.07%	—	—
Enron	**n(t)**	**e(t)**	**dn(t)/dt**	**de(t)/dt**	**e(n)**
NetTide	**1.51%**	**4.54%**	14.62%	**14.27%**	**4.62%**
SI	3.84%	—	20.74%	—	—
BASS	1.51%	—	**14.54%**	—	—
SPIKEM	1.63%	—	18.00%	—	—
PHR	1.99%	—	15.89%	—	—
Weibo	**n(t)**	**e(t)**	**dn(t)/dt**	**de(t)/dt**	**e(n)**
NetTide	**2.15%**	**2.15%**	14.93%	**14.90%**	**0.06%**
SI	14.19%	—	24.51%	—	—
BASS	2.31%	—	15.01%	—	—
SPIKEM	2.62%	—	**14.78%**	—	—
PHR	4.45%	—	17.53%	—	—

于 arXiv 和微博来说，Phoenix-R 和我们节点模型相比，对 arXiv 数据误差高出 4.8 倍，而对于 Enron 数据来说，误差高出 31.8% 。

总之，在刻画动态增长过程的绝对数值和增长形状方面，只有我们的模型准确地刻画了真实社交网络的节点和链接的动态增长过程。

2.4.2.3　参数分析

网潮模型的最佳拟合参数符合每个社交网络的实际增长动态特征（如表 2.5所示）。在微信的情况下，时间维度的衰减指数是 θ（我们模型给出 0.995 ）的价值非常接近 1，暗示节点的幂律增长，其幂律指数 $\approx \beta N = 2.16$ (接近真实 $n(t)$ 数据的

表 2.5　网潮模型在各个数据集最好拟合下的参数。

	N	βN	θ	β'	α	γ
WeChat	6.1B	2.16	0.995	0.03	0.14	0.47
arXiv	12584	8.81	1.35	7.56	0.28	0.74
Enron	458143	155.14	1.96	751.19	1.30	0.16
Weibo	18935	0.50	0.84	0.030	1.68	0.02

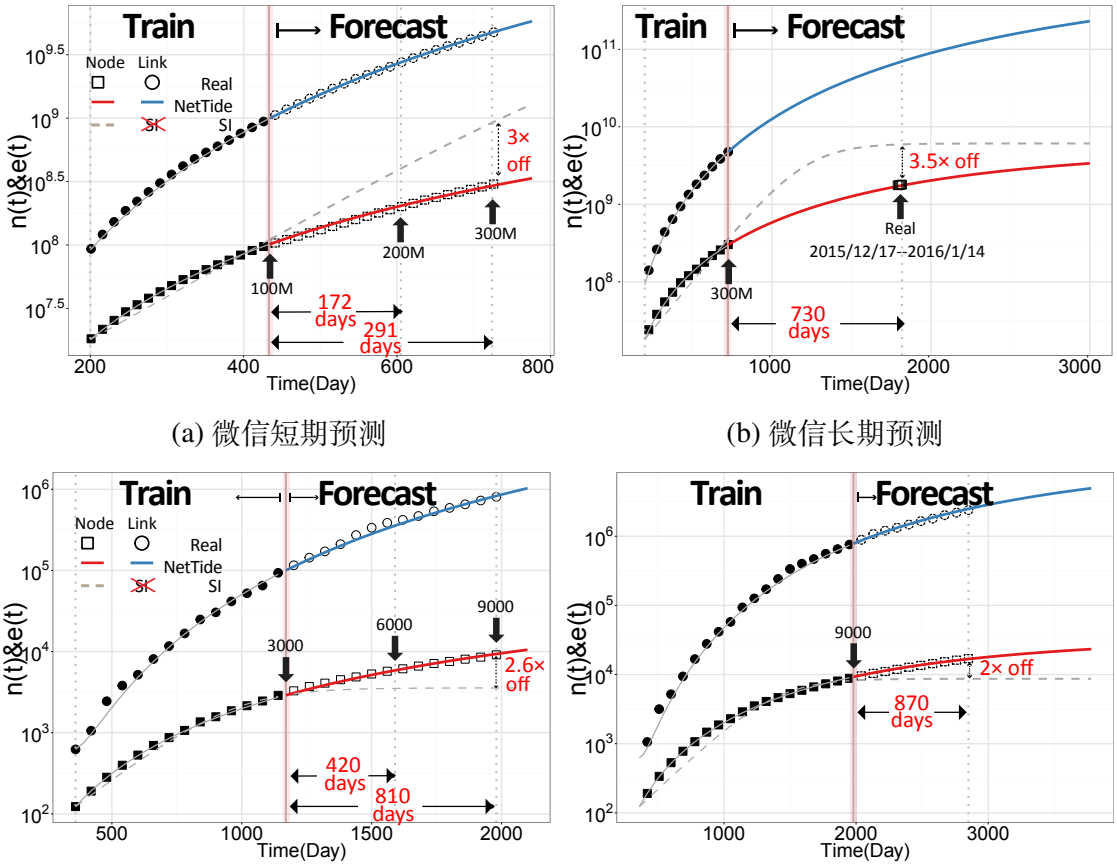
2.15，如图 2.1 a 所示)。边稠密幂律指数为 $1+\gamma = 1.47$（接近真实 $e(n(t))$ 的 1.41 显示，如图2.3 c）所示。对于 arXiv，学习得到的 $\theta = 1.35$ 意味着相比 $\theta = 1$ 情况下的 Log-Logistic 曲线，有着更快的时间衰减效果。arXiv 网络的发展在 4 个数据集中具有最大的边稠密幂律，其指数为 1.74，由 $\gamma = 0.74$ 刻画，这意味着高能物理社区的紧密结构。安然公司的时间衰减指数（1.96）是数据集中最大的一个，刻画了安然公司由于破产引起的人数停滞。安然公司的结构显示了像树一样的层次结构，因为与其他社交网络数据相比，致密化幂律的指数相对较小 1.17（由学习到的 $\gamma = 0.16$ 刻画）。微博的树形结构显示了链接相对于节点的线性增长，这使得 γ（$\gamma = 0.02$ ）非常接近于零。最合适的价值 $\theta = 0.84$，由于微博的快速增长特性，在早期阶段表现出快于幂律的增长。

2.4.3　预测

我们通过回答 Q2 来展示我们网潮模型的实用价值，即在短期和长期内对网络节点和链接的数量进行预测。

2.4.3.1　短期预测

在短期预测的实验设置中，我们通过验证对未来总体趋势的预测误差（整体预测任务）以及标记为里程碑的一些关键时刻点的到达时间预测（里程碑预测任务）来验证网潮模型的预测能力。以微信为例，通过在最初的 1 亿用户范围内训练节点的动态增长，整体预测任务是检查节点模型预测下之后的 2 亿用户的动态增长过程; 而里程碑预测任务是预测微信网络什么时候翻一倍，翻两倍，和翻三倍其用户数的里程碑日期。我们将 t_1，t_2，t_3 表示为里程碑的日期。在微信的情况下，它们是微信网络分别达到第一个 1 亿，2 亿，3 亿用户的日期，如图 2.5 a 所示。在 arXiv 的情况下，它们分别是图中的到达 3000，6000，9000 作者的日期，如图 2.5 c 所示。我们对链接模型 NetTide-Link 也进行了类似的实验设置，而之前的工作都从未对社交链接的数量进行过预测。



(a) 微信短期预测　(b) 微信长期预测

(c) arXiv 短期预测　(d) arXiv 长期预测

图 2.5　**网潮模型 NetTide 可以很好地预测未来**。图中点代表真实数据，黑色实心点表示训练数据部分，空心点表示预测数据部分。红线和蓝线分别是网潮-节点模型 NetTide-Node 和网潮-链接模型 NetTide-Link 的预测结果。灰色虚线是 SI 的结果。上面两张图是微信的结果，而下面的两张图是 arXiv 的结果。（a）和（c）是短期预测的结果，而（b）和（d）是长期预测的结果。

整体预测任务。网潮模型，包括网潮-节点模型 NetTide-Node 和网潮-链接模型 NetTide-Link，都可以非常准确地预测未来动态增长过程，覆盖微信未来 291 天和 arXiv 未来 730 天。在微信数据中，网潮模型对节点 $n(t)$ 和边 $e(t)$ 在 t_1 到 t_3 的时间段内预测的整体误差分别为 2.18% 和 0.44%（如图 2.5a 所示）。对于 arXiv 数据集，对节点 $n(t)$ 和链接 $e(t)$ 从 t_1 到 t_3 预测的整体误差分别为 2.86% 和 4.18%（如图 2.5 c 所示）。我们还比较了 SI 的预测结果：在微信情况下，S 形曲线严重高估了 nt，整体误差高达 134.62%，但是低估了 arXiv 的 $n(t)$，整体误差为 52.14%。SI 不适用于对 $e(t)$ 的预测（2.5图中没有用于链接的预测结果）。

里程碑预测任务。网潮-节点模型 NetTide-Node 和网潮-链接模型 NetTide-Link 都可以以很低的误差来预测未来里程碑日期的到来。具体来说，在图2.5 a 中所示的微信数据中，网潮节点模型预测 2 亿微信用户的时间早于实际日期 t_2 5 天（对

未来 172 天的预测），而对 3 亿用户的时间 t_3 预测仅晚 10 天（对未来 291 天的预测）。在 t_2 和 t_3，时刻，对 $n(t)$ 的预测误差分别为 1.67% 和 2.58%，对 $e(t)$ 的误差分别为 0.26% 和 0.33%。而对于 arXiv 网络，尽管 t_2（t_3）是在未来的 420（810）天，网潮-节点模型可以预测里程碑的到来 6000（9000）作者在正负一个月的误差内（对于 arXiv 和安然，我们选择的时间粒度就是一个月。）在 t_2 和 t_3 对 $n(t)$ 预测误差分别为 0.91% 和 2.47% ，而对 $e(t)$ 的误差为 11.32% 和 2.75%。相比之下，由 SI 模型预测的节点结果存在严重偏差：在微信数据下，对 t_2 的预测要早 93 天，对 t_3 的预测要早 167 天。随着时间的推移，偏差越来越大，对 t_3 的预测偏差超过 300%。至于 arXiv，SI 模型严重低估了里程碑节点的数量：对 t_3 的低估程度多于超过 260%。再次强调，没有对网络链接建模的动力学基准模型。

总之，我们的模型在短期内对节点和链接的增长预测都达到了很好的预测精度。

2.4.3.2 长期预测，未来两年

我们的网潮模型在长期预测上取得了很好的结果。具体而言，对微信未来 730 天和 arXiv 未来 870 的预测都取得了很好的结果。

对于微信数据，网潮-节点模型可以准确地预测未来 730 天节点的数量（图 2.5 b）。我们通过 t_3 之前的动态增长数据训练模型，然后我们通过微信社交网络的 5 个最新快照验证模型的预测结果。未来 5 个最新检查点跨越一个多月（2015 年 12 月 17 日，25 日和 2016 年 1 月 1 日，8 日，14 日）。由于隐私问题，我们不会报告注册用户的确切数量和链接数量。我们将初始总人口 N 设定为 61 亿，是由爱立信[①]报道的到 2020 年全球智能手机用户。因为一个用户只能通过验证他的电话号码成功注册微信。这五个检查点的 $n(t)$ 的误差为：2.86%，2.72%，2.68%，2.68% 和 2.64%。然而，SI 模型的节点增长曲线严重高估了实际节点的增长：饱和点更早到达，与 2016/1/14 的实际数据相差 350%。

对于 arXiv 数据，网潮模型可以长期准确地预测 $n(t)$ 和 $e(t)$ 未来 870 天，如图2.5 d 所示。我们在 t_3 之前通过实际动态增长数据训练节点和链接模型，我们对 $n(t)$ 的预测误差为 2.84% ，$e(t)$ 的预测误差 3.56%，为未来 870 的总误差。然而，SI 模型的节点数量的预测结果严重低估了实际数字，最高可达现实 200% 偏差。

① http://www.ericsson.com/mobility-report

2.4.4　微观随机生成过程

到目前为止，我们已经检查了网潮模型的准确性和实用性。在这里，我们通过两个微观随机过程生成器，即网潮-存活模型 NetTide-Survival 和网潮-过程模型 NetTide-Process，来产生逼真的宏观动态增长曲线。

2.4.4.1　NETTIDE-Survival 产生的随机动态增长

网潮-生存生成器 NetTide-Survival 可以直接生成节点和链路增长的随机轨迹。对于不同的模型参数组合 $(\beta, \theta, N, \beta', \alpha, \gamma)$，NetTide-Survival 可以生成不同的节点和链接轨迹。这里我们报告一个特定的实验设置 $(\beta = 2.45\times10^{-4}, \theta = 1, N = 10^4, \beta' = 0.5, \alpha = 0.9, \gamma = 0.52)$ 这个动力学增长设置类似于微信，它遵循幂律增长，节点和链接的幂律增长指数为 2.15 和 3.01。我们开源了代码，用于在任意参数设置中生成随机动态增长过程。

图2.6a-d 展示了 10 条 NetTide-Survival 产生的随机增长轨迹，每一条轨迹由不同颜色区分。虽然 NetTide-Survival 产生了不完全相同的 $n(t)$，$\frac{dn(t)}{dt}$，$e(t)$，和 $\frac{de(t)}{dt}$，且他们有着随机的波动，但是我们的网潮模型 NetTide 都很好地刻画了这些随机轨迹。由于 $\theta = 1$, NetTide-Survival 会产生类似幂律的动态增长，如图 2.6 e-f。我们通过 NetTide-Survival 生成了 1,000 个 n(t) 和 e(t) 实例，并通过最小二乘法拟合生成的幂律早期增长。图 2.6c&f 分别显示节点和链接幂律增长的幂律指数的分布，表明 NetTide-Survival 生成具有幂律指数接近 2.15 和 3 的 n(t) 和 e(t)，意味着 NetTide-Survival 成功再现了微信的动力学增长过程。

2.4.4.2　NETTIDE-Process 产生的随机动态增长

对于第二个随机生成器，我们试图从网络内的微观随机交互进行建模，即网潮过程生成器 NetTide-Process。给定基础组织结构 G_0，我们模拟所需的建模参数是 (β, θ, β')。正如在模型小节所讨论的，在随机图假设下，时间间隔 $(t, t + h)$ 中用户的感染概率 $p_n(t)$ 和链接概率 $p_e(t)$ 可以从等式2-19和2-20推导出来。我们再次在 $(\beta = 2.485 \times 10^{-4}, \theta = 1, \beta' = 0.48)$ 和 $G_0 = RandomGraph(\alpha = 0.01, N = 10^4)$ 的实验设置下报告我们的结果，即再次产生与微信数据相似的动态增长过程。为了在任意参数设置下生成随机轨迹，我们开源了代码实现。

NetTide-Process 产生的随机动力学增长曲线如图2.7所示。同样，实曲线几乎击中每个点，表明网潮模型很好的拟合了随机 $n(t)$，$\frac{dn(t)}{dt}$，$e(t)$，和 $\frac{de(t)}{dt}$。通过使用上述的参数设定，我们可以生成 1,000 条随机增长实例。我们发现生成的轨迹显示幂律早期增长（如图2.7 e 和 f 的插图所示），n(t) 和 e(t) 的幂律指数接近 2.15 和 3，

表明 NetTide-Process 产生了如微信数据的动态增长过程。

我们进一步检查网潮模型是否可以学习出 NetTide-Process 使用的“真实”参数。在拟合过程中，由于随机图设置，我们设置 $\gamma = 1$。我们发现网潮推断出建模参数，如图2.8所示。我们的网潮模型不仅可以找到控制行为的 β，θ 和 β' 的值，还可以发现结构参数 $N = 10^4$ 和 $\alpha = 0.01$ 。

总之，网潮-生存模型 NetTide-Survival 和网潮-过程模型 NetTide-Process 都可以生成节点和链接的逼真的随机增长过程，并且生成的随机增长轨迹可以很好地被网潮模型刻画。

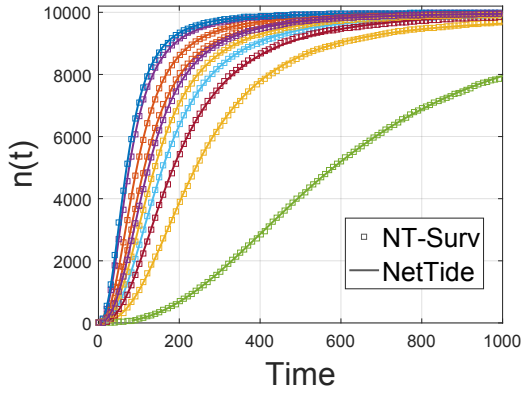
2.5 结论

在本文中，我们研究了现实世界中社交网络的动态增长，并提出了动力学模型来同时刻画节点和链接的动态增长过程。我们研究了一系列真实演化的社交网络，特别是中国最大的在线社交网络微信。我们发现真实社交网络中的节点和链接都随时间遵循幂律增长，而不是预期的指数增长或是均匀线性增长。因此，我们提出了网潮模型 NetTide ，包含节点和链接数量增长的微分方程，来刻画新的增长现象。我们的网潮-节点模型 NetTide-Node 给出了一个统一但简约的模型来刻画真实的社交网络增长，比如 Log-Logistic 增长的幂律早期增长，以及 Stretched-Logistic 增长的拉伸指数早期增长的更一般形式。我们的网潮-链接模型 NetTide-Link 是第一个捕捉链接动态增长的动力学模型。网潮模型准确地拟合了现实动态增长数据。此外，我们提出了两个微观随机生成过程，即网潮-存活生成器 NetTide-Survival 和网潮-过程生成器 NetTide-Process，它们分别从生存分析和网络内的微观随机交互的角度生成真实的随机动态增长过程。我们的模型再次准确地拟合了产生的随机动态增长过程，并很好地推断出其参数。总结我们的主要贡献如下：

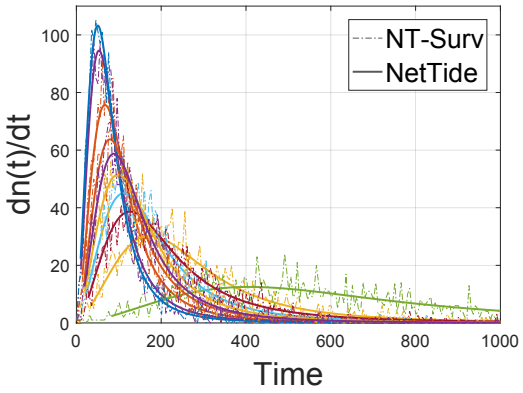
1. **网潮动力学模型**: 我们给出产生幂律动力学增长的机制，并提出网潮模型。网潮-节点模型 NetTide-Node 刻画了广泛的动力学增长过程，网潮-链接模型 NetTide-Link 是第一个刻画链接动力学增长的微分方程。这两个方程都是简约的，可以在微观层面上解释。
2. **准确性**: 我们在四个真实世界中不断演化的社交网络数据上进行了实验，特别是微信（3 亿个节点，47.5 亿个链接）。我们的网潮模型 NetTide 准确地刻画了现实世界的动态增长过程。
3. **实用性**: 我们的模型可用于短期和长期预测。我们通过大量实验验证了模型的预测能力，并表明它可以准确地预测短期甚至长期的节点和边的数量（对微信和 arXiv 的未来预测分别为 730 天和 870 天）。

4. **生成器**：我们提出了两个微观层面的随机过程生成器，即网潮-生存生成器 NetTide-Survival 和网潮-过程生成器 NetTide-Process，它们生成了逼真的随机动力学增长过程。

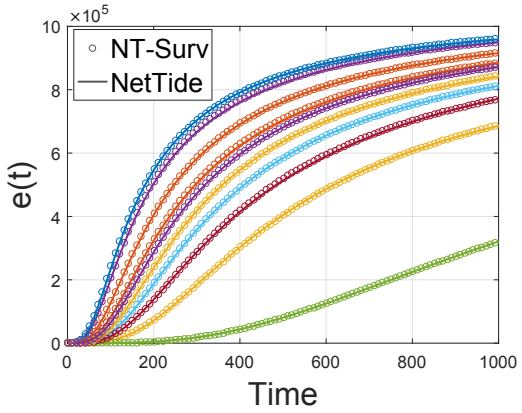
我们的工作存在着许多进一步研究的可能。首先，提出的增长模型可用于验证其他领域的普遍增长现象，如生态学，社会科学，人口学等。其次，受到模型建模参数的物理意义的启发，如何改善社交网络服务，提升社交行为体验是一个开放性的问题。第三，我们模型的一个主要限制是忽视外部影响。外部信号如何影响社交网络的增长或衰退动态仍有待研究。我们开源了网潮模型 NetTide 及两个生成器 NetTide-Survival 和 NetTide-Process 的代码，以及拟合/生成节点和链接的确定性/随机的动态增长轨迹数据，代码网址为：https://github.com/calvin-zcx/NetTide。



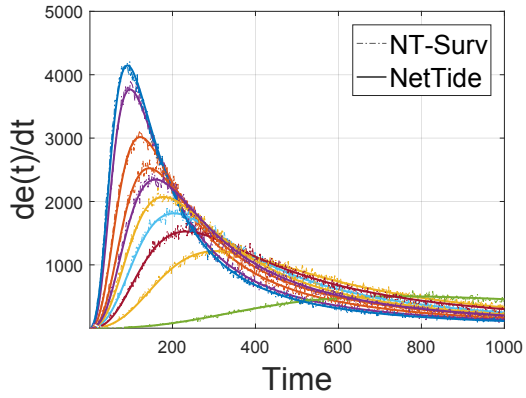
(a) n(t)



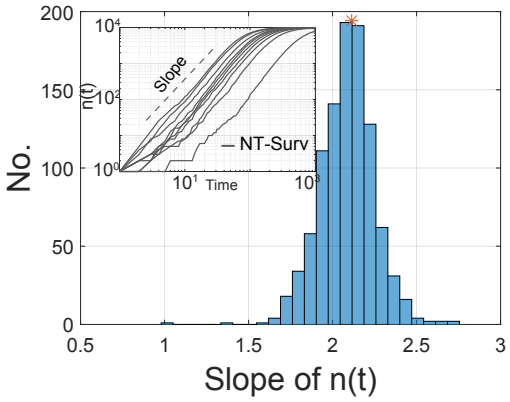
(b) $\frac{dn(t)}{dt}$



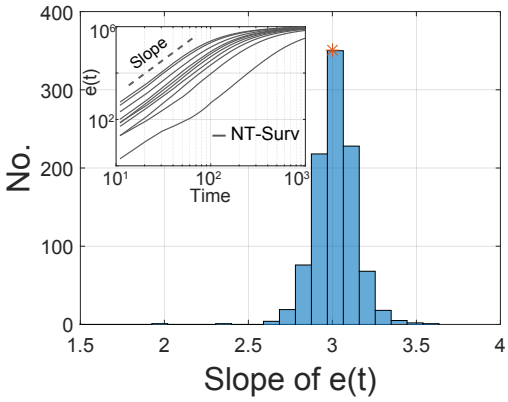
(c) e(t)



(d) $\frac{de(t)}{dt}$

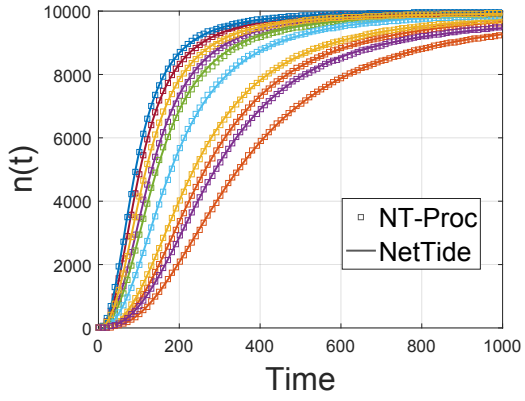


(e) Slope distribution of n(t)

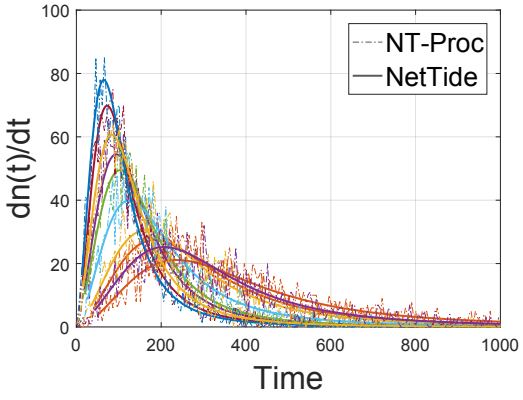


(f) Slope distribution of e(t)

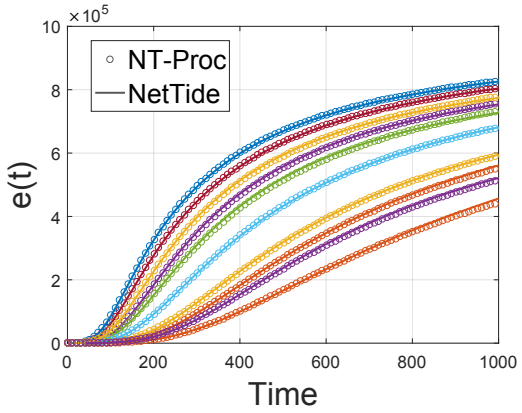
图 2.6　网潮-存活随机生成器 NetTide-Survival 产生了逼真的动力学增长过程。(a-b) NetTide-Survival 产生的 10 条 n(t) 累计增长曲线 (方块标记 □) 和 10 条 $\frac{dn(t)}{dt}$ 速率曲线 (点划线)。实线是我们模型拟合的结果。每条颜色的线代表着一个模拟样本。(c-d) 是对链接模拟的结果。(e-f) 是 n(t) 和 e(t) 幂律增长指数的直方图。由 NetTide-Survival 分别产生 1,000 条 n(t) 和 e(t) 曲线。红心代表着真实微信数据的结果。其中插图展现 n(t) 和 e(t) 在双对数坐标下的幂律初期增长，即类似直线。



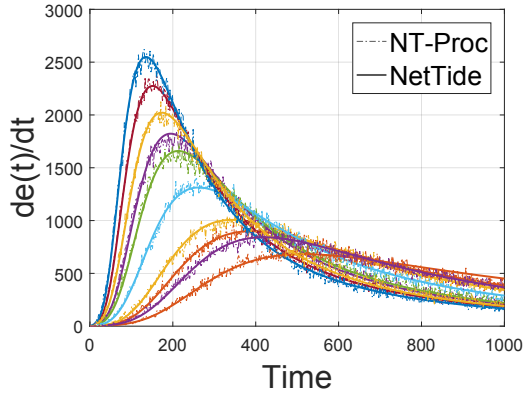
(a) n(t)



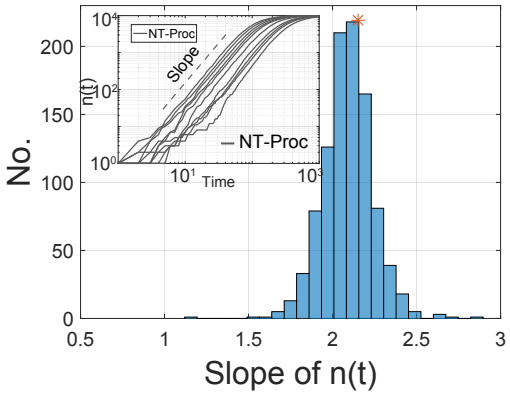
(b) $\frac{dn(t)}{dt}$



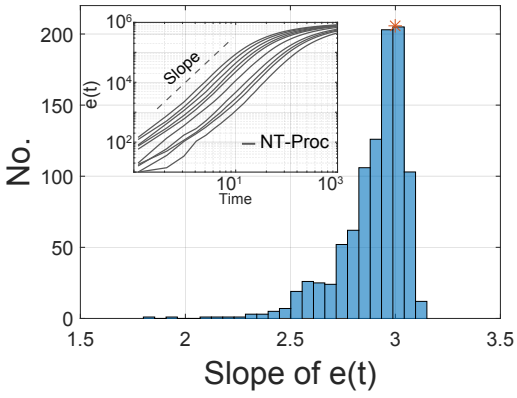
(c) e(t)



(d) $\frac{de(t)}{dt}$

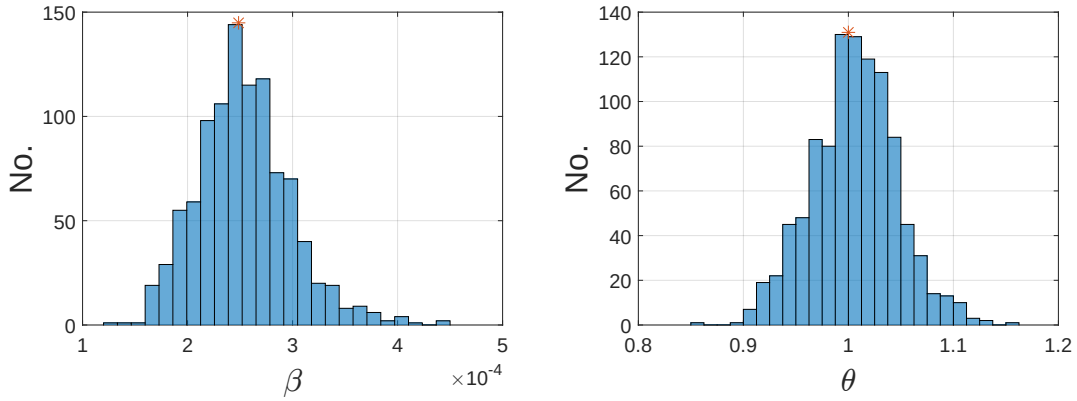


(e) Slope distribution of n(t)



(f) Slope distribution of e(t)

图 2.7　网潮-过程随机生成器 NetTide-Process 产生了逼真的动力学增长过程。(a-b) NetTide-Process 产生的 10 条 n(t) 累计增长曲线 (方块标记 □) 和 10 条 $\frac{dn(t)}{dt}$ 速率曲线 (点划线)。实线是我们模型拟合的结果。每条颜色的线代表着一个模拟样本。(c-d) 是对链接模拟的结果。(e-f) n(t) 和 e(t) 幂律增长指数的直方图。由 NetTide-Process 分别产生 1,000 条 n(t) 和 e(t) 曲线。红心代表着真实微信数据的结果。其中插图是双对数坐标，展现 n(t) 和 e(t) 的幂律初期增长。



(a) Node: β　(b) Node: θ

(c) Node: N　(d) Link: β'

(e) Link: α　(f) Link: γ

图 2.8　网潮-过程随机生成器 NetTide-Process 产生的模拟数据可以被网潮模型 NetTide 很好的学出。(a-f) 有生成的 1,000 个实例学习出的参数直方图，其实例如图2.6。红星代表参数真实数值，我们设置 $\gamma = 1$，代表着随机网络情形。

第 3 章　微观社交网络演化规律发现和建模

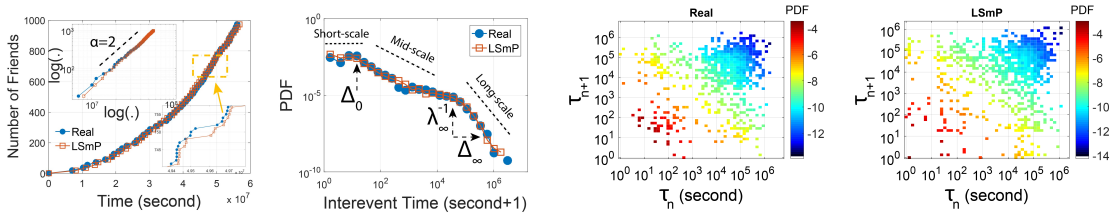
每个人在社交网络中如何动态地交朋友？你的社交关系数随着时间如何演化？控制这些时间维度模式形成的基本机制是什么？无论是虚拟在线社交网络还是物理社会系统，它们的结构和动力学特性都是由微观个人的动态社交行为所驱动。然而，由于缺乏真实的微观数据，对复杂社交系统微观社交行为层面的动力研究少之又少，更不用说探究这些微观动态模式的规律或模型了。

我们研究了中国最大的在线社交网络 “微信” 的详细增长过程，包含它拥有头 3 亿用户和 47.5 亿连接的两年。我们发现了广泛异构的用户动态加好友行为模式，但是他们在长期都遵循非线性幂律（随机）增长（Long-term Power-law Growth），在短期呈现随机爆发增长（Short-term Bursty Growth ）的普遍规律。我们提出三个关键因素，即平均效应（Average Effect），多尺度效应（Multi-scale Effect）和相关效应（Correlation Effect），们它控制着微观层面观察到的随机增长模式。基于此，我们提出了包含这些成分的长短记忆随机过程（Long Short Memory Process），它成功地再现了真实数据中观察到的复杂随机增长现象。通过分析模型参数，我们发现了个体社交行为背后的统计规律。我们的模型和发现为网络动态增长的微观社交机制提供了理论基础，而且其应用范围广泛，例如对人类动态行为的预测，聚类和异常值检测等等。

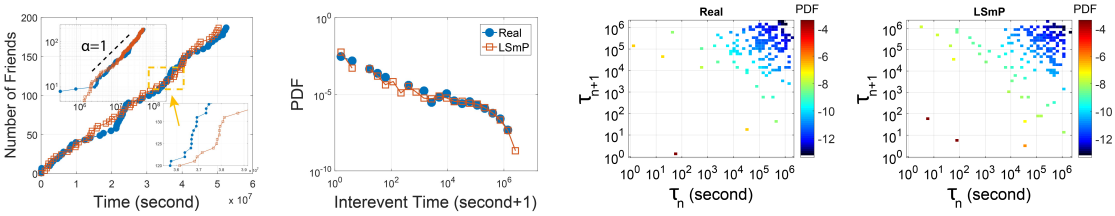
3.1　引言

近年来，很多科研工作致力于理解和建模社交网络的增长，包括复现社交网络的结构的统计特性[2] 到建模节点和链接的动力学增长过程[3]，旨在增强我们的对社交复杂系统的认识理解。从微观层面来讲，社交复杂系统的结构和动力学特性主要是由个人社交行为的动力学特性驱动的。然而，现有研究中对微观个人社交行为的动力学特性知之甚少，主要是由于缺乏大规模的详细记录社交网络演变的微观数据集。因此，如下很基本的问题在很大程度上是未知的：微观个体在社交网络中的社交行为的动态增长模式是什么？控制这些微观动态增长的机制是什么？我们能否用一个简单的数学模型将异构个体的社交动态行为统一刻画？对这些问题的回答不仅可以促进我们对社交复杂系统演变规律的理解，而且还提供了在微观层面刻画人类动态行为的模型。其应用广泛，例如预测个人未来增长现象，聚类人类的动态行为模式，进行用户画像，以及检测异常行为/用户等等。

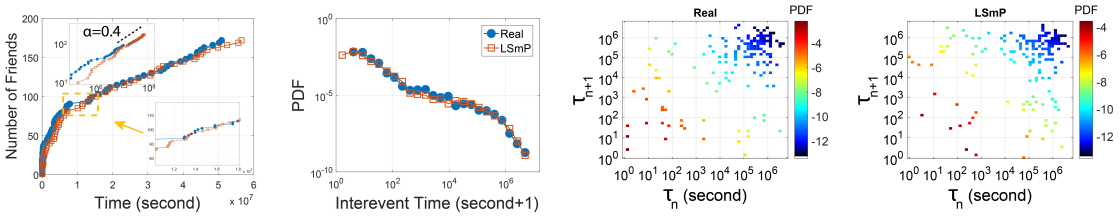
在现有文献中，关于微观个人社交链接动力学的研究主要是在理论上进行的。



(a) 加速幂律用户　(b) 加速幂增长律用户 (c) 加速幂增长律用户 (d) 加速幂增长律用户



(e) 线性增长用户　(f) 线性增长用户　(g) 线性增长用户　(h) 线性增长用户



(i) 减速幂律增长用户 (j) 减速幂律增长用户 (k) 减速幂律增长用户 (l) 减速幂律增长用户

图 3.1　微观个人的社交链接，在长期呈现出各种各样的非线性随机幂律增长，在短期呈现随机爆发增长。长短记忆随机过程模型 LSMP 很好地刻画了真实数据。（a）(e）(i）分别绘制了三个不同增长模式的实例：加速幂律增长，线性增长和减速幂律增长。相同曲线在双对数坐标系的图在上部插图中，而下部插图放大了短期爆发增长。每一行描述了相同动态增长的不同方面。（b）(f）(j）绘制了事件间间隔时间（IET）分布。（c）(g）(k）绘制连续 IET 的联合分布，而（d）(h）(l）是 LSMP 生成的联合分布。我们的模型在所有方面都很好地刻画了真实数据。

Barabási-Albert 模型[2] 生成了无标度网络，预测个体好友数与时间之间的幂律关系指数为 0.5。Fitness 模型[15] 将此指数扩展为一个范围 < 1 的 fitness 函数。然而，这些理论假设忽略了无处不在的边稠密化现象[9]，而真实动态增长的曲线可能具有幂指数值 ≥ 1。最近,[11] 发现构建社交链接的速率平均而言随着系统现有边数增长，表明平均而言边的超线性增长。然而，它忽略了不同用户之间的异质性。文献[50] 研究表明社交链接构建的速率是恒定的，表示平均而言边的线性增长，但其结果也是平均而言的，忽略了个人的异质性，并且连接时间戳是根据用户行为日志估计的，而不是明确记录的。综上所述，尚没有大规模数据支持微观个人社交行

为演化的研究。

在本文中，我们研究了微信从开始到 3 亿用户和 47.5 亿条社交链接横跨 2 年时间的详细演变过程。数据集明确记录谁和谁在何时构建社交关系的详细过程，是支持该研究的现有最大数据集之一。从这个数据集中，我们发现个人社交行为的动态增长表现出丰富的复杂性，远远超出了我们目前的理解。图3.1 展示了三个示例用户在两年内其好友数随时间的动态增长过程（之后实验部分我们进一步说明这三个用户在全局用户池所处的位置）。我们发现：i 在长时间尺度内广泛的非线性幂律随机增长，以及 ii 在短时间尺度内随机爆发增长。基于这些经验观察，我们提出了三个控制动态增长的因素，即平均效应（average effect），多尺度效应（multi-scale effect）和相关效应（correlation effect）。基于上述因素，我们提出一个随机过程模型，即长短记忆随机过程 **L**ong- **S**hort- **M**emory **P**rocess（**LSMP**）来刻画真实用户复杂社交行为的随机动力学特性。我们通过多个方面验证了长短记忆过程模型 LSMP 的有效性。通过分析模型的参数，我们进一步发现了用户社交链接动态增长背后的统计规律。通过在参数空间中聚类，我们发现了长期社交链接增长的三种典型模式，以及短期内的两种典型模式。我们还说明了在参数空间中可以很容易地检测到异常值。综上所述，我们总结贡献如下：

- **科学发现**：个人社交链接的动态增长在长期呈现非线性幂律随机增长，在短期呈现随机爆发增长。我们发现三个机制，即平均效应，多尺度效应和相关效应，来解释非线性幂律和突发性增长的动力学特性。
- **长短记忆过程模型**（LSMP）：一个随机过程模型来建模如上科学发现。它是简约的，参数具有明确的物理意义。
- **准确性**：LSMP 准确地拟合了个人动态社交链接的随机增长模式。
- **实用性**：LSMP 对人类动态行为提出了深刻理解，并可应用于行为预测，聚类，模式发现和异常值检测等。

本章的大纲如下：相关工作，模型，实验结果和最后的结论。为了保证实验的可复现性，我们的代码开源在了https://github.com/calvin-zcx/LSMP。

3.2　相关工作

由于所研究的问题与网络演化（Network evolution）和人类行为动力学（Human dynamics）密切相关，我们主要回顾这两个领域的相关工作。

网络演化。经典的 Barabási-Albert model (BA) 模型[2] 预测个人社交好友数随着时间呈现 $k(t) \sim t^{1/2}$，其幂指数为 $\alpha = 0.5$。进一步，fitness 模型[15] 扩展了 BA 模型来刻画异构的动力学过程，将幂指数由固定值设置为 $\alpha < 1$ 的范围。然而，我们

发现现实世界中链接的动态过程更加复杂：其动态幂律很多都大于 1，即 $\alpha >= 1$。最近，工作[11] 发现平均而言社交链接数随着现有边数的增长而增长，虽然其刻画了边稠密现象，但是忽略了不同用户的异质性。工作[50] 发现社交链接的增长率是常数，但是其结果也是人群的均值，而且其链接构建时间是通过对话行为推测的，并没有确切建边时间数据。工作[3] 研究了多个真实社交网络，发现宏观网络节点和社交链接随着时间都遵循幂律增长，暗示平均而言个人社交链接也遵循幂律动力学增长。综上所述，我们发现没有现有工作研究真实社交网络中大规模异构用户加好友行为的动态随机过程。

人类行为动力学。人类的各种行为通常展现爆发（bursts）和长尾（heavy-tailed）事件间隔时间（inter-event time，IET）分布。优先级队列[44] 和调制泊松过程[45] 被提出来建模 IET 的长尾动力学特性。最近，越来越多的证据表明了多尺度 IET 分布[46]，包括双模态分布[47] 和多模态分布[46] 等等。但是，这些工作虽然刻画了 IETs 分布，但是忽略了事件间的关联性[51]。有两个研究分支尝试刻画爆发性和关联性：自激励过程（Self-exciting Process），例如 Hawkes 过程[40,42,52]，和自反馈过程（Self-feeding Process）[53–55]。但是，他们都不能产生我们在真实数据中普遍观测到的任意幂指数的非线性幂律增长，也不能产生我们观测到的多尺度间隔时间 IET 分布。

3.3　长短记忆随机过程

在本节中，我们提出了控制长期随机幂律增长和短期随机爆发增长的机制，然后提出了基于这些机制的的随机过程模型。

3.3.1　建模基础

在这里，我们将介绍启发我们模型设计的直觉和背景。

动力学幂律增长的机制。 Zang et al.[3] 提出动力学模型来刻画社交网络中的节点幂律增长过程：$\frac{dn(t)}{dt} = \frac{\beta}{t^\theta} n(t)(N - n(t))$, 其中 N 是人口最大上限，β 是增长率。这个式子可以产生一系列动力学增长模式，包括幂律增长为其特例（当 $\theta = 1$ 和 $n(t) \ll N$ 的时候）。通过定义 $\alpha = \beta N$，我们可以得到产生幂律增长的极简动力学方程式：

$$\frac{dn(t)}{dt} = \frac{\alpha}{t} n(t) \text{ 。} \tag{3-1}$$

因此，当衰减指数 $\theta = 1$ 时，这个等式表示的微观层面的物理意义是：用户当前加好友的速率 $\frac{dn(t)}{dt}$ 与长期加好友的平均速率 $\frac{n(t)}{t}$ 成正比。这个长期记忆是由到

表 3.1　符号和定义

符号	定义
$n(t)$	特定用户在 t 时刻的累积事件数
t_i	事件 i 的发生时刻
τ_i	事件 i 和事件 $i-1$ 的事件间隔事件（inter-event time，IET）
$\mathcal{H}_t$	时刻 t 之前的事件历史，一系列事件事件数据
$\lambda(t\|\mathcal{H}_t)$	LSMP 模型的强度函数（Intensity Rate）
$\Phi_\infty(t)$	刻画长期记忆的事件强度函数
λ_∞	长期事件的最大发生率
α	长期增长指数
Δ_∞	长期记忆时间尺度
$\Phi_0(t)$	刻画短期记忆的事件强度函数
λ_0	短期记忆最大发生率
θ	短期记忆衰减系数
Δ_0	短期记忆时间尺度
m	短期记忆覆盖事件长度

目前为止所加好友 $n(t)$ 所刻画的。我们将其命名为平均效应（average effect）。正是平均效应产生了动力学幂律增长。但是如何产生随机的幂率增长?

爆发增长的机制。 人类行为通常表现出多尺度的连续事件时间间隔（Inter-Event-Time, IET）分布[46]，如图3.1 b，其在短时间尺度上呈现平台，中时间尺度呈现长尾形状，长时间尺度呈现指数分布形状。当我们 i.i.d. 从该多尺度分布中采样 IET 时，一系列值很小的 IET 短期在短期被采样，然后偶尔采样到中尺度和长尺度较大值的 IET，即一个是一系列密集活动，之后是几个长时间间隔，这样就产生爆发行为。因此，从多尺度 IET 分布刻画了自上次事件以来的短期记忆，这个多尺度短期记忆就可以产生爆发行为。我们将此机制称作多尺度效应（multi-scale effect）。如何产生人类多尺度的记忆行为?

给定一个泊松过程，会表现出均匀的线性性增长。如果我们按如下方式重新排列 IET 序列：使小 IET 后跟较小的 IET，并使大值 IET 后跟较大的 IET，然后我们也得到突发行为。因此，IET 之间的相关性也可以产生爆发增长。图片 3.1 cg &k 描绘了两个连续 IET 与实际数据的联合分布。如果 IET 是独立的，则联合分布上的点应该展开的，而实际上点是聚集的，例如红点，这支持了我们 IET 相关的猜想。实际上，对应联合分布图的动态增长图 3.1ae&i 和其下部插图，展示了短期内的爆发式增长。我们把这个产生爆发行为的机制叫做关联效应（correlation effect）。如何产生关联的随机爆发行为?

模型框架-随机点过程。用户在特定时间下添加朋友的过程是随机时间点过程。

在数学上，我们想要知道下一个事件的时间 t，定义让 $f(t|\mathcal{H}_t)$ 为第 $(n+1)^{th}$ 个时间发生时间的条件概率密度，以 t 时刻之前所有行为历史记录 $\mathcal{H}_t$ 作为条件。这个点过程，由事件强度函数 $\lambda(t|\mathcal{H}_t)$[56] 唯一刻画。事件强度函数 $\lambda(t|\mathcal{H}_t) = \mathbb{E}[n(dt)|\mathcal{H}_t]/dt$ 描述了给定历史记录 $\mathcal{H}_t = (t_1, ..., t_{n-1}, t_n)$ 条件下事件发生率。其中，$n(dt)$ 表示在 $[t, t+dt)$ 内发生事件数，$n(t)$ 代表在 $(-\infty, t)$ 时间内累计事件数。强度函数和条件概率密度函数之间的关系是 $\lambda(t|\mathcal{H}_t) = \frac{f(t|\mathcal{H}_t)}{1-F(t|\mathcal{H}_t)}$ 和 $f(t|\mathcal{H}_t) = \lambda(t|\mathcal{H}_t)e^{-\int_{t_n}^{t}\lambda(s|\mathcal{H})_t ds}$。

3.3.2 长短记忆随机过程模型

在此，我们提出长短记忆随机过程模型（**L**ong-**S**hort-**M**emory **P**rocess, LSMP），它是基于平均效应，多尺度效应和相关效应构建的模型，用来刻画微观个体用户加好友过程（亦或是其他点过程）长期的非线性幂律随机增长，和短期随机爆发式增长。刻画 LSMP 的强度函数为：

$$\begin{aligned}\lambda(t|\mathcal{H}_t) &= \underbrace{\lambda_\infty \alpha(\frac{t}{\Delta_\infty}+1)^{\alpha-1}}_{\text{长期记忆 } \Phi_\infty(t|\mathcal{H}_t)} + \underbrace{\sum_{i=n(t)-m+1}^{n(t)} \lambda_0(\frac{t-t_i}{\Delta_0}+1)^{-\theta}}_{\text{短期记忆 } \Phi_0(t|\mathcal{H}_t)} \\ &= \Phi_\infty(t|\mathcal{H}_t) + \Phi_0(t|\mathcal{H}_t)。\end{aligned} \tag{3-2}$$

3.3.2.1 长期记忆部分

我们首先表明平均效应会产生幂律**随机**增长，呈现出长期记忆效果。

引理 3.1：（**随机幂律增长机制**）均值效应，即当前的强度函数 $\Phi_\infty(t|\mathcal{H}_t)$ 是过去强度函数从 $-\Delta_\infty$ 到 t 的均值，产生了幂律随机增长，其幂律指数为 α。

证明 我们建模均值效应为：

$$\Phi_\infty(t|\mathcal{H}_t) = \frac{\alpha}{t+\Delta_\infty}\int_{-\Delta_\infty}^{t}\Phi_\infty(s|\mathcal{H}_t)ds\ , \tag{3-3}$$

并且替代积分项为：$\int_{-\Delta_\infty}^{t}\Phi_\infty(s)ds = \lambda_\infty\Delta_\infty(\frac{t}{\Delta_\infty}+1)^\alpha$，我们得到：

$$\Phi_\infty(t|\mathcal{H}_t) = \lambda_\infty\alpha(\frac{t}{\Delta_\infty}+1)^{\alpha-1}。 \tag{3-4}$$

□

强度函数 $\Phi_\infty(t|\mathcal{H}_t) = \lambda_\infty\alpha(\frac{t}{\Delta_\infty}+1)^{\alpha-1}$ 刻画了个人加好友行为的长期随机幂律增长过程，是动力学方程 3-1 刻画的动力学幂律增长的随机过程版本。

模型的进一步说明：

- **广谱的幂律增长**。 在第3.3.1小节所述, α 是在时间间隔 $(t, t+dt)$ 内由任意一个 $n(t)$ 中已有好友所介绍新好友平均数。当 $\alpha > 1$，加速的幂律增长出现，表明富者越富（rich-get-richer）的现象：你有越多的好友，就倾向于交更多的好友。当 $\alpha = 1$，$\Phi_\infty(t|\mathcal{H})$ 等于 λ_∞，表明 $n(t)$ 呈现匀速线性增长，没有长期记忆效应。相反，当 $\alpha < 1$ 时，$n(t)$ 表现出减速的幂律增长。上述三种情况在图 3.1 ae&i 中描述。
- **多尺度效应的长时间尺度部分**。长时间尺度部分由参数 λ_∞ 和 α 决定。当 $\alpha = 1$ 时，$\Phi_\infty(t|\mathcal{H}_t)$ 退化为泊松过程（Poisson process），刻画了均匀的例行加好友行为。具体来讲，一个事件预计以 λ_∞^{-1} 的时间间隔发生，其时间尺度由 Δ_∞ 控制，如图 3.1 b 所示。当 α 增加，长时间尺度的指数分布部分向左移动，表明越来越小的 τ，即加速增长的 $n(t)$。相反，当 α 减小时，该长时间尺度的指数分布向右移动，表明越来越大的 τ，即减速增长的 $n(t)$。
- IETs **的长期关联**。 当 $\alpha \neq 1$ 时候，长期记忆出现，呈现 IETs 的长期关联性。确实，当 $\alpha \geq 1$ 时候，IETs 越来越小，表明逐渐加速增长的朋友圈；当 $\alpha \leq 1$ 时，IETs 越来越大，表明逐渐减速和饱和的朋友圈。

3.3.2.2　短期记忆部分

短期记忆部分 $\Phi_0(t|\mathcal{H}_t)$ 刻画了多尺度效应和关联效应，他们一起产生了短期爆发式的增长。

引理 3.2：　LSMP 的强度函数 $\lambda(t|\mathcal{H}_t)$ 当 $\alpha = m = \theta = 1$ 时，其在短时间尺度产生常值的 IET 分布，在中时间尺度产生幂律分布，在长时间尺度产生指数分布。

证明　下个事件，例如第 $(n+1)$ 个事件，发生的时间间隔条件概率密度函数是 (参考第3.3.1小节):

$$
\begin{aligned}
f(t|\mathcal{H}_t) &= \lambda(t|\mathcal{H}_t) e^{-\int_{t_n}^{t} \lambda(s|\mathcal{H}_t) ds} \\
&= (\Phi_\infty(t|\mathcal{H}_t) + \Phi_0(t|\mathcal{H}_t)) e^{-\int_{t_n}^{t} \Phi_\infty(s|\mathcal{H}_t) ds} e^{-\int_{t_n}^{t} \Phi_0(s|\mathcal{H}_t) ds} \\
&= [\lambda_\infty \alpha (\frac{t}{\Delta_\infty} + 1)^{\alpha-1} + \sum_{i=n-m+1}^{n} \lambda_0 (\frac{t-t_i}{\Delta_0} + 1)^{-\theta}] \\
&\times e^{-\lambda_\infty \Delta_\infty [(\frac{t}{\Delta_\infty}+1)^\alpha - (\frac{t_n}{\Delta_\infty}+1)^\alpha]} \\
&\times \prod_{i=n-m+1}^{n} e^{-\int_{t_n}^{t} \lambda_0 (\frac{s-t_i}{\Delta_0}+1)^{-\theta} ds} 。
\end{aligned}
$$

当 $\alpha = 1$ 时，没有长期记忆效应。当 $m = 1$ 时，短期记忆完全由自上个事件开始的衰减效应控制。因此，$\lambda(t|\mathcal{H}_t)$ 刻画了一个更新过程（**renewal process**），即 $f(t|\mathcal{H}_t) = f(\tau_{n+1}) = f(\tau)$。当 $\theta = 1$ 时，我们得到：

$$\begin{aligned} f(t|\mathcal{H}_t) &= [\lambda_\infty + \lambda_0(\frac{t-t_n}{\Delta_0}+1)^{-1}]e^{-\lambda_\infty(t-t_n)}(\frac{t-t_n}{\Delta_0}+1)^{-\lambda_0\Delta_0} \\ &\approx \lambda_0(\frac{t-t_n}{\Delta_0}+1)^{-(1+\lambda_0\Delta_0)}e^{-\lambda_\infty(t-t_n)}, \quad \text{当 } \lambda_\infty \ll \lambda_0 \text{ 时。} \end{aligned}$$

因此，当 $\alpha = m = \theta = 1$ 时，IET 的分布是：

$$f(\tau) \approx \lambda_0(\frac{\tau}{\Delta_0}+1)^{-(1+\lambda_0\Delta_0)}e^{-\lambda_\infty\tau}, \tag{3-5}$$

其在短时间尺度（$\tau < \Delta_0$）有着常数速率 λ_0，中时间尺度有着幂律分布，其幂律指数是 $1 + \lambda_0\Delta_0$，长时间尺度部分有着指数分布（τ 在 λ_∞^{-1} 附近）。 □

模型的进一步说明：

- **多尺度 IET 分布**： IET 分布的多尺度模式也适用于其他的模型参数设置情形：
  - 短尺度 IET 分布的范围由 Δ_0 控制，如图3.1 b 所示。当 Δ_0 增大时，短尺度的范围向右扩展；如果 Δ_0 减小，短尺度的范围向左扩展。当 $\Delta_0 \to 0$ 时，IET 分布的短时间尺度部分消失。
  - 中时间尺度 IET 分布的斜率主要由 θ 控制。当 $\alpha = m = 1$ 时，$\theta = 1$ 的情况显示在引理3.2中证明。当 $\theta \neq 1$ 时，$e^{-\int_{t_n}^{t}\lambda_0(\frac{s-t_n}{\Delta_0}+1)^{-\theta}ds}$ 是一个拉伸的指数分布（stretched exponential distribution），在中等规模上展示了长尾 IET 分布。因此，θ 越大，斜率越大。
- IET **间的短期相关性**： 相关效应和多尺度效应都会影响 IET 的短期相关性：
  - 当 $\alpha = m = 1$ 时，LSMP 是个更新过程，表明 IETs 之间没有相关性。当 $m > 1$ 时，相关性出现。当 $m = \infty$，$\Phi_0(t|\mathcal{H}_t)$ 类似有着幂律衰减核的 Hawkes 过程，其超临界态（$\int \Phi_0(s|\mathcal{H}_t)ds > 1$）产生指数增长，表明最强的正相关性。
  - θ 的值越大，上个事件后对未来下个事件的影响随时间衰减越来越大，所以 IET 间的关联越来越小。
- **爆发**： 爆发由多尺度效应和关联效应共同决定：
  - 如图3.1 b 所示，Δ_0 是短时间尺度（小 IET，密集行为）和中时间尺度（长尾 IET 行为）的临界点。因此 Δ_0 越大，$n(t)$ 爆发性越强。θ 越小，$n(t)$

的爆发性越强。因为 θ 越小，IET 分布的中间尺度越高，所以 IET 被越来越多的相对长的时间间隔分开。

– 由于先前事件的叠加影响，较长的记忆长度引起较强的短期 IET 相关性，导致突然的大速率，即爆发。

3.3.3 模型参数估计

我们通过最大化对数似然函数来学习 LSMP 的参数。在时间段 $[0,T)$ 观察到点过程 $\{t_1,...,t_{n-1},t_n\}$ 的对数似然函数是：

$$\begin{aligned}\log L(t_1,...,t_n) &= -\int_0^T \lambda(t|\mathcal{H}_t)dt + \int_0^T \log\lambda(t|\mathcal{H}_t)dN(t)\\ &= -\lambda_\infty\Delta_\infty[(\frac{t_n}{\Delta_\infty}+1)^\alpha - 1]\\ &- \sum_{i=1}^{n}\frac{\lambda_0\Delta_0}{1-\theta}[(\frac{t_{i+m}-t_i}{\Delta_0}+1)^{1-\theta}-1]\\ &+ \sum_{i=1}^{n}\log[\lambda_\infty\alpha(\frac{t_i}{\Delta_\infty}+1)^{\alpha-1}+\lambda_0 A(i)]\ ,\end{aligned} \tag{3-6}$$

其中 $A(i)=\sum_{t_{i-m}\le tj<t_i}(\frac{t_i-t_j}{\Delta_0}+1)^{-\theta}$ 对于 $i\ge 2$，并且 t_i 表述第 i 个事件发生的时间，$A(1)=0$。为方便起见，我们假设 $T=t_n$。最大化如上对数似然函数来得到参数 $\{\lambda_\infty,\alpha,\Delta_\infty,\lambda_0,\theta,\Delta_0,m\}$，并满足 $\{\alpha,\Delta_\infty,\theta,\Delta_0,\lambda_\infty,\lambda_0\ge 0; m\in\mathcal{N}\}$ 的限制，可以得到我们想要的最优模型估计参数。

我们模型的一个很好的优点是所有参数都具有可推导出的梯度。长记忆部分的梯度是：

$$\frac{\partial\log L}{\partial\lambda_\infty} = -\Delta_\infty[(\frac{t_n}{\Delta_\infty}+1)^\alpha-1] + \sum_{i=1}^{n}\frac{\alpha(\frac{t_i}{\Delta_\infty}+1)^{\alpha-1}}{D(i)} \tag{3-7}$$

$$\begin{aligned}\frac{\partial\log L}{\partial\alpha} &= -\lambda_\infty\Delta_\infty(\frac{t_n}{\Delta_\infty}+1)^\alpha\ln(\frac{t_n}{\Delta_\infty}+1)\\ &+ \sum_{i=1}^{n}\frac{\lambda_\infty(\frac{t_i}{\Delta_\infty}+1)^{\alpha-1}[\alpha\ln(\frac{t_i}{\Delta_\infty}+1)+1]}{D(i)}\end{aligned} \tag{3-8}$$

$$\begin{aligned}\frac{\partial \log L}{\partial \Delta_\infty} =& -\lambda_\infty[(\frac{t_i}{\Delta_\infty}+1)^\alpha - 1] + \lambda_\infty \alpha \frac{t_n}{\Delta_\infty}(\frac{t_n}{\Delta_\infty}+1)^{\alpha-1} \\ & - \sum_{i=1}^{n} \frac{\lambda_\infty \alpha(\alpha-1)\frac{t_i}{\Delta_\infty^2}(\frac{t_i}{\Delta_\infty}+1)^{\alpha-2}}{D(i)}\end{aligned} \tag{3-9}$$

其中 $D(i) = \lambda_\infty \alpha(\frac{t_i}{\Delta_\infty}+1)^{\alpha-1} + \lambda_0 A(i)$。

当 $\theta \neq 1$，短期记忆部分的剃度是：

$$\frac{\partial \log L}{\partial \lambda_0} = -\frac{\Delta}{1-\theta}\sum_{i=1}^{n}[(\frac{t_{i+m}-t_i}{\Delta}+1)^{1-\theta} - 1] + \sum_{i=1}^{n}\frac{A(i)}{D(i)} \tag{3-10}$$

$$\begin{aligned}\frac{\partial \log L}{\partial \Delta_0} =& -\frac{\lambda_0}{1-\theta}\sum_{i=1}^{n}[(\frac{t_{i+m}-t_i}{\Delta_0}+1)^{-\theta}(\theta\frac{t_{t+m}-t_i}{\Delta_0}+1) - 1] \\ & + \sum_{i=1}^{n}\frac{\lambda_0 B(i)}{D(i)}\end{aligned} \tag{3-11}$$

$$\begin{aligned}\frac{\partial \log L}{\partial \theta} =& -\frac{\lambda_0 \Delta_0}{(1-\theta)^2}\sum_{i=1}^{n}[(\frac{t_{i+m}-t_i}{\Delta_0}+1)^{1-\theta} - 1 \\ & - (1-\theta)(\frac{t_{i+m}-t_i}{\Delta_0}+1)^{1-\theta}\ln(\frac{t_{i+m}-t_i}{\Delta_0}+1)] \\ & - \sum_{i=1}^{n}\frac{\lambda_0 C(i)}{D(i)}\end{aligned} \tag{3-12}$$

其中 $B(i) = \sum_{t_{i-m}\leq tj<t_i}\frac{\theta(t_i-t_j)}{\Delta_0^2}(\frac{t_i-t_j}{\Delta_0}+1)^{-\theta-1}$ 对于 $i \geq 2$ 和 $B(1) = 0$，且 $C(i) = \sum_{t_{i-m}\leq tj<t_i}(\frac{t_i-t_j}{\Delta_0}+1)^{-\theta}\ln(\frac{t_i-t_j}{\Delta_0}+1)$ 对于 $i \geq 2$ 且 $C(1) = 0$。当 $\theta = 1$ 时，对应的梯度可以很容易得到。在数值计算 $\theta = 1$ 情况时，我们可以复用以上公式，只需要给 θ 加一个很小的误差，比如 10^{-8}。

对应确定的 m，我们求解如上带约束的优化问题的优化算法时 trust-region-reflective algorithm[57]。我们通过枚举 m 的值来根据最大似然原则[11] 挑出最优的参数。为了实验的可重复性，我们开源了代码，参加章节 3.5。

3.3.4　模拟生成算法

我们通过 Inverse Method[56] (P260, Algorithm 7.4.III.) 设计了 LSMP 的模拟方法，由于当 $\alpha > 1$ 时 $\lambda(t|\mathcal{H}_t)$ 不是有界的。Inverse Method 的基本操作是通过求解等

式 $\log u + \int_{t_n}^{t} \lambda(s|\mathcal{H}_t)ds = 0$ 中的 t，其中 u 是由均匀分布 $U(0,1)$ 产生的。当 $\alpha \leq 1$ 时，我们也可以通过采用 Ogata's thining algorithm[58] 来得到加速的模拟过程。为简明起见，我们在这里只给出 Inverse Method。我们开源了我们的代码，包括模拟代码，详情参阅第 3.5节。

Input : Intensity function of LSMP

$$\lambda(t|\mathcal{H}_t) = \Phi_\infty(t|\lambda_\infty, \Delta_\infty, \alpha) + \Phi_0(t|\lambda_0, \Delta_0, \theta), \text{ total event number } N$$

Output : $\{t_1, ..., t_N\}$

Set current number of events $n = 1$, and current time $t = 0$;

while $n \leq N$ **do**

  Sample $u \sim Uniform([0, 1])$;

  Solve $\log u + \int_t^x \lambda(s|\mathcal{H}_t)ds = 0$ for x by Algorithm 2.;

  $t = t + x$;

  $t_n = t$;

  $n = n + 1$;

end

Algorithm 1: 长短记忆随机过程 Long-Short-Memory Process (LSMP) 的模拟算法

Input : Equation $F(x) = \log u + \int_{t_n}^{x} \lambda(s|\mathcal{H}_t)ds$

Output : x

Set $\epsilon = 10^{-8}$, $x = t_n - t_{n-1}$;

while $|F(x)| \leq \epsilon$ **do**

  $x = x - \frac{F(x)}{F'(x)}$;

end

Algorithm 2: 牛顿迭代法

3.4 实验结果

在本节中，我们将通过大规模真实数据来评估 LSMP 模型的有效性。我们在第3.4.1小节介绍数据集，在第3.4.2小节中，我们从三个方面验证了真实数据集的准确性。更重要的是，我们的模型可以应用于许多数据挖掘任务中。在第3.4.3小节

中，通过分析建模参数，我们发现了经验增长动力学背后的统计规律。在第3.4.4中，我们聚类了人社交连接增长的典型模式，并检测异常用户行为。

3.4.1　数据集

我们的实验是在微信上进行的。微信是中国最大的在线社交网络，截至 2018 年 6 月 30 日，每月活跃用户超过 10.58 亿。我们收集了微信的历史数据，包括从 2011 年 1 月 21 日（微信的发布日）开始完整记录的所有节点和链接数据，到 2013 年 1 月 16 日当注册用户达到 3 亿。总共有 3 亿注册用户和超过 4.75 亿美元的好友链接。我们的记录数据是 (u, t_u) 和 $(u, v, t_{u,v})$，其中 u 和 v 代表用户的 ID，t_u，$t_{u,v}$ 分别是节点和社交链接建立的时间戳。为了获得足够的个人数据来研究每个人的动态加好友行为，我们按度数排序并选择头 1 百万用户（好友数 ≥ 143），并随机选择此群体的 10% 用于以下实验。我们可以访问的所有微信数据都是匿名的，并遵守了其严格的隐私政策。

3.4.2　准确性

我们通过回答我们的 LSMP 模型是否可以准确刻画每个人的真实动态加好友过程。我们的实验验证了以下三个方面的准确性：一，增长形状，二，多尺度 IET 分布，以及 三，IET 的相关性。

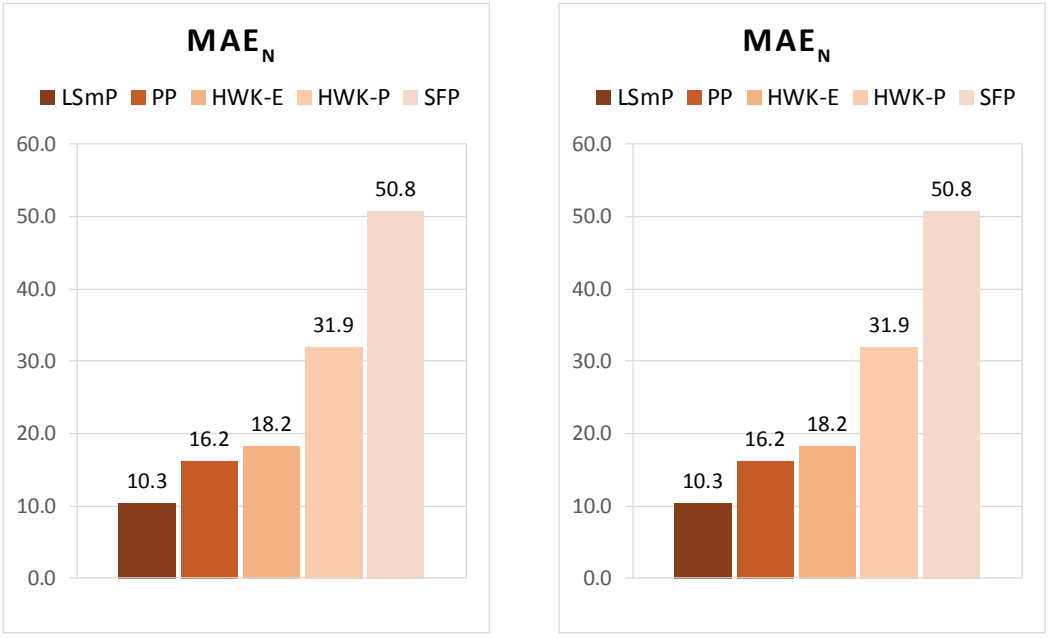
3.4.2.1　对比的基准模型

我们对比了以下四种最新的具有代表性的随机过程模型，其分布为：

1. 泊松过程（Poisson process，PP)），其有着常数强度函数 $\lambda(t|\mathcal{H}_t) = \lambda$，产生随机线性增长。
2. 指数核函数的 Hawkes 过程 (HWK-E)，其强度函数为 $\lambda(t|\mathcal{H}_t) = \mu + \sum_{t_i<t} \alpha e^{-\beta(t-t_i)}$ [59]。
3. 幂律核函数的 Hawkes 过程 (HWK-P)，其强度函数是 $\lambda(t|\mathcal{H}_t) = \mu + \sum_{t_i<t} \alpha(t - t_i)^{-\beta}$ [41]。HWK-P 有着临界相变现象[42]。
4. 自反馈过程（Self-feeding process，SFP)[53,55]，刻画了 IET 间的 Markovian 关联性。其强度函数是 $\lambda(t|\mathcal{H}_t) = \frac{1}{\mu/e+\tau_i}$。

3.4.2.2　增长形状的准确性

我们评价了对于每个人 i 真实数据 $n_i(t) = |\{t_j < t | t_j \in \mathcal{H}_t = (t_1, ..., t_{n_i})\}|$ 和模型生成数据 $\hat{n_i}(t) = |\{\hat{t_j} < t | \hat{t_j} \in \hat{\mathcal{H}_t} = (\hat{t_1}, ..., \hat{t_{n_i}})\}|$ 动态加好友过程的形状准确性。其



(a)　　(b)

图 3.2　在两种度量情况下，和比竞争对手相比，LSMP 都更准确地拟合了真实数据。(a)，对累计事件数 $n_i(t)$ 的平均均值误差（MAE）的中位数，即 $MAE_N(i)$；(b)，以及对 $n_i(t)$ 事件时刻 t 的平均均值误差的中位数，即 $MAE_T(t)$。

准确性由累积好友数的平均均值误差 $MAE_N(i) = \frac{\sum_{j=1}^{n_i} |n(t_j) - n(\hat{t}_j)|}{n_i}$，以及加好友事件时间的平均均值误差 $MAE_T(i) = \frac{\sum_{j=1}^{n_i} |t_j - \hat{t}_j|}{n_i}$ 刻画。对于每个人，我们计算 $MAE_N(i)$ 和 $MAE_T(i)$，图3.2 绘制了 $n_i(t)$ 的 MAE 的中位数。

我们提出的 LSMP 模型与基线模型相比，就所提出的两个度量指标而言，都有着大幅度提升，更加准确地刻画了真实数据。平均而言，LSMP 刻画 $n(t)$ 每个事件时间 t_i 的平均误差为 10.3。相比之下，PP，HWK-E，HWK-P，SFP 给出的 MAE_N 分别为 LSMP 的 1.8 倍，3.1 倍，4.9 倍。类似的结果与事件时间 t_i 的误差指标一致：PP，HWK-E，HWK-P，SFP 给出的 MAE_T 是 LSMP 误差的 1.7 倍，1.8 倍，3.1 倍，18.7 倍。最糟糕的结果是 SFP 给出的，因为它可能会产生高估的 IET。HWK-P 给出了第二大误差，因为具有幂律核函数的 Hawkes 过程很容易处于超临界状态，这会产生指数增长。PP 和 HWK-E 给出类似的结果，因为它们产生线性增长。

3.4.2.3　多尺度 IET 分布的准确性

然后我们验证了 LSMP 模型对随机动态增长过程的多尺度 IET 分布的刻画效果。我们比较每个人的动态加好友的 IET $\tau_i = \{\tau_1, ..., \tau_{n_i}\}$ 和模型生成的对应 $\hat{\tau}_i = \{\hat{\tau}_1, ..., \hat{\tau}_{n_i}\}$。通过假设检验 τ_i 和 $\hat{\tau}_i$ 是否来自同一个连续分布。双样本 Kolmogorov-

表 3.2　模型刻画多尺度 IET 分布的 KS-Test 结果。我们的 LSMP 很好地刻画了 IET 的多尺度分布。最好的结果由粗体标识。

	KS-Test (Sig. α = 5%)				
τ	LSMP	PP	HWK-E	HWK-P	SFP
Pass Rate	**87.1%**	7.8%	62.9%	37.0%	2.4%
Error Statistic	**0.098**	0.256	0.134	0.162	0.291
$\log(\tau)$	NetTide	PP	HWK-E	HWK-P	SFP
Pass Rate	**87.1%**	7.8%	62.9%	37.0%	2.4%
Error Statistic	**0.098**	0.256	0.134	0.162	0.291

Smirnov 检验（KS -Test）适用于此标准的假设检验任务。拟合效果通过两个指标测试：假设检验通过率（Pass Rate），和经验数据和拟合结果之间的差异程度（Error Statistics）。

表 3.2 给出了 LSMP 和对比模型对真实数据拟合的平均通过率和误差统计指标，我们设定假设检验的显著程度为 α = 5%。我们的模型在所有指标上都大幅度好于对比模型：我们模型产生的随机过程以 87.1% 的通过率刻画了真实数据，有着 0.098 的平均误差。如果我们设定显著程度为 α = 1% 我们模型的通过率会增加到 94.1%。我们还对对数转换的 IET $\log(\tau)$ 进行了同样的假设检验。我们得到的 $\log(\tau)$ 的结果与 τ 的相同。

但是，所有对比的基准模型都无法很好的刻画多尺度 IET 分布：

- PP 只能在长时间尺度上产生指数分布，在短时间尺度和中时间尺度都远远偏离实际。
- HWK-E 在短时间尺度和长时间尺度展现出两种指数分布的混合，但是其忽略了中等规模的长尾分布。
- HWK-H 不能刻画短时间尺度的平台，即密集行为部分。
- SFP 不能刻画短时间尺度的平台，并且在中时间尺度上也表现出较大的偏差。

3.4.2.4　多个 IET 间关联性的准确性

进一步实验表明，我们的 LSMP 模型很好的刻画了 IET 序列的关联性。我们通过二维版本的 Kolmogorov-Smirnov 测试（2D-KS-Test）[60] 来比较真实数据的联合分布 $p(\tau_{i,j}, \tau_{i,j+1})$ 和模型产生的 $p(\hat{\tau}_{i,j}, \hat{\tau}_{i,j+1})$ 的差异。我们将显着性水平设置为 5%。表 3.3记录了 LSMP 和其他对比模型的平均通过率和误差统计量。我们的模型再一次在两个指标上都显著地击败了所有对比模型。我们进一步研究了 2D-KS-Test 的拟合优度在多大程度上受边际一维分布的影响。我们随机打乱了真实数据

表 3.3　模型刻画多尺度 IET 联合分布的 2D-KS-Test 结果。我们的模型 LSMP 很好地刻画了 IET 的二维联合分布，最好的结果由粗体标注。

	2D-KS-Test (Sig. α = 5%)				
τ	NetTide	PP	HWK-E	HWK-P	SFP
Pass Rate	**85.5%**	7.9%	61.1%	34.1%	2.0%
Error Statistic	**0.145**	0.329	0.189	0.220	0.371
Shuffled τ	NetTide	PP	HWK-E	HWK-P	SFP
Pass Rate	**77.5%**	7.9%	57.2%	27.3%	1.9%
Error Statistic	**0.159**	0.329	0.194	0.231	0.384
$\log(\tau)$	NetTide	PP	HWK-E	HWK-P	SFP
Pass Rate	**85.4%**	6.7%	62.0%	34.0%	2.6%
Error Statistic	**0.145**	0.329	0.189	0.220	0.371

和模型拟合产生数据地 IET 序列，然后再一次比较打乱顺序后的 $p_s(\hat{\tau}_{i,j'}, \hat{\tau}_{i,j'+1})$ 和 $p(\hat{\tau}_{i,j}, \hat{\tau}_{i,j+1})$。我们发现 LSMP 得到最大的通过率 77.5%（如表 3.3的第二个面板所示），说明我们的模型对一维边缘分布的刻画优势。更重要的是，由于很好的刻画了 IET 之间相关性的表征，LSMP 还有额外 8%（= 85.5% – 77.5%）的增加。对数转换的 IET $\log(\tau)$ 的相同实验中，我们得出相同的结论。

综上所述，只有我们的 LSMP 模型准确地拟合了真实随机增长曲线的形状，多尺度 IET 分布，和 IET 的联合分布。

3.4.3　参数分析之社交规律发现

LSMP 模型的一个优点是所有建模参数都具有明确的物理意义。在本节中，我们通过参数分析找到了真实加好友模式背后的统计规律。**据我们所知，这是对人类微观加好友行为的首次数据驱动的验证**，例如对 α 和 θ 的研究是第一次实证研究。

图3.3 绘制了真实微信数据驱动的 LSMP 六个参数的分布。我们通过 Levenberg-Marquardt 算法[61] 拟合每个参数的分布，并用以置信区间 95% 报告它们的系数值。我们得到以下人们长期加好友行为的发现：

非线性幂律增长的复杂性和异质性。 图3.3 a 绘制了幂律增长的幂指数 α 的分布，我们发现它是三个子分布构成的混合分布，具有如下复杂性：

- 第一个子分布：偏正态分布（skewed normal distribution，$\mu = 0.70$，$\sigma = 0.57$，$s = 2.6$）接近 1，其中 μ 是平均位置，σ 是标准偏差和 s 是偏斜度。位置值表示

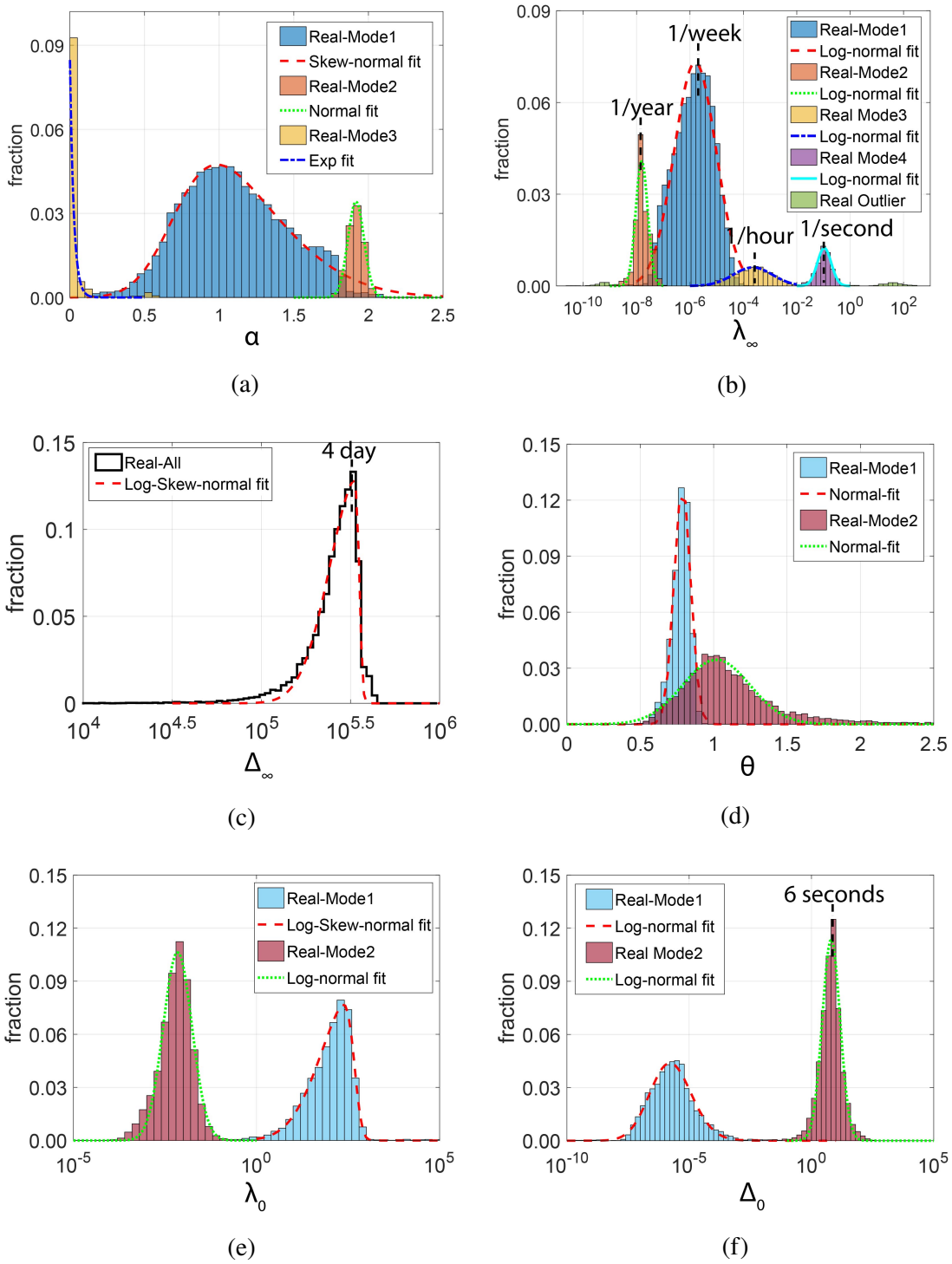


图 3.3　**社交行为参数分布**。短期记忆参数：(a) α，(b) λ_∞，(c) Δ_∞；长期记忆参数：(d) θ，(e) λ_0，(f) Δ_0。不同颜色代表不同子类，相同的颜色代表相同的子类。整体分布以黑线标志。每个簇的拟合曲线用彩色线表示。

平均而言用户加好友的动力学过程呈现减速幂律增长，幂律指数为 $\alpha \sim t^{0.7}$。正偏度值表明存在大量线性增长（$\alpha = 1$）和加速幂律增长（$\alpha > 1$）。相对

较大的方差意味着人们长期加好友动态模式的异质性。此外，这些微观层面的研究结果与我们之前的发现一致[3]，即在宏观层面上节点随着时间的推移按照 $n(t) \sim t^{2.15}$ 增长，社交链接随时间 $e(t) \sim t^3$ 增长，因此平均好友数按照 $\frac{e(t)}{n(t)} \sim t^{0.85}$ 增长。

- 第二个子分布：正态分布（normal distribution，$\mu = 1.9$，$\sigma = 0.077$）接近 2，意味着存在一些具有较大幂律指数的用户，其上限接近 ≈ 2。这种约束的原因可能源于这样的事实：在社交网络中要链接的最大链接数量与到目前为止的节点数量的平方成比例。
- 第三个子分布：指数分布（exponential distribution，$\alpha \propto e^{-39.4x}$）接近 0，表明该区域的动态增长过程在长期显著减速。有限的好友数，源于个人好友圈的迅速饱和，或是用户的退出。

长尺度加好友速率遵循混合 Log-normal 分布。人们平均一个星期加一个好友！图 3.3 b 绘制了长尺度加好友速率 λ_∞ 的分布图，我们发现了一个混合了四个 Log-normal 子分布的混合模型，其横跨超过 8 个数量级，范围从 1 /秒（模式四），1 /小时（模式三），1 /周（模式二）到 1 /年（模式一），对应模式如图 3.3 b 所示。大部分的人们长期加好友速率处于模式一，最好的 log-norm 拟合结果为（$\log(\lambda_\infty)$, $\mu = -5.80, \sigma = 1.11$），表明人们以平均而言每 7.3 天的周期添加一个朋友。模式二（$\mu = -3.65, \sigma = 1.04$）中较 de 小 λ_∞ 意味着在初始生命周期内的很长一段时间没有加好友。而在模式三（$\mu = -0.97, \sigma = 0.37$）和模式四（$\mu = -0.97$，$\sigma = 0.37$）拥有较大的 λ_∞，意味着在最初的生命周期中发生了爆发式的密集行为。我们将在下一小节中讨论 λ_∞ 和 α 之间的相互关系及其含义。

长期加好友的时间尺度是 4 天。 图3.3 c 描绘了 Δ_∞ 的分布。它被 Skewed Log-norma 分布（$\mu = 5.55, \sigma = 0.17, s = -12.00$）很好的拟合，表明人们加好友的典型时间尺度是 4 天（$= 10^\mu$ 秒）。其社会学含义为，比如，当你星期一开始一份新工作后，并逐渐与你的同事建立一系列好友关系，直到星期五告一段落了，正好是四天后！

在短时间尺度内，人们动态加好友的规律为：

高斯混合模型的衰减指数。 图3.3 d 绘制了短期记忆核函数的衰减指数 θ，它遵循两种模式的高斯混合分布：模式一（$\mu = 0.79, \sigma = 0.080$）和模式二（$\mu = 1.02, \sigma = 0.32$）。模式 2 中 $\theta = 1$ 的情况符合优先级队列模型的预测[44]，表明幂律衰减（参见引理3.2）。但是我们发现 $\theta = 1$ 只是一种临界态，$\theta \neq 1$ 为更为一般的情况。我们发现只有在临界状态 $\theta = 1$ 下才是幂律衰减，而非临界态，$\theta \neq 1$ 下是拉伸的指数衰减幂指数衰减（stretched exponential decay）。特别是我们发现对于模式

1，其平均值为 0.79，表明 IET 分布更加长尾。

双峰的短期增长速率。 图3.3 e 绘制了短期增长率 λ_0 的分布，其遵循双峰分布，其中偏斜的对数正态分布（skewed log-normal distribution，$\mu = 2.68, \sigma = 0.88, s = -5.63$）很好的拟合了模式一，而对数正态分布（log-normal distribution，$\mu = -2.16, \sigma = 0.56$）很好的拟合了模式二。分离的两种模式清晰地展示了人们加好友行为在短期 IET 分布上的两个子类，我们将在下一小节详细讨论这些子类。

短期内时间尺度呈双峰分布，加朋友最快仅需 6 秒。 图3.3 f 绘制了短期时间尺度 Δ_0 的分布情况，其也是遵循双峰分布。模式 2 很好的由对数正态分布（log-normal distribution，$\mu = 0.81$，$\sigma = 0.51$）刻画，其平均值为 6 秒，表明 IET 分布的短时间尺度下快速操作和中时间尺度下长尾模式之间的临界时间。实际上，微信提供了多种方法来添加朋友，例如“摇一摇”。通过摇动移动电话，用户可以添加同时摇动的陌生人，从而构建随机链接。模式一遵循对数正态分布（log-normal distribution，$\mu = -6.03$，$\sigma = 1.3$），平均值为 $10^{-5.77}$ （~ 0）秒，表明这个子类中的用户的 IET 分布只包括中时间尺度和长时间尺度（图 3.1 d）。此外，我们发现这三个短期参数之间存在很强的相关性：Δ_0 分布模式一中的人口对应于 λ_0 分布模式一中的较大的 λ_0 值，以及 θ 分布模式一中的 0.79 衰减指数部分。

3.4.4　动态社交行为聚类及异常检测

我们在模型参数的高维联合分布中进一步聚类人们加好友的行为模型。

由于我们数据中（偏斜/对数）正态分布的普遍存在，我们将高斯混合模型应用于长期记忆参数空间 $(\alpha, \lambda_\infty, \Delta_\infty)$ 和短期参数空间 $(\theta, \lambda_0, \Delta_0)$，发现如图 3.3所示的子类/簇。具有不同颜色的柱状图表示不同的子类，而不同边缘分布图中的相同颜色表示相同的簇的不同侧面。我们进一步在高维参数空间进行聚类模式[62]，如图 3.4 和图 3.5 中的联合分布所示。这些联合分布可以找到人们加好友长期和短期的不同模式。

长期增长模式。 图 3.4 描绘了我们将人们加好友行为投影到 LSMP 所刻画长期行为参数（λ_∞, α）空间，并且我们发现了正如我们 LSMP 所预测的三个主要的加好友行为模式：

- **线性增长**。线性增长位于 $\alpha = 1$ 区域附近，由 Cluster1 模式附近的人口组成。$\alpha = 1$ 具有线性增长，如图3.4 c 所示。
- **加速幂律增长**。 加速幂律增长在 $\alpha > 1$ 区域内，包含 Cluster1 的上半部分，以及 Cluster2 中的二次动态增长区域。二次增长（$\alpha = 2$）是加速幂律增长的特例，如图3.4 b 所示。

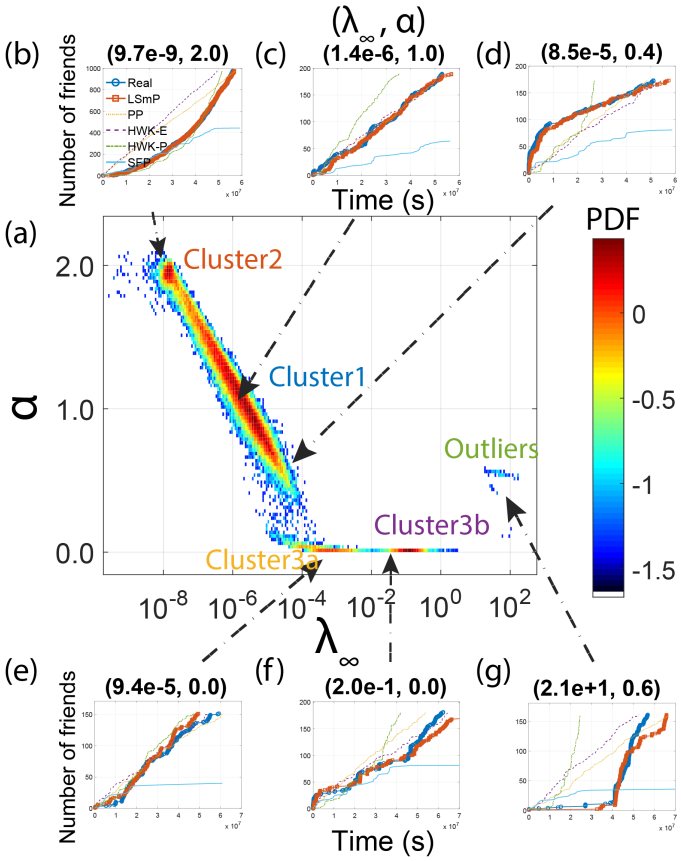


图 3.4　长期加好友的动态增长行为聚类。(a) 在对数-线性坐标系下长期加好友速率 λ_∞ 和幂律增长指数 α 的联合分布。其对应的一维边缘分布如图 3.3。不同子类中的样例在 b-g 图中展示。

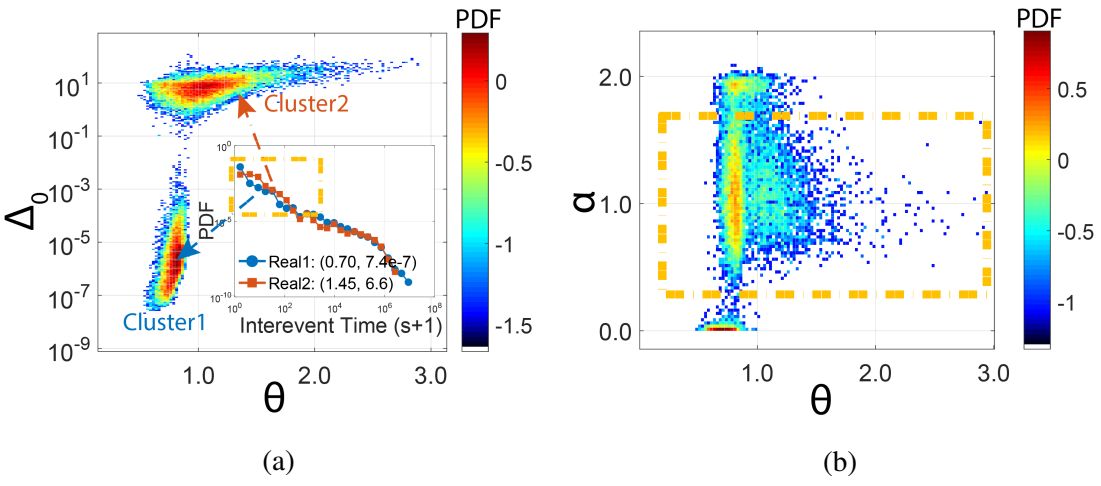


图 3.5　短期加好友的记忆行为聚类。(a) 在线性-对数坐标系下短期记忆衰减指数 θ 和短期时间尺度 δ_0 的联合分布。插图刻画了在子类 1 和 2 中典型的连个用户的 IET 分布。(b) 短期记忆衰减指数 θ 和长期幂律增长指数 α 在线性-线性坐标系下的联合分布。

- **减速幂律增长**。减速幂律增长位于 $\alpha < 1$ 的范围内，由 Cluster1 的下半部分（3.4d 的一个实例）和 Cluster3a 和 3b 组成。图3.4 e 和 f 分别在 Cluster3a 和 Cluster3b 中绘制两个实例。沿着 λ_∞ 轴，Cluster3a 在初始时间表现出很大的延迟，然后在小 λ_∞ 值区域之后减小幂律增长，而 Cluster3b 子 λ_∞ 值较大的区域，先是初始突发，然后是减速幂律增长。

短期增长模式。图3.5 a 绘制了短期记忆参数空间（θ, Δ_0）- 的联合分布图。我们发现短期动态的两种模式是由它们是否在 IET 分布中显示短尺度而分开，如图3.5 a 的插图所示。

长期增长模式和短期增长模式之间的关联。我们进一步研究了长期记忆参数和短期记忆参数之间的相关性。图3.5 b 描绘了短期记忆核函数衰减指数 θ 和长期幂律增长指数 α 的联合分布。我们发现大多数动态行为（在 Cluster1 中）就 θ 而言显示出很大的方差，这意味着长期记忆和短期记忆之间的相关性相对较小。

异常检测。另外，我们在参数空间中展示了一个异常检测应用。通过检查 λ_∞ 和 α 的联合分布，我们进一步找到了异常值。图3.4 g 绘制了在异常点云中的一个增长实例，具有长时间休假和突然的大爆发增长。这个用户在第一个 1.3 年内几乎没有任何行为动作，然后他/她在几分钟内增加了 50 个朋友。我们发现基线根本无法捕捉到这些现象，但我们的 LSMP 甚至在某种程度上很好的刻画这个异常值。此外，简约的特征维度和清晰的物理意义有助于为未来的工作设计更精细的异常值检测方法。

3.5 结论

在本文中，我们研究了复杂社交系统微信中微观个人加好友行为的随机动态过程。我们发现，对于人们好友数随时间增长曲线，在长期呈现随机非线性幂律增长，在短期呈现爆发式增长。我们进一步发现，平均效应可以解释幂律增长，多尺度效应以及相关效应可以解释爆发式增长。基于这些机制，我们提出长短记忆随机过程 LSMP，来刻画每个人加好友随机过程的动力学特性。我们的 LSMP 是一个统一的但简约的模型，它可以准确地刻画异构的用户行为，并全面地解释微观用户社交连接的增长动态过程。综上所示，我们的主要贡献归纳如下：

- **全新的发现**：我们在微观层面分析了真实社交网络中个人用户加好友的动态过程，我们发现人们好友数随时间增长曲线在长期呈现随机非线性幂律增长（stochastic power-law growth），在短期呈现随机爆发式增长（bursty growth）。我们进一步发现了三个机制，即平均效应（average effect），多尺度效应（multi-scale effect）和相关效应（correlation effect）来解释幂律增长和爆发性增长。

- **全新的随机过程** LSMP：我们提出了一个随机过程模型来刻画上述发现。LSMP 是简约的，所有参数都有明确的物理意义。
- **准确性**：我们在真实社交网络中验证了我么的模型。我们的 LSMP 模型可以准确地刻画真实的随机增长模式。
- **实用性**: 我们的 LSMP 模型进一步加深了对人们社交动态行为的深刻理解，它可以应用有模式发现如聚类，异常检测，行为预测等等应用。并且它是一个全新的随机过程，有理论研究价值。为了保证实验的可重复性，我们开源了所有实验代码，见网址：https://github.com/calvin-zcx/LSMP。

局限性和未来的工作。我们观察到的复杂的个体社交行为背后的社会学机制需要进一步研究。由于生长现象无处不在，我们的模型是否可以应用于其他系统的动力学增长也是一个有趣的问题，从线虫 C.elegans[63] 的连接增长到公司企业的成长等等。

第 4 章　信息流在网络中传播的复杂模式生成

模式生成（Pattern formation）是自然界中一种无处不在的现象，它描述了自组织系统产生有规律输出模式的过程。无论是在线社交网络或是物理社会中，社交互动行为模式主要由信息流驱动。尽管越来越多的研究旨在了解信息流的传播机制，但对于这些传播模式的几何形状以及它们在传播过程中是如何形成的却知之甚少。在本文中，通过探索从大规模在线社交媒体数据集中提取的 4.32 亿个信息流模式，我们在一个三维度量空间发现信息流传播结构的复杂几何特性。相比之下，对信息流传播结构的几何模式的现有理解仅限于扇形展开（fanning-out）或狭窄的树状（narrow tree-like）的几何形状。我们发现了控制复杂信息流复杂几何模式生成的三个关键因素。之后，我们提出了一个包含这些因素的随机过程模型，证明它可以成功复现真实信息流传播模式中发现的复杂几何形状。我们的发现为信息流的微观机制提供了理论基础，其可能的应用包括对信息的预测，控制，舆情分析和政策决策等等。

4.1　引言

模式生成（Pattern formation）是自然界中普遍存在的现象[64,65]，例如物理研究[66–69]，生物学研究[70,71]，化学反应[72]，和社交行为[73–77] 等等。在过去的几年中，很多科学研究工作致力于研究和建模物理世界中个人社交行为的信息流[78,79]，和在线社交网络中的信息传播[80,81]，旨在理解信息在社交系统中流动的机制。然而，人们对信息传播过程中形成的结构的几何模式知之甚少[3,31,82–86]。现有模型如传染病模型（epidemic models）[87–90] 和分支过程模型（branching process）[91–93] 易于产生扇形（或称作星形）模式（图 4.1C 和 D）。然而，现实世界的模式似乎更复杂。例如，Liben-Nowell 和 Kleinberg 在互联网连锁信（chain letter）数据[78] 中发现了一种具有狭窄且深层的树结构模式。这些发现提出了许多重要问题：复杂的传播模式可以在多大程度上形成？以及控制复杂模式形成的潜在机制是什么？回答这些问题不仅增强了我们对传播模式形成的理解，而且提供了预测信息流的计算工具，可能导致应用于新技术的传播[3,94–96] 以及公众舆论的形成[97–100]。

由于缺乏真实的大规模记录信息流传播结构的数据集，对信息流模式几何复杂性的研究还很缺乏。实际上，大多数研究的数据集通常缺乏显示的跟踪信息流从何而来的信息。例如，Twitter 数据集不会为每条推文提供明确的“转发”标记，从而难以推断出适当的信息流方向[101,102]。在本文中，我们探索了一个大规模的社

交媒体数据集，其中包括 7 天内用户中的 1.01 亿用户和 4.32 亿信息级联传播。当前数据跟踪创建或转发帖子时的所有步骤，是解决上述问题的最准确数据集之一。值得注意的是，我们从现实世界的扩散模式中发现了复杂的几何图形（图 4.1A），揭示了信息流模式形成过程中出现的复杂性。接下来，我们将介绍三个关键指标来表征模式的几何特性，发现现有模型与真实数据之间的系统性的差异。最后，我们发现了三种新的机制来控制信息流的模式形成，从而形成一个随机过程模型，成功地捕捉了现实世界中观察到的复杂的几何模式。

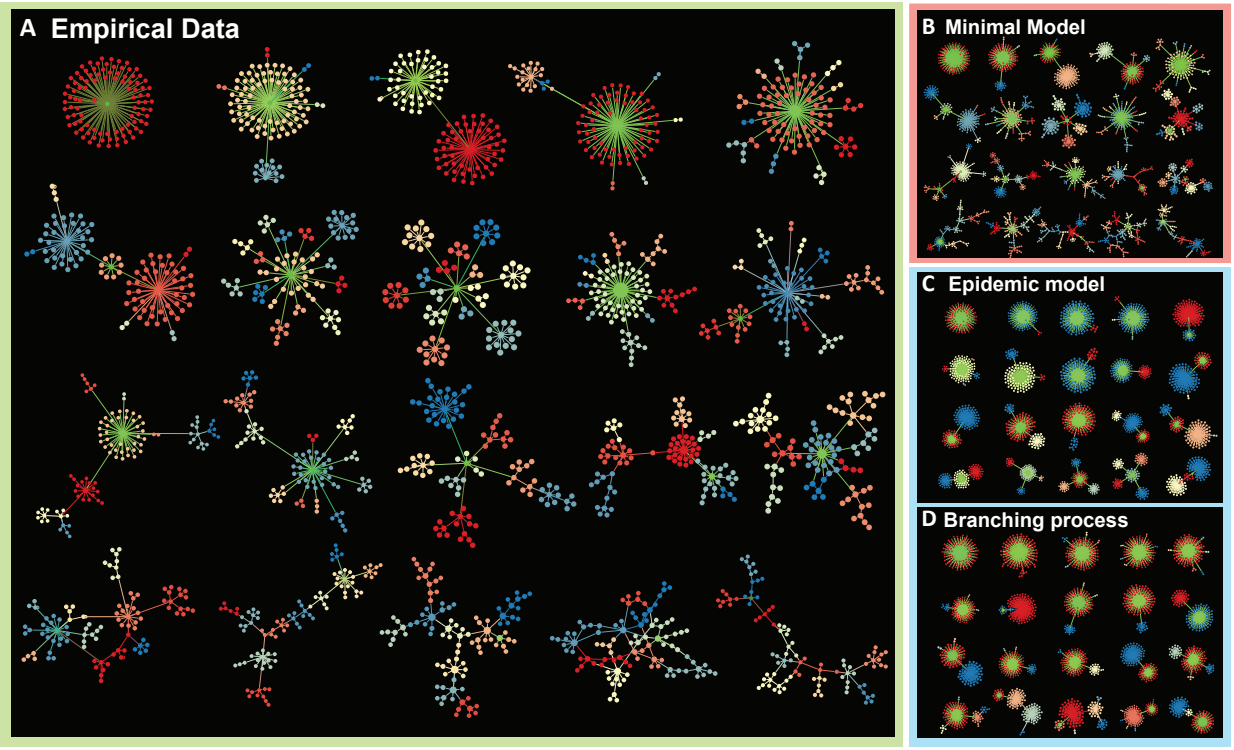


图 4.1　真实信息流传播数据和不同模型产生的模式。（A）二十个真实信息传播模式，每个大小为 100 ± 3，代表从中心节点（以绿色着色）开始的信息级联传播。节点和链接分别代表用户和转发。不同的颜色对应不同的社区组[103]。（B-D）分别由：（B）我们所提出的模型（C）流行病模型和（D）分支过程模型产生的具有相同大小的信息传播模式。

4.2　结果

4.2.1　量化信息流模式的几何特性

为了量化信息流模式的几何特性，我们使用三个代表性指标：

- **质量**（mass）n 刻画了参与一个信息传播的人数；

- **极性**（polarity）v 衡量了信息流在多大程度上是方向相关联的，这意味着不同方向的不同传播趋势，其由所有受感染个体之间的最短距离的方差刻画；
- **延伸性**（outreach）d 刻画了信息流可以传播的最远距离。

我们所提出的度量（n，v，d）提供了一个三维度量空间来刻画信息流传播的几何模式，其空间的不同位置对应于图 4.1A 中所示的真实信息流图案的不同几何形状。例如，观察到的具有窄且深结构的大的信息流传播模式对应于大的 n，v 和 d 值。实际上，一条完美的链式传播模式拥有极性值 $v \sim O(n^2)$ 和延伸性 $d \sim O(n)$，而完美的星形模式拥有 $v \sim O(1)$ 和 $d \sim O(1)$ 其和其质量 n 无关（参见支持材料 SI，S1：Polarity 度量）。图 4.2A-C 列举了由几何模式空间表示的所有典型几何形状。

我们测量了数据集中观察到的 4.32 亿个传播模式的 n，v 和 d，并绘制 n 与 v，n 与 d，d 与 v 的联合概率密度分布图，分别如图 4.2D-F 所示。我们发现大多数传播模式属于具有小 n，v 和 d 值的区域，表明绝大多数传播模式都有着星形模式。我们还观察到存在大质量模式（即 $n > 100$），这意味着扩散模式的质量分布的胖尾性质。此外，这些模式的极性 v 和延伸性 d 在很大范围内变化，揭示了现实世界中信息流几何模式的丰富性和复杂性。例如，图 4.2D-F 显示了中等大小（$n \sim 100$）模式的大 v 和 d 值，表明这些模式的几何形状最复杂，而最大的或最小的模式相对简单。我们还在图 4.2F 中发现极性和外展正相关，但这两个指标之间存在很大的方差。

4.2.2　现有模型

目前对信息流建模的框架主要分为两类：

- 传染病模型（Epidemic model）[90,104–106] 将信息流的传播视为个体之间的传染过程。简单传染病模型（Simple-contagion models）[83,87,90,107] 通常假设社交边上的独立的传播，而复杂传染模型（complex contagion models）[102,108,109] 考虑到易感染个体可能同时从多个受感染的邻居获得感染暴露的可能性；
- 分支随机过程（the branching process）假设每个人将消息转发给一组“后代”邻居，其中受感染后代大小是从预定的概率分布中随机采样的[92,93,110]。

为了证明现有模型与真实数据之间的差异，我们分别用分支过程模型[92,93]和两个传染病模型[86,107,111]，即 Susceptible-Infected-Susceptible (SIS) 模型和 Susceptible-Infected-Recovered (SIR) 模型对真实数据进行数值模拟。我们采用最大似然估计[110,111] 来从真实数据中学习估计出最可能的建模参数，之后生成 4.32 亿个信息级联传播，每个信息级联传播从真实数据观察到的相应原始发帖者开始模拟。图 4.1C 和 D 分别绘制了由流行病模型和分支过程模型产生的扩散模式。与

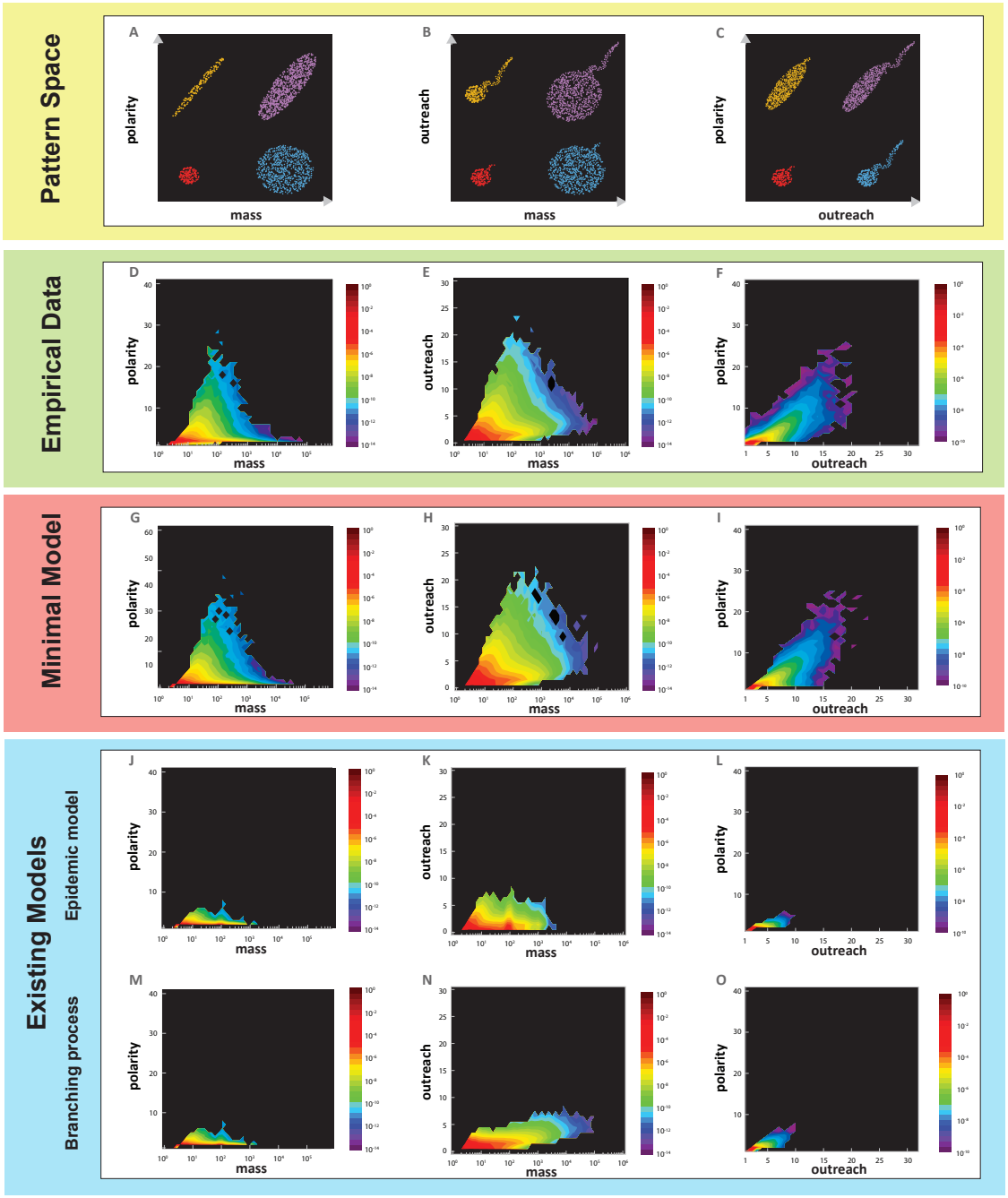


图 4.2　通过三维度量空间量化信息流传播模式的几何特性。真实信息流传播几何结构的（A）质量与极性，（B）质量与延伸性，和（C）延伸性与极性的二维联合分布。其中不同角落处的图案示出了对应度量值的典型几何形状。（D-F）热度图是真实 432, 101, 384 个信息级联传播数据所绘制的二维概率密度函数（对数变换的数值）。（G-I）是我们提出模型所产生的信息传播模式，（J-L）是现有 SIS 模型产生的信息流传播模式，（M-O）是现有分支过程模型产生的信息流传播模式。所有模型都产生与经验数据集相同的信息级联个数，其中建模参数通过实际数据的最大似然估计（更多细节参见 SI，S3：模型参数）。

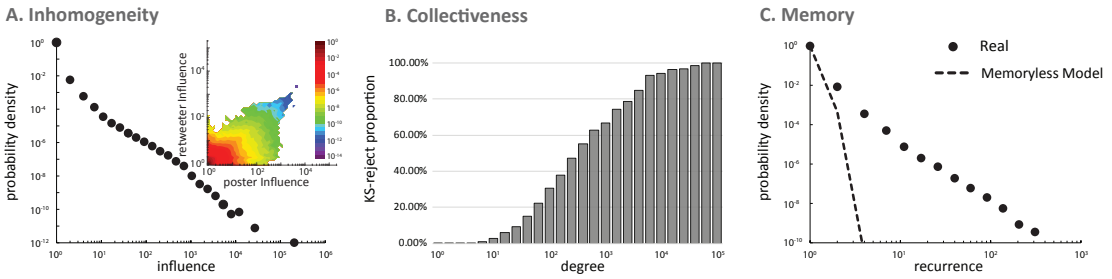


图 4.3　控制信息流传播的三个要素。（A）异质性。对数分布图上的影响分布，证明了整个用户群的异质性。插图描绘了对不同角色（即原始发布者与转发用户）的影响的热图，其中颜色代表概率密度值。（B）假设检验来测试信息传播的集体效应。我们的零假设是用户间传播没有集体效应，而柱状图显示了拥有不同粉丝的用户的零假设拒绝比例，我们采用双样本 Kolmogorov-Smirnov 检验设置了 5% 显着性水平。（C）重复出现次数 m 的分布，用于衡量用户在单个信息传播中发布的消息数量，显示具有幂律指数 $\gamma = 3.53 \pm 0.34$ 的胖尾分布，其中虚线绘制了没有记忆效应的零模型的结构，因此是窄尾的。

真实数据中展现的丰富的传播模式不同，现有模型只能产生简单的星形模式。为了进一步量化差异，我们测量了生成的信息传播模式的几何特性的质量，极性和延伸性的值。图 4.2J-O 绘制了这些三维度量的联合分布图。我们发现，与真实数据不同，两种模型的极性值都小于 10，且质量相关性很弱（图 4.2J 和 M），这意味着这些模型只能生成简单的模式。类似地，我们还观察到，对于两种模型，延伸性值都非常小，并且延伸性和质量之间的相关性非常弱，尽管分支过程可能存在一些轻微的相关性（图 4.2K 和 N）。然而，对于两种模型，极性和延伸性显示出强烈的正相关性（图 4.2L 和 O）与真实数据（图 4.2F）形成鲜明对比。

4.2.3　机制

现有模型刻画真实数据的失败意味着存在未知的信息流的模式形成机制，而这些模型不能刻画这些机制。实际上，我们发现控制信息流复杂模式形成有三个基本机制：

- **异质性**（Inhomogeneity）。在现实世界中，不同个体具有明显不同的感染他人的能力。例如，具有较多好友数的个体通常比拥有较少好友数的人能感染更多的邻居。为了量化个体传播影响力之间的差异，我们测量了每个用户 i 在所有其参与的信息级联传播中感染的邻居数量，即后代大小 b_i。用户 i 所参与的所有信息级联传播的平均后代大小 $\langle b_i \rangle$ 表征了单个 i 的传播影响力。图 4.3A 描绘了用户影响 $\langle b \rangle$ 遵循胖尾分布，证明个体具有高度不均匀的感染他人的能力。此外，个人在信息传播中扮演着不同的角色。例如，微博用户既可以作为启动级联的原始发帖者，也可以作为转发来自其他人的信息的转

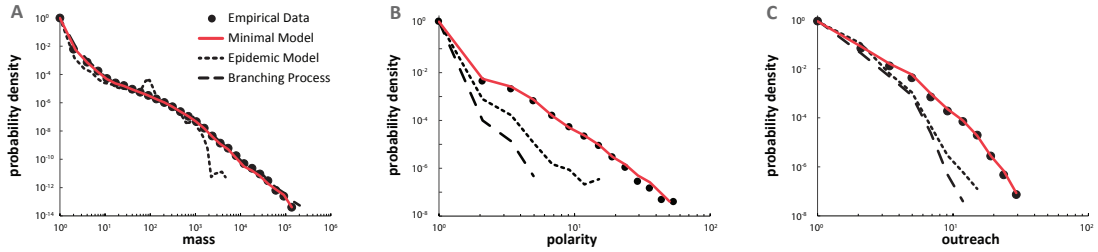


图 4.4　模型评价。三个度量（A）质量（mass），（B）极性（polarity）和（C）延伸性（outreach）的分布，分别由我们的模型模型（红色实线），SIS 模型（黑色虚线曲线）和分支过程（黑色虚线曲线）产生。而真实数据分布用（圆圈）标注。

发者。为了量化这一点，我们分别衡量每个人作为原始发布者的影响力 $\langle b\rangle_p$，和作为转发者的影响力 $\langle b\rangle_r$。图 4.3A 插图描绘了 $\langle b\rangle_p$ 与 $\langle b\rangle_r$ 的联合概率分布图，表明尽管这两个量之间存在正相关，但它们之间存在很大的差异。上述量化证据支持了信息传播中异质性的观测。

- **集体性**（Collectiveness）。现有的模型，如 SIS 模型和 SIR 模型，假设传播过程独立地发生在每对个体上。但是，我们观察到传播过程经常集体发生的证据，即一组用户可能同时被感染，导致长尾后代数（offspring size distribution）大小分布（参见 SI，图 S1）。相反，SIS 和 SIR 模型给出的后代大小遵循二项式分布。我们对每个真实个体传播后代大小分布 $p(b)$ 采用双样本 Kolmogorov-Smirnov 检验来假设检验零模型产生的 1 二项分布与真实数据之间的差异。图 4.3B 绘制了零假设的拒绝率，即 $p(b)$ 不遵循二项分布的概率随着用户好友数 k 的变化。我们发现拒绝率随着 k 增大而增加，表明具有较多好友的节点具有较强的集体效应。例如，超过 66% 好友数超过 1000 的用户拒绝了零假设，而对于好友数超过 10,000 的用户其拒绝率增加到 94% 以上。
- **记忆性**（Memory）。个体的传播行为在很大程度上取决于他/她的历史行为，导致信息流之间的长期时间相关性。为了捕获每个单独级别的记忆效应，我们测量每个个体出现在一个级联传播中的重复次数 m 的分布，并绘制图 4.3C 中所有信息级联传播中个体重复出现次数分布 $p(m)$。我们发现 $p(m)$ 遵循一个长尾的分布。相比之下，诸如 SIS 模型之类的无记忆模型显示出窄尾分布。例如，我们观察到一个用户在同一个信息级联传播中出现超过 300 次，而无记忆模型仅预测复发时间少于 3 次。

4.2.4　我们的模型

在这里，我们提出了一个包含上述三种观察机制的随机过程模型：

1. 从单个节点开始，作为原始发布者的 i 发布了一条信息，然后其后代节点因受该发布而可能被感染。
2. 从个体 i 的分布 p_i 中随机采样确定之后被感染的邻居数量 b，其中 p_i 是预先确定的建模原始发帖者 i 传播影响力的参数。
3. 从节点 i 的邻居中随机选择 b 个邻居节点。选择邻居 j 的概率与 w_{ijm} 成比例，其中 m 表示单个 j 的重现次数。我们假设选择概率可以被分解为乘积 $w_{i,j,m} = q_{i,j} \times \alpha_{j,m}$，其中建模参数 $q_{i,j}$ 捕获对 i 和 j 的感染率的异质性，而参数 $\alpha_{j,m}$ 捕获记忆效应。在每个步骤中有一组节点被感染，这自然地刻画了集体效应。
4. 对新感染的节点重复步骤 2 和 3，直到没有节点被进一步感染。请注意，我们使用 r_i 作为转发用户的后代大小分布，而不是使用表示初始发帖者的 p_i。

所有的参数 $(p_i, r_i, q_{i,j}, \alpha_{j,m})$ 是通过从真实数据的最大似然估计得到 (参考 SI, S3: 模型参数)。图 4.1B 描绘了由我们模型生成的信息传播模式，显示了与真实数据模式的很好的一致性，如图 4.1A 所示。

为了量化地比较我们的模型与真实数据，我们对真实社交网络上的信息传播进行了大量模拟（参见 SI，S4：基础社会网络），其中传播动态过程由我们所提出的模型确定。我们从如下三个度量测量如下信息传播模式，即质量，极性和延伸性。图 4.4A 分别绘制了真实数据，我们模型的结构，传染病模型和分支过程模型的质量分布，发现我们的模型和分支过程都很好地符合真实数据的质量分布，而传染病模型无法捕获分布的肥胖性质。图 4.4B 描绘了经验数据和所有模型的极性分布。我们的模型根据经验观察预测长尾分布。然而，先前的模型仅产生有限的极性值。类似地，图 4.4C 中所示的延伸性分布也表明我们的模型与真实数据之间的良好一致性，而现有模型不能产生大的延伸性模式。

此外，我们的模型不仅很好地刻画了这些指标的概率分布，还捕获它们的联合分布。图 4.2G-I 分别绘制了我们模型的质量与极性，质量与延伸性，以及延伸性与极性的联合密度分布。我们再次发现真实模式与我们提出的模型之间的完美契合。相反，虽然分支过程正确地预测了经验质量分布，但它无法捕捉质量与其他指标（极性和延伸性）之间的基本相关性。

4.3　讨论

综上所述，通过探索由 4.32 亿个信息级联传播组成的大规模真实信息流数据集，我们观察到复杂的在三维度量空间为特征的信息流模式，发现真实数据与传统传染病模型和分支模型的产生的结果存在系统的偏差。我们发现了三个机制，即异质性（inhomogeneity），集体性（collectiveness）和记忆效应（memory），它们控制着信息流的模式形成。最后，我们提出了一个包含这些成分的随机过程模型，可以重现现实世界中出现的复杂信息流模式。随着我们对信息流机制的理解随着越来越详细的数据的出现而加深，我们对三个基本成分的发现和所提出的模型为未来机制理解信息流模式形成提供了可能的基础。我们的模型可用于验证广泛的传播机制和现象，并可应用于营销，控制和推广方案。

4.4　数据集

我们从腾讯微博平台[①][8]，收集了信息流数据，该数据详细记录了每个人创建或转发帖子时每个人活动的全部信息。我们的社交网络由观察到的转发活动（SI，S4：基础社交网络）重建（跟随者 - 跟随者网络）。信息级联传播数据由从原始帖子开始并在关注者之间转发的信息传播来构建的。该数据覆盖 2012 年 6 月 20 日至 2012 年 6 月 26 日期间 7 天内的 563,331,392 条信息记录，101,802,707 个用户，并构成了 432,101,384 个信息级联传播。

4.5　支持材料

本小节为正文内容的补充支持材料（Supporting Information，SI）：

4.5.1　极性度量

极性（Polarity）v 测量信息流何种程度上是方向相关的，意味着沿着不同方向的不同传播趋势，其由一个信息级联传播中任意两个个体间的最短距离的方差刻画。完美的星形图案（star pattern）和完美的链条图案（chain pattern）在极性方面是两个极端。质量 n，S_n 的完美星型由一个中心节点和 $n-1$ 个卫星节点组成。任意两个节点之间的最短路径长度（l）分为两类：$\binom{n-1}{2}$ 个长度为 2 的卫星节点到卫星节点对，和 $\binom{n-1}{1}$ 个长度为 1 的卫星到中心节点对。

① t.qq.com

所有，一个星型图案的极性计算公式如下：

$$v(S_n) = \frac{(2-\mu)^2\binom{n-1}{2}+(1-\mu)^2\binom{n-1}{1}}{n-1} = 1 - \frac{2}{n} \tag{4-1}$$

其中 μ 是在无向图中任意两个节点距离的平均值，计算公式如下：

$$\mu(S_n) = \frac{2\binom{n-1}{2}+1\binom{n-1}{1}}{\binom{n}{2}} = 2 - \frac{2}{n}. \tag{4-2}$$

对于有 n 个节点的链型（chain），记为 C_n，其中有 $n-1$ 个距离为 1 的节点对，$n-2$ 个距离为 2 的节点对，...，1 个距离为 $n-1$ 的节点对。为了求得 $v(C_n)$，我们引入辅助函数标记：

$$A(n) = 1(n-1) + 2(n-2) + ... + (n-1)1 \tag{4-3}$$

$$B(n) = 1^2(n-1) + 2^2(n-2) + ... + (n-1)^2 1 \tag{4-4}$$

其中 $A(0) = B(0) = 0$ 和 $A(1) = B(1) = 1$。$A(n)$ 和 $B(n)$ 可以被写成递归的形式：

$$\begin{aligned} A(n) =& A(n-1) + (n-1) + ... + 1 \\ =& A(n-1) + \frac{(n-1)n}{2} \end{aligned} \tag{4-5}$$

和

$$\begin{aligned} B(n) =& B(n-1) + (n-1)^2 + ... + 1^2 \\ =& B(n-1) + \frac{(n-1)n(2n-1)}{6} \end{aligned} \tag{4-6}$$

通过求解以上两个等式，我们得到：

$$A(n) = \frac{n(n^2-1)}{6} \tag{4-7}$$

$$B(n) = \frac{n^2(n^2-1)}{12} \tag{4-8}$$

因此得到：

$$\mu(C_n) = \frac{A(n)}{\binom{n}{2}} = \frac{n+1}{3} \tag{4-9}$$

$$v(C_n) = \frac{B(n)}{\binom{n}{2}} - \mu(C_n)^2 = \frac{n^2 - n - 2}{18} \tag{4-10}$$

实际上，完美的链模式 C_n 有着极性 $v \sim O(n^2)$，而完美的星形模式 S_n 有着 $v \sim O(1)$，其独立于其质量 n。

4.5.2　后代大小分布

后代大小分布（Offspring size distribution）。

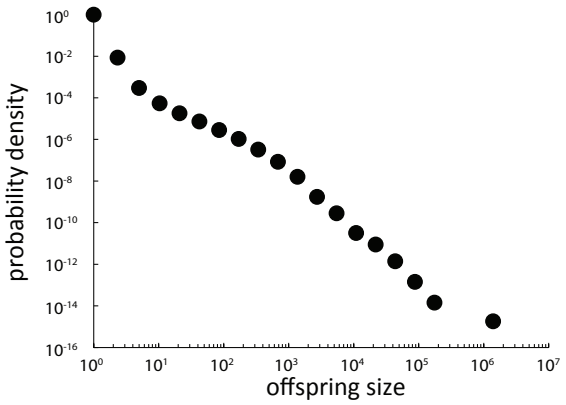


图 4.5　双对数坐标系下的后代大小分布。

我们测量每个用户的感染邻居的数量，即每个级联中的后代大小 b。通过汇总 4.32 亿个真实信息级联传播，我们得出后代大小分布 $p(b)$，如图4.5所示。后代大小分布的胖尾性质意味着存在大量用户被集体地感染。例如，虽然平均后代大小为 0.19，但可以一次性被集体地感染的用户可以超过 10^6。

4.5.3　模型参数估计

这里我们展示了估计我们模型参数的方法。我们首先估计用户 i 的后代大小分布 $p_i(b)$ 和 $r_i(b)$，其为当他/她分别作为原始发帖者和转发者时的情形。定义 $f_i^p(b)$（$f_i^r(b)$）是用户 i 作为原始发帖者（转发者）的信息级联的数量，在一个信息级联中其后代大小为 b，$p_i(b)$（$r_i(b)$）为当用户用户 i 作为原始发帖者（转发者）时感染 b 哥后代的概率。所有，对于作为原始发帖者的似然函数是:

$$L_i^p(b) = \prod_b p_i(b)^{f_i^p(b)}, \tag{4-11}$$

其对应的对数似然函数为：

$$\ln L_i^p(b) = \sum_b f_i^p(b) \ln p_i(b). \tag{4-12}$$

为了最大化受约束 $\sum_b p_i(b) = 1$ 的对数似然函数，我们使用拉格朗日系数：

$$\Lambda(b, i, \lambda) = \sum_b f_i^p(b) \ln p_i(b) + \lambda(\sum_b p_i(b) - 1). \tag{4-13}$$

我们要求

$$\frac{\partial \Lambda}{\partial p_i(b)} = \frac{f_i^p(b)}{p_i(b)} + \lambda = 0 \tag{4-14}$$

并且得到

$$\lambda = -\sum_b f_i^p(b) \tag{4-15a}$$

$$p_i(b) = -\frac{f_i^p(b)}{\lambda} \tag{4-15b}$$

最后，$p_i(b)$ 的最大似然估计（maximum likelihood estimation，MLE）值是当用户 i 作为原始发帖者时有着 b 个后代的比例：

$$\hat{p_i(b)} = \frac{f_i^p(b)}{\sum_b f_i^p(b)}. \tag{4-16}$$

此外，$r_i(b)$ 的推导过程和 $p_i(b)$ 类似。所有，$r_i(b)$ 的最大似然估计值为：

$$\hat{r_i(b)} = \frac{f_i^r(b)}{\sum_b f_i^r(b)}. \tag{4-17}$$

用户 i 和用户 j 之间的成对感染概率，由 $q_{i,j}$ 表示，由 $f_{i,j}$ 与 f_i 的比率得出，其中 $f_{i,j}$ 是用户 j 转发用户 i 的微博的总次数，f_i 是 i 发布或转发微博的次数：

$$q_{i,j} = \frac{f_{i,j}}{f_i} \tag{4-18}$$

记忆系数 $\alpha_{i,m}$ 正比于 $c_{i,m}$ 与 c_i 的比率，其中 c_i 是用户 i 参与的信息级联个数，而 $c_{i,m}$ 是用户 i m 次参与的信息级联数：

$$\alpha_{i,m} \propto \frac{c_{i,m}}{c_i} \tag{4-19}$$

4.5.4　社交网络构建

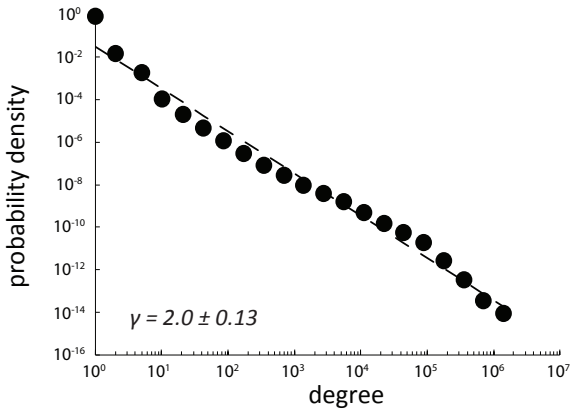


图 4.6　社交网络的度分布。对数 - 对数坐标系下，虚线具有斜率-2.0。

我们从信息流记录中重建底层社交网络。只有在两个人之间存在传播记录时，我们才会在两个人之间建立连接。例如，只有当有一条信息从用户 i 传播到用户 j 时，j 才会被添加到 i 的可能后代集合中。以这种方式，捕获任何两个个体之间的定向关系。总的来说，我们为参与我们的经验数据集的所有 101,802,707 个用户建立了后代集。为了表征网络的连通性，我们测量每个人的度数 k，即子孙的数量。图4.6绘制了底层社交网络的度分布图，显示虽然平均度数为 0.6，但非常大的节点（例如，$k > 10^7$）的存在，意味着度分布的长尾性质。

第 5 章　分布函数的动力学起源定理

许多科学研究都遵循如下模式：从对一个系统在特定时刻下的断面状态观察数据（cross-sectional states）来推断其动力学演化机制（dynamics）。但是，正式且系统地构建和学习出它们之间关系的研究却少之又少。在本文中，我们将复杂的截面状态观测数据视为由具有随机均匀输入信号的确定动态系统生成。我们构造了了概率分布函数与其动力学生成系统之间的一个等价关系，然后开发了一个框架来从截面状态分布，或截面样本数据中，推断其动力学生成机制。通过这样的框架，我们能够从各种分布函数中发现新的动力学生成机制，而且可以从各种动力学生成机制中发现新的概率分布函数。我们通过合成数据和真实数据验证了我们的框架。实验结果表明，我们的框架能够准确地发现和刻画各种数据分布的动力学生成过程。我们的研究有助于发现现实世界中复杂宏观截面现象的未知微观动力学演化机制。

5.1　引言

许多自然法则都有着微观的动力学起源，但是，大多数情况下我们只能在某个特定时间点观察到关于自然现象的宏观截面状态数据。从断面状态数据中推测其动力学演化机制是许多领域共同的研究目标，包括生物[112]，物理[35]，社会科学[113]和计算机科学[114]等等。例如，Barab á si 和 Albert[2] 从复杂网络节点链接数（度）的幂律度分布推断出无标度网络演化的动力学机制。Yoshida 等人[115] 从物种的横截面丰度数据推断出生态系统的动力学演化过程。Sinatra 等人[116] 和 Wang 等人[117] 从科学家论文的引文分布推断出控制学者科学影响的动力学演化机制。Oliveira 和 Barabási[36] 从达尔文和爱因斯坦的通信事件间的时间分布（或从在线合作[118] 事件间时间分布）推断出人类行为的动态决策过程。Pierson 等人[119] 从病人的横断面记录中推断出疾病随时间发展过程。在这些情况下，我们通常无法跟踪数百万年复杂系统（例如，生态系统）的演化轨迹。在某些情况下，随时间获取纵向演化数据也极其困难（例如，在医疗保健中[119]）。因此，如何从对一个复杂系统截面状态观察中推断其随时间演化的动力学过程，是一个至关重要却极具挑战性的问题。

关于理解观测数据背后的复杂系统的动力学过程的研究大多是个案分析的。据我们所知，没有工作试图推导出一般理论框架来揭示各种宏观数据分布与微观动态系统之间的内在联系。这种关系可以进一步用于直接从横截面数据集学习纵向动力学生成机制。该研究可以促进各种从现实世界中复杂横截面数据中发现其

未知动力学演化过程的科学发现，理解复杂系统演化运行的机制。

我们的目标是填补上述研究空白。具体来说，我们在理论上证明了截面状态分布的宏观统计特性（例如，重尾，窄尾，幂律或 S 形等），是由动态系统 $\frac{dx_i(t)}{dt}|_{x_0} = \frac{dF^{-1}(1-\frac{t_i}{t})}{dt} = \frac{d\Lambda^{-1}(\ln(\frac{S(x_0)}{t_i}t))}{dt}$（参见定理5.1和推论 5.1）产生的。通过我们的定理的构造，我们证明了概率分布函数与它们的动力学生成系统之间的一个等价关系，并开发了一套新工具，从观察到的数据样本的横截面分布来推断出它们可能的纵向动力学生成过程。

为了证明我们框架的能力，如表5.1 所示，我们推导出多个典型分布函数的动力学生成机制，这些分布函数包括从窄尾分布（narrow-tailed distributions）到重尾分布（heavy-tailed）。之后，我们示例了从一些典型动态系统中推导出其横截面状态分布，如表5.2。这些动态系统由诸如偏好依附（preferential attachment），生长竞争（growth competition），环境限制（environment limit）等可解释的机制组成。许多我们推导出的分布函数和动态系统都是全新的，没有在文献中进行研究过。此外，我们展示了一个包含典型动力学生成机制的模型，它可以直接从数据样本中学习复杂系统的潜在动力学生成过程。

我们在合成数据和真实数据集上验证我们的框架。对于合成数据集，我们的框架准确地发现了各种真实数据分布，包括窄尾（narrow-tailed），重尾（heavy-tailed）和混合（mixture）分布，及其动力学生成过程。然后，我们将我们的框架应用于各种现实世界的数据集，用来推断它们的动力学生成机制。即使现实世界数据集的分布表现出更多的复杂性，我们的框架也能够准确地刻画和再现真实世界的数据集。这意味着我们的方法可以给出观察到的数据的一种合理的动力学生成过程。我们归纳贡献如下：

- 我们将复杂的宏观数据分布视为在某时刻观察一个动态系统的微观截面状态获得。通过证明定理5.1 和推论5.1，我们连接概率分布函数，生存分析中的危险函数（hazard functions）及其动态生成系统，并给出参数学习和数据模拟算法（第5.2 小节）。
- 我们发现了几种全新的概率分布函数，危险函数及其相应的可以解释的动态生成系统（第5.3 小节和表5.1）。
- 我们提出了一个模型直接从宏观截面数据样本中学习动态系统。(第5.4小节)。

5.2　定理

符号。我们的符号来自概率论（probability theory），生存分析（survival analysis），点过程（point process）和动态系统（dynamic systems ）：

- **概率论**：定义 X 为根据累计分布函数 $F(X \leq x) = \int_{x_0}^{x} f(s)ds = 1 - S(X > x)$ 产生的随机变量，我们观测到 n 个样本 x_1，...，x_n。我们假设 $F(X)$ 是绝对连续的。进一步，$f(x)$ 和 $S(x) = 1 - F(x)$ 分别是 X 的概率密度函数（probability density function，pdf）和生存函数（survival function）（或称作 complementary cumulative distribution function，ccdf）。
- **存活分析**：X 的危险函数 $\lambda(x)$ 定义为：

$$\lambda(x) = \lim_{\Delta x \to 0^+} \frac{Pr(x \leq X < x + \Delta x | X \geq x)}{\Delta x} = \frac{f(x)}{S(x)} \tag{5-1}$$

  解释为在条件 $X \geq x$ 下 X 的采样值是 x^+ 的条件概率密度函数。我们定义 $\Lambda(x) = \int_{x_0}^{x} \lambda(s)ds$ 为累计危险函数。其中，x_{x_0} 是随机变量 X 所定义的最小值。由于 $\Lambda(x)$ 是单调递增的，因此 $\Lambda(x)$ 存在反函数 $\Lambda^{-1} : \mathscr{R}^+ \to \mathscr{R}$，$\Lambda^{-1}(\Lambda(x)) = x$。类似的，我们定义 $F^{-1} : \mathscr{R}^+ \to \mathscr{R}$, $F^{-1}(F(x)) = x$。
- **点过程**：我们用符号 $\mathscr{P}(t|\lambda_p) = \{t_1, ..., t_i, ...|0 < t_1 \leq ... \leq t_i \leq .. \leq t\}$ 来表示一个到 t 时刻的泊松过程（Poisson point process），其中每个事件点时间为 t_i，泊松过程的强度函数为 $\lambda_p > 0$，并且其等价于计数过程 (counting process) $N(t|\lambda_p) = \sum_{i \geq 1} \mathbb{1}_{(0,t]}(t_i)$。
- **动态系统**：我们定义在时刻 t 下的动态系统为：$\mathscr{D}(t) = \{x_i(t)|\frac{dx_i(t)}{dt}; x_i(t_i); i = 1, 2, ...\}$，其中 $x_i(t)$ 是第 i^{th} 个体在 t_i 时刻加入系统时的状态，并且其状态的变化由动态方程 $\frac{dx_i(t)}{dt}$ 和初态 $x_i(t_i)$ 决定。

我们的定理构造了任意分布函数 $F(x)$ 的动力学生成过程如下（证明请参阅第5.7.1小节）：

定理 5.1：给定一个动态系统 $\mathscr{D}(t) = \{x_i(t)|\frac{dx_i(t)}{dt}; x_i(t_i) = x_0; i = 1, 2, ...\}$，其中每个个体按照泊松过程 $\mathscr{P}(t|\lambda_p) = \{t_1, ..., t_i, ...|0 < t_1 \leq ... \leq t_i \leq ...\}$ 来到系统，而且个体 i 的状态根据动态方程 $\frac{dx_i(t)}{dt}|_{x_0}$ 改变，那么在时刻 t 观察该系统每个个体的状态值，即截面（cross-sectional）状态 $x(t) = \{x_1(t), ..., x_i(t), ...\}$，那么，$x(t)$ 遵循分布 $F(x(t))$ 当且仅当 $\frac{dx_i(t)}{dt}|_{x_0} = \frac{dF^{-1}(1-\frac{t_i}{t})}{dt}$。

推论 5.1：在定理5.1相同条件下，动态系统 $\mathscr{D}(t)$ 在 t 时刻的截面状态 $x(t) = \{x_1(t), ..., x_i(t), ...\}$ 遵循分布 $F(x(t))$ 当且仅当 $\frac{dx_i(t)}{dt}|_{x_0} = \frac{d\Lambda^{-1}(\ln(\frac{S(x_0)}{\tau_i}t))}{dt}$。

通常，存活函数在初始状态 x_0 时 $S(x_0) = 1$，那么我们得到动力学方程 $\frac{dx_i(t)}{dt}|_{x_0} = \frac{d\Lambda^{-1}(\ln \frac{t}{t_i})}{dt}$。

根据如上定理和引理，$F(x)$ 的统计特性，比如长尾，短尾，幂律，S 型等等，都由动力学增长过程 $\frac{dx_i(t)}{dt}|_{x_0} = \frac{dF^{-1}(1-\frac{t_i}{t})}{dt} = \frac{d\Lambda^{-1}(\ln(\frac{S(x_0)}{t_i}t))}{dt}$ 产生，而且每个个体的初态都为相同的 x_0。我们之后通过一个网络演化系统来解释如上动力学过程。

参数估计 通过定理5.1和引理5.1描述的动力学系统 $\frac{dx_i(t)}{dt}$ 和截面分布 $f(x)$ 的等价性，我们可以通过任一个方面的观测数据来估计它们相同的参数。通常情况下，我们往往只得到截面数据 $f(x)$。因此，我们在此仅展示如何从截面数据来估计参数的过程。从 $\frac{dx_i(t)}{dt}$ 数据来学习参数的过程请参考支持材料第5.7.2小节。

给定一组截面静态数据 $\{x_1, ..., x_{n-1}, x_n\}_t$，我们可以估计分布 $f(x|\Theta)$ 或 $\lambda(x|\Theta)$ 的参数，我们采用最大似然估计的方法：

$$\begin{aligned}\max_{\Theta} \ln L(\Theta|x_1, ..., x_n) &= \ln \prod_{i=1}^{n} f(x_i) \\ &= \ln \prod_{i=1}^{n} \lambda(x_i) e^{-\Lambda(x_i)} = \sum_{i=1}^{n} \ln \lambda(x_i) - \sum_{i=1}^{n} \Lambda(x_i)\end{aligned} \tag{5-2}$$

根据参数化 $f(x|\Theta)$ 和如下关系 $f(x|\Theta) = \lambda(x|\Theta)e^{-\Lambda(x|\Theta)}$ 得到。而估计得到的动力学方程为 $\frac{d\hat{x}_i(t|\hat{\Theta})}{dt}|_{x_0} = \frac{d\Lambda^{-1}(\ln(\frac{S(x_0|\hat{\Theta})}{t_i}t)|\hat{\Theta})}{dt}$. 我们发现有时候通过运用引理5.1中的生存函数和危险函数会更简洁，详细在第5.4 小节。

数据模拟器。 我们可以通过动态系统或是分布函数模拟生成样本数据。如果通过 $\frac{dx_i(t)}{dt}|_{x_0}$ 产生截面状态分布数据 $x_i(t)$ $i = 1, 2, ...$ ，我们可以直接运用定理5.1，即在 t 时刻观测每个个体的状态 $x_i(t) = \int_{t_0}^{t} \frac{dx_i(\tau)}{d\tau} d\tau$ 。

另一方面，如果从分布方生成数据样本，尤其是累积危险率，该方法基于求解 $\ln(u) + \Lambda(x) = \ln S(x_0)$ 方程来得到 x。其中 u 是从均匀分布 $U0, 1]$ 生成的。我们可以应用最通用的方法，比如 Newton 的迭代方法来解决上面的方程式。当我们有 $F(x)$，$S(x)$ 或 $\Lambda(x)$，我们可以通过运用支持材料第5.7.1小节中的引理5.2或引理5.3来取随机样本。

总体框架如上所示。我们在第5.4 小节给出了具体的例子来阐述如上框架。之后我们通过一个随机网络系统来进一步解释。

5.3　发现新动力学系统和分布函数

通过应用我们的定理5.1和推论5.1，我们可以从（i）分布函数中发现新的动力学系统，（ii）从给定动态系统中发现新的（截面状态）分布函数，（iii）归纳总结动态系统的可解释模式，即“机制”，以及（iv）有原则的设计具有所需动力学特性的新分布函数。

总之，我们试图找到宏观横截面状态分布函数，生存分析中的危险函数，及其微观动力学生成机制之间的内在联系。我们在表5.1中展示了我们的结果。

5.3.1　新动力学系统的发现

通过运用定理5.1和推理5.1 ，我们从典型的分布函数中发现了新的动力学系统，如表5.1所示。

- **幂律**（Power-law ）**和指数**（Exponential）**分布**。幂律分布 $f(x) = \alpha x_0^{\alpha} x^{-(\alpha+1)}$ 其中 $x \geq x_0$ 以其无尺度的缩放属性和动力学生成机制而闻名。它的危险函数是一个更简单的形式 $\frac{\alpha}{x}$。通过使用推论5.1，我们得到其动力学增长曲线为 $x_i(t) = x_0(\frac{t}{t_i})^{\frac{1}{\alpha}}$ 和动力学微分方程为 $\frac{dx_i(t)}{dt} = \frac{x_i(t)}{\alpha t}$（显示在表5.1中）。至于复杂网络（统计物理学）的研究，$x_i(t)$ 表示节点 i 的邻接节点数，即度数，$\frac{dx_i(t)}{dt} \propto x_i(t)$ 表示网络演化的偏好附加（preferential attachment）机制。αt 表示由于新节点到达而导致的增长竞争。在经济学中，$x_i(t)$ 表示财富，而 $\frac{dx_i(t)}{dt} \propto x_i(t)$ 表示富者更富现象（或马太效应）。由于新的竞争者涌入市场，αt 代表了竞争的增长。类似的动态思想适用于 Dirichlet 过程。相反，指数分布 $f(x) = \alpha e^{-\alpha x}$ 具有动力学生成过程 $\frac{dx_i(t)}{dt} = \frac{1}{\alpha t}$，没有偏好附加项 $x_i(t)$ 和由于增长竞争而线性衰减的增长率，这对应于随机图的生成机制。
- **拉伸指数**（Stretched exponential）**或威布尔**（Weibull）**分布**。对于拉伸指数 $f(x) = \frac{\alpha}{x^{\theta}} e^{-\frac{\alpha}{1-\theta}(x^{1-\theta})}$ 和威布尔 $f(x) = \alpha\lambda^{\alpha} x^{\alpha-1} e^{-(\lambda x)^{\alpha}}$ 分布，它们有着相同的动力学过程，即 $\frac{dx_i(t)}{dt} = \frac{x_i^{\theta}(t)}{\alpha t}$ 和 $\frac{dx_i(t)}{dt} = \frac{x_i^{1-\alpha}(t)}{\lambda^{\alpha}\alpha t}$。与幂律分布相比，它们具有非线性偏好附加动力学增长过程，例如由幂指数 θ 调整的 $x_i^{\theta}(t)$ 形式。
- Sigmoid **和** Log-logistic **分布**。 Sigmoid 分布 $f(x) = \frac{e^x}{(1+e^x)^2}$ 在深度学习中被广泛用作激活函数。我们的定理发现其动力学生成过程为 $\frac{dx_i(t)}{dt} = \frac{1}{t-t_i}$，在其初始时刻 t_i 时有一个瞬时爆发（速率）增长，然后增长过程遵循 $\ln(\frac{t}{t_i} - 1)$ ，其类似于指数函数 $\ln(\frac{t}{t_i})/\alpha$ 情况。至于 log-logistic 分布 $f(x) = \frac{\lambda\alpha(\lambda x)^{\alpha-1}}{[1+(\lambda x)^{\alpha}]^2}$，它的动力学生成过程为 $\frac{dx_i(t)}{dt} = \frac{x_i(t)}{\alpha(t-t_i)}$，它在出生时 t_i 也具瞬时爆发（速率）增长，但由于线性偏好附加项 $x_i(t)$ 而更快的增长改变状态。
- **对数正态**（Log-normal）**和正态**（Normal）**分布**。通过我们的定理，对数正态分布函数 $f(x) = \frac{1}{x\sqrt{2\pi}} e^{-\frac{(\ln x)^2}{2}}$ 和正态分布 $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ 分别具有动力学生成过程 $\frac{dx_i(t)}{dt} = x_i(t)\frac{d\Phi^{-1}(z)}{dz}\frac{t_i}{t^2}$ 和 $\frac{dx_i(t)}{dt} = \frac{d\Phi^{-1}(z)}{dz}\frac{t_i}{t^2}$，其中 $z = 1 - \frac{t_i}{t}$。它们都随着增长竞争的平方而衰退，但对数正态分布具有线性偏好附加项，表明有着长尾分布属性。

- **均匀分布**（Uniform distribution）。均匀分布 $f(x) = \frac{1}{b-a}$ 对应于动力学生成过程 $x_i(t) = b - (b-a)^{\frac{t_i}{t}}$，如后面所讨论的那样，它遵循受限限的指数增长（constrained exponential growth）。其微分方程为：$\frac{dx_i(t)}{dt} = \frac{b-x_i(t)}{t}$，表示具有环境限制项 $b - x_i(t)$ 和增长竞争项 $\frac{1}{t}$。

5.3.2 新（截面状态）分布函数的发现

接下来，我们推导出一些典型动态系统横截面状态的的分布函数（参见表5.2）：

- **指数**（Exponential）**增长**。由于非常快速的增长率 $\frac{dx_i(t)}{dt} = \frac{x_i(t)}{\alpha}$，均匀的输入信号通过指数增长系统的变换可以产生差异非常大的输出信号，表现出对数幂律分布（或称作 log-Cauchy 分布）$f(x(t_i)) = \frac{\frac{\alpha}{t_i}}{x(\frac{\alpha}{t_i}\ln\frac{x}{x_0}+1)^2}$。$f(x)$ 中的 t_i 意味着 $f(x)$ 随时间变化（不稳定）。指数增长可以通过分支过程（branching process）来解释，其中 $\frac{1}{\alpha}$ 是平均分支因子（average branching factor），完全树的增长以及早期细胞的分裂可以产生该动力学增长系统。
- **幂律**（Power-Law）**和拉伸指数**（Stretched-Exponential ）**增长**。幂律动力学增长系统 $\frac{dx_i(t)}{dt} = \frac{x_i(t)}{\alpha t}$ 和拉伸指数动力学增长系统 $\frac{dx_i(t)}{dt} = \frac{x_i(t)}{\alpha t^\theta}$ [120]，具有偏好附加项 $x_i(t)$ 和线性或非线性增长竞赛项。幂律分布 $f(x) = \alpha x_0^\alpha x^{-(\alpha+1)}$ 由幂律增长动力学产生，而对数幂律分布 $f(x) = \frac{\frac{\alpha}{t_i^{1-\theta}}}{x[\frac{\alpha(1-\theta)}{t_i^{1-\theta}}\ln\frac{x}{x_0}+1]^{\frac{2-\theta}{1-\theta}}}$ 由可调节指数 θ 的拉伸指数动力学系统产生。
- Sigmoid，Log-Logistic **和** Stretched Log-Logistic **增长**。遵循 logistic 增长框架（偏好附加项和环境限制项相乘：$x_i(t) * [N - x_i(t)]$），但他们具有不同的增长竞争，Sigmoid 增长为 $\frac{dx_i(t)}{dt} = \frac{x_i(t)[N-x_i(t)]}{\alpha}$，log-logistic 增长为 $\frac{dx_i(t)}{dt} = \frac{x_i(t)[N-x_i(t)]}{\alpha t}$，以及拉伸对数增长 $\frac{dx_i(t)}{dt} = \frac{x_i(t)[N-x_i(t)]}{\alpha t^\theta}$，生成复杂的 logistic 形式分布，如表5.2所示。以 log-logistic 增长 $\frac{dx_i(t)}{dt} = \frac{x_i(t)[N-x_i(t)]}{\alpha t}$ 为例，我们可以解释它作为一个不断增长的无标度网络模型（scale-free network model），但是我们限制中心节点（hubs）的度数，导致更多有着适度的好友数的中心节点产生，而不是只产生少数寡头中心节点。
- **受限指数**（Confined Exponential），**受限幂律**（Confined Power-Law）**和受限拉伸指数**（Confined Stretched-Exponential）**增长**。遵循限制增长框架（环境限制项 $N - x_i(t)$ 但增长竞争率不同，受限制指数增长 $\frac{dx_i(t)}{dt} = \frac{N-x_i(t)}{\alpha}$，受限制幂律增长 $\frac{dx_i(t)}{dt} = \frac{N-x_i(t)}{\alpha t}$，以及受限拉伸指数增长 $\frac{dx_i(t)}{dt} = \frac{N-x_i(t)}{\alpha t^\theta}$，生成复杂的约束形式分布，如表5.2所示。均匀分布是由有限幂律增长产生的特殊情况。

5.3.3　常见动力学生成机制

通过我们的定理，我们进一步总结了横截面状态分布的常见生成模式。例如，动力学系统增长越快，其截面状态分布越不均匀。我们发现偏好附加项 $x_i(t)$ 是产生长尾分布的常见成分，如幂律，拉伸指数（或 Weibull），对数正态，对数逻辑，对数幂律等如表 5.1所示。相比之下，如动力学系统 $\frac{dx_i(t)}{dt} = \frac{d\Phi^{-1}(z)}{dz}\frac{t_i}{t^2}$ 中的平方竞争所示，产生正态分布；而动力学系统 $\frac{dx_i(t)}{dt} = \frac{1}{\alpha t}$ 中的线性竞争产生指数分布，其动力学增长竞争越快，其截面状态分布尾部越窄，因此分布方差越小。

另一方面，截面状态分布函数可以通过它们的动力学生成机制来设计。例如，延时增长竞争（since-then-growth-competition ）$\frac{1}{t-t_i}$ 项可用于设计具有瞬时突发的激活函数，环境限制和增长竞争机制可用于限制截面状态的方差。其他成分如偏好附加用于生成长尾分布，非线性幂指数用于增加模型灵活性，上述机制的组合可用于构建具有所需动力学特性的全新的且复杂的分布函数。

5.4　从静态截面数据学习动力学系统

在上一节中，我们将展示如何从分布函数推断其动力学生成过程，以及如何从动力学生成过程中获得截面状态分布。在本节中，通过应用我们的定理，我们直接从截面状态数据样本中学习动力学过程，以及拟合分布和生成样本。我么展示了一个简单但通用的模型，并可以以类似的方式设计更复杂的参数化模型。

通过应用我们的定理5.1和推论5.1，我们展示了一个参数模型来拟合复杂的经验分布，并推断其动力学增长过程。我们给出参数化危险函数（hazard function）为：

$$\lambda(x) = \beta + \frac{\alpha}{(x+\Delta)^{\theta}} \tag{5-3}$$

其中 $X > -\Delta$。其相应的概率密度函数可由关系 $f(x) = \lambda(x)e^{-\int_{x_0}^{x}\lambda(s)ds}$ 导出。当 $\theta \neq 1$ 时，我们得到 pdf：

$$\begin{aligned} f(x) = {} & \beta e^{-\beta x - \frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]} \\ & + \alpha(x+\Delta)^{-\theta} e^{-\beta x - \frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]} \end{aligned} \tag{5-4}$$

当 $\theta = 1$ 时，我们得到 pdf 为：

$$f(x) = \beta e^{-\beta x}(\frac{x}{\Delta} + 1)^{-\alpha} + \frac{\alpha}{\Delta}(\frac{x}{\Delta} + 1)^{-(\alpha+1)} e^{-\beta x} \tag{5-5}$$

公式5-3 是一个长尾混合模型（*heavy-tailed mixture model*）。通过运用我们的定理5.1，其中 $x_0 = 0$ 和 $S(x_0) = 1$，我们得到：

$$\frac{dx_i(t)}{dt}|_{x_0} = \frac{(x_i(t) + \Delta)^\theta}{\beta(x_i(t) + \Delta)^\theta t + \alpha t} \tag{5-6}$$

是我们设计的混合长尾分布函数的动力学生成过程：

- 当 $\theta \neq 1$ 和 $\beta = 0$ 时，$f(x)$ 蜕化为拉伸指数（stretched exponential）或威布尔分布 (or Weibull)： $f(x|\beta = 0) = \alpha(x + \Delta)^{-\theta} e^{\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]}$ [①]，我们得到特例情况下的动力学生成过程为：

$$\frac{dx_i(t)}{dt}|_{x_0} = \frac{(x_i(t) + \Delta)^\theta}{\alpha t} \tag{5-7}$$

- 当 $\theta = 1$ 和 $\beta = 0$ 时，我们得到产生幂律分布 $f(x) = \frac{\alpha}{\Delta}(\frac{x}{\Delta} + 1)^{-(\alpha+1)}$ 的动力学过程：

$$\frac{dx_i(t)}{dt}|_{x_0} = \frac{x_i(t) + \Delta}{\alpha t} \tag{5-8}$$

- 当 $\alpha = 0$ 时，我们得到产生指数分布 $f(x|\alpha = 0) = \beta e^{-\beta x}$ 的动力学过程：

$$\frac{dx_i(t)}{dt}|_{x_0} = \frac{1}{\beta t} \tag{5-9}$$

为简洁起见，我们在支持材料5.7.5小节中的网络演化场景中进一步说明和解释了上述想法。

对于参数学习方法，当给出危险函数5-3时，我们得到对数似然函数：

$$\begin{aligned} \ln L(x_1, ..., x_n) = &\sum_{i=1}^{n} \ln\left[\beta + \alpha(x_i + \Delta)^{-\theta}\right] - \beta \sum_{i=1}^{n} x_i \\ &- \frac{\alpha}{1-\theta} \sum_{i=1}^{n} \left[(x_i + \Delta)^{1-\theta} - \Delta^{1-\theta}\right] \end{aligned} \tag{5-10}$$

为简便起见，请参阅支持材料第5.7.3小节以获取最优参数，包括目标函数梯度，优化算法等的详细步骤。

基于上述危险函数和微分方程生成截面状态样本模拟生成器如下：

① 当 $\theta = -1$ 时，其包括类似 Gaussian 分布为特例

Input : Hazard function of $\lambda(x) = \beta + \frac{\alpha}{x+\Delta}$, total event number n

Output: $\{x_1, ..., x_n\}$

Set current iteration $i = 1$;

while $i \leq n$ **do**

  Sample $u \sim U(0, 1]$;

  Solve $\log u + \int_{x_0}^{x} \lambda(s)ds = 0$ for x by Algorithm 7.;

  $x_i = x$;

  i += 1 ;

end

Algorithm 3: 从危险函数 Eq. 5-3中产生样本

Input : Dynamic equation Eq. 5-3, Poisson process

$$\mathscr{P}(t|\lambda_p) = \{t_1, ..., t_i, ...|t_1 \leq ... \leq t_i \leq .. \leq t\} \text{ with } N(t|\lambda_t) = n,$$

Output: $\{x_1(t), ..., x_n(t)\}$

Set current iteration $i = 1$;

while $i \leq n$ **do**

  if $\theta == 1$ **then**

    Solve $\beta x_i(t) + \alpha \ln(\frac{x_i(t)}{\Delta} + 1) = \ln \frac{t}{t_i}$ for $x_i(t)$ by Algorithm 7.;

  else

    Solve $\beta x_i(t) + \frac{\alpha[(x_i(t)+\Delta)^{1-\theta} - \Delta^{1-\theta}]}{1-\theta} = \ln \frac{t}{t_i}$ for $x_i(t)$ by Algorithm 7. (解方程参阅支持材料第5.7.4小节);

  end

  i += 1 ;

end

Algorithm 4: 通过动力学方程 Eq. 5-6产生样本

5.5　实验

我们通过在合成数据集和真实数据集上回答以下问题，来验证定理5.1和推论5.1以及参数学习和模拟算法：

- 动态系统能否生成由我们的定理预测的横截面状态分布?
- 我们可以学习出生成的分布的参数吗?
- 我们可以从截面状态反推出其动态生成系统嘛?
- 各种真实数据集的可能的动力学生成机制是什么?

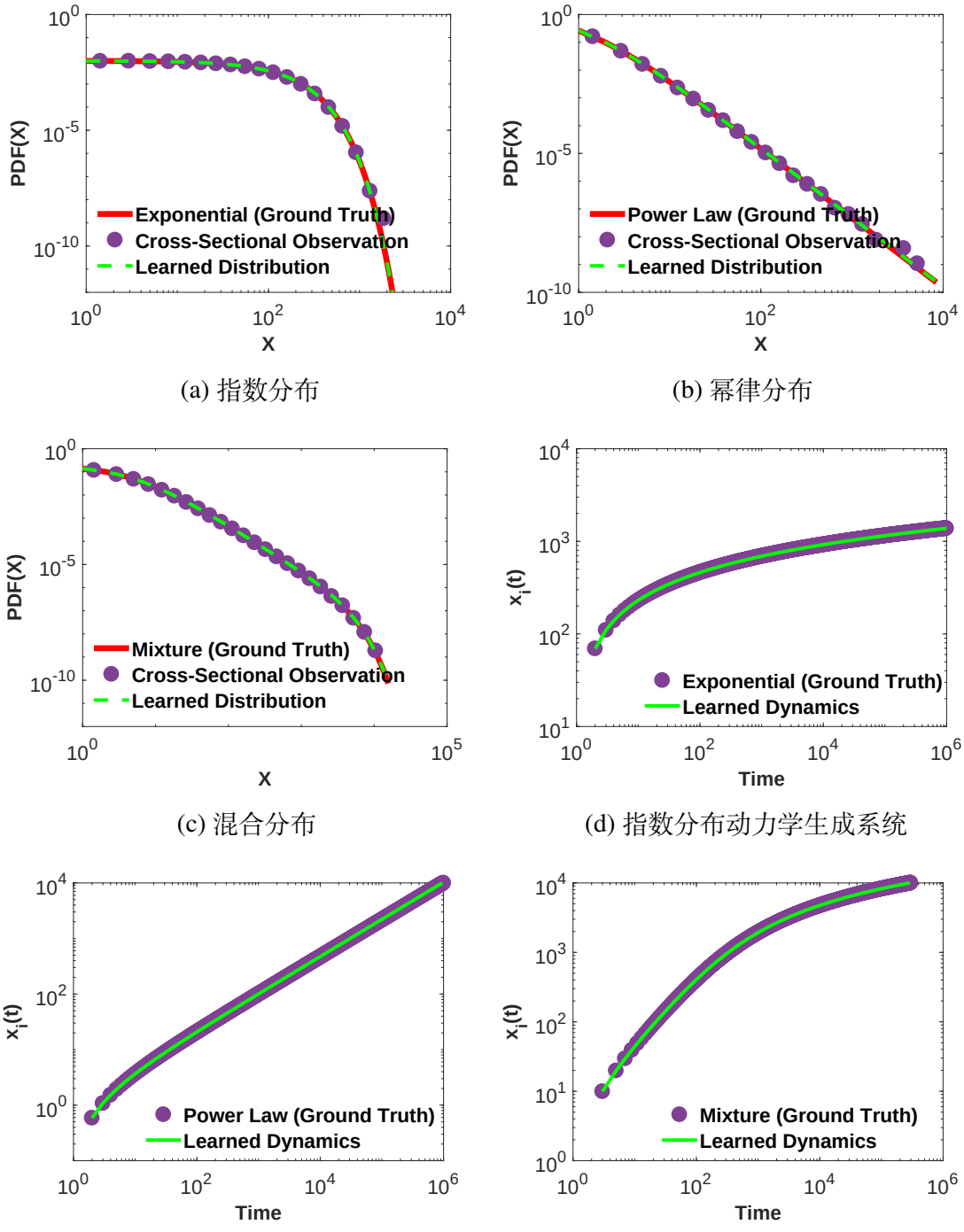
5.5.1　模拟数据实验

实验设置。 表5.3列出了由三个微分方程刻画的实验设置，用来产生具有代表性的动力学生成过程，包括指数分布（窄尾），幂律分布（重尾）和重尾混合分布。对于每个动态系统，我们在 $(0, 10^6]$ 的时间间隔内设置 $E[N(t|\lambda_t)] = 10^6$ 个个体，即 $\mathscr{P}(t|\lambda_p) = \{t_1, ..., t_i, ...|0 < t_1 \leq ... \leq t_i \leq .. \leq t\}$ 其中 $t = 10^6$ 和 $\lambda_t = 1$。其中 10^6 个个体根据动态方程 $\frac{dx_i(t)}{dt}|_{x_0=0}$ 更改其状态，如表5.3所示。我们观察他们在 $t = 10^6$ 时的截面状态。

结果。 动力学系统确实产生了我们定理5.1预测的截面状态分布。如图5.1 a-c所示，动力学系统生成的截面状态分布（紫色点），很好的和定理预测的分布重合（红线），包括窄尾分布（图5.1a），长尾分布（图5.1b），和长尾混合分布（图5.1c）。为了进一步量化分析观测到的截面状态分布和定理预测分布是同一个分布，我们通过 1% 显着性水平的双样本 Kolmogorov-Smirnov 检验方法[121]。我们在所设置的 1% 显着性水平下接受两个分布相同的假设，其 p 值（KS 距离）对于三种场景分布为：0.99（$4.7 * 10^{-4}$），0.51（$1.2 * 10^{-3}$）和 0.78（$6.9 * 10^{-4}$）。

我们的参数学习框架可以从截面数据样本中估计出建模参数。通过最大化对数似然函数6-5和6-7，我们得到指数分布的参数 $\hat{\beta} = 9.988 * 10^{-3}$，幂律分布参数 $\hat{\alpha} = 1.499, \hat{\Delta} = 0.999$，和混合模型参数 $\hat{\beta} = 4.969 * 10^{-4}, \hat{\alpha} = 1.000, \hat{\Delta} = 5.005$。这些估计的参数都很好的拟合了如表5.3所示的真实参数值。

我们确实反推出了动力学生成过程。通过从观察到的截面状态分布中学习参数，我们推断出动力学生成过程，然后生成估计得到的增长动力学过程 $\hat{x_1}(t)$。如图5.1 d-f 所示，估计出的的动力学曲线（绿线）准确地拟合了真实动力学曲线（紫色点）。

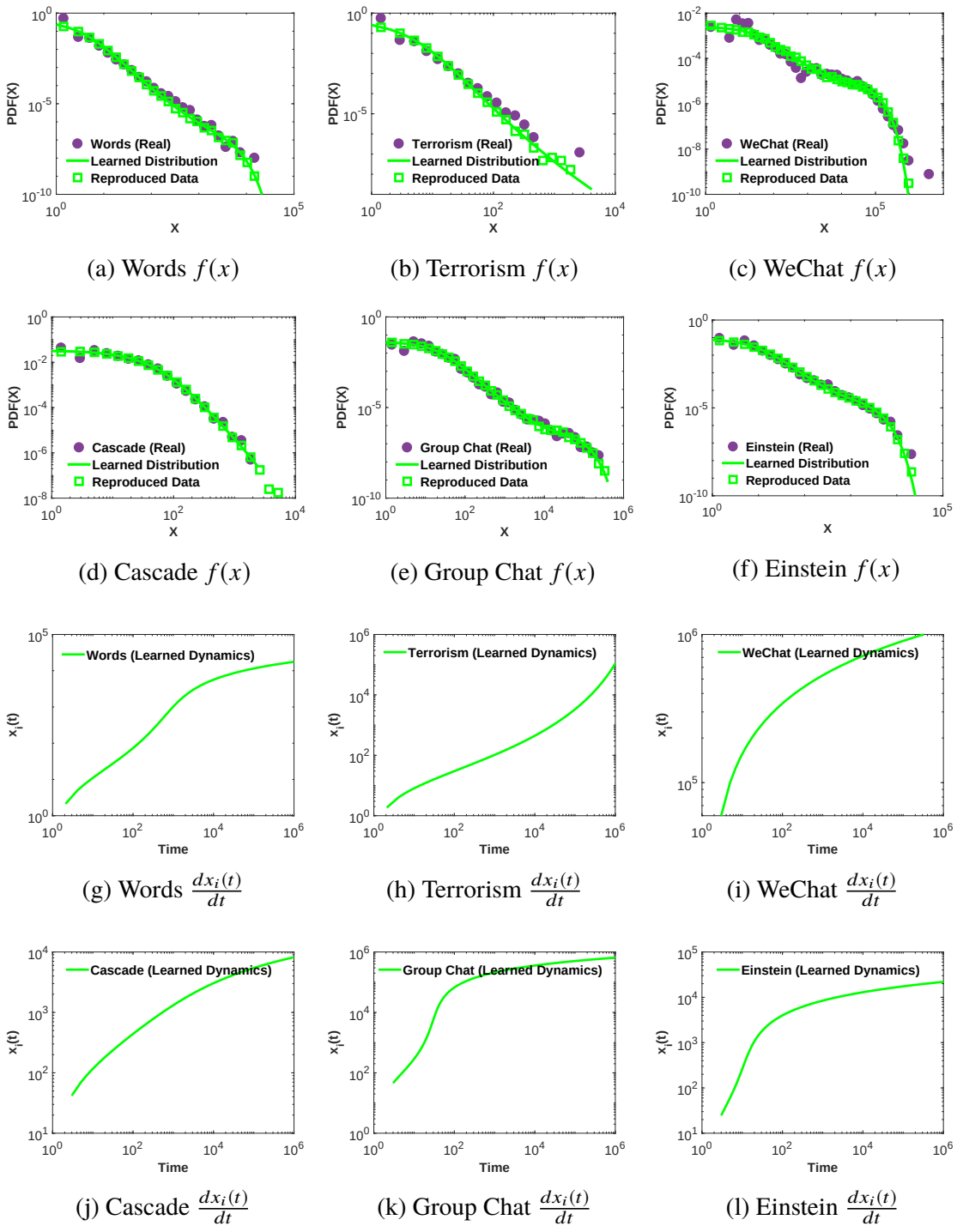


(a) 指数分布　(b) 幂律分布

(c) 混合分布　(d) 指数分布动力学生成系统

(e) 幂律分布动力学生成系统　(f) 混合分布动力学生成系统

图 5.1　所有动态系统都生成我们定理预测的截面状态分布。我们的模型准确地恢复了其动态增长过程参数，并且模拟产生了逼真的数据样本。Exp: Exponential, PL: Power Law, Mix: Mixture。



(a) Words $f(x)$　(b) Terrorism $f(x)$　(c) WeChat $f(x)$

(d) Cascade $f(x)$　(e) Group Chat $f(x)$　(f) Einstein $f(x)$

(g) Words $\frac{dx_i(t)}{dt}$　(h) Terrorism $\frac{dx_i(t)}{dt}$　(i) WeChat $\frac{dx_i(t)}{dt}$

(j) Cascade $\frac{dx_i(t)}{dt}$　(k) Group Chat $\frac{dx_i(t)}{dt}$　(l) Einstein $\frac{dx_i(t)}{dt}$

图 5.2　6 个真实数据的 PDF，它们分布的拟合结果，估计出的动力学生成过程，和模拟产生的样本。真实分布和动力学增长过程是复杂的，但我们的方法准确地拟合了它们。

5.5.2　真实世界数据实验

我们研究了来自不同学科的各种现实世界数据集，来探究它们可能的动力学生成过程是什么。实际上，矛盾的是，有时我们永远不会知道真正的生成动力学。

因此，我们通过检查我们的模型是否拟合并重现我们观察到的现实数据来评估我们的模型可能性。我们采用的数据集是：*(a)* 赫尔曼梅尔维尔在小说“白鲸记”中出现的词数[122]；*(b)* 1968 年 2 月至 2006 年 6 月全球恐怖袭击事件造成的死亡人数[123]；*(c)* 一个活跃微信（中国最大的社交网络）用户连续添加好友行为的时间间隔[124]；*(d)* 腾讯微博信息级联中两次转发的时间间隔[62];*(e)* 腾讯 QQ 在线群聊聊行为的时间间隔[46]；和 *(f)* 爱因斯坦一生中信件的回应时间[36]。其中 a-f 的数据编号与图5.2a-f 相对应。

实际上，现实世界中的分布函数表现出很大的复杂性，如图5.2所示。例如，图5.2f 中描绘的爱因斯坦一生发送邮件的分布函数，我们发现：x（响应时间）很小时 pdf 是平坦的，意味着短时间尺度上的泊松（类似）过程; 中等时间尺度的长尾 x 的 pdf，意味着（基于优先级队列的）决策过程; 而且在长时间尺度上是双峰的 pdf，意味着长期规模的泊松（类似）过程。此外，我们发现现实世界数据集推断出的动力学生成机制是复杂非线性的。对于爱因斯坦的情形，如图5.2l，加速动力学增长然后是饱和部分，暗示爱因斯坦的一生中发送信件的动力学过程是复杂的。

虽然各种数据集都存在内在的复杂性，但我们的框架准确地拟合了它们。在图5.2a-f 中，估计的分布（绿线）很好的拟合了所有这些数据集的复杂分布（紫色点）。进一步，我们在图5.2g-l 中展示了推断出的动力学系统。我们按照定理5.1，从这些估计出的动态系统重现截面状态数据样本，并绘制生成数据的分布，如图5.2a-f 中绿色方块所示。我们发现所生成的数据很好地拟合了这些广泛不同的数据集，表明我们给出了一种产生所观测数据的动力学系统。

5.6　结论

许多科学研究都遵循如下研究模式：从对一个系统在特定时刻下的断面状态观察数据来推断其动力学演化过程。原因是很多自然法则都有着动力学的起源，但是，我们可以观察到的只是在特定时刻下的横截面状态数据。在这里，我们尝试构建它们之间的理论关系。我们认为截面状态数据的宏观统计特性是由我们的定理5.1和推论5.1中描述的微观动态系统生成的，并从理论上构建了截面状态数据分布（或以危险函数的形式）与其动力学生成系统之间的等价关系。然后，我们提出系统参数学习和模拟算法。通过该框架，我们从分布函数发现了几种新的动力学系统，而且反过来，从给定动态系统推测任意时刻的截面状态生成的新分布。我们进一步展示了一个模型，直接从经验数据中学习它们的动力学生成过程。我们的模型框架准确地刻画了各种复杂的合成和真实数据集。我们的研究有助于发现现实世界中复杂动态系统的动力学演化机制。

限制和扩展。我们可以将我们的定理5.1和推论5.1 扩展到非绝对连续的情况，因为：累积危险函数可以推广为 $\Lambda(x) = -\int_{x_0}^{x} \frac{dS(t)}{S(t-)}$ 到非绝对连续的情况，由于 S(x) 是单调递减地[125]。非参模型的扩展可以基于 $\Lambda(x)$（例如 Nelson-Aalen 估计量）和 $S(x)$（例如 Kaplan-Meier 估计量)[125]。包含协变量的半参数模型仍有待探索。

5.7　支持材料

5.7.1　定理证明

我们基于引理5.1，5.2和 5.3 证明了定理5.1 和推论5.1

引理 5.1：[56] 给定一个泊松过程（Poisson process）$\mathscr{P}(t|\lambda_p) = \{t_1, ..., t_i, ...|t_1 \leq ... \leq t_i \leq .. \leq t\}$ 当 $N(t|\lambda_t) = n$ 时，那么当给定时刻 t，对于第 i 个时间的发生时间 t_i 的概率密度函数是 $f(t_i) = \frac{1}{t}$，是 $(0, t]$ 上的均匀分布。

证明 对于随机变量 $t_i, i = 1, ..., n$ 的联合概率密度函数是：

$$
\begin{aligned}
&Pr(t_i < T_i \leq t_i + \delta_i, i = 1, ..., n|N(t) = n) = \\
&\left[\frac{\begin{array}{r} Pr(N(t_i + \delta_i) - N(t_i) = 1, N(t_{j+1}) - N(t_j + \delta_j) = 0, \\ i = 1, ..., n, j = 0, ..., n, t_0 = 0, \delta_0 = 0) \end{array}}{Pr(N(t) = n)} \right] \\
&= \frac{\prod_{i=1}^{n} \lambda_p \delta_i e^{-\lambda_p \delta_i} e^{-\lambda_p (t - \sum_{i=1}^{n} \delta_i)}}{e^{-\lambda_p t} (\lambda_p t)^n / n!} = \frac{n!}{t^n} \prod_{i=1}^{n} \delta_i,
\end{aligned}
\tag{5-11}
$$

因此，

$$
\begin{aligned}
&f(t_i, i = 1, ..., n|N(t) = n) \\
&= \lim_{t_i \to 0, i=1,...,n} \frac{Pr(t_i < T_i \leq t_i + \delta_i, i = 1, ..., n|N(t) = n)}{\prod_{i=1}^{n} \delta_i} \\
&= \frac{n!}{t^n},
\end{aligned}
\tag{5-12}
$$

其中 $t_1 \leq ... \leq ti \leq ... \leq t$。对于有序统计量 $t_i, i = 1, ..., n$，$f(t_i) = \frac{1}{t}$。 □

引理 5.2：[126] 给定随机变量 u 服从 $U(0, 1]$ 上的均匀分布，那么 $x = F^{-1}(u)$ 遵循分布 $F(x)$。另一方面，如果 X 遵循分布 $F(x)$，那么 $F(x)$ 遵循 $U(0, 1]$ 上的均匀分布。

证明 随机变量 X 的累计概率密度函数是：

$$\begin{aligned} Pr(F^{-1}(u) \leq x) = Pr(u \leq F(x)) = F(x), \text{ 并且} \\ Pr(F(x) \leq u) = Pr(x \leq F^{-1}(u)) = F(F^{-1}(u)) = u \end{aligned} \tag{5-13}$$

□

引理 5.3：　给定随机变量 u 服从 $U(0,1]$ 上的均匀分布，那么 $x = \Lambda^{-1}(\ln \frac{S(x_0)}{u})$ 遵循分布 $F(x)$。另一方面，$S(x_0)e^{-\Lambda(x)}$ 遵循 $U(0,1]$ 上的均匀分布。

证明 根据危险函数 $\lambda(x)$ 的定义，

$$\lambda(x) = \frac{f(x)}{S(x)} = \frac{-S'(x)}{S(x)}, \tag{5-14}$$

通过对公式两边从 x_0 积分，我们得到：

$$\int_{x_0}^{x} \lambda(s)ds = \Lambda(x) = -\ln \frac{S(x)}{S(x_0)} \tag{5-15}$$

推出

$$S(x) = 1 - F(x) = S(x_0)e^{-\Lambda(x)}. \tag{5-16}$$

通过引理5.2的结论，我们得到：

$$x = \Lambda^{-1}(\ln \frac{S(x_0)}{1-u}) \tag{5-17}$$

服从分布 $F(x)$。由于 $1-u$ 也遵循 $U(0,1]$，我们得到结果 $x = \Lambda^{-1}(\ln \frac{S(x_0)}{u})$ 遵循分布 $F(x)$，而且 $S(x_0)e^{-\Lambda(x)}$ 遵循均匀分布 $U(0,1]$。 □

此时，我们可以证明定理5.1和推理5.1：

证明 对于动态系统 $\mathscr{D}$ 中的任意个体 x_i，其到达时间 t_i 遵循泊松过程 $\mathscr{P}(t|\lambda_p) = \{t_1, ..., t_i, ...|t_1 \leq ... \leq t_i \leq .. \leq t\}$ 且 $N(t|\lambda_t) = n$，那么 $\frac{t_i}{t}$ 遵循均匀分布 $U(0,1]$（引理5.1）。我们把 $F^{-1}(u)$ 中的 u 替换为 $u = 1 - \frac{t_i}{t}$，把 $\Lambda^{-1}(\ln \frac{S(x_0)}{u})$ 中的 u 替换为 $u = \frac{t_i}{t}$，

我们得到：

$$\begin{aligned} x_i(t) &= \int_{t_0}^{t} d\tau \frac{dF^{-1}(1-\frac{t_i}{\tau})}{d\tau} = F^{-1}(1-\frac{t_i}{t}) - F^{-1}(1-\frac{t_i}{t_0}). \\ x_i(t) &= \int_{t_0}^{t} d\tau \frac{d\Lambda^{-1}(\ln(\frac{S(x_0)}{t_i}\tau))}{d\tau} = \Lambda^{-1}(\ln(\frac{S(x_0)}{t_i}t)). \end{aligned} \tag{5-18}$$

服从分布 $F(x(t))$，而且 $F(x(t))$ 对于动力学过程：

$$\frac{dx_i(t)}{dt}|_{x_0} = \frac{dF^{-1}(1-\frac{t_i}{t})}{dt} = \frac{d\Lambda^{-1}(\ln(\frac{S(x_0)}{t_i}t))}{dt} \tag{5-19}$$

由引理5.2和引理5.3中的等价性保证。 □

当 $S(x_0) = 1$ 时，上式可化为：

$$\frac{dx_i(t)}{dt}|_{x_0} = \frac{dF^{-1}(1-\frac{t_i}{t})}{dt} = \frac{d\Lambda^{-1}(\ln(\frac{t}{t_i}))}{dt} \tag{5-20}$$

5.7.2　参数学习

根据定理5.1和推论5.1，我们发现动力学方程 $\frac{dx_i(t)}{dt}$ 和分布 $f(x)$ 有着相同的参数。因此，我们可以从横截面数据或纵向动态数据中学习它们的参数。

最常见的情况是我们只观察到复杂系统截面状态数据。给定一组横截面数据 $\{x_1, ..., x_{n-1}, x_n\}_t$，我们通过最大化对数似然函数来学习了 $f(x|\Theta)$ 或 $\lambda(x|\Theta)$ 的参数 Θ：

$$\begin{aligned} &\max_{\Theta} \ln L(\Theta|x_1, ..., x_n) \\ &= \ln \prod_{i=1}^{n} f(x_i) \\ &= \ln \prod_{i=1}^{n} \lambda(x_i) e^{-\Lambda(x_i)} = \sum_{i=1}^{n} \ln \lambda(x_i) - \sum_{i=1}^{n} \Lambda(x_i) \end{aligned} \tag{5-21}$$

如果我们随着时间的推移观察动态数据 $\hat{x}_i(t)$，其中 $t = t_i...T$ 和 $i = 1...n$，那么我们通过最小化以下目标函数来学习动态数据 $x_i(t|\Theta)$ 的参数：

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{t_j=t_i}^{T} (\hat{x}_i(t_j) - x_i(t_j))^2 \tag{5-22}$$

其中 $m_i = T - t_i + 1$。那么，估计得到的分布式：$\hat{F}(x(t)|\hat{\Theta})$。

这里我们展示了一般框架，无论分布或动态方程的具体选择如何。因此，一般的优化算法如内点法[127] 或一般非线性最小二乘算法，如 Levenberg-Marquardt 算法[128] 用于目标函数5-22。我们在下一小节中展示了具体的例子。我们可以基于分布函数或动力学系统的特性进一步定制优化算法。

5.7.3　从截面样本学习参数

给定危险函数 $\lambda(x) = \beta + \frac{\alpha}{(x+\Delta)^\theta}$，我们遵循生存分析中的最大似然估计框架，得到对数似然函数：

$$\begin{aligned}\ln L(x_1, ..., x_n) = &\sum_{i=1}^{n} \ln\left[\beta + \alpha(x_i + \Delta)^{-\theta}\right] - \beta \sum_{i=1}^{n} x_i \\ &- \frac{\alpha}{1-\theta} \sum_{i=1}^{n} \left[(x_i + \Delta)^{1-\theta} - \Delta^{1-\theta}\right]\end{aligned} \tag{5-23}$$

当 $\theta \neq 1$ 时，我们得到模型参数的梯度是：

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta} &= \sum_{i=1}^{n} \frac{1}{A(i)} - \sum_{i=1}^{n} x_i \\ \frac{\partial \ln L}{\partial \alpha} &= \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-\theta}}{A(i)} - \frac{1}{1-\theta} \sum_{i=1}^{n} \left[(x_i + \Delta)^{1-\theta} - \Delta^{1-\theta}\right] \\ \frac{\partial \ln L}{\partial \Delta} &= -\alpha\theta \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-\theta-1}}{A(i)} - \alpha \sum_{i=1}^{n} \left[(x_i + \Delta)^{-\theta} - \Delta^{-\theta}\right] \\ \frac{\partial \ln L}{\partial \theta} &= -\alpha \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-\theta} \ln(x_i + \Delta)}{A(i)} \\ &\quad - \alpha \sum_{i=1}^{n} \{\frac{\left[-(x_i + \Delta)^{1-\theta} \ln(x_i + \Delta) + \Delta^{1-\theta} \ln \Delta\right]}{1-\theta} \\ &\quad + \frac{(x_i + \Delta)^{1-\theta} - \Delta^{1-\theta}}{(1-\theta)^2}\}\end{aligned} \tag{5-24}$$

其中 $A(i) = \beta + \alpha(x_i + \Delta)^{-\theta}$。当 $\theta = 1$ 时，我们得到参数梯度为：

$$\begin{aligned}
\frac{\partial \ln L}{\partial \beta} &= \sum_{i=1}^{n} \frac{1}{B(i)} - \sum_{i=1}^{n} x_i \\
\frac{\partial \ln L}{\partial \alpha} &= \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-1}}{B(i)} - \sum_{i=1}^{n} \ln(\frac{x_i}{\Delta} + 1) \\
\frac{\partial \ln L}{\partial \Delta} &= -\alpha \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-2}}{B(i)} + \alpha \sum_{i=1}^{n} \frac{x_i}{x_i\Delta + \Delta^2}
\end{aligned} \tag{5-25}$$

其中 $B(i) = \beta + \frac{\alpha}{x_i + \Delta}$。

5.7.4　样本模拟的解方程算子

基于危险函数和动态微分方程的样本生成器在正文中给出。在这里，我们给出求解文中算法等式的解方程算子：

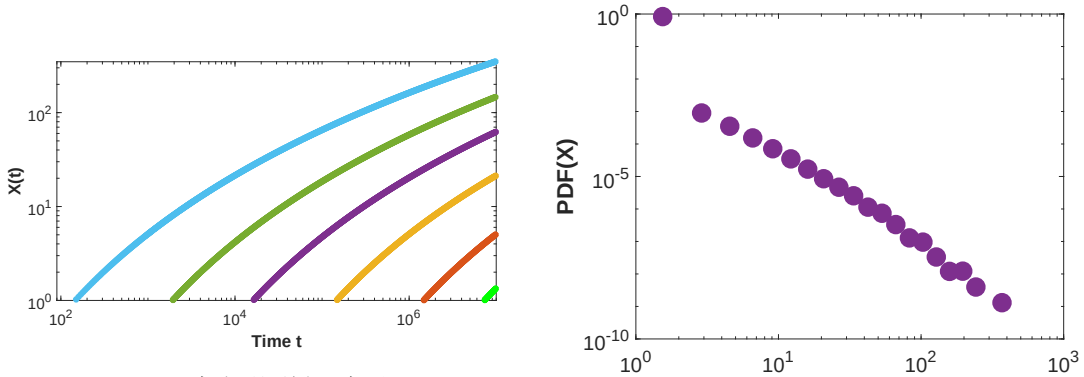
Input ： Equation $\Phi(x) = \log u + \int_{x_0}^{x} \lambda(s)ds$.
Output: x
Set $\epsilon = 10^{-8}$, $x = 0$;
while $|\Phi(x)| \leq \epsilon$ **do**
  if $\theta == 1$ **then**
    $\Phi(x) = \ln u + \beta x + \alpha \ln(\frac{x}{\Delta} + 1)$;
  else
    $\Phi(x) = \ln u + \beta x + \frac{\alpha}{1-\theta}[(x + \Delta)^{1-\theta} - \Delta^{1-\theta}]$;
  end
  $\Phi'(x) = \beta + \alpha(x + \Delta)^{-\theta}$;
  $x = x - \frac{\Phi(x)}{\Phi'(x)}$;
end

Algorithm 5: Equation solver: Newton's iterative method

5.7.5　一个网络系统的解释

在这里，我们在网络演化场景中解释以下动力学微分方程：

$$\frac{dx_i(t)}{dt} = \frac{(x_i(t) + \Delta)^{\theta}}{\beta(x_i(t) + \Delta)^{\theta}t + \alpha t} \tag{5-26}$$



(a) 动力学增长过程

(b) 截面状态分布

图 5.3　定理的示意图。a）每个节点根据公式5-26 改变其好友连接数，如每条彩色曲线所示。（b）在时间点 t，我们观察到横截面度分布。

我们将 $x_i(t)$ 视为网络中每个节点 i 在时间点 t 的好友数，即度数，刻画其动力学变化的机制包括物理机制非线性偏好附加项 $(x_i(t)+\Delta)^\theta$，增长竞争项 αt，以及偏好附加的偏差 Δ 和增长竞争的偏差 $\beta(x_i(t)+\Delta)^\theta t$。因此，我们将网络演化过程描述如下：

- 新节点 i 按照泊松过程在 $0t_i$ 时刻进入网络系统，其中 $0 < t_i < t$ 的之后，t 是最大观察时间（如图5.3a）所示；
- 节点 i 的度数，表示为 $x_i(t)$，随时间根据微分方程5-26增长（如图5.3b）所示。

然后，该网络在 t 时刻的横截面度分布遵循 $f(x) = \lambda(x)e^{-\int_{-\infty}^{x}\lambda(s)ds}$，其中 $\lambda(x) = \beta + \alpha(x+\Delta)^{-\theta}$。

5.7.6　进一步总结

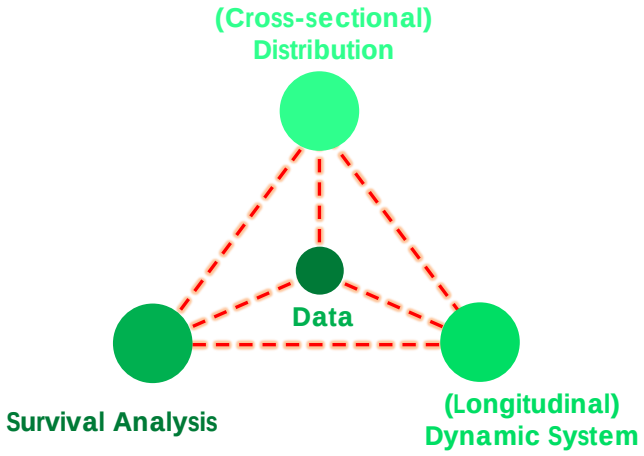


图 5.4　我们的定理和推论尝试连接图中孤立的节点。

通过应用我们的定理5.1 和推论5.1，我们试图找到系统横截面状态分布，生存分析中的危险函数和它们的动力学演化之间的内在关系，如图5.4所示。

表 5.1　分布函数的动力学起源之一。运用我们的定理5.1和推论5.1，我么示例了 9 个典型分布的动力学产生过程。其中 $f(x)$，$F(x)$，$\lambda(x)$ 和 $\Lambda(x)$ 分别表示随机变量 $x \geq x_0$ 的概率密度函数，累计概率密度函数，危险函数，和累计危险函数。*PA* 是 *P*referential *A*ttachment 的简称，*GC* 是 *G*rowth *C*ompetition 的简称，而 *EL* 是 *E*nvironment *L*imit 的简称。

	Data Distribution		**Survival Analysis**		**Dynamic System**		
	$f(x)$	$F(x)$	$\lambda(x)$	$\Lambda(x)$	$x_i(t)$	**Dynamics** $\frac{dx_i(t)}{dt}$	**Interpr.**
Exp.	$\alpha e^{-\alpha x}$	$1-e^{-\alpha x}$	α	αx	$\frac{\ln(\frac{t}{t_i})}{\alpha}$	$\frac{1}{\alpha t}$	GC
Powerlaw	$\alpha x_0^{\alpha} x^{-(\alpha+1)}$	$1-(\frac{x_0}{x})^{\alpha}$	$\frac{\alpha}{x}$	$a \ln \frac{x}{x_0}$	$x_0(\frac{t}{t_i})^{\frac{1}{\alpha}}$	$\frac{x_i(t)}{\alpha t}$	PA + GC
Stretched Exp.	$\frac{\alpha}{x^{\theta}} e^{-\frac{\alpha(x^{1-\theta}-x_0^{1-\theta})}{1-\theta}}$	$1-e^{-\frac{\alpha(x^{1-\theta}-x_0^{1-\theta})}{1-\theta}}$	$\frac{\alpha}{x^{\theta}}$	$\frac{\alpha}{1-\theta}(x^{1-\theta}-x_0^{1-\theta})$	$[\ln(\frac{t}{t_i})\frac{1-\theta}{\alpha}+x_0^{1-\theta}]^{\frac{1}{1-\theta}}$	$\frac{x_i^{\theta}(t)}{\alpha t}$	Non-linear PA + GC
Weibull	$\alpha\lambda^{\alpha}x^{\alpha-1}e^{-(\lambda x)^{\alpha}}$	$1-e^{-(\lambda x)^{\alpha}}$	$\alpha\lambda^{\alpha}x^{\alpha-1}$	$(\lambda x)^{\alpha}$	$\frac{(\ln\frac{t}{t_i})^{\frac{1}{\alpha}}}{\lambda}$	$\frac{x_i^{1-\alpha}(t)}{\lambda^{\alpha}\alpha t}$	Non-linear PA + GC
Log-logistic	$\frac{\lambda\alpha(\lambda x)^{\alpha-1}}{[1+(\lambda x)^{\alpha}]^2}$	$1-\frac{1}{1+(\lambda x)^{\alpha}}$	$\frac{\lambda\alpha(\lambda x)^{\alpha-1}}{1+(\lambda x)^{\alpha}}$	$\ln[1+(\lambda x)^{\alpha}]$	$\frac{(\frac{t}{t_i}-1)^{\frac{1}{\alpha}}}{\lambda}$	$\frac{x_i(t)}{\alpha(t-t_i)}$	PA + Since then GC
Sigmoid	$\frac{e^x}{(1+e^x)^2}$	$1-\frac{1}{1+e^x}$	$\frac{e^x}{1+e^x}$	$\ln(1+e^x)$	$\ln(\frac{t}{t_i}-1)$	$\frac{1}{t-t_i}$	Since then GC
Log-normal *	$\frac{1}{x\sqrt{2\pi}}e^{-\frac{(\ln x)^2}{2}}$	$\Phi(\ln x)$	$\frac{f(x)}{1-\Phi(\ln x)}$	$-\ln[1-\Phi(\ln x)]$	$e^{\Phi^{-1}(1-\frac{t_i}{t})}$	$x_i\frac{d\Phi^{-1}(z)}{dz}\frac{t_i}{t^2}$	PA + Square GC
Normal *	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$	$\Phi(x)$	$\frac{f(x)}{1-\Phi(x)}$	$-\ln[1-\Phi(x)]$	$\Phi^{-1}(1-\frac{t_i}{t})$	$\frac{d\Phi^{-1}(z)}{dz}\frac{t_i}{t^2}$	Square GC
Uniform	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$\frac{1}{b-x}$	$\ln\frac{b-a}{b-x}$	$b-(b-a)\frac{t_i}{t}$	$\frac{b-x_i(t)}{t}$	EL + GC

Notes: t_i in the distribution means that the generated distribution is time-varing as dynamic system evolves. * $z=1-\frac{t_i}{t}$; $A=\frac{N-x_0}{x_0}\frac{x}{N-x}$, and $B=\frac{x_0}{N-x_0}$.

表 5.2　分布函数的动力学起源之二。运用我们的定理5.1和推论5.1，我么示例了 9 个典型的动力学系统产生的截面状态分布函数。其余符合同上表。

	Data Distribution		Survival Analysis		Dynamic System		
	$f(x)$	$F(x)$	$\lambda(x)$	$\Lambda(x)$	$x_i(t)$	**Dynamics** $\frac{dx_i(t)}{dt}$	**Interpr.**
Gen. by Exp.Dyn.	$\frac{\frac{\alpha}{t_i}}{x(\frac{\alpha}{t_i}\ln\frac{x}{x_0}+1)^2}$	$1-\frac{1}{1+\frac{\alpha}{t_i}\ln\frac{x}{x_0}}$	$\frac{\frac{\alpha}{t_i}}{x(\frac{\alpha}{t_i}\ln\frac{x}{x_0}+1)}$	$\ln[\frac{\alpha}{t_i}\ln\frac{x}{x_0}+1]$	$x_0e^{\frac{t-t_i}{\alpha}}$	$\frac{x_i(t)}{\alpha}$	PA
Generated by Str-Exp. Dyn.	$\frac{\frac{\alpha}{t_i^{1-\theta}}}{x[\frac{\alpha(1-\theta)}{t_i^{1-\theta}}\ln\frac{x}{x_0}+1]^{\frac{2-\theta}{1-\theta}}}$	$1-[1+\frac{\alpha(1-\theta)}{t_i^{1-\theta}}\ln\frac{x}{x_0}]^{\frac{-1}{1-\theta}}$	$\frac{\alpha}{\alpha(1-\theta)x\ln\frac{x}{x_0}+t_i^{1-\theta}x}$	$\frac{1}{1-\theta}\ln[\frac{\alpha(1-\theta)}{t_i^{1-\theta}}\ln\frac{x}{x_0}+1]$	$x_0e^{\frac{t^{1-\theta}-t_i^{1-\theta}}{\alpha(1-\theta)}}$	$\frac{x_i(t)}{\alpha t^\theta}$	PA+ Non-linear GC
Gen. by Sigmoid Dyn.*	$\frac{d}{dx}(\frac{\frac{\alpha}{Nt_i}\ln A}{1+\frac{\alpha}{Nt_i}\ln A})$	$1-\frac{1}{1+\frac{\alpha}{Nt_i}\ln A}$	$\frac{d}{dx}(\ln[\frac{\alpha}{Nt_i}\ln A+1])$	$\ln[\frac{\alpha}{Nt_i}\ln A+1]$	$N\frac{Be^{\frac{N(t-t_i)}{\alpha}}}{1+Be^{\frac{N(t-t_i)}{\alpha}}}$	$\frac{x_i(t)[N-x_i(t)]}{\alpha}$	PA + EL
Gen. by LogLogistic Dyn.*	$\frac{\alpha(\frac{N-x_0}{x_0})^{\frac{-\alpha}{N}}x^{-\frac{\alpha}{N}-1}}{N(N-x)^{\frac{-\alpha}{N}+1}}$	$1-A^{-\frac{\alpha}{N}}$	$\frac{\frac{\alpha}{N}}{x(N-x)}$	$\frac{\alpha}{N}\ln A$	$N\frac{B(\frac{t}{t_i})^{\frac{N}{\alpha}}}{1+B(\frac{t}{t_i})^{\frac{N}{\alpha}}}$	$\frac{x_i(t)[N-x_i(t)]}{\alpha t}$	PA + EL + GC
Gen. by str.Logistic Dyn.*	$-\frac{d}{dx}[1+\frac{\alpha(1-\theta)}{Nt_i^{1-\theta}}\ln A]^{\frac{-1}{1-\theta}}$	$1-[1+\frac{\alpha(1-\theta)}{Nt_i^{1-\theta}}\ln A]^{\frac{-1}{1-\theta}}$	$\frac{d}{dx}\frac{\ln[1+\frac{\alpha(1-\theta)}{Nt_i^{1-\theta}}\ln A]}{1-\theta}$	$\frac{\ln[1+\frac{\alpha(1-\theta)}{Nt_i^{1-\theta}}\ln A]}{1-\theta}$	$\frac{NBe^{\frac{N}{\alpha}\frac{t^{1-\theta}-t_i^{1-\theta}}{1-\theta}}}{1+Be^{\frac{N}{\alpha}\frac{t^{1-\theta}-t_i^{1-\theta}}{1-\theta}}}$	$\frac{x_i(t)[N-x_i(t)]}{\alpha t^\theta}$	PA + EL + Non-linear GC
Gen. by Confined Exponential	$\frac{\frac{\alpha}{t_i}\frac{1}{N-x}}{(1-\frac{\alpha}{t_i}\ln\frac{N-x}{N-x_0})^2}$	$1-\frac{1}{1-\frac{\alpha}{t_i}\ln\frac{N-x}{N-x_0}}$	$\frac{\frac{\alpha}{t_i}\frac{1}{N-x}}{1-\frac{\alpha}{t_i}\ln\frac{N-x}{N-x_0}}$	$\ln[1-\frac{\alpha}{t_i}\ln\frac{N-x}{N-x_0}]$	$N-\frac{N-x_0}{e^{\frac{(t-t_i)}{\alpha}}}$	$\frac{N-x_i(t)}{\alpha}$	EL
Gen. by Confined Power law	$\frac{\alpha(N-x_0)^{-\alpha}}{(N-x)^{1-\alpha}}$	$1-(\frac{N-x}{N-x_0})^\alpha$	$\frac{\alpha}{N-x}$	$-\alpha\ln\frac{N-x}{N-x_0}$	$N-(N-x_0)(\frac{t}{t_i})^{\frac{-1}{\alpha}}$	$\frac{N-x_i(t)}{\alpha t}$	GC + EL
Gen. by Confined Str.Exp.	$\frac{\frac{\alpha}{t_i^{1-\theta}(N-x)}}{[1-\frac{\alpha(1-\theta)}{t_i^{1-\theta}}\ln\frac{N-x}{N-x_0}]^{\frac{2-\theta}{1-\theta}}}$	$1-[1-\frac{\alpha(1-\theta)}{t_i^{1-\theta}}\ln\frac{N-x}{N-x_0}]^{\frac{-1}{1-\theta}}$	$\frac{\frac{\alpha}{t_i^{1-\theta}(N-x)}}{1-\frac{\alpha(1-\theta)}{t_i^{1-\theta}}\ln\frac{N-x}{N-x_0}}$	$\frac{\ln[1-\frac{\alpha(1-\theta)}{t_i^{1-\theta}}\ln\frac{N-x}{N-x_0}]}{1-\theta}$	$N-\frac{(N-x_0)}{e^{\frac{t^{1-\theta}-t_i^{1-\theta}}{\alpha(1-\theta)}}}$	$\frac{N-x_i(t)}{\alpha t^\theta}$	Non-linear GC + EL
Gen. by Linear Dyn.	$\frac{\frac{\alpha}{t_i}}{[\frac{\alpha}{t_i}(x-x_0)+1]^2}$	$1-\frac{1}{\frac{\alpha}{t_i}(x-x_0)+1}$	$\frac{1}{(x-x_0)+\frac{t_i}{\alpha}}$	$\ln[\frac{\alpha}{t_i}(x-x_0)+1]$	$x_0+\frac{t-t_i}{\alpha}$	$\frac{1}{\alpha}$	Constant rate

Notes: t_i in the distribution means that the generated distribution is time-varing as dynamic system evolves. * $z=1-\frac{t_i}{t}$; $A=\frac{N-x_0}{x_0}\frac{x}{N-x}$, and $B=\frac{x_0}{N-x_0}$.

表 5.3　实验设置：动力学系统及其横截面状态分布。

	Dynamics $\frac{dx_i(t)}{dt}\|_{x_0}$	**PDF** $f(x)$	**Parameter**
Exponential	$\frac{1}{\beta t}$	$\beta e^{\beta x}$	$\beta = 0.01$
Power Law	$\frac{x_i(t)+\Delta}{\alpha t}$	$\alpha\Delta^{\alpha}x^{-(\alpha+1)}$	$\alpha = 1.5$ $\Delta = 1$
Mix model	$\frac{x_i(t)+\Delta}{\beta(x_i(t)+\Delta)t+\alpha t}$	$\beta e^{-\beta x}(\frac{x}{\Delta}+1)^{-\alpha}$ $+\frac{\alpha}{\Delta}(\frac{x}{\Delta}+1)^{-(\alpha+1)}e^{-\beta x}$	$\beta = 5\text{e}{-4}$ $\alpha = 1$ $\Delta = 5$

第 6 章　复杂分布函数生成及拟合

拟合真实数据的分布函数，然后解释他们的动力学生成机制，是许多学科中研究结构和动态数据的常见研究范式。然而，之前的工作主要是在不同学科背景的各种数据集上拟合和解释其数据分布，例如人类行为时间间隔分布函数，网络度结构分布函数，社会财富分布函数等等。但是，面对现实世界中不同语义下的复杂的数据分布，我们可以通过一个统一但简约的参数化模型来拟合和解释它们吗？对该问题的回答意义重大。

在本文中，我们将复杂的真实数据视为由一个动力学系统生成，该系统将均匀随机信号作为输入，通过确定的动态系统转化，产生复杂的数据。我们展示了一个有着四个参数的动力学模型，以及配套的参数学习和数据模拟算法，其能够拟合并生成一系列分布，包括高斯分布，指数分布，幂律分布，拉伸指数（威布尔 Weibull）分布，及具有多尺度复杂性的复杂混合分布。我们的模型不是一个黑盒模型，因为它可以通过一个统一的微分方程来解释，通过刻画动力学生成机制，来产生如上所述的复杂函数分布。更复杂的模型可以通过我们示例的推导框架有原则的得到。我们通过 16 个来自不同学科的真实数据验证了我们的模型。通过拟合这些数据，我们发现被广泛使用的已有统计模型产生了系统性的偏差，但是我们的模型准确地刻画了真实数据。简而言之，我们的模型提供了一个可能的框架，来拟合真实世界中的拥有复杂分布的数据，而且更重要的是，我们的模型尝试解释他们的动力学生成机制。

6.1　引言

通过参数化模型拟合真实数据分布然后以动力学方程解释其生成过程，是一种研究和理解数据的内在结构和动力学特性的重要科学研究范式。其被广泛应用于各个领域，包括生物学[112]，物理[35,129]，社会科学[45,113]，和计算机科学[46,114] 等等。例如，通过研究网络的幂律度分布[2]，物理学家在随机网络中发现了网络演化产生无尺度效应的动力学机制。通过研究达尔文和爱因斯坦[36] 的邮件回复时间分布，或者在线合作的编辑时间间隔分布[118]，社会科学家试图揭示人类决策行为的动态特性。通过拟合数据分布与高斯混合模型[130]，贝叶斯方法[131]，甚至深度生成模型[132]，计算机科学家试图找到观察数据集的聚类结构和其生成动力学机制。简而言之，这种科学研究范式适用于广泛的数据科学任务中。

然而，以前的工作主要是在不同学科不同数据下以个案研究的方式拟合或解

释复杂的真实数据分布。例如,高斯分布最广泛地用于拟合窄尾(narrow-tailed)数据分布。大量的文献尝试通过幂律分布[129], Weibull 分布 (或拉伸指数分布)[133] 等来模拟重尾 (heavy-tailed) 数据。特定混合模型也用于拟合复杂的多尺度分布[46,47,118]。像 GAN 这样的深度生成网络在拟合 1-D 参数分布[134] 方面表现出很一般的效果。因此，我们是否可以拥有一个统一且简洁的模型来拟合和解释现实世界中各种复杂的数据分布函数呢？对该问题的回答是至关重要的。

在本文中, 我们试图通过研究真实数据复杂分布的动力学产生机制, 来拟合分布并解释其产生机制。我们模型背后的直觉如下：我们将具有复杂分布的真实数据视为从一个动力学系统中生成，其采用均匀随机信号作为输入，然后经过系统变化产生复杂输出。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且可能很简约的动力学生成机制进行建模，从而用一个统一的模型来生成所有这些复杂的分布。表6.1 中展示了一个实例：我们并没有对高斯分布，指数分布，幂律分布，拉伸指数分布，以及它们在多尺度机制中具有复杂混合部分的变体分布进行分别的建模，而是通过一个具有四个参数的动力学模型，我们产生了所有这些看似不相关的分布。我们的框架可以有原则地构建更复杂的动态模型和分布函数。此外，我们给出了有效的参数学习方法和数据模拟生成算法。我们的模型不是一个黑盒模型，而是可以通过一个统一的微分方程来解释这些复杂分布的动力学生成机制。至于实验，我们首先通过各种模拟数据集分析了我们模型的性质，并通过来自很多学科的 16 个真实数据集进一步验证。我们的模型准确地拟合了所有这些复杂的数据分布函数（图 6.5）。我们的模型提供了一个可能的统一框架来拟合在现实世界中观察到的复杂分布函数，更重要的是，解释它们的动力学生成机制。我们的主要贡献总给如下：

- **统一模型**： 我们提出了一个通用的模型来拟合真实数据的各种复杂分布函数，并给出参数拟合和数据模拟算法。
- **简洁性**： 我们的模型只有四个参数来拟合真实数据分布中的多尺度混合模态的复杂性。我们展示了刻画动力学机制的优势--通过刻画相对简单动力学产生过程来简化刻画复杂生成现象的复杂度。
- **可解释性**：我们的模型由统一的动力学方程生成和解释，其所有参数都有明确的物理意义。
- **实用性**：我们的模型准确地拟合了各种真实数据集，并且可以以有原则的方式推广到更复杂的情况。而被广泛使用的拟合长尾分布的统计模型产生了系统性偏差。

该章节的大纲是：相关工作介绍，模型，物理学机制，实验，讨论和结论。

6.2　相关工作

我们主要通过回顾了以下两个方面的相关工作：

从真实数据的简单分布函数到复杂的分布函数。窄尾分布（narrow-tailed distributions），例如指数分布和高斯分布，可以通过它们的均值和方差很好得刻画，它们的结构和产生机制都有的详细研究。相反，重尾分布（heavy-tailed distributions），如幂律分布（power-law distribution），拉伸指数分布（stretched-exponential distribution），对数正态分布（log-normal distribution）等等，表现出更大的甚至无穷大的方差，这意味其有着着复杂的数据生成机制。在长尾分布中，幂律分布是最着名的，因为它的缩放的属性（scaling）[135] 和在网络中的生成机制[2]。关于幂律分布的广泛存在的证据和讨论可以在[122,136] 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[46,124]，网络数据[137] 到各种如图6.5 所示数据集。

拟合复杂函数的方法。最大似然估计法，可能额外使用先验或正则化因子，被广泛的用于拟合窄尾分布[131]。另一方面，像 GAN 这样的深度生成网络在拟合 1-D 参数分布的实验中，表现出较大的偏差[134]。相比之下，拟合复杂分布的理论，比如偏斜或长尾分布的理论[138]，并不是很完善。以最典型的情况 - 幂律分布 - 作为一个例子，最开始，人们通过视觉检查在双对数坐标系下最小二乘拟合的结果来判断对幂律分布的拟合好坏。后来，被广泛引用的工作[129] 显示了最小二乘拟合方法的系统偏差，然后提出了一个参数方法 $f(x) = \frac{\alpha_{PL}-1}{x_{min}}(\frac{x}{x_{min}})^{\alpha_{PL}}$，记为 PL 方法），通过最大似然法来拟合幂律分布。PL 方法在大量科学论文中被广泛用于拟合各种似是而非的幂律分布。然而，我们发现 PL 方法在检测现实数据中的幂律信号时显示出较大的偏差，如图 6.5所示。其系统偏差的根源在于 PL 方法忽略了现实世界数据分布的复杂性[46,47,124]。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。

6.3　模型

6.3.1　模型直觉解读

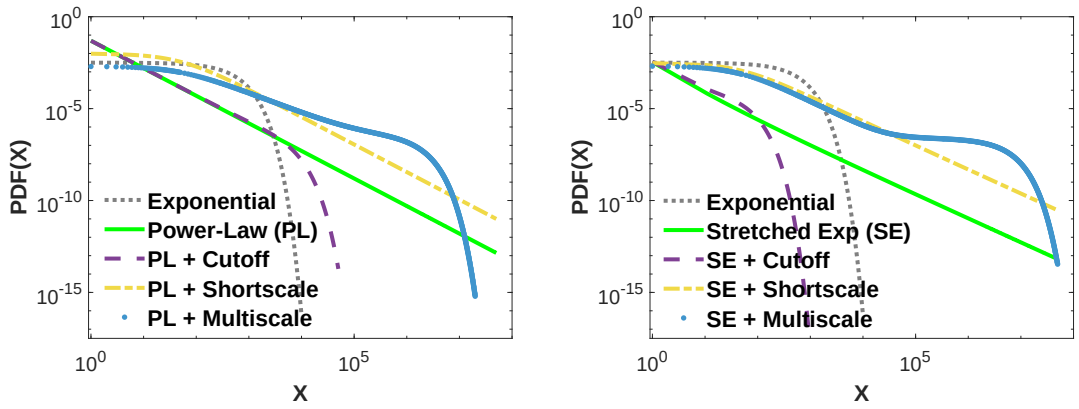
我们模型背后的直觉解读如下：我们将具有复杂分布的真实世界中的数据视为从一个（非线性）的动态系统中产生，该系统将均匀随机信号作为输入，产生复杂输出。与其刻画一个个复杂的输出，我们试图通过刻画统一且简单的动力学系统，来产生和建模复杂的输出（即各种数据分布）。简而言之，我们试图建模产生复杂现象的简单动力学生成模型。

表 6.1　模型能力表。我们的基本模型包含了如下所有分布。图示如 6.1。

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

* 对于具有截止情况的幂律分布和拉伸指数分布，其概率密度函数通过危险函数近似得到，参阅模型章节。

** 当 $\theta = -1$，模型的特例市近似正态分布，参阅模型章节。



(a) $\theta = 1$　　(b) $\theta \neq 1$

图 6.1　模型所产生的分布。我们的模型产生了一系列分布，包含幂律分布（power law，PL），幂律带截尾分布（PL with cutoff），幂律带短尺度复杂模态（PL with short-scale complexity），幂律带多尺度复杂模态（PL with multi-scale complexity），指数分布（exponential），拉伸指数分布（stretched exponential，SE），拉伸指数带短尺度复杂模态（SE with short-scale complexity），拉伸指数带多尺度复杂模态（SE with multi-scale complexity）等等。

我们的模型基于生存分析（survival analysis）[46]，随机点过程（point process）[124]，以及动态系统（dynamic systems）[3,139,140]。我们定义数据 $\mathcal{X} = (x_1, ..., x_{n-1}, x_n)$ 的概率密度函数（probability density function）为 $f_{\mathcal{X}}(x)$，其可以被危险函数（hazard function）$\lambda(x) = \frac{f_{\mathcal{X}}(x)}{S_{\mathcal{X}}(x)}$ 建模。危险函数刻画了一种条件概率，即产生随机变量（random variable）$\mathcal{X} = x$ 在条件 $\mathcal{X} \geq x$ 下的概率密度。其中 $S_{\mathcal{X}}(x) = 1 - \int_{-\infty}^{x} f_{\mathcal{X}}(s)ds$。我们定义 $\Lambda(x) = \int_{-\infty}^{x} \lambda(s)ds$ 为累计危险函数（cumulative hazard rate）。通过建模危险函数，我们可以产生复杂的概率密度函数，根据变换关系 $f_{\mathcal{X}}(x) = \lambda(x)e^{-\int_{-\infty}^{x} \lambda(s)ds}$。我们进一步将危险函数 $\lambda(x)$ 和动力学系统连接，来进一步解释复杂数据分布的动力学产生机制。

6.3.2　生存分析建模

在此我们提出我们的基本模型，它虽然简单但是多才多艺，可以产生一系列复杂的分布，如下表 6.1 所示，我们进一步通过图示 6.1 将分布在双对数坐标系下画出。刻画模型的危险函数（hazard function）是：

$$\lambda(x) = \beta + \alpha(x + \Delta)^{-\theta} \tag{6-1}$$

6.3.2.1　临界态：以幂律为基础的复杂分布

当 $\theta = 1$ 时，$\lambda(x|\theta = 1)$ 产生一族以幂律分布（power-law distribution）为基础的分布。

引理 6.1：危险函数 $\lambda(x|\theta = 1)$ 产生一族分布，包括：指数分布（exponential distribution），幂律分布（power-law distribution），有指数截尾的幂律分布（power-law distribution with exponential cut-off），复杂多尺度混合分布（complex multi-scale distributions）等等。

证明 随机变量 x 的概率密度分布（probability density function）是：

$$\begin{aligned} f(x|\theta = 1) &= \lambda(x|\theta = 1)e^{-\int_0^x \lambda(s|\theta=1)ds} \\ &= (\beta + \frac{\alpha}{x + \Delta})e^{-\beta x - \alpha \ln(\frac{x}{\Delta}+1)} \\ &= \beta e^{-\beta x}(\frac{x}{\Delta} + 1)^{-\alpha} + \frac{\alpha}{\Delta}(\frac{x}{\Delta} + 1)^{-(\alpha+1)}e^{-\beta x} \end{aligned} \tag{6-2}$$

指数分布（Exponential distribution）。当 $\alpha = 0$ 时，无论其他三个参数取值如何，危险函数 $\lambda(x|\alpha = 0)$ 产生指数分布函数，其概率密度函数是 $f(x|\alpha = 0) = \beta e^{-\beta x}$，如图6.1中灰线所示。

幂律分布（Power-law distribution）。当 $\beta = 0$ 和 $\Delta \ll x$ 时，$f(x|\theta = 1, \beta = 0) = \alpha\Delta^{\alpha}(x + \Delta)^{-(\alpha+1)} \propto x^{-(\alpha+1)}$。$\Delta$ 的另一个含义是 x 的最小取值，比如 x_0。因此，x 的 pdf 形式为：$f(x|\theta = 1, \Delta = 0) = \frac{\alpha}{x}e^{\int_{x_0}^{x} \frac{\alpha}{s} ds} = \alpha {x_0}^{\alpha} x^{-(\alpha+1)}$。

有指数截尾的幂律分布（power-law distribution with exponential cut-off）。当 $\beta \gg 0$ 且 $\Delta \ll x \ll \frac{\alpha}{\beta} - \Delta$ 时，x 的 pdf 为：$f(x|\theta = 1) = (\beta + \frac{\alpha}{x+\Delta})(\frac{x}{\Delta} + 1)^{-\alpha}e^{-\beta x} \approx \alpha\Delta^{\alpha}x^{-(\alpha+1)}e^{-\beta x}$。

复杂多尺度混合分布（Complex multi-scale distributions）。当 $\beta \to 0$ 时，复杂多尺度混合分布在短时间尺度趋近于常数，在中时间尺度呈现幂律分布，在长时间尺度呈现指数分布。当 $x \to 0$ 时，$f(x|\theta = 1) \to \frac{\alpha}{\Delta}$。在短时间尺度范围 $x \in (0, \Delta]$，$f(x|\theta = 1) \approx \frac{\alpha}{\Delta}(\frac{x}{\Delta} + 1)^{-(\alpha+1)}$，其 pdf 值慢慢下降成为中尺度幂律分布。当 $\beta e^{-\beta x}(\frac{x}{\Delta}+1)^{-\alpha} \gg \frac{\alpha}{\Delta}(\frac{x}{\Delta}+1)^{-(\alpha+1)}e^{-\beta x}$，即 $x \gg \frac{\alpha}{\beta} - \Delta$，$f(x|\theta = 1) = \beta e^{-\beta x}(\frac{x}{\Delta}+1)^{-\alpha} + \frac{\alpha}{\Delta}(\frac{x}{\Delta}+1)^{-(\alpha+1)}e^{-\beta x} \approx \beta e^{-\beta x}(\frac{x}{\Delta} + 1)^{-\alpha}$，其呈现长时间尺度下的指数分布。当 $\Delta \ll x \ll \frac{\alpha}{\beta} - \Delta$ 时，$f(x|\theta = 1) \approx \alpha\Delta^{\alpha}(x + \Delta)^{-(\alpha+1)} \propto x^{-(\alpha+1)}$，其呈现中时间尺度的幂律分布。　□

6.3.2.2　一般情况：以拉伸指数分布为基础的复杂分布

当 $\theta \neq 1$ 时，$\lambda(x|\theta \neq 1)$ 产生一族以拉伸指数分布（stretched exponential distribution）为基础的分布。

引理 6.2：　危险函数 $\lambda(x|\theta \neq 1)$ 产生一族概率分布函数，包括指数分布（Exponential distribution），拉伸指数分布（stretched exponential distribution，或称作 Weibull 分布），带有指数截尾的拉伸指数分布（stretched exponential distribution with exponential cut-off），复杂多尺度混合分布（complex multi-scale distributions）等等。

证明 和上一个引理证明类似，随机变量 x 的概率密度函数为：

$$\begin{aligned} f(x|\theta \neq 1) &= \lambda(x|\theta \neq 1) e^{-\int_0^x \lambda(s|\theta \neq 1) ds} \\ &= [\beta + \alpha(x+\Delta)^{-\theta}] e^{-\beta x - \frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]} \\ &= \beta e^{-\beta x} e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]} \\ &+ \alpha(x+\Delta)^{-\theta} e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]} e^{-\beta x} \end{aligned} \tag{6-3}$$

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

拉伸指数分布（Stretched exponential distribution），**或称作威布尔** (Weibull) **分布**。当 $\beta = 0$ 且 $\Delta = 0$ 时，危险函数 $\lambda(x|\theta \neq 1)$ 为：

$$f(x|\theta \neq 1) = \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta} x^{1-\theta}} \tag{6-4}$$

□

累计概率分布函数为：$F(x|\theta \neq 1) = 1 - e^{-\frac{\alpha}{1-\theta} x^{1-\theta}}$，为拉伸指数分布。一些特例：当 $\theta = 0$ 时，其为指数分布 $\alpha e^{-\alpha x}$；当 $\theta = -1$ 时，趋近于正太分布（Normal distribution）$\frac{\alpha}{x^2} e^{-\frac{\alpha x^2}{2}}$。

带有指数截尾的拉伸指数分布（Stretched exponential distribution with exponential cut-off）　当 $\beta \gg 0$ 和 $\Delta \ll x \ll (\frac{\alpha}{\beta})^{\frac{1}{\theta}} - \Delta$ 时，$f(x|\theta \neq 1) = [\beta + \alpha(x+\Delta)^{-\theta}] e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]} e^{-\beta x} \approx \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta} x^{1-\theta}} e^{-\beta x}$。

复杂多尺度混合分布（Complex multi-scale distributions）。当 $\theta \neq 1$ 时，复杂多尺度混合分布是基于拉伸指数分布而复杂的。具体来讲，当 $\beta \to 0$，复杂多尺度混合分布在短时间尺度呈现常数，在中时间尺度呈现拉伸指数分布，在长时间尺度呈现指数分布。当 $x = 0$，$f(x|\theta \neq 1) \approx \frac{\alpha}{\Delta^\theta}$。在短时间尺度范围内 $x \in (0, \Delta]$，$f(x|\theta \neq 1) \approx \alpha(x+\Delta)^{-\theta} e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]}$，其概率密度慢慢以拉伸指数分布衰减到中时间尺度。当 $\beta \gg \alpha(x+\Delta)^{-\theta}$ 时，即 $x \gg (\frac{\alpha}{\beta})^{\frac{1}{\theta}} - \Delta$，$f(x|\theta \neq 1) \approx \beta e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]} e^{-\beta x}$，其时长时间尺度的指数分布。当 $\Delta \ll x \ll (\frac{\alpha}{\beta})^{\frac{1}{\theta}} - \Delta$ 时，$f(x|\theta \neq 1) \approx \alpha(x+\Delta)^{-\theta} e^{-\frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]}$，其是中时间尺度的拉伸指数分布。

我们在图6.1中展示了如上结论。我们在图6.1 a 中展现了以幂律分布为基础的复杂分布；在图6.1 b 中展现了以拉伸指数分布为基础的复杂分布。再一次强调，所

有这些分布都是由我们简单的危险函数模型产生，它可以产生从简单经典分布到多尺度混合的复杂分布。

6.3.3 学习模型参数

我们模型的参数可以通过最大似然估计（MLE）框架来学习估计出。我们观察到一组数据 $\{x_1, ..., x_{n-1}, x_n\}$ 的对数似然函数为：

$$\ln L(x_1, ..., x_n) = \ln \prod_{i=1}^{n} \lambda(x_i) e^{-\Lambda(x_i)} = \sum_{i=1}^{n} \ln \lambda(x_i) - \sum_{i=1}^{n} \Lambda(x_i) \tag{6-5}$$

根据 θ 的不同取值，$\Lambda(x)$ 呈现不同的数学形式。当 $\theta \neq 1$ 时，对数似然函数是：

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

，当 $\theta = 1$ 时，其对数似然函数为：

$$\ln L(x_1, ..., x_n | \theta = 1) = \sum_{i=1}^{n} \ln [\beta + \alpha(x_i + \Delta)^{-1}] - \beta \sum_{i=1}^{n} x_i - \alpha \sum_{i=1}^{n} \ln(\frac{x_i}{\Delta} + 1) \tag{6-7}$$

最大化公式6-5 或6-6 来得到参数 $\{\beta, \alpha, \Delta, \theta\}$，且满足限制 $\{\beta, \alpha, \theta \geq 0; \Delta > 0\}$，就会得到我们估计的参数。然而，由于参数明确的物理含义，更多先验知识可以融入优化中，比如初始值设定等等。我们之后回展示这一点。

该模型的另一个好处是所有参数都具有显示可求解的梯度。我们给出当 $\theta \neq 1$ 情况下的梯度为（即基于拉伸指数的模型）：

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{A(i)} - \sum_{i=1}^{n} x_i \tag{6-8}$$

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-\theta}}{A(i)} - \frac{1}{1-\theta} \sum_{i=1}^{n} [(x_i + \Delta)^{1-\theta} - \Delta^{1-\theta}] \tag{6-9}$$

$$\frac{\partial \ln L}{\partial \Delta} = -\alpha\theta \sum_{i=1}^{n} \frac{(x_i + \Delta)^{-\theta-1}}{A(i)} - \alpha \sum_{i=1}^{n} [(x_i + \Delta)^{-\theta} - \Delta^{-\theta}] \tag{6-10}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \theta} &= -\alpha \sum_{i=1}^{n} \frac{(x_i+\Delta)^{-\theta}\ln(x_i+\Delta)}{A(i)} \\ &\quad -\alpha \sum_{i=1}^{n} \{\frac{[-(x_i+\Delta)^{1-\theta}\ln(x_i+\Delta)+\Delta^{1-\theta}\ln\Delta]}{1-\theta} + \frac{(x_i+\Delta)^{1-\theta}-\Delta^{1-\theta}}{(1-\theta)^2}\}\end{aligned} \tag{6-11}$$

其中 $A(i)=\beta+\alpha(x_i+\Delta)^{-\theta}$。

当 $\theta=1$ 时，我们给出基于幂律模型的梯度：

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{B(i)} - \sum_{i=1}^{n} x_i \tag{6-12}$$

$$\frac{\partial \ln L}{\partial \alpha} = \sum_{i=1}^{n} \frac{(x_i+\Delta)^{-1}}{B(i)} - \sum_{i=1}^{n} \ln(\frac{x_i}{\Delta}+1) \tag{6-13}$$

$$\frac{\partial \ln L}{\partial \Delta} = -\alpha \sum_{i=1}^{n} \frac{(x_i+\Delta)^{-2}}{B(i)} + \alpha \sum_{i=1}^{n} \frac{x_i}{x_i\Delta+\Delta^2} \tag{6-14}$$

其中 $B(i)=\beta+\frac{\alpha}{x_i+\Delta}$。我们可以通过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[127]。为了实验的可重复性，我们开源了代码，见第6.7小节。

6.3.4　数据模拟

从累积分布函数 $F(x)$ 生成随机数 x 的最简单和最优雅的方法是逆变换方法（inverse transformation method）[126]。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 $F(x)=u$ 来得到 x，那么 x 就是服从分布 $F(x)$ 的随机数。我们进一步将该方法扩展到生存分析中，即通过 $F(x)=1-e^{-\Lambda(x)}$ 的事实将这种逆变换方法扩展到危险率函数，其中 $\Lambda(x)=\int_{x_0}^{x}\lambda(s)ds$。因此，$F(x)=u=1-e^{-\Lambda(x)}$，我们可以通过求解 $\Lambda(x)=-\ln(1-u)$ 来得到 x。其中 u 和 $1-u$ 在从 $U(0,1]$ 抽样时没有差别。由于 $\Lambda(x)$ 是单调递增函数，因此具有反函数 Λ^{-1}，我们可以得到：

$$x=\Lambda^{-1}(-\ln u) \tag{6-15}$$

即使反函数 Λ^{-1} 没有显示可推导出的形式，我们也可以通过数值求解的方式得到。具体而言，通过求解等式 $\ln u + \Lambda(x) = 0$ 得到 x，其中 u 是从均匀分布 $U(0,1)$ 中产生。

Input : Hazard function of model $\lambda(x) = \beta + \alpha(x+\Delta)^{-\theta}$, total number N
Output : $\{x_1, ..., x_N\}$
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 7.;
    $x_n = x$;
end

Algorithm 6: 通过危险函数模型6-1来产生随机变量

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

Algorithm 7: Newton 迭代法

6.4　物理动力学机制

在本节中，我们将给出我们模型6-1 的动力学生成机制，其产生的各种分布如表6.1所示。我们将复杂数据分布视为由（非线性）动态系统生成，该系统将均匀

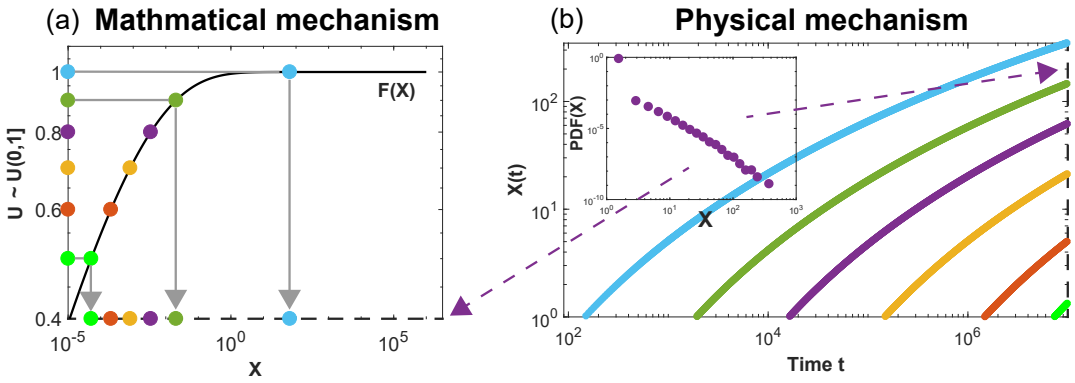


图 6.2　产生拉伸指数分布（如 b 中插图所示）的（a）数学方法，和（b）物理动力学过程。（a）中产生的数据在 x 轴上分布，（b）中产生的数据是一个在任意时刻 t 下的截面状态 x(t) 的分布。

表 6.2　动力学机制。表中展示了幂律分布和拉伸指数分布的动力学产生机制。动力学过程如图6.2 所示。

	幂律 Power law	**拉伸指数 Stretched exponential**
PDF	$\alpha\Delta^{\alpha}x^{-(\alpha+1)}$	$\alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$
Hazard rate	$\frac{\alpha}{x}$	$\frac{\alpha}{x^{\theta}}$
数学产生机制 $x=\Lambda^{-1}(-\ln u)$ *		
Inverse method	$x=\Delta e^{\frac{u}{\alpha}}$	$x=(\frac{1-\theta}{\alpha}u)^{\frac{1}{1-\theta}}$
动力学产生机制: $x_i(t)=\Lambda^{-1}(-\ln(\frac{t}{t_i}))$ *		
Growth	$x_i(t)=\Delta(\frac{t}{t_i})^{\frac{1}{\alpha}}$	$x_i(t)=(\frac{1-\theta}{\alpha}\ln(\frac{t}{t_i}))^{\frac{1}{1-\theta}}$
Preferential attachment	$\frac{dx_i}{dt}=\frac{x_i}{\alpha t}$	$\frac{dx_i}{dt}=\frac{x_i^{\theta}}{\alpha t}$

* 即使反函数 Λ^{-1} 没有显示可推导出的形式，我们也可以通过数值求解的方式得到

随机信号作为输入：

6.4.1　均匀输入信号和动力学增长

我们将从动力学系统的角度给出产生数据的过程，我们的第一步是通过随机点过程（Point process）和生存分析（Survival analysis）的角度构建系统的输入信号。我们从标准均匀分布 $U(0,1]$ 中产生 n 个符合目标分布随机数的过程可以看作是一个随机点过程。给定泊松过程 $N(t)=\{t_i|i=1,...,N(t)=n;0<t_1\leq t_2\leq ...\leq t_n\}$，然后 t_i 均匀分布 $U(0,1]$ 的区间内。如果我们将 t_i 通过除以 t 来标准化，然后 $u=\frac{t_i}{t}$ 遵循标准均匀分布 $U(0,1]$。我们把等式6-15 中的 u 替换成 $u=\frac{t_i}{t}$，得到每个个体 i

的动力学增长方程，每个个体 i 在 $(0,t]$ 时间段内到达时间 t_i 是随机均匀分布的:

$$x_i(t) = \Lambda^{-1}(-\ln u) = \Lambda^{-1}(-\ln(\frac{t_i}{t})) = \Lambda^{-1}(\ln(\frac{t}{t_i})). \tag{6-16}$$

例如，我们让 $\lambda(x) = \alpha x^{-\theta}$，当 $\theta = 1$ 时，其产生幂律分布 $f(x) = \alpha\Delta^{\alpha}x^{-(\alpha+1)}$，当 $\theta \neq 1$ 时，其产生拉伸指数分布 $f(x) = \alpha x^{-\theta}e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$ (如模型小节所述)。因此，$\Lambda(x|\theta = 1) = \alpha\int_{\Delta}^{x}\frac{1}{s}ds = \alpha\ln(\frac{x}{\Delta})$，而且 $\Lambda(x|\theta \neq 1) = \alpha\int_{\Delta}^{x}\frac{1}{s^{\theta}}ds = \frac{\alpha(x^{1-\theta}-\Delta^{1-\theta})}{1-\theta}$。我们可以得到他们的反函数 $\Lambda^{-1}(y|\theta = 1) = \Delta e^{\frac{y}{\alpha}}$ 和 $\Lambda^{-1}(y|\theta \neq 1) = (\frac{1-\theta}{\alpha}y + \Delta^{1-\theta})^{\frac{1}{1-\theta}}$。

通过对公式6-16 使用 $\Lambda^{-1}(y|\theta = 1) = \Delta e^{\frac{y}{\alpha}}$，我们得到动力学增长公式:

$$x_i(t) = \Lambda^{-1}(\ln(\frac{t}{t_i})|\theta = 1) = \Delta e^{\frac{\ln(\frac{t}{t_i})}{\alpha}} = \Delta(\frac{t}{t_i})^{\frac{1}{\alpha}}. \tag{6-17}$$

类似的，当 $\theta < 1$ 时，通过将 $\Lambda^{-1}(y|\theta \neq 1) = (\frac{1-\theta}{\alpha}y + \Delta^{1-\theta})^{\frac{1}{1-\theta}}$ 带入等式 6-16 ，我们得到动力学增长曲线:

$$x_i(t) = \Lambda^{-1}(\ln(\frac{t}{t_i})|\theta \neq 1) = (\frac{1-\theta}{\alpha}\ln(\frac{t}{t_i}) + \Delta^{1-\theta})^{\frac{1}{1-\theta}}. \tag{6-18}$$

6.4.2　动力学生成过程-基本模型

我们的第二步是通过连接生存分析和动力学系统来逆向工程其动力学生成机制。通过对公式6-17和6-18对时间求导，我们得到产生幂律分布和拉伸指数分布的动力学生成模型如下:

$$\frac{dx_i(t)}{dt} = \frac{d\Delta(\frac{t}{t_i})^{\frac{1}{\alpha}}}{dt} = \frac{\Delta}{t_i^{\frac{1}{\alpha}}}\frac{1}{\alpha}t^{\frac{1}{\alpha}-1} = \frac{x_i(t)}{\alpha t} \tag{6-19}$$

$$\frac{dx_i(t)}{dt} = \frac{d(\frac{1-\theta}{\alpha}\ln(\frac{t}{t_i}) + \Delta^{1-\theta})^{\frac{1}{1-\theta}}}{dt} = \frac{(\frac{1-\theta}{\alpha}\ln(\frac{t}{t_i}) + \Delta^{1-\theta})^{\frac{\theta}{1-\theta}}}{\alpha t} = \frac{x_i(t)^{\theta}}{\alpha t} \tag{6-20}$$

我们通过公式6-19 发现产生幂律分布的线性偏好依附（linear preferential attachment）机制，和通过公式6-20 发现产生拉伸指数分布的非线性偏好依附（non-linear preferential attachment）机制，和在网络科学中发现的随机图（random networks）产生的无尺度效应（scale-free observations）结论一致[2]，也间接证明了我们的理论。

6.4.3　动力学生成过程-扩展模型

在此我们给出我们模型的动力学生成过程。当 $\lambda(x) = \beta + \alpha(x+\Delta)^{-\theta}$ 时，$\Lambda(x) = \int_0^x \lambda(s)ds = \beta x + \frac{\alpha}{1-\theta}[(x+\Delta)^{1-\theta} - \Delta^{1-\theta}]$。通过我们的构造，我们得到：

$$\Lambda(x_i(t)) = \beta x_i(t) + \frac{\alpha}{1-\theta}[(x_i(t)+\Delta)^{1-\theta} - \Delta^{1-\theta}] = \ln\frac{t}{t_i} \tag{6-21}$$

通过对等式两边时间 t 求导，我们得到：

$$\frac{dx_i(t)}{dt} = \frac{(x_i(t)+\Delta)^{\theta}}{\beta(x_i(t)+\Delta)^{\theta}t + \alpha t} \tag{6-22}$$

因此，从动力学系统的角度来看，表现出复杂多尺度混合分布的复杂数据（由我们的模型6-1 所刻画）是从具有微分方程6-22 的动力学系统生成的。该系统包括物理机制如下：非线性偏好附件项 $(x_i(t)+\Delta)^{\theta}$，系统在不断增长项 αt，加上短时间尺度混合复杂性 Δ 和长时间尺度混合复杂性 $\beta(x_i(t)+\Delta)^{\theta}t$。我们的动力学系统包括系统6-19和系统6-20作为特例。

进一步，我们在随机网络场景中描述数据生成过程如下：

- 新节点 i 在时间段 $0 < t_i < t$ 内按照泊松过程到达系统，其中 t 是最大观察时间；
- 节点 i 的度数 $x_i(t)$，即临接边数，随着时间按照动力学方程6-22增长。

之后，在任意时刻 t 下观测系统每个节点的数度状态，即截面（cross-sectional）状态，符合分布 $f_X(x) = \lambda(x)e^{-\int_{-\infty}^{x}\lambda(s)ds}$，其中 $\lambda(x) = \beta + \alpha(x+\Delta)^{-\theta}$ 如模型6-1所示。

6.5　实验

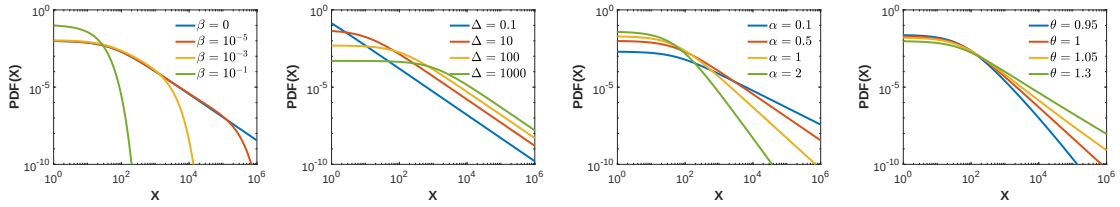
在本节中，我们将在模拟数据和真实数据集上评估我们的模型。

6.5.1　模拟数据分析

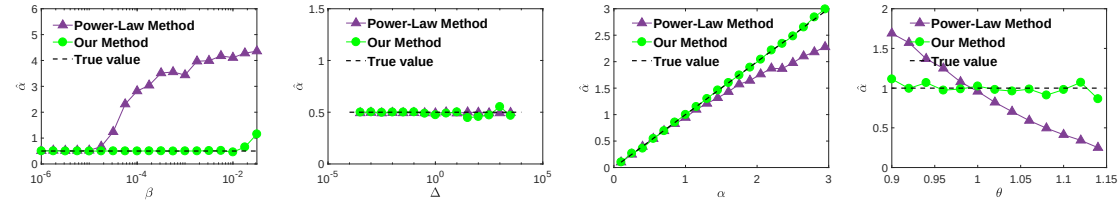
6.5.1.1　忽略现实复杂性引入的系统偏差

实际上，现实世界数据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，我们研究了将著名且被广泛使用的幂律拟合方法（表示为 PL 方法）[129] 应用于复杂分布而可能引入的偏差。

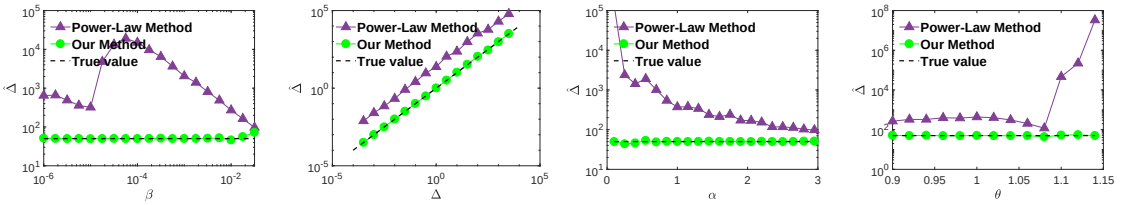
长时间尺度复杂性。 我们首先研究在长时间尺度引入复杂混合分布对幂律方法 PL 带来的系统误差。参数 β 是建模长期复杂性的最简单形式。通过改变



(a) Long-scale complexity: $PDF(X|\beta)$　(b) Short-scale complexity: $PDF(X|\Delta)$　(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:$\hat{\alpha}(\Delta)$　(g) Middle-scale complexity:$\hat{\alpha}(\alpha)$　(h) Middle-scale complexity:$\hat{\alpha}(\theta)$



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

图 6.3　在长时间尺度，短时间尺度和中时间尺度引入的复杂模态给先前的方法引入了系统偏差，而我们的方法（绿色圆点线）很好地拟合了现实。第一行显示了不同尺度范围内具有不同复杂性的分布。最后两行展示了模型参数的平均估计值，随着刻画相应尺度内复杂模态的参数变化而变化的函数。每列中的三个图像具有相同的参数设置。在所有情况下，统计误差都小于数据点，所以没有画出。真实参数值以虚线显示。PL 方法中描述的 pdf 是 $f(x) = \frac{\alpha_{PL}-1}{x_{min}}(\frac{x}{x_{min}})^{\alpha_{PL}}$，其中 $\alpha = \alpha_{PL} - 1$。

$\lambda(x|\beta, \alpha = 0.5, \Delta = 50, \theta = 1)$ 中的 β 的值，我们得到了一系列的分布，如图6.3所示。当 $\beta = 0$ 时，如图 6.3 a 中蓝色曲线的直线部分（幂律指数为 $1 + alpha = 1.5$）所示，表明其没有在长时间尺度下的复杂混合信号。随着 β 的增加，长时间尺度下的复杂混合信号将向左，即朝短时间尺度发展，直到两个部分重叠。实际上，长时间尺度的特征尺度是 $\frac{\alpha}{\beta}$，而短时间尺度的特征尺度是 Δ（参见模型小节）。我们通过限制 $\beta < 10^{-2}$ 来避免两个尺度范围的重叠。对于特定的 β 值，我们通过

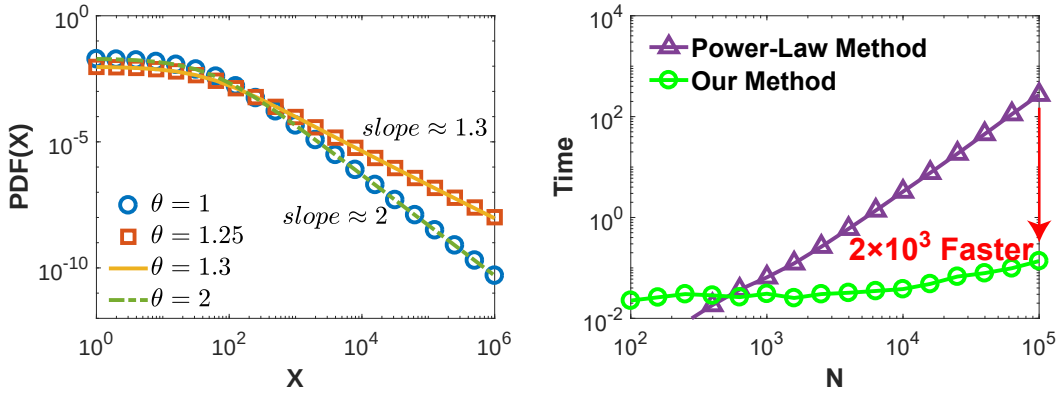
$\lambda(x|\beta, \alpha = 0.5, \Delta = 50, \theta = 1)$ 模型生成 10^4 个样本（一个相对较大可以获得合理的拟合结果的数据集，同时在基线对比方法的可扩展性范围内。我们将展示我们的模型的可扩展性，而对比模型不能扩展到大的数据集。），然后用 PL 方法和我们的模型来拟合样本来估计 α 和 Δ。

图 6.3 e 和 i 绘制了 α（缩放指数，scaling exponent）和 Δ 的平均估计值随着 β 变化而改变的函数。我们发现 PL 方法估计的幂律缩放指数 α 与虚线标记的真值之间的差异随着 β 增大而越来越大，如图6.3 e 所示。相比之下，我们的模型很好地估计了真正的幂律缩放指数的值。对于图6.3 i 中所示的刻画短期复杂信号的参数 Δ 的估计，PL 方法严重高估了真实价值，当 $\beta \approx 3 * 10^{-5}$ 时，PL 对其估计值高估达 450 倍！随着 β 的增加，长时间尺度和短时间尺度逐渐重合。因此通过 PL 方法估计的 Δ 从高估逐渐降低到真实值。

短时间尺度复杂性。 然后我们考虑短期复杂性的影响。参数 Δ 是刻画短时间尺度复杂信号的最简单形式。图 6.3 b 绘制了 $\lambda(x|\beta = 0, \alpha = 0.5, \Delta, \theta = 1)$ 随着 Δ 变化而变化的关系，其概率密度函数包括短时间尺度平台和之后的有着相同幂律指数的幂律分布。当 Δ 越大时，其短时间尺度的范围越大。类似地，对于特定的 Δ 的值，我们通过 $\lambda(x|\beta = 0, \alpha = 0.5, \Delta, \theta = 1)$ 生成 10^4 个样本，然后用 PL 方法和我们的模型拟合样本来估计 α 和 Δ 的值。如图 6.3 f 和 j 所示，我们的模型很好地估计了两个参数的真实值，对于这个实验设置，PL 方法很好地拟合了幂律缩放指数 α，但是 α 的好结果是以高估 Δ 为代价的，表明 PL 方法丢弃了在短时间尺度的很多样本数据样本。具体而言，短时间尺度的样本数占整个数据集的 $\geq 80\%$。说明 PL 方法并不能整体的刻画数据的分布，但是我们的模型可以。

中时间尺度复杂性。最后, 我们研究了中等时间尺度下复杂性的影响。当 $\theta = 1$ 时，中等时间尺度下的概率密度函数遵循幂律分布，其幂律缩放指数为 $1 + \alpha$。通过改变 $\lambda(x|\beta = 0, \alpha, \Delta, \theta = 1)$ 的 α 值并控制其他参数时，我们得到中等时间和长时间尺度下，有着不同幂律指数的幂律分布。图 6.3 c 绘制了当 α 变化时相应的 $\lambda(x|\beta = 0, \alpha, \Delta = 50, \theta = 1)$，$\alpha$ 越大时，曲线越陡峭。然而，我们发现当 α 增长时，PL 方法低估了幂律缩放指数 α，并且其差异变得越来越大，如图6.3 g 所示。此外，PL 方法同时严重高估了短时间尺度尺度参数 Δ，甚至高达 3 个量级，如图6.3 k 所示。相比之下，我们的模型始终如一地准确地估计了参数地真实值。

当 $\theta \neq 1$ 时，中等时间尺度下的概率密度函数遵循拉伸指数分布（stretched-exponential distribution）。图6.3 d 绘制了不同 θ 下的 $\lambda(x|\beta = 0, \alpha = 1, \Delta = 50, \theta)$。具有 $\theta = 1$ 的红色曲线是幂律分布（在中等和长时间尺度下），其 pdf 曲线尾部的斜率为 $\alpha + 1 = 2$，而其他曲线是拉伸指数分布。我们无法通过视觉检查来区分这些



(a) β　　(b) β

图 6.4　a）拉伸指数分布的渐近行为。(b）模型对数据规模的扩展性。我们的模型可以应用于大规模数据集，而幂律 PL 方法则不能。

幂律分布或拉伸指数分布曲线之间的差异。我们在下一个小节通过运用幂律工具和泰勒展开来分析幂律和拉伸指数分布之间的关系。在 $(-\infty, 1+\epsilon]$ 的范围内，当 $\epsilon \to 0$（参考渐近分析部分）时，θ 越大，曲线的尾部越高，如图6.3 d 所示。当 $\theta < 1$ 时，我们发现 PL 方法高估了幂律缩放指数 α，而当 $\theta > 1$ 时，PL 方法却会低估 α 的值，如图6.3 h 所示。与此同时，PL 方法一直高估了 Δ，如图6.3 l 所示。相比之下，我们的模型再次给出了更好的估计。

6.5.1.2　渐进分析

我们进一步研究了模型的渐进行为。以拉伸指数函数作为例子，当在概率概率分布尾部我们几乎无法从肉眼区分幂律分布和拉伸指数分布。具体来讲，拉伸指数分布的概率密度函数为：$f(x|\theta \neq 1) = \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$ 当 $x > 0$ 时。$\theta < 1$ 时，通过在 $x \to 0$ 时的 Taylor 展开，我们得到：

$$f(x \to 0|\theta < 1) = \frac{\alpha}{x^{\theta}}[1 - \frac{\alpha}{1-\theta}x^{1-\theta} + O(x^{2-2\theta})] \tag{6-23}$$

因此，$f(x \to 0|\theta < 1) \approx \frac{\alpha}{x^{\theta}}$，该信号很容易在 Δ 很大时在短时间尺度丢失。当 $x \to \infty$ 时，$f(x \to \infty|\theta < 1) = \alpha x^{-\theta} e^{-\frac{\alpha}{1-\theta}x^{1-\theta}}$ ，其比幂律分布 $\frac{\alpha}{x^{\theta}}$ 衰减得更快，但是比指数分布 $e^{-\frac{\alpha}{1-\theta}x^{1}}$ 衰减得更慢。

当 $\theta > 1$ 且 $x \to \infty$ 时，我们得到 Taylor 展开式：

$$f(x \to \infty|\theta > 1) = \frac{\alpha}{x^{\theta}}[1 + \frac{\alpha}{(\theta-1)x^{\theta-1}} + O(\frac{1}{x^{2\theta-2}})] \approx \alpha\frac{1}{x^{\theta}} + \frac{\alpha^2}{\theta-1}\frac{1}{x^{2\theta-1}} \tag{6-24}$$

，近似于幂律分布，其幂律缩放指数由 θ 和 $2\theta-1$ 交替控制。当 $\theta=1+\epsilon$ 其中 $\epsilon\to 0^+$ 时，系数 $\frac{\alpha^2}{\theta-1}\gg\alpha$，因此 $f(x\to\infty|\theta=1+\epsilon,\epsilon\to 0^+)\approx\frac{\alpha^2}{\theta-1}\frac{1}{x^{2\theta-1}}=\frac{\alpha^2}{\theta-1}\frac{1}{x^{1+2\epsilon}}$，表示其缩放指数为 $1+2\epsilon$ 当 $\epsilon\to 0^+$。相反，当 $\theta\gg 1$，$f(x\to\infty|\theta\gg 1)\approx\frac{\alpha}{x^\theta}$，表示其幂律缩放指数为 θ 当 $\theta\gg 1$ 时。因此，在数学上，我们得出结论，经验中的幂律结论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之前的状态并且更加陡峭和陡峭。上述渐近分析可以通过具有不同 θ 值的分布的重合来验证，如图 6.4 a 所示。

6.5.1.3　模型对数据规模的可扩展性

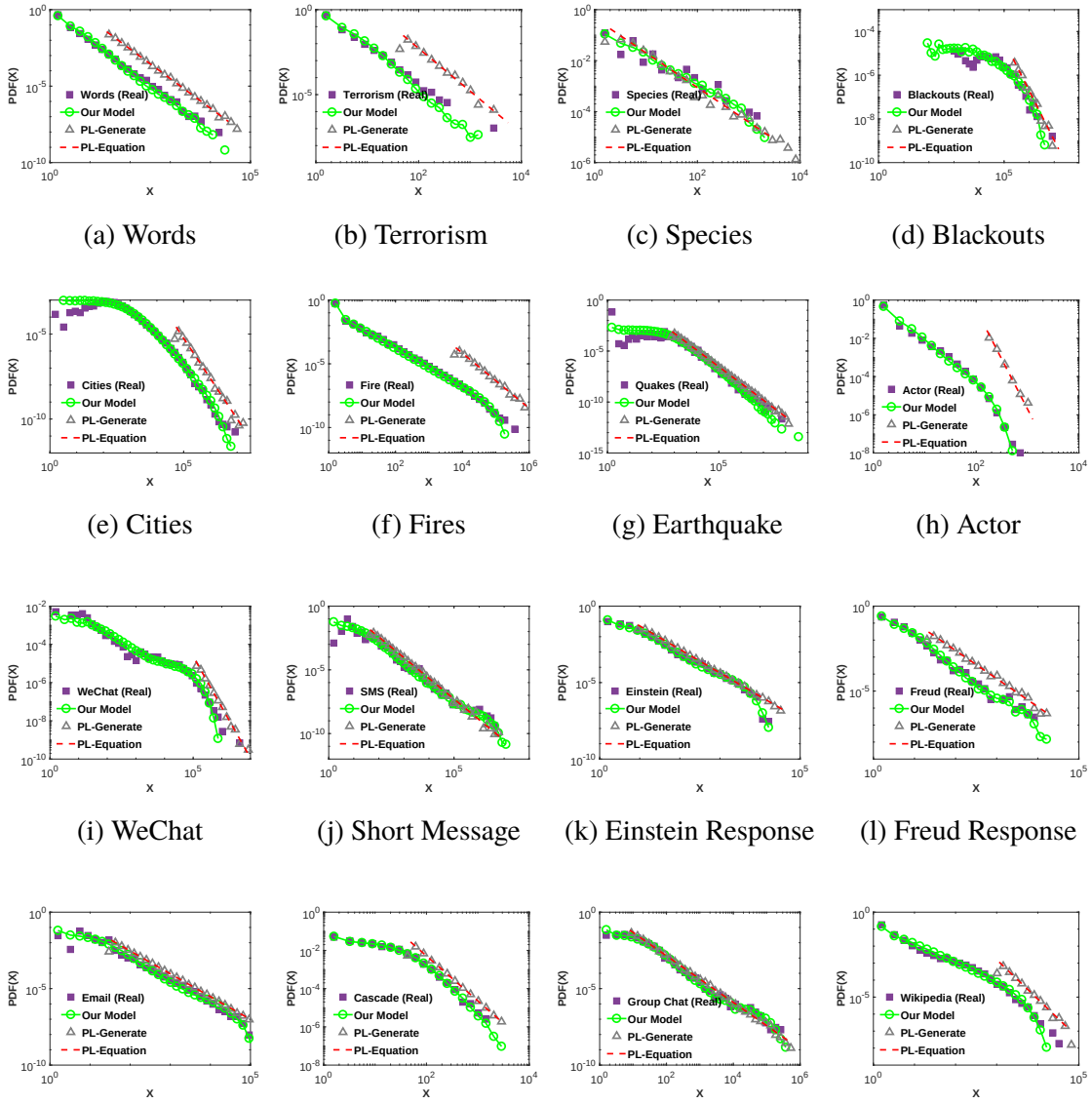
我们通过数值模拟对比了我们模型和幂律（PL）方法对大规模数据的的扩展性。我们在幂律分布配置 $\lambda(x|\beta=0,\alpha=1,\Delta=50,\theta=1)$ 下生成 N 个样本，并通过改变 N，我们绘制两种方法拟合数据所消耗的平均时间，如图 6.4 b 所示。幂律方法的复杂度为 $\approx O(N^2)$。幂律方法的样本大小的经验上限为 10^5。PL 方法的可扩展性更差是由于对 Δ[129] 的线性搜索导致。但是，我们的方法可以以更快的速度应用于更大的数据集。例如，当 $N=10^5$ 时，我们得到拟合结果的速度为 PL 方法的 $\approx 2*10^3$ 倍，并且对分布整体的拟合有更好的准确性。

6.5.2　真实世界数据分析

6.5.2.1　广泛的数据集

我们通过来自不同学科不同场景下的 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，我们将它们分类为静态截面数据（cross-sectional observation）和动态数据（dynamic observation）。我们使用 a-p 来对应如图6.5的编号。前八个数据集来自静态截面数据：

1. *(a)* 赫尔曼梅尔维尔（Herman Melville）在小说《白鲸记》（Moby Dick）中出现的单词的使用数[122]。
2. *(b)* 1968 年 2 月至 2006 年 6 月全球恐怖袭击事件造成的死亡人数[123]。
3. *(c)* 在地球上的每个分类群体中的哺乳动物数量[141]。
4. *(d)* 1984 年至 2002 年间受美国停电影响的客户数量[122]。
5. *(e)* 2000 年美国人口普查中的美国城市人口数[129]。
6. *(f)* 1986 年至 1996 年间美国发生的山林野火的英亩数大小[122]。
7. *(g)* 1910 年至 1992 年间在加利福尼亚发生的地震强度[122]。
8. *(h)* 电影-演员二分网络中的演员参演电影度数[2]。



(a) Words　(b) Terrorism　(c) Species　(d) Blackouts

(e) Cities　(f) Fires　(g) Earthquake　(h) Actor

(i) WeChat　(j) Short Message　(k) Einstein Response　(l) Freud Response

(m) Emails　(n) Weibo Cascade　(o) Group Chat　(p) Wikipedia

图 6.5　来自真实世界不同学科的 16 个数据。每张图绘制了真实数据的 PDF，我们模型拟合结果，和 PL 方法[129] 的拟合结果。真实的分布是复杂的多尺度混合分布，我们的方法（绿色圆圈）很准确的拟合了现实数据集（紫色方块），而被广泛采用的 PL 方法（其函数以虚线标记，其函数生成的样本以三角形刻画），则显示出较大的误差。

后八个数据集来自人类（human dynamics）或社会动态行为（social dynamics）的详细记录：

1. *(i)* 微信（WeChat）中一个活跃用户在添加连续好友事件的时间间隔（inter-event time，IET）[3,124]。
2. *(j)* 来自手机用户的连续发短消息的时间间隔分布[47]。
3. *(k)* 爱因斯坦一生中信函通信的回复时间[36]。

4. (l) 弗洛伊德一生中信函通信的回复时间[36]。
5. *(m)* 在一所大学内三个月所有用户间连续发送邮件的时间间隔[35,44].
6. *(n)* 腾讯微博（Tencent Weibo）中一条特定信息的级联传播（cascade）过程中中两次转发的时间间隔[6,62]。
7. *(o)* 腾讯 QQ 中多人在线群聊行为的时间间隔[46]。
8. *(p)* 连续修订一个维基百科（Wikipedia）项目的合作行为的时间间隔[118]。

我们编辑并公开所有数据集（参见6.7）以保证实验的可重复性，其中最后八个数据集为第一个人类和社会动态的数据集（human and social dynamics）。

6.5.2.2 结果

我们通过回答我们的模型是否可以刻画所有经验数据集来验证我们的方法。我们将我们的方法与[129] 中开发的最先进的方法进行比较，表示为 PL 模型，它被广泛用于拟合可能遵循幂律的胖尾（fat-tailed）分布。

图6.5绘制了真实数据集，PL 模型和我们的模型拟合的结果。我们发现真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。根据上一小节模拟数据分析，分布中的多尺度复杂性致使 PL 模型严重高估 x_{min}，我们也在真实数据集中观察到同样的偏差，如图6.5所示。由 PL 模型学习的 $f(x)$ 有着系统的偏差：参阅图6.5中表示 PL 结果的灰色三角形和表示实际数据的紫色方块的极大的差异。

然后我们进行了定量分析。给定真实数据 $X = \{x_1, ..., x_n\}$，通过 PL 方法和我们模型，我们估计了 pdf $f(x|\Theta)$ 的参数 Θ。然后，我们用我们的方法生成模拟数据样本标记为 $X_{our} = \{x_1, ..., x_{n'}\}$，和用 PL 方法产生的模拟数据 $X_{PL} = \{x_1, ..., x_{n'}\}$ 。我们通过双样本 Kolmogorov-Smirnov 距离（KS-Dist）评估拟合准确性，即 KS-Dist $= max_x|\hat{F}_i(x) - F(x)|$，该误差越低越好。$F(x)$ 是从实际数据中学习的非参数累积分布，而 $\hat{F}(x)$ 是从方法 i 产生的模拟数据集中学习的非参数累积分布。双样本 Kolmogorov-Smirnov 距离广泛用于这种假设检验任务。为了消除由于生成的样本数量较少而导致的误差，我们设置生成样本数为 $n' = 10 * n$。我们在表格6.3中总结了数据和结果。我们发现对于所有 16 个数据集，我们的方法较 PL 方法而言误差要小得多，这表明我们方法的优越性。

6.6　讨论

我们模型的设计原则是：保持简单，刻画复杂。我们只用一个参数 Δ 来刻画短时间尺度的复杂性，一个参数 β 刻画长时间尺度的复杂性，一个参数 θ 用于同时刻画中时间尺度的幂律分布和拉伸指数分布。进一步，通过我们的框架可以设计出更加复杂的模型。例如，可以进一步开发混合物重尾模型，对数正态分布等等。可以预期通过更复杂的模型我们可以得到更小的真实数据和拟合结果之间的误差。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。更多真实世界的数据集应该被验证。应重新检查以前基于将 PL 方法应用于复杂分布得到的所有结论。

6.7　结论

在本文中，我们发现各种真实数据的分布函数，从艺术，生物，物理，地质，社会科学到计算机科学，从横断面观察到动态记录，都遵循复杂的多尺度混合模型。我们开发了一个框架，来刻画现实世界中的复杂分布。通过对数据的生成过程的动力学建模，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。我们进一步给出了有效的参数学习方法和数据模拟生成算法。更重要的是，我们的模型不是一个复杂的黑盒模型，而是用一个统一的微分方程来解释所有这些复杂分布的动力学生成机制。我们通过各种各样的合成数据集分析模型的属性，并通过 16 个真实数据集验证了我们的模型。我们的模型很好地拟合了所有这些数据的复杂分布。我们的模型可能提供一个框架，来拟合真实数据中的复杂分布，并了解它们的动力学生成机制。简而言之，我们的贡献总结如下：

- **统一的模型**：我们提出了一个通用模型，来拟合真实数据中的各种复杂分布，并给出参数学习和数据模拟算法。
- **简洁性**：通过四个参数，我们的模型通过一个简单的形式来刻画真实数据分布中的多尺度复杂性。
- **可解释性**：我们的模型可以通过一个统一的动力学方程解释其动力学生成极致。模型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真实数据集分布，并且可以推广成更复杂的模型。我们将代码和数据集开源在：www.calvinzang.com。

表 6.3　真实数据集的拟合结果。16 个数据集的基本统计数据，PL 方法的结果和我们方法的结果。基于 KS-Dist 度量，我们的模型以比 PL 方法的误差大幅度缩小，更准确地拟合了真实数据集。PL 方法中描述的 pdf 是 $f(x) = \frac{\alpha_{PL}-1}{x_{min}}(\frac{x}{x_{min}})^{\alpha_{PL}}$。

	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**	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\Delta}$	$\hat{\theta}$	**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) **GroupChat**	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

第 7 章　结论与展望

7.1　结论

本文尝试将计算机科学和物理学理论融合，通过数据驱动的方式，对复杂社交系统演化机制进行了动力学建模，并通过微信（首次）和腾讯微博等大规模社交数据上进对所提出的研究方法在行了实验验证。具体而言，本文研究了复杂社交系统演化的三个核心子课题：一，社交网络多尺度演化规律发现和建模，其旨在回答复杂社交系统在不同尺度如何增长的问题；二，信息流在网络中传播的复杂模式生成，其旨在回答信息流在复杂网络上如何传播的问题；三，宏观分布函数的微观动力学起源定理，其旨在回答如何连接微观行为和宏观现象的问题。

- 在第二章中，我们研究了**宏观社交系统演化规律发现和建模**：我们研究了多个社交复杂系统的宏观演化过程，包括了中国最大社交网络微信从上线两年内网络演化的详细过程，覆盖 3 亿用户，47.5 亿条带创建时间的社交连接。我们发现许多社交系统用户数的增长不是指数增长，也不是线性增长，而是幂律增长。进一步，我们首先提出了社交连接的增长也是符合幂律增长的现象。我们给出了产生社交复杂系统节点和连接幂律增长的机制，并通过动力学方程和对应的微观随机过程建模。我们的动力学方程，能产生广泛的复杂动力学增长现象，并准确地拟合和预测了真实社交系统的增长规律。
- 在第三章中，我们研究了**微观社交系统演化规律发现和建模**：我们研究了驱动复杂社交系统宏观演化的微观个体动态行为，发现微观个体加好友行为呈现极大的随机性和异构性。我们发现微观个体行为在长期遵循非线性幂律随机增长过程，在短期呈现爆发随机增长过程。我们提出了三个机制，即平均效应，多尺度效应，相关效应，来控制不同尺度下的个体随机行为模式，并给出一个长短记忆随机过程建模。通过模型分析，我们进一步发现了微观用户加好友的统计规律和典型行为类型，并应用于用户画像聚类和异常检测等应用。
- 在第四章中，我们研究了**信息流在网络中传播的复杂模式生成**：我们研究了信息流在网络中传播产生有规律的复杂几何模式的过程。尽管越来越多的研究旨在了解信息流的传播机制，但对于这些传播模式的几何形状以及它们在传播过程中是如何形成的却知之甚少。通过探索了大规模在线社交媒体数据集中提取的 4.32 亿个信息流模式，我们在一个三维度量空间发现信息流传播结构的复杂几何模式。相比之下，对信息流传播结构的几何模式的现有理

解仅限于扇形展开或狭窄的树状的几何形状。我们发现了控制复杂信息流几何模式形成的三个关键因素：异质性，集体性，和记忆性。之后，我们提出了一个包含这些因素的随机过程模型，证明它可以成功复现真实信息流传播模式中发现的各种几何形状。我们的发现为信息流的微观机制提供了理论基础，其可能的应用包括对信息的预测，控制和政策决策等等。

- 在第五章和第六章，我们研究了**（宏观）分布函数的（微观）动力学起源定理**：我们总结了许多科学研究都遵循如下模式：从对一个系统的截面状态数据来推断其动力学生成/演化机制。但是，正式且系统地学习它们之间关系的研究却少之又少。我们将复杂的截面状态数据视为通过确定的动态系统生成，该系统以均匀的随机信号作为输入。我们构造了（截面状态）概率分布函数与其动力学生成系统之间的一个等价关系，然后开发了一个框架来从截面状态数据，或数据分布函数，来推断其动力学生成过程。通过这样的框架，我们能够从各种分布中发现新的动力学生成机制，而且可以从各种动力学生成机制中发现新的概率分布函数。我们通过合成数据和真实数据验证了我们的框架。实验结果表明，我们的框架能够准确地发现和拟合各种数据分布函数的动力学生成过程。我们的研究有助于发现现实世界中复杂截面数据的未知动力学生成机制进一步，在第六章中，我们给出一个统计模型，来拟合和解释真实世界中的复杂分布函数，而最常用的统计模型面对现实复杂数据表现出了系统误差。我们展示了刻画动力学机制的优势--通过刻画相对简单的动力学产生过程，来简化刻画复杂生成现象模型的复杂度。

7.2　展望

我们通过数据驱动的动力学建模研究方法，例如第一次对微信等数据的大规模研究，发现了很多全新的复杂现象，并进一步增强了原有物理动力学模型刻画复杂社交系统的能力。而相对于计算机对社交网络的研究工作，我们从复杂系统角度引入物理动力学模型的可解释性来解释复杂社交系统的运行机制。我们希望进一步沿着这个方向发展，通过融入机器学习模型，进一步增强物理模型对复杂社交系统数据的表达能力，进一步通过物理动力学模型（统计物理，网络科学等）融入可解释性，通过计算机科学（数据挖掘，机器学习等）提供可计算性，到达两个学科更好的融合。我们将未来研究思路总结如下图7.1：

具体而言，对于时间和空间维度的增长现象。无论是第二章中的宏观动力学增长，还是第三章中的微观随机增长，我们提出的增长模型可以被广泛用于其他研究领域，如生态学，社会科学，人口学等，从线虫 C.elegans 的连接增长到公司

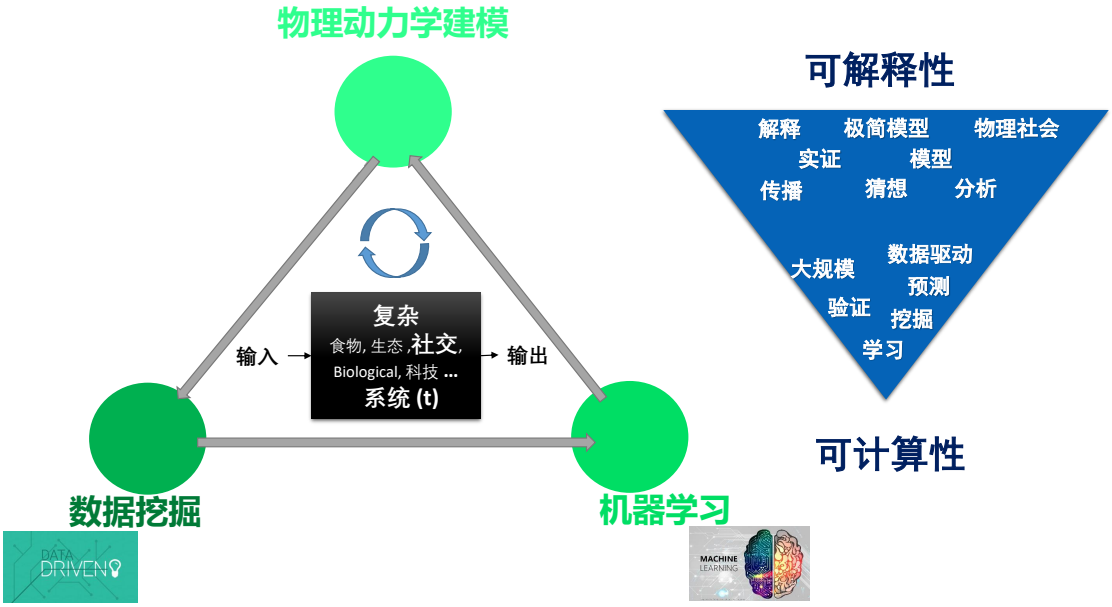


图 7.1　**计算科学与物理动力学理论的交叉研究范式**

企业的成长等等。此外，我们模型的一个主要限制是忽视外部影响。在开放环境中的外部信号如何影响社交网络的增长或衰退动态仍有待研究。无论是宏观还是微观，如何刻画网络增长的更多指标，如更多结构特征，抑或是提出新的模型框架，将网络结构在连续时间内演化直接建模，仍有待探索。

在第四章信息流在网络中传播的复杂模式生成中，我们探索了如何产生信息流复杂几何结构的动力学模型。进一步，我们是否可以给出刻画信息流传播的模型，在结构，时间和语义三个维度都可以很准确地模拟真实信息传播？这样，我们就可以给出模拟真实世界信息传播的微观系统，进一步模拟诸如群体事件的产生和发展过程，到达舆情监控的目标。

在截面状态的动力学起源定理一章，我们认为复杂的截面状态数据，是在某时刻观察确定的动态系统的输出状态得到的。而该系统输入的是均匀的随机信号。也就是，我们将宏观现象建模为确定的复杂系统对微观的随机输入信号的变换。我们希望连接微观随机性和宏观确定性，从静态截面状态数据反推动态演化机制。进一步，如何将描述系统的有语义的协变量融入动力学建模中，进一步加强动力学模型的准确性和可解释性，仍然值得探索。例如在医学研究中，协变量如年龄，性别，收入等社会经济状态及生活方式等对病理动力学演化至关重要，且与本文中探究的生存分析，尤其是半参数化生存分析模型密切相关。第七章中我们通过刻画简单的动力学产生机制，产生和拟合了复杂数据的分布。我们的思路是：与其直接建模复杂的输出现象，通过建模相对简单的动力学生成机制，可以极大地简化

模型复杂度。这也是动力学建模理解复杂社交系统机制的一个优势所在。我们可以沿着所提出的框架思路，给出更多模型，如刻画长尾混合模型等等的模型，来产生和刻画更复杂的输出数据。

总之，解释复杂社交系统演化机制需要计算机科学与物理学理论的进一步交叉融合，并需要进一步面向强应用。我们提出的数据驱动的动力学建模研究方法，试图为建模，理解，预测真实大规模复杂社交系统奠定了一定的理论基础。

插图索引

表格索引

公式索引

参考文献

[1] Mahajan V, Muller E, Bass F M. New product diffusion models in marketing: A review and directions for research[J]. The journal of marketing, 1990:1-26.

[2] Barabási A L, Albert R. Emergence of scaling in random networks[J]. science, 1999, 286 (5439):509-512.

[3] Zang C, Cui P, Faloutsos C. Beyond sigmoids: The nettide model for social network growth, and its applications[C]//KDD '16: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2016: 2015-2024.

[4] Zang C, Cui P, Faloutsos C, et al. Long short memory process: Modeling growth dynamics of microscopic social connectivity[C/OL]//KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2017: 565-574. http://doi.acm.org/10.1145/3097983.3098055.

[5] Zhang T, Cui P, Faloutsos C, et al. Come-and-go patterns of group evolution: A dynamic model[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.: s.n.], 2016: 1355-1364.

[6] Zang C, Cui P, Song C, et al. Structural patterns of information cascades and their implications for dynamics and semantics[J]. arXiv preprint arXiv:1708.02377, 2017.

[7] Zang C, Cui P, Song C, et al. Quantifying structural patterns of information cascades[C]// WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion. [S.l.]: International World Wide Web Conferences Steering Committee, 2017: 867-868.

[8] Yu L, Cui P, Wang F, et al. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics[C/OL]//IEEE International Conference on Data Mining, ICDM. 2015. http://dx.doi.org/10.1109/ICDM.2015.79.

[9] Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: Densification and shrinking diameters [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1):2.

[10] Kumar R, Novak J, Tomkins A. Structure and evolution of online social networks[C]//New York, NY, USA: ACM, 2006: 611-617.

[11] Leskovec J, Backstrom L, Kumar R, et al. Microscopic evolution of social networks[C]//[S.l.]: ACM, 2008: 462-470.

[12] Mislove A, Koppula H S, Gummadi K P, et al. Growth of the flickr social network[C]// Proceedings of the first workshop on Online social networks. [S.l.]: ACM, 2008: 25-30.

[13] Kossinets G, Watts D J. Empirical analysis of an evolving social network[J]. Science, 2006, 311(5757):88-90.

[14] Barabâsi A L, Jeong H, Néda Z, et al. Evolution of the social network of scientific collaborations [J]. Physica A: Statistical mechanics and its applications, 2002, 311(3):590-614.

[15] Bianconi G, Barabási A L. Competition and multiscaling in evolving networks[J]. EPL (Europhysics Letters), 2001, 54(4):436.

[16] Anderson R M, May R M, Anderson B. Infectious diseases of humans: dynamics and control: volume 28[M]. [S.l.]: Wiley Online Library, 1992

[17] arxiv hep-ph network dataset – KONECT[EB/OL]. 2015. http://konect.uni-koblenz.de/networks/ca-cit-HepPh.

[18] Klimt B, Yang Y. The Enron corpus: A new dataset for email classification research[C]//Proc. European Conf. on Machine Learning. [S.l.: s.n.], 2004: 217-226.

[19] Yu L, Cui P, Wang F, et al. Uncovering and predicting the dynamic process of information cascades with survival model[J]. Knowledge and Information Systems, 2016:1-27.

[20] Jiang M, Cui P, Faloutsos C. Suspicious behavior detection: current trends and future directions [J]. Special Issue on Online Behavioral Analysis and Modeling, IEEE Intelligent Systems Magazine, 2016.

[21] Li L, Prakash B A, Faloutsos C. Parsimonious linear fingerprinting for time series[J]. Proceedings of the VLDB Endowment, 2010, 3(1-2):385-396.

[22] Matsubara Y, Sakurai Y, Faloutsos C. Autoplait: Automatic mining of co-evolving time sequences[C]//Proceedings of the 2014 ACM SIGMOD. [S.l.]: ACM, 2014: 193-204.

[23] Huberman B A, Adamic L A. Internet: growth dynamics of the world-wide web[J]. Nature, 1999, 401(6749):131-131.

[24] Kumar R, Raghavan P, Rajagopalan S, et al. Stochastic models for the web graph[C]// Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on. [S.l.]: IEEE, 2000: 57-65.

[25] Berlingerio M, Bonchi F, Bringmann B, et al. Mining graph evolution rules[M]//Machine learning and knowledge discovery in databases. [S.l.]: Springer, 2009: 115-130

[26] Gonzalez R, Cuevas R, Motamedi R, et al. Assessing the evolution of google+ in its first two years[J]. Biological Cybernetics, 2016, 24(3):1813-1826.

[27] Albert R, Barabási A L. Topology of evolving networks: local events and universality[J]. Physical review letters, 2000, 85(24):5234.

[28] Ghoshal G, Chi L, Barabási A L. Uncovering the role of elementary processes in network evolution[J]. Scientific reports, 2013, 3.

[29] Weng L, Ratkiewicz J, Perra N, et al. The role of information diffusion in the evolution of social networks[C]//Proceedings of the 19th ACM SIGKDD'13. [S.l.]: ACM, 2013: 356-364.

[30] Antoniades D, Dovrolis C. Co-evolutionary dynamics in social networks: A case study of twitter[J]. Computational Social Networks, 2015, 2(1):1-21.

[31] Farajtabar M, Wang Y, Rodriguez M, et al. Coevolve: A joint point process model for information diffusion and network co-evolution[C]//Advances in Neural Information Processing Systems. [S.l.: s.n.], 2015: 1945-1953.

[32] Rozenfeld H D, Rybski D, Andrade J S, et al. Laws of population growth[J]. Proceedings of the National Academy of Sciences, 2008, 105(48):18702-18707.

[33] Zhu K, Li W, Fu X, et al. How do online social networks grow?[J]. PloS one, 2014, 9(6): e100023.

[34] Banks R B. Growth and diffusion phenomena: mathematical frameworks and applications: volume 14[M]. [S.l.]: Springer Science & Business Media, 1994

[35] Barabasi A L. The origin of bursts and heavy tails in human dynamics[J]. Nature, 2005, 435 (7039):207-211.

[36] Oliveira J G, Barabási A L. Human dynamics: Darwin and einstein correspondence patterns [J]. Nature, 2005, 437(7063):1251-1251.

[37] Lin Y, Raza A A, Lee J Y, et al. Influence propagation: Patterns, model and a case study[M]// Advances in Knowledge Discovery and Data Mining. [S.l.]: Springer, 2014: 386-397

[38] Figueiredo F, Almeida J M, Matsubara Y, et al. Revisit behavior in social media: The phoenix-r model and discoveries[M]//Machine Learning and Knowledge Discovery in Databases. [S.l.]: Springer, 2014: 386-401

[39] Ribeiro B. Modeling and predicting the growth and death of membership-based websites[C]// Proceedings of the 23rd international conference on World Wide Web. [S.l.]: ACM, 2014: 653-664.

[40] Hawkes A G. Spectra of some self-exciting and mutually exciting point processes[J]. Biometrika, 1971, 58(1):83-90.

[41] Laub P J, Taimre T, Pollett P K. Hawkes processes[J]. arXiv preprint arXiv:1507.02822, 2015.

[42] Crane R, Sornette D. Robust dynamic classes revealed by measuring the response function of a social system[J]. Proceedings of the National Academy of Sciences, 2008, 105(41): 15649-15653.

[43] Matsubara Y, Sakurai Y, Prakash B A, et al. Rise and fall patterns of information diffusion: model and implications[C]//[S.l.]: ACM, 2012: 6-14.

[44] Vázquez A, Oliveira J G, Dezsö Z, et al. Modeling bursts and heavy tails in human dynamics [J]. Physical Review E, 2006, 73(3):036127.

[45] Malmgren R D, Stouffer D B, Motter A E, et al. A poissonian explanation for heavy tails in e-mail communication[J]. Proceedings of the National Academy of Sciences, 2008, 105(47): 18153-18158.

[46] Zhang T, Cui P, Song C, et al. A multiscale survival process for modeling human activity patterns[J]. PloS one, 2016, 11(3):e0151473.

[47] Wu Y, Zhou C, Xiao J, et al. Evidence for a bimodal distribution in human communication[J]. Proceedings of the national academy of sciences, 2010, 107(44):18803-18808.

[48] Leskovec J, Chakrabarti D, Kleinberg J, et al. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication[M]//Knowledge Discovery in Databases: PKDD 2005. [S.l.]: Springer, 2005: 133-145

[49] Marquardt D W. An algorithm for least-squares estimation of nonlinear parameters[J]. Journal of the Society for Industrial & Applied Mathematics, 1963, 11(2):431-441.

[50] Arnaboldi V, Conti M, Passarella A, et al. Dynamics of personal social relationships in online social networks: a study on twitter[C]//Proceedings of the first ACM conference on Online social networks. [S.l.]: ACM, 2013: 15-26.

[51] Jo H H, Perotti J I, Kaski K, et al. Correlated bursts and the role of memory range[J]. Physical Review E, 2015.

[52] Ozaki T. Maximum likelihood estimation of hawkes' self-exciting point processes[J]. Annals of the Institute of Statistical Mathematics, 1979, 31(1):145-155.

[53] Vaz de Melo P O S, Faloutsos C, Assunção R, et al. The self-feeding process: a unifying model for communication dynamics in the web[C]//Proceedings of the 22nd international conference on World Wide Web. [S.l.]: ACM, 2013: 1319-1330.

[54] Ferraz Costa A, Yamaguchi Y, Juci Machado Traina A, et al. Rsc: Mining and modeling temporal activity in social media[C]//SIGKDD'15. [S.l.]: ACM, 2015.

[55] Alves R A d S, Assuncao R M, Vaz de Melo P O S. Burstiness scale: A parsimonious model for characterizing random series of events[C]//Proceedings of the 22nd ACM SIGKDD. [S.l.]: ACM, 2016: 1405-1414.

[56] Daley D J, Vere-Jones D. An introduction to the theory of point processes: volume i: Elementary theory and methods[M]. [S.l.]: Springer Science & Business Media, 2003

[57] Coleman T F, Li Y. An interior trust region approach for nonlinear minimization subject to bounds[J]. SIAM Journal on optimization, 1996.

[58] Ogata Y. On lewis' simulation method for point processes[J]. IEEE Transactions on Information Theory, 1981, 27(1):23-31.

[59] Dassios A, Zhao H, et al. Exact simulation of hawkes process with exponentially decaying intensity[J]. Electronic Communications in Probability, 2013, 18(62):1-13.

[60] Peacock J. Two-dimensional goodness-of-fit testing in astronomy[J]. Monthly Notices of the Royal Astronomical Society, 1983, 202(3):615-627.

[61] Moré J J. The levenberg-marquardt algorithm: implementation and theory[M]//Numerical analysis. [S.l.]: Springer, 1978: 105-116

[62] Zang C, Cui P, Song C, et al. Quantifying structural patterns of information cascades[C]// Proceedings of the 26th International Conference on WWW Companion. [S.l.: s.n.], 2017: 867-868.

[63] Nicosia V, Vértes P E, Schafer W R, et al. Phase transition in the economically modeled growth of a cellular nervous system[J]. PNAS, 2013, 110(19):7880-7885.

[64] Koch A, Meinhardt H. Biological pattern formation: from basic mechanisms to complex structures[J]. Reviews of Modern Physics, 1994, 66(4):1481.

[65] Ball P. Shapes: nature's patterns: a tapestry in three parts[M]. [S.l.]: OUP Oxford, 2009

[66] Ben-Jacob E, Godbey R, Goldenfeld N D, et al. Experimental demonstration of the role of anisotropy in interfacial pattern formation[J]. Physical review letters, 1985, 55(12):1315.

[67] Kessler D A, Koplik J, Levine H. Pattern selection in fingered growth phenomena[J]. Advances in Physics, 1988, 37(3):255-339.

[68] Makse H A, Havlin S, Ivanov P C, et al. Pattern formation in sedimentary rocks: connectivity, permeability, and spatial correlations[J]. Physica A: Statistical Mechanics and its Applications, 1996, 233(3):587-605.

[69] Angstmann C N, Donnelly I C, Henry B I. Pattern formation on networks with reactions: A continuous-time random-walk approach[J]. Physical Review E, 2013, 87(3):032804.

[70] Nakamasu A, Takahashi G, Kanbe A, et al. Interactions between zebrafish pigment cells responsible for the generation of turing patterns[J]. Proceedings of the National Academy of Sciences, 2009, 106(21):8429-8434.

[71] Kondo S, Miura T. Reaction-diffusion model as a framework for understanding biological pattern formation[J]. science, 2010, 329(5999):1616-1620.

[72] Turing A M. The chemical basis of morphogenesis[J]. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 1952, 237(641):37-72.

[73] Wachter K W, Hammel E A, Laslett P. Statistical studies of historical social structure[M]. [S.l.]: Elsevier, 2013

[74] Easley D, Kleinberg J. Networks, crowds, and markets: Reasoning about a highly connected world[M]. [S.l.]: Cambridge University Press, 2010

[75] Ahn Y Y, Bagrow J P, Lehmann S. Link communities reveal multiscale complexity in networks [J]. Nature, 2010, 466(7307):761-764.

[76] Jiang Z Q, Xie W J, Li M X, et al. Calling patterns in human communication dynamics[J]. Proceedings of the National Academy of Sciences, 2013, 110(5):1600-1605.

[77] Sekara V, Stopczynski A, Lehmann S. Fundamental structures of dynamic social networks[J]. Proceedings of the National Academy of Sciences, 2016, 113(36):9977-9982.

[78] Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using internet chain-letter data[J]. Proceedings of the National Academy of Sciences, 2008, 105(12):4633-4638.

[79] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena[J]. Science, 2013, 342(6164):1337-1342.

[80] Centola D. The spread of behavior in an online social network experiment[J]. science, 2010, 329(5996):1194-1197.

[81] Ugander J, Backstrom L, Marlow C, et al. Structural diversity in social contagion[J]. Proceedings of the National Academy of Sciences, 2012:201116502.

[82] Castro R, Coates M, Liang G, et al. Network tomography: recent developments[J]. Statistical science, 2004:499-517.

[83] Leskovec J, McGlohon M, Faloutsos C, et al. Patterns of cascading behavior in large blog graphs.[C]//SDM: volume 7. [S.l.]: SIAM, 2007: 551-556.

[84] Goel S, Watts D J, Goldstein D G. The structure of online diffusion networks[C]//Proceedings of the 13th ACM conference on electronic commerce. [S.l.]: ACM, 2012: 623-638.

[85] Goel S, Anderson A, Hofman J, et al. The structural virality of online diffusion[J]. Preprint, 2013, 22:26.

[86] Pei S, Muchnik L, Tang S, et al. Exploring the complex pattern of information spreading in online blog communities[J]. PloS one, 2015, 10(5):e0126894.

[87] Hethcote H W. The mathematics of infectious diseases[J]. SIAM review, 2000, 42(4):599-653.

[88] Pastor-Satorras R, Vespignani A. Epidemic spreading in scale-free networks[J]. Physical review letters, 2001, 86(14):3200.

[89] Goltsev A V, Dorogovtsev S N, Oliveira J, et al. Localization and spreading of diseases in complex networks[J]. Physical review letters, 2012, 109(12):128702.

[90] Pastor-Satorras R, Castellano C, Van Mieghem P, et al. Epidemic processes in complex networks [J]. Reviews of modern physics, 2015, 87(3):925.

[91] Harris T E. The theory of branching processes[M]. [S.l.]: Courier Dover Publications, 2002

[92] Kumar R, Mahdian M, McGlohon M. Dynamics of conversations[C]//Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. [S.l.]: ACM, 2010: 553-562.

[93] Wang D, Wen Z, Tong H, et al. Information spreading in context[C]//Proceedings of the 20th international conference on World wide web. [S.l.]: ACM, 2011: 735-744.

[94] Bass F. A new product growth model for consumer durables," management science, 15 (january), 215-227.(1980)[J]. The Relationship Between Diffusion Rates, Experience Curves, and Demand Elasticities for Consumer Durables Technical Innovations," Journal of Business, 1969, 53:51-67.

[95] Rogers E M. Diffusion of innovations[M]. [S.l.]: Simon and Schuster, 2010

[96] Bass F M. Comments on "a new product growth for model consumer durables the bass model" [J]. Management science, 2004, 50(12_supplement):1833-1840.

[97] Granovetter M. Threshold models of collective behavior[J]. American journal of sociology, 1978:1420-1443.

[98] Watts D J. A simple model of global cascades on random networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(9):5766-5771.

[99] Dow P A, Adamic L A, Friggeri A. The anatomy of large facebook cascades.[C]//ICWSM. [S.l.: s.n.], 2013.

[100] Del Vicario M, Bessi A, Zollo F, et al. The spreading of misinformation online[J]. Proceedings of the National Academy of Sciences, 2016, 113(3):554-559.

[101] Kwak H, Lee C, Park H, et al. What is twitter, a social network or a news media?[C]//Proceedings of the 19th international conference on World wide web. [S.l.]: ACM, 2010: 591-600.

[102] Romero D M, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter[C]//Proceedings of the 20th international conference on World wide web. [S.l.]: ACM, 2011: 695-704.

[103] Blondel V D, Guillaume J L, Lambiotte R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10):P10008.

[104] Moore C, Newman M E. Epidemics and percolation in small-world networks[J]. Physical Review E, 2000, 61(5):5678.

[105] Newman M E. Spread of epidemic disease on networks[J]. Physical review E, 2002, 66(1): 016128.

[106] Keeling M J, Eames K T. Networks and epidemic models[J]. Journal of the Royal Society Interface, 2005, 2(4):295-307.

[107] Bernardes D F, Latapy M, Tarissan F. Relevance of sir model for real-world spreading phenomena: Experiments on a large-scale p2p system[C]//Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). [S.l.]: IEEE Computer Society, 2012: 327-334.

[108] Centola D, Macy M. Complex contagions and the weakness of long ties1[J]. American Journal of Sociology, 2007, 113(3):702-734.

[109] Dodds P S, Watts D J. A generalized model of social and biological contagion[J]. Journal of Theoretical Biology, 2005, 232(4):587-604.

[110] Golub B, Jackson M O. Using selection bias to explain the observed structure of internet diffusions[J]. Proceedings of the National Academy of Sciences, 2010, 107(24):10833-10836.

[111] Goyal A, Bonchi F, Lakshmanan L V. Learning influence probabilities in social networks[C]// Proceedings of the third ACM international conference on Web search and data mining. [S.l.]: ACM, 2010: 241-250.

[112] Good B H, McDonald M J, Barrick J E, et al. The dynamics of molecular evolution over 60,000 generations[J]. Nature, 2017, 551(7678):45.

[113] Newberry M G, Ahern C A, Clark R, et al. Detecting evolutionary forces in language change [J]. Nature, 2017, 551(7679):223.

[114] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology[C]// ACM SIGCOMM computer communication review: volume 29. [S.l.]: ACM, 1999: 251-262.

[115] Yoshida T, Jones L E, Ellner S P, et al. Rapid evolution drives ecological dynamics in a predator–prey system[J]. Nature, 2003, 424(6946):303-306.

[116] Sinatra R, Wang D, Deville P, et al. Quantifying the evolution of individual scientific impact [J]. Science, 2016, 354(6312):aaf5239.

[117] Wang D, Song C, Barabási A L. Quantifying long-term scientific impact[J]. Science, 2013, 342(6154):127-132.

[118] Zha Y, Zhou T, Zhou C. Unfolding large-scale online collaborative human dynamics[J]. Proceedings of the National Academy of Sciences, 2016, 113(51):14627-14632.

[119] Pierson E, Koh P W, Hashimoto T, et al. Inferring multi-dimensional rates of aging from cross-sectional data[J]. arXiv preprint arXiv:1807.04709, 2018.

[120] Zang C, Cui P, Faloutsos C. Beyond sigmoids: The nettide model for social network growth, and its applications[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2016: 2015-2024.

[121] Wang J, Tsang W W, Marsaglia G. Evaluating kolmogorov's distribution[J]. Journal of Statistical Software, 2003, 8(18).

[122] Newman M E. Power laws, pareto distributions and zipf's law[J]. Contemporary physics, 2005, 46(5):323-351.

[123] Clauset A, Young M, Gleditsch K S. On the frequency of severe terrorist events[J]. Journal of Conflict Resolution, 2007, 51(1):58-87.

[124] Zang C, Cui P, Faloutsos C, et al. Long short memory process: Modeling growth dynamics of microscopic social connectivity[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2017: 565-574.

[125] Aalen O, Borgan O, Gjessing H. Survival and event history analysis: a process point of view [M]. [S.l.]: Springer Science & Business Media, 2008

[126] Devroye L. Sample-based non-uniform random variate generation[C]//Proceedings of the 18th conference on Winter simulation. [S.l.]: ACM, 1986: 260-265.

[127] Byrd R H, Gilbert J C, Nocedal J. A trust region method based on interior point techniques for nonlinear programming[J]. Mathematical Programming, 2000, 89(1):149-185.

[128] Lourakis M I. A brief description of the levenberg-marquardt algorithm implemented by levmar [J]. Foundation of Research and Technology, 2005, 4(1).

[129] Clauset A, Shalizi C R, Newman M E. Power-law distributions in empirical data[J]. SIAM review, 2009, 51(4):661-703.

[130] Reynolds D. Gaussian mixture models[J]. Encyclopedia of biometrics, 2015:827-832.

[131] Murphy K P. Machine learning, a probabilistic perspective[M]. [S.l.]: Taylor & Francis, 2014.

[132] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. [S.l.: s.n.], 2014: 2672-2680.

[133] Guo L, Tan E, Chen S, et al. The stretched exponential distribution of internet media access patterns[C]//Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing. [S.l.]: ACM, 2008: 283-294.

[134] Zaheer M, Li C L, Póczos B, et al. Gan connoisseur: Can gans learn simple 1d parametric distributions?[J]. 2017.

[135] West G. Scale: The universal laws of growth, innovation, sustainability and the pace of life in organisms and companies[M]. [S.l.]: London: Penguin Random House/Orion, 2017.

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

[137] Broido A D, Clauset A. Scale-free networks are rare[J]. arXiv preprint arXiv:1801.03400, 2018.

[138] Nolan J. Stable distributions: models for heavy-tailed data[M]. [S.l.]: Birkhauser New York, 2003

[139] Zhang T, Cui P, Faloutsos C, et al. Come-and-go patterns of group evolution: A dynamic model[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2016: 1355-1364.

[140] Zang C, Cui P, Faloutsos C, et al. On power law growth of social networks[J]. IEEE Transactions on Knowledge and Data Engineering, 2018.

[141] Smith F A, Lyons S K, Ernest S, et al. Body mass of late quaternary mammals[J]. Ecology, 2003, 84(12):3403-3403.

致 谢

衷心感谢我的导师清华大学朱文武教授和崔鹏副教授对本人多年来的精心指导。感谢朱老师的知遇之恩，在五年多来对我的细心指导，指方向，解疑惑，时时处处春风化雨的督促和包容。感谢两位老师带我走上了学术的道路。还记得二零一二年第一次来到 FIT 楼实验室的场景，有幸成为朱老师的第一批学生，时光如白驹过隙，实验室发展壮大如此迅速！虽然我即将开始一段新的征程，但我永远也不会忘记梦开始的地方。再次感谢两位老师多年来对我的包容和细心指导！

衷心感谢指导与帮助过我的清华大学杨士强教授，迈阿密大学的 Chaoming Song 副教授，CMU 的 Christos Faloutsos 教授，Northeastern 大学的 Albert-László Barabási 教授，Cornell 医学院的 Fei Wang 副教授，Cleveland Clinic 的 Feixiong Cheng 副研究员，Harvard Medical School 的 Yang-Yu Liu 副教授，Rensselaer Polytechnic Institute 的 Jianxi Gao 副教授。他们的言传身教将使我终生受益。

感谢我的母亲权晋丽，父亲臧和平，妻子张琮卉。正是我的家人，让我坚持在梦想的道路上。

感谢清华大学计算机系-媒体所-媒体与网络实验室全体老师和同学们的热情帮助和支持！本课题承蒙国家自然科学基金等基金的资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间发表的学术论文与研究成果

个人简历

1989 年 10 月 26 日出生于山西省太原市。

2009 年 9 月考入中山大学信息科学系计算机专业，2013 年 7 月本科毕业并获得理学学士学位。

2013 年 9 月免试进入清华大学大学计算机系攻读工学博士学位至今。

更多信息，请点击个人主页：

www.calvinzang.com

发表的学术论文

[1] Chengxi Zang, Peng Cui, Wenwu Zhu, and Fei Wang: Dynamical Origins of Distribution Functions. ACM Conference on Knowledge Discovery and Data Mining (KDD), 2019.

[2] Chengxi Zang, Peng Cui, Chaoming Song, Wenwu Zhu, and Fei Wang: Uncovering Pattern Formation of Information Flow. ACM Conference on Knowledge Discovery and Data Mining (KDD), 2019.

[3] Haoyang Li, Peng Cui, Chengxi Zang, Tianyang Zhang, Wenwu Zhu, and Yishi Lin: Fates of Microscopic Social Ecosystems: Keep Alive or Dead? ACM Conference on Knowledge Discovery and Data Mining (KDD), 2019.

[4] Yunfei Lu, Lingyun Yu, Peng Cui, Chengxi Zang, Renzhe Xu, Yihao Liu, Lei Li, and Wenwu Zhu: Uncovering the Co-driven Mechanism of Social and Content Links in User Churn Phenomena. ACM Conference on Knowledge Discovery and Data Mining (KDD), 2019.

[5] Zhixiu Liu, Chengxi Zang, Kun Kunag, Hao Zou, Hu Zheng, and Peng Cui: Causation-Driven Visualizations for Insurance Recommendation. IEEE International Conference on Multimedia & Expo (ICME), 2019, the Cross-media Analysis for Semantic Knowledge Reasoning and Transfer Workshop.

[6] Chengxi Zang, Peng Cui, Christos Faloutsos, Wenwu Zhu: On Power Law Growth of Social Networks. IEEE Transactions on Knowledge and Data Engineering (TKDE) 30(9): 1727-1740 (2018)

[7] Chengxi Zang, Peng Cui, Wenwu Zhu: Learning and Interpreting Complex Distri-

butions in Empirical Data. ACM Conference on Knowledge Discovery and Data Mining (KDD) 2018: 2682-2691

[8] Yunfei Lu, Linyun Yu, Tianyang Zhang, Chengxi Zang, Peng Cui, Chaoming Song, and Wenwu Zhu: Collective Human Behavior in Cascading System: Discovery, Modeling and Applications. IEEE International Conference on Data Mining (ICDM) 2018 –November 17-20, 2018 in Singapore

[9] Chengxi Zang, Peng Cui, Christos Faloutsos, Wenwu Zhu: Long Short Memory Process: Modeling Growth Dynamics of Microscopic Social Connectivity. ACM Conference on Knowledge Discovery and Data Mining (KDD) 2017: 565-574

[10] Chengxi Zang, Peng Cui, Chaoming Song, Christos Faloutsos, Wenwu Zhu: Quantifying Structural Patterns of Information Cascades. International World Wide Web Conference (WWW Companion Volume) 2017: 867-868

[11] Chengxi Zang, Peng Cui, Christos Faloutsos: Beyond Sigmoids: The NetTide Model for Social Network Growth, and Its Applications. ACM Conference on Knowledge Discovery and Data Mining (KDD) 2016: 2015-2024

综合论文训练记录表

学生姓名		学号		班级	
论文题目					
主要内容以及进度安排	指导教师签字： 考核组组长签字： 年　　月　　日				
中期考核意见	考核组组长签字： 年　　月　　日				

指导教师评语	指导教师签字：_____ 年　　月　　日
评阅教师评语	评阅教师签字：_____ 年　　月　　日
答辩小组评语	答辩小组组长签字：_____ 年　　月　　日

总成绩：_____

教学负责人签字：_____

年　　月　　日