



0000294

申请上海交通大学博士学位论文

利用移动网络数据的人类时空行为分析及建模研究

论文作者 陈夏明

学 号 010349025

导 师 金耀辉 教授

专 业 信息与通信工程

答辩日期 2016 年 08 月 22 日



0000294



0000294

Submitted in total fulfillment of the requirements for the degree of Doctor  
in Department of Information and Communication Engineering

# Analyzing and Modeling Human Spatio-Temporal Behaviors with Mobile Data

XIAMING CHEN

Advisor

Prof. YAOHUI JIN

SCHOOL OF ELECTRONIC INFORMATION AND ELECTRICAL ENGINEERING

SHANGHAI JIAO TONG UNIVERSITY

SHANGHAI, P.R.CHINA

Aug. 22, 2016



0000294



0000294

## 上海交通大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：\_\_\_\_\_

日期：\_\_\_\_\_年\_\_\_\_月\_\_\_\_日



0000294



0000294

## 上海交通大学 学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

保 密 ，在 \_\_\_\_\_ 年解密后适用本授权书。  
不保密 .

(请在以上方框内打√)

学位论文作者签名: \_\_\_\_\_

指导教师签名: \_\_\_\_\_

日期: \_\_\_\_\_ 年 \_\_\_\_\_ 月 \_\_\_\_\_ 日

日期: \_\_\_\_\_ 年 \_\_\_\_\_ 月 \_\_\_\_\_ 日



0000294



0000294

# 利用移动网络数据的人类时空行为分析及建模研究

## 摘要

随着移动技术的快速发展，人类通信的方式向便携式和多样化的方向发展，种类丰富的移动应用在满足了人们通信、休闲、社交等日常需求的同时，也积累了海量的移动网络数据。这些数据有的直接记录了用户的移动位置和服务使用行为，有的间接记录了行为发生时的周边环境和社交关系。本研究利用被动采集的方法，收集了不同空间尺度下的移动网络数据，并从时间和空间两个基本维度出发，对个体和群体粒度上的人类时空行为模式和规律进行分析和建模研究。本研究中提出的系统化的量化、分析和建模方法，不仅在疾病传播、城市管理、移动网络优化等方面具有直接的应用价值，而且对揭示人类时空行为的统计规律、网络科学、以及行为科学等理论研究具有一定的贡献。具体来讲，该研究的成果可归纳为以下四个方面：

本文采集并收集了三种不同空间尺度下的移动网络数据，包括校园 WiFi 网络、城市和国家移动网络，并基于此提出了时空数据质量的客观评估和提升方法。首先，从数据记录的准确性、采集时间的连续性、以及空间分布的合理性出发，本文提出一种结合时空数据点局部特征信息、和用户轨迹全局特征的数据质量量化方法。在该方法中，局部质量信息从单个数据点的动态和静态两个层面进行刻画，突破了传统方法中单纯对动态特征（如移动速度）的依赖，因此对数据质量的刻画更加准确。进一步，用户的移动轨迹特征代表了全局的数据质量信息，我们基于轨迹中连续采集点的时空分布异质性，并结合单个数据点的平局质量水平，从而克服了传统信息熵的方法误差较大的缺陷。

本文对传统个体微观模式和宏观统计模式进行了扩展，提出了介观行为模式的概念，并对介观模式的提取算法、实证分析、以及新型个体移动模型进行了系统性研究。首先，从时空耦合的角度出发，提出了个体介观时空行为模式的概念，这种模式不仅保留了行为序列中的基本特征，而且方便从时空结构角度对移动行为进行挖掘。为了从大量轨迹记录中得到介观模式，我们提出了拓扑和属性结合的图相似匹配算法，结合群组



0000294

内不同个体行为之间的相似性，提取出了属于不同群组的显著介观行为模式，并和传统的移动模序分析进行了比较研究。基于得到的介观时空行为模式，提出了一种鲁棒性更好的个体移动性模型，即流涌现模型。由于该模型建立在机遇资源的空间分布和干扰机遇的框架之上，从而摒弃了传统模型中微观和宏观统计一致的假设，为连接微观移动模式挖掘和宏观统计分析提供了基础。

通常人类时空行为的规律性，既表现在个体粒度上的地点偏好，也表现在群体粒度上的“潮汐效应”。由于介观模式体现了个体行为的时空特征，因此我们进一步对群体行为的时空关联关系进行了分析和建模研究。首先，我们利用协方差方程对群体的时空依赖关系进行描述，分别从时间和空间维度对群体行为的统计特征进行度量。其次，由于传统的群体时空行为研究基于单一的数据源，所得出的结论往往在另一个数据集上难以复现；因此本文采集了多空间尺度（包括校园、城市、国家）下的群体移动行为数据，对群体的时空行为规律在空间上作横向和纵向比较，从而使群体行为的分析结论更加可靠。最后，基于所观测到的群体时空关联关系，在考虑空间不同区域差异性的前提下，本文提出了基于盖内特方程的群体行为模型，并利用城市尺度下的人群分布预测对模型性能进行了验证和分析。

本文将物理空间的移动行为和网络空间的参与行为在形式上进行了统一，利用服务类型序列代替空间位置序列，对移动用户的参与行为模式进行挖掘。首先，针对移动用户参与网络服务的时空行为，提出一种基于被动测量的行为识别算法。通过与客户端采集的基准数据进行比较，该算法能够在大规模的用户网络行为监测中表现出较好的性能。其次，通过量化用户参与行为的重要指标，建立了参与行为和底层网络性能之间的联系，展示了不同硬件平台下用户参与行为随网络性能变化的模式。利用结构相关性分析的方法，对场景因素（如用户个性、应用类型、地点熟悉度等）影响用户参与行为的现象进行了细粒度的量化分析。最后，基于对参与行为时空特性的研究，提出了利用隐马尔可夫过程的参与行为建模，并对群体参与行为进行了聚类分析。

**关键词：**时空数据挖掘，移动行为，网络参与行为，时空关联，行为模型



0000294

# Analyzing and Modeling Human Spatio-Temporal Behaviors with Mobile Data

## ABSTRACT

With the evolution of wireless technologies, a bunch of mobile applications accumulate massive amount of data which conveys plentiful information about social behavior, use habits or personal preferences etc. Some of these datasets record directly the coordinates of whereabouts and engaging behavior, while others encode indirectly physical surroundings and social connections. By employing a passive collection technique, we in this paper obtain multi-source mobile network data under varying spatial granularities, and then perform analyses and modelling on the patterns and rules in human spatio-temporal behavior, which considers behavioral properties from both individual and group aspects simultaneously. The systematical approaches for quantification, analyzing and modelling of human behavior have potential values in multiple R&D areas, e.g. the epidemic prediction, urban management, and mobile network optimization. Meanwhile, our researches also make contributions for the theoretical development of behavioral and network sciences. Specifically, we summarize our achievements as following:

We propose a novel framework to evaluate the quality of spatio-temporal datasets, as well as a quality-enhancing algorithm for human mobility. This evaluation method focuses on the common form and properties in spatio-temporal data, by considering objective metrics for data quality from single data point, individual and group perspectives, respectively. For a single data point, we quantify the data quality with the static spatio-temporal resolution and dynamic transitions between successive records. For an individual's trajectory, we compress the quality of multiple data points into a single metric by calculating the heterogeneity of a sequence of spatio-temporal observations. This method shows remarkable performance efficiency when comparing



0000294

with traditional entropy metric. For trajectories of a group of people, we involve space splitting to combine the correlation between different blocks and feature distribution within the same block. We finally apply our evaluation on the real datasets, and propose a quality-enhancing method for human mobility data, which shows a great performance improvement when comparing with existing spatial or temporal interpolations.

We have studied the unified spatio-temporal mesostructures in human mobility. To discover the mesostructures from cellular data, we first introduce a topology-attributes coupling similarity algorithm to derive the elementary (nodes and edges) similarities for two attributed graphs. With the construction of individual profiles from mesostructure analyses, we provided a novel mobility model from a process-driven perspective, which reduced the dependence of many existing models on the consistency between local and global mobility statistics. We gained some insights on the dominating mesostructures in human mobility by leveraging mobile data in a large city. The statistical distribution of mesostructures is found to be determined by the intrinsic heterogeneity of spatio-temporal properties in human behavior. Our model evaluation showed that a process with basic rules could demonstrate the key statistical properties in mobility mesostructures. We believe that these approaches and observations would be a good reference for management of human mobility in mobile networks and transportation systems.

Beyond the spatio-temporal dependence in individual's mobility as revealed by mesostructures, we also investigate the dependence at population level. Despite recent progress in revealing temporal dynamics and spatial inhomogeneity of group mobility, limited knowledge about spatio-temporal dependence is gained. One of challenges comes from the absence of sustained observations at varying spatial scales. We characterize the group dynamics with correlation functions and measures the spatial and temporal properties statistically. Different from previous observations with single data source, we compare the group mobility dynamics with three varying granularities, i.e., campus, city and country. We eventually model the spatio-temporal dependence, whose evaluation results suggest connections between spatio-temporal dependence of group mobility and the organization of human lives. Region differences and spatial scales are observed to impact spatio-temporal dependence to a great extent. Additionally, interactive



0000294

knowledge between space and time enhances population prediction with a decrease in root-mean-square error of 2.8%~25.2%. We believe that these achievements will benefit multiple research and development areas such as network deploying and simulation researches.

Although passive measurements of mobile traffic have been conducted in previous literature, they mostly address protocol and traffic properties, rather than responsive user behaviour consequences. In this paper, we perform a characterization of mobile traffic and engaging behaviours from end-user's view. The proposed concurrence index equipped by the model is more powerful to capture delicate difference of user-perceived application performance than previous volume-based metrics. Then we profile the behavioural dynamics of user participation in mobile usage and its interaction with user-perceived application performance. And finally, we perform a unique modelling of individual engaging trajectories and a model-based clustering to explore user behavioural patterns. We find that user engaging behaviour is primarily governed by a small portion of latent states, and the behavioural patterns regarding principle engaging states illustrate distinctive properties in discovered user clusters.

**KEY WORDS:** Spatio-temporal data mining, human mobility, mobile engaging behavior, spatio-temporal coupling, behavioral models



0000294



0000294

# 目 录

目录	x
插图索引	xiii
表格索引	xv
算法索引	xvii
主要符号对照表	xix
<b>第一章 绪论</b>	<b>1</b>
1.1 人类时空行为的研究背景 . . . . .	1
1.1.1 时空行为的定义 . . . . .	2
1.1.2 时空行为的研究价值 . . . . .	4
1.1.3 时空行为的研究趋势 . . . . .	6
1.2 人类时空行为挖掘的国内外研究进展 . . . . .	8
1.2.1 人类移动行为研究进展 . . . . .	8
1.2.2 用户参与行为研究进展 . . . . .	16
1.2.3 进展总结及分析 . . . . .	18
1.3 人类时空行为挖掘中的关键研究问题 . . . . .	18
1.3.1 面临的挑战 . . . . .	18
1.3.2 关键算法研究 . . . . .	21
1.3.3 关键分析与建模研究 . . . . .	22
1.4 本文的主要研究工作和创新点 . . . . .	24
1.5 本文的结构安排 . . . . .	27

<b>第二章 时空行为数据采集及质量管理</b>	<b>29</b>
2.1 移动网络中的时空行为数据采集 . . . . .	29
2.1.1 移动流量采集系统 . . . . .	29
2.1.2 时空行为数据分析平台 . . . . .	35
2.2 多空间尺度的时空行为数据集 . . . . .	37
2.3 时空数据的质量管理 . . . . .	41
2.3.1 时空数据质量的关键点 . . . . .	41
2.3.2 时空数据质量的量化评估与提升 . . . . .	42
2.3.3 移动网络数据的质量分析 . . . . .	48
2.4 本章小结 . . . . .	51
<b>第三章 个体移动行为的时空模式挖掘</b>	<b>53</b>
3.1 移动模式挖掘的方法介绍 . . . . .	53
3.2 个体移动行为的介观模式挖掘 . . . . .	56
3.2.1 拓扑-属性耦合的图相似算法 . . . . .	58
3.2.2 显著介观模式提取 . . . . .	63
3.2.3 介观模式数据分析 . . . . .	68
3.3 基于介观模式的个体移动模型 . . . . .	74
3.3.1 现有模型的时空分布假设冲突 . . . . .	74
3.3.2 基于介观模式的流涌现模型 . . . . .	75
3.3.3 模型性能验证 . . . . .	82
3.4 本章小结 . . . . .	87
<b>第四章 群体移动行为的时空分布研究</b>	<b>89</b>
4.1 群体行为的研究背景 . . . . .	89
4.2 群体行为的时空统计分析 . . . . .	93
4.2.1 群体行为的时空分布描述 . . . . .	93
4.2.2 群体行为的时空相关性 . . . . .	94
4.3 多空间尺度下的时空分布特征 . . . . .	96



0000294

4.3.1 空间分布特征 . . . . .	97
4.3.2 时间分布特征 . . . . .	103
4.4 群体行为的时空关联建模 . . . . .	107
4.4.1 无时空关联的行为模型 . . . . .	107
4.4.2 融合时空关联的行为模型 . . . . .	108
4.4.3 模型性能验证 . . . . .	109
4.4.4 数据分析及结果 . . . . .	111
4.5 本章小结 . . . . .	115
<b>第五章 用户参与行为的时空建模</b>	<b>117</b>
5.1 用户参与行为研究介绍 . . . . .	117
5.2 移动流量中的用户参与行为识别 . . . . .	119
5.3 用户参与行为的量化分析 . . . . .	123
5.3.1 用户参与行为的量化特征 . . . . .	123
5.3.2 用户参与行为的统计分析 . . . . .	127
5.3.3 用户参与行为的关联分析 . . . . .	130
5.4 用户参与轨迹的时空模型 . . . . .	136
5.4.1 用户参与行为建模 . . . . .	138
5.4.2 群体参与行为聚类 . . . . .	141
5.4.3 数据分析及验证 . . . . .	142
5.5 本章小结 . . . . .	146
<b>第六章 总结和展望</b>	<b>147</b>
6.1 工作总结 . . . . .	147
6.2 工作展望 . . . . .	148
<b>参考文献</b>	<b>151</b>
<b>致 谢</b>	<b>163</b>



0000294

攻读学位期间发表的学术论文 **165**

攻读学位期间申请的专利 **167**

攻读学位期间参与的项目 **169**



0000294

## 插图索引

1-1 全球移动网络流量和注册用户量趋势预测 . . . . .	2
1-2 2005 至 2015 年间发表的有影响力国际会议和期刊论文 . . . . .	7
1-3 基于数据分析流程的时空行为挖掘成果总结 . . . . .	19
1-4 本文主要研究内容及创新点概括 . . . . .	25
2-1 被动式的移动网络流量数据采集系统架构 . . . . .	30
2-2 网络流量特征实时提取工具 HTTP-SNIFFER 设计结构 . . . . .	34
2-3 基于 Lambda 架构的时空行为数据分析平台 . . . . .	36
2-4 WIFI-T 数据中的 TCP 性能参数和 HTTP 会话特征示例 . . . . .	38
2-5 原始系统日志和 WIFI-M 数据集示例 . . . . .	39
2-6 数据记录质量的计算和影响因素示意图 . . . . .	43
2-7 用户轨迹质量与数据采集点的时空非均衡性的关系 . . . . .	45
2-8 时空行为数据质量提升算法流程图 . . . . .	47
2-9 不同空间尺度下的时空数据质量对比 . . . . .	48
2-10 数据点质量和用户轨迹质量与传统量化指标的比较 . . . . .	49
2-11 不同空间尺度下数据点质量 $Q_P$ 的熵分布及其与 $Q_I$ 的比较 . . . . .	50
3-1 个人时空行为轨迹和不同分析粒度上的行为模式 . . . . .	54
3-2 CITY-M 数据集中的个体移动图示例 . . . . .	57
3-3 移动图节点和边的 1 阶邻居示意图 . . . . .	59
3-4 对顶点和边相似矩阵分别应用匈牙利算法的介观模式结构 . . . . .	63
3-5 CITY-M 数据集中用户的移动图在工作日和周末的特征分布 . . . . .	69
3-6 具有不同图基数的分组的层级聚类结果可视化 . . . . .	70
3-7 不同聚类簇里的最显著介观模式和频次最高模序对比 . . . . .	71
3-8 自距离和聚类簇内平均距离的多模关系 . . . . .	73



0000294

3-9	自距离和用户移动时空结构之间的关系展示 . . . . .	73
3-10	干扰机会框架和基于转移代价的个体机会地图示例 . . . . .	77
3-11	CITY-M 数据集中不同时间段的全局机会地图 . . . . .	78
3-12	个体的回转半径与时间和移动距离的分布关系 . . . . .	82
3-13	个体停留时间分布的直方图和拟合概率密度分布 . . . . .	83
3-14	群体停留时间在工作日和周末的概率密度分布 . . . . .	84
3-15	纯时间角度的个体停留时间模型的参数分布 . . . . .	85
3-16	时空结合角度的个体停留时间模型的参数分布 . . . . .	85
3-17	流涌现模型的性能验证 . . . . .	86
4-1	个体和群体行为的时空分布特征对比 . . . . .	90
4-2	不同角度的群体时空行为研究方法比较 . . . . .	91
4-3	CITY-M 数据集中群体行为的潮汐效应热力图 . . . . .	92
4-4	不同空间尺度下观测到的群体分布热力图 . . . . .	97
4-5	不同空间尺度下网络基站/热点粒度下的人群分布 . . . . .	98
4-6	网络基站粒度下人群和网络流量的空间异质性对比 . . . . .	99
4-7	不同空间尺度下人群移动的空间结构对比 . . . . .	100
4-8	不同空间尺度下的人群移动的空间相关性分析 . . . . .	102
4-9	城市尺度下人群分布的双模模型参数与时间的关系 . . . . .	104
4-10	国家尺度下人群分布的双模模型参数与时间的关系 . . . . .	104
4-11	校园尺度下人群分布的双模模型参数与时间的关系 . . . . .	104
4-12	城市尺度下不同功能区域的空间结构示意图 . . . . .	106
4-13	城市尺度下不同功能区的人群移动趋势和时间相关性对比 . . . . .	106
4-14	利用时空关联特征进行人群分布预测的模型示意图 . . . . .	110
4-15	基于时空关联的 ST-Model 人群预测值与观测值的对比 . . . . .	112
4-16	不同模型在考虑区域差异下的协方差分布 . . . . .	114
5-1	移动网络用户参与行为建模及量化分析框架 . . . . .	119
5-2	传统 Web 流量模型和新型移动 HTTP 流量模型对比 . . . . .	120



0000294

5-3 CLICK 数据集中用户行为和网络流量的关系示例 . . . . .	122
5-4 $\mathcal{AID}$ 算法性能评估和参数 $\tau_L$ 估计 . . . . .	123
5-5 不同设备平台上用户参与行为特征对比 . . . . .	128
5-6 时间因素对移动用户参与行为的影响 . . . . .	128
5-7 地点熟悉度对用户参与行为的影响 . . . . .	129
5-8 不同设备平台和应用类型下的服务质量特征 . . . . .	130
5-9 移动网络中用户参与行为与应用质量特征的关系 . . . . .	132
5-10 移动网络中不同设备平台对异常行为率的影响 . . . . .	133
5-11 移动用户的参与行为相关系数分布图 . . . . .	134
5-12 用户访问不同类型移动应用的概率分布 . . . . .	134
5-13 空间偏好（地点熟悉度）对用户参与行为的影响 . . . . .	138
5-14 移动用户参与行为的隐马尔可夫过程示例 . . . . .	139
5-15 基于模型的参与轨迹序列聚类分析 . . . . .	141
5-16 不同用户的特征分布示例及全体用户的熵分布 . . . . .	143
5-17 用户参与行为的模型选择及隐状态时间序列 . . . . .	144
5-18 用户参与行为的不同隐状态及对应特征分布 . . . . .	144
5-19 模型距离的快速近似计算以及基于模型的参与轨迹聚类示例 . . . . .	145
5-20 不同聚类簇中的用户参与行为特征分析 . . . . .	145
5-21 不同聚类簇和显著状态下的参与行为特征分布 . . . . .	146



0000294



0000294

## 表格索引

2-1 OmniPerf 采集主要数据类型说明 . . . . .	32
2-2 网络流量特征实时提取工具 HTTP-SNIFFER 性能验证 . . . . .	35
2-3 不同空间尺度移动网络数据集特征及比较 . . . . .	41
3-1 模序分析和介观模式挖掘对比 . . . . .	67
4-1 不同空间尺度下人群移动图节点度分布规律 . . . . .	100
4-2 不同模型在考虑区域和时段差异下的预测性能对比 . . . . .	111
5-1 参与行为特征之间的显著系数 . . . . .	135
5-2 不同应用语义对参与行为特征相关性的影响 . . . . .	137



0000294



0000294

## 算法索引

3-1 拓扑-属性耦合的图相似算法 TACSim . . . . .	62
3-2 显著模式提取 PPM 算法 . . . . .	66
3-3 流涌现模型 FEM 生成算法 . . . . .	81
5-4 用户网络参与行为的识别算法 $\mathcal{AID}$ . . . . .	122



0000294



0000294

## 主要符号对照表

$\mathcal{B}$	个体时空行为序列
$\mathbf{s}$	空间位置向量
$t$	时间标量
$\mathbf{c}$	场景信息向量
$Q_P$	时空数据点的质量
$Q_I$	用户轨迹的数据质量
$H_p$	历史轨迹序列
$G_p$	用户移动图
$M_{AB}$	介观时空模式
$\mathbf{A}$	移动图的邻接矩阵
$\mathbf{A}_s, \mathbf{A}_t$	移动图的顶点-边邻接矩阵
$g_{ij}$	邻居顶点的连接强度
$h_{ij}$	邻居元素对的强度一致性
$\mathbf{X}_k, \mathbf{Y}_k$	移动图的顶点和边相似矩阵
$\mathbf{Z}$	转移相似矩阵
$\mathcal{M}$	显著介观模式集
$\delta_{ij}$	移动图 $G_i$ 和 $G_j$ 的结构距离
$\rho_i$	机会资源密度
$Y(\mathbf{s}, t)$	群体的时空分布函数
$D_s, D_t$	时间和空间观测范围
$C(\mathbf{h}, u)$	时空协方差函数
$\rho(\mathbf{h}, u)$	时空相关性函数
$\mathbf{h}$	空间跨度
$u$	时间跨度



0000294

---

$\hat{C}(\mathbf{h}, u)$	经验时空协方差函数
$\hat{\rho}(\mathbf{h}, u)$	经验时空相关性函数
$h_X$	最大空间相关距离
$\mathcal{L}$	对数似然函数
$\mathbf{M}_i$	时刻 $t_i$ 的人群分布矩阵
$\mathcal{A}$	移动用户的参与行为
$E$	网络对象实体
$\tau_L$	参与动作时间间隔阈值
$D_e$	参与会话时长
$f_v$	访问频次
$r_{int}$	行为异常率
$d_{\mathcal{A}}$	感知操作时长
$w_{\mathcal{A}}$	感知等待时间
$b_{\mathcal{A}}$	感知吞吐率
$I_c$	并发指数
$R_{ij}$	结构相关系数
$\phi$	参与行为的显著系数
$\mathbf{O}_b$	参与行为特征向量
$\mathbf{O}_p$	参与行为的应用质量特征向量
$\mathbf{O}_c$	参与行为的场景特征向量
$\mathbf{O}_{1:T}$	参与行为的观测序列
$\mathbf{S}_m$	参与行为的隐藏状态序列
$H_{ij}$	参与行为的模型距离



0000294

# 第一章 绪论

随着移动技术的飞速发展，人类社会积累了大量的通信数据。这些数据借助于移动设备搭载的微型传感器，有的直接记录了用户的移动位置、网络服务的使用，有的间接记录了周边的环境信息、以及用户的社交关系。和采用传统调研方式采集的数据集相比，移动网络数据具有样本规模大、数据更新度高、采集代价小的特点。在本文中，我们简称这类数据为“移动大数据”。移动大数据以其自身的规模优势，减少了传统分析方法对先验分布假设的依赖，使得行为科学领域（Behavior Science）以前难以直接研究的问题变得可行，从而推动了近年来对大规模人类行为规律、及其产生机理的研究。

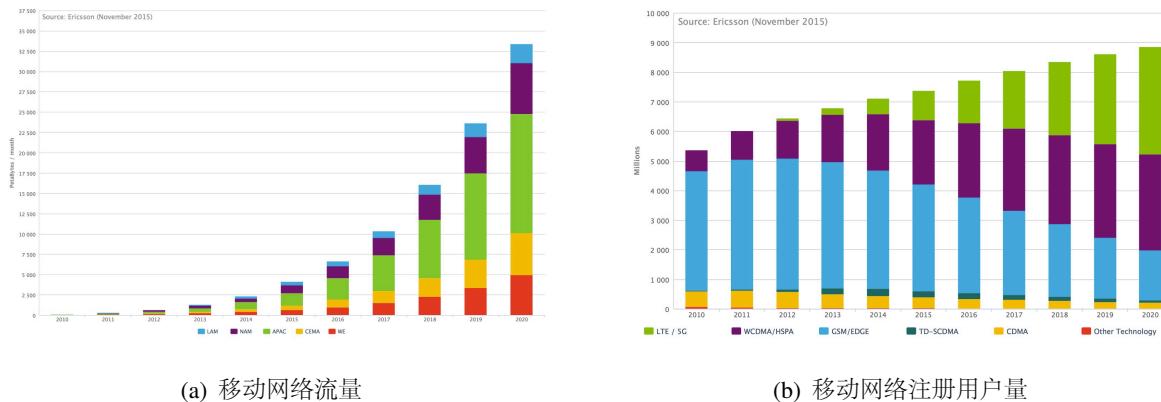
人类的时空行为<sup>1</sup>记录了个体或群体在特定空间、时间范围内的移动性。作为近年来行为科学领域的热点问题之一，人类时空行为研究，不仅有助于移动网络运营商优化网络性能和服务体验<sup>[1]</sup>，而且在流行疾病传播控制<sup>[2,3]</sup>、城市交通系统规划<sup>[4]</sup>、以及群体性事件的检测和预防<sup>[5]</sup>等有着重要的应用价值。另一方面，利用移动大数据对时空行为进行分析和建模，不仅对正在发展中的大数据研究理论具有参考价值，而且新的分析方法和算法模型，对于类时空数据挖掘问题（即通过转换可等价为时空行为分析的领域问题）具有较强的普适作用。

本章作为全文的基础，主要阐述了人类时空行为研究的背景。在给出时空行为定义的基础上，讨论了时空行为挖掘的应用和理论价值，以及国内外近年来的研究趋势。通过对有影响的国际会议、期刊上的相关工作进行深度调研，分别从时空规律、算法模型、以及方法论角度进行了归纳总结。在此基础上，梳理了人类时空行为挖掘中仍待解决的关键问题，以及本文的主要研究工作及创新点。最后给出了全文的结构安排。

## 1.1 人类时空行为的研究背景

自本世纪初智能手机设备普及以来，人类的通信方式不断向便携式和多样化发展。种类丰富的移动应用，不仅满足了人们在通信、休闲、消费、社交等诸多方面的需求，

<sup>1</sup>如无特殊说明，本文中时空行为特指人类的时空行为，二者在表述上具有相同的含义。



(a) 移动网络流量

(b) 移动网络注册用户量

图 1-1 全球移动网络流量和注册用户量趋势预测，数据来自爱立信，2015

Fig 1-1 Prediction of mobile network traffic and subscription globally, Ericsson, 2015.

也推动了移动互联网的用户数量呈指数增长。据中国互联网信息中心（CNNIC）第 36 次《中国互联网发展状况统计报告》，截止 2015 年 6 月，我国的互联网普及率为 48.8%，其中手机用户规模达到了 5.49 亿，占互联网用户总人数的 88.9%。丰富的移动应用，促使人们的日常通信不再局限于传统的语音通话，而是逐渐被社交网络、即时消息等新型通信方式所取代。在这两方面因素的共同作用下，移动网络流量呈现爆炸式增长。据爱立信最新的移动市场报告<sup>1</sup>，2015 年全球移动网络流量（图 1-1(a)）约为 4090PB/月，较 2014 年增加了 77.8%，以亚太地区（APAC）为代表的世界各区域将以 47%~65% 的速率逐年（至 2020 年）递增。移动用户群体和网络流量的激增，潜在地促进了无线通信技术的升级换代。传统 GSM/GPRS 网络的低速率移动通信技术，正逐步被传输效率更高的 4G 以及 LTE/5G 技术所取代。如图 1-1(b)所示，2015 年 GSM 网络注册用户 36.08 亿，是 LTE/5G 用户的 367%；到 2020 年，这一比例将缩小到 35%，从而更好地满足人们日益增长的网络资源和服务体验需求。这些国内外不同机构的统计数据表明，我们正在步入一个真正的“移动大数据”时代。

### 1.1.1 时空行为的定义

移动网络应用在提供服务的同时，还在以多种不同的形式数字化着人类的生活。现代的智能设备不但硬件体积小、携带方便，而且搭载了功能丰富的微型传感器，例如

<sup>1</sup><http://www.ericsson.com/mobility-report>

GPS 和网络模块能够辅助进行空间定位，加速度传感器可以感知用户的运动状态。基于硬件传感器开发的移动服务，为用户提供便捷生活的同时，也在人类历史上首次实现了大规模的、低成本的数据采集<sup>[4]</sup>。例如风靡全球的基于位置定位的游戏 Ingress<sup>1</sup> 由 Google 公司开发，成功实现了人们以娱乐的方式共享真实时空数据。另外一些服务（如 Foursquare）利用基于地理位置的签到数据，为用户提供更加个性化的服务和推荐内容，这些内容反过来帮助服务提供商理解特定位置对于用户的重要性。在网络空间 (Cyber Space) 中，如果将不同的网络服务（以域名为标识）看作独立的“空间位置”，超文本链接为连接不同位置的道路，那么用户在网络空间的使用偏好和浏览行为，则被海量的移动网络数据记录着。除此以外，现实生活中人类时空行为，还被其他网络系统或服务记录着<sup>[6]</sup>，如城市公共交通的电子刷卡系统、自行车租赁服务、以及出租车 GPS 系统等，以起止 (OD) 点的形式在较粗空间粒度上记录了乘客的通勤行为；货币流通网络<sup>[7]</sup> 从货币交换的过程感知人类长距离的旅行。由于移动网络通常覆盖的空间范围广、用户群体大、采集成本低，本文以多尺度下的移动网络数据为基础，通过给出时空行为的一般形式，使提出的分析方法及理论模型，可以在相似数据集或时空问题上同样适用。

**定义 1.1** (个体时空行为). 个体时空行为指在给定时间、空间范围内，观测对象表现出来的一系列空间位置转换的序列，表示为  $\mathcal{B} := \{(\mathbf{s}_i, t_i, \mathbf{c}_i) | i = 0, 1, 2, \dots\}$ ，式中  $\mathcal{B}$  表示观测到的时空行为序列，其中每个元素  $\mathbf{b} = (\mathbf{s}, t, \mathbf{c}) \in \mathcal{B}$ ，称作时空行为向量，且  $\mathbf{s} \in \mathcal{D}^{|\mathbf{s}|}$  表示  $|\mathbf{s}|$  维的空间位置向量， $t \in \mathcal{T}$  表示时间标量， $\mathbf{c} \in \mathcal{X}^{|\mathbf{c}|}$  表示  $|\mathbf{c}|$  维的场景信息向量。

在实际研究中，针对不同来源、不同质量的时空数据集，定义1.1中的时空行为向量和序列会有不同的形式。通过将不同背景下的时空行为用统一的形式进行表征，不但可以利用相似的算法和模型对人类时空行为进行研究，而且从不同角度揭示了人类时空行为的普遍规律和内在联系。本文着重讨论两种不同的时空行为，即物理空间中的移动行为和网络空间中的用户参与行为。对于移动行为，空间向量  $\mathbf{s}$  记录了用户的物理位置，如经纬度，从而可以通过欧氏距离、球面距离、或者路网距离对用户的移动行为进行量化分析；场景向量  $\mathbf{c}$  包括空间和时间的相关属性，如地点热度、POI 特征、日期等。对于用户参与行为，如果将网络服务映射为网络空间中的位置，则空间向量  $\mathbf{s}$  表示用户在

<sup>1</sup><https://www.ingress.com/>



0000294

时刻  $t$  所处的网络空间位置，超链接关系和服务使用次序构成了空间中的转移行为；相应的场景向量  $\mathbf{c}$  表示用户参与网络服务时的环境信息，如物理位置、网络状况等。

### 1.1.2 时空行为的研究价值

对人类的时空行为规律进行挖掘的重要价值，不仅体现在对已有的社会活动现象（如大规模群体事件）的形成、消失过程进行定量分析，充分了解其发展脉络；而且能够结合历史事件规律，和已有部分观测信息，对未来同类事件发生的可能性、规模、以及趋势等提前把握，帮助管理员和决策者做出合理而准确的规划。本节从应用和理论两个角度，结合实际应用场景和案例，对时空行为的研究价值进行了阐述。

#### 1.1.2.1 应用价值

传播型疾病通常以个体接触的形式进行扩散，因此和人类的时空行为紧密相关<sup>[2,3,8-10]</sup>。例如作为人类历史上爆发的最严重的传染疾病之一，埃博拉（Ebola）病毒<sup>1</sup>截止 2015 年 9 月已造成 11,306 人失去了生命。该病毒主要通过接触被感染人或动物的体液进行传播，从初次爆发的非洲中部，一直蔓延到西非（如几内亚）和非洲北部（如苏丹），并在美国、西班牙等地区发现疑似和确诊病例。这种“跨洋越海”的病毒传播方式，正是借助于人类的移动行为实现的。Wesolowski 等<sup>[3]</sup> 基于移动运营商的手机通话记录（Call Detail Record, CDR），对埃博拉疫区以及周围若干国家的人口分布、行为模式等进行了分析。研究结果表明，对于埃博拉病毒的肆虐，除了和当地政府的卫生条件和政策制定有关以外，地区内、地区间的人口流动是另一决定性因素。因此，对人口流动强度的分析，有助于对下一次爆发的时间和地点实现前瞻性把握，及时做好预防措施。例如，研究者发现，西非地区的人口比目前较爆发集中的中非地区流动性更强，有着潜在的大规模爆发风险。

随着现代城市化进程的加快，城市所承担的功能不再是满足人们的日常所需，如商品和工作机会，更重要的是实现生活便捷的同时，提供高质量的生活环境和城市智慧。高效率的城市交通、城市污染监控、群体性应急事件预防等一直是“智慧城市”领域<sup>[11]</sup>关注的热点问题。利用城市系统产生的各种时空数据，如公交刷卡、移动网络数据等，

<sup>1</sup><https://en.wikipedia.org/wiki/Ebolavirus>



0000294

对城市居民的出行方式和行为从大尺度上进行把握，不仅可以优化城市交通资源<sup>[4]</sup>，降低居民出行成本，而且结合居民的时空行为规律，能够实现智慧化的生活服务，如针对下班后要去菜场购物的上班族推出的“一站式”配送服务。另一方面，对城市居民的时空行为进行长期分析，能够对区域土地的使用、人口通勤分布提供量化指标<sup>[5]</sup>，为城市管理者的未来建设规划提供有力的依据。Woodcock 等<sup>[5]</sup> 利用伦敦市自行车共享系统的 740 万租赁记录，从空气污染和交通事故伤害角度，量化分析了自行车共享系统对居民健康程度的影响；结果从数据角度肯定了自行车共享系统对减小居民死亡风险的好处，同时发现男人较女人、老人较年轻人风险降低程度更高。

从移动网络自身来看，研究网络中的用户移动模式和上网行为，对优化网络资源和开发新型网络协议都大有裨益。对网络服务提供商（ISP）而言，用户的网络体验是网络优化的首要目标。但是从用户端到服务端，中间涉及的硬件和软件组件众多，一旦用户反馈问题很难从单个用户记录里诊断出故障所在。通过对网络中大量用户的时空行为数据进行采集，分析在相似约束条件下（如空间位置、时间段、网络服务类型、终端硬件、软件等）批量用户的网络体验数据，便能以较高的可信度诊断出用户故障反馈的真实性、以及故障原因所在。对网络组件和协议开发者而言，用户的时空行为研究有助于从大尺度（如城市、国家）上了解网络负载的时空分布特性，开发出符合用户使用规律的动态资源调度的网络器件；同时，结合用户移动行为和网络参与行为的移动网络流量模型，因为减小了对流量先验分布特征假设的依赖，在开发新型网络协议（如机会网络协议<sup>[1]</sup>）时，较纯统计的流量模型<sup>[12]</sup> 更准确、更灵活。

### 1.1.2.2 理论价值

作为行为科学的一个子领域，时空行为研究从时间和空间维度对人类的行为进行探索，并从理论上对时空行为规律进行总结。2005 年 Barabási<sup>[13]</sup> 对人类行为在时间上的非泊松特性进行了研究，修正了以往理论研究中对人类行为符合泊松分布的假设。在空间上，人们发现了莱维飞行（Lévy Flight）特性<sup>[14]</sup>，并且在停留地点数<sup>[15]</sup>、地点偏好<sup>[16]</sup>、停留时间<sup>[15]</sup>、接触时间<sup>[17]</sup> 等分布中观测到了重尾分布的特点。通过对一段时间内人群在空间上的分布数据进行分析，人们在传统重力模型<sup>[18]</sup>、介入机会模型<sup>[19]</sup> 的基础上提出了辐射模型<sup>[20]</sup>。这些统计规律和理论模型的一般性，为行为科学自身的不断发展、



0000294

以及与其他学科的交叉融合上提供了基础。

时空行为分析的另一个贡献是对网络科学理论的发展。人类的社会活动随着时间的积累，会演化出不同空间下的网络结构，如社交行为演化出了社交网络、日常的通勤行为构成了通勤网络（即描述了城市里不同地点之间的人群流动方向和数量）。这些特定场景下的网络结构，为探索复杂网络的新特性的发现提供了基础。例如，通过对大规模社交网络中用户的交友关系进行研究<sup>[21,22]</sup>，人们发现了网络科学里称之为“小世界”（Small World）<sup>[17]</sup>的新特性，即单个节点到达其他节点的平均距离一般较小；对应到现实生活中，两个陌生人之间以较高的概率通过数目不多的熟人连接在一起。另一方面，人类的行为随着时间在变化着，形成的网络结构也在相应地发生着改变，因此衍生出对动态复杂网络理论的研究<sup>[23,24]</sup>。Holme 等<sup>[24]</sup> 基于包括个人通信行为在内的多种网络结构，总结了时序网络（Temporal Network）分析的度量指标和理论模型。

此外，人类时空行为与经济学的交叉研究，也为经济理论的发展提供了新的视角<sup>[25-28]</sup>。在区域经济关系的研究中，Eagle 等<sup>[25]</sup> 利用国家尺度的手机网络通信数据，研究了网络多样性和空间经济结构之间的关系；分析结果发现区域通信模式的多样性（包括社交网络和空间网络）是区域经济健康程度的一个“指示器”。在城市经济的研究中，Arcaute 等<sup>[26]</sup> 分析了国家尺度上城市的空间分布和各项经济指标之间的关系，并提出了用功能性的“Urban” 定义代替传统上以行政单位为基础的“City” 划分。综上所述，时空行为研究中的理论方法和模型，对行为科学、网络科学以及经济学的研究和发展起到了补充和推进作用。

### 1.1.3 时空行为的研究趋势

自从人类时空行为的非泊松特性被首次发现<sup>[13]</sup> 以来，国内外涌现出一批长期致力于时空行为研究的组织。在国际上，美国西北大学物理系的 Barabási 课题组<sup>1</sup> 从物理学角度对人类时空行为进行理解，研究特色是基于统计力学理论，将个体看作物理粒子、人群看作受特定约束的粒子系统，进而从不同尺度上对粒子的行为进行研究。以计算科学为背景的麻省理工学院（MIT）多媒体实验室人类行为课题组<sup>2</sup>，结合计算机和大数据

<sup>1</sup><http://barabasi.com/>

<sup>2</sup><http://hd.media.mit.edu/>



0000294

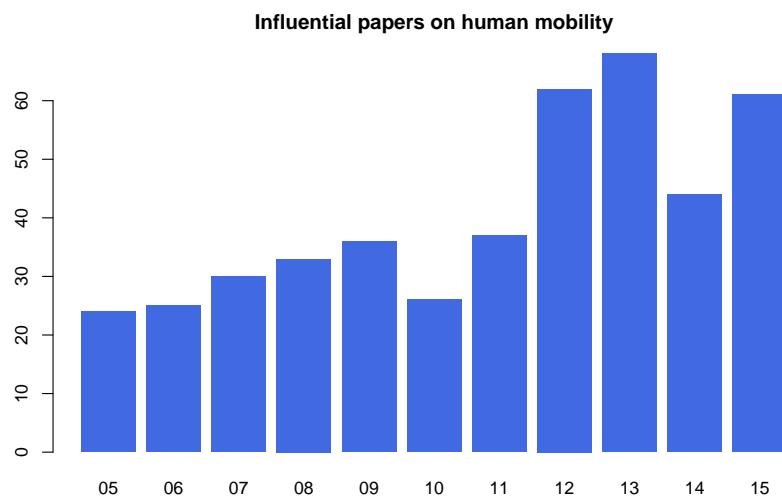


图 1-2 2005 至 2015 年间发表的有影响力国际会议和期刊论文

Fig 1-2 Influential papers on human mobility that are published during 2005 and 2015.

的处理优势，对人类在交通系统、健康、能源、以及金融系统里的行为进行研究。此外，MIT 环境工程系 González 领导的 HumNet 课题组<sup>1</sup>，从复杂网络角度对人类的时空行为进行研究，在移动模式挖掘问题上有着长期的积累。

同一时期国内也涌现出一批优秀的研究组织。北京城市实验室（Beijing City Lab<sup>2</sup>）通过虚拟组织的形式将一批研究人员聚集在一起，研究特色是结合多种类型的数据，对城市系统里的重要环节（如区域规划、城市经济等）进行量化的分析。微软亚洲研究院的城市计算组<sup>3</sup>致力于利用异构的城市数据和机器学习的理论，对城市生活中的实际问题（如交通资源优化、空气污染、房产价格等）进行了模型统计和实证分析。电子科技大学的周涛所在课题组，是国内人类行为动力学研究的代表，同时利用物理模型和复杂网络理论对时空行为规律进行研究。

图1-2展示了 2005 至 2015 十年间发表在有影响力的国际会议和期刊的论文数量。图中数据检索自计算机科学专业论文数据库 DBLP，按照中国计算机协会推荐的会议和期刊进行过滤。从中可以看出，人类时空行为的有影响力的研究成果数量，在最近五年

<sup>1</sup><http://humnetlab.mit.edu/>

<sup>2</sup><http://www.beijingcitylab.com/>

<sup>3</sup><http://research.microsoft.com/en-us/projects/urbancomputing>



0000294

得到了 80%~93% 的提升。这类现象的产生，一方面受人类行为新成果的驱动，另一方面受益于近年来后智能化移动设备的普及，为大规模采集匿名行为数据提供了可能。

## 1.2 人类时空行为挖掘的国内外研究进展

作为一个交叉领域，时空行为研究者来自计算机科学、通信工程、物理学、地理信息学等多个领域，从各自领域的视角揭示着人类行为的内在规律。本节对 2005~2015 年间有代表性的研究工作进行了调研，并从数据集、研究方法、行为规律、理论模型等角度进行了对比分析。

### 1.2.1 人类移动行为研究进展

#### 1.2.1.1 时空统计特征

时间特征：人类长期以来遵循着“日出而作，日落而息”的生活规律。随着社会的发展，这样的作息规律和生活习惯、地理分布、职业种类等结合起来，衍生出复杂的时空移动行为。尽管如此，时间特征依然为我们揭示了人类生活中的一些普遍规律。Barabási<sup>[13]</sup>从动力学角度出发，对人类行为在时间上满足泊松分布的假设提出了质疑，并在通信数据中发现了非泊松分布以及阵发的行为特征。为了探究产生这一新特性的机制，作者分析了三种队列服务模型：先进先出、随机服务、以及基于优先权的服务，并利用基于优先权的队列服务模型，对人类日常行为中表现出来的幂律（Power-Law）现象进行了解释。Goh 等<sup>[29]</sup>提出了阵发参数  $\Delta$  和记忆参数  $\mu$  用来衡量人类行为的间隔时间分布，并通过阵发-记忆象限图，对现有的时间行为模型是否能完全捕捉人类行为的阵发与记忆特征进行了研究。Barabási 模型和 Goh 参数均是对事件间隔时间的刻画。地点的停留时间通常也满足幂律的分布特征<sup>[14,15]</sup>。连续时间的随机游走模型<sup>[14]</sup>（CTRW）捕捉了单个地点的停留时间分布：针对每一个由随机游走模型产生的新地点，对应的停留时间由给定参数的幂律分布采样得出。

另一个重要的时间统计特征是接触时间间隔（Inter-contact time, ICT），即不同个体在连续两次接触时的间隔时间。Mei 等<sup>[17]</sup>从实际数据中观测到 ICT 通常具有二分性（Dichotomy），即 ICT 的概率分布函数在低值部分满足幂律分布，在高值部分满足指数

分布。这样的二分性被认为是由于移动个体在特定时间内，其移动范围随着时间的推移趋向一个特定的边界值，因此高值部分的概率随着 ICT 增大概率急剧下降。Karagiannis 等<sup>[30]</sup>发现 ICT 二分性的临界特征值位于 0.5 天附近（即超过 0.5 天概率分布将按照指数分布下降）并将这种二分性归因于移动个体对不同地点的偏好差异以及返回时间的幂律分布上。从本文的研究视角来看，Mei 和 Karagiannis 对 ICT 二分性的解释本质上是一致的，均基于移动个体对不同地点的访问偏好不同，而移动范围的边界性和返回时间的幂律分布是访问偏好产生的不同统计结果。

**空间特征：**社会资源在空间上的分布和人类行为的空间特征有着紧密的联系。在早期的移动模型中，个体的下一地点被认为是随机选择的，也就是所谓的随机游走模型。后来随着位置服务和应用的兴起，人们通过真实地理位置数据分析，发现个体的移动位置通常表现出所谓的“莱维飞行”特性<sup>[7,14,31]</sup>：虽然在大多数情况下，人们倾向于短途旅行，但依然存在一定的概率进行长距离远行。莱维飞行可以被认为是布朗运动的一般形式，同时属于无标度的分形随机过程<sup>[14]</sup>。莱维飞行是一种相邻地点的转移距离  $\lambda$  满足幂律分布的马尔可夫随机模型，这意味着  $\lambda$  概率分布的二次矩是发散的，并且任意长转移距离都有可能产生。这样过程被称为“简单莱维飞行”。

事实上，和现实观测对比分析后发现，简单的莱维飞行模型是不完备的<sup>[14,16]</sup>。简单莱维飞行的转移距离满足无标度的分布，从而使空间扩散形成一个超扩散过程（Super-diffusion process）<sup>[14]</sup>。而在实际观测中，个体的移动范围有边界性，意味着移动行为的空间扩散是不断减弱的。González 等通过对十万移动网络用户的匿名轨迹分析，发现空间转移距离满足结尾的幂律分布，即在  $\lambda$  超过阈值  $\lambda_0$  后其概率分布满足指数分布。用数学形式可以简洁地表达为， $p(\lambda) \sim (\lambda + \lambda_0)^{-b} \cdot e^{\lambda/k}$ ，其中  $0 < b \leq 2$ ， $\lambda_0$  在不同应用中取值不同。我们称这样的过程为“截尾莱维飞行”。

CTRW 模型是结合了简单莱维飞行和停留时间幂律分布的随机游走模型<sup>[14]</sup>。虽然其包含了空间转移分布，但访问地点演化以及访问频率与实际观测数据不符。针对这些不足，Song 等<sup>[15]</sup>基于 CTRW 模型并考虑了以下空间分布特征：i) 个体的独立地点数目  $L(t)$  满足幂律的增长过程， $L(t) \sim t^s, s < 1$ 。和单纯的随机游走模型和莱维飞行的随机游走模型<sup>[14]</sup>相比，Song 的模型预测了更加准确的地点分布。ii) 地点的访问频率满足齐夫（Zipf）定律<sup>[32]</sup>，即人类访问地点的频率不是均匀分布的，而是倾向于频繁访问一小

部分常去的地点。利用数学形式表达,  $f_k \sim k^{-\zeta}$ , 其中  $\zeta = 1.2 \pm 0.1$ ,  $k$  为地点按照访问频次逆序排列的整数序号。iii) 空间距离满足超低扩散过程 (Ultraslow diffusion)。在 CTRW 模型中, 观测时间越长, 个体的访问地点越远离初始点。而实际观测显示, 个体的空间移动范围随着时间推移, 将趋近一个特定的空间边界, 即均方转移距离  $MSD^{[15]}$  满足对数增长分布。

回转半径  $R_g^{[16]}$  是对个体空间移动能力的有效度量, 其定义为  $R_g^2 = \frac{1}{n} \sum_{k=1}^n (\mathbf{r}_k - \bar{\mathbf{r}})^2$ 。近年来研究人员从多种角度对  $R_g$  的特性进行了研究。González 等<sup>[16]</sup> 首次将回转半径引入到人类时空行为分析中, 并发现个体的  $R_g$  分布满足截尾的幂律分布, 同时群体的转移距离分布  $P(\lambda)$  由个体的转移距离分布和群体  $R_g$  的异质结构共同决定, 即  $P(\lambda) = \int_0^\infty P(\lambda|R_g)P(R_g)dR_g$ 。Bagrow 等<sup>[32]</sup> 研究了不同活动区域里人们的回转半径分布特征, 并发现在主要活动区域内,  $R_g$  的分布特征符合对数分布, 较全局分布的增长较缓。Park 等<sup>[33]</sup> 研究了回转半径的时变特征, 通过马尔可夫过程得出了与实际观测一致的时间序列, 并将马尔可夫矩阵与时空行为的特征模式分析关联在了一起。

近年来, 人类行为的其他空间特征规律逐渐被揭示了出来。Calabrese 等<sup>[34]</sup> 利用手机通信数据对人们的“共现性”(即手机通话时处在同一基站范围内)进行了研究, 分析结果显示超过 90% 的用户即使实际居住地相距较远, 但都出现过“共现”行为, 为从单纯的通信行为角度量化研究人类的空间移动提供了支撑。Hossmann 等<sup>[35]</sup> 对移动网络用户的共现性从复杂网络角度进行了研究。通过将多源数据集的用户空间关系转换成“接触图”(Contact Graph), 分别对接触网络的度分布、小世界特性、以及社区结构等进行了量化和对比分析。Kang 等<sup>[36]</sup> 从城市形态学角度(包括城市紧凑程度和规模)对人类的移动模式进行了研究, 结果发现城市内的空间移动距离满足指数分布, 并且对应的指数参数与城市的紧凑程度和规模密切相关。Frank 等<sup>[37]</sup> 基于带有空间位置标签的 Twitter 数据, 将情感分析和空间移动模式关联起来, 分析结果发现人们的快乐程度和与平均位置的距离呈对数相关。Bora 等<sup>[38]</sup> 同样利用社交网络数据, 对种族隔离与人们的移动性之间的关联关系进行了研究, 结果发现所有种族都倾向于经常访问同一种族聚集的区域, 非洲裔、亚裔、以及西班牙裔较少出现在白人聚集区域, 并且非洲裔聚集的区域被其他种族访问的可能性最小。Kung 等<sup>[39]</sup> 利用手机通话记录、从不同空间尺度上对城市内的通勤模式进行了研究, 数据分析显示通勤时间的分布具有一般性, 和观测的空间尺

度相关性较小。

### 1.2.1.2 行为模式挖掘

时空行为模式挖掘是一类从历史轨迹数据中发现移动模式（Mobility patterns）的方法。行为模式符合人类认知客观世界的一般规律，即从多个观测对象中发现其共有部分。研究人类的时空行为模式具有普遍的应用价值和意义。一方面，从杂乱无章的日常轨迹中提取出群体的出行规律，有助于及时检测出突发性群体事件的发生，以及为事后疏散方案的制定提供有效支撑。另一方面，从空间尺度上来看，一些人造网络系统，如城市交通网络、电网、移动网络等，承受着来自人群自由移动带来的潜在系统性风险。模式挖掘有利于将用户的行为规律，嵌入到网络系统的资源调度中，从而从底层降低系统化风险的产生。本节从可预测性分析、序列模式挖掘、时空行为结构角度回顾了时空行为模式挖掘的相关研究成果。

可预测性<sup>[40]</sup>分析首次从信息量角度对时空行为的可预测性进行了理论研究。可预测性指利用“最合适”的预测算法正确预测用户下一位置的概率<sup>[40]</sup>。研究人员利用三个月的匿名手机通话数据，首先基于 Lempel-Ziv 压缩算法对单个用户的轨迹信息熵进行了测定，进而利用 Fano 不等式推导出可预测性的理论上限，并将规律性（Regularity）作为其下限。数据分析显示可预测性在熵等于 0.8 附近取得峰值，这意味着一个典型用户的下一位置信息量非常小，约为 1.74 左右。观测用户的平均可预测性为 0.93 左右。

序列模式是时空行为分析中最基本、也是最常用的轨迹模式之一。按照是否考虑轨迹序列的时间信息为原则，可以分为无时间序列模式和时间序列模式挖掘。无时间序列模式挖掘来源于机器学习理论中的一类经典算法，例如 Apriori 算法和 PrefixSpan 算法<sup>[41-43]</sup>。通过将用户轨迹抽象为一组符号序列，算法输出满足给定支持度的一组序列模式。该类模式具有算法复杂度低、结果直观等优点，在实际场景中得到了广泛应用。例如 Gong 等<sup>[43]</sup>利用 WiFi 网络数据、研究了校园用户的轨迹模式随时间的变化，以及模式变化和外部公共事件之间的联系。Tiakas 等<sup>[41]</sup>在考虑空间网络的前提下，提出了一系列衡量轨迹序列距离的量化指标，为相似轨迹聚类、用户画像等提供了支撑。

无时间序列模式的挖掘结果拥有两个基本特征：1) 模式的子序列也是模式；2) 子模式的支持度不小于父模式的支持度。这决定了该方法在行为分析中的鲁棒性较低：a)



0000294

由于序列中的每个位置符号是无权重的，因此位置的模糊和缺失将严重影响长模式的数量；b) 序列模式的本质是一个多阶马尔可夫过程。虽然包含序列信息，但缺少时间和空间上的结构信息。因此对于序列不同、时空结构相似的用户模式区分能力较弱。

时空序列模式<sup>[41,44-47]</sup>在保留空间马尔可夫过程的同时，添加了更多时间信息。其中时间信息可以以多种形式集成到算法当中，如停留时间、移动间隔时间、到达时间点等。时间信息的类型不同，对应的模式挖掘算法设计则不同。Patel 等<sup>[46]</sup>从停留时间角度，基于相似轨迹段和最小描述长度原则，将多个轨迹数据融合到一个时间加权的轨迹网络中，并实现了对网络中 Top-K 轨迹进行分类标注。这个方法的优势是将停留时间作为轨迹段聚类的基础，从而形成了统一的轨迹网络；不足之处在于，最小描述长度原则下的最优并非都是实际情况下的最优。Giannotti 等<sup>[45]</sup>和 Tiakas 等<sup>[41]</sup>从移动间隔时间角度对用户的轨迹序列进行了研究。前者<sup>[45]</sup>对无时间序列模式中的序列包含关系进行了一般性拓展，定义了时间和空间结合的轨迹包含关系，并在此基础上提出了兴趣区域（Region of Interest, ROI）检测的方法。这个方法的不足在于，轨迹包含关系缺少位置之间的空间距离信息。后者<sup>[41]</sup>首先从单纯的时间角度给出了轨迹的时间距离，然后通过加权和的方法将空、时距离结合定义了轨迹的时空距离。这个方法的缺陷在于未考虑时间和空间维度的交互信息。Chen 等<sup>[47]</sup>研究了包含到达时间点的轨迹数据，建立了用户的轨迹画像，并对无时间和时空结合的轨迹序列模式进行了对比研究。这个方法的优势是考虑了序列和语义角度的空间距离、并将时空交互信息融合到了相似度计算当中，不足之处在于缺少单个轨迹内部的时空结构信息。

时空结构是近年来逐渐得到重视的概念，指从用户的行为语义<sup>[48-50]</sup>和特征模式<sup>1</sup>角度理解人类时空行为规律。一方面，人类的移动行为表现出不同的语义状态，例如，Zheng 等<sup>[48]</sup>利用 GPS 数据和监督学习算法对行人的交通方式（即步行、私家车、公共汽车）进行研究，并对不同的特征进行了性能分析。Farrahi 等<sup>[49]</sup>借鉴了文本分析当中的隐含狄利克雷分布（LDA）技术，对带有时间戳的轨迹序列数据进行分析，从中识别出四种行为状态，并基于这些行为状态对用户的移动规律进行了量化研究。

另一方面，研究者发现人类的移动行为拥有一组特征模式<sup>[33,51-53]</sup>，这组特征模式各自之间的相关性较小，但是相互之间的组合构成了多样的行为模式。通过对特征模式的

<sup>1</sup>在本文中，特征行为<sup>[51]</sup>、特征状态<sup>[33]</sup>、以及特征模式<sup>[52]</sup>表示同一概念，并统称为特征模式。



0000294

解读，有助于把握人类移动的一些本质规律。例如 Qin 等<sup>[52]</sup> 将用户的轨迹数据表示成时空矩阵的形式，然后应用聚类算法识别出用户的特征模式，并分析了特征模式和轨迹信息熵之间的关系。Eagle 等<sup>[51]</sup> 基于二值化后的时空矩阵，提出特征行为 (EigenBehavior) 的概念，利用主成分分析 (PCA) 方法对转换后的时空矩阵进行分解，发现 15 个特征模式即可以 98% 的准确率表达用户的时空行为。Park 等<sup>[33]</sup> 首先建立了用户轨迹的马尔可夫转移矩阵，然后对该矩阵进行模分析 (Eigenmode analysis)，并对特征模式和特征值的物理意义进行了解释。这类方法的优势是摆脱了序列模式挖掘对马尔可夫性质的依赖，并结合了不同时间段内的行为信息；不足之处在于分解得到的特征模式数目需要人为确定，且各模式的可解释性有限。与此不同的是，Schneider 等<sup>[53]</sup> 通过将用户轨迹表达成有向无权图的形式，引入了网络科学中“模序” (Motif) 的概念，结果发现超过 90% 的用户轨迹可以用 17 种模序进行表达。但该研究仅考虑了时空行为中的空间结构信息，以至于缺少时间维度和时空交互信息。

### 1.2.1.3 个体移动行为建模

行为模型<sup>[54-57]</sup> 是对人类移动的时空特征和行为模式的理论抽象和概括。行为模型在客观洞察人类移动规律的基础上，使用带参数的数学模型、物理过程等方式对行为产生的机理进行模拟，从而达到量化分析和重用的目的。例如，在新型移动协议开发中，模拟用户在真实场景下的移动行为对协议设计、性能测量等起着至关重要的作用；对于需要感知用户行为的网络协议<sup>[54]</sup>，将用户行为以模型的形式嵌入到协议当中，协议便可以根据真实数据对用户行为进行学习，从而利用行为的产生和演化特征对网络协议性能进行优化。本小节对现有行为模型进行回顾，并根据模型特点分四个类别进行介绍<sup>1</sup>：

随机游走模型是用户移动性管理中最简单的一种模型，该模型因模拟气体粒子在给定空间里的随机运动规律而得名。随机游走模型<sup>[56]</sup> (RW) 描述了一个完全随机的运动过程：单个粒子或个体以随机速度  $v \in [V_{min}, V_{max}]$  和随机方向  $\theta \in [0, 2\pi]$ 。随机停靠点模型<sup>[58]</sup> (RWP) 在随机游走模型的基础上添加了停留时间的限制，即粒子在到达下一地点的时候，随机停留时间  $t \in [0, T]$ 。为了捕捉实际观测中的时空统计特性，莱维随机游走模型<sup>[59]</sup> (LRW) 和连续时间随机游走模型<sup>[14]</sup> (CTRW) 为空间转移距离和停留时间分

<sup>1</sup>更多模型可参考 Hess 调研<sup>[57]</sup> 的表 1，Pirozmand 调研的表 2，以及 Karamshuk<sup>[54]</sup> 和 Gorawski<sup>[56]</sup> 的调研文献。



0000294

别增加了重尾分布的特性。这类模型的不足在于将运动个体看作随机粒子，意味着粒子的运动过程是无记忆的，也就是下一方向和速度的选择不依赖于以前的状态。

位置偏好模型基于实际观测中，人们倾向于访问为数不多的若干位置，并且位置之间的组合构成了时空行为的主要模式。**SLAW** 模型<sup>[60]</sup> 基于分形驻留点模型和最小代价路径规划策略，实现了实际观测中的时空统计特性，如转移距离、停留时间、ICT 的幂律分布特性；结合基于空间簇的个体游走模型，从而捕捉了人们日常生活中对常去地点的偏好、以及出行路线偏好等。**SWIM** 模型<sup>[17]</sup> 实现了社交网络中的“小世界”特性，通过模拟现实生活中人们结合距离和受欢迎程度选择目的地的特点，将社交关系引入到了行为模型中；同时作者从理论上证明了模型中 ICT 二分性的存在。**TVCM** 模型<sup>[61]</sup> 对随机游走模型进行了增强，基于时变的群落（Community）结构实现了模拟个体的地点偏好特性和周期行为。这类模型往往基于随机游走模型，通过叠加额外的约束条件（如周期性），使得模型产生更加接近实际观测的统计特征。这些特征也使得该类模型在实际中得到广泛的应用<sup>[54,55]</sup>。

行程规律模型强调了移动个体的社会属性，移动轨迹的生成是受每天要参与的社会活动和活动属性所决定的。该类模型的思想来源于实际生活中，人们的行为总是按照一定的行程规划进行的，如早上八点吃早点，然后在九点前必须到达工作场所；又如晚上五点去学校接孩子下学。这样的行程规划一般包含三个要素：时间、地点、活动内容。各个要素受一定的条件约束，且活动内容和属性因人而异，从而能够按照一定的规则产生多样的个体移动行为。**ADMM** 模型<sup>[62]</sup> 是该类模型的典型代表，研究人员以模块的形式对行程要素进行定义，并利用 NHTS 调研数据生成各要素的具体内容；模型性能也通过 Ad-Hoc 网络仿真和网络协议性能分析进行了验证。另一方面，行程规划决定了观测周期内的时空行为结构，如行为模序<sup>[53]</sup> 从空间拓扑角度反映了行程活动之间的相互关系。**PBM** 扰动模型<sup>[53]</sup> 利用行程规律生成了移动行为的扰动过程，从而较好重现了实际观测中的 17 种典型行为模序。

社交关系模型基于人们的社交活动和位置变化有着潜在的关联。**Cho** 等<sup>[63]</sup> 利用带有位置标签的社交网络数据研究了这种内在的关联性，数据分析发现虽然人们在多数情况下在较小的空间范围内活动，但是依然有较高的概率去远距离的朋友居住地附近；换句话说，社交关系对远距离移动较近距离移动的影响更大。从量化角度来看，已有社交



0000294

关系对移动性的影响是移动性对新建社交关系影响的两倍。尽管如此，研究人员发现单纯依赖社交关系对用户的移动性进行预测，其性能依然有限。**PSMM** 模型<sup>[63]</sup> 基于时变的一阶马尔可夫过程，将移动行为的时空周期性特征和社交关系结合在了一起，并从平均签到似然、预测准确率、平均距离误差角度对模型性能进行了评判。与 **PSMM** 中社交关系以条件概率的形式出现不同，**HCMM** 模型<sup>[64]</sup> 的主要思想在于用户的位置依赖于拥有社交关系的其他用户的位置，且用户社交关系越强，影响力越大；同时 **HCMM** 模型融合了转移距离满足幂律分布的空间统计特征。

#### 1.2.1.4 群体移动行为建模

个体移动行为从细粒度上研究人类的时空行为规律，在小尺度、个性化应用场景中有着重要的价值。但是对于大尺度的应用，如移动网络资源优化、城市规划等，需要能够对群体移动行为规律和特点有整体性的把握。以移动网络为例，为了满足移动用户对网络性能和应用体验的需求，网络运营商通常会在人口密度高的区域部署数目较多的、高带宽的基站，反之在人口密度稀疏的区域部署较少的、覆盖范围广的基站。同时随着用户在一天内的移动，骨干网资源（如光传输波长）也会随着用户密度的变化（亦称“潮汐效应”）而进行调整。这样的场景需要我们从大空间尺度上理解人群的移动规律和时空分布特点。**De Mongis** 等<sup>[65]</sup> 研究了意大利撒丁岛各自治区间的人类移动行为，从复杂网络角度对人群移动网络的群聚系数、边权重、节点强度等基本特征进行了分析，从较大尺度上对城市的人群交互提供了观测依据。**Jiang** 等<sup>[66]</sup> 利用芝加哥的旅行调研数据（TTS）研究了城市区域之间的群体移动模式；作者基于主成分分析的 K-Means 聚类，识别出 8 种人群类型和 5 种区域交互模式。**Tanahashi** 等<sup>[67]</sup> 使用匿名的 CDR 数据分析了纽约人口的空间分布和时间动态性，并通过朴素贝叶斯模型对移动概率矩阵和可预测性进行了研究。**Deville** 等<sup>[68]</sup> 从数据角度论证了人口普查、遥感信息和匿名手机记录对国家级别、不同时间尺度上人口分布评估的可靠性，研究结果展示了匿名手机数据对时变应用场景，如区域冲突、疾病扩散等的价值。

对群体移动行为更加深入的理解方式是建立群体行为模型，从量化的角度对群体行为的时空分布进行研究。**Schilcher** 等<sup>[69]</sup> 介绍了一种统一的、客观的指标来衡量给定空间区域内点分布的不均衡（Inhomogeneity）程度，该指标基于点密度的局部方差来定义，衡



0000294

量结果和线上主观调研具有较好的一致性。此外，从网络流量分布角度，Michalopoulou 等<sup>[70]</sup> 提出移动流量的空间分布能够用混合的对数正态分布进行拟合；同时作者预言，在国家尺度范围内，网络流量的时空分布和群体时空分布紧密相关，而在城市尺度上，出于对城市环境和网络业务的考虑，二者的相关性并不显著。Lee 等<sup>[71,72]</sup> 提出利用高斯随机场，对具有对数正态分布特点的网络流量空间分布进行建模。但是，这些成果在研究群体时空分布上依然具有局限性：首先，正如 Lee 等指出，网络流量和人群数量的空间分布，既有关联也有差别，二者的模型形式和物理意义仍需进一步研究；其次，即使将现有的网络流量模型应用在人群时空分布研究上，依然只能反映空间的静态分布，而缺少时间维度的动态信息。

从群体行为的动态性角度出发，另一类模型研究人群移动的起止(Origin-Destination)规律，即对给定空间内任意两点之间的人口迁移数量进行建模，本文称作“OD 模型”。这类模型利用地点对、以及周边环境的局部信息<sup>[73]</sup> 对人口迁移量进行建模。这里介绍三种典型的 OD 模型：干扰机会模型 (Intervening Opportunity Model, IOM)<sup>[19]</sup>，重力模型 (Gravity Model, GM)<sup>[74]</sup> 和辐射模型 (Radiation Model, RM)<sup>[20]</sup>。干扰机会模型表述成，地点 A 向地点 B 迁移的人群数量，正比于两地点的机会数量，反比于 A-B 之间的干扰机会的总量。重力模型表述成，地点 A 和地点 B 之间的人口迁移量（无方向），正比于 A 和 B 的人口总量，反比于二者之间的距离，因其最终的数学形式和牛顿的重力公式类似而得名。辐射模型是最新提出的、仅利用局部人口数据对两地点之间的人口迁移量进行建模的方法，其在国家和城市尺度上都较重力模型的性能较好。Palchykov 等<sup>[73]</sup> 用两地之间的通信量代替人口数量、提出了类似于重力模型形式的预测模型。综上所述，OD 模型虽然具有对人群动态性刻画的能力，但是这种能力建立在地点的局部信息之上，因此缺少大尺度上的空间分布特点、以及不同空间点上人群密度的相关性。这些不足构成了本文利用移动网络数据，对人群时空分布特点以及依赖性进行研究的动机。

### 1.2.2 用户参与行为研究进展

在以人为中心的技术和服务模型中，用户体验（UX）是一个重要的考量方面。从用户角度来看，一个成功的人机交互服务设计不仅能够激发人们即刻的使用兴趣，而且能够让人们保持较为持久的黏着度。用户参与行为描述了人类在网络服务使用过程中



0000294

表现出来的时空行为，是人类行为规律在网络空间的一种表现形式。因其和用户体验测量、服务质量评估等有着密切的联系，近年来也受到行为科学研究领域的重视<sup>[75-79]</sup>。例如，O'Brian 等<sup>[80]</sup> 将参与行为的概念引入到 Web 测量中，并提出基于时间阶段的参与行为框架对一般的人机交互行为进行描述。另一方面，Attfield 等<sup>[81]</sup> 从特性角度（如持久度）将参与行为分解成不同的描述维度，并在每个维度上将理论描述和实际观测指标联系在一起。除此以外，由于用户参与行为和移动行为具有相似的时空数据结构，将二者结合起来进行类比研究，不仅可以验证时空行为模式挖掘、时空依赖性模型等的有效性，还能够对人类时空行为的内在规律从不同应用角度进行理解。

用户参与行为的测量对于了解用户体验有着重要的价值。无论是商品服务、人工服务、还是网络服务，服务提供者都会注重对用户体验的把握，而用户体验表现在用户的参与行为上。我们首先对用户参与行为的相关测量方法和成果进行介绍。传统上的直接测量方法是主观调研，通过问卷<sup>[82]</sup> 或用户打分<sup>[83]</sup> 的形式收集用户对于某项服务的直接评价，这样的方法一般数据质量高，但是数据规模小，单样本的成本较高。另外一类方法是利用被动的方法，对用户参与服务的过程进行客观测量。通过提取、分析相关的测量指标，简介获得用户在服务过程中的体验。这样的方法在计算机游戏<sup>[84]</sup>、Web 服务分析<sup>[75,76]</sup>、和视频服务体验感知<sup>[77-79]</sup> 等领域获得广泛应用。本文将这种被动的测量思路引入移动体验测量中，进而对移动用户的参与行为进行量化分析。

随着无线技术的普及，商用 WiFi 网络和高速移动网络为研究移动用户的参与行为提供了较好的平台。Afanashev 等<sup>[85]</sup> 利用 Google 的商用 WiFi 网络数据，从时间动态性、空间异构性角度研究了用户服务流量和移动行为的内在模式。Trestian 等<sup>[86]</sup> 将用户的服务参与行为和空间移动行为进行关联研究，发现移动模式和服务兴趣有着显著的关联，如驻留用户和移动性较强的用户倾向于使用更多的服务类型。同时，由于地点类型和场景因素的影响，特定地点（如休闲场所）和特定类型的网络服务关联在一起，表现了用户地点偏好对网络服务偏好的映射。Gember 等<sup>[87]</sup> 分析了在不同硬件平台上（移动和非移动）用户的使用状态和物理空间环境对感知到的网络性能的影响。虽然这些成果对移动网络中用户的参与行为进行了刻画，但是面临着两方面的不足：1) 网络因素和场景因素较多，对用户参与行为的影响也各有不同，孤立的因素分析并不能捕捉不同因素之间的相互作用。2) 虽然已有成果对用户的参与行为进行了量化分析，但是缺少对量化



0000294

关系的模型建立，进而降低了研究成果在其他应用中的复用程度。

### 1.2.3 进展总结及分析

最后，从数据分析流程的角度，我们对时空行为挖掘的已有成果进行可视化分析。如图1-3所示，行为数据分析通常包括数据采集和预处理→数据特征分析与刻画→理论模型建立三个阶段，其中每个阶段都包括行为分析（图1-3左侧/绿色部分）和质量控制（图1-3右侧/黄色部分）两个方面的内容。可以看出，如果时空行为挖掘仅仅重视分析方法和现象的展示，极易忽略对行为数据质量、现象显著性、以及模型有效性的控制。具体而言，在数据采集和预处理阶段，数据修复的好坏决定了后续行为表征与真实用户行为之间的差异大小，而数据置信度的测量有助于研究人员从宏观上对数据质量有所把握，从而对挖掘结果的显著程度有所预判。在数据特征分析阶段，分析现象的显著程度也应该具体而客观地衡量，以帮助研究人员判断数据现象是偶然所致，还是时空行为的普遍规律。在理论模型建立之后、投入领域应用之前，应该经过不同场景下的数据集的验证。由此可见，在行为科学领域，我们不但要不断发展行为挖掘的算法性能、模型准确度，还应该在分析结果的质量控制上开发出更加有效的方法和工具。

## 1.3 人类时空行为挖掘中的关键研究问题

人类时空行为挖掘的广泛应用，标志着高效的模式挖掘算法、准确的行为模型、以及坚实的理论支撑，有着重要的研究价值。本文基于将时间和空间维度的信息统筹考虑的思路，首先分析了在移动大数据背景下，进行时空行为挖掘所面临的多重挑战，然后从中总结出关键研究问题，并针对性地提出解决这些问题的思路和方法。

### 1.3.1 面临的挑战

虽然移动大数据为时空行为挖掘带来了便捷，但这种便捷同时伴随着新的挑战：

1) 数据质量影响了时空行为分析算法和模型的性能。在移动网络数据的采集过程中，通常采用被动采集（Passive Collection）的方式，在最小化对在线业务的影响前提下，实现采集样本数量的最大化。但是通常一个网络中有多种类型的组件，组件间又有着复杂的交互行为。其中任何组件的宕机、错误，都有可能导致原始数据的质量降低，

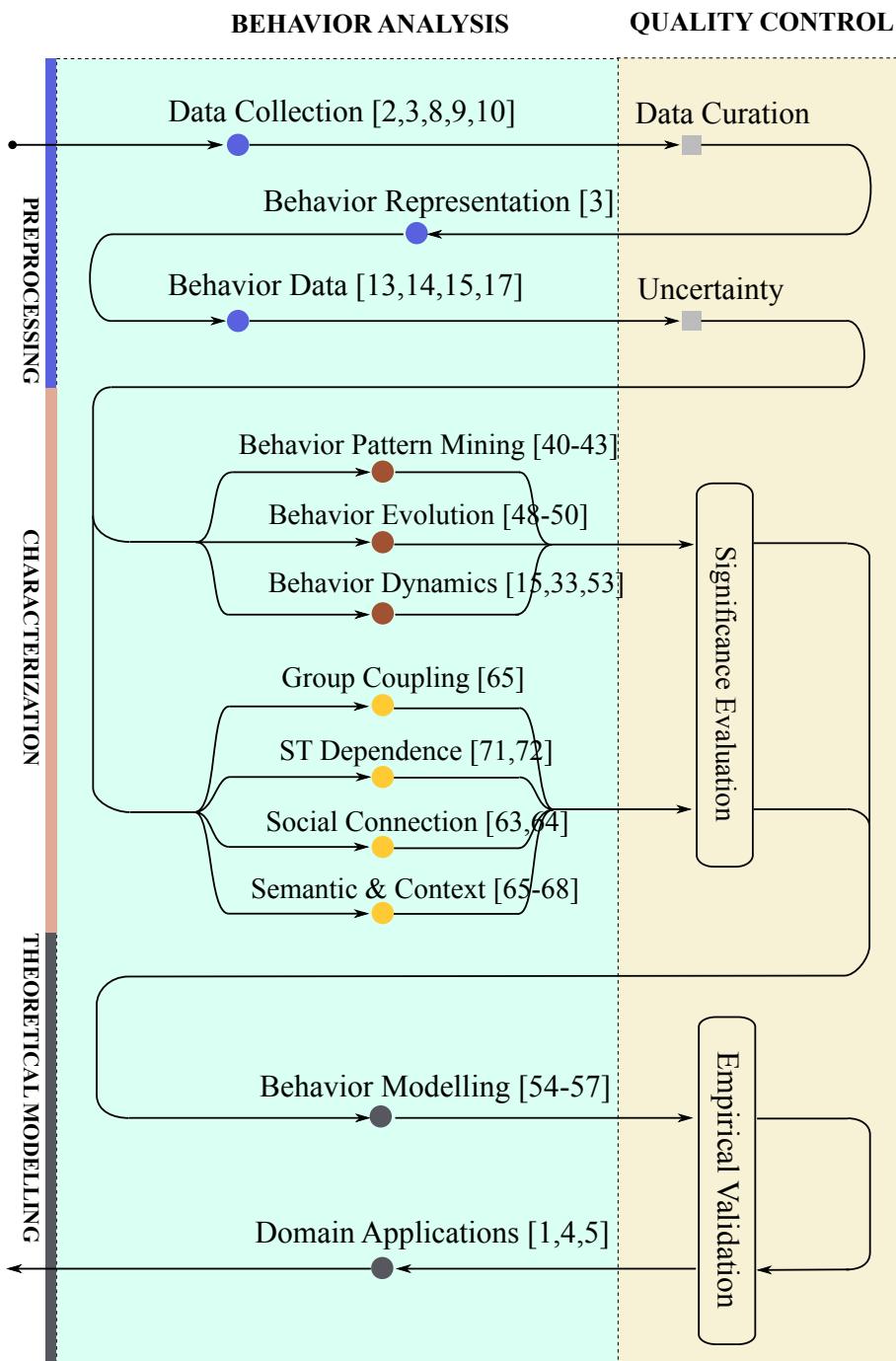


图 1-3 基于数据分析流程的时空行为挖掘成果总结

Fig 1-3 The refined framework of mining spatio-temporal human behaviour.

进而影响到时空行为挖掘算法的性能<sup>[3,67,88]</sup>。如果底层数据质量不高，上层的数据分析算法和模型便失去了可靠的基础。产生这一挑战的根本原因在于，通常的数据分析算法



0000294

对数据质量作了较为理想的假设，例如认为数据点的质量是均匀分布的<sup>[53]</sup>。虽然这样的简化有利于降低算法的复杂度，但是从另一个角度来讲，也减弱了算法对异常观测的容错能力，即鲁棒性。综上所述，在时空行为研究当中，数据质量问题对数据采集、数据清洗、行为分析、以及行为建模等环节的鲁棒性提出了更高的要求。

2) 时空行为模式的挖掘算法缺少多维度特征的关联关系。个体时空行为的差异，不仅体现在空间停留位置的变化，还体现在人们在不同时间段拥有不同的出行计划。例如，在空间上，出租车司机的出行轨迹和普通上班族有着较大的不同；在时间上，人们在工作日和周末的出行方式和目的地也有着较大的差异。在传统行为模式挖掘中，序列模式<sup>[41-43]</sup> 和移动模序<sup>[53]</sup> 算法，分别从空间位置序列和无权重的轨迹拓扑角度，揭示了个体时空行为的共同特征。但是，这两类算法由于缺失时间维度的信息，所表达的人类行为规律具有一定的局限性。在实际应用中，除了时间和空间特征，用户的时空行为与行为发生时的现实场景有着密切的联系。通常场景信息体现在位置周边区域的兴趣点 (Point of Interest, POI) 的分布上。例如，城市商业区的 POI 类型繁多，包括大型商场、电影院、各类消费设施等；而住宅区的 POI 类型以居民小区、学校、周边超市等生活服务设施为主。一个区域的兴趣点分布不同，人们在该区域的移动行为往往具有较大的差异。由以上分析可见，在时空行为模式挖掘中，不仅要时间和空间上的关联性，还应对周围场景因素的影响作系统性的量化分析。

3) 时空行为的挖掘结果缺少对可靠性的客观、有效的评估。在一个测量系统中，我们通常不仅要获得观测的量值，还需要对测量值的可靠程度进行后验式评估。这样当使用这些观测值作为另一个系统输入的时候，人们便能够有迹可循，根据输入的可靠性对输出结果的误差进行估计。与此同理，无论是从时空数据中得到的行为模式，或是行为模型，当进一步应用所得结论的时候（如利用行为模型进行新型移动协议的开发），我们都需要了解所采用的行为模式或模型的可靠性如何。这个性质具有重要的意义，然而在以往的工作中较为缺乏，因此近年来受到行为科学领域越来越多的重视<sup>[40,68,89]</sup>。总体来看，以往工作中缺乏的主要原因来自三个方面：i) 缺乏质量一致的多空间尺度下的时空行为数据集。多尺度的数据集有利于对行为挖掘结果进行横纵向比较，但是由于大规模的网络数据掌握在运营商手中，进行不同尺度的数据采集、尤其是在长时间尺度上，本身具有较大的挑战。ii) 缺少鲁棒性较高的个体移动行为模型。人类的移动行为在不



0000294

同尺度下观测，既有相似的、微观的移动行为模式（如移动序列模式<sup>[42]</sup>），也有相异的、宏观的统计特征（如行为间隔时间的分布<sup>[13]</sup>）。结合合多空间尺度的时空数据进行研究，便需要能够将微观模式和宏观统计特征在统一的框架下进行分析，而不是当前形成的两股不同的研究力量。但是将微、宏观进行统一，便需要能够兼具二者核心特性的个体行为模型。iii) 对时空行为挖掘的理论推导和实证研究不足；将时空行为的分析结果和已有的行为科学理论统一起来，不仅使得时空行为的分析结果更加完备，而且为分析方法的普适性提供了保障。

### 1.3.2 关键算法研究

自动化的时空数据挖掘和用户行为分析离不开高效的程序算法，本节对人类时空行为研究中的关键算法进行分析和介绍：

1) 时空数据的质量评估和提升算法。在挑战分析中，读者对数据质量的重要性有了初步的理解，这里对时空数据质量的算法作进一步介绍。时空数据质量的问题归根结蒂来源于数据层面上，行为在时间和空间上的不连续性。通常受到移动设备电量的限制、或用户隐私保护的考虑，数据采集模块（如 GPS）并非持续、永远在线运行，而是进行特定时间间隔的采样。时间的不连续性导致用户位置在没有数据记录的时间内是不确定的<sup>[40]</sup>，从而造成数据记录的准确度降低。另一方面，由于手机蜂窝网基站的覆盖范围通常在 500m~2km 之间，在移动网络数据中，即使用户在某时刻有数据记录，用户在空间上的位置精度也受限于单个基站的覆盖度。由此可见，时空数据质量算法的核心，便是通过量化的手段对数据的时间和空间不连续性进行客观概括。而时空数据质量提升算法是对质量评估算法的补充，在对数据时空特征度量的基础上，对缺失和错误的数据记录进行补充和修复。这类算法在数据样本极为稀少的情况下尤为重要。数据质量的提升算法，一方面可以利用行为数据的总体或局部统计特征进行插值处理，但是处理的结果趋于平滑，失去了较多的细节信息；另一方面也可以基于个体或群体的行为模式进行增强，这样的处理结果保留了用户行为的个性化信息，但是对算法设计和性能提出了更高的要求。这部分内容将在后续的第二章作进一步阐述。

2) 移动网络数据中的用户行为识别算法。移动网络数据通常采集自底层设备信息，如系统维护日志、原始网络流量等，导致所采集到的数据与上层用户的行为距离较远。



0000294

这里“距离”指用户的原始操作行为，经过多层软件和硬件的处理到达网络数据层，使得用户的行为信息变得极其稀疏。以手机浏览网页为例，当用户点击一个超链接时，设备在向主服务器请求网页文本的同时，也会从其他服务器获取网页的内嵌内容，如图片、广告、分析报告等。这意味着原始的用户行为（即一次点击），被分割到不同的网络数据包和网络流上。从网络流量中识别用户的这种服务参与行为，需要算法从原始移动网络流量中，检测出用户使用网络服务时的独立点击行为、以及相应的网络状态和场景因素。用户参与行为识别算法的挑战，不仅在于网络流量中行为信息的稀疏性，还在于移动大数据样本量大、数据传输速率高带来的处理瓶颈，这便需要在算法性能和复杂度之间取得较好的平衡。因此高效的用户行为识别算法，对移动用户的时空行为分析起到至关重要的作用。这部分内容将是第五章研究工作的基础。

3) 包含时空特征的个体行为模式挖掘算法。1.3.1节介绍了多维度特征对于时空模式挖掘的必要性。在利用程序对行为模式进行自动化处理时，需要在算法设计上实现时间、空间、以及场景特征（如POI类型）的“有机”融合。在大多数的传统模式挖掘算法<sup>[45-47]</sup>中，用户的行为模式表示为带有时间戳的符号序列。这类算法的特点是将用户的空间转移行为看作马尔可夫过程，下一次停留地点受最近的停留地点所决定，而时间仅仅以附加权重的形式出现，导致时间和空间的交互信息的缺失。尽管时间和空间表征了用户行为的不同维度，但是二者之间的交互作用已经得到了实验的证实，如不同时段人们的通勤行为<sup>[66]</sup>，以及时间段和网络服务偏好之间的关联<sup>[86]</sup>。同时包含时间和空间特征，并且将二者有机融合在一起的算法，我们称作时空耦合的模式挖掘算法。这类算法旨在通过将时间特征（如停留时间）和空间特征（如转移距离）同等对待，结合二者之间的依赖信息，提取出个体行为的时空模式。算法设计的核心，是如何实现时空信息的“有机”融合，这需要对人类行为的时空规律有更深层次的理解和刻画。在第三章和第五章中，我们介绍了两种不同类型的发现时空耦合的行为模式算法。

### 1.3.3 关键分析与建模研究

时空行为挖掘通常有两类目标，一是利用新型的数据源，从前人未曾观察过的角度，对个体或群体的行为特征进行探索，进而对人类的行为规律进行解读。二是在需要以人类行为模型为基础的研究（如新型移动网络协议开发）中，利用第一类研究中得到



0000294

的行为规律，对人类的时空行为进行模拟，从而在理论或实验环境下分析现实中的情形；甚至通过改变模型的参数，构造出现实观测的可能变种，对所研究的问题分析得更加全面。从这两个目标出发，本文总结出了时空行为挖掘中的三个关键子问题：

1) 时空行为的量化分析。在统计分析理论中，大多数的理论和方法（如著名的大数定理）都是在讨论，如何从有限的部分观测结果中得出可靠的分析结论。受益于移动大数据中丰富的样本，人类的时空行为分析克服了小数据上的统计规律的偏差。但同时新的问题出现，即如何利用客观、有效的指标对海量的时空记录进行总结，使得大量的观测数据变成可重复使用、或者进行比较的信息。本文提出，人类的时空行为需要从三个层次进行有效的量化，即时间动态性、空间异构性、以及时空关联性。时间动态性表示某个观测值随着时间的变化规律；而空间异构性指该观测值在空间上分布的均匀程度。在人类行为动力学中，人们已经找到了多种有效的量化指标对其进行刻画（参见1.2.1.1节）。然而，我们对于时空关联性的认识依然有限。和前两类指标相比，时空关联性反映了人类行为的深层次性质，因为这需要结合人类在空间位置和时间范围的内在联系。除此以外，外在的场景因素也是影响时空关联性指标的重要因素，例如，用户使用移动网络服务时，不仅受到个人日程安排和停留场所限制，也受到客观的网络状态、用户性格、操作习惯等的影响。在后续的第三至五章中，我们从不同观测粒度上（如个体和群体）对人类的时空行为进行量化，并构成了理论分析和建模的基础。

2) 时空行为模式的理论方法研究。人类时空行为分析的潜在价值之一，是其分析方法能够实现理论化，并推广到能够转化成类时空行为研究的问题上。实现分析方法理论化的同时，有利于从不同领域的时空问题中提取出共性部分，利用数学工具进行推演并使之完备，从而从另一个层面上加深我们对时空行为的认知。以传统时空序列挖掘为例，虽然人们在多年间发现了多种类型的人类时空行为模式<sup>[41,45,51]</sup>，并发现利用行为模式能够较好地对下一位置进行预测，但是，Song 等<sup>[40]</sup>从信息熵角度对时空行为的可预测性进行了研究，从理论上给出了任意的时空序列预测算法性能的上限和下限。这样的理论方法不仅使时空序列的研究系统化，而且对实际应用具有很强的指导意义。在时间和空间融合的人类行为研究中，系统化的理论研究尚不充分，而本文着重探索了时空行为理论方法中的时空耦合模式分析、多维度量化指标之间的结构化分析等，这些内容将在第三和第五章中进行展开。



0000294

3) 基于时空模式的行为模型研究。在对人类时空行为进行量化和理论分析的基础上,建立行为模型有助于将挖掘结果应用于实际场景中,如优化城市内的交通管理。但是,选择从个体粒度还是群体粒度建模,模型的特点和适用场景是有所差异的。对于个体模型而言,由于在大规模研究中个体数量多、各自之间的差异大,因此模型一方面要包含个体的个性化特征,另一方面又需要在统一的模型中对不同个体进行描述。一般来讲,模型对个性化特征描述得越准确,模拟生成的用户行为就越接近实际观测,对多样性的表达能力就相应地变弱,所得的模型通常也较复杂。相反地,模型对不同个体之间的共性信息越多,鲁棒性便越高,模拟生成用户的总体分布和实际便越接近,而个体化信息的损失就越大。同时做到这两方面是比较困难的,因此需要在相互制约的两个因素之间寻找到平衡点。本文在第二章就介绍了一种新型的平衡二者的个体行为模型。对于群体行为模型而言,人群在空间上的分布特征、以及时间的动态变化是其核心因素。虽然已有的群体行为模型<sup>[70]</sup>对空间的分布特征能够较好地捕获,但是缺少时间上的动态变化信息。另一方面,OD模型<sup>[20]</sup>刻画了移动网络上两点之间的人群迁移量,虽然这类模型包含了人群的动态变化,但主要是网络链路上的局部动态信息,缺少全局特征,例如空间范围内的人群分布的关联关系<sup>[90]</sup>。针对以上不足,本文第四章提出了一种融合了空间分布和时间动态性的群体移动模型。

## 1.4 本文的主要研究工作和创新点

面对时空数据挖掘的诸多挑战,本研究基于以多空间尺度数据为支撑、时空特征关联为核心、理论方法为保障的解决思路,对人类时空行为进行了系统性的量化分析和理论建模研究,内容框架如图1-4所示。具体而言,主要创新点包括:

1) 基于多空间尺度的移动网络数据,提出一种时空行为数据质量的评估和提升方法。该方法基于时空数据的一般特点。从单数据点、单用户样本、和群体观测角度,对数据质量进行客观量化。首先,针对单数据点在时间和空间上离散的特点,分别从静态的时空分辨率和动态的转移过程对数据采集的质量进行评估,从而有利于对不同数据集、以及同一数据集的不同时刻进行比较分析。接下来,从单数据点的质量指标引申出用户轨迹的数据采集质量,从时空异质性角度对单用户的轨迹质量进行量化,并和传统信息熵的度量进行了比较。最后,从群体观测角度,我们基于用户的轨迹样本在空间上

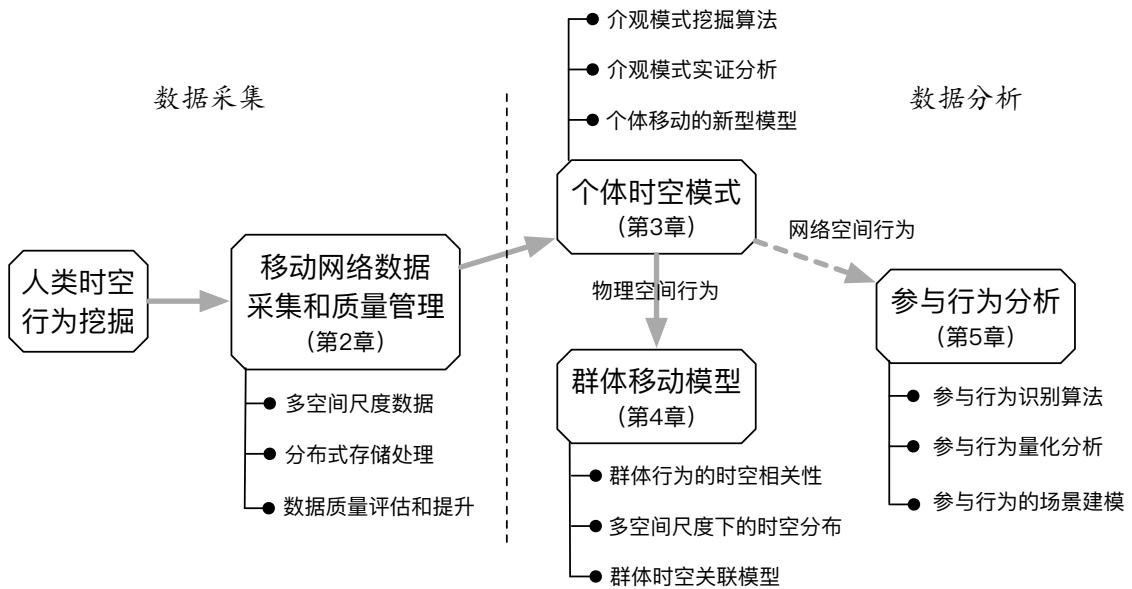


图 1-4 本文主要研究内容及创新点概括

Fig 1-4 The diagram of key contributions in this work.

的分布特征，利用用户轨迹的质量指标代替传统方法中的数据点规模，提出了针对弱数据质量的提升算法，并和传统的数据点规模的评估方法进行了比较，其准确度得到了较大提高。详细内容参见第二章。

2) 从网络结构出发，提出了个体移动行为的介观模式，并对介观模式的提取算法、实证分析、以及一种新型的个体移动模型进行了系统性研究。本文对传统个体微观序列模式<sup>[46,47]</sup>和宏观统计模式<sup>[14,15]</sup>进行了扩展，结合个体移动的网络拓扑和时空属性特征，提出了个体移动行为的介观模式（Mesostructure）的概念。从而将网络分析方法引入到个体时空行为的研究当中，为探索人类行为规律提供了新的视角。为了从大量原始记录中提取介观模式，我们提出了拓扑和属性结合的图相似匹配算法 TACSim。结合不同个体行为之间的相似性，进而提出一种带有修剪技术的显著介观模式提取算法 PPM，实现了对用户群组的介观模式提取。利用城市尺度的观测数据，对介观模式进行了实证分析，并和传统的移动模序分析<sup>[53]</sup>进行了比较。在介观模式的自距离分析中，我们发现了介观模式的自距离与移动行为的结构异质性紧密相关，并表现出四种相关关系，即零模式、对数模式、线性模式、以及随机模式。最后，基于得到的介观行为模式，提出了



0000294

一种鲁棒性更好的个体移动模型，即流涌现模型 FEM。由于该模型建立在机遇资源的空间分布和干扰机遇的框架<sup>[19,20]</sup>之上，从而摒弃了传统模型中微观和宏观统计一致的假设<sup>[14]</sup>，为连接微观移动模式挖掘和宏观统计分析提供了基础。详细内容参见第三章。

3) 同时考虑空间分布与时间动态特征，对不同空间尺度下的群体时空行为进行实证分析和建模研究。本文利用三种不同空间尺度（校园、城市、国家）下的移动网络数据，对人群在较大尺度下表现出来的“潮汐效应”进行了实证研究。首先，我们利用协方差方程对群体的时空依赖关系进行描述，分别从时间和空间维度对群体行为的统计特征进行度量。这样的建模方法既包含空间上的分布特征，也包括时间上的时律性。我们发现，国家尺度上的资源分布和校园尺度上的群体构成，对人群时空分布具有相似的影响，而城市尺度上的区域功能差异则表现出不同的影响特征。在较大空间尺度下（如城市和国家），人群聚集度较高的区域动态变化范围反而相对较小，其原因在于人群移动的“莱维飞行”特性更加突出，倾向于以非常小的概率进行长途的城际旅行。基于所观测到的群体时空关联关系，在考虑空间不同区域差异性的前提下，本文提出了基于盖内特分布的群体行为模型，并利用城市尺度下的人群分布预测对模型性能进行了验证和分析。实验结果证明，融合了时空关联信息的模型，在不同观测时间段内均表现出较好的预测性能，且预测准确度提高了约 3.7%~23.6%。详细内容参见本文第四章。

4) 提出一种被动的用户行为识别方法，对用户参与行为进行结构化分析，并结合场景因素进行建模研究。本文将物理空间的移动行为和网络空间的参与行为在形式上进行了统一，利用服务类型序列代替空间位置序列，对移动用户的参与行为模式进行挖掘。针对移动用户参与网络服务的时空行为，提出一种基于被动测量的行为识别算法 *AID*。该算法充分利用网络访问请求之间的逻辑约束条件，克服了移动网络流量引用关系缺失的挑战。通过与客户端采集的基准数据进行比较，算法的识别准确度比已有的流结构算法提高了 10% 以上。通过量化用户参与行为的重要指标，建立了参与行为和底层网络性能之间的联系，展示了不同硬件平台下用户参与行为随网络性能之间的变化关系。进而提出了一种结构相关性分析（Structured Correlation Analysis）的方法，对场景因素（即用户个性、应用类型、地点熟悉度等）如何用户的参与行为进行了细粒度的量化分析。最后，基于对参与行为时空特性的研究，提出了利用隐马尔可夫过程的参与行为建模，并对群体参与行为进行了聚类分析。详细内容参见本文第五章。



0000294

## 1.5 本文的结构安排

本文主要内容结构安排如下：第二章首先介绍了本研究使用到的多种来源、不同空间尺度上的移动网络数据集，并对网络流量的采集处理平台、用户行为识别算法进行了介绍；结合提出的数据质量管理框架，对本文使用的时空行为数据集质量进行分析和比较。第三章对个体移动行为进行研究，介绍了新型的个体介观模式挖掘算法，以及和传统模序分析的比较；基于介观模式提出了一种鲁棒性更好的个体移动模型。第四章作为对移动行为分析的延续，从群体角度对不同空间尺度上的时空分布特征进行了研究，并对群体的时空依赖性建立了统计模型。第五章将多维度融合分析的思想应用在用户参与行为上，在对用户参与行为进行量化分析的基础上，建立了相应的参与行为模型。第六章是对本文研究工作的总结、及未来研究方向的展望。



0000294



0000294

## 第二章 时空行为数据采集及质量管理

随着移动技术的快速发展，移动服务和应用已经深入人们的生活。作为丰富的用户行为信息来源，移动网络流量由于体积大、速度快、结构多样等特点。从网络流量角度来看，用户原始的行为特征或多或少受到了损失，并被编码成字节流在网络中传输，因此从网络流量中“逆向工程”用户的行为信息，不但需要高性能的采集处理平台，还需要高效的用户行为识别算法。高质量的用户行为画像不仅取决于行为识别算法，还受到数据源本身质量的限制，比如在同等分析需求的前提下，较高的空间分辨率往往能提供更多的行为信息。本章节作为后续几章行为分析内容的基础，主要介绍本文采用的数据采集分析平台、多空间尺度的数据集、以及时空数据集质量的量化评估方法。

### 2.1 移动网络中的时空行为数据采集

#### 2.1.1 移动流量采集系统

##### 2.1.1.1 系统架构

移动网络流量的采集方式分为主动式和被动式两种。主动式采集通过向网络注入流量，获得对目标协议、网络链路的性能评估数据。这种方式是按需测量的一种体现，具有明确的目的性，因此能够实现参数控制下的多次测量；不足之处在于注入流量对网路来说是无差别的，容易对网络的正常服务造成影响，因此大规模的单点测量是比较困难的。被动式采集也称网络监听、网络嗅探，是通过在网络流量端口处旁置软件或硬件探针，实现对网络流量无干扰的采集和分析。由于这种方式对网络的正常服务没有干扰，且易于实现大规模部署，因此本研究主要采用被动的方式对移动网络流量进行采集。

通信网络中有许多不同类型的组件，被动探针的放置位置不同，所需的软件和硬件性能也将有所差异。移动网路流量的采集点通常有三种：移动终端、网络中间件、和服务器端。1) 移动终端采集需要研究人员开发网络嗅探工具，并在用户终端设备上进行安装。这种采集点的优势在于可以实现与用户的交互，获取用户在设备上的各类操作、

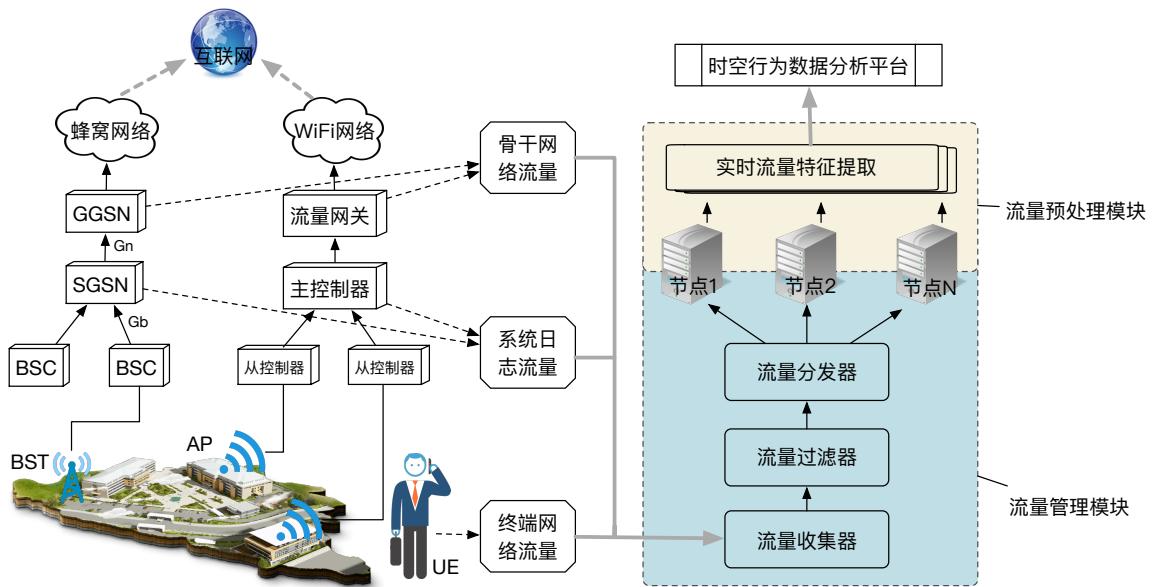


图 2-1 被动式的移动网络流量数据采集系统架构

Fig 2-1 The architecture of passive collection system for mobile network traffic.

各类服务的使用数据，从而得到立体的用户行为画像。同样地，这将带来对用户隐私泄露的担忧，如恶意代码的嵌入可能导致探针对用户敏感数据的收集，进而终端采集的初期部署较为困难，为大规模数据收集带来困难。2) 服务器端采集是将探针部署在数据中心内服务器访问端口附近、或直接读取服务器日志信息。因为通常需要服务提供商支持，这种采集方式的优势在于可以获取用户对于某类服务的完整行为信息，以及从网络设计逻辑推断出的用户行为逻辑。不足之处在于一般情况下研究人员很难获得对数据中心的访问权限，因此这类采集多适用于对某类服务拥有访问权的机构；另一方面，服务器端采集因为只针对某种服务的采集数据，因此给进行多服务类型的行为分析带来挑战。3) 网络中间件采集是将探针部署在网络链路或网关端口附近，实现对流量的无差别监听。这种方式的优势在于用户和服务多样性大大提高，加密协议也防止了用户敏感信息的泄露，同时利于大规模的网络数据采集。不足之处在于用户的行为信息被隐藏在原始的网络数据包之中，因此需要开发高效的识别算法将用户的行为过程（部分）恢复出来。综上所述，本研究中除了少数基准数据集采用终端采集的方式以外，主要网络流量数据采用网络中间件采集的方式。

图2-1展示了研究中采用的移动网络流量采集系统的架构。该系统整体基于被动式

采集的原理，图左侧所示为不同无线网络场景下的关键节点设备，细虚线箭头表示探针所在位置及数据流向，其中骨干网络流量通常采集自网关镜像端口，系统日志流量来自网络管理设备，终端网络流量采集自用户的移动终端；图右侧为研究中采用的流量采集系统内部模块结构，主要包括流量管理模块和流量预处理模块，其中流量管理模块分为流量收集器、流量过滤器、以及流量分发器，流量预处理模块基于并行处理的思想，利用多进程或多节点完成对流量特征的实时提取。这部分我们首先对图左侧的两种无线网络原理作基本介绍，下一小节将对采集系统的关键组件进行详细说明。

蜂窝网络是商用无线网络中普及度最高的技术，因网络中各基站的空间模型类似于蜂窝状六边形而得名。蜂窝网络基站的覆盖范围通常为 0.5~35km，其名称在不同的网络技术中称呼不同，如 **GSM** 网络中为基站收发信机（**BST**），而在 **WCDMA** 系统中称为 **NodeB**。这里以 **GSM** 网络为例、从数据传输角度对蜂窝网数据采集原理进行介绍。在数据传输过程中通常涉及四个部分的组件：用户终端设备（**UE**）、无线接入网络（**RAN**）、核心网（**CN**）和公共网络（**PN**），其中无线接入网络和核心网是移动服务提供商拥有和维护的主要部分。无线接入网络包括基站收发信机和基站控制器（**BSC**），核心网包括 **GPRS** 服务支持节点（**SGSN**）和 **GPRS** 网关支持节点（**GGSN**）。基站收发信机负责接收用户终端设备的通信数据，并通过基站控制器传送到核心网中；核心网负责移动网络内、外部数据传输、路由、计费等，其中 **GPRS** 支持节点从 **Gb** 口接收基站控制器传送的数据信号，并通过 **Gn** 口发送到 **GPRS** 网关支持节点。在本研究中，被动探针位于 **GPRS** 网关支持节点的 **Gn** 端口附近，将双向的 IP 网络流量镜像到数据采集系统中。

**WiFi** 网络是局域网构建在无线技术上的一种典型技术，实质上是一种商业认证<sup>1</sup>，即“无线相容性认证”。**WiFi** 网络的末端的设备称为接入点（**Access Point, AP**, 或 **WiFi** 热点），在这个接入点电波覆盖的有效范围内可使用无线保真的连接方式进行联网。**WiFi** 接入点的信号覆盖范围通常为 0.01~0.5km，因此在家庭、企业、校园的内部网络设施中常见，在实际应用中对流量资费较高、室内信号不稳定的蜂窝网络形成了互补。本研究以上海某高校的校园 **WiFi** 网络为采集对象，主要部分包括：**WiFi** 接入点、主从控制器、流量网关设备等。该网络中 **WiFi** 接入点部署在校园主要建筑和活动场所内，负责将用户终端设备发送的流量传输给控制器。主从控制器协调工作，负责用户身份认证、IP 分

<sup>1</sup><http://www.wi-fi.org/>

配、计费、管理、日志等功能，最终将外部通信数据通过网关设备传送到公网上。因为目前的 WiFi 接入点采用“瘦 AP”的设计方式，其性能和软件环境受到限制，无法直接安装数据探针，因此和蜂窝网络中类似，骨干网路流量从网关设备的镜像端口传送到数据采集系统，同时用户身份认证、IP 分配等系统日志流量通过主控制器进行收集、并转发给数据采集系统。

### 2.1.1.2 系统关键组件

移动网络流量数据采集系统离不开软件、硬件的支持，尤其面对骨干网络的高速流量，软件的处理流程设计对数据采集的实时性能起到关键作用。本节对采集系统中的关键组件、以及基于廉价商用服务器的各组件软件设计进行了介绍。

**流量收集器**负责将镜像流量、系统日志等采集存储下来，或转发给后续组件对流量作进一步处理。通常情况下，移动终端和网络中间件的被动采集原理是相同的，但工具形式不尽相同。对于移动终端采集，本研究中作者开发了基于 Android 系统的 OmniPerf<sup>1</sup>数据采集软件，实现对用户位置移动和上网行为的基准数据采集。该软件采集到的主要数据类型如表2-1所示。

表 2-1 OmniPerf 采集主要数据类型说明

Table 2-1 Data specification for OmniPerf of Android.

数据类型	来源	说明
PCAP 流量	网络端口	利用移植的 libpcap 和 tcpdump 对原始流量被动监听
位置数据	GPS	调用 LocationListener 获取位置经纬度
屏幕点击	系统调用	新建 WindowManager 获取用户点击动作 ( $\leq 2.3$ )

对于网络中间件采集，主要挑战之一是高速率的原始网络流量，解决方案一般有：

- 1) 专用网络流量设备，为商业公司所开发，如 EmbedWay<sup>2</sup>的 ExProbe 系列，具有性能好、针对性强、软件升级及时的优点，不足在于升级成本高、功能定制化小、与其他软件兼容性低；
- 2) 网络处理器，如 Netronome 公司<sup>3</sup>的 NFE 系列网络适配器，以 PCIe 接

<sup>1</sup>OmniPerf 源码：<https://github.com/caesar0301/OmniPerf>

<sup>2</sup><http://www.embedway.com/>

<sup>3</sup><http://www.netronome.com/products/>

口的形式接入商用服务器，成本相对较低，但是需要特定的软件开发包（如 NFM）支持，因此需要对其他底层库（如 libpcap）进行定制、重新编译后才能使用；3) 普通万兆网卡 + 开源软件，和前两者相比，普通万兆网卡的成本最低，而且通过 PCI/PCIe 接入廉价商用服务器后，对上层程序和软件包透明，无需进行定制或重编译；不足之处在于性能受到 CPU 和软件处理性能限制。本研究从硬件成本和软件复用度考虑，采用第三种方案进行骨干网络流量采集。

**流量过滤器**的作用在于，在数据流进入分析工具前，通过过滤掉不必要的信息和数据实现提高采集系统性能的目的。该组件包括三种类型的过滤功能：1) IP 地址范围过滤，实现匹配网络数据包中符合目标网络的 IP 地址段；2) 网络协议过滤，本研究关注用户的网络移动和网络参与行为，而这些信息主要包含在 HTTP 流量当中，因此通过 TCP 协议端口<sup>1</sup>对 HTTP 流量进行过滤；3) 日志类型过滤，对于系统管理日志，该组件主要过滤出包含连接管理、用户身份认证、IP 分配、空间移动等类型的日志信息。

**流量分发器**利用“分而治之”的思想，对过滤后速率依然较高的数据源，按照一定的规则进行分组并行处理。分组功能包括镜像和分割，前者表示将单一数据源复制成多份，并按照不同的需求进行特征解析，后者表示将同一数据源分成无交叠的部分，并将各部分分配到不同的线程或节点并行处理。针对这样的功能，本研究中作者开发了面向日志数据的 RELOGGER<sup>2</sup>程序，该工具采用 RFC 822 标准的配置文件格式，实现了对服务器、端口、文件等日志源、目的的灵活配置。RELOGGER 从多个源地址读取数据，并通过多线程的方式进行规则匹配和数据转发，从而将较小的数据流发送到不同的节点上进行处理（图2-1）。

**流量预处理**以实时的方式对分发到不同处理节点上的网络数据进行行为特征抽取，并传送给后端的数据处理平台，以达到对移动网络流量持续采集的目的。在数据不断累积的情况下，如果原始数据处理不充分，冗余信息增多，对数据存储以及数据处理系统的后期处理复杂度就会增加；如果特征提取过于复杂，又会影响实时处理的性能。因此在处理性能和特征提取复杂度上应该有所权衡。本研究中作者开发了 HTTP-SNIFFER<sup>3</sup>程序对原始网络流量的 TCP 和 HTTP 协议特征进行解析。

<sup>1</sup>过滤 HTTP 流量主要端口包括 80, 8080, 3128, 8081, 9080, 8000, 8001 等。

<sup>2</sup>RELOGGER 源码：<https://github.com/caesar0301/relogger>

<sup>3</sup>HTTP-SNIFFER 源码：<https://github.com/caesar0301/http-sniffer>

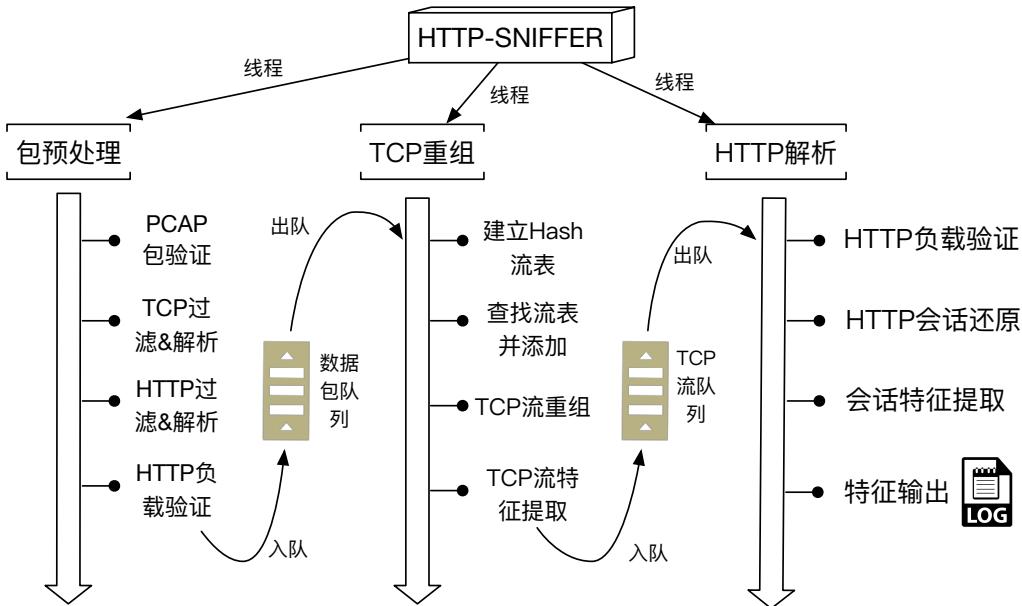


图 2-2 网络流量特征实时提取工具 HTTP-SNIFFER 设计结构

Fig 2-2 The design architecture of HTTP-SNIFFER for online network traffic processing.

**HTTP-SNIFFER** 是一个面向高速网络流量的多线程处理工具，其内嵌的算法能够高效地对 TCP 进行流重组、特征提取、HTTP 会话还原等，并支持多种格式化文本输出。**HTTP-SNIFFER** 的处理流程和设计结构图2-2所示。该设计充分考虑了网络流量高速率特点，将资源消耗较多的 TCP 流重组和 HTTP 会话还原分别利用独立的线程来处理。HASH 流表采用 TCP 流的四元组（源、目的 IP 地址和源、目的端口）作为主键，并利用双向链表解决 HASH 冲突问题。对于每个 HTTP 数据包，对数据包的统计特征记录以后，HTTP 协议头部以后的其他字节将被丢弃以提高内存的利用率。在 HTTP 会话还原中，请求和应答消息按照时间顺序依次匹配，即新的应答消息匹配处于空闲状态的最早的请求消息。而没有请求消息的应答消息将被丢弃，反之则保留下。

**HTTP-SNIFFER** 的性能数据由表2-2给出。实验利用校园 WiFi 网络中真实的 HTTP 流量和网络流量生成软件，合成不同速率（1Gbps 和 10Gbps）、不同包大小（64、512、1518 字节）的源数据流，然后利用 **HTTP-SNIFFER** 对 HTTP 流量进行分析，并记录性能数据。由表中可以看出，对于低速率的网络流量，**HTTP-SNIFFER** 能够以 100% 的处理速率对流量特征进行提取。对于较高的网络流量速率，**HTTP-SNIFFER** 依然能够



0000294

表 2-2 网络流量特征实时提取工具 HTTP-SNIFFER 性能验证

Table 2-2 The performance evaluation of HTTP-SNIFFER under different real-time traffic scenarios.

流量速率	包长度(字节)	包发送速率	包处理速率	处理比例
1Gbps	1518	6,612,327	6,612,327	100%
1Gbps	512	8,839,346	8,839,346	100%
1Gbps	64	26,609,540	26,609,540	100%
10Gbps	1518	20,675,640	20,675,640	100%
10Gbps	512	32,091,505	31,910,392	99.44%
10Gbps	64	94,528,167	32,586,237	34.47%

以 99% 以上的比例分析较大的数据包（即 512 和 1518 字节）；而对于较小的数据包，HTTP-SNIFFER 在 10Gbps 的链路上丢包速率达到 65.53%。这样的原因在于，小体积的数据包带来较大的包数目，对 TCP 的流重组带来较大的压力；当数据包队列达到设置上限时，便出现丢包现象。在研究当中，校园 WiFi 网络网关万兆端口的 HTTP 流量峰值约为 3.5Gbps/s（下行）和 530Mbps/s（上行），且现实中 HTTP 流量数据包平均体积大于 512 字节，因此 HTTP-SNIFFER 能较好满足本研究中的数据采集需求。

## 2.1.2 时空行为数据分析平台

移动流量采集系统（图2-1）以实时的方式输出用户移动行为和上网行为数据，我们的时空行为数据分析平台的主要功能便是对这样的流式数据进行存储、质量管理、以及分析。如图2-3所示，数据分析平台主要包括三个部分：流式数据缓冲队列、实时和离线处理引擎、以及基于 HDFS 的数据仓库。接下来我们对各部分的角色、功能、以及设计考虑进行介绍。

为了实现对采集到的实时流量灵活利用，我们在 Lambda 架构<sup>1</sup>中添加了基于 Apache Kafka 的实时数据的缓冲组件，并实现对流式数据的话题管理（每个话题代表一个独立的流式数据源）。话题管理首先能够对数据接入源进行监控和质量管理，对非法数据源进行拦截，对合法数据源进行数据质量监控，并及时反馈给上游数据提供者。另一方面，话题管理能够实现对外部数据应用的授权，这样外部应用能够利用现有的话题，并将处

<sup>1</sup>Lambda 架构由 Nathan Marz 提出来解决大数据平台的可扩展性问题，参考 “How to beat the CAP theorem”，2011。

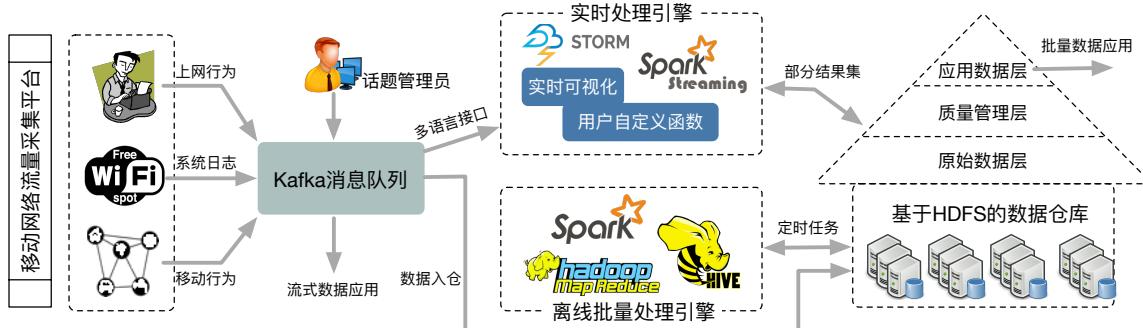


图 2-3 基于 Lambda 架构的时空行为数据处理平台

Fig 2-3 The Lambda architecture of spatio-temporal data analysis platform.

理后产生的新话题加入到消息队列当中供其他应用使用。

计算引擎是行为数据分析平台的核心，该平台目前支持在线流式处理和离线批量处理两种模型。流式处理实现对数据的不间断处理，通常采用单数据条目或时间窗的形式对数据进行解析和分析，适用于实时分析应用如网路状态监控等。本研究基于 Apache Spark Streaming 构建实时处理引擎，该组件采用时间窗和小批量处理的方式对数据进行近实时（如 1s）分析，适合对实时性要求不高的场景。批量处理是采用离线的方式对数据进行定时处理，适用于需要有一定数据积累的应用，如统计报表、行为模式挖掘等。本研究采用 Apache Spark 构建离线处理引擎。

数据仓库<sup>1</sup>是行为数据分析平台的基础，其构建通常需要考虑到可扩展性和数据应用方式。从扩展性角度来讲，由于移动网络流量数据体量大、持续时间长，因此采用可扩展性和容错性较好的分布式数据库进行存储较为理想。本研究采用 HDFS 作为行为数据仓库的存储介质，由于 HDFS 有着丰富的软件接口，因此对于非大数据软件也较为友好。数据仓库的数据 IO 有三种类型：流式数据入仓、实时分析的中间结果存储、离线分析输入输出。流式数据入仓是对原始数据的沉淀，将数据从缓冲队列中单向存储到 HDFS 介质上，作为数据的持久化和后续分析的基础。实时分析的中间结果指将在线分析中的中间结果、生成小数据等存储在仓库中，其中中间结果存储是双向 IO 操作。离线分析入输出是将有一定时间积累的数据进行批量处理，通常 IO 资源消耗较高。

数据仓库的另一个方面便是数据的应用方式，这也构成了数据仓库的组织逻辑。如

<sup>1</sup> 数据仓库源代码：<https://github.com/OMNILab/OmniDataHouse>



0000294

图2-3右侧所示，从逻辑上来讲，本研究中的数据仓库分为三个层次：原始数据层（L0层）、质量管理层（L1层）、以及应用数据层（L2层），各层的数据格式和内容均不相同，信息量也逐层递减。L0层是对原始流量数据的持久化存储层；该层数据是原始数据的直接存储，因此信息量是最大的一层，通常数据仓库管理员具有访问权限。L1层是对原始数据进行清洗、修复、添加标签等质量管理工作，单一数据源的不同特征并列存储，形成一张逻辑上的“宽表”，通常数据仓库的核心开发人员具有访问权限。L2层是通过数据分析和挖掘的方法（如关联分析、模式挖掘等）得到的针对不同应用场景的小数据集，形成许多张逻辑上的“窄表”；这样的应用数据集相互之间因为目标不同会有所冗余，但其使用较为便捷，通常L2层数据以接口的形式向外部应用开发人员所开放。

**软硬件配置：**该平台部署了Cloudera公司的CDH（v5.4.4）社区版，由3台主节点（Master node）和8台从节点（Slave node）组成，节点间通过40Gbps交换网路连接。主节点主要负责任务调度、数据目录存储、任务提交等，从节点负责HDFS中数据的实际存储以及计算任务的运行。主节点和从节点分别拥有48GB/台和128GB/台的内存空间。外置存储方面，每个从节点拥有12TB/台的机械硬盘存储；同时整个集群配置300GB×2的SSD存储，以满足中间生成数据对IO的高吞吐需求。

## 2.2 多空间尺度的时空行为数据集

这部分对本研究中使用的多尺度时空行为数据集进行介绍。从采集方式和尺度上来分，包括四个主要的数据集：终端采集的用户点击数据集、校园WiFi网络数据集、城市移动网络数据集、以及国家移动网络数据集。下面对各数据集的采集环境和数据特征进行详细说明。

**用户点击数据集（CLICK）**是利用OmniPerf采集工具从用户终端上被动监听到的数据集，由于包含精确的用户点击行为信息，因此成为验证下面的用户参与行为识别算法性能的基准数据。该数据集采集自OMNILab<sup>1</sup>的内部成员，包括12位学生以及3位研究员。数据采集流程为：每位参与人员在签署隐私协议的前提下将OmniPerf软件安装在个人手机设备上，并开启自动运行模式；程序在后台运行并将采集到的数据保存在本地，然后以一定的频率（如24小时）将采集数据上传到数据收集器上。数据采集持

<sup>1</sup><http://omnilab.sjtu.edu.cn/>



图 2-4 WIFI-T 数据中的 TCP 性能参数和 HTTP 会话特征示例

Fig 2-4 The illustration of WIFI-T data entries.

续时间为四个星期，每个用户平均每两天上传一份数据，每份数据平均记录了5~10分钟的使用行为。

**校园 WiFi 网络数据集**采集自上海某高校的校园 WiFi 网络，目标网络是覆盖校园内的主要建筑、马路、以及部分开放区域，拥有 2.7K 个 WiFi 接入点，服务于约 80K 的校园用户。网络对校园用户免费开放，由于覆盖了学校的主要区域，因此用户能够在校园内漫游（Roaming）使用。该数据集自 2013 年 4 月起至今持续采集，已累计采集~10TB 的行为数据，分为上网流量（WIFI-T）和用户移动（WIFI-M）两个子数据集。

**WIFI-T** 数据集包含用户的上网使用行为，即 **HTTP-SNIFFER** 解析获得的 **TCP** 流数据和 **HTTP** 会话特征（如图2-4所示），每天数据采集量为 1.5 亿条。**TCP** 流数据记录了用户使用网络服务时的网路性能状态，即网络质量（QoS）指标，记录字段内容如 RTT、带宽等。**HTTP** 会话特征记录了负载在 **HTTP** 协议上的应用使用行为，如应用域名、请求内容类型、内容地址、以及衡量 **HTTP** 会话的性能参数。每条流数据和会话特征以独立的行式条目存储，每个条目包含了用户身份标识：用户账号标识 **ID** 和用户设备 **ID**。**账号标识 ID** 来源于匿名化的网络账号，其作用反映出了个体用户的连续行为，即使使用不同的网络设备；同时账号标识 **ID** 关联了部分用户特征，即性别、年级、年龄、入学年等。**设备 ID** 来源于匿名化后的用户设备地址（**MAC**），其作用在于反映了对不同类型、性能的硬件设备的行为，即使公共设备被多个用户账号使用。同时结合设备的软件指纹（即 **HTTP** 头部的 **User-Agent** 字段）和硬件指纹（即设备机构为亿标识符，**OUI**），我们对用户设备的类型（移动和非移动）进行了识别<sup>1</sup>。

<sup>1</sup>特别的，由于平板设备用户的平均停留地点数目较少 ( $\bar{n} = 2.3$ )，因此分析中将其归类为非移动设备。

### 原始系统日志

```
1449963445646 <141>Dec 13 07:29:18 2015 SJTU-Local3 mobileip[2209]: <500010> <NOTI> <@> <@> Station 60:fe:c5:6b:fa:4b,
10.188.71.9: Mobility trail, on switch 10.190.3.1, VLAN 1003, AP XH-ZY-3F-04, SJTU/6c:f3:7f:34:9f:18/a
1449963445741 <141>Dec 13 07:18:42 2015 SJTU-Local5 mobileip[2161]: <500010> <NOTI> <@> Station 00:ee:bd:88:8c:c3, : Mobility
trail, on switch 10.190.5.1, VLAN 1005, AP D3ST-1F-01, SJTU/6c:f3:7f:5a:cc:61/g
```

### WiFi-M数据集

38591392e5,2013-04-23 23:53:27,5,111.186.25.18 6ce873c244c4,2013-04-23 23:53:07,5,111.186.35.110 ec852f74582d,2013-04-24 00:00:26,6,111.186.4.109 c46ab78a4592,2013-04-23 23:53:07,5,111.186.33.47 c46ab78a4592,2013-04-23 23:53:07,5,111.186.33.47	00127b63f31d,1366877578000,1366877578000,YXL-4-A-3F-06,31.03714,121.449829,, 00127b63f31d,1366877578000,1366877578000,YXL-4-A-4F-06,31.03714,121.449829,, 00127b63f31d,1366877589000,1366877589000,YXL-1-4F-03,31.035960,121.4450403,, 00127b63f31d,1366877589000,1366877589000,YXL-4-A-4F-06,31.03714,121.449829,, 00127b63f31d,1366877619000,1366877619000,XXZL-A-4F-05,31.03403,121.447545,
---	--

图 2-5 原始系统日志和 WiFi-M 数据集示例

Fig 2-5 The illustration of WiFi-M data entries.

WIFI-M 数据集包含用户在校园 WiFi 网络中的移动行为信息，通过对 WiFi 控制器系统日志的解析获得（如图2-5所示）。我们用正则匹配的方式从原始日志数据中解析获得用户移动行为相关的信息，每条信息记录了用户的账号标识 ID、设备 ID、时间、WiFi 连接点。从时间角度上，我们从记录条目中提取出连接到不同 AP 上的停留时间，并以此表示用户在 AP 对应的地理位置的停留时间。多个 AP 序列构成了 WiFi 会话，且两个 WiFi 会话（Session）之间间隔系统设定的超时时间（如 30 分钟），或用户主动发起断开连接的动作。为了对网络参与行为和移动行为进行联合分析，我们通过分配到的 IP 地址将 WIFI-T 和 WIFI-M 子数据集关联在一起，即同时获得用户在不同时间和地点的网络服务使用行为。

**城市移动网络数据集**反映了城市尺度上用户的移动行为和参与行为特征。我们采集并分析了中国某大型移动网络服务商在中国东部某发达城市的网路流量数据。该网络覆盖城、乡区域面积约为  $50\text{km} \times 60\text{km}$ ，其中城市商业及住宅区基站部署较为稠密，郊区及县乡区域较为稀疏。原始数据的采集时间为 2012 年 8 月 16 日 ~8 月 31 日，记录了用户连续两周内的移动和网络使用行为。每个独立用户通过国际移动注册标识（IMSI）进行区分。为了重点分析城市用户的行为模式，我们选取了市区周边的  $28\text{km} \times 35\text{km}$  的活动区域，其独立用户数为 377,566，占原始数据集中用户总数的 87%，以及该城市 2012 年常住居民总数的 25%。该数据集包含 62,486,319 条记录，以及 36,633 个网络基站。

和校园 WiFi 网络数据集类似，城市移动网络从上网行为和移动行为角度分为 CITY-T 和 CITY-M 两个子数据集，且采用和 WiFi 网络类似的数据格式。由于 CITY-M 中用户



的数据点较稀少，我们通过以下质量控制条件对其进行预处理：a) 每个用户至少观测到两个不同的基站，且观测期间内超过 75% 天数拥有 HTTP 记录；b) 每个用户平均每天有 5 条 HTTP 记录，且分布在不同的半小时段内。但是和校园 WiFi 数据集相比，该城市数据集有两个方面的限制：1) 由于移动网路基站的覆盖范围更大（约为 500m~2km），因此用户移动行为分析的空间解析度最大为基站覆盖范围；2) 该数据集主要记录了用户的 HTTP 上网行为，缺少系统日志的相关信息，因此其 HTTP 流量对用户数时空分布的估计略低于实际用户数。这些限制也表现了结合多空间尺度的数据集对用户时空行为进行分析的必要性。

**国家移动网络数据集 (Senegal)** 来源于法国 Orange 移动网路服务商在非洲 Senegal 的移动网络数据<sup>[91]</sup>。该数据集基于 2013 年 1 月 1 日 ~12 月 31 日间 9 百万用户的通话记录 (CDR) 和文字消息数据生成，覆盖 1609 个独立基站和 123 个省市级行政区。研究人员对原始日志进行了预处理：首先，每个用户在观测期间内超过 75% 的时间拥有数据记录；其次，每个用户每周的数据记录不超过 1000 条，超过该数值的用户被认为是自动设备或多人共享的移动设备；最后在不同空间分辨率上将用户移动行为分为基站级别 (Senegal-S) 和行政区域级别 (Senegal-A) 的两个子数据集。

为了保护用户的隐私，研究人员从时间和空间上对数据进行了模糊处理。在时间上，研究人员以两周为最长连续观测时间，将基站级别的用户行为分割成不同的时间段，即出现在两个不同时间段内的同一用户，将拥有不同的匿名 ID；而行政区域级别的用户行为连续观测时间为一年。在空间上，研究人员通过 Voronoi 分割对基站级别的空间区域进行划分，并在不同基站区域内随机选择新的基站点作为该区域的经纬度坐标；行政区域级别则以该区域的中心作为其经纬度坐标。经此处理后，Senegal-S 的每个时间段包含约 15 万用户、5.6 亿条的用户行为记录；Senegal-A 在一年的观测时间内包含约 32 万用户、12 亿条的用户行为记录。

综上所述，本研究从不同空间尺度上对移动网路数据进行了采集，来研究用户时空行为模式的一般性和内在规律。表 5-2 从时空覆盖率以及基本维度对数据集进行了对比，虽然采集时间和用户规模有所差异，但是这些数据集在空间尺度和分辨率上互为补充。从研究客观性来讲，不同数据来源、空间尺度下的时空行为规律面临着数据质量不统一的挑战，进而破坏了行为的关联分析的可靠性，因此下一节中我们将对时空数据质量的

表 2-3 不同空间尺度移动网络数据集特征及比较

Table 2-3 Comparison of data dimensions at different spatial scales.

类型	空间范围	分辨率	时间范围	用户数	基站/AP 数	数据量
用户终端	-	1m~10m	2014.03.10 至 03.24	15	-	270MB
校园 WiFi	2.7km×1.4km	10m~50m	2014.09 至 2015.06	80K	2.7K	3TB
城市移动网	28km×35km	500m~2km	2012.08.16 至 08.31	377K	36K	500GB
国家移动网	500km×390km	500m~2km	2013.01 至 2013.12	9M	1609	50GB

量化分析进行研究。

## 2.3 时空数据的质量管理

### 2.3.1 时空数据质量的关键点

在数据挖掘任务中，分析结果的好坏，不仅决定于后期的算法和模型，而且受到前期数据质量的影响。实际的数据采集场景千差万别，是导致数据质量不一致的主要原因。对数据质量进行评估的好处在于，1) 让数据分析师在分析数据之前，便对数据的总体特征有所把握，从而对数据能分析什么、不能分析什么做到心中有数；2) 在多源数据的联合分析中，客观的时空数据质量评估，能够帮助研究人员选择出质量一致的多源数据。这里我们从时空数据的一般性出发，对时空数据质量管理中的关键点进行了概括。这里所述的关键点，不仅适用于该研究中采用的移动网络数据，也适用于其他类型的时空数据，如正在发展中的物联网数据等。

**数据记录的准确性：**数据记录的可靠程度往往和采集方法联系在一起。因为人们随身携带着手机设备，所以手机配置的传感器可以作为人类活动数据的来源。在享受这种便捷性的同时，我们同样付出了代价，即数据质量受到数据源本身的限制。采集方法和数据源本身所带来的测量误差，我们称之为系统误差。在不同的系统配置下，系统误差不同，所得到的数据质量也不一样。在时空行为分析中，用户定位信息的准确性至关重要。在单点定位技术中，用户位置被当前连接基站或热点的位置所代替，而位置精度决定于单点的覆盖范围。例如，手机蜂窝网基站的覆盖范围通常在 2~5km 之间，而 WiFi 网络热点的覆盖范围在 10~50m 之间。在多点定位技术（如蜂窝网中的多基站定位，和



0000294

GPS 网络的多卫星联合定位) 中, 虽然其精度比单点定位技术高出许多, 但其代价在于需要在客户端设备进行信息采集及定位, 使得大规模的数据收集成本显著提高。

**采集时间的连续性:** 在时空数据采集过程中, 测量方法的设计需要尽可能多地获取观测数据。对于单一测量值, 数据点采集的频率越高, 对观测指标的描述越完整, 但是所付出的代价是需要更高的传感器功耗和传输带宽。在被动采集中, 采集频率还受到用户使用行为规律的影响, 例如在移动网络中, 大多数用户每天产生 CDR 数据不超过 5 条通话记录<sup>[67]</sup>。如此稀疏的数据点, 为个体的移动行为分析增加了更多的不确定性。此外, 有效的数据采集频率也与用户的移动状态紧密相关, 在状态发生剧烈变化的时候应适当提高采集频率, 如移动速度快的时刻比速度慢的需要更多的数据点。虽然从数据挖掘角度期望更高的采集频率, 但是出于隐私保护的考虑, 或采集方法本身的制约, 我们总是难以得到具有统一时间特征的多源数据集。因此需要客观的时间连续性的量化方法, 以便对行为挖掘的算法性能和结果有效性进行把握<sup>[92-95]</sup>。

**空间分布的合理性:** 表示数据集合对观测指标的真实空间分布的描述质量。数据样本分布越接近真实过程, 则数据质量越高, 而其中采样 (Sampling) 是决定空间分布的关键因素之一。数据采样技术通常贯穿时空数据挖掘的整个生命周期。在数据采集阶段, 采集到的个体和潜在的整体相比, 本身便是进行了一次采样, 这二者之间的差异和偏差近年来也得到了研究人员的注意<sup>[96]</sup>。在数据预处理阶段, 当部分个体的数据点过于稀疏, 直接使用会引入较大的误差, 我们往往会对群体进行抽样, 例如随机抽取一部分用户<sup>[16]</sup>, 或者选择数据点数目满足一定阈值的用户<sup>[40]</sup>。在模型训练阶段, 交叉验证的思想本质上是基于训练样本的采样, 来提高模型性能的鲁棒性。虽然已有的研究工作中或多或少采用了数据采样技术, 但是不同的采样方法对数据分布有怎样的影响, 依然是时空数据质量管理中需要考量的一个重要方面。

### 2.3.2 时空数据质量的量化评估与提升

针对以上数据质量的关键点, 本文提出一种多层次的时空数据质量评估方法, 该方法结合单数据记录的局部信息, 和用户移动轨迹的全局信息, 对时空数据质量进行客观量化。进而根据质量指标的分布特征, 提出了移动网络数据质量的提升算法。

**时空数据点质量:** 时空数据记录作为数据集中最小的单元, 如果想要准确刻画数据



0000294

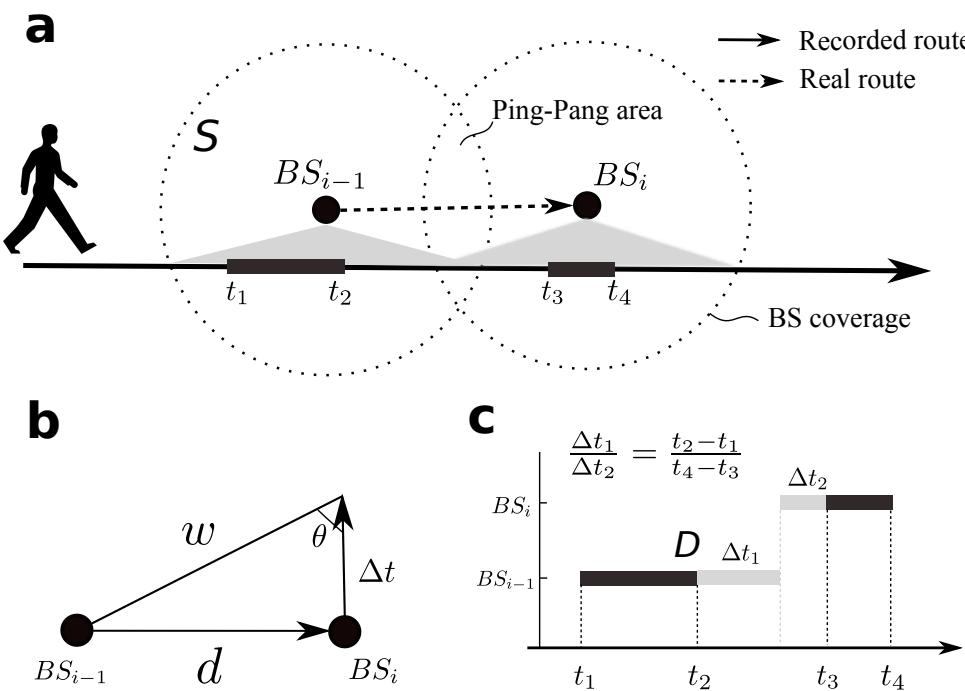


图 2-6 数据记录质量的计算和影响因素示意图

Fig 2-6 The illustration of calculation and impacting factors for data record quality.

质量，便需要对单条数据记录的质量进行量化。我们对时空行为分析中的单条记录的数据质量从动态和静态特征两个层面进行刻画<sup>1</sup>。动态特征代表观测个体在移动过程中所表现出的特征。图2-6a展示了用户在两个相邻基站之间移动的过程。用户进入基站的覆盖区域（BS Coverage）后，设备便自动连接到基站。在两个基站交叠的区域，往往会发生网络连接在基站间来回切换，我们称之为“乒乓效应”，这将在质量提升算法中进一步处理。

当我们从移动过程分析时空数据质量时，在相邻的两个数据点之间（如  $t_2$  和  $t_3$ ），空间距离越远、间隔时间越长，则这样的数据点质量越低。这是因为对于较大的时空距离，所采集的数据点越少，关于用户移动的位置信息丢失越多。从这一角度出发，Iovan 等<sup>[97]</sup>提出了一种基于移动速度的数据质量指标，这里我们采用这一方法对动态特征进行描述。如图2-6b 所示，对于相邻时刻的两条数据记录，它们的空间距离和时间间隔分别为  $d$  和  $\Delta t$ 。保留二者的量值关系而忽略物理单位，可以得到几何意义上的速度指数

<sup>1</sup>本节提到的数据质量评估和提升算法的实现代码: [https://raw.githubusercontent.com/caesar0301/movr/master/R/data\\_quality.R](https://raw.githubusercontent.com/caesar0301/movr/master/R/data_quality.R)

$\theta = \arctan \frac{d}{\Delta t}$ , 以及标准化的间隔时间为  $w = \frac{\Delta t}{\cos \theta}$ 。由于  $\theta \in [0, \frac{\pi}{2})$ , 因此添加相应的常量因子进行调节, 并将动态角度的数据质量表示为:

$$Q_a = \exp\left(-\frac{2\theta}{\pi}w\right) \in (0, 1]. \quad (2-1)$$

特别地, 由于在初始时刻数据记录没有更早的历史记录, 因此我们规定  $Q_a|_{t=0} = 1$ 。

在公式2-1中, 一个潜在的假设是在每个时刻, 数据记录的用户空间位置是完全准确的。如图2-6a所示, 由于基站有一定的覆盖范围, 用户的位置记录本身存在误差, 而这部分信息是指标  $Q_a$  中所缺失的。为了衡量这类时空上的静态信息(如基站覆盖)对数据质量的影响, 我们提出了数据质量指标  $Q_b$  进行补充, 其定义为

$$Q_b = \frac{2}{\pi} \arctan \frac{D}{S\sqrt{\rho}} \in [0, 1) \quad (2-2)$$

公式2-2来源于对以下观测经验的总结: 在某一时刻, 用户的数据记录显示其连接到了覆盖面积为  $S$  的移动基站, 因此基站覆盖面积越大, 确定该用户的具体位置便越困难, 数据质量也越低, 即  $Q_b \propto 1/S$ 。从用户之间的差异性分析, 对于来源于同一个用户、在覆盖面积相同的两个基站下的数据, 则在相同时间段内, 观测到的独立用户越多, 从中区分该用户的可能性便越低, 因此数据质量也越差, 即  $Q_b \propto 1/N = 1/S\rho$ , 其中  $N$  表示观测到的独立用户数,  $\rho$  表示独立用户密度。从时间覆盖范围来分析, 对于某一时刻的数据记录, 如果用户在观测位置的停留时间  $D$  越长, 则在观测时间点附近, 该用户处于所记录的位置的可能性越高, 数据质量也越高, 即  $Q_b \propto D$ 。综合以上空间不确定性、个体差异、以及时间不确定性, 我们利用反正切函数作归一化处理, 并得到公式2-2的表达形式。

在该式中, 基站覆盖面积  $S$  和用户密度  $\rho$  比较容易获得, 而用户在基站下的停留时间  $D$  较为困难, 这是因为在被动的网络流量中, 用户只有使用基站服务才会产生数据记录, 因此在两条数据记录之间的时段里, 用户的位置(在统计上)是随机的。在本文中, 我们基于以下两个假设来计算停留时间  $D$ : 在不同基站的两条数据记录之间, 1) 用户处于两个基站之一的覆盖范围内, 2) 两个基站的停留时间正比于已知数据记录的时间跨度值, 如图2-6c所示。最后, 将  $Q_a$  和  $Q_b$  进行调和平均, 最终得到衡量单点时空数据记录的质量指标为:

$$Q_P = F_1(Q_a, Q_b) = \frac{2Q_a Q_b}{Q_a + Q_b} \in (0, 1) \quad (2-3)$$

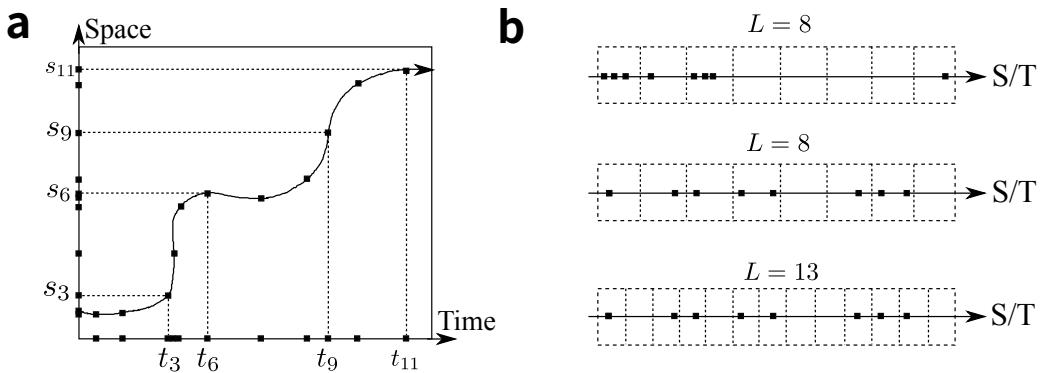


图 2-7 用户轨迹质量与数据采集点的时空非均衡性的关系

Fig 2-7 The relationship of user trajectory quality with spatio-temporal inhomogeneity.

**用户轨迹质量：**数据点的质量是从局部上对时空数据集的质量进行度量，但是在时空行为分析中，通常是对个体的移动行为进行研究，因此需要了解单个用户的历史轨迹的质量，即从全局对数据质量进行评估。Iovan 等<sup>[97]</sup>提出利用用户轨迹对应的  $Q_a$  值序列，和信息熵 (Entropy) 对轨迹质量进行测量。可以看出这样的方法有两个明显的缺陷：一是  $Q_a$  仅包含用户移动导致的数据质量下降的局部信息；二是信息熵<sup>[97]</sup>的误差较大，对于

$$H(Q_a) = -\frac{1}{n} \sum_n Q_a \log_2(Q_a) \quad (2-4)$$

其中  $Q_a$  满足归一化处理， $\sum_{i=1}^n Q_a = 1$ ；这样即使单个点的质量较高，当均匀分布时所得的熵依然很高。例如，数据质量序列  $Q_a^{(i)} = 0.9, i = 1, \dots, 10$ ，与  $Q_a^{(i)} = 0.1$ ，具有相同的熵值。针对已有方法的不足，本文提出利用  $Q_P$  代替  $Q_a$  进行单点数据质量评估，进一步结合数据点质量的期望和时空分布特征，计算用户轨迹的数据质量。如图2-7a 展示了某用户移动过程中空间和时间距离的变化曲线，以及数据点在时间和空间维度上的投影。这里我们基于以下假设量化时空分布特征对用户轨迹质量的影响：当给定数据点数目时，假如用户以均匀速度行驶，则时间维度上的投影点越均匀，轨迹的整体信息丢失越少，数据质量越高；类似地，假如用户以非均匀速度行驶，则空间上的投影点越均匀，数据质量越高。由此得出用户轨迹的数据质量为：

$$Q_I = E(Q_P) \exp\left(-\frac{2H_s H_t}{H_s + H_t}\right) \quad (2-5)$$

$$E(Q_P) = - \sum_{i=1}^m p_i \cdot Q_P^{(i)} \quad (2-6)$$

$E(Q_P)$  表示用户数据点质量的平均值, 且观测到的用户数据点质量为  $\{Q_P^{(1)}, \dots, Q_P^{(m)}\}$ , 及其频率分布  $p_1, \dots, p_m$ 。 $H_s$  和  $H_t$  分别表示用户轨迹中, 空间和时间上数据点分布的非均衡性 (Heterogeneity<sup>[69]</sup>) 系数 (简称为非均系数)。非均系数越大, 用户轨迹的数据质量越低。本文使用栅格法对非均系数进行度量。图2-7b 展示了在单一维度 (时间或空间) 上数据点的非均衡性差异。给定栅格  $\mathcal{L}$ , 当  $L = |\mathcal{L}| = 8$  时, 上图数据点倾向于聚集在轴的两端, 因此中图的均衡性更高; 对于中图和下图, 由于当  $L = 13$  时数据点分布在不同的栅格内, 因此下图的栅格方案更好。具体而言, 给定具有  $n$  个数据点的某用户轨迹, 以及栅格  $\mathcal{L}$ , 每个栅格内的数据点数目的期望为

$$\bar{l}(\mathcal{L}) = \frac{n}{L} \quad (2-7)$$

则对于栅格  $\mathcal{L}$  的非均衡系数为

$$h(\mathcal{L}) = \frac{1}{2n} \sum_{i \in [1, L]} |l_i - \bar{l}(\mathcal{L})| \quad (2-8)$$

给定栅格序列  $\{\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_{L_*}\}$ , 及其对应的栅格数序列  $\{1, 2, \dots, L_*\}$ , 其中  $\mathcal{L}_{L_*}$  表示可以将每个数据点放到不同格子内的最小分割方案, 换句话说, 利用最小的  $L$  满足每个格子内最多包含一个数据点。通过给予栅格不同的惩罚因子  $\sigma^{1-i}$  (栅格数目越多, 惩罚因子越大), 我们得出单维度上的非均衡系数为

$$H := \frac{1}{L_*} \sum_{i \in [1, L_*]} \sigma^{1-i} h(\mathcal{L}_i), \sigma < 1, H \leq 1 \quad (2-9)$$

其中惩罚因子  $\sigma$  也可作为归一化常量, 使得  $H$  的最终取值范围位于  $[0, 1]$  之间。针对图2-7中时间和空间维度的投影点, 利用公式2-9分别计算出  $H_s$  和  $H_t$  并代入公式2-5, 便得出个体移动轨迹的数据质量  $Q_I$ 。

**质量提升算法:** 在对时空数据质量进行量化评估的基础上, 我们提出了数据质量提升算法。如图2-6a 所示, 移动网络数据的特征之一是存在明显的“乒乓效应”, 即当用户处于基站的交叠区域时, 由于信号的强弱变化, 用户的设备在基站之间频繁进行切换。

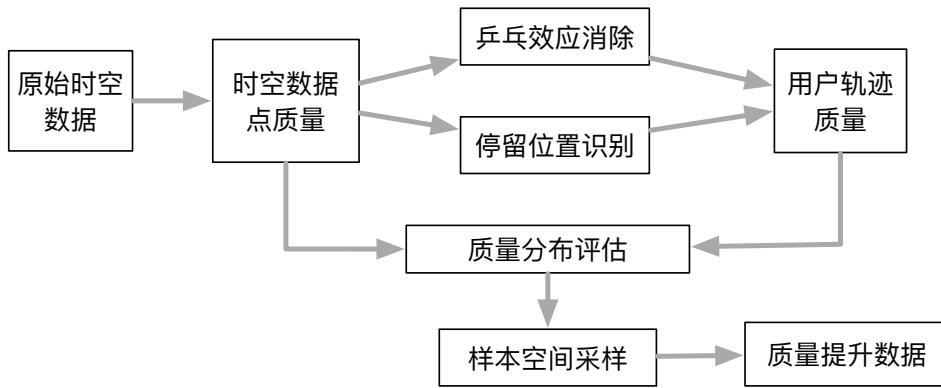


图 2-8 时空行为数据质量提升算法流程图

Fig 2-8 The illustration of data quality enhancement algorithm.

即使用户的位置没有发生变化，也会产生多条位置变化的记录，因此首先利用单点数据质量评估消除其影响。

图2-8展示了质量提升算法的完整处理流程。原始的时空数据记录了用户在特定时间点所连接的网络基站或热点。首先基于公式2-1~2-3，对单个数据点的质量进行评估。为了消除数据的“乒乓效应”，我们通过以下两个步骤来完成：1) 根据经验中用户乘坐不同交通工具的移动速度上限，获得  $Q_a$  的阈值上限  $Q_a^*$ ，从而将数据质量  $Q_a > Q_a^*$  的数据点作为候选的“乒乓效应”记录。2) 检查连续的候选数据点序列，如果序列满足如  $ABAB\dots$  的形式，则认为用户在两个基站间频繁地切换，将累积连接时间最长的基站作为用户在当前时间的停留位置；否则，将候选数据点从原始数据中删除。

对于停留位置的识别，我们将用户在  $t$  时刻所连接基站的位置，作为用户在该时刻的位置，对应的停留时间通过公式2-2中  $D$  的计算方法获得。对于基站的覆盖面积  $S$ ，我们利用维诺图（Voronoi Diagram）将所有基站划分到连续多边形中，并将每个多边形的面积作为基站的覆盖面积  $S$ 。这样，在网络基站部署密集的区域， $S$  值较小；反之，在基站部署稀疏的区域， $S$  值较大。

在获得时空数据点  $Q_P$  的基础上，进一步通过公式2-5计算出用户轨迹质量  $Q_I$ ，并对  $Q_P$  和  $Q_I$  的分布特征进行评估，以及对用户样本进行采样。这里基于累积分布函数（CDF）对用户轨迹的样本空间进行采样。通常针对不同的分析目标，所采用的采样策略有所差异。这里我们采用更具一般性的采样策略。这一策略基于不同质量的采样函数

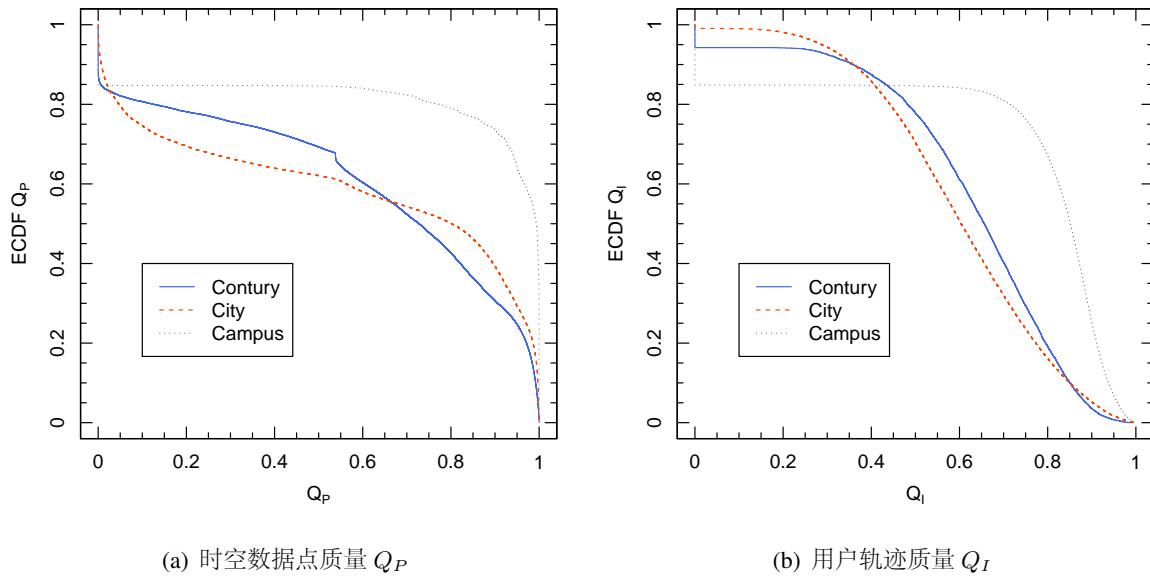


图 2-9 不同空间尺度下的时空数据质量对比

Fig 2-9 The comparison of data qualities at varying spatial scales.

$p = f(Q_I)$ , 即用户轨迹的质量不同, 采样时被选中的概率则不同。当函数  $f(Q_I)$  为均匀函数时, 对应的采样策略为随机采样<sup>[16]</sup>。当函数  $f(Q_I)$  为阶跃函数时, 即  $p = 0, \forall Q_I < Q_I^*$ ;  $p = 1, \forall Q_I \geq Q_I^*$ , 对应的采样策略为阈值采样<sup>[40]</sup>。更灵活地, 可以根据分析需求对函数  $f(Q_I)$  进行灵活设置, 从而获得不同质量分布的数据集。在后续章节的分析中, 我们采用阈值采样的方法对用户轨迹质量进行控制, 并根据经验选择  $Q_I^* = 0.05$ 。

### 2.3.3 移动网络数据的质量分析

这部分我们利用图2-8的数据质量提升方法, 分别对不同空间尺度下的时空数据集(即 WIFI-M, CITY-M, Senegal)进行量化比较:

图2-9展示了在不同空间尺度下, 时空数据点和用户轨迹的数据质量分布曲线, 其中 Contury, City 和 Campus 分别对应 Senegal, CITY-M, 以及 WIFI-M 数据集。从图中可以看出, 代表局部信息的数据点质量在不同空间尺度下具有较大的差异。和移动蜂窝网相比, 校园 WiFi 网络数据集由于热点密集、定位精度较高, 因此数据质量也最好, 其中 80% 以上的数据点的质量  $Q_P > 0.8$ 。对比城市和国家尺度下的移动网络数据, 在国家尺度下的数据集中, 当  $Q_P > 0.5$  时, 符合条件的数据点比例呈快速下滑趋势, 表

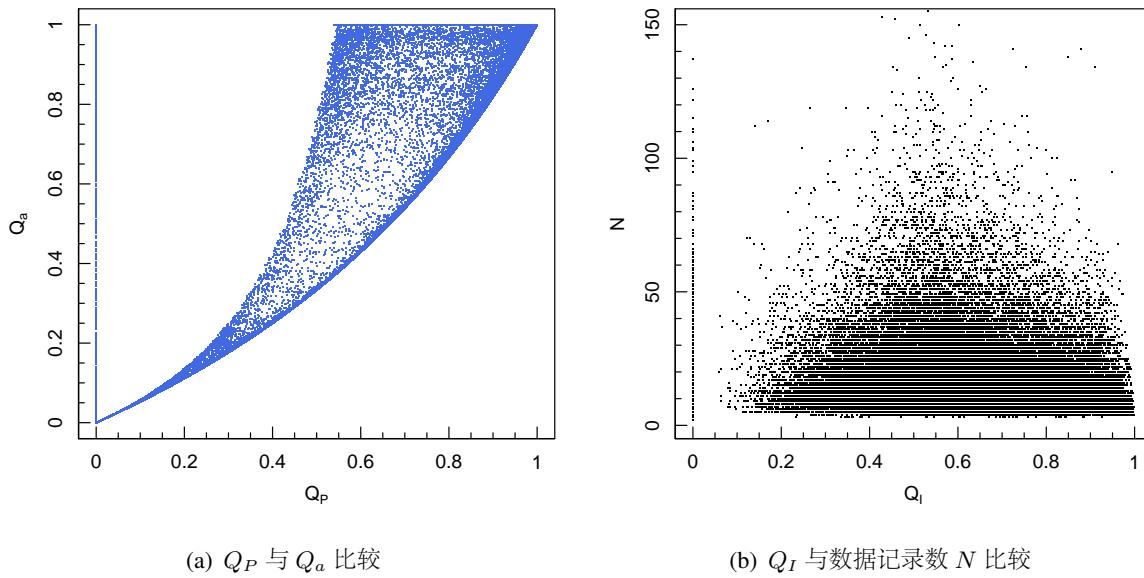


图 2-10 数据点质量和用户轨迹质量与传统量化指标的比较

Fig 2-10 The comparison of novel spatio-temporal data quality measurement and traditional methods.

明城市网络具有相对较高的数据点质量。这潜在地得益于城市内部较为均衡的基站部署，进而减小了公式2-2中基站的覆盖面积  $S$ ，以及标志着区分独立用户的能力特征  $\rho$ 。图2-9给出了代表全局信息的用户轨迹质量，可以看出，校园 WiFi 网络用户的轨迹质量普遍较高，其中 80% 以上的用户轨迹质量  $Q_I > 0.75$ ；而国家和城市尺度下的用户轨迹质量并无明显差距，其 80% 以上的用户轨迹质量  $Q_I > 0.5$ 。

为了进一步展示上述数据质量评估的有效性，我们分别与已有的量化方法进行了比较。首先，Iovan 等<sup>[97]</sup>首次提出利用移动的过程信息（如速度、间隔时间等）对数据点质量进行衡量，即公式2-3中的  $Q_a$  指标。图2-10a 给出了  $Q_a$  与本文中提出的  $Q_P$  之间的变化关系。可以看出，即使对于相同的  $Q_a$ ，对应的  $Q_P$  的取值也存在较大的差异，并且差异随着  $Q_a$  的变大而增多。这种差异的主要来源在于  $Q_b$  对时空数据点静态特征的反应，并且静态特征的变化范围随着随着数据点质量的提高而相应变大。对于用户轨迹质量，传统常用的方法为选择轨迹数据点数目  $N$  满足一定阈值的用户<sup>[67]</sup>，这种方法虽然操作起来比较方便，但是由于缺少用户的行为特征，以及数据点  $N$  的底层分布，因此所选择的阈值往往具有主观随意性。图2-10b 展示了  $N$  和我们提出的用户轨迹质量  $Q_I$  之间的关系，可以看到，虽然随着  $N$  变大， $Q_I$  的取值范围不断减小，但是依然较

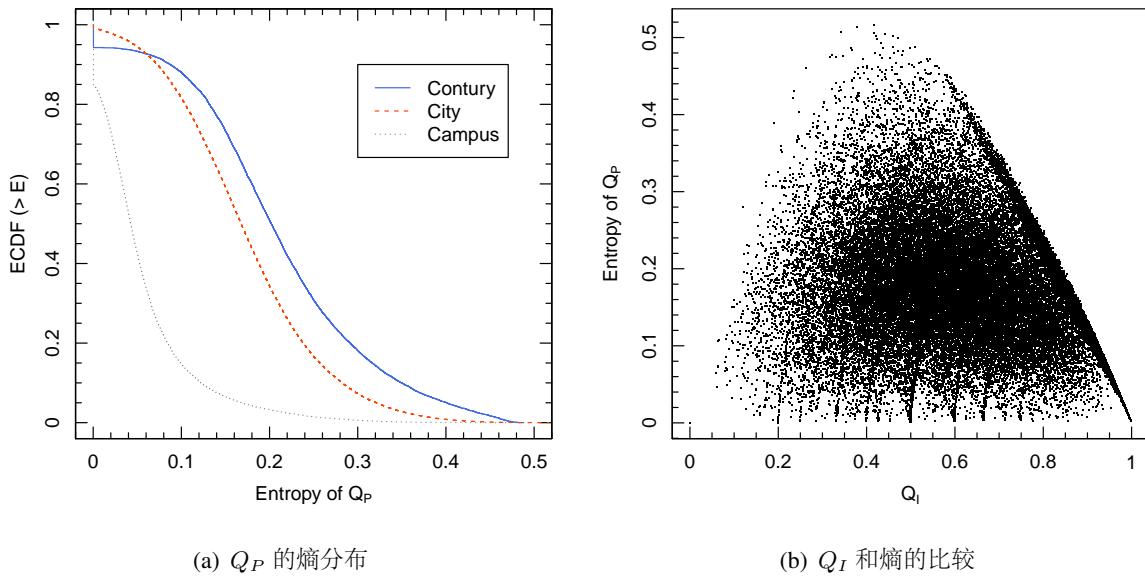


图 2-11 不同空间尺度下数据点质量  $Q_P$  的熵分布及其与  $Q_I$  的比较

Fig 2-11 The entropy distribution of data quality  $Q_P$  at varying spatial scales and its comparison with  $Q_I$ .

宽, 如  $N = 50$  时,  $Q_I$  所处的范围约为  $[0.2, 0.9]$ 。另一方面, 当  $Q_I < 0.6$  时,  $N$  与  $Q_I$  呈现较弱的正相关 (Pearson 相关系数为  $\rho_p = 0.25$ ), 而当  $Q_I \geq 0.6$  时, 二者呈现较弱的负相关 (Pearson 相关系数为  $\rho_p = -0.27$ )。由此可见, 单纯的数据点数目  $N$  对于分析时空数据质量依然具有较大的局限性。

信息熵是 Iovan 等<sup>[97]</sup> 用来衡量用户轨迹质量的方法。如公式2-4所示, 信息熵更多地反映了一条用户轨迹中, 多个数据点质量分布的均匀程度, 因此对于均匀分布的、质量较高的多个数据点组成的用户轨迹, 会出现数据质量反而较低的矛盾。图2-11分别展示了本文的数据质量评估方法和信息熵之间的比较。由图 a 可以看出, 在不同空间尺度下, 数据点质量的信息熵分布差异较大。具体而言, 和移动网络数据相比, 校园 WiFi 网络数据的信息熵最小 (或数据质量最高); 同时, 城市尺度的信息熵 (或数据质量) 比国家尺度小 (高)。图 b 展示了对于用户轨迹质量, Iovan 的方法和本文提出的  $Q_I$  之间的差异。从整体上而言, 两种方法具有相同的结论, 即信息熵越小, 数据质量越高 (Pearson 相关系数为  $\rho_p = -0.55$ )。但是图中的扇形分布表明, 随着信息熵的减小, 用户的轨迹质量  $Q_I$  有着更宽的变化范围, 这是由于和信息熵相比,  $Q_I$  包含更多的时空分布特征, 因而也能够更加准确地行为观测的数据质量。



0000294

## 2.4 本章小结

本章工作的背景是移动大数据。移动大数据作为一类覆盖范围较广的时空数据类型，其包含了丰富的时空行为数据，是一种理想的信息来源。但是由于其体积大、速度快、结构多样，因此在数据采集、存储、处理、以及数据质量评估上都有着潜在的挑战。本文介绍了一种适合于大规模采集、处理用户行为数据的被动测量方法，通过不同功能模块的抽象，实现了对移动网络流量高效、灵活的处理。同时，与之相配合的是实时和批处理相结合的数据处理平台，其内部自下往上的逻辑数据仓库结构，为数据的预处理、存取、以及质量控制带来了较大的便捷。基于这套方法和系统，本文采集并收集了三种不同空间尺度下的移动网络数据，并在数据格式上进行了统一，为后续章节的分析研究工作打好了基础。

作为时空数据分析的一个重要方面，数据质量管理是本章的另一重点内容。针对以往研究工作中数据质量评估的简单粗糙，本文从数据记录的准确性、采集时间的连续性、以及空间分布的合理性出发，对时空数据质量进行了多角度的刻画，并通过量化的方法对其进行评估。随着时空数据研究的不断发展，数据质量的重要性将不断被验证。另一方面，本文的数据质量评估方法主要考虑了时空行为数据的特征，对于一般的时空数据类型，如物联网观测数据，依然具有一定的局限性，因此也是本文的质量评估方法，在未来工作中仍能够进一步扩展，从而覆盖更丰富的时空数据类型。



0000294



0000294

## 第三章 个体移动行为的时空模式挖掘

在行为科学和相关社会科学研究中，观测到的人类行为通常是复杂的<sup>[13,16,63]</sup>。例如，对于人类的移动行为，一方面受到外在环境和工作制度的约束，另一方面又受个人内在需求的支配，而后者巨大差异导致了人类行为较高的不确定性。而行为模式的概念表示在人类行为观测中重复出现的部分<sup>[45]</sup>，其分析方法在多个领域有着潜在的应用价值，如大型紧急事件监测、传染性疾病的扩散预测、以及城市交通资源的优化。同时，随着近年来带有 GPS 的便携式设备的普及，大尺度的个体移动行为分析及研究成为了可能。

人类移动行为的研究通常有两个大的方向，一是宏观的统计规律研究，即平滑掉用户个体的行为差异，对行为的共性特征（如转移距离）进行统计分布研究；而是微观的个体序列模式研究，例如通过历史的移动位置信息，对用户在下一时刻的可能位置进行预测。虽然宏观统计比较符合人类从概念上把握事物的认知规律，但是其不足在于统计参数的物理含义不明确，从而导致对其成果进一步应用的可能性降低。同时，即使人们从宏观统计上观测到了移动规律的参数发生了变化，但是对于到底发生了何种变化依然是无从知晓的，即缺少了与底层个体移动模式的关联。以此为研究动机，本章在前人研究行为序列模式和时空性质的基础上，提出了一种介观尺度上的时空行为模式，将属性图的分析方法引入到人类时空行为分析中，并基于新的时空模式提出一种鲁棒性和准确度都较好的个体移动模型。本章工作的创新性在于，所提出的研究方法和所得结论，为连接宏观统计规律和微观个体行为模式提供了基础。

### 3.1 移动模式挖掘的方法介绍

移动轨迹模式挖掘是从一组轨迹序列或轨迹数据库中，利用特定的模式提取算法，分析得到满足一定规律的行为现象。图3-1首先展示了 CITY-M 数据集中三位用户的轨迹时空序列，每个点表示用户在特定时间点的特定空间位置，水平面按照 Voronoi 网格进行了划分，可以看出这三位用户在常住地点、活动范围、规律性上有着很大的区别。目前研究人员通常采用两种方法对多位用户的轨迹数据进行分析：一种是从微观

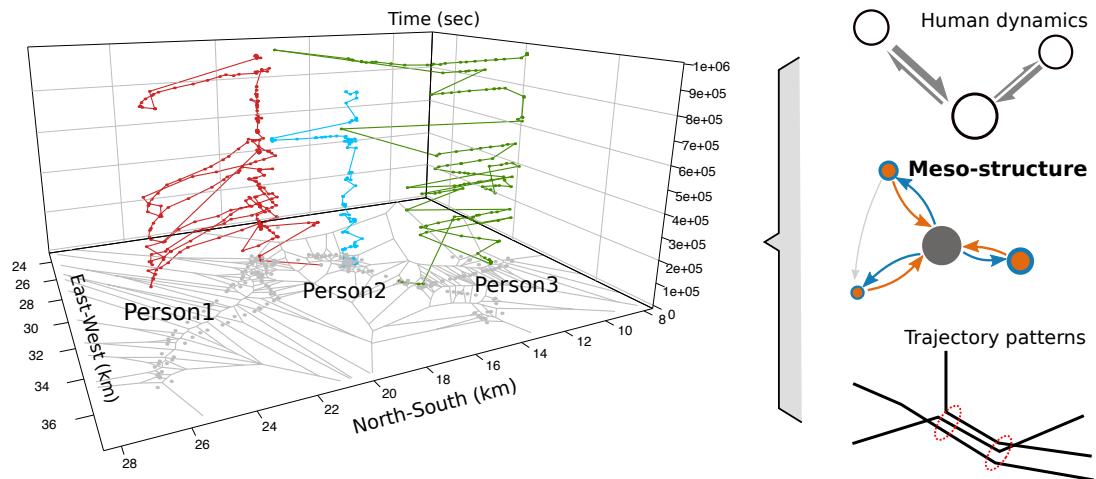


图 3-1 个人时空行为轨迹和不同分析粒度上的行为模式

Fig 3-1 The exemplification of individual trajectories and mobility patterns at different granularities.

(Microscope) 相似性的角度，采用序列模式挖掘<sup>[98]</sup> 的方法，从多个轨迹序列中提取出满足一定出现频次的子序列。另一种是从人类动力学角度<sup>[13]</sup>，采用宏观 (Macroscopic) 统计的方法，对不同地点之间的人数迁移规律进行研究。我们首先介绍这两种常用模式挖掘方法的基本原理。

**微观角度的序列模式**研究对象通常为轨迹序列  $\mathcal{I} = \{i_1, i_2, \dots, i_N\}$ ，其中  $i_x$  表示停留位置的符号条目 (Item)；多个符号条目的组合构成了  $\mathcal{I}$  的条目子集 (Itemset)。一条  $m$  长度的轨迹序列包含  $m$  个这样的条目子集，如  $\alpha = \langle \alpha_1, \alpha_2, \dots, \alpha_m \rangle$ ，其中  $\alpha_i = (i_{j_1}, \dots, i_{j_k})$  and  $i_{j_l} \in \mathcal{I}$ 。给定另一个轨迹序列  $\beta = \langle \beta_1, \beta_2, \dots, \beta_n \rangle$ ，如果存在整数索引  $1 \leq k_1 < \dots < k_m \leq n$ ，使得  $\alpha_1 \subseteq \beta_{k_1}, \alpha_2 \subseteq \beta_{k_2}, \dots, \alpha_m \subseteq \beta_{k_m}$ ，则  $\alpha$  成为  $\beta$  的子序列，或反过来， $\beta$  称为  $\alpha$  的父序列。对于一组组序列来说， $\alpha$  的父序列的数目定义为该子序列的支持度；当支持度高于某一给定阈值的时候，则称该子序列为一个模式。

通常时间是影响微观序列模式挖掘的重要因素，模式结果的数目和支持度取决于观测轨迹历史记录的时间窗口大小<sup>[43]</sup>。时间窗口越大，出现频次较少的模式会被较多模式淹没，反之，模式数据将减少。因此通常以参数的方法控制时间窗口  $\mathbf{w} = (u, t_s, t_e)$  的属性，若窗口宽度为  $wid(\mathbf{w}) = |t_e - t_s|$ ，其大小为  $vol(\mathbf{w}) = |u|$ ，则采用阈值的时间窗口调节机制为  $\lfloor 0.5w_\theta \rfloor \leq wid(\mathbf{w}) \leq w_\theta$  且  $\lfloor 0.5v_\theta \rfloor \leq vol(\mathbf{w}) \leq v_\theta$ 。

微观序列模式的另一个重要性质是模式的相似性。对于符号序列，常用的表示方法

有编辑距离、字符串核、以及最长公共子序列 (LCS) 等。以 LCS 为例，假设两个轨迹序列  $\alpha$  和  $\beta$  的公共部分为  $\theta$ ，则它们的公共因子为

$$R(\alpha, \theta) = \frac{\sum_{i=1}^{|\alpha|} \sum_{j=1}^{|\theta|} h_j \cdot \frac{|\alpha_i \cap \theta_j|}{|\alpha_i|}}{|\alpha|}, \quad (3-1)$$

其中  $h_j = e^{-\delta_j}$  衡量不同地点位置之间的不连续性，其中指数  $\delta_j$  可以通过模式条目子集的归一化频率表示，即

$$\delta_j = \begin{cases} 0 & j = 1; \\ \frac{|u-v|-1}{|\alpha|} & j > 1, \theta_{j-1}, \theta_j \text{ matching to } \alpha_u, \alpha_v \end{cases} \quad (3-2)$$

**宏观角度的行为模式**，与微观上的序列模式不同，指从大量用户的时空轨迹上获得人群移动的统计规律。以辐射理论<sup>[20]</sup> 为例，在一个分布有发射和吸收源的空间里，单个用户被看作是从某个初始出发点发射出的随机粒子。假设粒子  $X$  从地点  $i$  发射出时，带有量值为  $z_x^{(i)}$  的能量，表示该粒子被吸收的最低能量阈值，且  $z_x^{(i)}$  取值为对分布  $p(z)$  进行  $m_i$  次采样后的最大采样值，其中  $m_i$  表示地点  $i$  的人口总数。另一方面，周围的地点  $j$  拥有  $n_j$  的人口总数，因此构成了一个吸收源，能够以一定的概率对粒子  $X$  进行吸收，吸收的能量阈值为  $z_x^{(j)}$ ，且该取值为对分布  $p(z)$  进行  $n_j$  次采样后的最大采样值。同时，每个地点  $i$  发射的粒子能够被距离其最近的、且  $z_x^{(j)} \geq z_x^{(i)}$  的地点  $j$  所吸收。这些条件和过程构成了人类移动的辐射理论的基础，从而地点  $i$  发射的粒子被地点  $j$  吸收的概率为

$$\begin{aligned} P(1|m_i, n_j, s_{ij}) &= \int_0^\infty dz P_{m_i}(z) P_{s_{ij}}(< z) P_{n_j}(< z) \\ &= m_i \int_0^\infty dz \frac{dp(< z)}{dz} [p(< z)^{m_i+s_{ij}-1} - p(< z)^{m_i+n_j+s_{ij}-1}] \\ &= \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \end{aligned} \quad (3-3)$$

其中  $s_{ij}$  表示地点  $i$  和  $j$  之间（不包括二者本身）的人口分布总数， $P_{m_i}(z)$  表示对  $p(z)$  进行  $m_i$  次采样后取得的最大值等于  $z$  的概率。由此可见，从宏观上来看，两个地点之间人流的迁移量仅由两个地点、以及二者之间区域的人口分布数有关。

虽然序列模式挖掘和辐射理论，分别从微观和宏观层面揭示了人类移动的规律，但是这两种方法对理解人类移动的时空行为结构依然有限。首先，宏观统计方法通过带参

数的分布形式给出群体级别特征，虽然这些特征（如两地点之间的人口迁移量）理解较群体差异有帮助，但是由于缺乏个体的移动性信息，对实际应用场景的指导性不足。尤其在小群体的比较当中，个体行为较大的差异性甚至会被群体统计的相似性所掩盖。其次，微观上的序列模式挖掘由于强调行为的顺序化信息，即马尔可夫特性，因此相同的用户时空行为结构将可能产生不同的移动序列。基于这些考虑，我们需要一种鲁棒性更好的行为时空结构表示，而研究这样的时空结构不仅对现有的行为分析提供了补充，而且对理解个体微观的多样性和群体宏观统计的差异之间的关联。

### 3.2 个体移动行为的介观模式挖掘

在本节当中，我们从介于微观和宏观之间的角度出发，提出一种新的个体移动行为模式，即介观时空模式（Mesostructure）；该结构基于周期性的行为观测数据，将一个周期内的用户移动序列，抽象为融合时间和空间特征的有向属性图，而介观时空模式为多个有向属性图结构的共享部分。这样的时空行为结构表示通常有以下三个方面的好处：1) 空间依赖性：由于采用有向属性图的形式，因此介观时空模式不仅包含了个体移动序列的相似性，也包含了轨迹拓扑中的地点之间的相关性信息。2) 多重时间约束：介观时空模式拥有的时间上限为人类行为的观测周期，如“日出而作，日落而息”的通勤规律；其时间下限为人们由于位置偏好而表现出来的单地点停留时间。这些时间约束表现了人们在行动力上的局限性。3) 算法鲁棒性：通常由于数据采集源的不精确或人类移动行为的内在复杂性，模式挖掘算法通常既要发现有价值的行为模式，又要抵抗来自多方面的噪声干扰。介观时空模式通过将时间和空间的行为属性以统一的形式表达，因此结合最大相似结构挖掘算法，对数据的缺失或行为内在的突发性具有较好的鲁棒性。我们这里首先介绍介观时空模式相关的概念和定义。

在时空数据库中，用户的移动性通过一定时间段内产生的离散时空点序列体现。假设一段轨迹序列为  $H_p = \{l_t | t \in \mathbb{N}\}$ ，其中  $l_t$  表示用户在时间  $t$  时刻的位置坐标。在本研究当中，我们考虑一种时间限定的轨迹序列，即  $H_p = \{l_0, \dots, l_T\}$ ，且  $l_0 = l_T$ ,  $l_i \neq l_j$ ,  $\forall i, j \in (0, T)$ 。在这样的前提下，定义3.1给出了个体的移动图定义，图3-2给出了CITY-M数据集中真实用户的移动图示例。在实际分析中，由于移动网络依靠基站或AP热点的位置对用户进行定位，用户的位置转移表现为在基站间的跳跃。为了更加精确的评估用

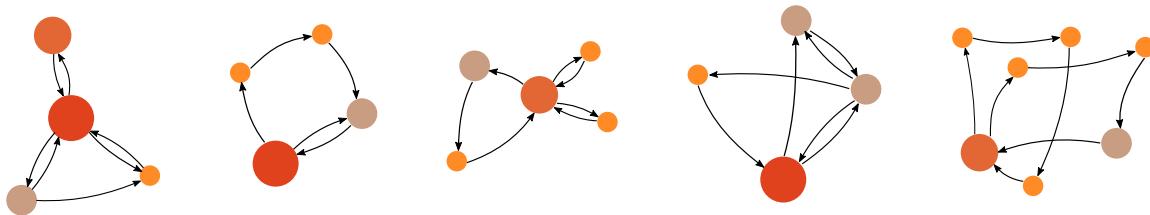


图 3-2 CITY-M 数据集中的个体移动图示例

Fig 3-2 The examples of mobility graphs from CITY-M dataset.

户位置转移的代价，我们结合路网的物理限制，并利用最短路径距离算法计算出位置之间的转移距离，以此修正基站定位对用户在城市中出行代价的评估误差。

**定义 3.1** (移动图 Mobility Graph). 给定用的轨迹序列  $H_p = \{l_0, \dots, l_T\}$ ，移动图是由此轨迹生成的有向属性图，且至少拥有一个环；数学形式表示为  $G_p = (V_p, E_p, \phi_p, \psi_p)$ ，其中顶点集合  $V_p$  表示  $H_p$  出现的所有地点的，边集合  $E_p$  表示轨迹中相邻地点之间的位置转移，顶点权重  $\phi_p = \{d_i\}$  表示用户在某个地点的累计停留时间，边权重  $\psi_p = \{c_j\}$  表示用户在相邻地点之间转移所付出的代价。

**定义 3.2** (介观时空模式 Mesostructure). 给定两个移动图  $G_A$  和  $G_B$ ，他们的介观时空模式（简称为介观模式）定义为

$$M_{AB} := \arg \min_{M_i} \sum_{k \in \{A, B\}} \Gamma(M_i, G_k | f'), \quad (3-4)$$

其中函数  $f': G_k \rightarrow M$  给出移动图  $G_k$  到介观模式候选对象  $M_i$  之间的映射，函数  $\Gamma$  衡量这种映射带来的误差。

介观时空模式反映了两个或多个移动图在考虑时间、空间结构条件下的移动相似性。由定义3.2可以看出，介观时空模式将用户移动的拓扑结构和时空属性融合在一起，从而为人类行为规律的挖掘和可解释性提供了良好的基础。从用户大量的移动轨迹中挖掘介观时空模式时，需要对模式显著性和模式有效性进行有效的评估。模式显著性的评估存在于介观模式挖掘算法当中，但是由于用户的移动拓扑和时空属性有着很大的偏差，加之数据之外的因素对用户行为存在潜在影响，因此在提取显著模式的时候应该尽可能降低时空信息的损失。模式有效性有助于更好地理解介观模式对用户行为的表达，

行之有效的方法便是根据观测到的模式对用户行为进行建模，并结合经验感测和已有的结论、对模型性能进行对比研究。本节的后续内容中，我们首先介绍一种拓扑-属性耦合的图相似算法，然后对显著介观模式的提取方法进行介绍，最后将介观模式和模序分析进行对比研究。

### 3.2.1 拓扑-属性耦合的图相似算法

图相似计算在图数据结构和网络科学中有着重要的应用价值，一般情况下算法可以分为两类：一类是特征向量法，通过提取图结构的多个特征组成特征向量，然后将向量之间的距离作为图相似计算的基础；另一类是元素迭代法，通过两个图中元素之间的相似性计算，进而得出元素组成的整个图的相似性，如 PageRank 算法<sup>[99]</sup> 和 SimRank 算法<sup>[100]</sup>。第一种方法基于统计特征，虽然计算指标对衡量整个图的相似性较方便，但是缺少图内部结构的细节信息。因此在移动图的分析中，我们基于元素迭代的思想提出拓扑-属性耦合的图相似 (TACSim) 算法。

TACSim 算法的主要功能是对两个移动图的相似性进行计算。首先，对于移动图  $G = \{V, E, \phi, \psi\}$ ，顶点-顶点邻接矩阵 (Adjacency Matrix) 表示为  $\mathbf{A}$ ，其中元素  $a_{ij} = 1$  表示顶点  $v_i$  和  $v_j$  相邻，且边由  $i$  指向  $j$ 。顶点-边邻接矩阵  $\mathbf{A}_s$  中的元素  $a_{ij}^{(s)} \in \{0, 1\}$  表示顶点  $v_i$  是否是边  $e_j$  的源顶点，用函数表示为  $s(e_j) = v_i$ ；反之，顶点-边邻接矩阵  $\mathbf{A}_t$  的元素  $a_{ij}^{(t)}$  表示顶点  $v_i$  是否是边  $e_j$  的目的顶点，用函数表示为  $t(e_j) = v_i$ 。这几种邻接矩阵之间的关系为

$$\mathbf{A} = \mathbf{A}_s \mathbf{A}_t^T, \quad \mathbf{A}^{(e)} = \mathbf{A}_s^T \mathbf{A}_t \quad (3-5)$$

其中  $\mathbf{A}^{(e)}$  表示移动图  $G$  的边-边邻接矩阵，其元素  $a_{ij}^{(e)} = 1$  表示  $e_i$  和  $e_j$  通过顶点  $a_{ij}$  进行连接，用函数表示为  $s(e_i) = t(e_j)$ 。

由于 TACSim 算法的核心思想是通过检查目标元素的邻居元素，邻居元素的相似性越高，则目标元素相似性越高。而邻居元素的相似性既和邻居元素的影响范围有关，又决定于邻居元素之间的相互作用关系。我们首先对元素的影响范围进行定义：

**定义 3.3. L 阶邻居顶点：**给定有向图  $G = \{V, E\}$ ，顶点  $v_i \in V$  的  $L$  阶入邻居元素定义为  $L$  组顶点-边对，即  $\{(v_k, e_k)\}_{i-L \leq k < i}$ ，其中  $t(e_k) = v_{k+1}$ ，且  $s(e_k) \neq t(e_{k+1})$ 。同

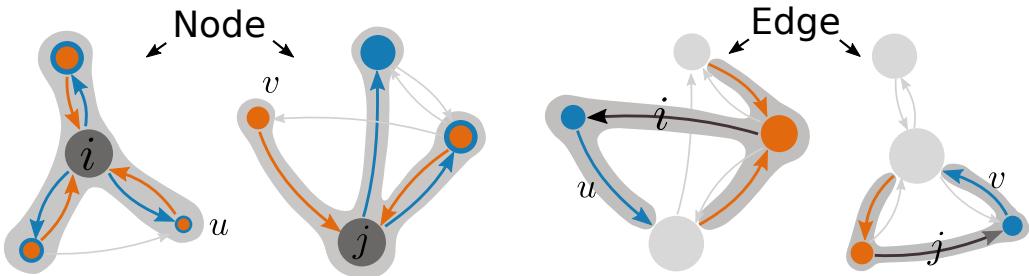


图 3-3 移动图节点和边的 1 阶邻居示意图

Fig 3-3 Illustration of 1-order neighbors for node and edge in mobility graphs.

理, 顶点  $v_i \in V$  的  $L$  阶出邻居元素定义为  $L$  组顶点-边对, 即  $\{(v_k, e_k)\}_{i < k \leq i+L}$ , 其中  $s(e_k) = v_{k-1}$ , 且  $t(e_k) \neq s(e_{k-1})$ 。元素下标  $i < j$  表示顶点  $v_i$  指向  $v_j$  或其某个源顶点。

**定义 3.4. L 阶邻居边:** 给定有向图  $G = \{V, E\}$ , 边  $e_i \in E$  的  $L$  阶入邻居元素定义为  $L$  组边-顶点对, 即  $\{(e_k, v_k)\}_{i-L \leq k < i}$ , 其中  $s(e_{k+1}) = t(e_k)$ , 且  $s(e_k) \neq v_{k+1}$ 。同理, 边  $e_i$  的  $L$  阶出邻居元素定义为  $L$  组边-顶点对, 即  $\{(e_k, v_k)\}_{i < k \leq i+L}$ , 其中  $s(e_k) = t(e_{k-1})$ , 且  $s(e_{k-1}) \leq v_k$ 。元素下标  $i < j$  表示边  $e_i$  指向  $e_j$  的源顶点。

元素的邻居关系有助于将移动图的拓扑结构和属性进行耦合, 从而元素的相似性能够根据网络路径进行传播, 最终到达一个稳定的状态。图3-3展示了  $L = 1$  条件下的顶点邻居和边邻居关系。其中红色表示入邻居元素, 蓝色表示出邻居元素。在这种邻居模型下,  $L$  越大表示某个元素的影响范围越大, 对元素之间的本地相似性刻画也就越细致。基于图3-3示例, 我们接下来对邻居顶点  $v_u, v_v$  影响  $v_i, v_j$  的关系进行量化处理。

为了清晰表达 TACSim 算法的内在结构, 以及简化算法复杂度, 我们在本研究中主要考虑 1 阶邻居的情况。假设顶点  $v$  的停留时间权重为  $d$ , 边  $e$  的转移代价权重为  $c$ , 我们引入指标  $g_{ui}$  对邻居顶点之间的连接强度 (Strength) 进行量化, 即

$$g_{ui} = \frac{d_u d_i}{c_{ui}^2} \quad (3-6)$$

该量化指标同时考虑地点之间的时间和空间属性, 表示用户在两个地点之间移动的吸引程度。连接强度正比于用户在地点的停留时间, 而反比于地点之间的转移代价, 其数学形式和万有引力定律相近, 这是因为人类行为的趋向性在观测上和物体之间的引力较为类似。在宏观的移动性研究中, 重力模型<sup>[20,74]</sup> 在相对较小的尺度上对人群流动规律

能够较好地表达，如两城市之间的流动人数正比于城市的人口总数，而反比于城市之间的距离。在3-6中，为了获得统一的连接强度单位，我们在计算中首先要对  $d$  和  $c$  进行归一化处理。接下来对两个移动图中邻居元素对之间的强度一致性（Strength Coherence）进行度量，即

$$h_{uv} = \frac{2\sqrt{g_{ui}g_{vj}}}{g_{ui} + g_{vj}} \in (0, 1]. \quad (3-7)$$

其中  $g_{ui}$  和  $g_{vj}$  表示图3-3中顶点  $v_i$  和顶点  $v_j$  分别与其入邻居  $v_u$  和  $v_v$  之间连接强度， $h_{uv}$  衡量  $g_{ui}$  和  $g_{vj}$  的算术平均值和几何平均值之间的偏离程度。对于顶点的出邻居而言，将公式3-6和3-7中  $g_*$  的下标交换位置即可。

同理，对于移动图中的边，我们首先对邻居边之间的连接强度进行量化，即

$$g'_{iu} = \frac{d_{iu}^2}{c_i c_u}, \quad (3-8)$$

其中， $d_{iu}$  表示图3-3中边  $e_i$  与其出邻居  $e_j$  的连通点处用户的停留时间权重， $c_i$  和  $c_u$  分别表示边  $e_i$  和  $e_u$  的转移代价权重。从物理意义上讲，分子部分表示如果用户在某地点的停留时间越长，则该地点的出入路径便有更强的连接关系；而分母部分基于实际观测中用户行为的“莱维飞行”特性<sup>[7,14,31]</sup>，即用户倾向于（或以较大的概率）进行短途移动；因此当两条临近的边各自距离（即转移代价）越远时，两者之间的连接关系也越弱。和上述顶点邻居对相似，边邻居对的强度一致性定义为

$$h'_{uv} = \frac{2\sqrt{g'_{iu}g'_{jv}}}{g'_{iu} + g'_{jv}} \in (0, 1]. \quad (3-9)$$

在移动图相似的计算中，我们采用迭代的方法使得顶点和边的相似性沿着图拓扑进行传播，这里我们介绍移动图拓扑和属性之间的耦合以及迭代步骤。给定两个移动图  $G_A = \{V_A, E_A, \phi_A, \psi_A\}$  和  $G_B = \{V_B, E_B, \phi_B, \psi_B\}$ ，它们的顶点相似矩阵和边相似矩阵分别表示为  $X$  和  $Y$ ，且  $x_{ij} \in X$ ,  $y_{ij} \in Y$ 。假设第  $k$  次迭代后， $G_A$  的顶点  $v_u$  和  $G_B$  的顶点  $v_v$  的相似度为  $x_{uv}^{(k)}$ ，顶点  $v_u$  合  $v_v$  相关联的边相似度为  $y_{uv}^{(k)}$ ，则在第  $k+1$  次迭代后顶点的相似矩阵性为：

$$2x_{ij}^{(k+1)} = \sum_{(u,v) \in P} h_{uv}^{(k)} (x_{uv}^{(k)} + y_{uv}^{(k)}) \quad (3-10)$$

其中  $P = P_s \cup P_t$  表示顶点邻居对的组合，且  $P_s = \{(u, v) | e_{ui} \in E_A, e_{vj} \in E_B\}$ ,  $P_t = \{(u, v) | e_{iu} \in E_A, e_{jv} \in E_B\}$ 。

同理从边角度来看，第  $k+1$  次迭代后边的相似性矩阵为：

$$2y_{ij}^{(k+1)} = \sum_{(u,v) \in Q} h'_{uv}^{(k)} (y_{uv}^{(k)} + x_{uv}^{(k)}) \quad (3-11)$$

其中  $Q = Q_s \cup Q_t$  表示边邻居对的组合，且  $Q_s = \{(u, v) | s_A(e_i) = t_A(e_u), s_B(e_j) = t_B(e_v)\}$ ,  $Q_t = \{(u, v) | t_A(e_i) = s_A(e_u), t_B(e_j) = s_B(e_v)\}$ 。

为了清晰表达 TACSim 算法中相似矩阵和移动图结构之间的关系，我们用矩阵的形式对上述迭代关系进行表达。由于元素属性需要以更高维度的矩阵进行表示，下式中我们假设  $h = h' = 1$ ，即

$$\mathbf{X}_k = \mathbf{A}\mathbf{X}_{k-1}\mathbf{B}^T + \mathbf{A}^T\mathbf{X}_{k-1}\mathbf{B} + \mathbf{A}_s\mathbf{Y}_{k-1}\mathbf{B}_s^T + \mathbf{A}_t\mathbf{Y}_{k-1}\mathbf{B}_t^T \quad (3-12)$$

$$\mathbf{Y}_k = \mathbf{A}_s^T\mathbf{X}_{k-1}\mathbf{B}_s + \mathbf{A}_t^T\mathbf{X}_{k-1}\mathbf{B}_t + \mathbf{A}_s^T\mathbf{A}_t\mathbf{Y}_{k-1}\mathbf{B}_t^T\mathbf{B}_s + \mathbf{A}_t^T\mathbf{A}_s\mathbf{Y}_{k-1}\mathbf{B}_s^T\mathbf{B}_t \quad (3-13)$$

为了准确衡量不同迭代阶段相似性的传播过程，在实际当中我们需要在进入下一次迭代以前对  $\mathbf{X}_k$  和  $\mathbf{Y}_k$  进行正规化处理。基于这样的形式，传统的仅结合节点性质的图相似算法 SimRank 便成为了 TACSim 的一个特例，即  $\mathbf{X}_k = \mathbf{A}\mathbf{X}_{k-1}\mathbf{A}^T$ 。

接下来我们对 TACSim 算法的性能从时间和空间角度进行分析。首先，采用传统二维矩阵的存储方法，在计算过程当中节点和边的相似性矩阵需要占用  $|V_A||V_B|$  的存储空间。假设每个顶点的 1 阶邻居平均数为  $\bar{n}$ ，且每个邻居以相同的概率成为“入”或“出”邻居，因此包含元素属性信息的  $h_{uv}$  和节点相似性矩阵的最小存储空间为

$$S_v = (1 + \frac{\bar{n}^2}{2})|V_A||V_B| \quad (3-14)$$

假设算法在  $K$  次迭代后收敛，其运行时间为

$$\mathcal{O}(T_v) \sim \mathcal{O}((K\bar{n} + \bar{n}^2)|V_A||V_B|) \quad (3-15)$$

相似的，对于边相似性计算的最小存储空间和运行时间分别为：

$$S_e = (1 + \frac{\bar{m}^2}{2})|E_A||E_B| \quad (3-16)$$

$$\mathcal{O}(T_e) \sim \mathcal{O}((K\bar{m} + \bar{m}^2)|E_A||E_B|) \quad (3-17)$$

**算法 3-1 拓扑-属性耦合的图相似算法 TACSim**

**Input:** Two mobility graphs  $G_1$  and  $G_2$ , maximum iteration number  $C_{iter}$ , and convergence threshold  $tol$ .

**Output:** The similarity matrix  $\mathbf{X}$  and  $\mathbf{Y}$  for nodes and edges.

```

for  $k \leftarrow 1 : C_{iter}$  do
     $\mathbf{X}_{prev} \leftarrow \mathbf{0}; \mathbf{X} \leftarrow \mathbf{1}$                                 # Initialize node similarity to pass convergence test
     $\mathbf{Y}_{prev} \leftarrow \mathbf{0}; \mathbf{Y} \leftarrow \mathbf{1}$                                 # Initialize edge similarity to pass convergence test
    if  $|\mathbf{X} - \mathbf{X}_{prev}| \leq tol, |\mathbf{Y} - \mathbf{Y}_{prev}| \leq tol$  then
        break                                         # The calculation converged and quit the loop
    end if
     $\mathbf{X}_{prev} \leftarrow \mathbf{X}, \mathbf{Y}_{prev} \leftarrow \mathbf{Y}$ 
    for  $i, j \leftarrow 1 : |V_1|, 1 : |V_2|$  do
        for  $u, v \leftarrow P_s \cap P_t$  do
             $x_{ij}^{(k+1)} \leftarrow \frac{1}{2} h_{uv}^{(k)} (x_{uv}^{(k)} + y_{uv}^{(k)})$           # Update node similarity
        end for
    end for
    for  $i, j \leftarrow 1 : |E_1|, 1 : |E_2|$  do
        for  $u, v \leftarrow Q_s \cap Q_t$  do
             $y_{ij}^{(k+1)} \leftarrow \frac{1}{2} h'_{uv}^{(k)} (y_{uv}^{(k)} + x_{uv}^{(k)})$           # Update edge similarity
        end for
    end for
     $\mathbf{X} \leftarrow normalized(\mathbf{X}), \mathbf{Y} \leftarrow normalized(\mathbf{Y})$ 
end for

```

其中  $\bar{m}$  表示每个边的 1 阶邻居平均数目。在复杂网络（如无标度网络）分析场景中，由于通常顶点和边数远大于平均邻居元素的数目，即  $K, \bar{n} \ll |V|$ ，且  $K, \bar{m} \ll |E|$ ，因此 TACSim 算法的时间和空间复杂度主要由两个移动图中的元素数目的乘积所决定。

在本研究中，我们对 TACSim 算法进行了实现<sup>1</sup>，其伪代码如算法3-1所示。在算法实现中，连接强度和强度一致性的计算位于多层循环内部，复用率很高，因此为了提高计算效率，利用空间换时间的方法对强度一致性进行预处理并存储下来。在空间存储方面，移动图的邻接矩阵在复杂图分析中占用空间较大，有效的解决方法是采用稀疏矩阵

<sup>1</sup>TACSim 源代码：<https://github.com/caesar0301/graphsim>

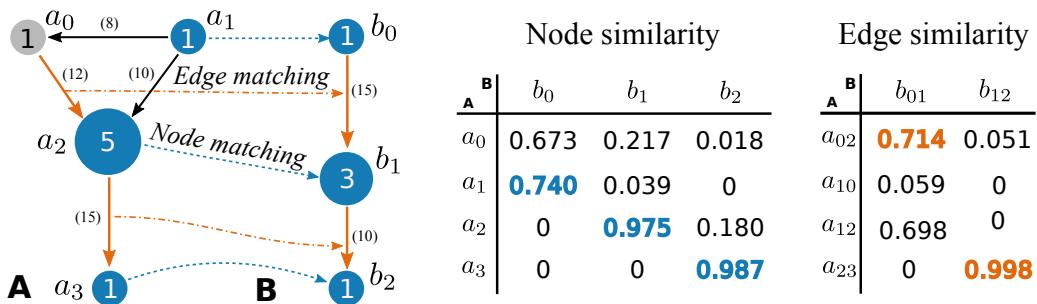


图 3-4 对顶点和边相似矩阵分别应用匈牙利算法的介观模式结构

Fig 3-4 A toy example of inconsistent matching paths for nodes (blue) and edges (orange), respectively.

存储，并对数据进行压缩，从而大幅降低算法基本数据的存储开销。

### 3.2.2 显著介观模式提取

在获得移动图之间的相似矩阵 ( $X$  和  $Y$ ) 的基础上，我们接下来讨论如何提取出两个移动图的共同部分，即介观模式。根据定义3.2，介观模式的提取是一个最优化的过程，即通过遍历两个移动图的元素组合，并找到使得其中与原移动图的累积差异最小的一种组合。基于相似矩阵，这一过程等价于，通过对相似矩阵内元素对应关系的排列组合，找到一种对应关系使得全局相似性最大，这样双移动图的介观模式提取便转换成为一个分配问题 (Assignment Problem)<sup>[101,102]</sup>。而对于一个一般的分配问题，目前性能最好的算法为匈牙利 (Hungarian) 算法<sup>[102]</sup>，该算法性能的复杂度为  $\mathcal{O}(n^3)$ 。因此，本研究中采用此算法对 TACSim 算法得到的相似矩阵作进一步处理。

基于相似矩阵的最优化方法，其优势在于既保留了用户移动行为中的时空细节信息，又同时对行为中的差异性表现出较强的鲁棒性。例如，某个用户拥有两个停留时间较长的点，即使某天用户轨迹中新增了一个路过点，该路过点对时空结构的影响也是有限的。但是，从信息量角度来讲，顶点和边的相似矩阵分别从不同侧面反映了两个移动图的结构和属性特点，带来的挑战是在某些情况下，我们通过通过匈牙利算法甚至得到不一致的结论。如图3-4所示，从顶点相似的角度出发，我们得到的移动图的最大相似结构为  $a_1 \rightarrow a_2 \rightarrow a_3$ ，而从边相似的角度出发，我们得到  $a_0 \rightarrow a_2 \rightarrow a_3$ 。出现这种差异的本质原因在于，将有向图的顶点和边进行互换，图的结构往往发生较大的变化，因此我们需要对模式的一致性进行修正。

在本文中，我们引入一个新的相似矩阵概念，即转移相似矩阵（Transition Similarity Matrix），该矩阵的计算方法为

$$\mathbf{Z} = \mathbf{Y} + \gamma \mathbf{A}_s^T \mathbf{X} \mathbf{B}_s + (1 - \gamma) \mathbf{A}_t^T \mathbf{X} \mathbf{B}_t, \gamma \in [0, 1] \quad (3-18)$$

其中  $z_{ij}$  表示用户相邻两次位置变化之间的相似度，参数  $\gamma$  作为线性调节因子对出发和目的顶点之间的比重进行调节。转移相似矩阵融合了停留地点和转移轨迹的相似性信息，通过  $\gamma$  参数可调节不同分析中对停留点的选择，比如  $\gamma = 1$  侧重于提取出出发点相似的模式。将转移相似矩阵输入到匈牙利算法中，我们得到最优的分配矩阵  $\mathbf{Z}'$ ，其元素

$$z'_{ij} = \begin{cases} 1 & \text{when } e_i \text{ matching to } e_j \\ 0 & \text{otherwise} \end{cases} \quad (3-19)$$

通过分配矩阵  $\mathbf{Z}'$  给定匹配得到的转移集合  $E_M$ ，移动图  $G_A$  和  $G_B$  中对应  $e_i \in E_M$  的源或目的顶点构成了匹配顶点的集合。换句话说，介观模式通过相似转移路径的匹配获得用户经过的空间位置点，而未被匹配的路径则不被保留。在这个意义上，我们得到了介观模式的一个普遍性质，称作“无保边”特性（Non-edge-preservation）。无保边的特性是介观模式鲁棒性的一种体现。最后，结合用户移动的时空特征完成介观模式的提取。其中，停留时间  $d_M$  和转移代价  $c_M$  表示为对应匹配顶点和边的特征的算术平均值，也就是最终获得的介观模式为  $M_{AB} = \{V_M, E_M, d_M, c_M\}$ 。例如在图3-4A 中，当选择调节参数  $\gamma = 0.5$  时，所得的介观模式顶点和边集合分别为  $V_M = \{c_1, c_2, c_3\}$  和  $E_M = \{c_{12}, c_{23}\}$ ，以及相应的时空属性为  $d_M = \{1, 4, 1\}$  和  $c_M = \{12.5, 12.5\}$ 。

在实际应用中，我们经常面临大量的属性有向图。如在移动网络中有大量的网络用户，且每个用户在连续观测中会产生多种不同的移动图。因此模式挖掘算法需要能够从多个移动图中，有效地提取出具有代表性的介观时空模式，或称为显著介观模式。我们首先分析一种朴素的模式提取算法：给定一组移动图  $\mathcal{G} = \{G_1, \dots, G_N\}$  和相似性阈值  $s^*$ ，1) 对  $\mathcal{G}$  中的移动图两两分析得到介观模式，然后通过下面定义的距离指标得出移动图对的相似性  $s$ ；2) 将相似性小于  $s^*$  的介观模式过滤掉，得到  $\mathcal{M} = \{M_{ij} | s_{ij} \geq s^*\}$ ；3) 更新  $\mathcal{G} \leftarrow \mathcal{M}$  并重复步骤一和二，直到  $\mathcal{M}$  中的模式数目连续两次保持不变为止。这样的朴素算法较为简单灵活，但是面临着几个方面的挑战：a) 即使用户的移动图相对来说结构比较简单，但是大量的图相似计算依然是一件计算密集型的任务，对计算资源

的要求较高；b) 该算法的收敛性随着  $\mathcal{G}$  中移动图数目和结构的不同而差异较大，且其收敛性有待进一步研究；c) 该算法未考虑移动图之间的差异，导致在某些情况下信息损失较为严重，例如，对于规模相差较大的两个移动图，即使提取的介观模式相似值较高，但是由于“无保边”特性的存在，所得模式对于规模较大的图信息损失较为严重。

为了克服朴素算法中的这些不足，我们提出一种基于修剪的显著模式提取算法 (Pruning-based Principle Mesostructures, PPM)。PPM 算法主要包括三个步骤：一是将原始的移动图按照结构特征分成不同的同质 (Homogeneous) 组中；二是利用聚类算法对每个组中的移动图进行聚类；三是完成对不同组内的介观模式的提取。其中修剪过程体现在步骤一和步骤三中，首先，利用图结构进行分组的过程，能够避免上述挑战中相差较大的移动图之间信息损失严重的问题，而且降低了两两移动图之间的计算次数；其次，步骤三中对模式显著性进行排序，在保留模式信息的前提下减少了总体计算量。接下来，我们对算法步骤分别进行介绍：

我们首先依据结构特征对移动图进行分组。一般来讲，在用户的移动轨迹中，代表用行为复杂度的关键特征之一便是停留地点的数目。在这里，移动图的停留点数目称作图基数 (Cardinality)。图基数越大，用户的移动行为越趋于复杂。因此我们首先采用文献<sup>[48]</sup>提出的停留点检测算法对图基数进行计算，并将基数相同的移动图分到同一组当中。这里停留点检测算法有助于降低用户移动过程中短暂的经过点对时空结构的影响。

对于同一组当中的移动图，我们在第二步中对其进行无监督的学习，即将移动图按照相似度大小进行聚类操作，因此首先需要对移动图之间的相似性进行度量。利用提出的转移相似矩阵 (3-18式)，我们定义了移动图的结构距离 (Structural Distance)，即给定移动图  $G_i$  和  $G_j$ ，它们的结构距离为

$$\delta_{ij}^2 = 1 - \|\mathbf{Z}_{ij} \circ \mathbf{Z}'_{ij}\|_2, \quad (3-20)$$

其中符号  $\circ$  表示矩阵当中各元素之间的乘积。特别地， $\delta_{ii}$  表示移动图  $G_i$  的自距离 (Self-Distance)。在聚类过程中，由于传统 K-Means 算法需要预先设定聚类簇的数目、并通过欧式 (Euclidean) 距离计算实例点之间的相似程度，因此不适合这里的移动图聚类问题。本研究中采用更为灵活的层级聚类 (Hierarchical Clustering) 算法<sup>[103]</sup>；该算法中以分类点之间的距离矩阵为基础，通过层级树高度表示实例点之间的距离，从而可以根据

**算法 3-2 显著模式提取 PPM 算法**

**Input:** A list of mobility graphs  $\mathcal{G} = \{G_1, \dots, G_N\}$  in the same group by performing stay-point detection.

**Output:** A list of mesostructures  $\mathcal{M} = \{M_1, \dots, M_K\}$  for the specific group.

```

 $simMat[N][N] \leftarrow \mathbf{0}$ 
for  $i, j \leftarrow \mathcal{G} \otimes \mathcal{G}$  do
     $simMat[i][j] \leftarrow$  Derive the structural distance  $\delta_{ij}$  after performing TACSim( $G_i, G_j$ )
end for
 $\{C_i\} \leftarrow$  Obtain clusters by HierarchicalClustering( $simMat$ )
for  $C_i \leftarrow \{C_i\}$  do
     $simMat_c \leftarrow simMat[C_i][C_i]$  # Select columns and rows in sim. matrix for each cluster
     $\Delta \leftarrow AverageByRow(simMat_c)$  # Record the order index according to mean distance
    for  $p \leftarrow \Delta.length - 1$  do
         $M \leftarrow M_{p,p+1}$  # Extract mesos. for mobility graph  $G_p$  and  $G_{p+1}$ 
    end for
end for

```

实际需求调节树枝截断的高度来调节聚类簇数目。

最后，我们对不同簇进行显著介观模式的提取。值得注意的是，在上一步聚类操作当中，我们已经完成了对组内移动图的介观模式提取，因此不需要对图相似性进行重复计算。在 PPM 算法中，我们根据两两介观模式的信息对移动图进行排序。假设任意聚类簇中包含  $L$  个移动图，这些移动图按照簇内平均距离进行排序，即

$$\hat{\delta}_i = \frac{1}{L} \sum_{j=1:L} \delta_{ij}, \quad \delta_0 \leq \hat{\delta}_1 \leq \dots \leq \hat{\delta}_L \quad (3-21)$$

其中  $\hat{\delta}_i$  给出了簇内元素到序列中第  $i_{th}$  个移动图的平均距离，而  $\delta_0$  表示簇内元素到簇的真实中心的平均距离。由聚类算法的原理可知， $\hat{\delta}_i \geq \delta_0$ 。换句话说，在一个聚类簇内，平均距离越小，移动图距离聚类簇的真实中心越接近，在结构上包含的相似性信息越多。将簇内的移动图按照  $\hat{\delta}_i$  值进行升序排列，则我们提取出的显著介观模式为

$$\mathcal{M} := \{M_{i,i+1} | \hat{\delta}_i \leq \hat{\delta}_{i+1}, i \in [1, L]\} \quad (3-22)$$

且  $M_{i,i+1}$  的簇内平均距离为  $R_i = (\hat{\delta}_i + \hat{\delta}_{i+1})/2$ 。这里  $R_i$  越小，表示该介观模式对于代



0000294

表簇内时空结构的显著性越高。最终  $\mathcal{M}$  便构成一个聚类簇内的显著模式集合。PPM 算法的伪代码如算法3-2所示。

**模序 v.s. 介观时空模式：**模序<sup>[53]</sup>和介观模式都是对用户行为结构的描述，但二者在概念、分析方法和应用场景上均有所不同（表3-1）：

表 3-1 模序分析和介观模式挖掘对比

Table 3-1 Comparison of motif analysis and meso-structure mining

维度	模序	介观模式
移动性	空间拓扑	时空结构
模型	图同构	属性有向图匹配
边保留	是	否
匹配方式	精确	非精确
复杂度	P	介于 P 和 NP 之间
应用	空间信息挖掘	带属性的时空行为分析

在概念上，模序将用户的位置抽象为无属性的停留点，位置转移通过点之间的有向连接表示，从而用户在一天内的行为利用有向图表示。这样的表示方法仅仅包含了位置之间的空间关系，而忽略了位置之间存在的差异，如路过的一个便利店和停留五个小时的工作地点在模序中都用单个点表示。介观模式是对用户移动行为过程以及时空结构的抽象表示，其中位置和位置转移的权重信息不仅衡量了不同地点对于用户的重要性差异，而且刻画了地点与地点、以及转移行为之间的依赖关系强度。

在分析方法上，因为模序仅包含移动位置的空间拓扑关系，因此其理论模型为图同构问题，即一个有向图经过有限步变化后能够完全等同于另一个有向图。图同构在匹配中同时保留顶点和边的相对关系，属于精确的匹配方法，因此对于时空数据质量以及用户行为多样性的鲁棒性有限。介观模式既包含移动位置的空间拓扑，又包含时空属性，其理论模型为属性有向图匹配问题。和图同构相比，图匹配从局部到整体对两个属性有向图的相似性进行衡量，具有无保边的特性，属于非精确的匹配方法，因此鲁棒性较高。

从应用场景上来讲，模序分析结果有助于解释在不同场景下人类行为的一般规律，但是由于仅仅包含空间拓扑而缺失属性的量化信息，因此适合于大规模行为的统计分



0000294

析，而不适用更细粒度上的个体行为预测或建模。介观模式不仅对模序概念进行了延伸，而且我们提出的属性图匹配 TACSim 算法和显著模式提取 PPM 算法，适用于能够抽象成属性有向图形式的其他用户时空行为，如移动用户的参与行为。

### 3.2.3 介观模式数据分析

#### 3.2.3.1 移动图统计特征

在这部分中，我们利用 CITY-M 数据集对城市尺度上的介观时空模式进行分析。首先，对于个体用户，我们依照人类普遍的生活规律将轨迹记录分割到不同的时间段（上午 03:00 至第二天上午 03:00），并对每个时间段建立用户的移动图。转移代价权重  $c$  通过起、止点映射在路网地图上的最短路径（即在仅已知停留点数据时用户最有可能选择的路径）距离表示，停留时间权重  $d$  通过每个地点的累计停留时间表示（根据最大熵原理，相邻地点之间的间隔时间被平均分割）。由于 TACSim 算法基于移动图结构的迭代过程，因此移动图的复杂程度对模式挖掘的算法性能有较大影响。图3-5展示了 CITY-M 数据集中用户移动图的停留地点数和顶点度分布。如图所示，数据集中超过 75% (95%) 的用户访问了不超过 10 (25) 个独立地点，类似的结论也在其他数据源的研究中<sup>[16,40,53]</sup> 得到了印证。同时，我们利用顶点的平均度数评估不同地点之间的依赖程度，发现超过 60% (15%) 的用户的平均顶点度数大于 2 (3)，从而表明这些用户至少拥有一次重返已访问地点的行为。

#### 3.2.3.2 显著介观模式提取

我们利用利用 PPM 算法提取 CITY-M 数据集中用户的介观时空模式，并和模序分析的结果进行了比较分析。图3-6首先给出了使用层级聚类对不同组的移动图的聚类结果。为了清晰表达的需要，我们从原始数据中随机选择了 400 个移动图进行分析，图中颜色表示3-20式的结构距离（颜色越接近红色则距离越大），参数  $C$  代表图基数， $K$  代表聚类簇的数目， $p$  表示不同  $C$  值的移动图占样本总体的比例。由于人类行为具有较高的返回概率，因此我们着重分析了  $C \leq 20$  的分组。从图中可以看出，不同组的聚类中的最佳团簇数处于 2~5 之间，且簇之间有着较清洗的边界。我们同时观察到，不同个体之间的时空结构具有较大的相似性，例如，大约  $p = 42\%$  的用户由于行为规律较为简



0000294

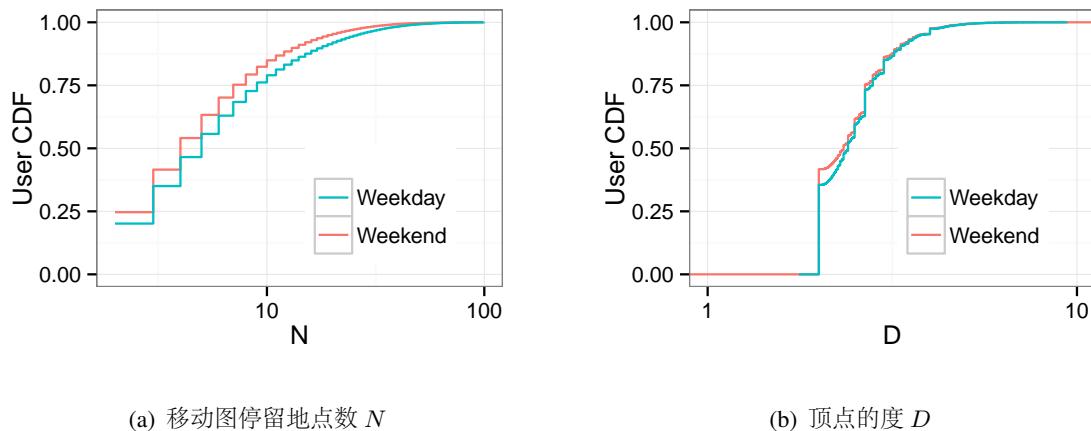
(a) 移动图停留地点数  $N$ (b) 顶点的度  $D$ 

图 3-5 CITY-M 数据集中用户的移动图在工作日和周末的特征分布

Fig 3-5 The scales of mobility graphs for both weekdays and weekends

单 ( $C = 2$ ) 而共享一种时空结构；随着图基数  $C$  的增加，图中蓝色的区域开始逐渐减小，这样的规律和我们的实际观测较为一致，即随着停留地点数的增加，用户行为保持规律性的可能性在不断降低。

接下来，我们着重对图3-6中的蓝色区域进行分析。从经验直觉上来讲，用户行为的复杂程度应该是和停留地点总数呈正相关，因为地点数越多，用户越有较高的概率访问当前地点以外的新地点。但是数据结果显示与经验有所冲突的现象，即个体的行为结构在图基数  $3 \leq C \leq 5$  呈现出轻度的“杂乱”现象，而在区域之外的两端 ( $C = 2$  或  $C = 6$ ) 表现更加紧凑清晰。类似的现象和趋势同样在 Schneider 等<sup>[53]</sup> 的模序分析中观测到。但是模序和介观模式的观测分别从不同角度解释了人类移动行为的性质：模序分析从时间角度表现出越是行为活跃的个体，随着时间的推移这种活跃程度依然存在；而介观模式分析表明，个体行为的活跃性和复杂性二者并没有直接的联系，如活跃用户虽然经历了更多的停留地点，但是其行为结构的复杂程度和潜在的可能构成相比，仍然具有较少的变化，即活跃性较低。

对于每个簇中的移动图，我们分别提取了最具代表性的介观模式和模序，并从量化角度对二者在行为特征上的表达进行了分析。图3-7展示了三种图基数 ( $C = 3, 5, 7$ ) 对应的模式分析结果，其中  $k$  表示各组内聚类簇的个数。在每个聚类簇中，对最具代表性的介观模式（对应于  $R_i$  最小）和模序（对应最高的出现频次  $f$ ）进行了对比，其中模



0000294

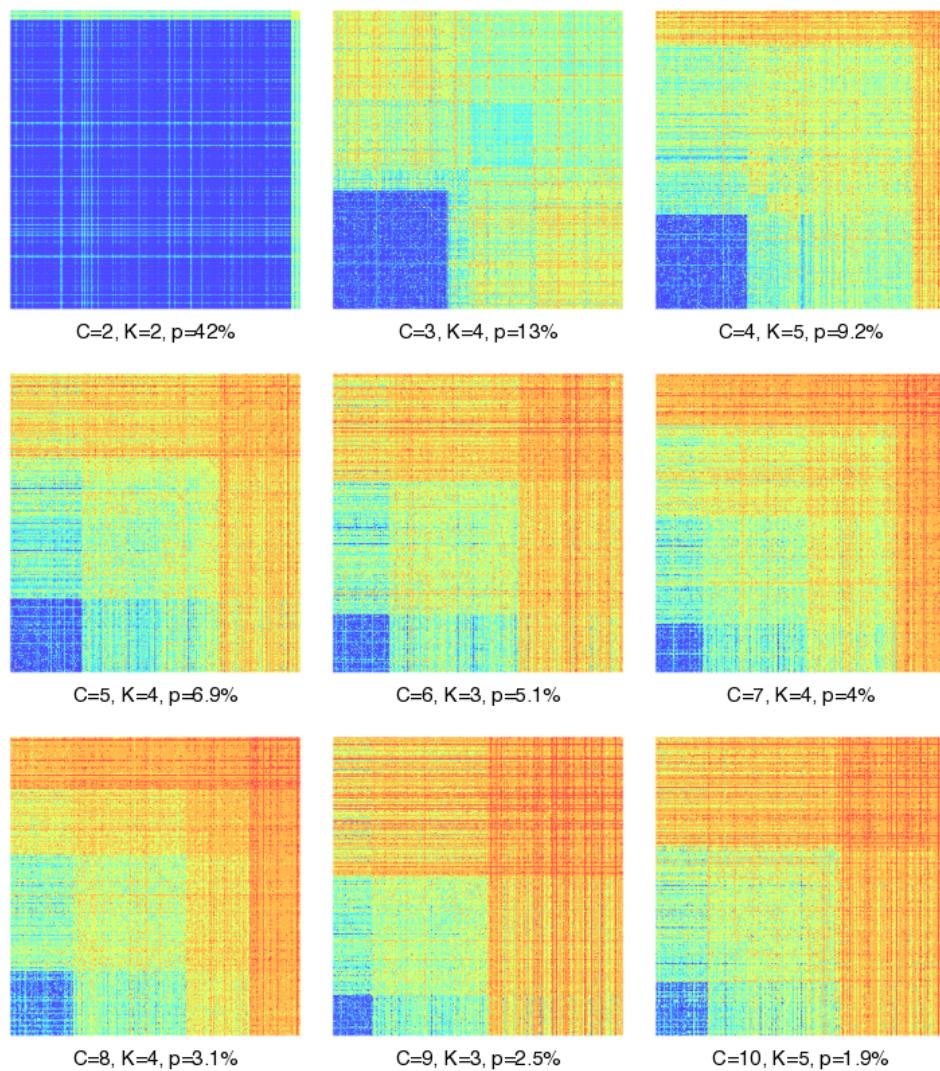


图 3-6 具有不同图基数的分组的层级聚类结果可视化

Fig 3-6 Visualization of hierarchical clustering with varying group cardinality ( $C$ ).

序位于对应簇的嵌入图内，且介观模式的点大小和边长度分别正比于停留位置的累计停留时间和转移距离。通过比较可以看出，和代表纯空间结构的模序相比，介观模式比较直观地体现了人类行为的时空结构，比如，虽然聚类簇 ( $C = 5, k = 2, f = 0.44$ ) 和 ( $C = 5, k = 3, f = 0.37$ ) 具有完全相同的模序和相近的出现频次，但是介观模式揭示了更多行为含义：在聚类簇 ( $C = 5, k = 2$ ) 中，用户行为类似于 “Home → Work → … → Home” 的移动方式，然而在簇 ( $C = 5, k = 3$ ) 中，两个停留时间较长的点之间的路径

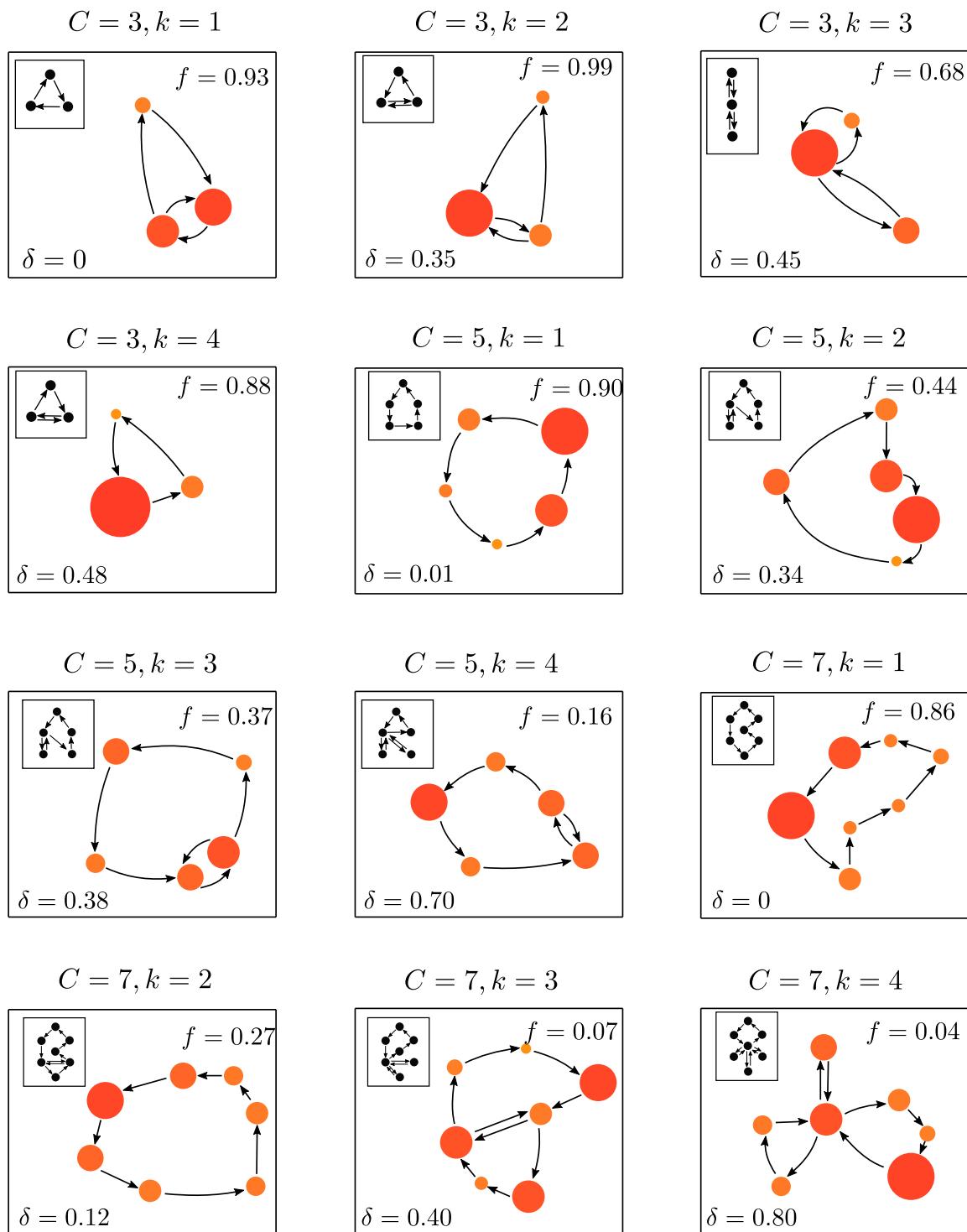


图 3-7 不同聚类簇里的最显著介观模式和频次最高模序对比

Fig 3-7 Illustration of top mesostructures extracted from each cluster.



0000294

圈 (Circle)，暗示了这两个停留点对于研究对象来说是紧密关联在一起的。

为了研究图基数  $C$  对于介观模式分析结果的影响，我们将介观模式的相似度度量和模序分析的频率进行了对比研究。从图3-7中可以看出，这两个指标在图基数较小的范围内（如  $C \in \{3, 5\}$ ）表现出较好的一致性，而对于较大的取值（如  $C = 7$ ）刻画的用户行为差异性较大。对于这一现象的一种解释是，当  $C$  取值较小时，无论行为的拓扑结构还是时空属性的变化范围都不大，且介观模式中潜在的包含了模序的空间拓扑信息，因此二者在行为刻画上的量化指标趋于一致。但是，另一方面，由于模序分析通过图同构的精确匹配，因此随着  $C$  值的增大，微小的空间结构变化都会导致频率  $f$  的下降；而介观模式由于采用鲁棒性较好的相似图匹配，因此即使顶点或边的数目、属性发生一定量的变化，TACSim 算法依然能够对两个移动图的公共部分进行很好的识别。

最后，我对不同组内得到的介观模式进行了分析。例如，对于图3-7中图基数  $C = 5$  的小组，相似度最高的介观模式， $\delta = 0.01$ ，来源于图3-6中对应小组的蓝色簇中，其原因在于该簇中的用户有着顺序访问的行为模式，且与其他小区的行为差异较大。但是当我们减少层级聚类算法的簇数目  $K$ ，结构较接近或边界较模糊的介观模式之间的差异性将变小，如  $C = 5$  小组中的  $k = 2$  和  $k = 3$  将合并成更加粗略的介观模式。相反的，当我们增加参数  $K$  的值，介观模式将包含更多时空行为的细节信息。总之，聚类簇数目的选择应该和实际应用较好地匹配，例如拥有两个以上的停留时间超过 30% 的地点的行为模式，而不仅仅从算法最优的角度进行选择。

### 3.2.3.3 自距离分析

自距离 (Self-Distance) 是我们在介观模式分析中发现的一种新的移动行为属性。从定义上来讲，自距离衡量了属性有向图（如移动图或介观模式）与自身的相似程度，这部分对自距离在介观模式发现、用户行为表征、以及时空行为分析中的应用进行研究。

图3-8首先展示了自距离  $\delta_{ii}$  和聚类簇内平均距离  $R_i$  之间的关系，我们发现二者在量值上存在多模式的相关关系。具体而言，我们将相关关系分为四种类型，即零模式 (28.3%)、对数模式 (7.5%)、线性模式 (60.8%) 和随机模式 (3.4%)，其中括号内的百分数表示特定模式的观测点占总体的比例，且线性和对数模式的量化参数如图所示。通过对模式在不同组内的分布特点进行分析，我们发现零模式和线性模式在  $C \geq 3$  的各

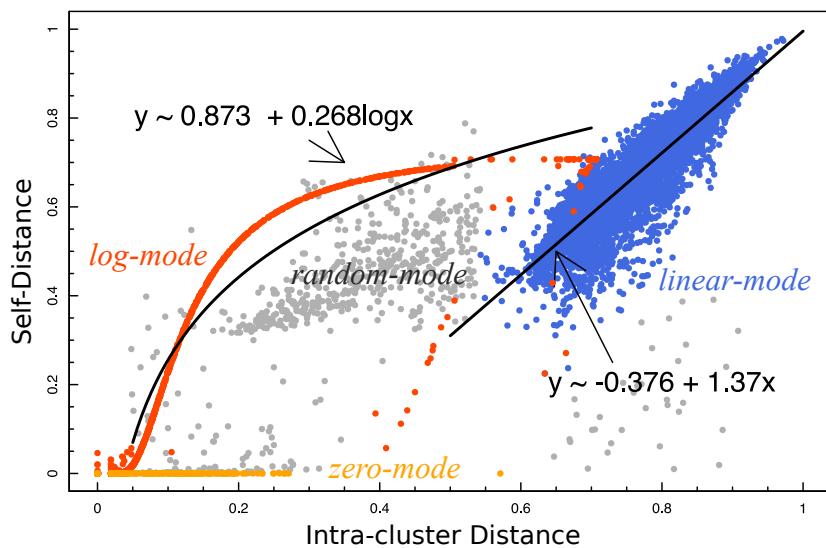


图 3-8 移动图自距离和聚类簇内平均距离的多模关系

Fig 3-8 The multi-mode relationships between self-distance and intra-cluster distance.

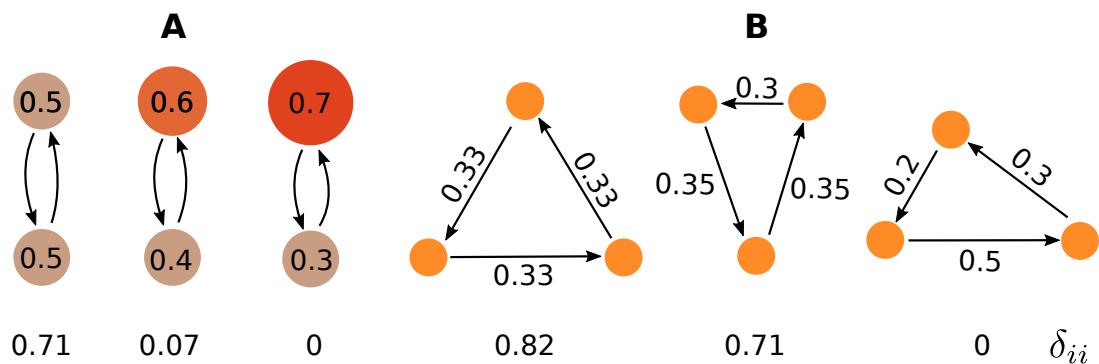


图 3-9 自距离和用户移动时空结构之间的关系展示

Fig 3-9 Characterization of self-distance on the heterogeneity of mobility structure.

组内几乎是均匀分布的，而对数模式主要来自于  $C = 2$  的小组中。

为了对不同模式的形成原因进行分析，我们研究了自距离对于用户行为特点的表征能力。在图3-9中，我们展示了自距离的变化趋势和地点停留时间以及转移距离之间的关系，进而发现自距离和用户移动行为的结构异质性（Heterogeneity）紧密相关，即

$$\delta_{ii} \propto \text{Structural Heterogeneity} \quad (3-23)$$

换句话说，在移动过程中，用户对的地点的偏好性越高、转移距离的偏离程度（Skewness）



0000294

越高，则其移动性的唯一性或可识别性越好。图3–9A展示了用户对地点的偏好变化；当地点的停留时间对称时， $\delta_{ii}$  取值为 0.71，且随着用户倾向于某一地点而逐渐趋于 0。图3–9B展示了空间转移距离对  $\delta_{ii}$  的影响，随着转移距离的对称性减弱， $\delta_{ii}$  的取值从 0.82 递减到 0。自距离这样的性质由 TACSim 算法的基本原理以及结构距离的定义所决定：在 TACSim 算法中，移动图元素的相似性受到邻居属性以及拓扑关系的影响，因此自距离大小与移动图元素的内部关系紧密相关；同时，根据结构距离的定义（3–20式），由于  $G_A$  中单个顶点和边元素的相似值通过  $G_B$  的元素来计算，因此  $G_B$  中元素对称性越高，所得的最优匹配元素相似值越低，进而自距离的取值越大。结合图3–8进行分析，零模的介观模式通常与簇内元素的平均距离较小，因此具有较高的可识别性；对数模的介观模式由于集中在  $C = 2$  的小组中，因此意味着停留时间对行为结构的异质性影响在  $R_i = 0 \sim 0.65$  的范围内呈现指数增长的关系；而线性模的介观模式融合了空间和时间上的多样性，和模式本身的复杂程度相比， $\delta_{ii}$  在  $R = 0.5 \sim 1$  范围内的线性变化趋势显得很有趣，因此值得在多尺度的分析中进一步研究。

自距离表示属性有向图的内在特性，以城市结构分析为例，将城市内的交通枢纽或兴趣点抽象为顶点，顶点之间的人群流动方向和数目抽象为边，则城市的人群移动可通过属性有向图表示。在这样的场景中，自距离分析通过挖掘地点之间的对称性信息，在约束边属性的前提下，能够挖掘出具有相似通勤模式的位置点，如不同的景点在上、下午时段出行和返程方向上的人数差异。因此自距离分析对于其他能够转换成相似模型的行为分析有着潜在的应用价值。

### 3.3 基于介观模式的个体移动模型

#### 3.3.1 现有模型的时空分布假设冲突

在前面的研究当中，介观模式揭示了人类行为中普遍存在的时空依赖规律，而对观测规律的重现以及应用则需要移动行为的理论模型。移动模型是通过对经验观测的理论总结、量化，从而形成对用户行为分析的系统化方法。但是现有的移动模型中依然欠缺人类行为表现出来的时间和空间相关性。如基于模序分析的扰动模型<sup>[53]</sup> 缺少空间依赖的信息，并且时间上的相关性通过经验式的宏观概率函数给出；而在连续时间的随机游



0000294

走 CTRW 模型中<sup>[14,15]</sup>，随机游走的个体拥有停留时间和转移距离的特征，但是这些特征的取值是从宏观统计分布中采样得出。

无论是 CTRW 模型还是扰动模型，均拥有一个前提假设，即个体移动行为的时间和空间统计特征与宏观观测是一致的。基于这个假设，如果群体的停留时间服从无标度的重尾分布，则个体行为被认为也是服从重尾分布的。但是从人们近些年来的行为分析中发现，这一假设存在一定的局限性。例如，Hidalgo 等<sup>[104]</sup> 通过理论分析得出，一群具有泊松特征的个体在统计上也能够呈现出重尾分布的规律。González 等<sup>[16]</sup> 发现个体的移动行为在空间上有很大的差异，但是这种差异却被群体行为的异质性所掩盖。在对 CITY-M 数据集的介观模式挖掘中，我们观测到个体行为和群体行为的时空特征满足不同的统计规律。从已有的观测和分析出发，本研究对上述假设提出修正，即个体移动行为的时空特征与群体的宏观观测拥有不同的统计分布，且二者有所联系。基于此假设，我们在干扰机会框架<sup>[19]</sup> 内提出了融合介观模式信息的新型个体移动模型。在这种模型中，个体根据机会地图上的资源分布和自身需求进行移动，从而行为的时空特征基于物理过程生成，而不是纯粹的宏观统计分布的采样。

从应用角度来讲，该模型为联系微观移动模式和宏观统计规律提供了一种有效的途径。通常情况下，微观上个体移动行为随着环境而改变，如城市道路改建维修或新设施建造。宏观统计分析的不足在于对行为模式的变化缺少解释性，如个体停留地点数的分布函数  $L(t) \sim t^s, s < 1$  中，当指数从 0.4 变化为 0.8 时微观的个体模式变化无从得知。基于物理过程的模型由于摆脱了对宏观分布的依赖，因此在验证模型有效性的前提下，可以在对模型的生成过程进行改变的同时对宏观分布进行测量，从而深入理解微观移动模式和宏观统计规律之间的联系。

### 3.3.2 基于介观模式的流涌现模型

在提出的模型当中，我们基于干扰机会（Intervening opportunity, IO）框架<sup>[19]</sup> 对个体的移动过程进行模拟。具体而言，在一定的空间范围内，不同地点拥有不同的机会分布，且密度分布随着时间的推演发生变化。在建模过程中，移动个体在满足时间、空间和机会资源限制的前提下，对停留地点进行选择并生成移动轨迹。由于在此过程中，不同用户的移动性包含了介观尺度上的行为模式，且生成过程如同在机会地图上产生新

的移动流，因此我们称这种模型为“流涌现”模型（Flow Emergence Model, FEM）。下面首先对模型中的基本概念和模块进行介绍，然后对模型的生成算法进行说明，最后利用 CITY-M 数据集，对群体观测特征进行实证分析，并对流涌现模型的有效性进行验证。

### 3.3.2.1 干扰机会框架介绍

干扰机会框架由 Stouffer 等<sup>[19]</sup> 在社会学研究领域首次提出，用于从机会分布而非距离特征角度理解人类迁移行为。最初框架的理论验证是通过交通调研数据完成的，从机会角度对城市宏观的人群流动进行了很好的预测。该框架对人类行为的刻画能力又在近期的辐射模型研究中<sup>[20]</sup> 得到了体现。干扰机会框架的基本观点认为：从给定地点到一定距离以外的另外一个地点的迁移人数，正比于另一个地点的机会量，而反比于两个地点之间的干扰机会总量。如图3-10A 图所示，当给定距离  $s$  的时候，以用户为中心、 $s$  为半径的区域内的机会分为两类，即用户感知机会和干扰机会。用户感知机会指在距离用户  $s$ 、宽度为微分量  $\Delta s$  的圆环区域内的机会数量，而干扰机会则为圆环和用户之间区域的机会总量。假设用户所在位置的人数为  $y$ ，向  $\Delta s$  区域的迁移人数为  $\Delta y$ ，则干扰机会框架表示为

$$\frac{\Delta y}{\Delta s} \propto \frac{1}{x} \cdot \frac{\Delta x}{\Delta s} \quad (3-24)$$

其中  $\Delta x$  表示圆环区域  $\Delta s$  内的机会量，且  $x$  表示该区域内的干扰机会总量。如果已知机会量和距离的分布关系为  $x = f(s)$ ，则上式表示为

$$\Delta y \propto \frac{f'(s)}{f(s)} \Delta s \quad (3-25)$$

进一步积分得

$$y = a \ln f(s) + c \quad (3-26)$$

3-26式表明  $s$  距离内停留的总人数正比于该区域内的机会总数的对数值。

干扰机会框架的优势在于以机会资源为媒介，替代了用户移动性对空间距离的直接依赖关系。在实际应用中，该框架的局限性在于机会资源的空间分布函数  $f(s)$  难以直接测量，尤其在较大空间范围内，如城市和国家。另一方面，干扰机会框架源自于对群体迁移问题的研究，即使最近的辐射模型<sup>[20]</sup> 也依然延续了这样的思想。本研究将干扰机会框架引入到个体行为研究中，并利用介观模式的时空特征对干扰机会框架进行增强，



0000294

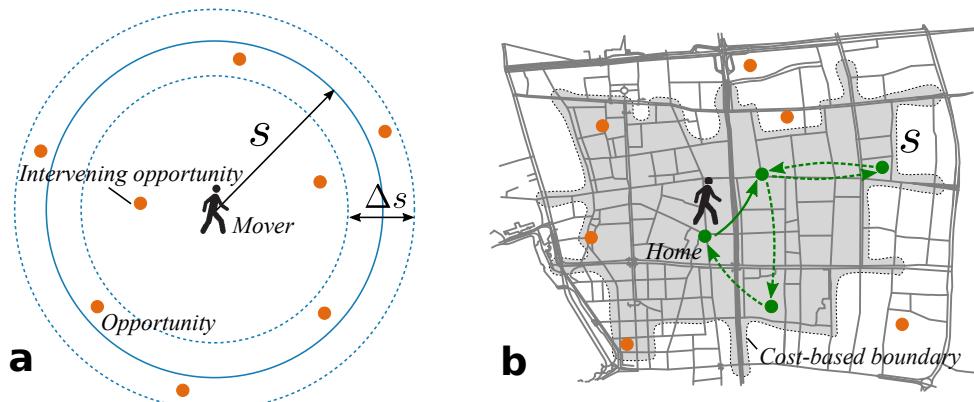


图 3-10 干扰机会框架和基于转移代价的个体机会地图示例

Fig 3-10 Illustration of the Intervening Opportunity Framework and cost-based opportunity map.

通过物理过程驱动而非统计特征采样描述用户的移动行为。除此以为，针对  $f(s)$  函数难以直接测量的问题，我们在模型中引入了个体机会地图的概念，并基于移动网络数据动态生成个体机会地图，从而增强了产生个体行为轨迹的多样性。接下来介绍流涌现模型中的基本组成模块。

### 3.3.2.2 流涌现模型的基本模块

**个体机会地图：**在干扰机会框架中，机会首先被定义为在特定空间范围内移动式所获得的“收益”，如工作机会或满意程度。在实际场景下，也许会存在多种类型的机会、以及不同类型的机会会带来不同的个体收益。在本研究中，为了阐述模型的基本思想，我们假定只有一种类型的机会，且给定地点的机会量正比于该地点在观测时间内的平均人口数。在复杂场景中，该模型应得到必要的修正，例如个体针对不同类型的机会设置不同的收益权重，或采取不同的移动策略。利用移动网路数据，我们首先生成全局的机会地图，其生成方法为：通过扫描数据集中的用户的转移记录，统计出在过去  $T$  时间（如 24 小时）内，区域内任意两个位置之间的转移总人数，并生成带属性的有向图  $G_m = (V_m, E_m, \phi_m, \psi_m)$ ，其中  $V_m$  表示区域内所有地点的集合，且每个地点的权重为到达该地点的总人数  $\phi_m$ ； $E_m$  表示有转移记录的地点对的集合，且每条边的权重为基于路网匹配的最短路径距离  $\psi_m$ 。图3-11给出了 CITY-M 数据中不同时间段内的全局机会地图，可以看出不同地点的机会分布是不均匀的，且少数区域之间的转移路径负载了大多



0000294



图 3-11 CITY-M 数据集中不同时间段的全局机会地图

Fig 3-11 Illustration of the global opportunity map using CITY-M dataset.

数的人流量。个体机会地图构建在全局机会地图之上，其范围是以用户停留点为中心、路网距离  $S$  为半径的子区域，且随着时间的持续，用户的停留地点发生变化，其个体机会地图的区域范围也发生变化。图3-10B 展示了用户从初始地点“Home”出发时的机会地图，并在该区域内选择了实现箭头的指向点作为下一停留地点。

**个体移动性画像是**对个体移动特征进行概括而得的结果。为了获得更多的移动多样性和更好的模型鲁棒性，我们分别从时间和空间角度对个体行为进行总结和描述。在空

间特征上，我们利用回转半径  $R_g$ <sup>[16]</sup> 刻画个体的移动能力，其计算方法为

$$R_g^2 = \frac{1}{n} \sum_{k=1}^n (\mathbf{r}_k - \bar{\mathbf{r}})^2 \quad (3-27)$$

式中  $\mathbf{r}$  表示  $n$  个地点向量  $\mathbf{r}_k$  的平均值。从已有研究中可以看到，回转半径的分布特征中包含着不同个体的活跃性程度<sup>[16]</sup>、以及个体在不同活动区域内的潜在差异<sup>[32]</sup>。在流涌现模型中， $R_g \sim p(r)$  代表了移动个体在当前机会地图上发现潜在的机会资源的能力，从而驱动个体通过简单的物理规则生成多样化的移动轨迹。

在时间特征上，我们对个体在不同地点的停留时间进行建模。通过数据分析实验我们发现，个体的停留时间近似满足双模对数正态混合分布 (Two-mode lognormal mixture distribution)。假设两种模式分别拥有权重  $\lambda_i$  ( $\sum_{i=1,2} \lambda_i = 1$ )，则个体的停留时间分布可以表示为

$$p(d|\mu, \sigma, \lambda) = \sum_{i=1,2} \lambda_i \cdot p(d|\mu_i, \sigma_i), \quad (3-28)$$

其中参数  $\mu_i, \sigma_i \sim \Theta(\mu, \sigma)$  描述了单个用户的时间画像。可以看出，在流涌现模型中，因为个体的时间画像参数各不相同，我们没有直接利用宏观统计特征，而是建立了时间模型的参数模型  $\Theta$ ，即“模型的模型”。为了更好地对齐上式中的时间序列，模型需要根据人们的生活规律，将观测时间范围  $T$  (如 24 小时) 等分成相同数目的时间片段序列，如 1:120 或 1:240。

### 3.3.2.3 流涌现模型生成算法

这部分我们对流涌现模型中如何生成单个用户的轨迹进行介绍。具体而言，个体移动行为建模是个  $N$  步的规划过程，该过程受限于三个方面的因素：机会资源分布、个体的空间移动能力  $R_g$ 、以及移动时间  $T$ 。在每一步当中，个体依据自身对当前机会分布的了解和机会选择的规则，对下一步停留的地点和时间进行规划；且在进行局部规划的同时，需要满足全局上的时间和空间约束条件。

假设全局的资源地图为  $G_m$ ，拥有回转半径  $R_g$  的移动个体在第  $k$  步时的个体机会地图由以下顶点子集 (及相应的边) 给出：

$$V^{(k+1)} := \{\mathbf{r}_j \in V_m \mid \sum_{\mathbf{r}_i \in V^{(k)} \cap \{\mathbf{r}_j\}} (\mathbf{r}_i - \bar{\mathbf{r}})^2 \leq (k+1)R_g^2\} \quad (3-29)$$

其中  $\{\mathbf{r}_i, 1 \leq i \leq k\}$  表示已确定的前  $k$  次停留地点向量， $\mathbf{r}_{k+1}$  代表所有地点中除了前  $k$  个地点以外的其他可能的停留点，而  $\bar{\mathbf{r}}$  表示  $k+1$  个地点的平均位置向量。该式的物理意义在于，在确定个体移动能力的前提下， $V^{(k+1)}$  包含了用户可能感知到的周边潜在所有的机会资源。实质上，3-29式是对仅考虑已知地点的方法的修正，在这种被替代的方法中，个体根据当前所在的地点和回转半径探索附近的机会资源，即  $V^{(k+1)} = \{\mathbf{r}_j \in V_m | \sqrt{|\mathbf{r}_j - \mathbf{r}_k|} \leq R_g\}$ ，从而导致生成地点的回转半径  $R'_g \leq R_g$ ，使得模型低估了用户的真实移动能力；另一方面，仅考虑当前停留地点信息，缺少了用户已访问地点之间的空间依赖性信息。基于这些原因，我们采用3-29式更加准确地描述了个体周边的潜在资源分布，使得模型生成结果更加接近用户的真实行为。

在第  $k$  步移动中，个体根据机会地图上已知和潜在机会资源的分布，搜索符合自身需求的下一个停留地点。在搜索方法中，我们从用户当前所在地点  $\mathbf{r}_k$  出发，将机会地图沿着回转半径的方向划分为等宽的搜索带（Band area），并用参数  $\omega \in (0, 1)$  控制搜索带的密度。具体而言，假设共有  $L = \lceil \frac{1}{\omega} \rceil$  个搜索带，则第  $i$  个搜索带  $E_i$  包含宽度为  $\omega R_g$  的空间范围内的机会资源，且  $E_i$  内的机会资源密度为

$$\rho_i = \frac{V_i}{\omega R_g} \quad (3-30)$$

其中  $V_i$  表示搜索带  $E_i$  内包含的机会资源总量。这里我们假设在相邻两个搜索带（如  $E_i$  和  $E_{i+1}$ ）之间，机会资源存在较强的相关性；不相邻的搜索带则相关性较弱，且这里为简化模型而不予考虑。

接下来模型根据不同搜索带内的资源分布决定下一停留位置及时间。在这个过程中，移动个体首先从最近的搜索带进行搜索，即从  $E_1$  到  $E_L$ ，并且依据下式对相邻搜索带内机会资源进行评估：

$$p_i = \frac{2}{\pi} \arctan\left(\frac{\rho_{i+1}}{\rho_i} \cdot C\right) \in [0, 1] \quad (3-31)$$

其中  $C$  为调节因子。当个体到达搜索带  $E_i$  时，以概率  $p_i$  进入下一搜索带  $E_{i+1}$ ，或以概率  $(1 - p_i)$  选择  $E_i$  作为下一步最终停留的区域。这样一种机会驱动的生成过程也可以和物理粒子的跃迁模型进行类比。假设存在某种物理粒子能够在  $E_1$  到  $E_L$  中的相邻能级间进行跃迁，并且能够以较高的概率被具有高能量的能级吸收。在迁移过程中，粒子以概率  $p_i$  从能级  $E_i$  跃迁到  $E_{i+1}$ ，或以  $(1 - p_i)$  的概率停留在能级  $E_i$  并保持稳定。

**算法 3-3 流涌现模型 FEM 生成算法**

**Input:** A global opportunity map  $G_m$  of past  $T$  time, the number of emulated users  $N$ .

**Output:** A list of generated users  $U$ .

Initialize user list  $U \leftarrow \{\}$  and set time limit  $B_0 = \lceil \frac{T}{240} \rceil$ ;

**for**  $l \leftarrow 1 : N$  **do**

    Create new user  $u_l$  with spatial profile  $R_g \sim p(r)$  and temporal profile  $\mu_i, \sigma_i \sim \Theta(\mu, \sigma), i \in \{1, 2\}$ ;

    Generate a random starting location  $H_1 \in V_m$ ;

$k = 1, B = B_0$

**while**  $B > 0$  **do**

        Generate opportunity map  $G_l^{(k+1)}$  according to Eq. (3-29);

        Assign dwelling time  $T_k \sim p(d|\mu, \sigma, \lambda)$  for current location  $H_K$ ;

        Move  $u_l$  to  $H_{k+1} \in V_l^{(k+1)}$  according to the opportunity-driven strategy indicated by Eq. (3-31);

$B \leftarrow B - T_k, k \leftarrow k + 1$ ;

**end while**

    Move  $u_l$  back to  $H_1$  and update  $U \leftarrow u_l$ ;

**end for**

在干扰机会框架内，这种移动策略利用了较近搜索带内的机会对较远机会的干扰作用，并利用概率因素增加了用户移动行为中的随机因素。但是，和原始的干扰机会模型相比，我们的移动策略有两方面的改进：1) 利用机会资源密度代替机会总量，从而去掉了因为空间分布的不均匀性对模型性能的影响；2) 利用搜索带方法代替图3-10A中 $s$ 中的整个空间范围，从而克服机会资源在空间上的累积效应。

如果粒子在第  $k$  步中停留在搜索带  $E_i$  内，模型根据已经到达过的  $k$  个停留位置、以及他们之间的空间关系信息决定下一位置。对于每一个候选的停留位置，我们通过该候选到达已知  $k$  个地点的平均距离决定其空间排序值  $r_s$ ，并通过比较各候选的机会数量得出其机会资源排序值  $r_o$ ；对于与这两个指标，排序值越小，则对个体是否选择该候选地点的影响越大。在流模型中，我们选取最小的  $\hat{r} = 0.5 * (r_s + r_o)$  作为个体在第  $k$  步中搜索带  $E_i$  内的最终停留地点。

对于停留时间，FEM 模型在生成每个个体轨迹之前，首先从时间画像的分布函数获得一组参数  $\mu_i, \sigma_i \sim \Theta(\mu, \sigma)$ ，根据此参数和3-28式得到当前个体的停留时间分布函

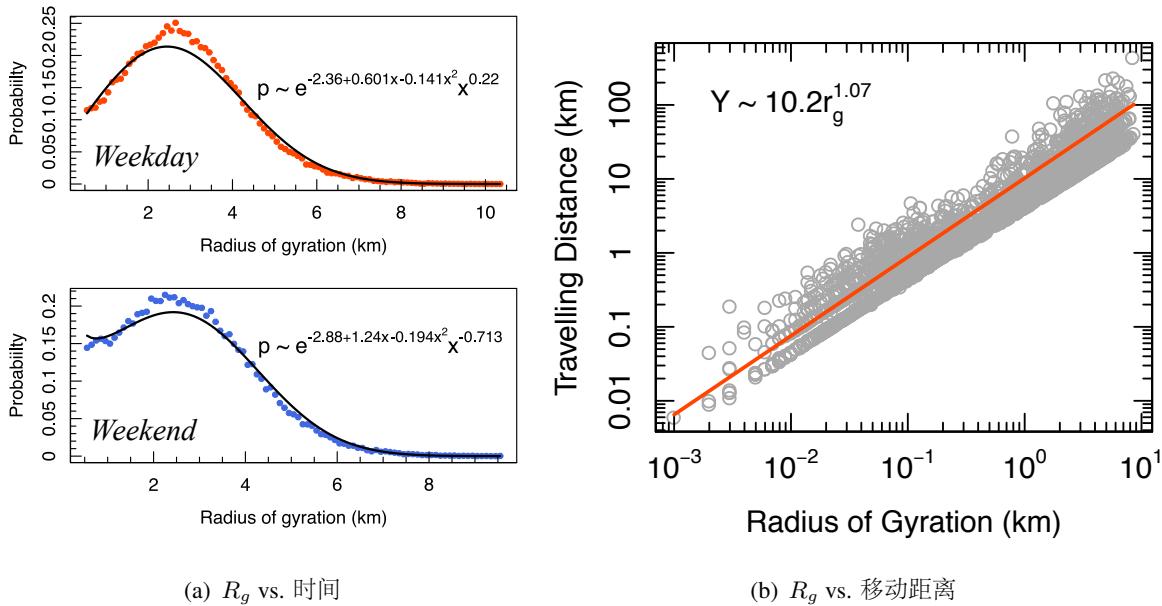


图 3-12 个体的回转半径与时间和移动距离的分布关系

Fig 3-12 The relationship of radius of gyration and traveling distance.

数。假设个体的总移动时间为  $B$ , 则在第  $k$  步是获得的停留时间  $T_k \sim p(d)$ , 且剩余时间资源为  $B \leftarrow B - \sum_k T_k$ 。不断重复上述移动策略和时间分配规则, 直至时间消耗完为止, 个体最终返回出发地点并结束自己的移动行为。该过程的完整过程由算法3-3给出。

### 3.3.3 模型性能验证

本节对提出的流涌现模型的性能进行验证和分析。我们首先利用 CITY-M 数据从实验角度证明了个体和群体行为特征在统计上满足不同的分布; 进而对流涌现模型的性能进行了验证, 并与随机游走模型和最大机会模型进行了比较。

#### 3.3.3.1 空间特征实证分析

我们首先对用户个体的  $R_g$  分布特征与不同用户之间的分组和聚类结构进行了关联分析, 但是实验结果并未显示出介观模式的不同对  $R_g$  分布特征的显著影响, 这意味着个体移动能力的大小和行为的时空结构并没有直接联系。但是在时间上, 个体的移动时间越长则会有较高的概率产生较大的回转半径。图3-12A 展示了一周内工作日和周末不同时间段内的  $R_g$  分布, 可以观测到回转半径在 2 ~ 3 公里附近存在明显的特征值, 这

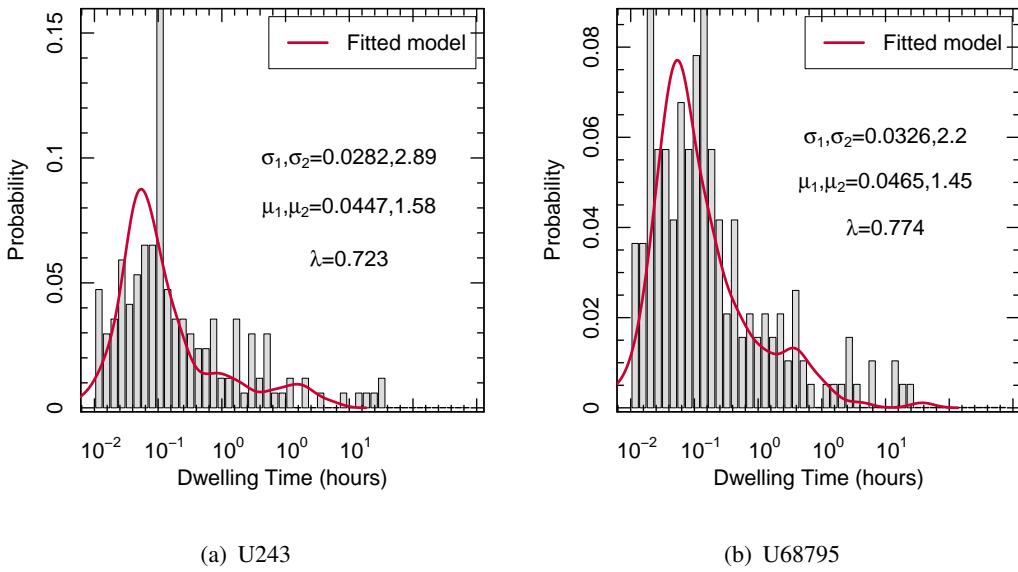


图 3-13 个体停留时间分布的直方图和拟合概率密度分布

Fig 3-13 The distributions of dwelling time at the individual level.

表示大多数用户倾向于在选择较近的机会资源满足自身的需求。在量化角度上，回转半径的分布特征较好地拟合  $p \sim e^{a+bx+cx^2} x^{-\lambda}$ ，且工作日和周末的 K-S 距离为  $d_{ks} = 0.09$ ，这表明个体的工作状态并未对分布造成明显的影响。图3-12B 给出了一天时间内回转半径和用户的旅行距离之间的关系，且二者关系满足  $y \sim LR_g^\lambda$ ；该结果从实验角度证明了我们选择  $R_g$  作为个体移动能力标识的有效性，即  $R_g$  越大用户的移动能力越强。从机会分布特征上看，移动能力越强的用户便以较大的概率接触到更多潜在的机会资源。

### 3.3.3.2 时间特征实证分析

我们首先对个体的时间画像和群体宏观观测在统计分布上的差异性。从个体尺度上来看，图3-13展示了 CITY-M 数据集中随机选择的两个用户（U243 和 U68795）在观测时间段内的停留时间直方图。我们可以看出个体时间特征可以被双模的对数正态分布所拟合，其中两种模式分别对应较长和较短的停留时间特征值。例如，用户 U243，两种模式的平均停留时间为  $\mu_1 = 0.0447h$  和  $\mu_2 = 1.58h$ ，二者的调和参数为  $\lambda = 0.723$ 。这样的个体行为特征的物理意义为，当用户移动时，在机会资源较多的地点选择停留较长的时间，而较少的地点则成为移动过程中的路过地点，相对停留时间较短。从群体尺

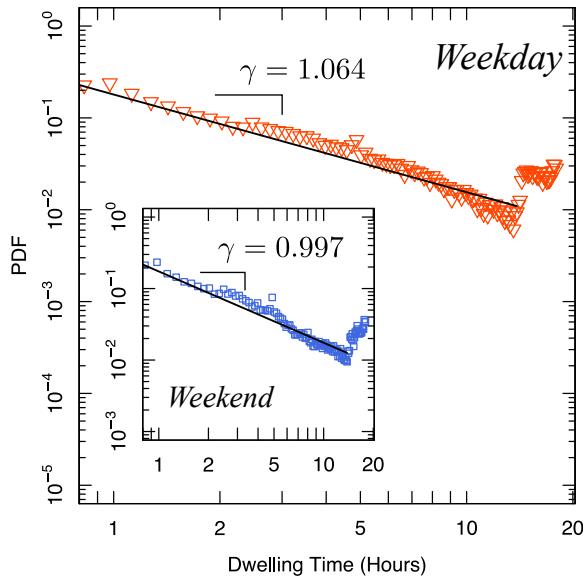


图 3-14 群体停留时间在工作日和周末的概率密度分布

Fig 3-14 The distributions of dwelling time at the population level (weekday vs. weekend).

度上看，图3-14展示了不同时段内的停留时间分布，可以看出工作日和周末都近似满足形式为  $p \sim t^{-\gamma}$  的幂律分布 ( $\gamma_{wd} = 1.064$  和  $\gamma_{we} = 0.997$ )，但是工作的下降速度和周末比较为平缓，二者的 K-S 距离为  $d_{ks} = 0.125$ 。对比个体和群体的行为特征，我们发现二者并不满足相同的分布规律，且和 Hidalgo 等<sup>[104]</sup> 分析结论类似，大量双模的对数正态分布叠加可产生重尾分布的特征。因此已有模型（如 CTRW 模型）将群体的宏观特征作为个体时空特征来源的假设存在一定的局限性，且这一局限性在我们的 FEM 模型中得到了克服。

接下来我们从时空依赖性角度出发，分析用户的位置偏好对时间特征的影响。图3-15和图3-16分别展示了纯时间和时空结合角度的停留时间模型（3-28式）的参数空间。纯时间模型中用户在同一地点的停留时间作为两个观测值，而时空结合模型中，我们对同一地点的停留时间求取平均值作为其特征值。如图所示，在纯时间模型中，参数  $\mu$  和  $\sigma$  都具有两个不同的峰值，而考虑了时空依赖性信息以后，时空模型的参数  $\mu$  和  $\sigma$  均表现得更加紧凑，且具有一个显著的峰值。由此可以看出，考虑了空间偏好性以后，个体用户的停留时间模型更加简洁，如图3-16中，平均停留时间的特征值为  $\mu_1 = 10^0 h$  和  $\mu_2 = 10^1 h$ ，对应方差的特征值为  $\sigma_1 = 0.25 h$  和  $\sigma_2 = 25 h$ 。因此我们在 FEM 模型中采用

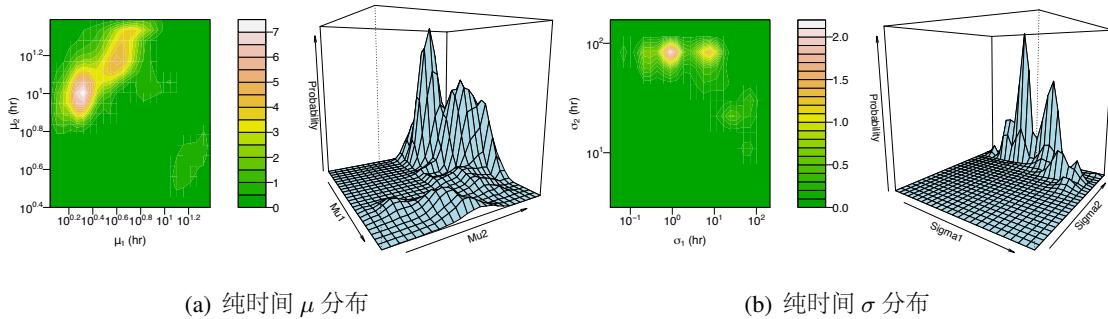


图 3-15 纯时间角度的个体停留时间模型的参数分布

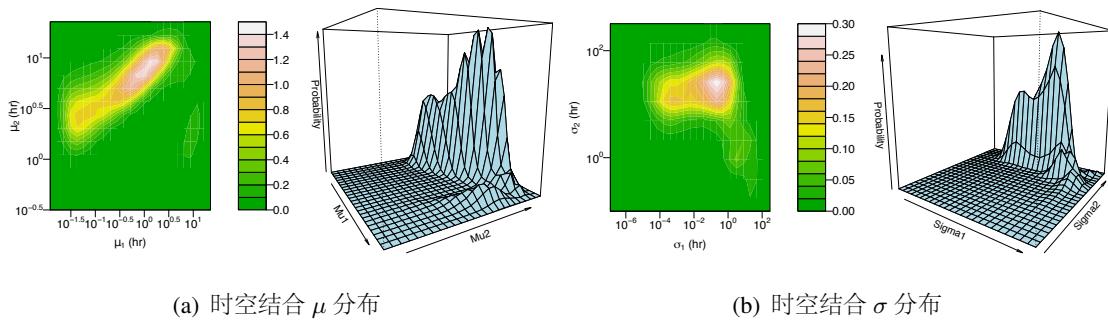
Fig 3-15 The parameter space for individual dwelling time models *without* location information.

图 3-16 时空结合角度的个体停留时间模型的参数分布

Fig 3-16 The parameter space for individual dwelling time models *with* location information.

时空结合的参数分布对停留时间进行建模。

### 3.3.3.3 流涌现模型性能验证

这部分我们对提出的 FEM 模型性能进行分析和对比研究。为了表现模型集成的机会驱动过程的优势，我们同时对两种不同的模型进行了对比试验，分别为随机游走模型和最大机会模型。另种模型均是对生成个体在机会地图上选择下一地点的策略进行了改变。随机游走模型（Random Walk Model, RWM）中，个体在第  $k$  步的机会地图节点集合  $V^{(k+1)}$  上随机选择一点作为下一停留位置，即在该模型中，个体忽略不同地点的潜在机会资源和物理距离，而赋予各地点相同的权重和移动概率。最大机会模型（Most Opportunity Model, MOM）中，个体在第  $k$  步选择  $V^{(k+1)}$  中尚未到达过的地点中，选取机会资源最多的一点作为下一停留位置。在验证实验中，我们利用图3-12A 中的工作日分布函数对用户的空间行为进行建模，并利用时空结合的停留时间参数分布对时间行为

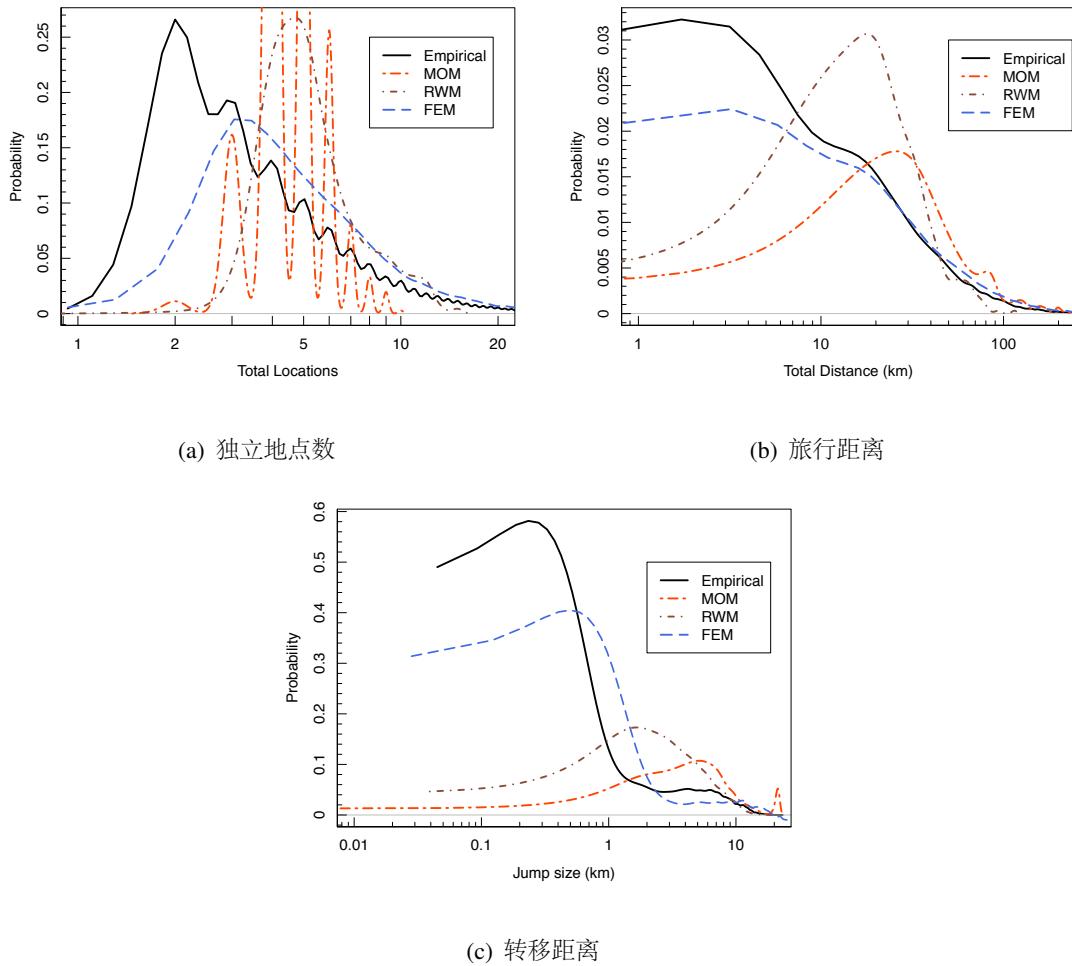


图 3-17 流涌现模型的性能验证（独立地点数、旅行距离、转移距离）

Fig 3-17 The evaluation of FEM performance w.r.t. total locations, traveling distance and jump size.

建模；实验结果分析如下：

图3-17A展示了不同模型产生的用户独立访问地点数的分布。我们发现MOM模型概率密度函数呈锯齿状分布。这是由于MOM模型优先选择搜索带内机会最多的地点作为下一停留位置，而在城市尺度上，不同地点的人数分布符合无标度的重尾分布<sup>[20]</sup>，导致当回转半径 $R_g$ 较小时，移动个体以较高的概率被限制有限的空间范围内，从而形成具有不同地点数目的峰值。对于RWM模型，分布曲线的右侧呈现出和经验数据分布一样的尾巴，但是该模型的特征值位于 $N = 5$ 左右，而经验数据分布的特征值为 $N = 2$ 。在我们的FEM模型中，产生了接近于经验分布的概率曲线，其特征峰值为 $N = 3$ 左右，



0000294

和另外两种模型相比，最接近于经验数据分布。

图3-17B 从累计旅行距离角度对比了三种模型和经验数据的概率密度分布。其中 **MOM** 和 **RWM** 模型倾向于产生较长距离的旅行，且 **MOM** 模型的特征值更大，即  $d_{rwm} = 20km$  和  $d_{mom} = 30km$ 。这是因为 **RWM** 模型对于机会地图上的潜在地点赋予相同的概率，而 **MOM** 模型倾向于选择最后欢迎的地点。**MOM** 的倾向性可能导致生成个体在机会数较多、且距离较远的少数几个位置之间来回移动，从而产生较大的移动距离。相比之下，**FEM** 模型同时结合已访问地点的距离和每个搜索带内的机会的局部密度信息，并根据概率  $p_i$  对选择下一位置或进入下一个搜索带。这样的策略使得个体倾向于选择附近比较受欢迎的地点，从而产生相对较小的旅行距离特征值。

图3-17C 对比了三种模型的转移距离概率密度分布。和旅行距离的分析类似，**MOM** 和 **RWM** 模型的转移距离特征值分别在  $c_{mom} = 6km$  和  $c_{rwm} = 2km$ ，而 **FEM** 模型的特征距离远小于前两种模型。和经验数据相比，虽然 **FEM** 表现出拥有较大转移距离的趋势，但是它捕获了经验分布中大多数用户拥有较小的转移距离、而一小部分用户拥有较大转移距离的特征。结合图3-17A~C 的结果，我们发现 **FEM** 模型在三种模型中具有最接近于经验数据分布的宏观统计特征，从而证明了利用介观模式中的时空结构信息对用户行为建模的有效性。

### 3.4 本章小结

本章从时空依赖性角度入手对个体的移动行为进行了研究。为了对传统个体行为序列模式挖掘和宏观的 **OD** 模式进行补充，我们提出了一种融合了个体移动行为的空间拓扑和时空性质的新模式，即介观时空模式。本章的研究方法和结果不但将图模式分析引入到个体时空行为研究当中，而且提供了一种新的角度对人类的时空行为进行理解。具体而言，在将个体行为轨迹转换成移动图的基础上，我们提出了 **TACSim** 算法衡量两个移动图之间的相似性，并提取其公共部分；进而提出一种带有修剪技术的显著介观模式提取算法 **PPM**，实现了对一组用户移动图的介观模式提取。我们提出了结合介观模式的时空特征的个体流涌现模型，该模型摆脱了传统模型中对个体行为和群体行为的统计特征一致的假设，且与经验数据分布在独立地点数、旅行距离、转移距离等多个宏观特征上表现出较好的一致性。



0000294

从一般性上讲，由于人类行为的动态性和复杂性，发现人类行为的时空结构依然是一项具有挑战的任务。虽然我们的方法成功提取了介观时空模式，但其依然具有两个方面的局限性：1) TACSim 算法中移动图的属性经过了归一化处理，从而使得介观模式忽略了行为属性的绝对差异。例如，假设两个用户的移动图具有属性相同的两个顶点，虽然二者的绝对移动距离不同，但是对于 TACSim 算法，二者具有相同的介观结构。这样的结果导致介观模式中丢失了两个用户行为在物理意义上的差异。2) 从大量移动图中提取介观模式依然面临着性能的挑战，这是由于显著介观模式的提取需要对移动图进行两两运算。虽然我们的 PPM 算法在损失一定信息的基础上提高了模式提取的效率，但是对大规模移动图的直接处理依然需要开发更高效的算法。

从应用角度来讲，个体的介观模式提取有助于交通网路优化和城市结构分析。介观模式分析能够解释城市内不同区域之间的时空连通性；如果结合更多的外部特征，如 POI (Position of interests)，我们便能够从功能和用户需求角度检测出城市内具有相似功能的区域，或者揭示出交通路网上不同的拥堵路段和拥堵状况在路网上的传播。对于流涌现模型，一个潜在的应用是用来分析移动个体的微观模式和宏观统计特征之间的联系。目前尚没有完整的理论和方法解释当微观个体的移动模式如何变化时，宏观统计上的参数（如幂律分布的幂指数）将发生显著变化。由于流涌现模型依赖介观模式的本地特征信息，因此我们可以通过改变生成个体的移动模式和群体组成来分析对宏观统计规律的影响。从而反过来利用流涌现模型更加深入地理解人类的时空行为。



0000294

## 第四章 群体移动行为的时空分布研究

前一章从个体介观模式角度揭示了人类时空行为具有普遍的时空依赖关系。当我们去掉个体差异、从更大的空间尺度上观测时，用户的群体行为同样也表现出较强的时空关联。例如在城市的交通网络中，由于区域功能的差异性，人群的分布和移动行为具有不对称的特点，上下班高峰路段拥堵严重，且随着时间的推移，这种趋势沿着道路网络不断传播。实际生活中的类似现象，来源于群体移动行为本身在时间和空间上并不完全独立，而是相互依赖，相互影响。但是在移动网络技术普及之前，群体行为的研究依赖人口普查机构的调研数据，即使这类数据覆盖的空间范围可以很广（如芝加哥旅行调研数据<sup>1</sup>），但是由于样本数较小，且用户记录和描述的误差较大，为测量和分析群体移动行为的时空关联带来了挑战。本章借助于被动的移动网络流量采集，从不同空间尺度，即校园（WIFI-M 数据集）、城市（CITY-M 数据集）、以及国家（Senegal-S 和 Senegal-A 数据集）尺度上对群体行为的时空关联进行分析和建模。通过对不同尺度上群体行为的对比研究，发现了人类行为中时空关联的一般规律，进而对群体行为时空规律的产生机理建立了理论模型，并通过时空预测的方法对模型进行了验证。该研究的成果可以在多个领域得到应用，如在移动网络部署中，最优的基站规划需要考虑所在地点的人群移动特点，从而以最小的资源投入获得最大的用户体验；在网络仿真领域，群体的时空分布特点有助于开发出更加真实的网络流量模型，为新协议的开发提供了便捷。

### 4.1 群体行为的研究背景

群体性的时空行为在自然和人工领域有着丰富的形式，如自然界里的动物季节性迁徙、植被覆盖范围的变化，人类社会中的移动网络用户移动、交通网络流量的变化、以及供电网络中的用电量变化等。与个体粒度的行为分析不同，群体行为的研究帮助人们从宏观尺度上把握复杂系统的动态变化，在对系统变化规律理解的基础上，反过来产生更有效的系统设计和管理策略。本文以此为出发点，从第三章观测到的个体时空模式中

<sup>1</sup>芝加哥旅行调研数据：url <http://www.cmap.illinois.gov/data/transportation/travel-tracker-survey>

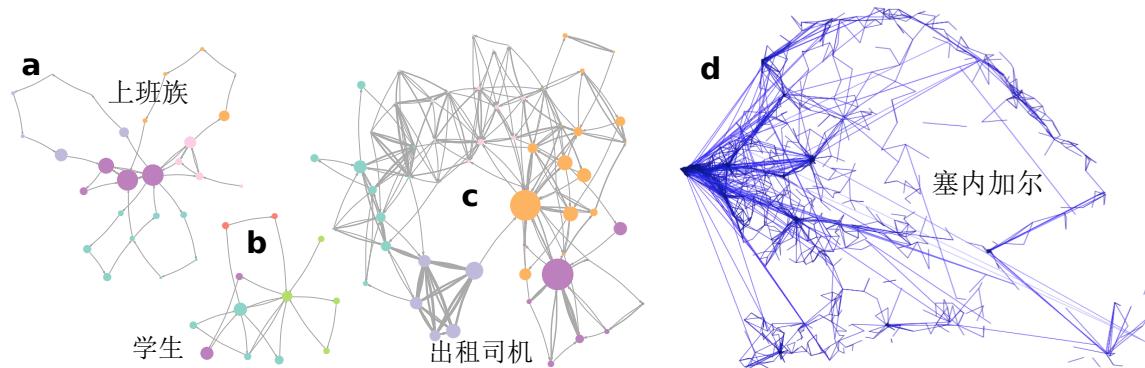


图 4-1 个体 (a~c) 和群体 (d) 行为的时空分布特征对比

Fig 4-1 The comparison of investigating mobility from individual and group perspectives.

引出群体移动行为的时空依赖性，并在多空间尺度下对群体移动行为进行了实证分析和理论建模研究。

虽然群体由大量独立的个体构成，但是在时间和空间观测角度上，二者的特点和产生机理并不完全相同。对于个体而言，行为上的时空依赖性更多来源于人们自身的差异和偏好，如年龄、社会角色等。图4-1中 a~c 展示了三种不同身份的个体所产生的时空模式，其中节点代表不同的停留地点，大小和颜色分别表示累积停留时间和同一个聚类簇，边方向和粗细分别表示移动方向和路径出现的频繁程度；可以看出个体的行为分布决定于个体的移动能力、位置偏好、以及不同地点的时间需求的不同，如图中出租司机较普通上班族表现出更加复杂的时空结构。对于群体而言，由于空间尺度远大于单个个体的活动范围，行为的时空关联决定于地理空间上的资源分布（如工作机会）和区域功能上的差异（如市区和郊区、居住地和道路等）。图4-1d 展示了 Senegal-S 数据集所描述的塞内加尔用户在一周内的时空行为分布，可以看出首都和其他主要城市构成了分布图的主要节点，国家级交通网络和航班路线构成了主要的转移路径。由于存在这些差异，基于个体介观模式揭示的时空依赖性，群体移动行为的研究需要从宏观尺度上对时间和空间特点进行关联分析。

在一定空间范围内，个体的移动行为导致群体的时空分布不断发生着变化，那么人群在时间和空间上的特征是如何关联在一起的呢？图4-2展示了三种不同的分析方法和思路。方法 a 以时间为导向，具体来讲，对于不同的空间位置，随着时间的推移特定空



0000294

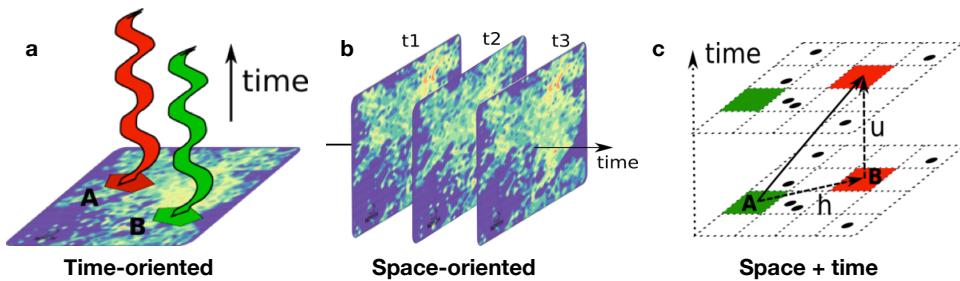


图 4-2 不同角度的群体时空行为研究方法比较

Fig 4-2 The comparison of investigating group mobility from spatial and temporal dimensions separately.

间点（如手机基站）产生一个独特的观测时间序列。通过对单个时间序列的分析，能够把握每个地点的变化特点及趋势；通过对两个或多个时间序列进行关联分析，便可以对地点之间的相互影响进行衡量。这种方法的优势在于纯时间序列的分析方法较为成熟，一般处理效率也较高；不足之处在于虽然时间序列观测是空间移动的结果，但是并没有包含显式的时空交互特征，尤其在不同的区域（如城区和郊区）这种交互特征有所不同。方法 b 以空间为导向，即在不同的时间片段里，人群移动形成了空间上的分布特征。通过对相邻时间片段内的空间分布进行对比，能够得出不同区域之间人群变化的差异。和方法 a 类似，这种方法的不足之处依然是缺少直接的时空依赖性信息。方法 c 克服了前两种方法的不足，即同时考虑群体行为在时间和空间上的相关性，如图所示地点 B 在  $t+1$  时刻的观测不仅与自身  $t$  时刻的值有关，而且依赖于  $t$  时刻地点 A 的观测。该方法立足于对群体移动过程的直接观测，具有显式的时空交互信息，因此是本节研究的主要方面。下面对时空交互信息的来源和具体形式进行举例说明。

从形成机制上分析，群体行为的时空交互信息主要来自于两个方面：**网络效应**和**时律性**。首先，网络效应是空间依赖性的主要影响因素。在各种各样的人造网络中，如移动网络、交通网络、以及电力网络，网络节点之间存在着较强的相关性，且这种基本属性容易导致节点之间的级联失效。例如在交通网络中，路段通行量接近饱和极限而发生拥堵，甚至瘫痪，从而将过剩的交通压力转移到空间关联的其他路段，并进一步使得其他路段发生饱和而拥堵。以此类推，网络中某处的状态会随着拓扑关系而发生扩散，从而形成不同地点之间的依赖关系。移动网络与此类似，一个基站的拥塞升高或体验下降同样会对其他基站造成影响。另一方面，人类行为具有较强的时律性，在宏观尺度上

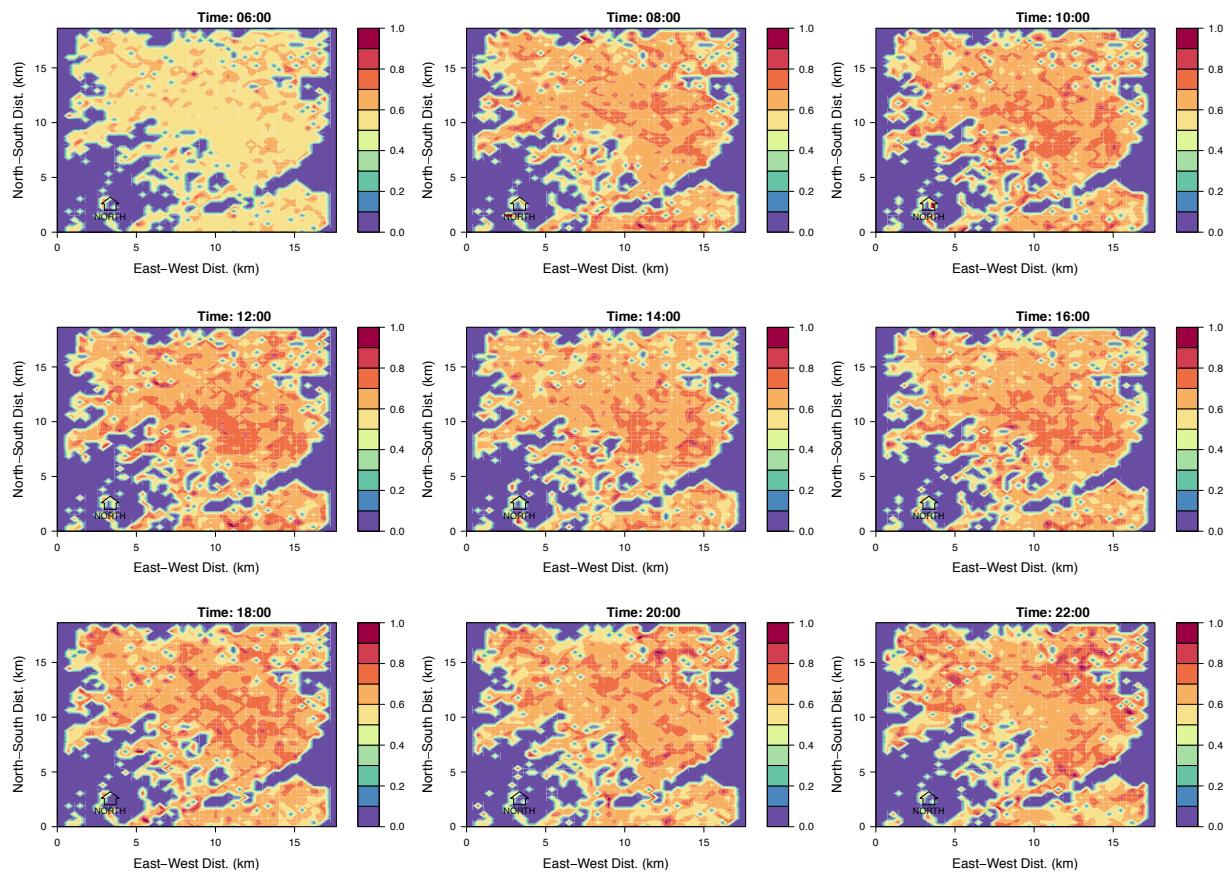


图 4-3 CITY-M 数据集中群体行为的潮汐效应热力图 (06AM~22PM)

Fig 4-3 Demonstration of spatial-temporal distribution of cellular traffic at the city scale.

表现为群体行为的“潮汐效应”，即人群有规律地在不同的空间分布之间进行切换（如图4-3），而这样的分布通常由区域功能所决定，如城市中的住宅区和商业区。当网络效应和时律性叠加时，例如移动网络基站人群的时律变化，会沿着移动网络拓扑进行传播，进而与其他基站的时律变化进行叠加，使得群体行为呈现出较强的时空依赖性。

本章利用不同空间尺度下的移动网络数据，对用户在宏观尺度上体现出来的时空依赖性特征进行分析，并对其形成过程进行理论建模研究。在研究过程中需要解决以下两个挑战：1) 人群分布值在时间和空间维度上都可以进行量化（如图4-2a 和4-2b），但是由于空间具有各向异性，而时间没有，因此二者的变化值难以直接进行比较分析；2) 时空关联分析中需要同时考虑用户行为在时间和空间上的依赖性，从而增加了模型



0000294

的复杂程度。尽管如此，对群体移动行为的时间和空间关联特性进行研究，不但可以缩小仅考虑时间或空间因素的理论模型和实际观测之间的差距，而且在网络调度和资源优化中具有重要的应用价值。例如在移动网络中，移动信号的干扰和基站的距离有关，从宏观上同时考虑网络资源的空间配置和时间动态性，能够实现对珍贵的频谱资源进行有效利用。

## 4.2 群体行为的时空统计分析

### 4.2.1 群体行为的时空分布描述

我们首先从过程角度对群体用户的时空分布进行描述。由于网络效应的存在，人群在时空上的分布具有连续性，即距离较近的空间点往往具有较高的相似度。因此对人群的时空分布特征进行描述，不仅要保持不同观测点的差异，还应该保留在观测维度上的连续性。具体来讲，我们对时间和空间维度建立索引，并将人群在特定时、空范围内的分布进行量化。因此在某个时间点上的人群数用标量随机变量进行表示，即  $y(\mathbf{s}, t)$ ，其中独立的随机变量  $\mathbf{s}$  和  $t$  分别表示空间和时间上的索引值，则生成该随机变量的统计过程由定义4.1给出。

**定义 4.1. 群体时空分布：**假设观测  $y(\mathbf{s}, t)$  由随机过程  $Y(\mathbf{s}, t)$  产生，我们将群体的时空分布过程定义为

$$\{y(\mathbf{s}, t) \sim Y(\mathbf{s}, t) | \mathbf{s} \in D_s, t \in D_t\}, \quad (4-1)$$

其中  $D_s = \{D_s^{(k)} : k = 1, \dots, N\}$  和  $D_t = \mathbb{Z}$  分别表示观测的空间和时间范围。

虽然4-1式中的  $D_s$  和  $D_t$  为连续的定义域，实际观测和分析中由于数据源的限制，往往造成定义域离散化。在我们的分析中，时间域被等分为相同的时间间隔，并对同一间隔内的人群分布特征进行统计汇总；空间域则被分割成两两不相交空间区块，每个区块中的点对于整个该区块来说是等概率的。对于每个区域的空间坐标，如果区域内有单个基站则记录为该基站坐标，否则为基站坐标的平均值。

本文中我们重点研究移动网络中人群分布的时空统计模型。通常情况下，统计模型有两种不同的类型：一种是基于动态的物理过程描述，对统计过程产生的因素进行量化和关联，如附近地点对当前空间点的影响。这样的描述方法包含了底层过程的演化规律

和因果关系，通常建立在条件概率分布（Conditional probability distribution）之上，利用附近地点的过去观测值

$$\{Y(\mathbf{x}, r) : \mathbf{x} \in D_s, r \leq t\} \cup \{Y(\mathbf{x}; t) : \mathbf{x} \neq \mathbf{s}\} \quad (4-2)$$

对当前地点的值  $Y(\mathbf{s}, t)$  进行估计。但是条件概率分布的得出，需要对物理过程的不同阶段、以及不同部分有定量的描述，而这个条件在实际中往往难以满足。如在群体移动中，某工厂区域对商业区的人群移动，不仅和两地区的人口分布有关，而且受到交通网络、人们的生活需求等具体因素影响。对动态的物理过程建模，需要对不同因素、以及因素之间的影响进行建模，而这在仅已知宏观观测数据的前提下是难以获得的。另一种方法是在过物理过程的细节信息缺少的情况下，采用统计描述的方法，利用统计分布的一阶或二阶矩（如均值、方差、协方差等）特征对分布进行建模。与物理过程模型的条件概率分布不同的是，统计描述模型基于边缘概率分布（Marginal probability distributions）。在物理过程模型已知的条件下，我们也能够导出统计描述模型，因此统计描述模型具有更好的一般性和普适性。因此，在本文中，我们采用第二种方法，利用移动网络数据研究群体的时空分布特征，以及基于这些特征的统计描述模型。

#### 4.2.2 群体行为的时空相关性

在移动网络中，不同基站或建筑附近的人群分布随着时间推移不断发生着交互，因此在时间和空间上具有较强的相关性。我们用时空协方差函数（Spatio-Temporal Covariance Function, STCF）对不同时空点的人群观测进行描述。假设  $Y(\mathbf{s}, t)$  是一个在时间和空间上都具有二阶矩的稳定随机过程，其期望和方差分别满足  $E[Y(\mathbf{s}, t)] = \mu$  和  $Var[Y(\mathbf{s}, t)] = \sigma^2 < \infty$ ，给定两个时空点  $(\mathbf{s}_i, t_i)$  和  $(\mathbf{s}_j, t_j)$ ，他们的时空协方差函数表示为：

$$C(\mathbf{s}_i, \mathbf{s}_j; t_i, t_j) = Cov[Y(\mathbf{s}_i, t_i), Y(\mathbf{s}_j, t_j)] = C(\mathbf{h}, u) \quad (4-3)$$

其中  $\mathbf{h} = \mathbf{s}_j - \mathbf{s}_i$  和  $u = t_j - t_i$  分别表示空间和时间上的跨度（Lag）。时空协方差函数描述了一个时空点的观测值对另一个时空点的影响信息，且协方差值越大，表示二者之间的相关程度越强。给定零点的方差值  $\sigma^2 = C(\mathbf{0}, 0)$ ，4-3式的时空相关性函数表示为  $\rho(\mathbf{h}, u) = C(\mathbf{h}, u)/\sigma^2$ 。特别地，当空间或时间跨度为零时，则得到时间或空间维度上的

协方差函数。如  $C(\mathbf{0}, u)$  表示图4-2a 中单个地点的观测序列在时间维度上的相关程度，而  $C(\mathbf{h}, 0)$  描述了图4-2b 中单个时间片段上不同地点之间的相关程度。

从理论上来讲，时空协方差函数为半正定函数，反之亦然。一个函数的半正定性由下面的定义给出：

**定义 4.2. 半正定性：**对于定义在  $D \times D$  区间上的函数  $\{f(h, u) : h, u \in D\}$ ，其半正定性指对于任意复数  $\{a_i : i = 1, \dots, m\}$ ，变量  $\{u_i : i = 1, \dots, m\} \in D$ ，以及整数  $m$ ，满足

$$\sum_{i=1}^m \sum_{j=1}^m a_i \bar{a}_j f(h_i, u_j) \geq 0$$

其中  $\bar{a}_i$  为  $a_i$  的复共轭。

上述定义4-3式揭示了时空协方差函数的一个重要性质，即时空稳定性。这里我们给出时空稳定性的一般定义，即

**定义 4.3. 时空稳定性 (Spatio-Temporal Stationarity) :** 定义在实数区间  $\mathbb{R}^d \times \mathbb{R}$  上的函数  $f$  为稳定时空协方差函数的充分条件是：1) 函数  $f$  是半正定的，2) 满足关系  $f(\mathbf{s}_i, \mathbf{s}_j; t_i, t_j) = C(\mathbf{s}_i - \mathbf{s}_j, t_i - t_j)$ ，其中  $\mathbf{s}_i, \mathbf{s}_j \in \mathbb{R}^d, t_i, t_j \in \mathbb{R}$ 。

进一步，我们可以分别考虑空间和时间维度上的稳定性。如果时空协方差函数满足

$$f(\mathbf{s}_i, \mathbf{s}_j; t_i, t_j) = C(\mathbf{s}_i - \mathbf{s}_j; t_i, t_j) \quad (4-4)$$

则称函数  $f$  具有空间稳定性。类似地，如果时空协方差函数满足

$$f(\mathbf{s}_i, \mathbf{s}_j; t_i, t_j) = C(\mathbf{s}_i, \mathbf{s}_j; t_i - t_j) \quad (4-5)$$

则该函数称为具有时间稳定性。

在我们的观测维度里，时间往往以标量的形式出现，且标量值的增大意味着时间的增长。但是对于空间来说，即使在不考虑高度的情况下，我们依然需要两个坐标维度（如经纬度）对空间点进行测量，因此以矢量的形式出现。矢量意味着空间具有方向性（或各向异性），如4-3式中的空间跨度  $\mathbf{h}$ 。虽然矢量形式的空间跨度在理论分析中具有方便的优势，但是在实证研究中，由于不同空间尺度下用户移动的物理环境不同，导致

难以对群体的时空规律进行横向比较。基于这样的考虑，我们在群体移动行为中引入空间各向同性。如定义4.4所示，空间各项同性意味着人群分布的差异在不同观测方向上看遵循相同的规律或关系。

**定义 4.4. 空间各向同性 (Spatial Isotropy)**：定义在区间  $\mathbb{R}^d \times \mathbb{R}$  上的时空协方差函数  $f$  具有空间各向同性指满足

$$f(\mathbf{s}_i, \mathbf{s}_j; t_i, t_j) = C(||\mathbf{s}_i - \mathbf{s}_j||; t_i, t_j) \quad (4-6)$$

其中  $\mathbf{s}_i, \mathbf{s}_j \in \mathbb{R}^d, t_i, t_j \in \mathbb{R}$ 。

利用移动网络的观测数据，我们可以对经验的时空协方差值进行计算。假设随机过程  $Y(\mathbf{s}, t)$  的一阶矩仅与空间相关（即在时间上具有稳定性），且二阶矩仅与空间和时间的跨度有关，则经验的时空协方差通过下式计算：

$$\hat{C}(h, u) = \frac{1}{|N_h|} \frac{1}{|N_u|} \sum_{N_h} \sum_{N_u} (Y(\mathbf{s}_i, t_i) - \hat{\mu}(\mathbf{s}_i))(Y(\mathbf{s}_j, t_j) - \hat{\mu}(\mathbf{s}_j)), \quad (4-7)$$

其中  $\hat{\mu}(\mathbf{s}_i) = \frac{1}{T} \sum_{t=1}^T Y(\mathbf{s}_i, t)$ ， $N_h$  表示距离点  $(\mathbf{s}_i, t_i)$  的空间跨度为  $h$  的所有时空点  $(\mathbf{s}_j, t_j)$  的集合， $N_u$  表示距离点  $(\mathbf{s}_i, t_i)$  的时间跨度为  $u$  的所有时空点  $(\mathbf{s}_j, t_j)$  的集合。这里我们假设空间跨度  $\mathbf{h}$  与方向无关，则空间跨度向量可以用标量代替，即  $C(\mathbf{h}, u) = C(h, u)$ ，其中  $h \in \mathbb{R}$  表示空间跨度的欧式距离。类似的，经验的时空相关性定义为  $\hat{\rho}(h, u) = \hat{C}(h, u)/\hat{C}(0, 0)$ 。接下来，我们对移动网络中群体行为的时空分布特征、以及时空相关性进行实证分析研究。

### 4.3 多空间尺度下的时空分布特征

这部分我们基于多空间尺度下的移动网络数据，对人群移动和分布特点进行分析和量化。其中，WIFI-M 数据集记录了某大学校园内人群在各建筑和活动场所间的移动行为，CITY-M 数据集从城市尺度上记录了用户在各手机基站之间的转移行为，Senegal-S 和 Senegal-A 数据集在国家尺度上分别从手机基站和行政区规划粒度上记录了人群的移动行为。我们分别对人群的时空分布以及关联特征进行挖掘和分析。



0000294

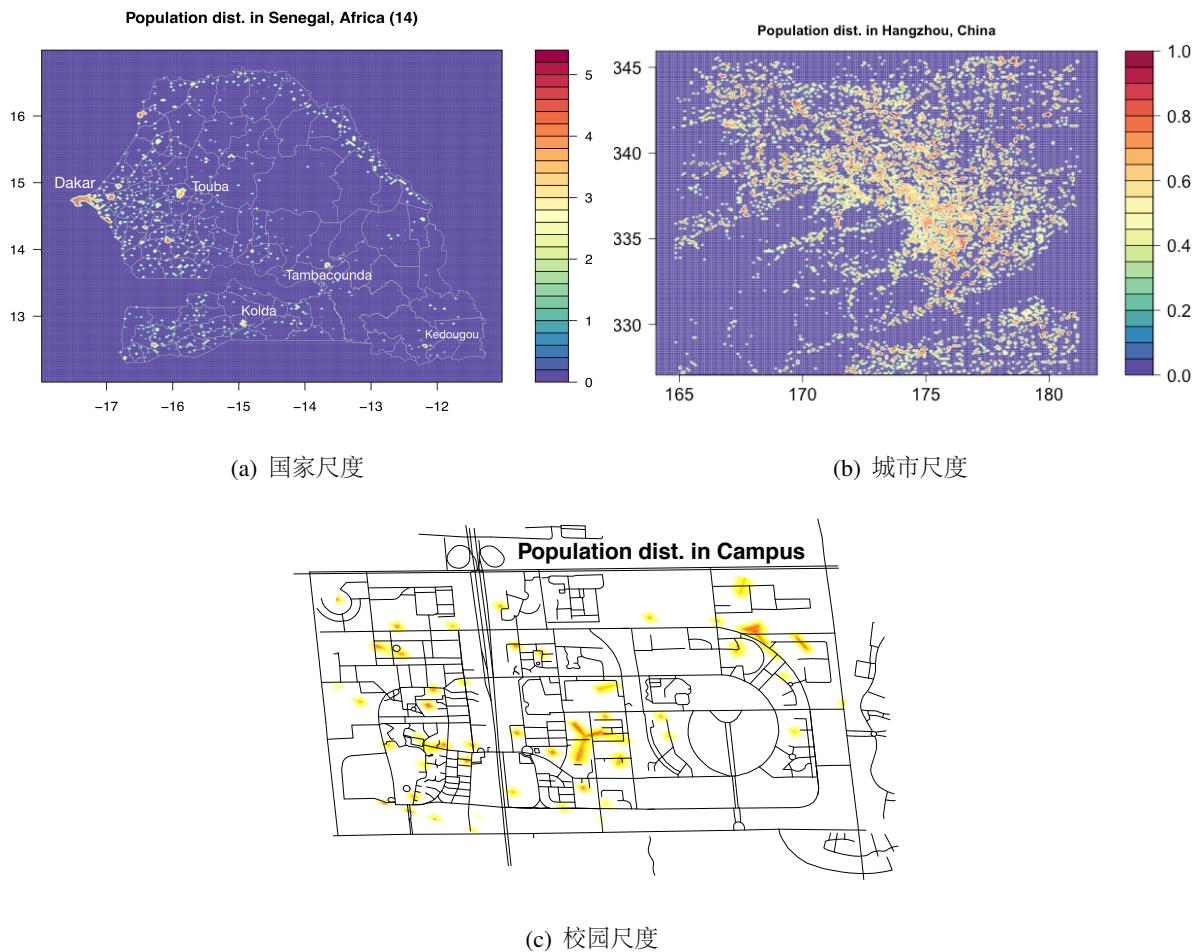


图 4-4 不同空间尺度下观测到的群体分布热力图（下午 2 点）

Fig 4-4 Heatmaps of population distributions observed at varying spatial scales (2:00 PM).

### 4.3.1 空间分布特征

在研究人群分布的时空相关性以前，我们首先对其空间特征，即4.1式中  $t = t_0$  时的分布  $Y(\mathbf{s}, t_0)$  进行分析。图4-4给出了不同空间尺度下人群分布的热力图，其中颜色越趋近于红色表示所在位置的人群密度越高。在国家尺度上（图4-4a），人群热点对应于塞内加尔的主要城镇区域以及道路网络，且人群密度和城市规模表现出较强的相关性，如人口数分别居于前两位的首都达喀尔（Dakar）和中部城市图巴（Touba）。由于人类个体的移动行为具有“莱维飞行”的特性，同时倾向于聚集在工作机会较多的经济发达区域，因此造成了空间大尺度上的较强的不均衡性。另一方面，便利的交通为人群大范围移动提供了可能，从而导致北部和中部的道路主干网周围拥有可观的人群分布。在城市

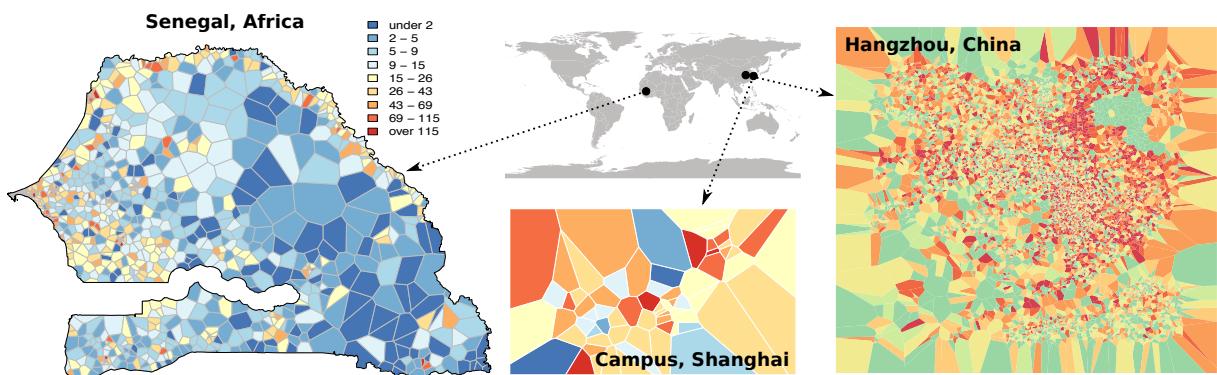


图 4-5 不同空间尺度下网络基站/热点粒度下的人群分布（下午 2 点）

Fig 4-5 Comparison of population distributions observed at base-station granularity (2:00 PM).

尺度上（图4-4b），人群分布的不均衡性较国家尺度上变小，这是由于城市范围内资源（如工作机会等）分布更加均衡，且发达的交通网络为个体长距离移动提供了便捷。但是，城市内不同功能区域之间依然有所差异，如图中坐标 [175, 355] 及周边区域为杭州市主要商业圈，而坐标 [170, 340] 及周边区域为住宅区之一；可以看出，城市内人群存在明显的“潮汐效应”，随着时间的推移，人群对不同地理区域的依赖程度也在发生着相应的变化，从而导致在不同功能区域之间的周期性迁移。在更小的校园尺度上，我们观测到用户群体在特定区域内的分布、以及与功能建筑的关系，从而形成对国家和城市尺度上观测的补充。通常大学校园内具有完善的生活设施和功能单位，从而在一定程度上包含了较完整的师生移动行为。如图4-4c 所示，在工作日高峰时段，校内的图书馆、教学楼、以及科研楼等聚集了大量的人群，这样的分布特点主要取决于个体的位置偏好，以及不同位置在个体生活中所扮演的功能角色。综上所述，群体移动的空间分布决定于地理空间上的资源分布、区域功能、以及群体构成上的差异，如国家尺度上的不均衡取决于资源分布，城市尺度上的分布主要受区域功能的影响，而校园尺度上的差异主要来自于不同的群体组成。尽管如此，不同尺度上观测的差异依然来源于个体的移动规律、以及与物理空间的相互作用。为了进一步对这种分布特征进行研究，我们接下来从网络基站粒度上对人群分布进行量化研究。

由于我们的数据采集点对应于离散的网络基站或热点，因此首先将连续的观测范围划分为不重叠的空间区域，如图4-5所示，单个区域内的点到该区域中心的距离最短。

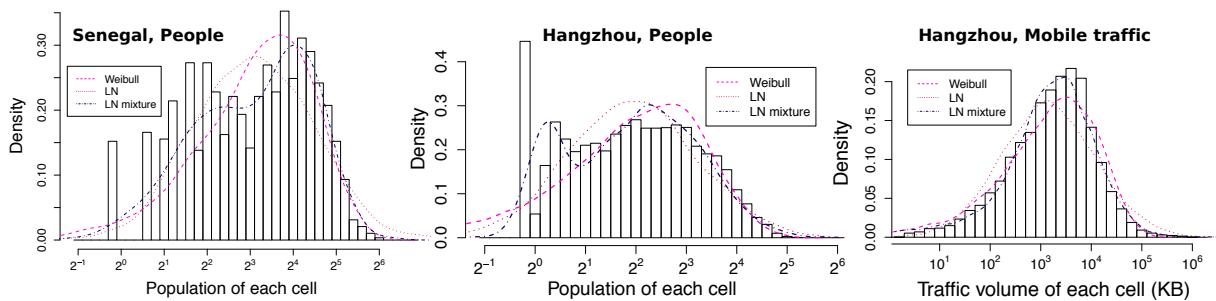


图 4-6 网络基站粒度下的人群和网络流量空间异质性对比（下午 2 点）

Fig 4-6 The heterogeneity comparison of population and mobile traffic distributions (2:00 PM)

一方面，该图从基站/热点粒度上验证了上述人群分布与地理空间因素之间的相互关系；另一方面，在不同的空间尺度下，可以看到高密度的人群分布聚集在较少数的基站/热点周围，而大多数区域的人数较少，即表现出较强的空间异质性 (Spatial Heterogeneity)。为了从量化角度捕捉这种异质性，我们对给定时间点的人群分布进行统计建模，这里考虑了三种不同的重尾分布关系，即

- 韦伯 (Weibull) 分布： $f(x; \lambda, k) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, x \geq 0;$
- 对数正态 (Log-normal) 分布： $f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/2\sigma^2}, x \geq 0;$
- 混合对数正态 (Log-normal mixture) 分布： $p(x) = \lambda_1 f(x; \mu_1, \sigma_1) + \lambda_2 f(x; \mu_2, \sigma_2),$   
其中  $\lambda_1 + \lambda_2 = 1, x \geq 0$ ，且  $f(x; \mu_1, \sigma_1)$  为参数  $\mu_1$  和  $\sigma_1$  的对数正态分布；

图4-6展示了人群在网络基站粒度下的经验分布以及上述统计拟合。可以看出，与韦伯和对数正态分布相比，混合分布更能准确捕获经验数据分布，这是由于人群具有明显的空间异质性，而这种异质性通过具有不同特征值的正态分布子模式体现出来，例如对于塞内加尔，不同子模式的特征值分别为  $y = 2^2$  和  $y = 2^4$ ，分别对应密度差异较大的两类基站群。另一方面，同样对于混合对数正态分布，国家和城市尺度下有明显的双峰特点，而校园尺度下较弱，这是因为人类移动性的限制，在较大空间范围内观测时人群分布的不均衡性较强，而对于小空间的校园环境，个体之间的交互增多，群体结构上的差异并不明显，因此表现出单模特征。此外，Michalopoulou 等<sup>[70]</sup> 预言在城市尺度上，出于对城市环境和网络业务的考虑，网络流量和人群分布的相关性并不显著。我们

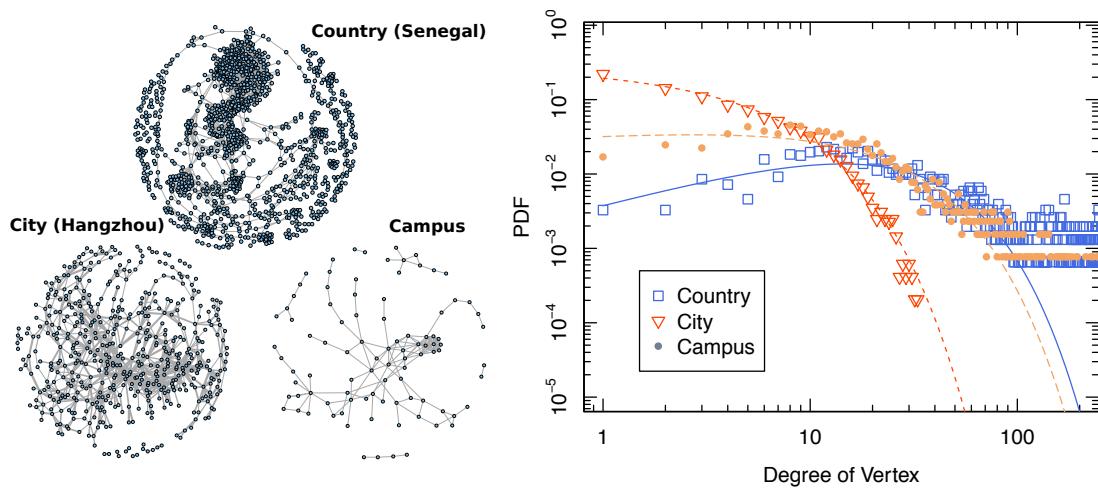


图 4-7 不同空间尺度下人群移动的空间结构对比

Fig 4-7 The comparison of mobility structures for group population at varying scales.

表 4-1 不同空间尺度下人群移动图节点度分布规律

Table 4-1 Parameter comparison of truncated power-laws over different spatial granularities.

TPL 参数	国家尺度	城市尺度	校园尺度
$a$	0.004	0.233	0.034
$\lambda$	-0.746	0.119	-0.150
$k$	19.3	5.57	18.1

同样对此假设进行实证分析。图4-6中比较了杭州市在基站粒度上的人群和网络流量分布，可以看出人群分布的双模特征比网络流量明显，即人群空间分布的异质性比网络流量要高。这主要是因为，在移动网络中，用户的上网行为受空间约束较小，即使拥有少数用户的基站，随着应用类型的的不同（如视频、音乐），也可能产生较高的流量比例，从而降低了不同空间点上流量分布的差异。

除了人群在空间上的静态分布特征以外，相同时段内的不同地点的人群交互（即动态性）在不同空间尺度下也有较大的差异。如图4-7所示，我们生成不同空间尺度下的人群移动结构图：其中每个节点表示不同的网络基站或热点，连接节点的边表示对应地点对之间存在人群交互，且交互量用边的粗细进行标示。从图左侧部分可以看出，在国家尺度上，人群移动在各大主要城市中形成了簇类结构，且最大的点簇表示塞内

加尔首都达喀尔，其他较小的独立点簇表示距离较远、人群交互较小的其他城市区域；在城市尺度上，图中大部分节点处于连通状态，表示城市内部区域之间的关联紧密，不同区域之间的人群交互较频繁；在校园尺度上，不同功能的建筑相互连通，其中中心度(Centrality)最高的节点为少数的公共建筑区域，如图书馆、校园餐厅、活动中心等。从统计角度来看，图4-7右侧展示了不同个空间尺度下的节点度分布满足截尾的幂律分布(Truncated Power-Law)规律：

$$D \sim a \cdot x^{-\lambda} e^{-\frac{x}{k}}, -1 \leq \lambda \leq 1. \quad (4-8)$$

其中  $\lambda$  为幂律分布参数， $k$  为截尾的指数分布参数。比较有趣的是，国家和校园尺度上的节点度分布较为接近，且二者与城市尺度上的差异较大。如表4-1所示，国家和校园尺度上幂律分布参数分别为  $\lambda = -0.746$  和  $\lambda = -0.150$ ，且截尾参数  $k$  以及调节因子  $a$  也较为接近。这意味着，国家尺度上的资源分布和校园尺度上的群体构成差异，对人群时空分布具有相似的影响，而城市尺度上的区域功能差异则表现出不同的特征。

最后，我们对不同空间尺度下人群分布的空间相关性进行分析。根据4-7式，观测的纯空间相关性为  $\hat{\rho}(h, 0) = \hat{C}(h, 0)/\hat{C}(0, 0)$ 。我们利用不同时间段内一个小时的观测数据计算空间相关性，其中最小空间跨度的变化量  $\Delta h$  根据观测尺度的不同而有所不同（国家尺度为 5km，城市尺度为 0.05km，校园尺度为 0.001km）。图4-8中展示了不同空间尺度下的人群移动空间相关性分布，在上午和下午我们分别选取了高峰期和非高峰期时段的两个小时进行对比研究。如图所示，在国家尺度上，空间相关性满足截尾的幂律分布(4-8式)；无论在上午和下午时段，高峰期和非高峰期都具有明显的差异，且不同时段内具有类似的空间相关性。例如，高峰期上、下午的幂律参数分别为  $\lambda_{AM}^{(p)} = 0.638$  和  $\lambda_{PM}^{(p)} = 0.821$ ，而非高峰期为  $\lambda_{AM}^{(np)} = 0.475$  和  $\lambda_{PM}^{(np)} = 0.421$ ；这表明和非高峰期相比，高峰期的空间相关性随空间跨度的增大而减弱较快。这是由于在高峰时段内，主要城市区域的群体移动性较强，造成与其他人口稀少区域的空间异质性增大，进而相关性随空间跨度的递减速度也增大。

城市尺度上则表现出与国家尺度不同的特征：首先，城市尺度的空间相关性满足幂律分布，即

$$\hat{\rho}(h, 0) \sim b \cdot h^{-\gamma} \quad (4-9)$$

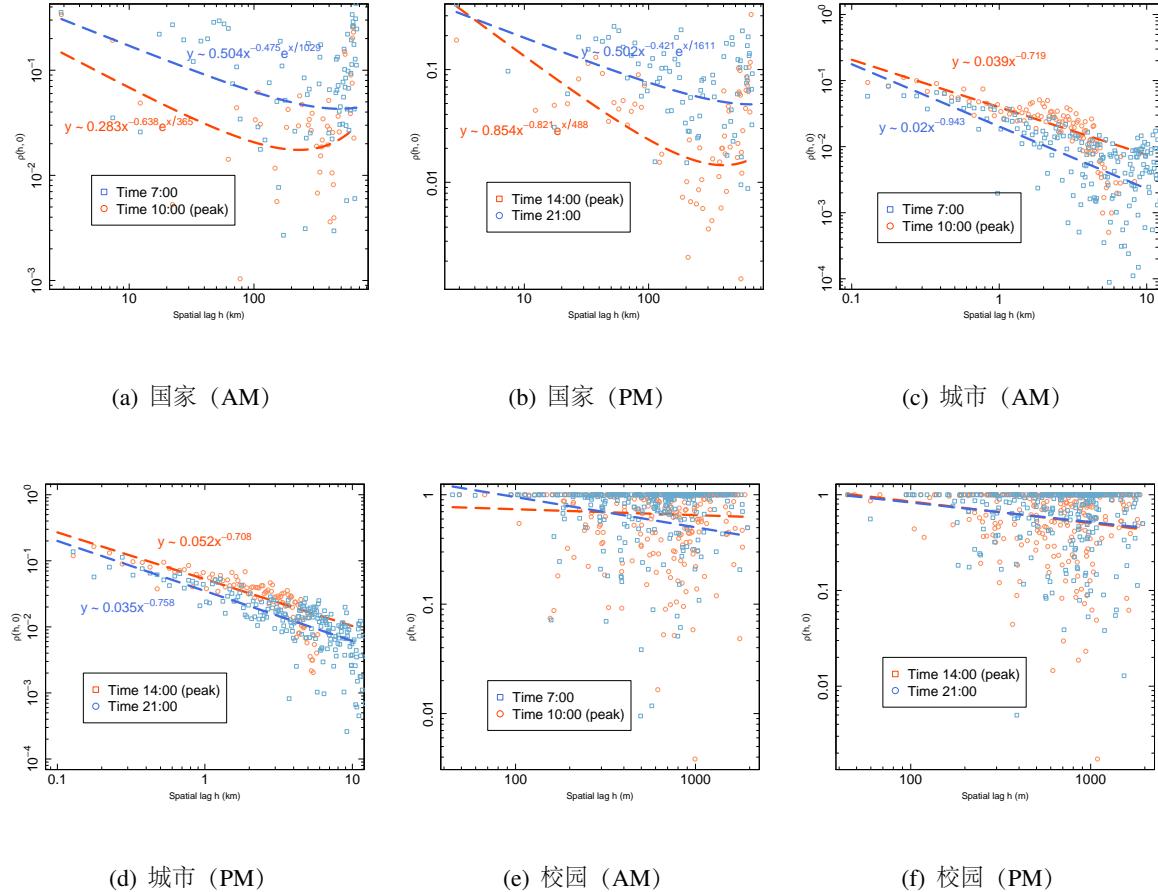


图 4-8 不同空间尺度下的人群移动的空间相关性分析 (杭州)

Fig 4-8 Spatial correlation analysis of population distribution at varying scales.

其中,  $\gamma$  为幂律指数; 这表明城市尺度的人群分布和国家尺度类似, 即具有空间长相关性 (Spatial long-range dependence)。对于上午时段, 高高峰期和非高峰期的幂律指数分别为  $\gamma_{AM}^{(p)} = 0.719$  和  $\gamma_{AM}^{(np)} = 0.943$ ; 而下午时段, 高高峰期和非高峰期其的幂律指数为  $\gamma_{PM}^{(p)} = 0.709$  和  $\gamma_{PM}^{(np)} = 0.758$ 。同时, 对于最大空间相关距离 (定义4.5), 高高峰期为  $h_X \approx 5\text{ km}$ , 而非高峰期为  $h_X \approx 10\text{ km}$ 。从上述观测可以看出, 和高峰期相比, 非高峰期的空间相关性下降速度较快, 且具有较强的长相关性。这样的差异和观测城市中居民的移动状态紧密联系在一起: 在城市尺度观测, 人们通常在高峰期向聚集度高的城区 (最远跨度约为 5km) 进行社交或商业活动, 而在非高峰期返回较分散的住宅区 (最远跨度约为 10~15km), 从而造成了不同时段最大空间相关距离以及幂律分布的差异。



0000294

**定义 4.5. 最大空间相关距离 (Maximum Correlative Distance, MCD)** 指特定的空间跨度  $h_X$ , 即当  $h > h_X$  时, 人群分布的空间相关性  $\hat{\rho}(h, 0)$  的减小速度比幂律分布快。

在校园尺度上, 人群分布在空间上的相关性不再具有幂律特征, 且高峰期和非高峰期不再有明显的异同点, 如图4-8e 和图4-8f 所示。综上所述, 在较大尺度上, 人群分布的空间相关性具有明显的幂律分布特征, 具体而言, 国家尺度满足截尾的幂律分布, 城市尺度满足幂律分布, 且高峰期和非高峰期在相关性下降速度和最大相关距离上有所不同, 而造成这种现象的主要原因在于大尺度上资源分布和区域功能的差异; 在较小尺度上, 如校园环境, 人群分布的空间相关性不再具有幂律分布, 产生这种特征的一种潜在原因是, 在较小空间尺度上, 人群的结构差异成为了决定空间异质性的主要因素, 而这些因素主要受人为规划的建筑分布的影响、比受人群动态移动的影响更大。

### 4.3.2 时间分布特征

这部分我们对不同尺度下人群分布的时间特征进行分析和研究。虽然人群分布具有空间异质性, 但是这种异质性随着时间的变化而发生改变, 且不同尺度上观测变化的规律有所不同。图4-9、4-10、4-11分别从城市、国家、校园尺度上对空间异质性的时变特征进行了展示。首先在图4-9a 中, 我们测量了人群分布的经验观测与拟合的三种不同的重尾分布之间的 K-S 距离, 及其随时间的变化关系。可以看出在所有时间段中, 混合对数正态分布都具有最小的 K-S 距离, 即该分布能够较好地捕捉底层的人群空间分布特征。在时间方向上, 韦伯和对数正态分布在白天 (7:00~16:00) 的拟合误差较大, 而晚上的误差较小; 这是由白天人群移动的动态性更高、空间分布的异质性更强导致的。虽然混合对数正态分布的拟合误差随时间的变化幅度较小, 但其在夜晚的性能有所降低, 并处在与对数正态分布相当的拟合误差范围内。

混合对数正态分布  $p(x) = \lambda_1 f(x; \mu_1, \sigma_1) + \lambda_2 f(x; \mu_2, \sigma_2)$  通过不同的模式捕捉具有差异较大的两组人群分布的空间特征, 因此表现出较好的平均拟合性能。我们对其两组组成模式的特征参数进行分析, 即对数正态分布的期望  $\mu$  和方差  $\sigma$ 、以及调和因子  $\lambda$ 。图4-9b 展示了期望  $\mu$  (准确地讲, 该期望表示人群分布的特征值为  $y = 2^\mu$ ) 随时间的变化关系。可见在城市尺度上存在两种不同的人群空间分布模式, 其中模式 1 包括聚集程度较高的城市区域, 如高峰期的商业区; 而模式 2 包括其余人群聚集度小的区域,

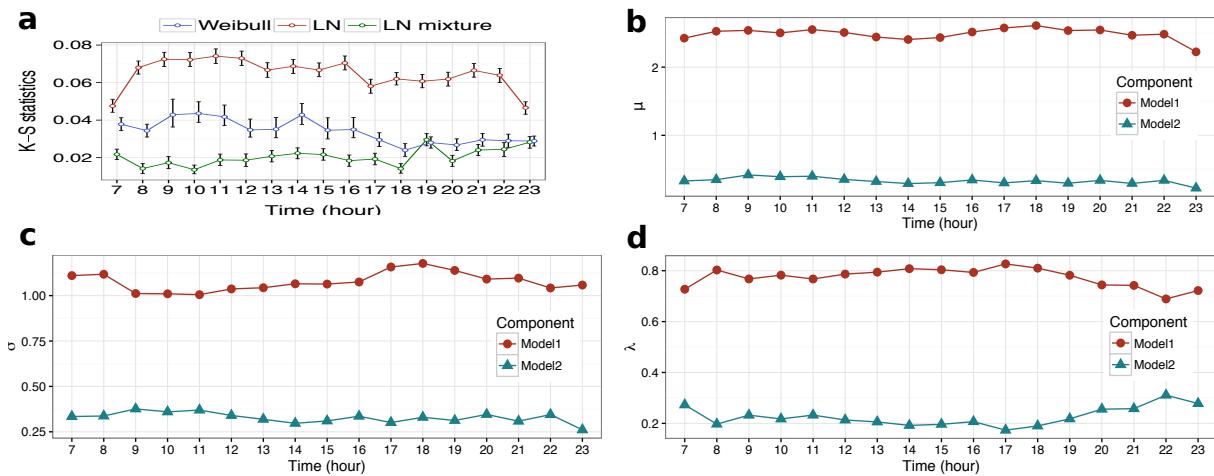


图 4-9 城市尺度下人群分布的双模模型参数与时间的关系（杭州）

Fig 4-9 The temporal tendency of model parameters for population distributions at city scale.

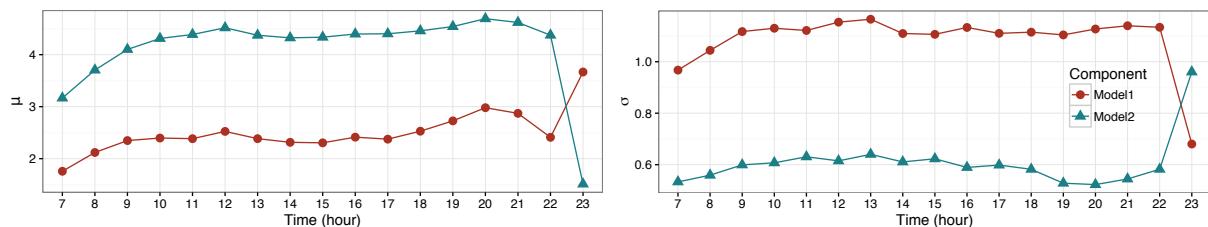


图 4-10 国家尺度下人群分布的双模模型参数与时间的关系（塞内加尔）

Fig 4-10 The temporal tendency of model parameters for population distributions at nation scale.

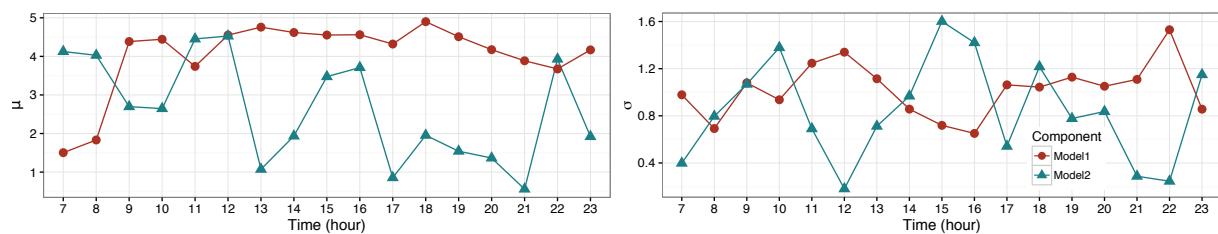


图 4-11 校园尺度下人群分布的双模模型参数与时间的关系

Fig 4-11 The temporal tendency of model parameters for population distributions at campus scale.



0000294

如高峰期的住宅区。由于模式 1 中包含的区域人群移动性强，因此具有较大的方差（如图4-9c 所示）。更进一步，我们通过量化的方法测量出模式 1 的期望和方差特征值分别为  $\mu_1 = 2^{2.1} \sim 2^{2.7}$  和  $\sigma_1 = 2^{0.9} \sim 2^{1.25}$ ，模式 2 的期望和方差特征值为  $\mu_2 = 2^{0.2} \sim 2^{0.5}$  和  $\sigma_2 = 2^{0.25} \sim 2^{0.4}$ 。图 4-9d 从概率角度展示了不同成分在经验观测中的比重，可以看出在所有观测时段内，模式 1 覆盖了更大的观测范围（参考图4-6）。

接下来，我们对国家和校园尺度上的人群分布模式进行了对比研究。图4-10首先展示了国家尺度上人群分布模式的期望和方差分布。可以看出，与城市观测相同的是，国家尺度上由两种差异较大的模式构成；与城市观测不同的是，这里模式 1 虽然具有较大的方差，但是期望值比模式 2 小，代表聚集性较小的人群分布区域。具体而言，模式 1 的期望和方差特征值分别为  $\mu_1 = 2^{1.5} \sim 2^3$  和  $\sigma_1 = 2^{0.96} \sim 2^{1.18}$ ，模式 2 的期望和方差特征值为  $\mu_2 = 2^3 \sim 2^5$  和  $\sigma_2 = 2^{0.51} \sim 2^{0.65}$ 。这表明国家与城市尺度观测的最显著差异在于，人群聚集度较高的区域动态变化范围反而相对较小，其主要原因在于大尺度上观测时，人群移动的“莱维飞行”特性更加突出，倾向于在特定城市范围内活动，而以非常小的概率进行长途的城际旅行。图4-11展示了校园尺度上人群分布的模式特征。可以看出，与大空间尺度（即城市和国家）不同，人群分布的模式特征不再有明显的聚类特征，代之以不同程度的交叠。例如在上午 8:00~10:00 之间，人群分布模式具有相近的方差和不同的期望，而在中午 11:00~12:00 具有相近的期望和不同的方差。这表明校园尺度下人群分布的空间异质性较弱，群体结构上的差异并不明显，更加倾向于单模特征。从时间维度上看，随着时间的变化，模式 1 的期望值呈递增趋势，而模式 2 呈递减趋势。这表明晚上比白天的空间异质性强，这是由于白天人群的移动性较强，在校园内各区域之间的活动较频繁，而晚上对地点的偏好选择更强，因此两种分布模式也更加明显。

除了不同时间片段内的人群空间分布的差异以外，特定地点或区域在时间上的相关性也具有明显的特征。这里我们对城市尺度下的不同功能区域和基站观测进行研究。首先，图4-12a 展示了我们选择的城市中三个不同的空间区域：从功能上讲，区域 A 覆盖杭州市的主要商业区域，包括大型商场和写字楼建筑。区域 B 和区域 C 分别表示不同地理位置的住宅区域；可见杭州市是一个单中心结构的城市，住宅区域主要分布在商业区的周边地区。图4-12b 和图4-12c 分别从基站粒度上展示了区域 A 和区域 B 在 14:00 的人群分布特征，其中颜色越接近红色表示人群密度越大；可以看出商业区域的基站密

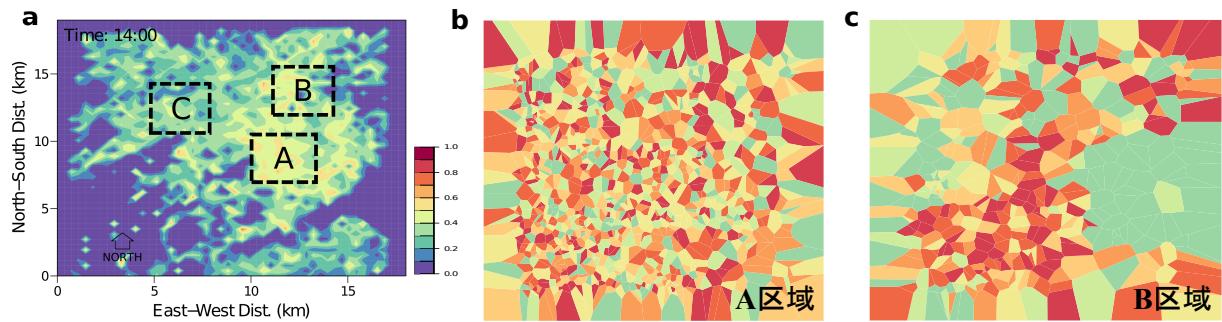
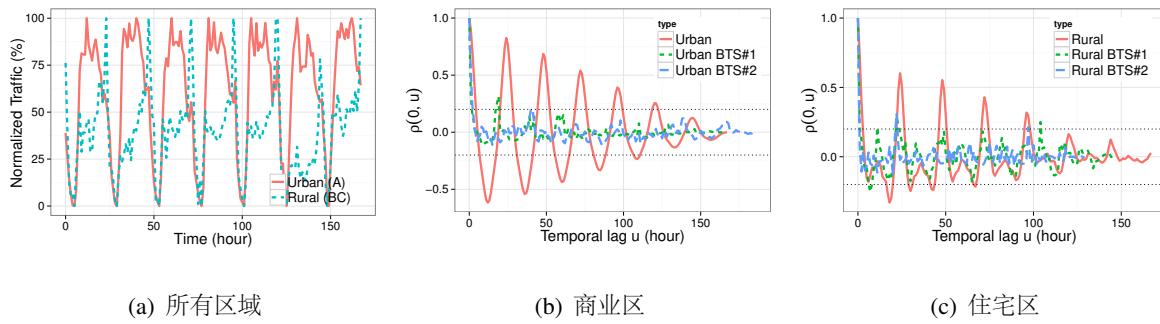


图 4-12 城市尺度下不同功能区域的空间结构示意图（杭州）

Fig 4-12 Visualization of spatial structures of different regions at city scale.



(a) 所有区域

(b) 商业区

(c) 住宅区

图 4-13 城市尺度下不同功能区的人群移动趋势和时间相关性对比（杭州）

Fig 4-13 The comparison of population tendencies and temporal correlation at varying areas.

度远高于住宅区域。

对于给定的区域或基站，我们对其人群分布的纯时间相关性进行计算，根据4-7式，时间相关性计算为  $\hat{\rho}(0, u) = \hat{C}(0, u)/\hat{C}(0, 0)$ 。图4-13a首先展示了不同区域在一周内的时间观测序列，可以得出两个直接结论：一是从短时间范围来看，验证了城市内人群分布的“潮汐效应”，例如在晚上商业区域和住宅区域的人群分布形成了明显的互补；二是在长时间范围内，不同区域均具有和人类生活习惯一致的周期性，且周末时段住宅区域的人群聚集明显高于工作日。图4-13b 和4-13c 分别给出了商业和住宅区域的时间相关系数  $\hat{\rho}(0, u)$  随时间的变化曲线。从总体上来看，人群分布具有较强的时间长相关性（Temporal long-range dependence）以及人类生活节律带来的周期模式。商业区和住宅区的时间相关性，在时间跨度  $u = 12h$  处表现出不同的特征（如  $\hat{\rho}_A(0, u) = -0.6$ ，



0000294

$\hat{\rho}_B(0, u) = 0$ ），而在时间跨度  $u = 24h$  处却具有相似的特征（如  $\hat{\rho}_A(0, u) = 0.76$ ,  $\hat{\rho}_B(0, u) = 0.6$ ）。这主要源自于超吸效应下人群在不同区域的移动特点，例如在我们的观测中，商业区在每天中下午时段达到人数高峰期，而在上午和晚上人数最少；住宅区却在一天的时间段内呈现出递增的趋势。但是，从基站粒度上观测，人群分布的时间相关性并不明显，这主要由于我们的观测基于移动网络数据，而用户使用手机的行为往往具有即席收发的特点。

综上所述，由于城市内不同区域的功能定位以及长期观测中的人群移动性差异，人群分布的时间相关性和区域类型有着紧密的联系。这启发我们在对人群时空分布的建模中，不仅要考虑其在空间和时间上的分布特点，还需要考虑不同功能区域之间的差异。下一节中，我们利用本节观测到的时间和空间分布特点，采用统计建模的方法对人群分布进行理论和实证分析。

## 4.4 群体行为的时空关联建模

虽然个体的时空行为模型得到了广泛的关注，但是宏观尺度下的群体时空统计模型依然未能得到深入的研究。时空统计分析针对时间和空间维度上多变量的观测序列，在环境和行为科学领域得到了较多应用。我们的目标是对4-7式所得的经验分布建立时空统计模型，从而能够重现移动网络中群体分布在时空上的变化过程（如图4-2）。时空统计模型由于方便融合时间和空间上的边缘分布信息，因此为理解群体移动的时空过程提供了有效途径。这部分我们利用时空统计的方法对群体时空分布进行建模，和传统的时间序列模型相比，所提出的新模型融合了时空关联性特征；然后我们通过群体分布的预测对模型性能进行了验证，并对比了不同时段和区域功能等因素对预测性能的影响。

### 4.4.1 无时空关联的行为模型

在引入时空关联信息之前，我们首先介绍并分析一种简单的行为模型。该模型中，时间和空间上的信息采用相乘或相加的方式相互作用，在相关文献中也称作时空分离模型（Separable Models）<sup>[105]</sup>。虽然这类模型具有有限的时空关联信息，模型结果往往与实际差异较大，但是该类模型构成了我们分析时空相关性的基础。通过比较，我们可以从量化的角度，衡量时空相关性对群体移动预测性能的提高程度。

一般来讲，由于相乘（ $\times$ ）和相加（ $+$ ）包含了相同的时空信息量，这里我们采用前者进行分析，即时空协方差函数表示为

$$C(\mathbf{h}, u) = \sigma^2 \rho(\mathbf{h}) \rho(u), \quad (4-10)$$

其中  $\rho(\mathbf{h})$  和  $\rho(u)$  分别表示空间和时间维度上的协相关函数，其隐含条件为  $C(\mathbf{s}_i, \mathbf{s}_j; t_i, t_j)$  分别具有空间和时间稳定性。由于  $\sigma^2 = C(\mathbf{0}, 0)$ ，4-10式只需对相应的相关函数建模。如定义4.2可知，时空协方差函数的充分必要条件是满足半正定性，因此所选择的时间和空间的相关函数乘积也需满足该条件。根据4.3节所得的空间和时间特征，我们考虑两类带参数的模型：

- 指数模型 (Exponential Mode):  $\rho(x) = e^{-x}$ , 其中参数满足  $x \geq 0$ ;
- 柯西模型 (Cauchy Model):  $\rho(x) = (1 + x^\alpha)^{-\frac{\beta}{\alpha}}$ , 其中  $\alpha \in (0, 2]$ ,  $\beta > 0$ , 且  $\geq 0$ 。

无时空关联的模型由于形式简单，且纯空间的矩阵处理或时间序列处理都较为方便，因此在以往的研究中经常被采用<sup>[105]</sup>。但是这类模型的缺陷也很明显，即缺少在经验数据中可直接观测到的时空依赖性信息。如4-10式所示，乘数因子  $\rho(\mathbf{h})$  意味着  $\rho(\mathbf{h}; u_1) \propto \rho(\mathbf{h}; u_2)$ ，即不同时间片段里的空间相关性满足相同的性质和规律；同理  $\rho(u)$  意味着  $\rho(\mathbf{h}_1; u) \propto \rho(\mathbf{h}_2; u)$ ，即不同地点的时间序列具有相似的自相关特征。而这样假设的不足可以从 4.3 节的时空相关性分布中反映出来。

#### 4.4.2 融合时空关联的行为模型

这部分我们基于时空关联信息对群体行为分布进行建模。建模方法不同，时空关联信息的融合形式也有所差异。在物理过程模型中，时空关联信息往往体现在群体移动的过程描述当中，例如，对于给定地点  $\mathbf{s}_i$ ，不同地点  $\mathbf{s}_j$  的关联强度不同，因此能够以较为清晰的规则形式添加在模型当中。但是在时空统计模型当中，其核心是对时空协方差函数进行统计建模，而挑战在于对于模型半正定性的验证。本研究基于 Gneiting 范式<sup>[106]</sup> 和实证分析的结果，提出了一种描述群体分布的时空行为模型。该模型通过对移动网络中用户移动的数据分析，得出时间特征以及时空交互特征的统计规律，进而利用 Gneiting 范式在保证模型半定性的前提下，将时间和空间特征进行关联融合。

对于时空协方差函数的半正定性，Cressie 等<sup>[105]</sup>提出利用傅里叶变换后的频域特征进行判断。这种方法由于依赖傅里叶变换，因此仅对极少数的积分收敛的函数有效。但是在实际分析中，人群分布的时空分布难以满足这样严苛的收敛条件。针对这样的不足，我们采用 Gneiting 范式对半正定性进行判断，由于避免了傅里叶分解，从而将函数范围扩大至更多的带参数的分布函数，也构成了我们对群体行为建模的基础。

我们首先回顾函数单调性判断。假设定义在  $t \geq 0$  上的连续函数  $\phi(t)$ ，其为严格单调函数的条件为  $n$  阶导数满足

$$(-1)^n \phi^{(n)}(t) \geq 0 \quad (4-11)$$

其中  $t > 0, n = 0, 1, 2, \dots$ 。则 Gneiting 范式描述为：假设  $\phi(t), t \geq 0$  是严格的单调函数， $\psi(t), t \geq 0$  为正定函数且其积分为严格单调函数，则函数

$$\rho(\mathbf{h}, u) = \frac{1}{\psi(u^2)^{d/2}} \phi\left(\frac{\|\mathbf{h}\|^2}{\psi(u^2)}\right) \quad (4-12)$$

是有效的时空协方差函数，其中  $d \geq 0$ ，且  $(\mathbf{h}, u) \in \mathbb{R}^d \times \mathbb{R}$ 。

在4-13式中，函数  $\psi(x)$  对时间分布特征进行建模，而函数  $\phi(x)$  对时间和空间的关联关系进行描述。对于人群时空分布来讲，时间和空间交互具有特定的分布规律；为了尽可能多地捕获人群特征，我们根据实证观测提出以下特征函数：

- **时间特征：**  $\psi(x) = (1 + x^\alpha)^\beta, \alpha, \beta \in (0, 1]$ ；
- **时空关联特征：**  $\phi(x) = \exp(-x^\lambda), \lambda \in (0, 1]$ 。

将上述两个特征函数代入4-13式中得到群体分布的时空协方差函数为：

$$\rho(\mathbf{h}, u) = \frac{1}{(1 + u^{2\alpha})^{\frac{d\beta}{2}}} \exp\left(-\frac{\|\mathbf{h}\|^{2\lambda}}{(1 + u^{2\alpha})^{\beta\lambda}}\right) \quad (4-13)$$

#### 4.4.3 模型性能验证

前面两节中，我们利用移动数据分析的时空特征，提出了无时空关联的模型（简称为 ExpModel 和 CauchyModel）和融合时空交互信息的模型（简称为 ST-Model），其中前者作为比较和分析的基准。对于每一个带参数的候选模型，我们采用期望最大

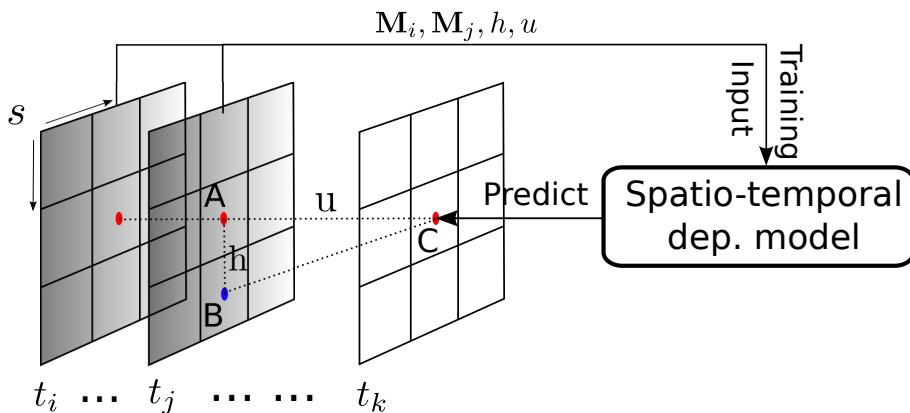


图 4-14 利用时空关联特征进行人群分布预测的模型示意图

Fig 4-14 A brief illustration of population prediction machine with spatio-temporal dependence models.

(Expectation–maximization) 算法对模型参数进行估计，即最大化相应的对数似然函数  $\mathcal{L}(\Theta|\mathbf{O}) = p(\mathbf{O}|\Theta)$  其中  $\Theta$  表示模型的参数集合  $\mathbf{O}$  表示数据集中的时空观测值。

接下来，我们通过时空分布预测的方法对模型的性能进行量化和比较。由于群体行为的时空协方差函数包含时空关联信息，因此预测模型性能的高低可以用来衡量时空协方差函数与实际观测的相符程度。图4-14展示了利用时空协方差函数的预测器。具体来讲，我们将观测区域划分为不重叠的块，并利用矩阵  $\mathbf{M}_i$  记录在时刻  $t_i$  时的人群空间分布观测。随着时间推移，多个这样的观测矩阵形成了矩阵序列。在模型训练阶段，我们利用观测序列  $\langle M_i, \dots, M_j \rangle$  对模型参数进行估计。在预测阶段，我们利用观测历史中不同时空点（如点 A 和 B）的线性组合对未来时空点（如点 C）进行预测，即

$$y(s', t') = \mu(s', t') + \sum_{i=1}^N w_i \cdot (y(s_i, t_i) - \mu(s_i, t_i)) \quad (4-14)$$

其中  $s \in \mathbb{Z}_N$ ,  $t \in \mathbb{Z}_T$ , 期望函数  $\mu(s, t) = E(Y(s, t))$ , 不同历史观测点的权重  $w_1, \dots, w_N$  由  $Y(s, t)$  的时空协方差函数决定，这里我们取权重为协相关值。在时空稳定性的前提下，我们进而获得预测方差为

$$\sigma(s', t') = \sum_{i=1}^N \sum_{j=1}^N w_i w_j C(h, u) \quad (4-15)$$

其中  $h = s_i - s_j$  和  $u = t_i - t_j$  分别为空间和时间跨度。

表 4-2 不同模型在考虑区域和时段差异下的预测性能对比

Table 4-2 The RMSE statistics for prediction performance of different models.

Day	Rural Area			Urban Area		
	ExpModel	CauchyModel	ST-Model	ExpModel	CauchyModel	ST-Model
TUE	0.182	<b>0.165</b>	0.181	0.180	0.198	<b>0.166</b>
WED	0.208	0.158	<b>0.156</b>	<b>0.186</b>	0.192	0.198
THU	0.175	0.150	<b>0.147</b>	0.187	0.185	<b>0.151</b>
FRI	0.213	0.206	<b>0.154</b>	0.180	0.193	<b>0.137</b>
SAT	<b>0.205</b>	0.231	0.207	0.238	0.246	<b>0.237</b>
SUN	0.208	<b>0.152</b>	0.157	0.224	0.225	<b>0.172</b>

利用上述模型，我们同时在时间和空间域上对群体行为的分布进行了预测。对于模型在时间段内的预测值，我们利用均方误差（RMSE）对预测性能进行评估，即

$$RMSE = \left[ \frac{1}{NT} \sum_s \sum_t (\hat{y}(s, t) - y(s, t))^2 \right]^{\frac{1}{2}} \quad (4-16)$$

其中  $\hat{y}(s, t)$  和  $y(s, t)$  分别表示在时空点  $(s, t)$  上的预测值和经验观测值。均方误差越小，意味着预测模型性能越好，因而相应的群体分布模型与实际则越接近。

#### 4.4.4 数据分析及结果

这部分利用 CITY-M 数据集对群体时空分布模型的性能进行验证。为了区分空间因素对预测准确度的影响，我们对城市内不同功能区（即商业区和住宅区）进行了分析；同时为了考察时间因素的影响，我们对一周内的不同时间天进行了独立分析，从而获得了时间维度上时空关联信息对群体分布建模和预测性能的影响。

我们首先给出了在一周观测数据（2012.08.19~2012.08.25）的基础上，所获得的群体时空分布模型的预测结果，以及区域和时间因素对预测性能的影响。对于每种模型，我们采用期望最大算法对模型参数进行估计，而较大的时间和空间跨度带来了参数估计过程中的计算开销。依据4.3节中的观测，由于高峰时段人们移动的特征距离约为 5 公里，因此我们将模型可见的最大空间跨度为  $h_{max} = 5km$ ，同理，由于人们移动行为的时间相关性的震荡周期约为 12 小时，因此我们选择半周期作为模型可见的最大时间跨度，即  $u_{max} = 6h$ 。这里模型可见指在给定的时间和空间跨度范围内，模型能够利用

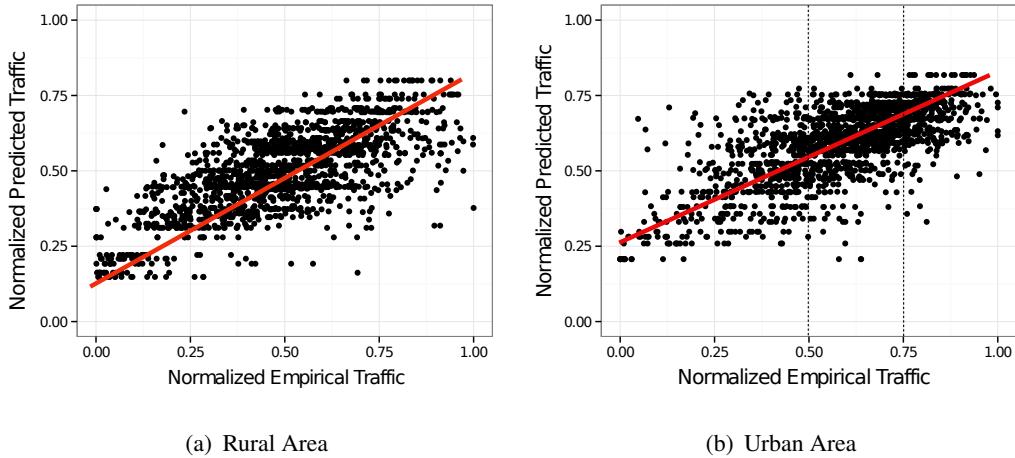


图 4-15 基于时空关联的 ST-Model 人群预测值与观测值的对比

Fig 4–15 Comparing population prediction with empirical observations (e.g., Friday).

历史的信息对未来进行准确的预测；否则，由于模型未包含范围之外的时空关联信息而无法得到有效预测。在模型训练阶段，我们利用过去 24 小时的历史数据对模型进行学习，然后利用得到的模型对未来 6 小时的群体分布进行预测。由于试验中模型需要冷启动，因此我们最开始的训练阶段基于周一的观测数据完成；接着重复上述步骤，直到余下的天数都被完整预测。同时，我们选取了观测城市中不同类型的区域，其中商业区范围为  $p_1 = [120.167722, 30.255391]$  和  $p_2 = [120.197546, 30.281592]$  确定的矩形区域，住宅区由  $p_1 = [120.16678, 30.300872]$  和  $p_2 = [120.200125, 30.336536]$  确定，且两者的边长距离约为  $4km \sim 5km$ 。在空间上，我们对每个区域分割成  $10 \times 10$  的网格，平均每个网格的面积与单个基站的覆盖范围相当。从而我们得到 4-16 式中的时间和空间索引分别为  $N = 100$  和  $T = 144$ 。

表4-2展示了利用上述实验环境的 RMSE 数据。可以看出，在住宅区域，融合了时空关联信息的行为模型（即 ST-Model）在星期三、四、五表现出较好的性能，且比无时空关联的模型（即 ExpModel 和 CauchyModel）性能提高了约 2.8%~25.2%；但是在周二和周末三天，后者表现出稍微的性能优势。在商业区域内，融合了时空关联信息的模型在大多数观测时间内都取得了较好的预测性能，且预测准确度提高了约 3.7%~23.6%。由此可见，时间和空间的关联信息在预测群体时空分布行为上起着重要的作用，从而也表明研究群体行为的时空依赖关系具有重要的价值。

为了解释融合时空关联信息的行为模型的细节特征，我们在图4-15中展示了在周四的24小时内，模型的预测值和实际观测值的散点分布（注：为了对比的需要，我们将不同区域的观测量值进行了正规化）。在这组散点图中，越接近于对角线的点表示预测得越准确，其中红色直线表示数据点的线性回归线。这里我们可以得出两个显著的结论：1) 融合了时空依赖信息的行为模型倾向于产生比较平滑的预测结果，这意味着，和实际观测相比，低值范围倾向于获得较大的预测结果，而高值范围则较容易获得较小的预测值。这个特点来自于行为模型本身，如4-14式所示，由于预测结果由历史时空点的线性组合而成，因此和一维的线性回归类似，表现出一定的平滑作用；且时空相关性越强，平滑效果越明显。2) 预测区域的功能差异对模型的预测性能具有一定的影响，如商业区较住宅区的RMSE数据小，即0.137 vs. 0.154。这是由于不同功能的区域，移动网络基站根据该区域人们的生活习惯部署方式也不同，进而表现在不同观测点的人群分布差异较大；而这样的差异与结论1中的模型预测机制相互作用，便出现结果的不同。例如，在4-15b中，由于商业区域的模型在0.5~0.75范围内预测性能较好（即预测和观测值接近），而该区域的多数观测点处于该范围内，因此其RMSE数值比住宅区域小。

另一方面，ST-Model的性能优势体现在对时空关联特征有效利用的基础上，而模型的机制不同，模型训练后所得的时空相关性特点也因此不同。这里我们展示了ExpModel、CauchyModel和ST-Model在考虑区域差异下（时段为THU），不同模型所捕获的时空关联性的差异。如图4-16所示，在每对特征组合下（如ExpModel-Rural），三个子图分别表示时空协方差函数 $C(h, u)$ ，纯时间协方差分布 $C(0, u)$ ，以及纯空间协方差分布 $C(h, 0)$ 。一般而言，曲线（面）的曲率越大，所在时间或空间跨度的相关程度越强。可以看出，在住宅区域中，ExpModel的时间和空间跨度特征值（即曲率最大的点）分别为 $u = 2h$ 和 $h = 1km$ ；CauchyModel的特征值分别为 $u = 1h$ 和 $h = 0.7km$ ；而ST-Model的特征值介于两者之间，且倾向于ExpModel的特征，即 $u = 1.5h$ 和 $h = 0.9km$ 。在大于特征值的区域，时空相关性逐步递减，且递减速度 $v(\text{CauchyModel}) > v(\text{ST-Model}) > v(\text{ExpModel})$ 。然而，在商业区中，时空相关性表现出与住宅区不同的特征，其中ExpModel的时间特征曲率较均匀，因此具有较大的时间特征值 $u = 6h$ ，且空间特征值为 $h = 0.8km$ ；CauchyMode的时间和空间特征值分别为 $u = 2h$ 和 $h = 0.7km$ ；而该区域中ST-Model与前两者的差异较大，其时间特征值在三类模型中最小，即 $u = 1h$ ，且空间特征值为

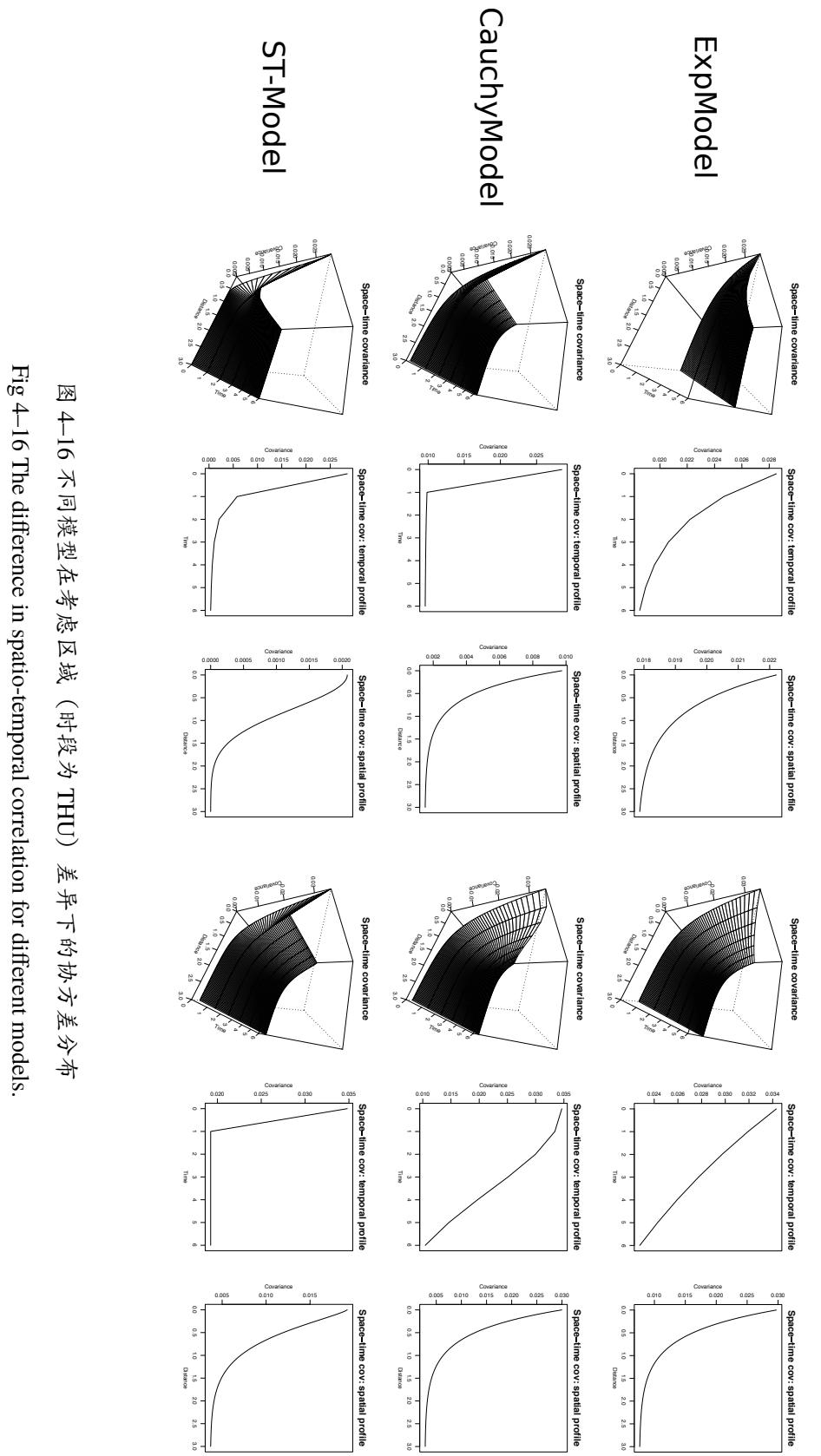


图 4-16 不同模型在考虑区域（时段为 THU）差异下的协方差分布

Fig 4-16 The difference in spatio-temporal correlation for different models.



0000294

$h = 0.8km$ 。在大于特征值的区域，时空相关性的递减趋势也表现出与住宅区的差异，即  $v(ST\text{-Model}) > v(Cauchy\text{Model}) > v(Exp\text{Model})$ 。和表4-2进行对比分析可以发现，在住宅区域，由于 CauchyModel 和 ST-Model 的时空相关性趋势较为接近，因此其预测性能也相当，分别为  $CM_{rmse} = 0.150$  和  $STM_{rmse} = 0.147$ ；同理，在商业区中，由于 ExpModel 和 CauchyModel 均与 ST-Model 的时空相关性趋势差异较大，因此前两者的预测性能也远低于融合了时空相关性的行为模型。

综上所述，利用时空统计模型对群体行为的分布特征进行研究，其准确度主要体现在两个方面：一是时间和空间的协方差函数能够捕获实际观测数据的特征，二是时空相关性信息能够被模型有效地利用。虽然 ExpModel 和 CauchyModel 利用单纯的时间或空间相关性具有一定的预测能力，且在某些取值范围内具有较好的性能优势，但是由于 ST-Model 在以上两个条件都得到了满足，因此总体上性能优势明显。

## 4.5 本章小结

本章基于第三章中介观模式所揭示的个体移动的时空依赖性，进而对群体移动的时空分布特点进行实证及建模研究。与传统的时间序列分析不同，本研究在把握不同尺度（校园、城市、国家）下人群分布特点的基础上，找出时间和空间的交互关系，进而对时空相关性进行统一建模。理论上来讲，这样的建模方法符合人群移动的内在规律，即同时包括空间上的网络效应和时间上的时律性。利用多空间尺度的观测数据，我们对探究人群时空分布的特点提供了一个新的视角。从空间角度来看，空间异质性在较大空间尺度（如城市和国家）下普遍存在，从而突破了以往研究和模型中对人群均匀分布的假设，例如，可以帮助移动网络网络设计人员避免网络资源的浪费，以及降低网络级联故障的发生概率。从时间角度来看，了解人群在基站和区域范围内的动态变化，也有助于研究人员开发考虑时间相关效应的预判模块（Proactive modules），从而提前对网络拥塞进行调节和控制。

同时，本研究的一些方法和结论可以在未来的工作中得到进一步发展和探索。其中一项有趣的扩展工作是将我们研究的时空统计特征和个体移动行为的语义特征（如交通方式、出行目的等）进行关联分析，从而将人群分布的空间异质性、时空依赖性与个体移动的时空模式进行融合，进而将微观和宏观观测统一起来。另一项有趣的扩展是，本



0000294

研究的时空统计模型虽然能够捕捉不同区域和时间段内的差异，但是一个潜在的挑战在于开发出高效的自动化算法，帮助检测出符合相同分布模式的连续时间和空间范围。这里提到的任何一项扩展都有助于我们更加深刻地理解人群在时空上的分布特征，进而开发出更加准确、实用的群体移动模型。



0000294

## 第五章 用户参与行为的时空建模

前面的章节中，我们分别从个体和群体尺度，研究了人类在物理空间的移动行为。随着当今移动通信技术的飞速发展，人们对移动应用和服务的依赖性普遍增强，网络空间成为丰富人们日常生活的一个重要维度。随着场景的变化，人们利用网络服务进行交友、订餐、获取资讯等多种活动，并通过移动网络流量记录了下来。这样的行为数据和移动行为类似，从时空角度记录了个人的使用习惯、兴趣偏好、行为模式等，但其揭示的一些人类时空行为特征，是纯粹的移动行为研究中所缺失的。这些活动有时伴随着物理空间的位置转移发生，有时是用户停留在某个地点时产生的；其表现出来的时空行为模式随着使用不同的硬件设备、不同的网络状况也有所差异。

本章节将人类时空行为分析从物理空间拓展到网络空间，利用移动网络数据，对大规模的用户参与行为进行了量化分析和建模研究。这里用户参与行为指使用移动网络服务的行为，是人类在时间和网络空间维度表现出来的行为特征。在从移动流量中识别出用户参与行为的基础上，我们首先从行为学角度，提出一组量化分析用户参与行为的指标，研究了参与行为和场景因素，如网络状况、物理位置、使用时间、硬件平台等的关联关系，并利用结构化分析的方法，对各因素之间的相互影响进行了研究。然后基于得到的参与行为模式和量化分析结果，我们将用户的参与行为轨迹抽象成多观测的时间序列，利用隐马尔可夫过程对行为序列进行建模，并从模型角度对显著的群体参与行为特征进行了研究。这些研究方法和成果在多个新兴的科学领域（如移动网络新型协议开发、用户体验测量和优化等）有着重要的借鉴和应用价值。

### 5.1 用户参与行为研究介绍

在以人为中心的技术和服务模型中，用户体验（UX）是一个重要的考量方面。从用户角度来看，一个成功的人机交互服务设计，不仅能够激发人们即刻的使用兴趣，而且能够让人们保持较为持久的黏着度。但是，随着产品或服务推向市场，最初的设计者和生产者难以及时获得珍贵的用户体验信息，从而将反馈改进到下一次的设计当中。



0000294

近年来，移动设备和应用的普及，为普适计算<sup>[84]</sup> 和行为科学<sup>[86]</sup> 研究提供了新的途径。借助于大范围和多样化的用户行为数据，研究人员能够便捷地从网络流量当中“感知”人们使用行为的变化。在最初的市场策略研究领域，参与行为的概念被提出<sup>[107]</sup> 来衡量新策略使用后的市场反应效果。O'Brian 等<sup>[80]</sup> 将参与行为的概念引入到 Web 测量中，并提出基于时间阶段的参与行为框架，对一般的人机交互行为进行描述。另一方面，Attfield 等<sup>[81]</sup> 从特性角度（如持久度）将参与行为分解成不同的描述维度，并在每个维度上将理论描述和实际观测指标联系在一起。

虽然上述两种理论框架具有较强的描述能力，但是对于移动网络参与行为的分析仍具有一定的局限性，主要表现在三个方面：1) 无论从时间还是特性角度，两个框架均未从时空结构上对用户参与行为进行研究，而时空结构是挖掘用户行为模式的重要依据之一。2) 未考虑场景因素对参与行为的影响；在移动网络中，场景因素从多个方面影响着用户的参与行为，如时段因素<sup>[87]</sup>，应用类型<sup>[86]</sup>，地理位置<sup>[85]</sup>，以及公共事件<sup>[43]</sup> 等。3) 缺少参与行为量化指标之间的关联关系挖掘；定量的关联关系分析有助于开发优化用户体验的策略配置，如流媒体社区对提高观影粘度的研究<sup>[78,79,82]</sup>，但是在移动网络中依然缺少大规模和系统化的研究工作。

针对已有框架中的不足，本文采用被动感知的方法对移动网络的参与行为进行研究。具体而言，我们的用户参与行为定义为：用户在特定场景和应用质量情况下，所表现出来的与网络服务交互及参与的总体观测行为。和主动测量的方法相比，被动测量不需要通过采访或调查问卷<sup>[80]</sup> 的方式，因而克服了个体的主管偏见或低样本数带来的实验偏差。我们的研究对参与行为进行量化测量，从而补充了上述理论框架的不足。但是，客观的被动测量受到了现实因素的影响，例如不同的网络接入选择导致行为数据不完整。为了在较为完整的行为数据上验证我们的方法，我们采用来源于同一网络类型、持续较长观测时间的 WiFi 数据集（2.2节）。由于我们的目标网络对注册者免费开放，因此保证了其在用户的候选接入方法中具有较高的优先级。

移动网络流量的用户行为模型是进行参与行为分析的基础。从流量角度来看，用户的时空行为信息隐藏在不同的协议层<sup>[108-110]</sup> 和应用特征当中，本文中我们提出了 *AID* 算法，对网络空间中的用户行为进行识别。图5-1A 展示了一个多维度的参与行为分析框架，分别从时间、物理空间、网络空间三个维度对用户参与行为进行了记录。在网络

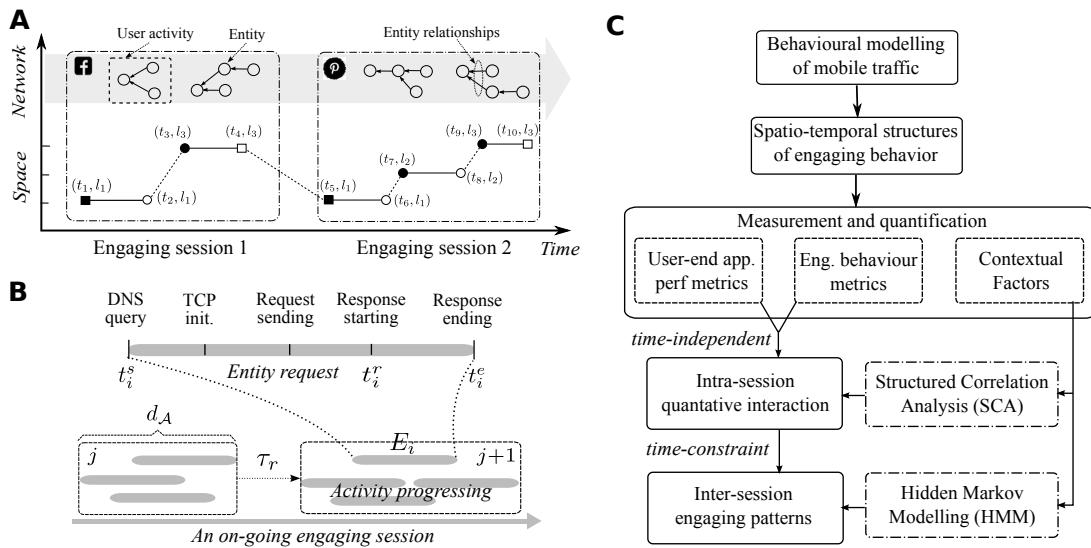


图 5-1 移动网络用户参与行为建模及量化分析框架

Fig 5-1 Illustration of user engaging behaviour in mobile networks and analysis framework.

维度上，*AID* 算法识别出带有语义信息（如应用名称和类型）的用户操作行为；在物理空间维度上，将用户的参与行为与网络接入点进行了关联；在时间维度上，我们记录了每次操作的时间戳，并基于时间阈值的方法识别出不同的参与会话。图5-1B 展示了从应用质量（Quality of Applications, QoA）角度对用户操作进行客观测量的过程。从网络角度观察，用户的单次动作（Activity）由若干网络实体（Entity）组成，因此我们分别从动作和实体粒度上对 QoA 性能进行了量化。图5-1C 以鸟瞰图的方式给出了本文研究用户参与行为的量化和分析框架。我们分别从应用质量、参与行为、以及场景因素角度，对个体的参与行为进行了量化测量，并利用结构相关分析的方法，对各因素的相互作用进行了研究。最后，在考虑参与会话之间的时间约束的前提下，对多观测的参与时间序列进行了建模研究，分别从个体和群体角度对参与行为的模式进行了挖掘。

## 5.2 移动流量中的用户参与行为识别

从移动流量角度观察用户行为，对评估用户体验、个性化服务推荐等起着重要的作用。由于接近网络底层，用户的时空行为信息隐藏在不同的协议层和应用特征当中，因此需要建立合理、准确的移动流量行为模型。

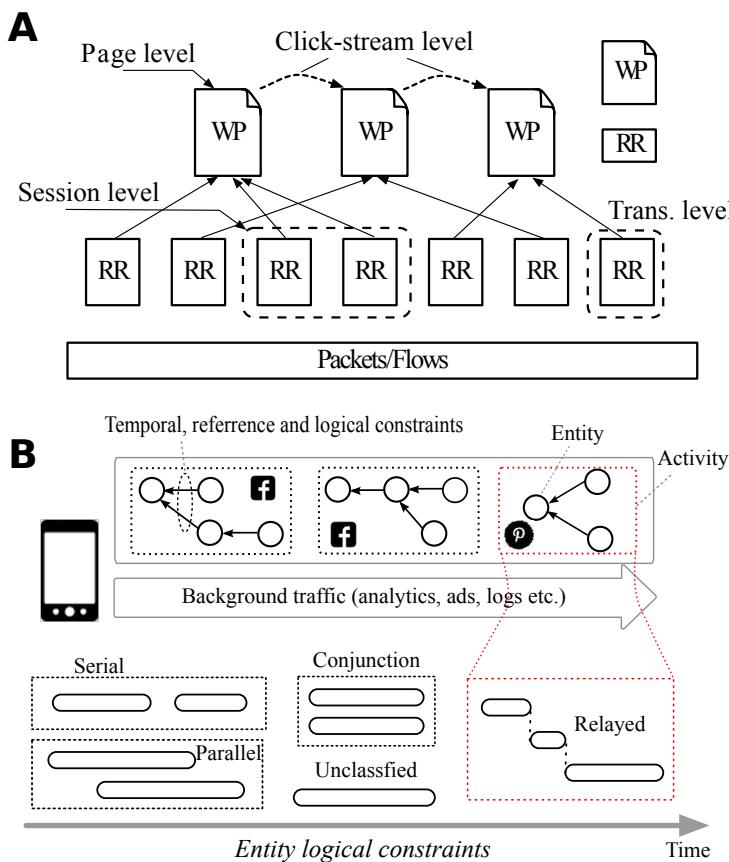


图 5-2 传统 Web 流量模型 (A) 和新型移动 HTTP 流量模型 (B) 对比

Fig 5-2 The model comparison between traditional web traffic and mobile HTTP traffic.

在移动应用兴起以前，传统的 Web 流量描述主要采用网页浏览模型<sup>[108-110]</sup>；在这类模型中，网页是内容组织的基本单元。通常网页建立在文本框架（HTML, XHTML 或 XML）上，页面包含若干内嵌的网页元素，如图片、视频、以及其他网页。当用户点击网页链接时，浏览器首先请求网页框架，然后解析、请求其他内嵌元素，最终通过软件协议栈发出网络数据包（Network packets），这种引用关系（Reference）如图5-2A 所示。因此 Web 流量的行为模型<sup>[109,110]</sup>，通过反向解析这种引用关系获得用户点击网页的动作序列。这类模型的代表性算法是 Ihm 等<sup>[110]</sup> 提出的 StreamStructure 算法，其识别精度达到 90% 以上，较以往算法提高 20%~40%。但该算法除了依赖元素引用关系外，还需要通过页面的 Google 分析标识对网页结束点进行判断，而这在许多网络场景下都受到限制。

和传统 Web 流量相比，移动网络流量有两个方面的特征：一是现代的移动操作系统（如 Android, iOS 等）拥有丰富的本地化应用<sup>1</sup>，这类应用采用更加灵活的 RESTful 接口代替网页形式，从而使网络角度的用户行为边界变得模糊。二是出于应用架构灵活和安全性考虑，传统模型中的引用关系在移动 HTTP 流量中比例大幅减少，如 CLICK 数据集中 70% 和 WIFI-T 中 63% 的 HTTP 记录中没有 Referer 字段。因此移动网络需要一种新型的流量模型来捕捉用户行为。

如图5–2B 所示，我们从用户行为逻辑角度提出了移动网络流量模型。通常，用户在移动应用上的点击行为会出发多个网络请求，如音乐播放会触发声频、歌词和图片等关联请求。这些请求虽然没有显示的引用关系，但是在逻辑上是关联在一起的。基于此我们提出了用户动作实体模型（即 Activity-Entity Model, AEM），其中动作表示用户主动的点击操作，记作  $A$ ，实体表示动作出发的网络对象请求，记作  $E$ ，且  $E \in A$ ，则用户在一段时间内的连续行为表示为  $\{A_t\}_{T_l, T_h}, T_l \leq t \leq T_h$ 。从一般性上来讲，这样的表示对传统 Web 流量同样适用。新模型的核心是从逻辑角度、而不是引用角度表示实体之间的关系。在本研究中考虑了以下四种逻辑关系：1) Parallel 关系 ( $E_i || E_j$ )，即不同实体是相互独立的，因此二者可以独立被触发请求，如页面内的不同图片；2) Conjunction 关系 ( $E_i \wedge E_j$ )，即实体间是彼此依赖的，需要同时请求成功才能进行下一动作，如视频及其声音文件；3) Serial ( $E_i \rightarrow E_j$ )，即其中一个实体依赖于另一个实体的完成，且二者组合起来实现一个动作的完成，如用户访问受限内容需要首先完成身份验证的动作；4) Relayed ( $E_i \leftrightarrow E_j$ )，即一个实体通过代替前一个实体完成一个共同的动作，如 HTTP 流媒体中的数据块请求。在这四种逻辑关系中，Parallel 和 Serial 表示动作之间较松散的关联，而 Conjunction 和 Relayed 表示动作之间较强的关联关系。

基于这样的行为模型，我们提出了用户网络参与行为的识别算法 ( $AID$ )，如算法5–4所示。该算法的核心思想是充分利用实体之间的逻辑关系<sup>2</sup>，包括传统模型中的时间约束和引用关系，构建网络实体树，然后利用时间约束将实体树分割成不同的子树，每个子树表示一次独立的用户点击行为。在移除背景流量（如广告、统计等）以后，算

<sup>1</sup>本地化应用（即 Native Apps）通常由第三方组织进行开发，并以软件包的形式在移动操作系统上进行安装使用。

<sup>2</sup>本研究中实体之间的逻辑关系通过静态规则的方法加入到算法中，如 Serial 关系存在于 URL 中包含 ‘auth’，‘login’ 等词的授权实体和紧邻访问的受限内容实体之间。

---

**算法 5-4 用户网络参与行为的识别算法  $\mathcal{AID}$** 


---

**Input:** 给定 {用户, 应用类型} 对, 按时间排序的实体序列  $\mathbf{E}$  和动作时间间隔阈值  $\tau_L$ ;

**Output:** 识别出无重叠的实体簇, 每个簇表示一个用户行为动作;

**for all**  $E_i \in \mathbf{E}$  **do**

**关联:** 通过关系库匹配获得两个实体  $E_{i-1}$  和  $E_i$  之间的逻辑关系  $LR(i-1, i)$ ;

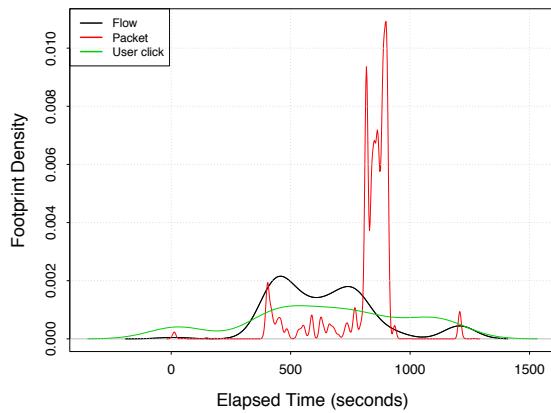
**匹配:** 当  $LR(i-1, i) \in \{\parallel, \wedge, \rightarrow, \leftrightarrow\}$ , 且  $\tau_{i-1, i} \leq 2\tau_L$  时, 将实体  $E_i$  匹配并连接到  $E_{i-1}$  上; 否则, 以  $E_i$  为根, 建立新的实体树;

**重连:** 如果  $E_i$  的引用字段 Referer 为空, 则将  $E_i$  直连 to  $E_{i-1}$  上; 否则, 断开  $E_{i-1}$  的连接, 按时间逆向回溯已经连接的实体, 并将  $E_{i-1}$  和  $E_i$  同时连接到时间最近的 Referer 指向的实体  $E_{ref}$ ;

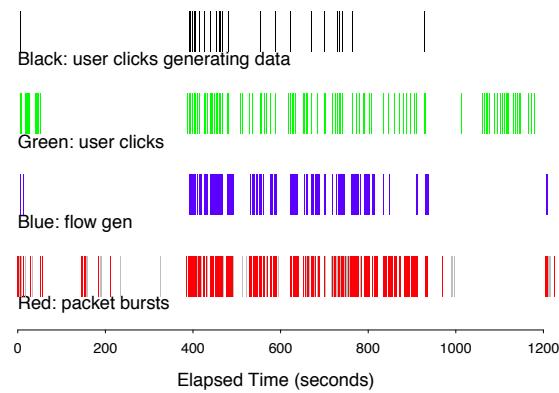
**end for**

**剪切:** 以深度优先方式遍历连接好的实体树, 如果第  $i$  个实体满足  $\tau_{ref, i} > \tau_L$ , 则断开  $E_{ref}$  和  $E_i$  之间的连接; 每个独立的实体子树构成一个实体簇。

---



(a) 用户点击行为和网络流量的突发性比较



(b) 用户点击行为和网络请求关联示意图

图 5-3 CLICK 数据集中用户行为和网络流量的关系示例

Fig 5-3 The relationship between user engaging behavior and network traffic bursts.

法输入为按时间排列的实体序, 且实体序有同一用户使用同一应用的过程中产生。其中参数  $\tau_{i,j}$  和  $\tau_L$  分别表示实体之间和动作之间的空闲时间。由算法5-4的性能评估可以看出, 实体之间的逻辑关系在重连阶段, 较时间约束和引用关系起到了主导作用。

接下来我们对  $\mathcal{AID}$  算法的性能进行评估, 并和传统模型中的 StreamStructure<sup>[110]</sup> 算法进行了比较。首先基于 CLICK 数据集, 我们生成了用户点击行为的基础, 如图5-3所示。图5-3(a)从突发性角度展示了用户行为和网络流量之间的关系, 可以看出网络包的

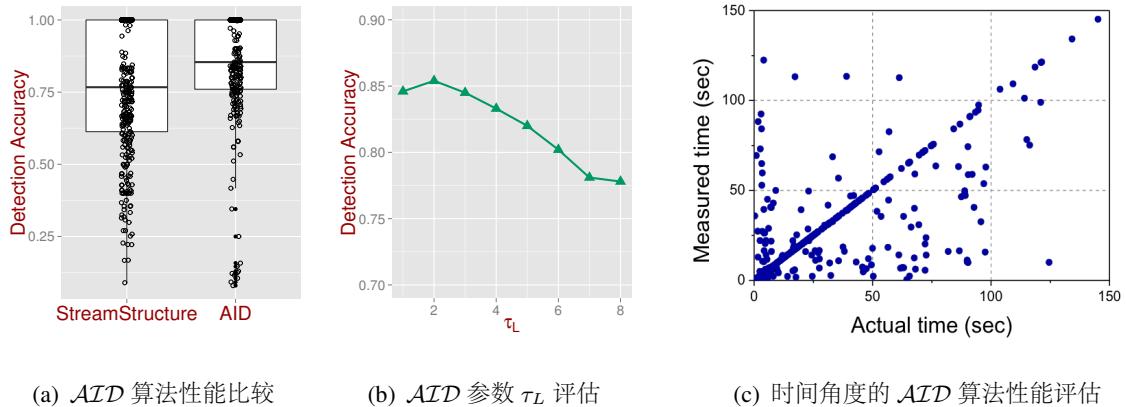


图 5-4 AID 算法性能评估  
Fig 5-4 The performance evaluation of AID algorithm.

突发性最强，且与点击行为的关联较弱；而网络流的突发性较为接近，不同的是，在网络数据请求完成后 ( $t = 1000s$ )，用户的点击行为依然持续，如翻页等操作。结合网络流和用户点击行为的突发性，我们生成用户参与行为识别的基准数据，即如图5-3(b)中的黑色点击序列所示。

在此基础上，我们对 AID 算法的性能进行了评估。图5-4(a)展示了和 StreamStructure 的比较，可以看出 AID 算法在识别准确率上有约 10% 的性能提升，且从数据点的聚集性上看，AID 的识别性能较为稳定。由于 WIFI-T 数据集中的 HTTP 请求约 70% 是没有 Referer 字段的，因此这样的提升主要来自于逻辑关系在关联网络实体上的应用。为了衡量参数  $\tau_L$  对算法性能的影响，我们选择了 1~8s 不同的取值对算法性能进行评估，从图5-4(b)可以看出最佳参数选择为 1~3s。此外，图5-4(c)从完成行为操作的延时角度评估了算法识别性能，可以看出 AID 算法检测出的行为延时和实际观测较为一致。

### 5.3 用户参与行为的量化分析

#### 5.3.1 用户参与行为的量化特征

虽然现有的理论框架<sup>[80,81]</sup>帮助我们从行为特征层面理解用户的参与行为，但是将分析结果应用到网络服务设计和优化中，则需要对用户的参与行为进行准确、客观的量化测量。在本节中，我们从行为过程、应用质量、以及场景因素等多个维度提取特征，

实现对用户参与行为的量化和分析。

### 5.3.1.1 行为特征

我们首先从用户与网络服务的交互过程对参与行为进行测量。这样的行为特征需要具备以下两个特点：1) 行为特征的量化指标需要具有较强的可解释性，即能够帮助人们从服务交互的角度理解用户的参与行为；2) 行为特征可直接用于实践性的分析和优化，即量化指标能够与现有的网络模型进行集成分析和研究，并对实践操作具有较强的指导作用。基于图5-1A 中的移动流量行为模型，我们的行为特征针对单个参与会话进行测量。给定具有  $m$  个操作行为的会话  $Q = \{\mathcal{A}_i\}_m$ ，行为特征包含：

**参与会话时长 (Engaging session duration)** 用于刻画用户在单个会话中的主动参与过程，是衡量参与持久度的基本指标。该特征定义为

$$D_e = \max\{T_i^e\} - \min\{T_i^s\}, i \in [1, m] \quad (5-1)$$

其中  $T^s$  和  $T^e$  分别为同一个用户动作的开始和结束时间（如图5-1B 所示）。在实际应用中，通过优化策略鼓励用户进行较长时间的会话，不但能够积累丰富的用户偏好画像，而且有助于增加用户参与过程中的潜在收益点，如提高广告投放的有效性。

**访问频率 (Visit frequency)** 包含了丰富的用户行为习惯信息。在 Attfield 等<sup>[81]</sup> 的理论框架中，用户参与过程与持久性 (Endurability)、交互性 (Interactivity)、以及受益度 (Pleasure) 紧密相关。我们使用访问频率来刻画用户参与行为的活跃和流畅程度。在 Web 研究领域中，人们利用类似的网页点击次数表示用户在一次浏览过程中的点击深度<sup>[75]</sup>。给定  $m$  个用户动作、以及相邻动作之间的间隔时间为  $\tau_r$ ，访问频率定义为用户在单位时间内产生的平均动作次数，即

$$f_v(H) = \mathbf{E}[\tau_r]_{m-1}^{-1}. \quad (5-2)$$

**行为异常率 (Ratio of interruption activities)** 是对访问频率特征的补充，从另一个角度描述了用户参与过程的流畅程度。在用户使用移动服务的过程中，网络性能的波动（如人群拥挤导致的服务体验下降）易导致中断、重载、放弃访问等异常操作行为<sup>[80]</sup>。但是在网络协议层面上，这类操作的观测特征容易与 TCP 协议的控制信号、以及浏览器

的实现方式产生混淆<sup>[111]</sup>。本研究中，我们采用了基于数据包负载的异常行为检测方法，从而实现了更加可靠的异常动作识别。如定义5.1所示，异常动作的第一个条件表示观测到客户端发出的重置（RST）数据包，同时服务器端没有发出结束（FIN）或重置信号；条件二表示结束标识符的空闲时间小于一个RTT的量值，换句话说，即用户在等待下一个数据包到来之前，主动结束数据流传输。行为异常率表示为在一个参与会话中，异常动作占所有用户动作的比例，即  $r_{int} = n_{int}/|Q|$ 。

**定义 5.1. 异常动作** (Interruption Activity)：通常情况下，移动服务的 HTTP 请求-应答对通过 TCP 流进行网络传输，因此异常动作基于 TCP 协议测量进行定义。一个异常动作的 TCP 流需要同时满足一下两个特征：

- 1)  $\neg(FIN_s \vee RST_s) \wedge RST_c$
- 2)  $\frac{t_{FIN}}{\mu_{RTT} + \sigma_{RTT}} \leq 1$

其中  $t_{FIN}$  表示成功传输的最后一个数据包和流结束标识之间的空闲时间； $\mu_{RTT}$  和  $\sigma_{RTT}$  分别给出了 TCP 流 RTT 时间的期望值和方差。

### 5.3.1.2 应用质量特征

应用质量特征是通过用户感知的移动应用性能对参与行为进行刻画。同样基于图5-1A 中的移动流量行为模型，以客观的方式对用户参与体验进行量化。和传统网络质量 (QoS) 分析<sup>[112-114]</sup> 相比，应用质量从应用使用行为、而不是网络协议层面上对用户体验进行评估。网络质量指标（如比特率）接近底层协议，是对网络传输数据质量的测量。在实际观测中，即使是在相同的网络质量条件下，不同的用户由于使用习惯、兴趣偏好的差异，也会在应用质量上表现出不同的特征。因此，为了更加准确地衡量用户参与过程中的服务体验，我们提出以用户为中心的应用性能量化指标：

**感知操作时长** (Perceived activity duration) 表示用户完成一次应用点击操作所需要的总体等待时间，是对用户感知性能的基本度量。给定包含  $n$  个实体的用户动作  $\mathcal{A} = \{E_i\}_n$ ，每个实体的完成过程通过三个时间值进行表示，即  $(t_i^s, t_i^r, t_i^e)$ ，分别表示请求发出时间、应答开始及结束时间，如图5-1B 所示。用户感知的应用操作时长为

$$d_{\mathcal{A}} = \max\{t_i^e\} - \min\{t_i^s\}, i \in [1, n]. \quad (5-3)$$

**感知等待时间** (Perceived waiting time) 是对感知操作时长  $d_{\mathcal{A}}$  的补充，从网络实体 (Entity) 粒度对用户的应用体验进行细粒度的测量。感知等待时间代表了用户在参与网络服务过程中忍耐度的消耗程度<sup>[11]</sup>，即等待时间越长，用户的忍耐度消耗越多。具体而言，我们将实体的请求过程分为两个过程，即静默等待 ( $|t_i^r - t_i^s|$ ) 和数据加载 ( $|t_i^r - t_i^e|$ )。前一过程表示客户端发出请求之后、第一个字节的应答数据到来之前的等待过程；后一过程表示数据开始接收至发送完成，通常情况下移动应用以进度条的方式将进度反馈给用户。由于一般情况下静默等待对用户忍耐度的消耗较大<sup>[80]</sup>，我们通过衡量静默等待过程表示用户感知到的等待时延，即

$$w_{\mathcal{A}} = \frac{1}{n} \sum_i |t_i^r - t_i^s|. \quad (5-4)$$

**感知吞吐率** (Perceived throughput) 由于移动应用种类丰富，因此单纯的时间度量难以充分量化用户感知的应用性能。例如从时间角度来看，持续一分钟的流媒体应用和持续 30 秒的微博操作可能产生相近的用户体验，这是因为流媒体数据消费时间长，而微博相对较短。为了衡量这样的差异，我们计算了感知吞吐率，即从用户角度对参与服务的数据传输速率进行衡量，即

$$b_{\mathcal{A}} = I_c \sum_i r(i) \quad (5-5)$$

其中  $I_c$  和  $r(i)$  表示并发指数和第  $i_{th}$  个实体的平均吞吐率。如定义 5.2 所示，并发指数同时包含了同一个动作里的实体关系，也包含了网络传输的时间信息。

**定义 5.2. 并发指数** (Concurrency Index)：给定用户动作  $\mathcal{A} = \{E_i\}_n$ ，并发指数定义为

$$I_c = \frac{t_{\mathcal{A}}}{n \cdot d_{\mathcal{A}}} \in (0, 1] \quad (5-6)$$

其中  $t_{\mathcal{A}} = \sum_i |t_i^s - t_i^e|$  表示所有实体的累计传输时长， $d_{\mathcal{A}}$  表示用户感知的操纵总体时延。

### 5.3.1.3 场景特征

在已有的 Web 参与行为分析中<sup>[81]</sup>，因为用户通常在固定终端浏览网页，研究人员较多关注网页的内容结构而不是参与时的外部场景因素。但是在移动网络中，用户的参



0000294

与行为可能在移动过程中发生，且较高的移动性伴随着较大的网络状态、以及外部场景因素变化。因此衡量场景特征对用户参与行为的影响有着重要的价值。本研究中，我们考虑了四个方面的外部场景特征，分别为**用户偏好**，**时间因素**，**地点偏好**，以及**应用语义**。

用户偏好指从单个用户的角度对参与行为进行分析，表现了由于使用习惯、历史经验等产生的个性化行为模式；该因素同时也对用户行为的其他三个维度产生较大的影响。虽然时间因素在已有网络质量工作中<sup>[112-114]</sup>得到较多研究，我们这里分别从天 (Time of day, ToD) 和周 (Time of week, ToW) 时段对用户的参与行为进行分析。在空间维度上，我们使用 PlaceRank 参数衡量用户对于某个地点的熟悉程度。该参数通过计算单个用户在特定地点的累计停留时间，并通过时间值逆向排序得出，即参数值越小，表示用户对地点的熟悉程度越高。对于应用的语义信息，我们利用移动流量的语义字段 (如主机名、终端类型) 对用户的单次操作行为进行分类。由于每个动作包含多个请求实体，我们将最先请求的实体作为应用语义分类的依据。

### 5.3.2 用户参与行为的统计分析

上述参与行为的特征，从不同方面对用户在网络上的时空行为进行刻画，且用户行为随着时空和场景因素的变化而有所不同。本节利用统计分析的方法，对上述参与行为的特征分布进行研究。这里以行为和应用质量特征为主要目标，同时从定性和定量的角度，对比了不同设备平台和场景因素对用户参与行为的影响。

在移动网络中，用户具有较高的移动性和数据即席 (Ad-hoc) 收发的特点，从而导致用户的参与行为表现出较大的变化。图5-5A 首先展示了不同硬件平台上用户参与会话数的差异，用户会话数较好地拟合一个截尾的对数正态分布 ( $p < 0.01$ )；该分布具有较大的偏度值 (Skewness)，即移动和非移动设备分别为  $\sigma = 3.79 \pm 0.20$  和  $\sigma = 4.14 \pm 0.40$ ，这意味着一小部分活跃的用户贡献了较大部分的移动网络流量。对于单个会话中的参与动作数，图5-5B 中可以看出不同平台上的分布具有很大的差异性，两条曲线的 KS 距离为  $ks = 0.694$ ；同时图5-5C 从时间角度确认了这样的差异性，且  $ks = 0.642$ ，其中移动和非移动用户会话时长的特征峰值为  $D_{mob} = 12$  和  $D_{nmob} = 78$ 。参与行为的这些差异性可能根植于不同硬件平台自身的特点：移动设备为了具有较高的便携性，往往在功

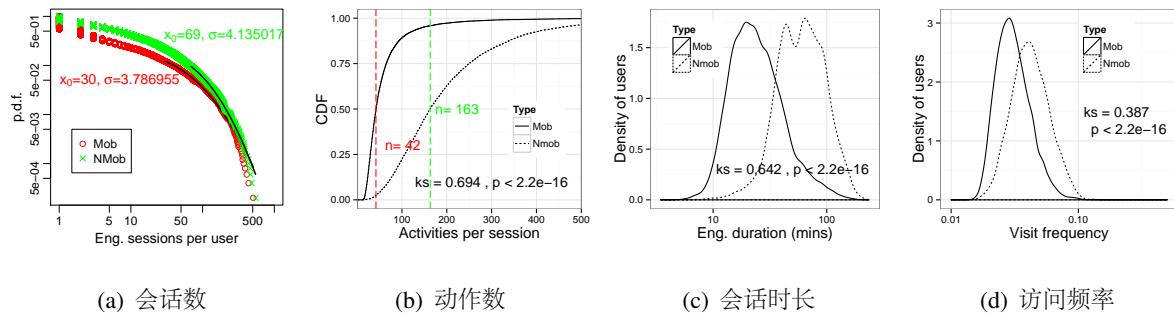


图 5-5 不同设备平台上用户参与行为特征对比

Fig 5-5 The diversity of user engaging behaviour concerning platform difference.

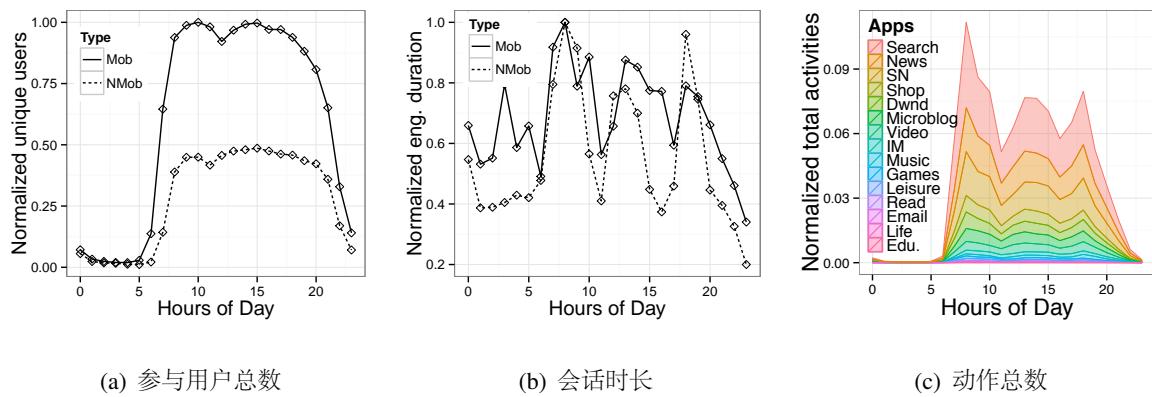


图 5-6 时间因素对移动用户参与行为的影响

Fig 5-6 The impact of temporal dynamics on mobile engaging behaviour.

耗和计算能力上有所折衷，从而导致用户的参与次数较多，但每次的持续时间和动作数较小。与其他特征不同的是，图5-5D中用户的访问频次并未因平台差异而有太大不同( $ks = 0.387$ )，进一步分析发现参与行为的差异与时间、地点等因素也无关，这表明访问频率的差异最可能来自于用户本身使用习惯和偏好上不同。在后面的关联分析中，我们将对访问频率的特征作深入的分析。

时间因素对参与行为的影响，决定了网络管理者如何制定策略是的网络资源的利用率达到最优，从而尽可能满足用户在不同时段对网络资源的需求。图5-6A给出了一天内不同时段的参与用户数分布，除了在总数上有所差异以外，不同平台上的用户行为较为相似。图5-6B分析了会话时长的差异，可以看出对于移动和非移动用户而言，在 $t = \{9h, 12h, 18h\}$ 等时刻出现明显的峰值；但是和非移动用户相比，移动用户的峰值

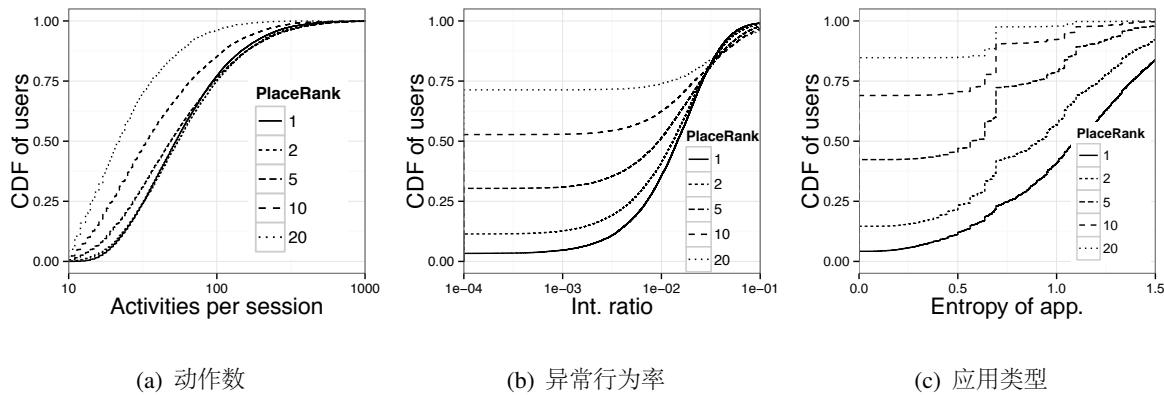


图 5-7 地点熟悉度对用户参与行为的影响

Fig 5-7 The impact of spatial preference on mobile engaging behaviour.

相对变化较小，这意味着空间因素对移动用户的参与行为影响较小。进一步，我们利用应用的语义信息对用户参与动作进行分类。需要注意的是，由于一些电子书应用（如 Kindle App）的流行，我们将 *reading* 和 *news* 分到不同的类别当中。如图5-6C 所示，在该网络中热门应用包括搜索、新闻、社交网络、视频以及微博等服务，但是从网络流量角度来看，搜索和视频占据了主导地位。这些时变规律表明移动用户的参与行为具有较强的偏好和模式。

在动态场景下，对参与行为影响的另一个重要因素便是空间位置。我们分析了位置偏好（即地点熟悉度）对参与行为的影响。如图5-7A 给出了不同熟悉度级别下，用户的参与会话时长分布，可以看出动作数与用户对地点的熟悉程度呈正相关，这是由于实际生活中，用户在熟悉的地点更倾向于进行长时间的参与活动。从经验角度来看，我们同样期望用户在熟悉的地点表现出更小的行为异常率，因为用户如果需要在某地点停留很长时间，便会表现出更多的忍耐性。但是相反的是，图5-7B 说明用户在参与过程中，在不熟悉的地点反而表现出对应用质量下降更高的容忍度。这种反常数据观测的一种合理解释是，当用户在一个陌生的区域时，由于不熟悉而降低了自身对应用质量的期望，从而表现出较低的异常行为率。

为了深入分析用户参与行为的模式，我们这里对用户感知到的应用质量特征进行研究。图5-8展示了三个具有代表性的应用，即邮件、微博和音乐，在不同设备平台上的用户体验。其中微博虽然具有较小的吞吐率，但是总体的动作时长较短。一方面因为这

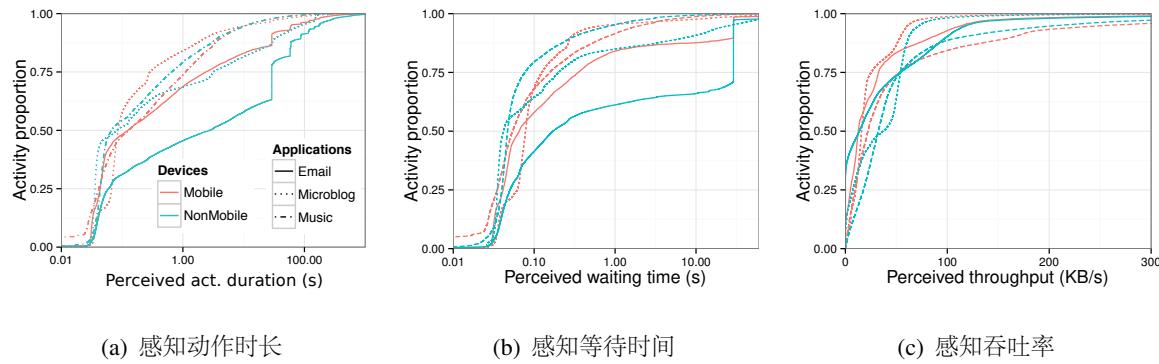


图 5-8 不同设备平台和应用类型下的服务质量特征

Fig 5-8 Visualizing the diversity of user-perceived application performance.

类应用的传输数据量较小，通常少数的 HTTP 应答即可完成数据传输，因此动作延时较小；另一方面由于 TCP 协议的慢启动机制，导致传输窗口未完全打开已结束通信，因此具有较小的吞吐率。特别的，邮件应用具有较高的概率经历较长的动作时长，如图5-8所示，为了验证图中曲线阶梯形状的产生原因，我们对不同浏览器的 TCP 协议超时行为进行了测量，其中 IE-10、Firefox-28 和 Opera-22 的超时设定分别为 60s、115s 和 120s，这些数据和观测值较好的拟合。综上所述，用户感知的应用质量差异，来源于硬件平台和应用协议特点的双重影响。

### 5.3.3 用户参与行为的关联分析

这部分我们着重研究用户的参与行为特征，与应用质量和场景因素之间的相互作用关系。一般来讲，对于变量关系的度量，一种潜在的分析方法是互信息（Mutual Information）分析，即利用随机变量  $X$  和  $Y$  之间的互信息来度量二者之间的关联程度。但是由于这种方法缺少变量作用关系的方向性信息，因此我们基于结构化分析的思想（即一个分布序列被分解或压缩成不同的映射子分布或条件分布），提出一种结构化相关系数来量化用户参与行为不同特征之间的关系。假设给定观测集合  $I$ ，每个观测由一个  $q$  维的特征向量表示；我们将观测特征空间  $O = \{m_i\}_q$  分为两部分，一部分称作观测特征，即研究相互作用关系的变量，记作  $T \subseteq O$ ；另一部分称作结构特征，即影响变量相互关系的其他因素，记作  $S \subseteq O$ 。在本研究中，我们的观测特征包括参与行为和应用质量特征，即  $m_i \in \{D_e, f_v, r_{int}\}$  和  $m_j \in \{d_A, w_A, b_A\}$ ；而结构特征指场景因素特征，即

$S = \{usr, app, loc\}$ 。给定结构特征集合  $X, Y \subseteq S$ , 观测特征  $m_i$  和  $m_j$  的相关系数为

$$R_{ij}(Y|X) = \{r_{ij}(x, y) | \forall y \in Y, X = x\}, \quad (5-7)$$

其中  $r_{ij}(x, y)$  表示但给定结构特征  $x, y$  时,  $m_i$  和  $m_j$  之间的相关系数。在结构化相关分析中, 相关系数的显著程度对实际应用和策略制定有着重要的作用。这里我们计算了显著系数

$$\phi = \frac{|R_{p<0.05}|}{|R_{p \geq 0.05}|} \quad (5-8)$$

表示在所有可能的观测中, 单个观测是显著相关的事件、而非偶然观测的概率。

在5-7式中, 我们使用 Spearman 相关系数 ( $r_s$ ) 来量化参与行为特征之间的相关程度。由于我们的观测数据代表了用户一部分行为过程, 因此直接利用经验数据计算 Spearman 系数具有较大的误差。这里我们将已有的观测数据看作底层用户行为特征分布的采样, 利用贝叶斯估计的方法对真实的相关系数进行计算, 即

$$P(r_s | m_i, m_j) \propto P(r_s) \frac{(1 - r_s^2)^{(n-1)/2}}{(1 - r_s \times r'_s)^{n-3/2}}, \quad (5-9)$$

其中  $r'_s$  表示从已有数据里计算得来的经验相关系数。利用因子代换  $r_s = \tanh \xi$  和  $r'_s = \tanh \eta$ , 我们发现  $\xi$  满足均值为  $\eta$ 、方差为  $1/n$  的正态分布。其中  $r_s$  的先验分布满足

$$P(r_s) \propto (1 - r_s^2)^c. \quad (5-10)$$

由于我们对先验分布的信息有限, 这里采用最简单的分布形式, 即  $c = 0$ 。最后, 结构化相关系数的符号含义和 Spearman 系数一致, 即  $r_s > 0$  表示变量之间单调递增的相互作用关系, 反之亦然。在特殊情况下,  $r_s = 1 / -1$  表示二者具有严格的单调关系, 且  $r_s = 0$  表示二者没有显著的相关性。

我们首先利用 WIFI-T 数据挖掘应用质量对移动用户参与行为的影响。如图5-9所示, 横轴对应应用质量特征, 纵轴对应行为特征; 其中黑色实线和蓝色虚线分别表示作了本地平滑以后的分布关系。总体来看, 这两组特征同时具有线性和非线性的相关关系。对于感知操作延时, 当  $0 \leq d_A \leq 10s$  时, 由于处在用户通常可以忍耐的范围内, 会话时长呈单调递增的趋势; 当  $d_A > 10s$  时, 会话时长的趋于平缓, 而此时行为异常率依然呈现逐渐增加的趋势。这表明, 用户对单次操作的延时忍耐阈值大约为 10s, 超

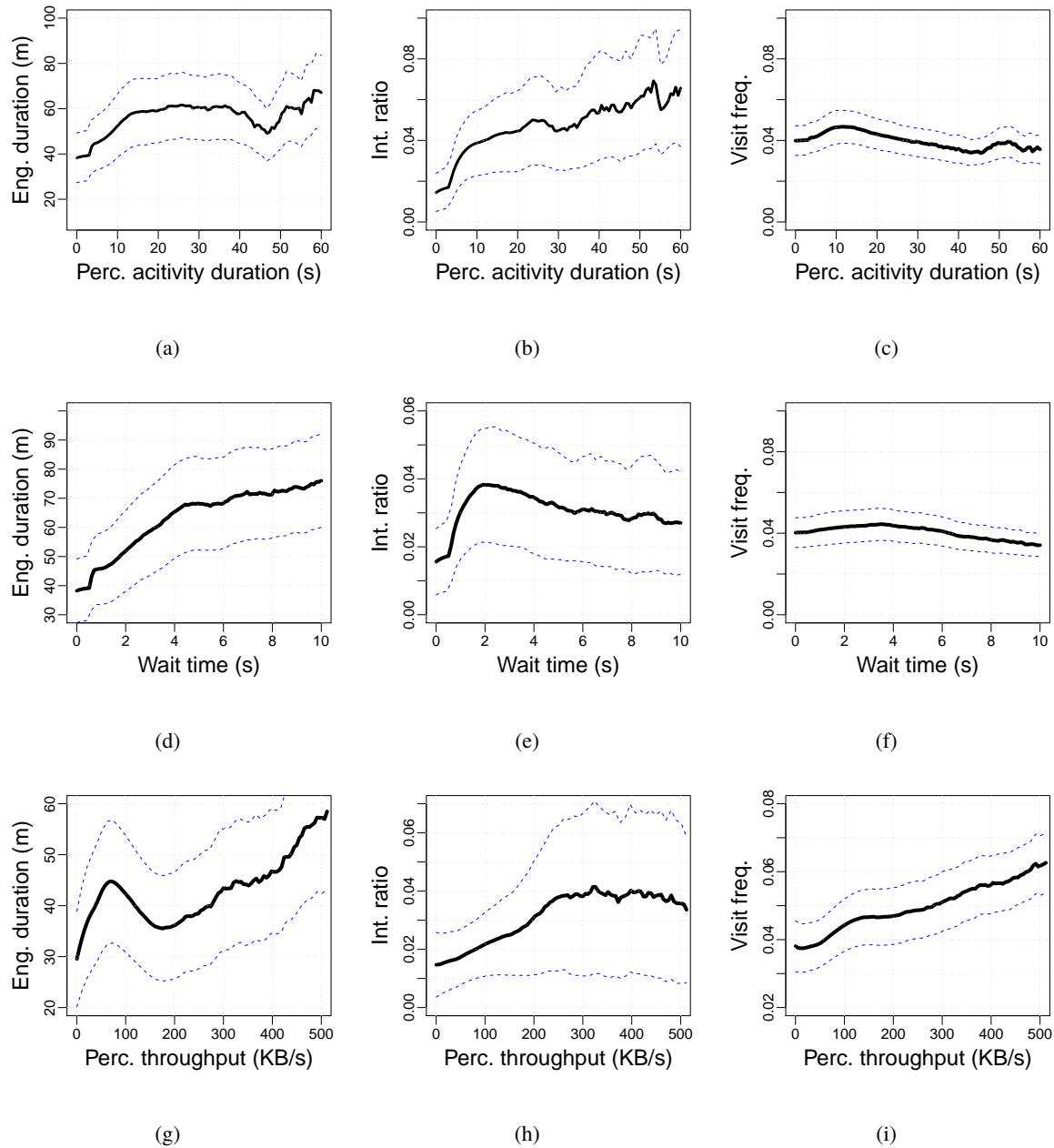
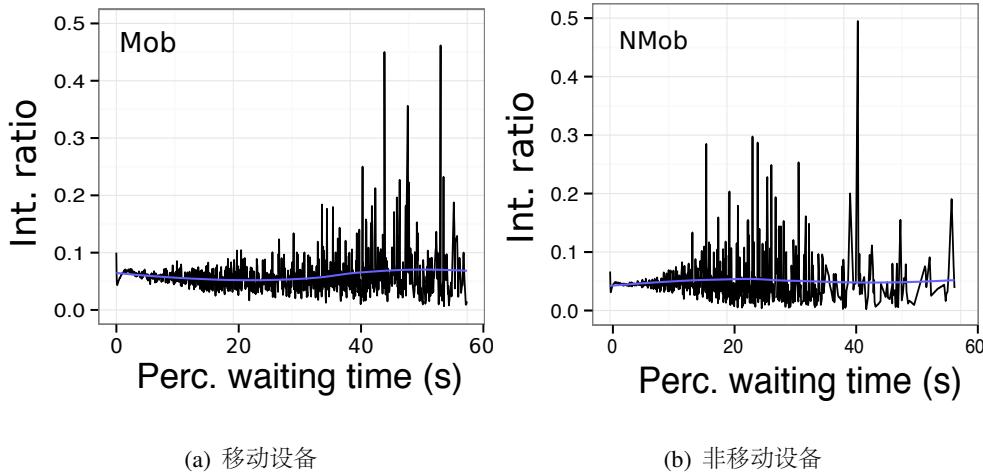


图 5-9 移动网络中用户参与行为与应用质量特征的关系

Fig 5-9 Qualitative relationships of mobile engaging behaviour and perceived application performance.

过此值以后，用户由于失败操作的比率升高而降低参与活跃度。而  $d_A$  对用户访问频次的影响并不显著。从感知等待时间来看， $w_A$  与  $d_A$  的趋势较为相似，不同的是前者在  $w_A = 2s$  附近出现一个特征峰值，在此峰值以后用户的异常行为率不再升高、且略有下



(a) 移动设备

(b) 非移动设备

图 5-10 移动网络中不同设备平台对异常行为率的影响

Fig 5-10 The influence of mobile platforms on interruption ratios.

降。这表明当  $w_A < 2$  时，用户的参与行为对网络延迟较为敏感，倾向于通过不断的重复操作恢复参与服务的过程；超过此范围后，用户对延迟的忍耐度增强。对于感知吞吐率，我们发现吞吐率越高 ( $b_A > 200kB/s$  时)，用户参与服务的会话时间越长，且访问频次越高。但是行为异常率却随着  $b_A$  的增加不断升高，且稳定在相对较高的水平。为了分析行为异常率的这种反常的行为模式，我们对其进行了深入的研究。

异常行为率是用户在服务参与过程中体验下降的信号。我们首先分析了吞吐率和单个动作数据量之间的关系，发现当操作的数据较少时，由于 TCP 协议的慢启动原理，在协议传输窗口未完全打开之前已经完成了传输，因此平均吞吐率较小。进一步分析发现，对于数据较小的用户动作，异常操作的吞吐率是完成操作的  $\sim 1.42$  倍，从而验证了上述观测中行为异常率随着吞吐率的增加不断升高的现象。图5-10展示了在不同硬件平台上用户的异常行为率随着感知等待时间的分布。虽然移动用户由于网络状态的变化而以较高的概率产生异常操作（移动和非移动用户的行为异常率三分位值分别为 6.25% 和 4.49%），但是他们比非移动用户对网络延迟的忍耐程度更高。例如，图5-10中非移动用户产生异常行为的峰值位于  $[15s, 45s]$  范围内，而移动用户的峰值区间更远，即  $[35s, 60s]$ 。这些观测结果，将有助于应用和服务开发人员设计出更加灵活的网络适配策略。

接下来，我们利用前面介绍的结构相关性分析方法，对用户参与过程中的行为特征

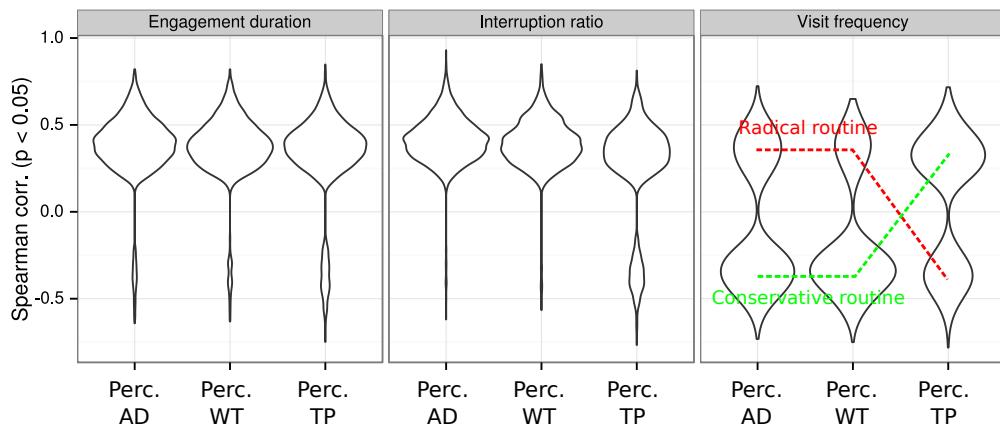


图 5-11 移动用户的参与行为相关系数分布图

Fig 5-11 The distribution of Spearman coefficient ( $p < 0.05$ ) for mobile users ( $n \geq 20$ ).

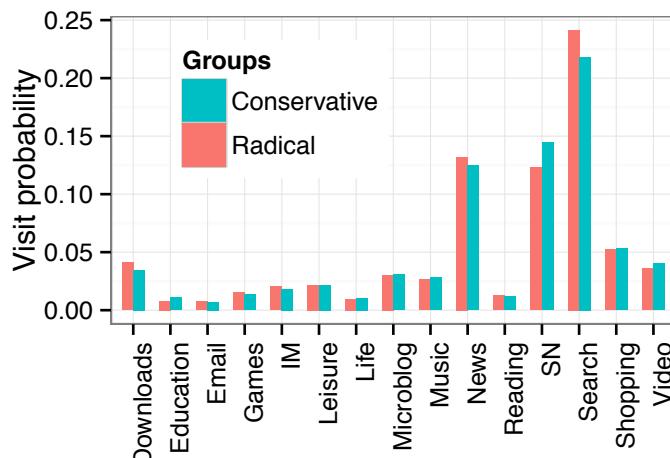


图 5-12 用户访问不同类型移动应用的概率分布

Fig 5-12 The probability distribution of visiting different applications by users in identified two groups.

和应用质量特征进行量化研究，同时考虑了以下三种不同的场景因素：

**用户偏好**侧重于用户个性化对参与行为的影响。根据5-7式，我们所要研究的相关性分布为  $R_{ij}(\{usr\})$ 。图5-11以小提琴图的形式给出了不同特征对的用户密度分布，其中垂直于小提琴轴的距离对应密度的大小。可以看出，对于大部分用户而言，会话时长和异常率都与应用质量特征呈正相关。比较有意思的是访问频次，我们观测到两种差异较大的行为模式，即一部分用户访问频次随着延迟的增加而变大；另一部分则相

表 5-1 参与行为特征之间的显著系数

Table 5-1 Statistics of users with significant correlation ( $p < 0.05$ ).

	参与会话时长	行为异常率	访问频率
感知操作延时	6144 (0.37)	12590 (1.29)	7139 (0.46)
感知等待时间	7099 (0.46)	9562 (0.75)	4438 (0.25)
感知吞吐率	4731 (0.27)	5340 (0.32)	7757 (0.53)

反。为了解释这样的现象，我们提出了两种假设：1) 应用类型不同导致参与行为的差异；2) 用户本身使用习惯的差异所致。为了验证第一种假设，我们分别计算了两组用户访问不同类型应用的概率，如图5-12所示，两组用户几乎无差别的访问概率分布帮助我们拒绝了第一种假设。相反地，如果是用户习惯的差异所致，我们应该有更加接近于经验直觉的观测结果。这里我们将两种不同的用户行为模式分类为平和型（Conservative）和激进型（Radical）：前一种类型的用户倾向于当应用质量低于自身期望时，便暂时离开进行其他任务（如在多任务操作系统中进行其他中断的任务），当应用质量恢复时则返回继续；后一种类型的用户则相反，当应用质量较差时，通过更多的访问动作以尽快完成所参与的会话，质量恢复时则能够以较少的访问次数完成任务。为了验证我们所观测的相关性现象的显著程度，表5-2给出了各特征对之间具有显著相关性的用户数和显著系数。这些量化结果有助于服务提供商根据需求优化服务性能，例如，优化用户操作的平均时长能够减少超过 50% ( $\phi = 1.29$ ) 的用户的行为异常率，而优化感知吞吐率最多只能让 29% 的用户收益 ( $\phi = 0.32$ )。

**应用语义**这部分我们研究应用类型的不同对参与行为的影响。为了区分用户和应用因素各自的影响，我们考虑两个分布： $R(\{app\})$  和  $R(\{usr\}|\{app\})$ ，分别从群体和个体粒度分析参与行为的特征。表5-2展示了两种粒度下的相关性大小，每个单元中前一个数值表示单纯应用语义的影响，括号中的数值表示应用-用户因素的联合影响。我们发现和参与会话时长与行为异常率相比，访问频次特征表现出更加复杂的模式：移动用户使用娱乐应用（如游戏、休闲、新闻等）的访问频次和应用质量特征呈正相关；社交应用（如即时通信、生活、社交网络等）随着感知动作延时和等待时间的增加而变大，而随着感知吞吐率的增加而减小。有趣的是，即时通信、阅读、视频等应用的行为异常率

与感知吞吐率的负相关关系更加明显，即  $r_{ij} = -0.25 \sim -0.37$ 。这些量化关系表明，如果服务开发者想要增加用户的参与活跃性，需要针对不同类型的应用采取有差异的优化措施，如减小社交应用类的感知等待时间，而增加阅读和视频类应用的感知吞吐率。

上述分析中的存在一个潜在的问题，即应用类型和用户偏好因素如何相互影响？对此我们将表5-2单元格中的数值进行比较，可以看出单独考虑应用类型时，用户的参与行为与应用质量的相关性较低，但是结合用户偏好因素时相关性显著增强。为了对两个因素的独立性进行检验，我们对两个分布进行了比较，即  $R(\{usr\})$  和  $R(\{app\}|\{usr\})$ ，其中后者比前者对于所有的特征对均拥有更大的相关系数，这表明用户偏好和应用类型两个因素相互独立、且互为补充。对于用户群组和个体应用偏好的关系，表面上看，表5-2和图5-11在访问频次上给出了不同的结论。实质上，这种统计上的差异来自于我们观测用户行为的尺度不同，即在个体行为粒度上的统计分布可分解成不同应用类型的子分布，且不同子分布表现出不同的相关关系。

**地点偏好**是人类行为的一个重要特征，且随着空间位置的变化对移动应用的偏好也有所改变<sup>[86]</sup>。在图5-13中，我们利用微博数据分析了地点熟悉度对用户参与行为的影响。首先，我们发现用户和地点偏好因素起到相互增强的作用，例如，对于用户偏好，特征对  $(r_{int}, d_A)$  分布的中值为 0.43，而相应的 PlaceRank-1 分布为 0.51。其次，对于较不熟悉的地点用户参与行为的相关性增强，这意味着在不常去的地点，移动用户的行为对应用质量的下降更加敏感。进一步，我们对 PlaceRank-X 和纯用户偏好的分布进行了 KS 测试。具体而言，对于行为异常率和感知动作时长的相关性，KS 指数分别为  $ks_1 = 0.384$ ,  $ks_2 = 0.472$ ,  $ks_3 = 0.578$ ；对于访问频次和感知吞吐率之间的相关性，KS 指数分别为  $ks_1 = 0.363$ ,  $ks_2 = 0.444$ ,  $ks_3 = 0.513$ 。有趣的是，通过结合图5-5D 和图5-13A，我们发现对于不常去的地点，尽管行为异常率受感知动作时长的影响较大，但是移动用户却在这些地点表现出对应用质量更高的容忍度。

## 5.4 用户参与轨迹的时空模型

在前两节的内容当中，我们看到移动用户的参与行为特征具有较强的时空依赖性，且与应用质量和场景因素（如用户偏好等）有着紧密的联系。通过客观的量化分析方法，我们得以对移动用户的参与行为进行测量，但是为了将参与行为的量化特征应用到

表 5-2 不同应用语义对参与行为特征相关性的影响

Table 5-2 The impact of application category on user engaging behaviors.

	Engaging session duration				INTERRUPTION ratio				Visit frequency		
	Perc. A.D.	Perc. W.T.	Perc. T.P.	Perc. A.D.	Perc. W.T.	Perc. T.P.	Perc. A.D.	Perc. W.T.	Perc. T.P.	Perc. T.P.	
Downloads <sup>†</sup>	0.34 / 0.49	0.25 / 0.46	0.18 / 0.44	0.40 / 0.48	0.32 / 0.46	0.14 / 0.45	-0.01 / 0.26	-0.03 / -0.42	- / -0.49		
Games	0.25 / 0.40	0.24 / 0.37	0.18 / 0.35	0.35 / 0.44	0.24 / 0.42	0.17 / 0.40	- / -0.26	0.02 / -0.29	-0.14 / -0.28		
IM	0.27 / 0.47	0.23 / 0.45	0.05 / 0.46	0.25 / 0.47	0.25 / 0.47	0.04 / -0.37	-0.02 / -0.40	-0.04 / -0.43	- / 0.33		
Leisure	0.18 / 0.38	0.18 / 0.42	0.13 / 0.40	0.28 / 0.46	0.31 / 0.37	0.16 / 0.37	-0.06 / -0.18	-0.04 / -0.20	0 / -0.36		
Life	0.29 / 0.52	0.23 / 0.57	0.26 / 0.42	0.35 / 0.48	0.21 / 0.37	0.15 / 0.40	-0.07 / -0.43	-0.11 / -0.50	-0.04 / 0.51		
Microblog	0.24 / 0.44	0.20 / 0.40	0.30 / 0.44	0.38 / 0.43	0.31 / 0.42	0.14 / 0.40	0.17 / 0.38	0.12 / 0.35	0.13 / 0.38		
Music	0.25 / 0.44	0.21 / 0.43	0.21 / 0.41	0.43 / 0.59	0.35 / 0.53	0.15 / 0.45	0.03 / -0.58	0.05 / 0.02	- / -		
News	0.31 / 0.46	0.27 / 0.45	0.20 / 0.43	0.33 / 0.44	0.30 / 0.42	0.18 / 0.41	-0.02 / -0.30	-0.02 / -0.37	-0.02 / -0.39		
Reading	0.07 / 0.42	0.11 / 0.55	0.20 / 0.41	0.30 / 0.54	0.26 / 0.41	- / -0.25	-0.05 / 0.43	-0.05 / -0.33	- / -0.66		
Search	0.20 / 0.44	0.29 / 0.43	0.27 / 0.44	0.33 / 0.44	0.28 / 0.42	0.21 / 0.41	-0.03 / -0.26	-0.02 / -0.35	-0.01 / -0.35		
Shopping	0.31 / 0.50	0.15 / 0.43	0.16 / 0.46	0.33 / 0.49	0.24 / 0.44	0.13 / 0.41	-0.07 / -0.38	-0.11 / -0.40	0.06 / 0.36		
SN	0.26 / 0.41	0.24 / 0.38	0.10 / 0.40	0.23 / 0.39	0.17 / 0.38	0.16 / 0.35	-0.03 / -0.37	-0.04 / -0.36	0.08 / 0.33		
Video	0.15 / 0.48	0.25 / 0.49	0.24 / 0.49	0.29 / 0.48	0.22 / 0.49	-0.03 / -0.31	-0.02 / -0.23	0.05 / 0.44	0.11 / 0.25		

<sup>†</sup> Applications and pairs with failed estimation (e.g., due to insufficient observations) of the real correlation are denoted by ‘-’ in the table.

<sup>‡</sup> For the right side of each cell, the median of zero means that roughly two halves of the whole population have opposite correlative tendencies.

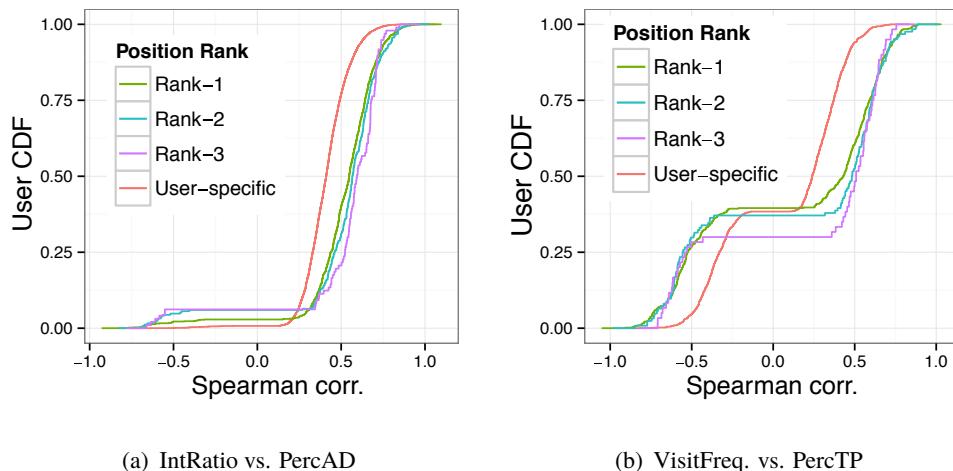


图 5-13 空间偏好（地点熟悉度）对用户参与行为的影响

Fig 5–13 The impact of location familiarity on engaging behaviour ( $p < 0.05$ ).

其他领域当中，需要对参与行为的过程建立理论模型。本节从两个角度对移动用户的参与行为建模：一是在个体粒度上对参与行为的时间序列进行建模，二是在群体粒度上为参与行为的时间序列建立聚类模型。

#### 5.4.1 用户参与行为建模

与5.3节中基于单个参与会话的分析不同，我们这里添加了对参与行为观测的时间约束，同时空间约束包含在场景特征当中。具体而言，对于每个参与会话，我们有来自3个维度的10个观测特征，分别是行为特征  $\mathbf{O}_b = \{D_e, f_v, r_{int}\}$ ，应用质量特征  $\mathbf{O}_p = \{d_A, w_A, b_A\}$ ，以及场景特征  $\mathbf{O}_c = \{App, PR, ToD, ToW\}$ ，其中 PR 表示用户对地点的熟悉程度，即 PlaceRank。从时间序列的角度来看，用户的参与行为可表示为一个具有多类型特征的多观测时间序列，定义5.3给出了具有  $q$  个特征的一般性定义，在本文研究中， $\mathbf{O} = \mathbf{O}_b \cap \mathbf{O}_p \cap \mathbf{O}_c$ 。

**定义 5.3. 参与轨迹序列 (Engaging Trajectory) :** 给定更具有  $q$  个变量的参与会话  $\mathbf{O}_t = (O_t^1, \dots, O_t^q)$ , 用户的参与轨迹序列表示为按时间排序的一组参与会话的集合, 即

$$\mathbf{O}_{1:T} = \langle \mathbf{O}_1, \dots, \mathbf{O}_T \rangle.$$



0000294

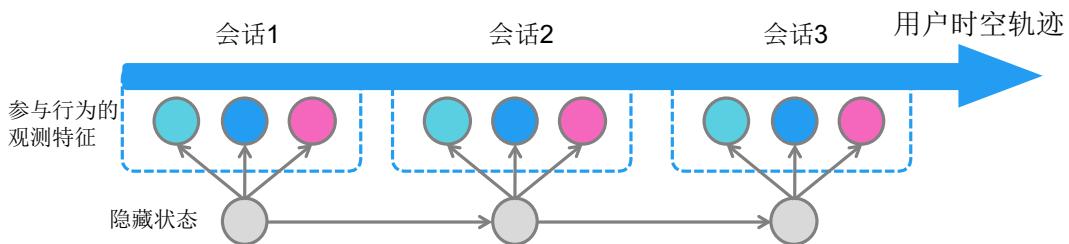


图 5-14 移动用户参与行为的隐马尔可夫过程示例

Fig 5-14 The illustration of Hidden Markov Model for mobile engaging behavior.

对用户的参与轨迹序列进行分析，我们通常有两种不同的思路：一种思路是采用多观测时间序列的分析方法，在变量相关性分析（如5.3.3节）的基础上，对整个时间序列的分布规律进行建模。但是这样的方法面临着特征类型不一致的挑战，例如在对群体轨迹序列聚类的分析中，由于单个观测点既有类别型变量（如应用类型），也有连续型变量（如异常行为率），而类别型和连续型变量具有不同的计算方法，且二者表达不同的物理意义，因此很难对具有不同类型特征的多变量观测点进行距离比较。第二种思路则基于隐藏状态的思想，即假设用户在时间  $t$  的观测  $\mathbf{O}_t$  由隐藏状态  $S_t$  决定，而不同特征对应着以隐藏状态为变量的不同激发函数。这样则避免了多观测时间序列分析中类型不一致的问题。因此在本文中，我们采用第二种思路对用户的参与轨迹序列进行研究。

假设用户隐藏的参与状态只和上一次的状态有关，即满足马尔可夫的特性，比较直观的方法便是采用隐马尔可夫模型对用户的参与过程建模。如图5-14所示，在单个参与会话中，用户行为的隐藏状态（如实际的需求）决定了用户在不同场景和应用质量条件下产生的行为特征；随着时间的变化，用户行为的隐藏状态也在不断地发生着变化。这样模型的合理性在于，在被动测量中用户真实的需求信息是缺失的，因此用隐藏状态对用户真实的需求进行表示，可以通过外部观测合理地捕捉用户的行为状态。

隐马尔可夫模型能够高效地捕获时间序列背后的产生机制，其衍生模型已经在语音识别和生物信息学领域得到广泛应用。和标准的隐马尔可夫模型相比，我们的研究中每个参与行为的观测由有限个隐藏状态叠加产生，其中由于场景因素对行为和应用质量特征之间的相关性影响较大，我们将场景因素作为单独的一组特征考虑。对于给定的参与轨迹序列  $\mathbf{O}_{1:T}$ ，隐藏状态和场景特征的观测序列分别为  $\mathbf{S}_{1:T}$  和  $\mathbf{c}_{1:T}$ 。假设  $L$  是所有隐藏状态的集合， $V$  是所有可能的观测的集合，即  $L = \{l_1, \dots, l_N\}$ ， $V = \{v_1, \dots, v_M\}$ ，

其中  $N$  和  $M$  分别为可能的状态数和观测数，则隐马尔可夫模型  $\lambda$  需要三组参数进行确定：初始转移概率  $\pi$ ，状态转移矩阵  $A$ ，以及观测激发概率矩阵  $B$ ，即

$$\lambda = (A, B, \pi) \quad (5-11)$$

其中  $a_{ij} \in A$  表示状态  $l_i \rightarrow l_j$  的转移概率， $b_i(k) \in B$  表示由状态  $l_i$  产生观测  $v_k$  的概率， $\pi_i$  表示用户在  $t = 1$  时处于状态  $l_i$  的概率。因此对用户的参与轨迹序列建模，需要根据观测对模型参数  $\lambda$  进行确定，即确定参数使得观测序列的似然（Likelihood）最大

$$\hat{\lambda} := \arg \max_{\lambda} P(\mathbf{O} | \lambda, \mathbf{c}). \quad (5-12)$$

这里采用非监督的学习算法来确定参数  $\hat{\lambda}$ ，根据 Baum-Welch<sup>[115]</sup> 判定原则可得

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (5-13)$$

$$b_i(k) = \frac{\sum_{t=1, o_t=v_k}^T \gamma_t(i)}{\sum_{t=1}^T \gamma_t(i)} \quad (5-14)$$

$$\pi_i = \gamma_1(i) \quad (5-15)$$

其中  $\gamma_t(i) = P(i_t = l_i | \lambda, \mathbf{O}, \mathbf{c})$ ，且  $\xi_t(i, j) = P(i_t = l_i, i_{t+1} = l_j | \lambda, \mathbf{O}, \mathbf{c})$ 。虽然上述过程能够根据观测序列确定出最优的模型参数，但是需要给出隐藏状态的数目。通常情况下，状态数目越多，模型的复杂度越高，对观测数据的拟合度越好；反之，状态数目越少，模型的复杂度相应降低，但是对数据的拟合度也越差。本文中采用 AIC (Akaike's Information Criteria) 信息度量指标对隐藏状态的数据进行选择，从而对模型的复杂度和拟合度进行平衡。由于参与轨迹序列的观测长度一般较少（即  $\bar{T} = 50.87$ ），我们采用一种修正的 AIC 指标来克服小样本数带来的过拟合风险，即

$$AIC_c = AIC + \frac{2k(k+1)}{T-k-1} \quad (5-16)$$

其中  $k$  和  $T$  分别表示模型参数的数目和观测序列的长度。由此看见，模型的参数越多、序列观测数目越少，修正的  $AIC_c$  值越高。

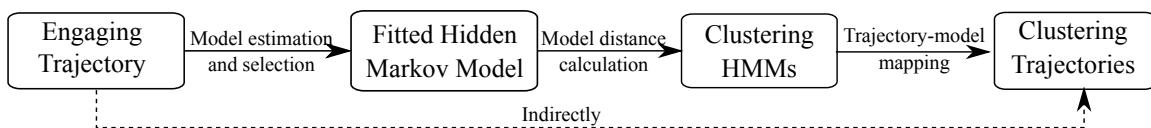


图 5-15 基于模型的参与轨迹序列聚类分析

Fig 5-15 A model-based clustering for user engaging trajectories.

#### 5.4.2 群体参与行为聚类

在个体参与行为建模的基础上，我们对群体的参与行为进行聚类研究。如前所述，由于观测中数据类型的不一致，对多观测时间序列的直接聚类，需要寻找可靠的距离指标对时间序列进行计算。与此不同的是，我们采用基于模型聚类的思想，如图5-15所示，如果两个参与轨迹序列的生成模型比较类似，则这两个序列类似。

聚类分析通常基于相似矩阵对元素进行无监督地分类，并对每个类别分配唯一的标识。为了计算两个隐马尔可夫模型之间的距离，一种潜在的直接方法是将模型参数序列化为参数向量，然后利用向量距离的计算方法对模型距离进行评估。但是者面临着两个挑战：1) 选择的隐藏状态不同，模型的复杂度不同，因此参数向量的长度不同。虽然能够采用补零的方法将参数向量对齐，但是这么做降低了对象的可分性，难以达到较好的分类效果。2) 由于各参数的物理意义不同，因此将不同参数在同一尺度内进行向量距离计算，所得的聚类结果从参与行为分析的角度来讲可解释性较低。

基于这些原因，而观测到  $P(\mathbf{O}|\lambda_i)$ <sup>1</sup> 为  $\lambda_i$  在整个轨迹序列空间内的条件概率分布，因此我们采用基于概率分布的散度（即 Hellinger 距离）对模型的距离进行衡量。给定具体的参与轨迹序列  $\mathbf{O}_{1:T}$ ，产生此观测序列的最可能隐藏状态序列为：

$$\mathbf{S}_m := \arg \max_{\mathbf{S}} P(\mathbf{O}|\mathbf{S}, \lambda_i)P(\mathbf{S}|\lambda_i) \quad (5-17)$$

因此轨迹序列模型的分布可以表示为最优隐藏状态序列的分布，即

$$f(\lambda_i) = P(\mathbf{O}, \mathbf{S}_m | \lambda_i) \quad (5-18)$$

换句话说，最优隐藏状态序列以最大的概率产生我们观测到的用户参与轨迹序列。对于

<sup>1</sup>为了后续表示的简洁，我们省略掉条件变量里的场景因素  $\mathbf{c}$ 。

模型  $\lambda_i$  和  $\lambda_j$ , 他们的距离  $H$  可以通过对概率分布的距离进行衡量, 即

$$H^2(f(\lambda_i), f(\lambda_j)) = 2 \int [\sqrt{f(\lambda_i)} - \sqrt{f(\lambda_j)}]^2 d\mathbf{O}. \quad (5-19)$$

5-19式的积分限为整个可观测到的轨迹序列空间。在现实中这是难以满足的, 因为我们无法在给定的模型参数空间里, 对所有可能出现的轨迹序列进行穷举。大数定理告诉我们, 一种可行的替代方案是采用蒙特卡罗 (Monte Carlo) 采样, 从完整的搜索空间里以等概率获得一份较小的样本集 ( $\sum_i T_i \rightarrow \infty$ ), 然后进行距离计算。由于我们无法获得模型参数的先验约束, 因此导致这样的方法需要进行大量的采样和高密度的计算才能获得趋近于  $n^{-1/2}$  的误差。

本研究中, 我们提出一种快速算法获得模型距离的近似解。该算法基于的假设为, 轨迹序列的所有观测以较高的概率分布在我们所观测到的样本周围。基于此假设我们便只需要在优先的观测集上采用蒙特卡罗采样, 从而缩小了积分的样本数目  $S_N$ :

$$H_{ij} = \sqrt{2 \int_{S_N} [\sqrt{f(\lambda_i)} - \sqrt{f(\lambda_j)}]^2 d\mathbf{O}} \quad (5-20)$$

这一假设的合理性在于, 实际生活中, 用户在相似网络环境和场景下所表现出来的行为具有较高的同质性。随着蒙特卡罗采样数目  $N$  的增加, 模型之间的距离  $H_{ij}$  将趋于一定稳定值。后续的数据分析实验显示较少的样本数即可产生稳定的模型距离估值。最后, 我们采用层级聚类 (Hierarchical Clustering) 算法<sup>[103]</sup> 对模型进行聚类, 从而达到对用户参与轨迹聚类的目的。

### 5.4.3 数据分析及验证

我们结合 WIFI-T 和 WIFI-M 数据集对用户的参与行为建模进行验证, 单个观测值包含一个参与会话的行为特征  $\mathbf{O}_b = \{D_e, f_v, r_{int}\}$ , 应用质量特征  $\mathbf{O}_p = \{d_A, w_A, b_A\}$ , 以及场景特征  $\mathbf{O}_c = \{App, PR, ToD, ToW\}$ 。图5-16左侧首先展示了两个随机用户的应用和地点访问时间分布, 右侧展示了所有用户的特征熵分布, 其中点画线表示随机情况下的特征熵; 可以看出所示特征 (尤其是应用和地点偏好) 均原理随机行为, 从而表明用户在网络和物理空间的行为都具有较强的规律性。

为了获得误差较小的用户行为数据, 我们选择了在一周内平均产生 20 个参与会话 (序列观测点) 的用户。图5-17展示了对于用户 U39856 的模型选择参数及对应的隐状

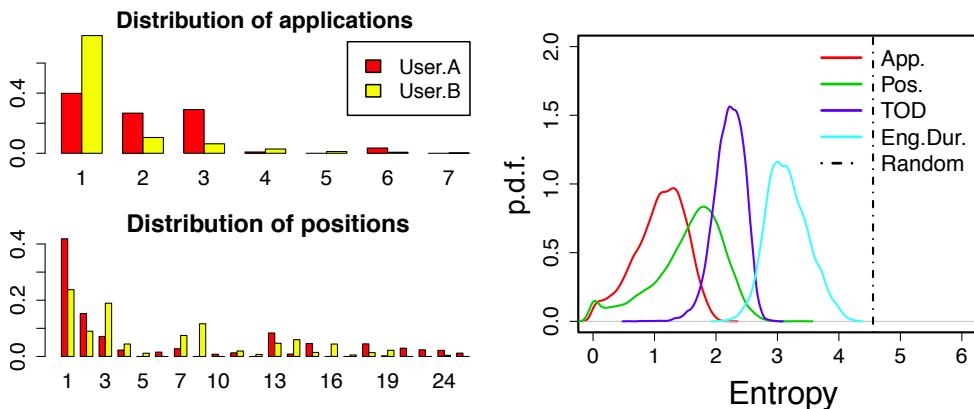


图 5-16 不同用户的特征分布示例及全体用户的熵分布

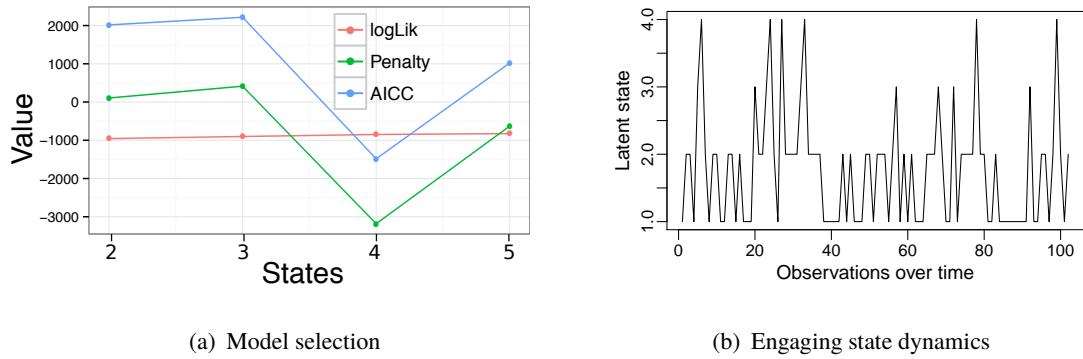
Fig 5-16 The distributions of user behavioral features and the feature entropy.

态时间序列；可以看出对于该用户，在模型拟合度保持不变的情况下， $AIC_c$  对复杂的模型进行了惩罚，最终选择平衡了模型拟合度和复杂度的  $N = 4$  个隐状态。图5-17B 表明该用户的参与状态主要停留在状态 1 和 2（即  $p_0 = 0.3$  时的显著状态），并以较小的概率转换到状态 3 和 4。为了分析用户在不同隐状态下的参与行为特征差异，我们在图5-18中给出了具有不同状态数的两位用户的特征分布；可以看出，对于用户 U39856，感知动作延时和等待时间在显著状态 1 和 2 上的方差和非显著状态 3 和 4 相比较小，这表明非显著状态代表了用户参与过程中的异常行为。对于其余的四个特征，显著状态 2 比状态 1 的取值较小且较为集中，例如状态 2 的参与会话时长中值为 27.5 分钟，而状态 1 为 13.2 分钟。而用户 U39874 的显著状态在相应特征维度上的分布差异较小。从以上观测可以得出两点结论：一是用户的参与轨迹序列通常受数目较少的隐状态支配着，即以较大的概率停留在这些少数状态上；二是尽管行为特征的分布有所差异，但是个体的隐状态分别对应着较窄的特征范围，即每个隐状态捕捉了特定的参与行为模式。

接下来，我们利用模型聚类的方法对群体的参与行为进行分析。图5-19A 首先展示了提出的快速近似算法的  $H_{ij}$  计算；随着蒙特卡罗采样数  $N$  的增大，用户 U39856 和 U39874 的相似距离在  $N = 100$  附近开始收敛，且  $H_{ij}$  约为 0.065，此时平均观测数为

$$\bar{T} * 100 \sim 2^{12} \quad (\bar{T} = 50.87) \quad (5-21)$$

这与 Dugad 等<sup>[115]</sup> 的观测一致，即当观测值增加到  $2^{10} \sim 2^{12}$  量级时，隐马尔可夫模型的距离收敛于一个稳定值。由于  $N = 100$  的采样标准对数据集中的其他用户同样适用，

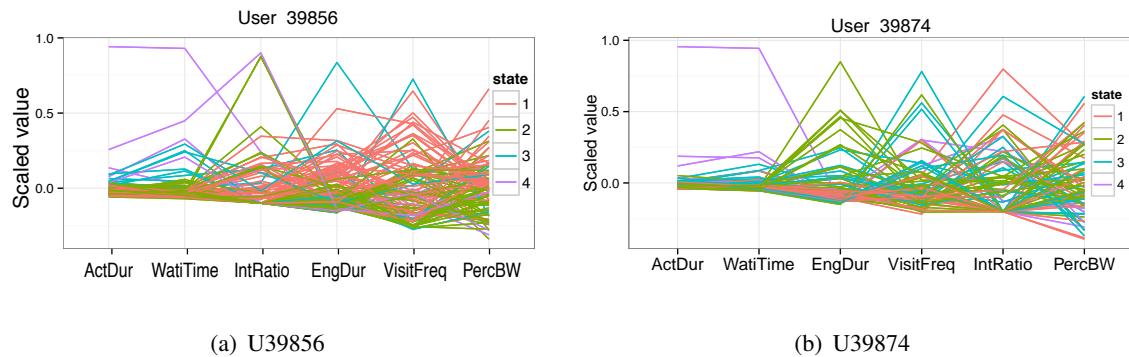


(a) Model selection

(b) Engaging state dynamics

图 5-17 用户参与行为的模型选择及隐状态时间序列

Fig 5-17 Illustration of model selection and series of hidden state (U39856).



(a) U39856

(b) U39874

图 5-18 用户参与行为的不同隐状态及对应特征分布

Fig 5-18 Illustration of feature distributions across behavioral features for different users.

因此本研究中采用此标准进行聚类分析。图5-19B~D 给出了聚类分析的结果，可以看出，无论从簇内距离还是簇间距离分析，用户的参与轨迹序列都呈现出明显的聚集性。通过手动调节簇数目 2~8，我们最终选取分类效果较为显著的  $k = 2$ ，同时 Shepard 图（图5-19D）也以较高的相关系数  $r_s = 0.677$  支持了我们的聚类结果。

这种聚集性表明移动网络中的用户参与行为具有潜在的模式。为了了解产生这些模式的根本原因，我们研究了不同聚类簇下的参与行为特征分布。如图5-20A 所示，我们发现聚类簇在单一特征上的分布无明显差别。接下来我们分析了行为和应用质量特征的相关性随着聚类簇不同而发生的变化（例如图5-20B），同样未发现明显的证据表明聚类簇的形成是由于相关性强弱导致的。

从显著状态的角度分析，我们以  $p_0 = 0.3$  为阈值解析出不同用户的显著状态，并研

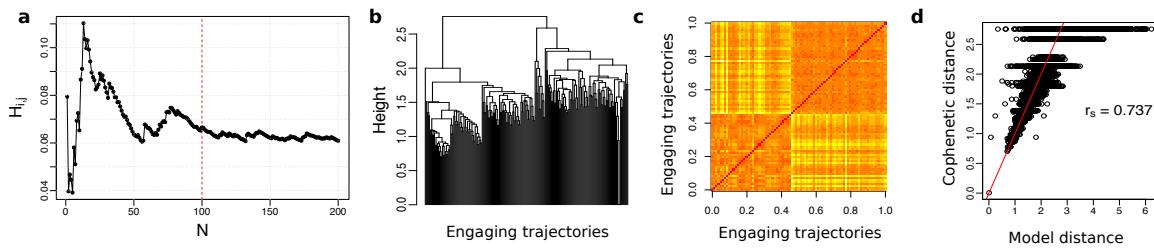


图 5-19 模型距离的快速近似计算以及基于模型的参与轨迹聚类示例

Fig 5-19 The illustration of distance calculation and model-based clustering.

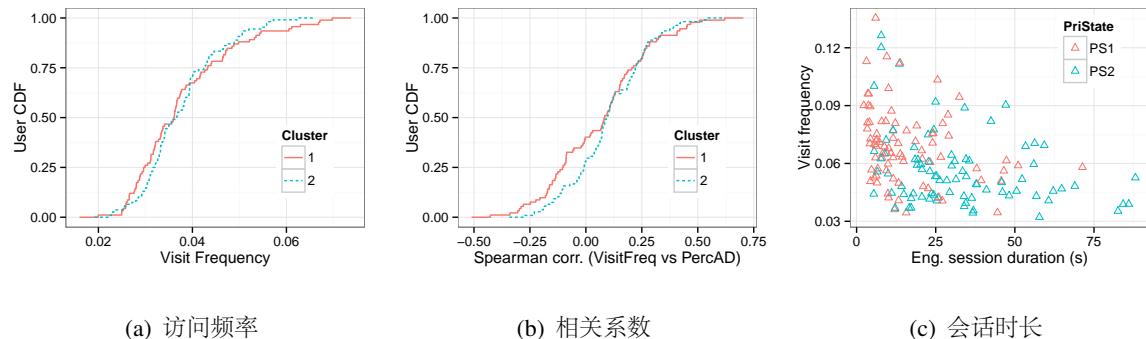


图 5-20 不同聚类簇中的用户参与行为特征分析

Fig 5-20 Investigation of trajectory clusters with the distribution of behavioural metrics in different clusters.

究了不同显著状态下的参与行为特征分布。由图5-20C可以看出，用户的参与行主要集中在两种模式之下，即较长的参与会话时长和较小的访问频率，如PS2状态，以及较短的会话时长和较大的访问频率，如PS1状态。这意味着显著状态不仅在时间维度上表现了用户的行为模式，也在群体粒度上揭示了用户行为的共性。为了检验我们的这一结论，我们展示了不同聚类簇和显著状态下的行为特征分布。如图5-21所示，由于其他特征的分布类似，我们仅展示了访问频率特征；可以看出不同簇内显著状态下的行为特征表现出较大的差异，且同一个簇内的群体参与行在不同特征维度上具有较强的一致性。具体而言，访问频次和感知等待时间特征对在簇1中表现出不同的趋势，但是在簇2中较为一致。综上所述，显著状态包含了用户参与行为的模式信息，且通过聚类的方法揭示了群体参与行为的规律。不同聚类簇的分布和性质有助于网络运营者开发出更加个性化的网络服务和盈利策略。从网络诊断的角度讲，通过识别显著状态以外的其他参与行为状态，从而获得一种被动获取用户体验状态的新途径。



0000294

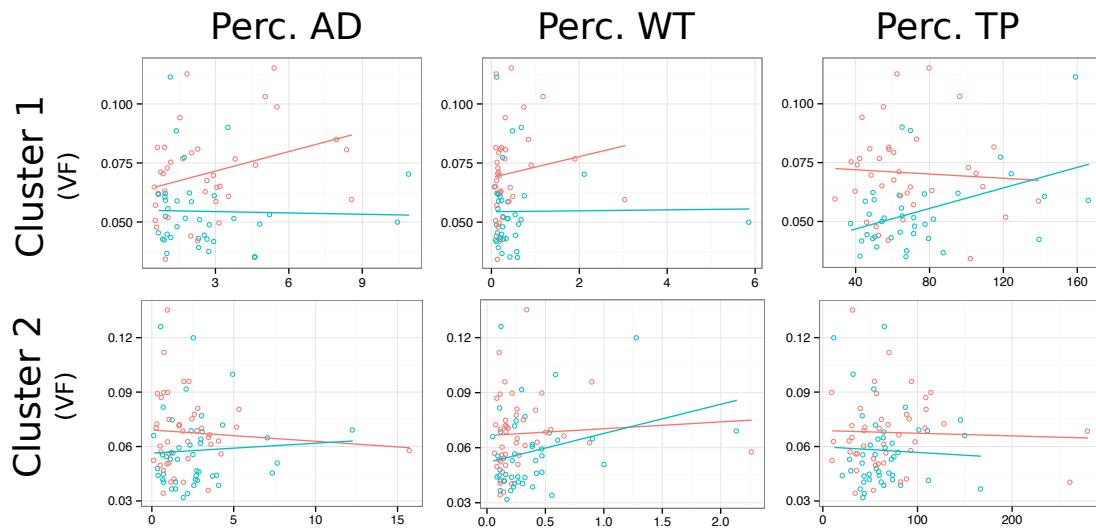


图 5-21 不同聚类簇和显著状态下的参与行为特征分布

Fig 5-21 Demonstration of the interaction between principle engaging states of different clusters.

## 5.5 本章小结

用户在网络空间的参与行为是人类时空行为的一个重要方面。本章从被动测量和量化的角度出发，对移动网络参与行为的时空模式、关联关系、以及行为模型进行了系统化的研究。在我们的分析当中存在一条潜在的假设，即移动用户无论是在有意识还是无意识情况下，用户的参与行为会根据所在的场景和当前的应用质量进行适配。由于我们采用被动的研究方法，且实际中很难对影响用户行为的所有因素进行研究，因此未能对上述假设进行全面的验证，例如，对于用户参与服务的心理状态或健康状况。这也启发我们在未来的分析中，为了获得对用户参与行为更精确的测量，可以在小范围实验中采用被动和主动测量结合的方法，同时包括其他的辅助方式对用户的主观因素进行测量，以完善模型对用户参与行为的描述能力。

尽管如此，该研究中对于用户参与行为规律的洞察和发现，在实际中有着广泛的应用。所提出的参与行为模型，为网络仿真中的流量生成、以及移动操作系统中的用户体验感知与优化提供了直接便利。由于该模型中包含了用户参与行为和应用质量之间的依赖信息，因此有助于开发出客观的移动体验量化标准。虽然本研究中的数据采集自 WiFi 网络，但是其中的移动流量模型、结构相关分析方法、以及参与轨迹序列建模等具有一般性，能够方便地迁移到 3G 和 LTE 网络当中，进行大规模的测量和研究。



0000294

## 第六章 总结和展望

### 6.1 工作总结

本文利用不同空间尺度下的移动网络数据，对人类行为中的时间和空间模式、以及时空相关性进行了实证分析和建模研究。主要贡献包括以下内容：

1) 本文采集并收集了三种不同空间尺度下的移动网络数据，包括校园 WiFi 网络、城市和国家移动网络，并基于此提出了时空数据质量的客观评估和提升方法。首先，从数据记录的准确性、采集时间的连续性、以及空间分布的合理性出发，本文提出一种结合时空数据点局部特征信息、和用户轨迹全局特征的数据质量量化方法。在该方法中，局部质量信息从单个数据点的动态和静态两个层面进行刻画，突破了传统方法中单纯对动态特征（如移动速度）的依赖，因此对数据质量的刻画更加准确。进一步，用户的移动轨迹特征代表了全局的数据质量信息，我们基于轨迹中连续采集点的时空分布异质性，并结合单个数据点的平均质量水平，从而克服了传统信息熵的方法误差较大的缺陷。

2) 从网络结构出发，本文提出了个体移动行为的介观模式，并对介观模式的提取算法、实证分析、以及一种新型的个体移动模型进行了系统性研究。本文将网络分析方法引入到个体时空行为的研究当中，结合个体移动的网络拓扑和时空属性特征，提出了个体移动行为的介观模式，形成对传统个体微观序列模式和宏观统计规律的扩展。针对介观模式的特征，本文设计并验证了一种拓扑和属性结合的模式匹配算法 *TACSim*，并结合不同个体行为的相似性，提出一种基于修剪技术的显著介观模式提取算法 *PPM*，实现了对群组用户的介观模式提取。进一步对介观模式的实证分析发现，介观模式的自距离与移动行为的结构异质性紧密相关，并表现出四种相关关系，即零模式、对数模式、线性模式、以及随机模式。基于分析得到的个体行为的介观模式，本文提出了一种鲁棒性更好的、描述个体时空行为的流涌现模型 *FEM*，该模型的优势在于摒弃了传统模型中微观和宏观统计的一致性假设，为连接微观移动模式挖掘和宏观统计分析提供了基础。

3) 本文利用三种不同空间尺度（校园、城市、国家）下的移动网络数据，对群体移动行为的时间和空间关联性进行了实证分析和建模研究。首先，我们利用协方差方程



0000294

对群体的时空依赖关系进行描述，分别从时间和空间维度对群体行为的统计特征进行度量。这样的建模方法既包含空间上的分布特征，也包括时间上的时律性。我们发现，国家尺度上的资源分布和校园尺度上的群体构成，对人群时空分布具有相似的影响，而城市尺度上的区域功能差异则表现出不同的影响特征。在较大空间尺度下（如城市和国家），人群聚集度较高的区域动态变化范围反而相对较小。基于所观测到的群体时空关联关系，在考虑空间不同区域差异性的前提下，本文提出了基于盖内特分布的群体行为模型，并利用城市尺度下的人群分布预测对模型性能进行了验证和分析。实验结果证明，融合了时空关联信息的模型，在不同观测时间段内均表现出较好的预测性能，且预测准确度提高了约 3.7%~23.6%。

4) 本文提出一种被动的用户行为识别方法，对用户参与行为进行结构化分析，并结合场景因素进行建模研究。将物理空间的移动行为和网络空间的参与行为在形式上进行了统一，利用服务类型序列代替空间位置序列，对移动用户的参与行为模式进行挖掘。针对移动用户参与网络服务的时空行为，提出一种基于被动测量的行为识别算法  $\mathcal{AID}$ 。该算法充分利用网络访问请求之间的逻辑约束条件，克服了移动网络流量引用关系缺失的挑战。通过与客户端采集的基准数据进行比较，算法的识别准确度比已有的流结构算法提高了 10% 以上。通过量化用户参与行为的重要指标，建立了参与行为和底层网络性能之间的联系；进而提出了一种结构相关性分析的方法，对场景因素（即用户个性、应用类型、地点熟悉度等）如何用户的参与行为进行了细粒度的量化分析。最后，基于对参与行为时空特性的研究，提出了利用隐马尔可夫过程的参与行为建模，并对群体参与行为进行了聚类分析。

## 6.2 工作展望

人类的时空行为分析是一个新兴领域，当前的大数据技术、以及不断发展的数据挖掘工具和算法，都对这个领域的发展起到了促进作用。虽然本文从时间和空间关联关系的角度，对个体和群体的移动行为、网络参与行为等进行了研究，但是其中依然尚有许多有趣的问题值得进一步研究，具体包括：

1) 个体介观模式挖掘在城市数据分析中的应用。本文提出的用户介观模式挖掘算法及相关规律，有助于将个体行为模式的分析引入到城市数据的分析当中，如交通网络



0000294

优化和城市结构治理的。介观模式分析能够解释城市内不同区域之间的时空连通性；如果结合更多的外部特征，如位置兴趣点 POI，我们便能够从功能和用户需求角度出发，检测出城市内具有相似功能的区域，以及揭示交通路网上拥堵状况的潜在传播方式。

2) 另一项有趣的研究是，个体行为模式发生变化的成因及演化分析，从而打破个体行为模式和宏观统计特征之间的屏障。虽然宏观统计规律更加有助于我们人类理解，但是由于分析中信息丢失太多，导致其物理意义变得模糊。因此利用个体行为模式的成因分析，能够从更长的时间尺度和更大的空间范围，对宏观统计规律进行解读。而当前这方面的理论方法，以及计算机算法研究依然存在较大的不足。

3) 用户行为和场景信息的关联分析。本文的主要特色是充分利用时间和空间特征上的相关性，虽然结合了用户所在的场景信息，但由于数据集限制，相关研究尚需要作进一步扩展。随着当前越来越多的穿戴式、便携式设备得到普及，精细化的场景信息将被更多地获取。而丰富的场景化应用需求建立在一个基本的科学问题之上，即用户的时空行为规律和场景因素的交互机制。这一领域问题的研究，也将有助于人工智能技术更好地识别用户所在场景，从而为我们的现实世界带来更大的想象空间。



0000294



0000294

## 参考文献

- [1] A. Chaintreau, Pan Hui, J. Crowcroft *et al.* “*Impact of Human Mobility on Opportunistic Forwarding Algorithms*”. *IEEE Transactions on Mobile Computing*, **2007**, 6(6): 606–620.
- [2] BD Dalziel. “*Human mobility patterns predict divergent epidemic dynamics among cities*”. In: *Proceeding of the royal society*, **2013**.
- [3] Amy Wesolowski, Caroline O. Buckee, Linus Bengtsson *et al.* “*Commentary: Containing the Ebola Outbreak - the Potential and Challenge of Mobile Network Data*”. *PLoS Currents*, **2014**.
- [4] M. Congosto, D. Fuentes-Lorenzo and L. Sanchez. “*Microbloggers as Sensors for Public Transport Breakdowns*”. *IEEE Internet Computing*, **2015**, 19(6): 18–25.
- [5] J. Woodcock, M. Tainio, J. Cheshire *et al.* “*Health effects of the London bicycle sharing system: health impact modelling study*”. *BMJ*, **2014**, 348(feb13 1): g425–g425.
- [6] Fereshteh Asgari, Vincent Gauthier and Monique Becker. “*A survey on Human Mobility and its applications*”. *arXiv:1307.0814 [physics]*, **2013**.
- [7] D. Brockmann and F. Theis. “*Money Circulation, Trackable Items, and the Emergence of Universal Human Mobility Patterns*”. *IEEE Pervasive Computing*, **2008**, 7(4): 28–35.
- [8] Stephen Eubank, Hasan Guclu, V. S. Anil Kumar *et al.* “*Modelling disease outbreaks in realistic urban social networks*”. *Nature*, **2004**, 429(6988): 180–184.
- [9] M. U. G. Kraemer, T. A. Perkins, D. a. T. Cummings *et al.* “*Big city, small world: density, contact rates, and transmission of dengue across Pakistan*”. *Journal of The Royal Society Interface*, **2015**, 12(111): 20150468.



0000294

- [10] Andrew J. Tatem, Youliang Qiu, David L. Smith *et al.* “*The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents*”. *Malaria Journal*, **2009**, 8(1): 287.
- [11] M. d’Aquin, J. Davies and E. Motta. “*Smart Cities’ Data: Challenges and Opportunities for Semantic Technologies*”. *IEEE Internet Computing*, **2015**, 19(6): 66–70.
- [12] U Paul. “*Understanding traffic dynamics in cellular data networks*”. *Proceedings IEEE INFOCOM*, **2011**.
- [13] Albert-László Barabási. “*The origin of bursts and heavy tails in human dynamics*”. *Nature*, **2005**, 435(7039): 207–211.
- [14] D Brockmann, L Hufnagel and T Geisel. “*The scaling laws of human travel*”. *Nature*, **2006**, 439(7075): 462–5.
- [15] Chaoming Song, Tal Koren, Pu Wang *et al.* “*Modelling the scaling properties of human mobility*”. *Nature Physics*, **2010**, 6(10): 818–823.
- [16] Marta C. González, César A. Hidalgo and Albert-László Barabási. “*Understanding individual human mobility patterns*”. *Nature*, **2008**, 453(7196): 779–782.
- [17] Alessandro Mei and Julinda Stefa. “*SWIM: A simple model to generate small mobile worlds*”. In: *INFOCOM 2009, IEEE*, **2009**: 2106–2113.
- [18] Morton Schneider. “*Gravity Models and Trip Distribution Theory*”. *Papers in Regional Science*, **1959**, 5(1): 51–56.
- [19] SA Stouffer. “*Intervening opportunities: a theory relating mobility and distance*”. *American sociological review*, **1940**, 5: 845–867.
- [20] Filippo Simini, Marta C. González, Amos Maritan *et al.* “*A universal model for mobility and migration patterns*”. *Nature*, **2012**, 484(7392): 96–100.
- [21] Jameson L. Toole, Carlos Herrera-Yaqüe, Christian M. Schneider *et al.* “*Coupling human mobility and social ties*”. *Journal of The Royal Society Interface*, **2015**, 12(105): 20141128.

- [22] Dashun Wang, Dino Pedreschi, Chaoming Song *et al.* “*Human Mobility, Social Ties, and Link Prediction*”. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, **2011**: 1100–1108.
- [23] Charu Aggarwal and Karthik Subbian. “*Evolutionary Network Analysis: A Survey*”. *ACM Comput. Surv.* **2014**, 47(1): 10:1–10:36.
- [24] Petter Holme and Jari Saramäki. “*Temporal networks*”. *Physics Reports*, **2012**, 519(3): 97–125.
- [25] Nathan Eagle, Michael Macy and Rob Claxton. “*Network diversity and economic development*”. *Science*, **2010**, 328(5981): 1029–31.
- [26] Elsa Arcaute, Erez Hatna, Peter Ferguson *et al.* “*Constructing cities, deconstructing scaling laws*”. *Journal of The Royal Society Interface*, **2015**, 12(102): 20140745.
- [27] Jameson L. Toole, Yu-Ru Lin, Erich Muehlegger *et al.* “*Tracking employment shocks using mobile phone data*”. *Journal of The Royal Society Interface*, **2015**, 12(107): 20150185.
- [28] Luís M. A. Bettencourt, José Lobo, Deborah Strumsky *et al.* “*Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities*”. *PLoS ONE*, **2010**, 5(11): e13541.
- [29] K.-I. Goh and A.-L. Barabási. “*Burstiness and memory in complex systems*”. *EPL (Europhysics Letters)*, **2008**, 81(4): 48002.
- [30] Thomas Karagiannis. “*Power law and exponential decay of intercontact times between mobile devices*”. *Mobile Computing, IEEE Transactions on*, **2010**, 9(10): 1377–1390.
- [31] Armando Bazzani, Bruno Giorgini, Sandro Rambaldi *et al.* “*Statistical laws in urban mobility from microscopic GPS data in the area of Florence*”. *Journal of Statistical Mechanics: Theory and Experiment*, **2010**, 2010(05): P05001.
- [32] James P. Bagrow and Yu-Ru Lin. “*Mesoscopic Structure and Social Aspects of Human Mobility*”. *PLoS ONE*, **2012**, 7(5): e37676.



0000294

- [33] Juyong Park, Deok-Sun Lee and Marta C González. “*The eigenmode analysis of human motion*”. *Journal of Statistical Mechanics: Theory and Experiment*, **2010**, 2010(11): P11021.
- [34] Francesco Calabrese, Zbigniew Smoreda, Vincent D. Blondel *et al.* “*Interplay between Telecommunications and Face-to-Face Interactions: A Study Using Mobile Phone Data*”. *PLoS ONE*, **2011**, 6(7): e20814.
- [35] Theus Hossmann, Thrasyvoulos Spyropoulos and Franck Legendre. “*A complex network analysis of human mobility*”. *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, **2011**: 876–881.
- [36] Chaogui Kang, Xiujun Ma, Daoqin Tong *et al.* “*Intra-urban human mobility patterns: An urban morphology perspective*”. *Physica A: Statistical Mechanics and its Applications*, **2012**, 391(4): 1702–1717.
- [37] Morgan R Frank, Lewis Mitchell, Peter Sheridan Dodds *et al.* “*Happiness and the patterns of life: a study of geolocated tweets*”. *Scientific reports*, **2013**, 3: 2625.
- [38] Nibir Bora, Yu-Han Chang and Rajiv Maheswaran; ed. by William G. Kennedy, Nitin Agarwal and Shanchieh Jay Yang. “*Mobility Patterns and User Dynamics in Racially Segregated Geographies of US Cities*”. In: *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer International Publishing, **2014**: 11–18.
- [39] Kevin S. Kung, Kael Greco, Stanislav Sobolevsky *et al.* “*Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data*”. *PLoS ONE*, **2014**, 9(6): e96180.
- [40] Chaoming Song, Z Qu, N Blumm *et al.* “*Limits of predictability in human mobility*”. *Science*, **2010**, 1018(2010).
- [41] E. Tiakas, A.N. Papadopoulos, A. Nanopoulos *et al.* “*Searching for similar trajectories in spatial networks*”. *Journal of Systems and Software*, **2009**, 82(5): 772–788.



0000294

- [42] Hoyoung Jeung, Man Lung Yiu and Christian S. Jensen. “*Trajectory pattern mining*”. In: *Computing with Spatial Trajectories*. Springer, **2011**: 143–177.
- [43] Wenjuan Gong, Xiaming Chen, Siwei Qiang *et al.* “*Trajectory Pattern Change Analysis in Campus WiFi Networks*”. In: *Proceedings of the Second ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*. ACM, **2013**: 1–8.
- [44] Huiping Cao, Nikos Mamoulis and David W. Cheung. “*Mining frequent spatio-temporal sequential patterns*”. In: *Data Mining, Fifth IEEE International Conference on*. IEEE, **2005**: 8–pp.
- [45] Fosca Giannotti, Mirco Nanni, Fabio Pinelli *et al.* “*Trajectory pattern mining*”. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, **2007**: 330–339.
- [46] D. Patel, Chang Sheng, W. Hsu *et al.* “*Incorporating Duration Information for Trajectory Classification*”. In: *2012 IEEE 28th International Conference on Data Engineering (ICDE)*, **2012**: 1132–1143.
- [47] Xihui Chen, Jun Pang and Ran Xue. “*Constructing and Comparing User Mobility Profiles*”. *ACM Trans. Web*, **2014**, 8(4): 21:1–21:25.
- [48] Yu Zheng, Quannan Li, Yukun Chen *et al.* “*Understanding mobility based on GPS data*”. In: *Proceeding of the 10th International Conference on Ubiquitous Computing*, **2008**: 312–321.
- [49] Katayoun Farrahi and Daniel Gatica-Perez. “*Discovering routines from large-scale human locations using probabilistic topic models*”. *ACM Transactions on Intelligent Systems and Technology (TIST)*, **2011**, 2(1): 3.
- [50] Raghu Ganti, Mudhakar Srivatsa, Anand Ranganathan *et al.* “*Inferring Human Mobility Patterns from Taxicab Location Traces*”. In: *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, **2013**: 459–468.



0000294

- [51] Nathan Eagle and Alex Sandy Pentland. “*Eigenbehaviors: identifying structure in routine*”. *Behavioral Ecology and Sociobiology*, **2009**, 63(7): 1057–1066.
- [52] Shao-Meng Qin, Hannu Verkasalo, Mikael Mohtaschemi *et al.* “*Patterns, Entropy, and Predictability of Human Mobility and Life*”. *PLoS ONE*, **2012**, 7(12): e51353.
- [53] CM Schneider. “*Unravelling daily human mobility motifs*”. *Journal of the Royal Society, Interface / the Royal Society*, **2013**, (May).
- [54] Dmytro Karamshuk, Chiara Boldrini and Marco Conti. “*Human Mobility Models for Opportunistic Networks*”. **2011**, (December): 157–165.
- [55] Poria Pirozmand, Guowei Wu, Behrouz Jedari *et al.* “*Human mobility in opportunistic networks: Characteristics, models and prediction methods*”. *Journal of Network and Computer Applications*, **2014**, 42: 45–58.
- [56] Michał Gorawski and Krzysztof Grochla; ed. by Dr Aleksandra Gruca, Tadeusz Czachórski and Stanisław Kozielski. “*Review of Mobility Models for Performance Evaluation of Wireless Networks*”. In: *Man-Machine Interactions 3*. Springer International Publishing, **2014**: 567–577.
- [57] Andrea Hess, Karin Anna Hummel, Wilfried N. Gansterer *et al.* “*Data-driven Human Mobility Modeling: A Survey and Engineering Guidance for Mobile Networking*”. *ACM Computing Surveys*, **2015**, 48(3): 1–39.
- [58] Anh Dung Nguyen, Patrick Sénac, Victor Ramiro *et al.* “*STEPS—an approach for human mobility modeling*”. In: *NETWORKING 2011*. Springer, **2011**: 254–265.
- [59] Injong Rhee, Minsu Shin, Seongik Hong *et al.* “*On the Levy-Walk Nature of Human Mobility*”. *IEEE/ACM Transactions on Networking*, **2011**, 19(3): 630–643.
- [60] Kyunghan Lee, Seongik Hong, Seong Joon Kim *et al.* “*SLAW: A New Mobility Model for Human Walks*”. In: *IEEE INFOCOM 2009*, **2009**: 855–863.



0000294

- [61] Wei-Jen Hsu, Thrasyvoulos Spyropoulos, Konstantinos Psounis *et al.* “*Modeling Spatial and Temporal Dependencies of User Mobility in Wireless Mobile Networks*”. *IEEE/ACM Trans. Netw.* **2009**, 17(5): 1564–1577.
- [62] Qunwei Zheng, Xiaoyan Hong, Jun Liu *et al.* “*Agenda Driven Mobility Modeling*”. *Ad Hoc and Ubiquitous Computing*, **2010**, 5: 22–36.
- [63] Eunjoon Cho, SA Myers and J Leskovec. “*Friendship and mobility: user movement in location-based social networks*”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, **2011**: 1082–1090.
- [64] Chiara Boldrini and Andrea Passarella. “*HCMM: Modelling spatial and temporal properties of human mobility driven by users’ social relationships*”. *Computer Communications*, **2010**, 33(9): 1056–1074.
- [65] Andrea De Montis, Marc Barthélemy, Alessandro Chessa *et al.* “*The structure of inter-urban traffic: A weighted network analysis*”. *arXiv preprint physics/0507106*, **2005**.
- [66] Shan Jiang, Joseph Ferreira Jr and Marta C. Gonzalez. “*Discovering urban spatial-temporal structure from human activity patterns*”. In: *Proceedings of the ACM SIGKDD international workshop on urban computing*. ACM, **2012**: 95–102.
- [67] Yuzuru Tanahashi, James R. Rowland, Stephen North *et al.* “*Inferring human mobility patterns from anonymized mobile communication usage*”. In: *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia*. ACM, **2012**: 151–160.
- [68] Pierre Deville, Catherine Linard, Samuel Martin *et al.* “*Dynamic population mapping using mobile phone data*”. *Proceedings of the National Academy of Sciences*, **2014**, 111(45): 15888–15893.
- [69] U. Schilcher, M. Gyarmati, C. Bettstetter *et al.* “*Measuring Inhomogeneity in Spatial Distributions*”. In: *IEEE Vehicular Technology Conference, 2008. VTC Spring 2008*, **2008**: 2690–2694.



0000294

- [70] M. Michalopoulou, J. Riihijarvi and P. Mahonen. “*Towards characterizing primary usage in cellular networks: A traffic-based study*”. In: *2011 IEEE Symposium on New Frontiers in Dynamic Spectrum Access Networks (DySPAN)*, **2011**: 652–655.
- [71] Dongheon Lee, Sheng Zhou and Zhisheng Niu. “*Spatial modeling of Scalable Spatially-correlated Log-normal distributed traffic inhomogeneity and energy-efficient network planning*”. In: *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, **2013**: 1285–1290.
- [72] Dongheon Lee, Sheng Zhou, Xiaofeng Zhong *et al.* “*Spatial modeling of the traffic density in cellular networks*”. *IEEE Wireless Communications*, **2014**, 21(1): 80–88.
- [73] Vasyl Palchykov, Marija Mitrović, Hang-Hyun Jo *et al.* “*Inferring human mobility using communication patterns*”. *Scientific Reports*, **2014**, 4.
- [74] R. A. Cochrane. “*A Possible Economic Basis for the Gravity Model*”. *Journal of Transport Economics and Policy*, **1975**, 9(1): 34–49.
- [75] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov *et al.* “*Models of user engagement*”. In: *User Modeling, Adaptation, and Personalization*. Springer, **2012**: 164–175.
- [76] Eric T. Peterson and Joseph Carrabis. “*Measuring the immeasurable: Visitor engagement*”. *Web Analytics Demystified*, **2008**.
- [77] Florin Dobrian, Vyas Sekar, Asad Awan *et al.* “*Understanding the impact of video quality on user engagement*”. *ACM SIGCOMM Computer Communication Review*, **2011**, 41(4): 362–373.
- [78] Muhammad Zubair Shafiq, Jeffrey Erman, Lusheng Ji *et al.* “*Understanding the Impact of Network Dynamics on Mobile Video User Engagement*”. In: *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*. ACM, **2014**: 367–379.



0000294

- [79] Athula Balachandran, Vyas Sekar, Aditya Akella *et al.* “Developing a predictive model of quality of experience for internet video”. In: *Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM*. ACM, **2013**: 339–350.
- [80] Heather L. O’Brien and Elaine G. Toms. “What is user engagement? A conceptual framework for defining user engagement with technology”. *Journal of the American Society for Information Science and Technology*, **2008**, 59(6): 938–955.
- [81] Simon Attfield, Gabriella Kazai, Mounia Lalmas *et al.* “Towards a science of user engagement (Position Paper)”. In: *WSDM Workshop on User Modelling for Web Applications*, **2011**.
- [82] S. Ickin, K. Wac, M. Fiedler *et al.* “Factors influencing quality of experience of commonly used mobile applications”. *IEEE Communications Magazine*, **2012**, 50(4): 48–56.
- [83] “ITU-T P.800. Methods for Subjective Determination of Transmission Quality - Series P: Telephone Transmission Quality; Methods for Objective and Subjective Assessment of Quality”. **1996**.
- [84] Alessandro Febretti and Franca Garzotto. “Usability, playability, and long-term engagement in computer games”. In: *CHI’09 Extended Abstracts on Human Factors in Computing Systems*. ACM, **2009**: 4063–4068.
- [85] Mikhail Afanasyev, Tsuwei Chen, Geoffrey M Voelker *et al.* “Analysis of a mixed-use urban wifi network: when metropolitan becomes neapolitan”. In: *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, **2008**: 85–98.
- [86] Ionut Trestian, Supranamaya Ranjan, Aleksandar Kuzmanovic *et al.* “Measuring serendipity: connecting people, locations and interests in a mobile 3G network”. In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, **2009**: 267–279.



0000294

- [87] Aaron Gember, Aditya Akella, Jeffrey Pang *et al.* “*Obtaining in-context measurements of cellular network performance*”. In: *Proceedings of the 2012 ACM conference on Internet measurement conference*, **2012**: 287–300.
- [88] Michele Coscia, Salvatore Rinzivillo, Fosca Giannotti *et al.* “*Optimal spatial resolution for the analysis of human mobility*”. In: *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. IEEE Computer Society, **2012**: 248–252.
- [89] Andrea Cuttone, Sune Lehmann and Jakob Eg Larsen. “*Inferring Human Mobility from Sparse Low Accuracy Mobile Sensing Data*”. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, **2014**: 995–1004.
- [90] Li Daqing, Jiang Yinan, Kang Rui *et al.* “*Spatial correlation analysis of cascading failures: Congestions and Blackouts*”. *Scientific Reports*, **2014**, 4.
- [91] Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart *et al.* “*D4D-Senegal: the second mobile phone data for development challenge*”. *arXiv preprint arXiv:1407.4885*, **2014**.
- [92] Fosca Giannotti, Mirco Nanni, Dino Pedreschi *et al.* “*Unveiling the complexity of human mobility by querying and mining massive trajectory data*”. *The VLDB Journal*, **2011**, 20(5): 695–719.
- [93] Andrew Kirmse, Tushar Udeshi, Pablo Bellver *et al.* “*Extracting Patterns from Location History*”. In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, **2011**: 397–400.
- [94] Hoyoung Jeung, Heng Tao Shen and Xiaofang Zhou. “*Mining trajectory patterns using hidden Markov models*”. In: *Data Warehousing and Knowledge Discovery*. Springer, **2007**: 470–480.



0000294

- [95] Sahar Hoteit, Stefano Secci, Stanislav Sobolevsky *et al.* “*Estimating human trajectories and hotspots through mobile phone data*”. *Computer Networks*, **2014**, 64: 296–307.
- [96] Amy Wesolowski and Nathan Eagle. “*The impact of biases in mobile phone ownership on estimates of human mobility*”. *Journal of the Royal Society, Interface / the Royal Society*, **2013**, (February).
- [97] Corina Iovan, Ana-Maria Olteanu-Raimond, Thomas Couronné *et al.* “*Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies*”. In: *Geographic Information Science at the Heart of Europe*. Springer International Publishing, **2013**: 247–265.
- [98] Yu Zheng. “*Trajectory Data Mining: An Overview*”. **2015**.
- [99] Lawrence Page, Sergey Brin, Rajeev Motwani *et al.* “*The PageRank citation ranking: bringing order to the Web.*” **1999**.
- [100] Glen Jeh and Jennifer Widom. “*SimRank: A Measure of Structural-context Similarity*”. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, **2002**: 538–543.
- [101] D. Conte, P. Foggia, C. Sansone *et al.* “*Thirty years of graph matching in pattern recognition*”. *International Journal of Pattern Recognition and Artificial Intelligence*, **2004**, 18(03): 265–298.
- [102] H. W. Kuhn and Bryn Yaw. “*The Hungarian method for the assignment problem*”. *Naval Res. Logist. Quart*, **1955**: 83–97.
- [103] Leonard Kaufman. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, **2005**.
- [104] César A. Hidalgo R. “*Conditions for the emergence of scaling in the inter-event time of uncorrelated and seasonal systems*”. *Physica A: Statistical Mechanics and its Applications*, **2006**, 369(2): 877–883.
- [105] N Cressie and CK Wikle. *Statistics for spatio-temporal data*, **2011**.

- [106] Tilmann Gneiting. “Nonseparable, stationary covariance functions for space – time data”. *Journal of the American Statistical Association*, **2002**, (063).
- [107] Alex Wang. “Advertising Engagement: A Driver of Message Involvement on Message Effects”. *Journal of Advertising Research*, **2006**, 46(4): 355.
- [108] B.A Mah. “An empirical model of HTTP network traffic”. In: , *Proceedings IEEE INFOCOM '97. Sixteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Driving the Information Revolution*, **1997**: 592–600 vol.2.
- [109] Hyoung-Kee Choi and John O. Limb. “A behavioral model of web traffic”. In: *Proceedings of Seventh International Conference on Network Protocols*. IEEE, **1999**: 327–334.
- [110] Sunghwan Ihm and Vivek S Pai. “Towards Understanding Modern Web Traffic”. In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, **2011**: 295–312.
- [111] D Rossi, M Mellia and C Casetti. “User patience and the Web: a hands-on investigation”. In: *Global Telecommunications Conference*, **2003**: 4163–4168.
- [112] Junxian Huang, Qiang Xu, Birjodh Tiwana *et al.* “Anatomizing application performance differences on smartphones”. In: *Proceedings of the 8th international conference on Mobile systems, applications, and services*, **2010**: 165–178.
- [113] Xian Chen, Ruofan Jin, Kyoungwon Suh *et al.* “Network performance of smart mobile handhelds in a university campus WiFi network”. In: *Proceedings of the 2012 ACM conference on Internet measurement conference*, **2012**: 315–328.
- [114] Aaron Gember, Ashok Anand and Aditya Akella. “A Comparative Study of Handheld and Non-handheld Traffic in Campus Wi-Fi Networks”. In: *Passive and Active Measurement*, **2011**: 173–183.
- [115] Rakesh Dugad and U.B. Desai. “A Tutorial on Hidden Markov Models”. *Indian Institute of Technology*, **1996**: 16.



0000294

## 致 谢

直博六年在我看来是一段人生修行。回首走过的路、看过的风景，有收获的喜悦，也有探索的迷茫。但是，很庆幸人生能够被这一段只求温饱后、对科学问题深沉思索的时期所填满，从一个对科学世界的憧憬者，成长为利用逻辑思维和方法论、看待这个世界所充满的未知领域的好奇者。我相信好奇是探索的导师和创新的来源。读博期间，有幸见证了所在的 OMNILab 从无到有的发展过程，在好奇心的驱使下开始了对数据世界的探索。虽然经历了研究起步阶段一没硬件平台、二没数据来源的困顿，但是当从另一个角度来看的时候，读博修行本身堪比一次“创业”，一次从无到有的历练，而完成这次历练的唯一捷径便是坚持。在即将完成这一次阶段性修行之时，首先需要感谢的是内心的自我，感谢当初毅然决然地选择了博士研究这一方向，并为之努力创业，从技能和思考上都得到了极大的磨炼，未辜负大好青春时光。

虽然修行本身需要孤独和激情的陪伴，但是没有家人和爱人的支持，这样的修行也是不能尽善尽美地完成的。在社会依然飘荡些许躁动的环境下，他们没有从眼前利益出发而对我的人生选择提出质疑。也许家境需要我做出一些快速或“稳重”的选择，但是他们给了我最大限度的尊重，让我能够从心出发，追寻内心所向往的未知与不安。这样的宽容和鼓励是我所得到的最大支持。当研究进展遇到困顿或挫折之时，他们的陪伴带给我休息并重新起航的动力。他们的贡献虽然在上述研究中未能直接体现，但是对于研究成果的产生是不可或缺的。

感谢在博士研究生期间对我悉心指导的金耀辉教授，他严谨的治学治研精神让我在入门之初便有所感触，尤其在对细节问题的注意和把握上，常常告诫我辈，“魔鬼在细节”。虽然我们在科研方向和研究问题上也有过争执，但是对科学的研究的执着、以及为 OMNILab 科研实力提高的出发点是一致的。金耀辉教授活跃的思维和广泛的信息面，让我辈不但能够及时接触到研究的最前沿，同时激发了更多对新问题和新角度的思考。在学术交流上，鼓励多走出去交流、少些闭门造车，让我辈有更多的机会和胆识拿出自己的成果，与国际上领域内的其他小组进行切磋。最后，感谢他不遗余力地找来多样的



0000294

研究数据，从而让我们能够进行交叉领域、多尺度多角度的时空行为研究。

感谢网络信息中心的罗萱博士和王永坤博士，他们都是实践精神的践行者，常常鼓励我在理论分析的基础上，多动手实践，利用客观的实验和数据分析问题。他们在数据中心网络、数据处理系统上有着深厚的积累，在历次的研究问题讨论、论文手稿修改上都给予了我很大的帮助。感谢网络信息中心的姜开达老师，他平时为人平和，工作中却一丝不苟，在网络安全领域造诣颇深。本研究中城市和校园尺度上的网络数据，在姜开达老师的协助和努力之下，才能够得以完整地采集。在此过程中的算法设计、网络安全管理等方面，他慷慨地分享了已有的软件工具和安全管理经验，让我受益良多。

感谢 OMNILab 的所有同学，他们热情积极的讨论，让我在研究过程中受益匪浅。尤其在网络数据采集和研究过程中，从底层硬件配置、网络管理，到上层数据处理平台的维护、数据分析工作等，都在我遇到问题的时候给与了热情的帮助和解决。感谢区域光纤通信网与新型光通信系统国家重点实验室的诸位老师，他们精心组织的学术交流和团队活动，既让我们感受到了科研的乐趣和魅力，也让我们的研究生活变得丰富多彩。

最后，对于诸多研究过程中帮助过我、启发过我的人们，深深道一声，感谢！



0000294

## 攻读学位期间发表的学术论文

- [1] **CHEN, XIAMING**; QIANG, SIWEI; WEI, JIANWEN; JIANG, KAIDA; JIN, YAOHUI. Passive Profiling of Mobile Engaging Behavior via User-End Application Performance Assessment[J]. *Pervasive and Mobile Computing (SCI)*, 2015, vol. 26: 95-112, WOS:000376728900006.
- [2] **CHEN, XIAMING**; JIN, YAOHUI; QIANG, SIWEI; HU, WEISHENG; JIANG, KAIDA. Modeling Spatio-Temporal Dependence of Cellular Traffic at City Scale[C]. *IEEE International Conference on Communications (SCI)*, 2015: 3585-3591, WOS:000371708103133.
- [3] **CHEN, XIAMING**; WANG HAIYANG, QIANG, SIWEI; WANG, YONGKUN; JIN, YAOHUI. Discovering and Modeling Meta-Structures in Human Behavior from City-Scale Cellular Data[J]. Submission to *Pervasive and Mobile Computing (SCI)*, 2016 (**minor revision**).
- [4] WANG, HAIYANG; **CHEN, XIAMING**; QIANG, SIWEI; JIN, YAOHUI. Early Warning of City-scale Unusual Social Event on Public Transportation Smartcard Data[C]. *UIC*, 2016, under print.
- [5] LI, ZHENHUAN; WANG, HAIYANG, **CHEN, XIAMING**; WANG, YONGKUN; JIN, YAOHUI. Discovering Mass Activities Using Anomalies in Individual Mobility Motifs[C]. *UIC*, 2016, under print.
- [6] GONG, WENJUAN; **CHEN, XIAMING**; QIANG, SIWEI; JIN, YAOHUI. Trajectory pattern change analysis in campus WiFi networks[C]. *Proceedings of the Second ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*, 2013, 1-8.
- [7] YANG, XIN; **CHEN, XIAMING**; JIN, YAOHUI. A high-speed real-time HTTP performance measurement architecture based on network processor[C]. *IEEE International Conference on ICT Convergence (EI)*, 2011, 744-745, .



0000294

- [8] 石开元, 陈夏明, 强思维, 王海洋, 孙莹, 金耀辉. 多源数据融合的危害公共安全事件关联关系挖掘 [J]. 计算机研究与发展, 2015.
- [9] 孙莹, 陈夏明, 王海洋, 强思维. 城市尺度下基站人群的时空预测模型 [J]. 计算机应用研究, 2016 年 12 期.



0000294

## 攻读学位期间申请的专利

- [1] 陈夏明, 金耀辉, 杨鑫, 韦建文, 叶伟, “被动网络性能测量系统及页面识别方法”,  
专利号 ZL201110186461.9, 授权日期 2012 年 06 月。



0000294



0000294

## 攻读学位期间参与的项目

- [1] 国家“973”基金项目：“Pbits 级可控管光网络基础研究：基于业务感知的光汇聚”  
(项目编号：2010CB328200)
- [2] 国家自然科学基金项目：“云计算环境中虚拟网络的性能可预见原理研究”(项目编号：61371048)
- [3] 华为公司合作研究项目：“移动网络流量挖掘与人群移动性分析研究”(实施时间：  
2012 年 1 月至 2013 年 1 月，负责移动用户的参与行为分析研究)
- [4] 上海电信合作研究项目：“智能管道中基于超文本传输协议的业务质量与客户体验  
研究与评估系统开发”(实施时间：2011 年 1 月至 2011 年 12 月，负责行为测量算  
法设计和研发)

# 上海交通大学

## 学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权上海交通大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于

保 密 ，在 \_\_\_\_\_ 年解密后适用本授权书。  
不保密

(请在以上方框内打√)

学位论文作者签名: 陈夏明

指导教师签名: 张伟华

日期: 2016 年 8 月 22 日

日期: 2016 年 8 月 22 日



0000294

## 上海交通大学 学位论文原创性声明

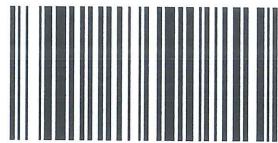
本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确式标明。本人完全意识到本声明的法律结果由本人承担。



0000294

学位论文作者签名: 陈夏明

日期: 2016 年 8 月 22 日



# 上海交通大学博士学位论文答辩决议书

0100349025

姓名	陈夏明	学号	0100349025	所在学科	信息与通信工程
指导教师	金耀辉	答辩日期	2016-08-22	答辩地点	校本部华山路校区浩然高科技大厦4楼会议室
论文题目	利用移动网络数据的人类时空行为分析及建模研究				

投票表决结果: 5/5/5 (同意票数/实到委员数/应到委员数) 答辩结论: 通过 未通过

评语和决议:



0000294

2016 年 8 月 22 日

答 辩 委 员 会 成 员 签 名	职务	姓名	职称	单位	签名
	主席	王新	教授	复旦大学	<u>王新</u>
	委员	吕岳	教授	华东师范大学	<u>吕岳</u>
	委员	薛广涛	教授	上海交通大学电子信息与电气工程学院(计算机系)	<u>薛广涛</u>
	委员	肖石林	教授	上海交通大学电子信息与电气工程学院(电子系)	<u>肖石林</u>
	委员	郭薇	教授	上海交通大学电子信息与电气工程学院(电子系)	<u>郭薇</u>
	秘书	张颖(02947)	工程师	上海交通大学	<u>张颖</u>

移动网络的大数据分析近年来是信息领域的研究热点，陈夏明同学的研究论文利用不同空间尺度下的移动网络数据，对人类行为中的时间和空间模式以及时空相关性进行了建模研究和实证分析。论文选题具有创新性和先进性，有很好的应用和理论价值。

本文主要创新如下：

1. 基于多空间尺度的移动网络数据，从单数据点、单用户样本和群体观测角度，对时空行为数据质量进行量化评估，提出了针对弱数据质量的提升算法，成效显著。
  - . 从网络结构出发，提出了个体移动行为的介观模式，设计了介观模式的提取算法并进行了实证分析，具有更好的鲁棒性。
3. 考虑空间分布与时间动态特征，对校园、城市和国家三种不同空间尺度下的“潮汐效应”群体时空行为进行建模研究和人群聚集预测的实证分析，准确度有效提升。
4. 提出一种被动用户行为识别方法，对用户参与行为进行结构化分析和聚类分析，并结合场景因素进行建模研究，比已有算法明显提高了准确度。

论文结构合理，图表规范，论述清楚。答辩过程中，回答准确。表明该生具有较好的理论基础和独立从事科研工作能力。经答辩委员会无记名投票表决，一致同意通过陈夏明同学的学位论文答辩，并建议校学位委员会授予其博士学位。

王新  
2016. 08. 22