申请上海交通大学硕士学位论文

基于信息熵的数据交易定价研究

论文作者 李希君

学　　号 115037910069

导　　师 姚建国副教授

专　　业 软件工程

答辩日期 2014 年 12 月 17 日

Submitted in total fulfillment of the requirements for the degree of Master in Software Engineering

Information Entropy-based Data Pricing

Xijun Li

Advisor
Prof. Jianguo Yao

School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, P.R.China

Dec. 17th, 2014

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

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

基于信息熵的数据交易定价研究

摘 要

上海交通大学是我国历史最悠久的高等学府之一，是教育部直属、教育部与上海市共建的全国重点大学，是国家“七五”、“八五”重点建设和“211 工程”、“985 工程”的首批建设高校。经过 115 年的不懈努力，上海交通大学已经成为一所“综合性、研究型、国际化”的国内一流、国际知名大学，并正在向世界一流大学稳步迈进。

十九世纪末，甲午战败，民族危难。中国近代著名实业家、教育家盛宣怀和一批有识之士秉持“自强首在储才，储才必先兴学”的信念，于 1896 年在上海创办了交通大学的前身——南洋公学。建校伊始，学校即坚持“求实学，务实业”的宗旨，以培养“第一等人才”为教育目标，精勤进取，笃行不倦，在二十世纪二三十年代已成为国内著名的高等学府，被誉为“东方 MIT”。抗战时期，广大师生历尽艰难，移转租界，内迁重庆，坚持办学，不少学生投笔从戎，浴血沙场。解放前夕，广大师生积极投身民主革命，学校被誉为“民主堡垒”。

新中国成立初期，为配合国家经济建设的需要，学校调整出相当一部分优势专业、师资设备，支持国内兄弟院校的发展。五十年代中期，学校又响应国家建设大西北的号召，根据国务院决定，部分迁往西安，分为交通大学上海部分和西安部分。1959 年 3 月两部分同时被列为全国重点大学，7 月经国务院批准分别独立建制，交通大学上海部分启用“上海交通大学”校名。历经西迁、两地办学、独立办学等变迁，为构建新中国的高等教育体系，促进社会主义建设做出了重要贡献。六七十年代，学校先后归属国防科工委和六机部领导，积极投身国防人才培养和国防科研，为“两弹一星”和国防现代化做出了巨大贡献。

改革开放以来，学校以“敢为天下先”的精神，大胆推进改革：率先组成教授代表团访问美国，率先实行校内管理体制改革，率先接受海外友人巨资捐赠等，有力地推动了学校的教学科研改革。1984 年，邓小平同志亲切接见了学校领导和师生代表，对学校的各项改革给予了充分肯定。在国家和上海市的大力支持下，学校以“上水平、创一流”为目标，以学科建设为龙头，先后恢复和兴建了理科、管理学科、生命学科、法学和人文学科等。1999 年，上海农学院并入；2005 年，与上海第二医科大学强强合并。至此，学校完成了综合性大学的学科布局。近年来，通过国家“985 工程”和“211 工程”的建设，学校高层次人才日渐汇聚，科研实力快速提升，实现了向研究型大学的转变。与此同时，学校通过与美国密西根大学等世界一流大学的合作办学，实施国际化战略取得重要突破。1985 年开始闵行校区建设，历经 20 多年，已基本建设成设施完善，环境优美的现代化大学校园，并已完成了办学重心向闵行校区的转移。学校现有徐汇、闵行、法华、七宝和重庆南路（卢湾）5 个校区，总占地面积 4840 亩。通过一系列的改革和建设，学校的各项办学指标大幅度上升，实现了跨越式发展，整体实力显著增强，为建设世界一流大学奠定了坚实的基础。

交通大学始终把人才培养作为办学的根本任务。一百多年来，学校为国家和社会培养了 20 余万各类优秀人才，包括一批杰出的政治家、科学家、社会活动家、实业家、工程技术专家和医学专家，

如江泽民、陆定一、丁关根、汪道涵、钱学森、吴文俊、徐光宪、张光斗、黄炎培、邵力子、李叔同、蔡锷、邹韬奋、陈敏章、王振义、陈竺等。在中国科学院、中国工程院院士中，有 200 余位交大校友；在国家 23 位“两弹一星”功臣中，有 6 位交大校友；在 18 位国家最高科学技术奖获得者中，有 3 位来自交大。交大创造了中国近现代发展史上的诸多“第一”：中国最早的内燃机、最早的电机、最早的中文打字机等；新中国第一艘万吨轮、第一艘核潜艇、第一艘气垫船、第一艘水翼艇、自主设计的第一代战斗机、第一枚运载火箭、第一颗人造卫星、第一例心脏二尖瓣分离术、第一例成功移植同种原位肝手术、第一例成功抢救大面积烧伤病人手术等，都凝聚着交大师生和校友的心血智慧。改革开放以来，一批年轻的校友已在世界各地、各行各业崭露头角。

截至 2011 年 12 月 31 日，学校共有 24 个学院/直属系（另有继续教育学院、技术学院和国际教育学院），19 个直属单位，12 家附属医院，全日制本科生 16802 人、研究生 24495 人（其中博士研究生 5059 人）；有专任教师 2979 名，其中教授 835 名；中国科学院院士 15 名，中国工程院院士 20 名，中组部“千人计划”49 名，“长江学者”95 名，国家杰出青年基金获得者 80 名，国家重点基础研究发展计划（973 计划）首席科学家 24 名，国家重大科学研究计划首席科学家 9 名，国家基金委创新研究群体 6 个，教育部创新团队 17 个。

学校现有本科专业 68 个，涵盖经济学、法学、文学、理学、工学、农学、医学、管理学和艺术等九个学科门类；拥有国家级教学及人才培养基地 7 个，国家级校外实践教育基地 5 个，国家级实验教学示范中心 5 个，上海市实验教学示范中心 4 个；有国家级教学团队 8 个，上海市教学团队 15 个；有国家级教学名师 7 人，上海市教学名师 35 人；有国家级精品课程 46 门，上海市精品课程 117 门；有国家级双语示范课程 7 门；2001、2005 和 2009 年，作为第一完成单位，共获得国家级教学成果 37 项、上海市教学成果 157 项。

关键词： 上海交大　饮水思源　爱国荣校

Information Entropy-based Data Pricing

ABSTRACT

An imperial edict issued in 1896 by Emperor Guangxu, established Nanyang Public School in Shanghai. The normal school, school of foreign studies, middle school and a high school were established. Sheng Xuanhuai, the person responsible for proposing the idea to the emperor, became the first president and is regarded as the founder of the university.

During the 1930s, the university gained a reputation of nurturing top engineers. After the foundation of People's Republic, some faculties were transferred to other universities. A significant amount of its faculty were sent in 1956, by the national government, to Xi'an to help build up Xi'an Jiao Tong University in western China. Afterwards, the school was officially renamed Shanghai Jiao Tong University.

Since the reform and opening up policy in China, SJTU has taken the lead in management reform of institutions for higher education, regaining its vigor and vitality with an unprecedented momentum of growth. SJTU includes five beautiful campuses, Xuhui, Minhang, Luwan Qibao, and Fahua, taking up an area of about 3,225,833 m2. A number of disciplines have been advancing towards the top echelon internationally, and a batch of burgeoning branches of learning have taken an important position domestically.

Today SJTU has 31 schools (departments), 63 undergraduate programs, 250 masters-degree programs, 203 Ph.D. programs, 28 post-doctorate programs, and 11 state key laboratories and national engineering research centers.

SJTU boasts a large number of famous scientists and professors, including 35 academics of the Academy of Sciences and Academy of Engineering, 95 accredited professors and chair professors of the "Cheung Kong Scholars Program" and more than 2,000 professors and associate professors.

Its total enrollment of students amounts to 35,929, of which 1,564 are international students. There are 16,802 undergraduates, and 17,563 masters and Ph.D. candidates. After more than a century of operation, Jiao Tong University has inherited the old tradition of "high starting points, solid foundation, strict requirements and extensive practice." Students from SJTU have won top prizes in various competitions, including ACM International Collegiate Programming Contest, International Mathematical Contest in Modeling and Electronics Design Contests. Famous alumni include Jiang Zemin, Lu Dingyi, Ding Guangen, Wang Daohan, Qian Xuesen, Wu Wenjun, Zou Taofen, Mao Yisheng, Cai Er, Huang Yanpei, Shao Lizi, Wang An and many more. More than 200 of the academics of the Chinese Academy of Sciences and Chinese Academy of Engineering are alumni of Jiao Tong University.

KEY WORDS: SJTU, master thesis, XeTeX/LaTeX template

目 录

插图索引

表格索引

算法索引

主要符号对照表

v	数据商品价值
v_l	数据商品估价的下限
v_h	数据商品估价的上限
D	待出售的数据商品
$s(D)$	数据商品的大小
$H(D)$	数据商品的信息熵
i	第 i 个参加拍卖的竞拍者
a_i	第 i 个竞拍者的到达时间
b_i	第 i 个竞拍者的出价
B	竞拍者集合
v_i	拍品对第 i 个竞拍者的真实价值
λ	单位时间内竞拍者到达率
p_i	第 i 个竞拍者的支付价格
q_i	第 i 个竞拍者赢得竞拍的概率
k	单次拍卖拍出的商品数
T	竞拍持续时间
$\mathbf{R}$	竞拍规则
r	竞拍保留价
c	数据商品收集和清洗的固定成本
R_s^{online}	在线拍卖卖家的期望收益
λ_t	在线拍卖开始时间 t 后单位时间内的竞拍者到达率
α	拍卖估值上下限放大系数

第一章 研究背景

近些年来，数据量的巨幅增长，尤其是从 2005 年到 2010 年，全球产生的数据量增长了 10 倍(从 130 艾字节增长到了 1227 艾字节)，目前仍在继续增长[1, 2]。数据交易的市场也以惊人的速度发展着，预计从目前到 2020 年，大数据和其商业分析的市场规模会从 1301 亿美元增长到 2030 亿美元[3]。如今，高质量和可信赖的数据商品及其相关的分析业务有着巨量的市场需求[1, 2, 4, 5]。

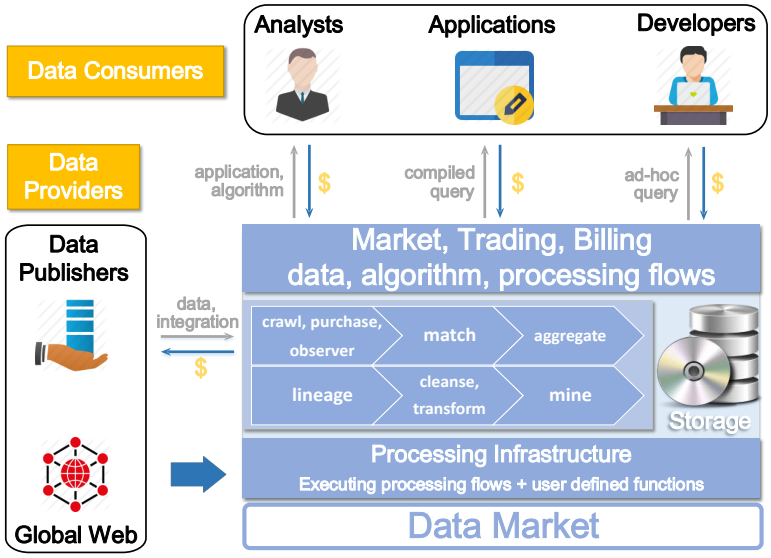


图 1–1 数据市场数据流动示意图

Fig 1–1 Data Flow in Data Market

现在，数据商品和相关分析业务主要是由在线数据市场提供的。这些市场从数据发布者和全球网络收集数据，并对其进行清洗、挖掘和整理，然后出售给不同的消费者，如图1–1所示。具体来说，数据消费者主要由开发者和中小企业主构成。这些数据消费者需要在线数据市场提供的数据和相关分析业务来帮助他们做商业决策。至于在线数据市场，他们在整理过的、有价值的数据的基础上，提供分析、商业应用和算法等服务给消费者。现在，国外主要有三家数据交易平台，分别是 Microsoft Windows Azure Data Marketplace[6], Inforchimps[7] 以及 Factual[8]。而国内也主要有三家平台，分别是贵阳大数据交易所[9]，武汉长江大数据交易中心和武汉东湖大数据交易中心[10]。然而，在这些国内外数据交易平台，并没有一个统一的定价机制来指导整个市场。因此，当前数据产品的定价处于一个比较混乱的阶段。不同的数据交易平台采用的是不同的定价机制，而其中最普遍采用的有四种种机制：基于订阅的定价机制、基于查询的定价机制，捆绑销售定价机制以及私下协商定价机制。选择什么样的定价机制具体取决于消费者使用模式以及数据提供商之间的竞争差异。但是，需要指出的是，目前没有任何一种定价机制将数据本身所含有的信息量考虑为定价因素。从消费者角度考虑，目前很多数据消费者通常只对市场上数据集的某些子集感兴趣，他们并不需要购买完整的数据集，而交

易平台往往给出的是完整数据集的价格。当消费者购买这些子集时，他们需要知道这些子集所含信息占完整数据集的比例从而评估交易平台给出的子集定价是否合理。从数据交易平台的角度看，如果他们能给出更多的子集数据价格以及它们之间的信息量关系的话，就能给消费者提供一个更加透明的定价关系，从而吸引更多的消费者进行消费。另一方面，目前已有的定价机制并不能产生最大交易剩余，这样会降低买卖双方的交易信心，从而使原本能发生的交易而没有发生，这将会对数据及其相关交易带来极大的经济损失。因此，我们进行了基于信息熵的数据交易的研究课题，期望以数据产品本身的信息量作为定价指标，然后基于这一指标探索更合适的交易机制，从而更好地促进数据交易市场的发展。

第二章 国内外研究现状

2.1 国外主流在线数据交易平台

Microsoft Windows Azure Data Marketplace[6] 在 2010 年正式成立，微软将其作为 Azure 云平台的一部分。Azure Data Marketplace 非常善于将数据买家与数据发布者联系在一起。因此，几个主要的国外数据提供商比如 ESRI, Dun and Bradstreet 都能在这个平台被找到。在 Azure Data Marketplace 上的所有数据集大致可以分为两类：免费和收费。这两种数据集都是需要消费者订阅才能有权限查阅或使用这些数据集的。对于免费的数据集，数据买家可以每月无限次数的访问它们，但在它们上定义的查询事务数量是有限的。而对于收费的数据集来说，数据买家需要根据每月访问次数来支付一定费用才能访问它们。通过使用一个标准的数据协议 Odata，这个平台给数据买家提供了一个定义良好的网络接口来访问平台中的数据。然而，对于 Azure Data Marketplace 上的数据发布者，微软并没给出具体的定价建议，尤其是多个数据发布者存在竞争的局面时。

Infochimps[7] 数据交易平台成立于 2009 年，它最初的目标是去收集尽可能多的公开的商业数据集。Factual[8] 成立于 2007 年，它主要致力于在地理信息数据的收集和售卖。与 Azure Data Marketplace 存在同样的问题是，这二者都没有为平台上的数据发布商提供一个明确的定价策略建议。在 Infochimps 和 Factual 上，数据买家在购买自己想要的数据时只能去直接联系平台上的数据发布商，经协商后确定数据价格。这么看来，数据交易平台只是起到了一个中介的作用，最终的数据交易价格并不是统一和透明的。在表2–1中，我们给出上述三个交易平台的详细比较。

表 2–1 三个国外主要数据市场的比较

Table 2–1 Comparison for three major data markets

	Azure[6]	Infochimps[7]	Factual[8]
数据类型	多种类型	主要是地理数据	主要是地理、社交以及网络数据
数据免费	是	是	—
付费数据的免费试用	是	—	是，但仅供 API 免费试用
交付方式	OData API	API, 下载	API，为重度用户提供下载
应用部署	Windows Azure	Infochimps 自建平台	—
数据发布	通过网络服务或者连接数据库	上传	上传
成立时间	2010	2009	2007

2.2 国内主流在线数据交易平台

贵阳大数据交易所[9] 成立于 2014 年 12 月 31 日，2015 年 4 月 14 日正式挂牌运营，是我国乃至全球第一家大数据交易所。秉承“贡献中国数据智慧释放全球数据价值”发展理念，志在成为全球最重要的交易所，旨在推动政府数据公开、行业数据价值发现。截至 2016 年 9 月 1 日，交易额累积突破 1 亿元，交易框架协议接近 3 亿元，发展会员超过 500 家，可交易数据产品接近 4000 个，可交易的数据总量超过 60PB。贵阳大数据交易所交易的并不是底层数据，而是基于底层数据，通过数据的清洗、分析、建模、可视化出来的结果，彻底解决了数据如何保护隐私及数据所有权的问题。贵阳大数据交易所将成为永不休市的交易所，将实行 7×24 小时的交易时间。其涉及的数据类型有金融、政府、医疗、社会、海关、能源、社交、商品、水电煤、法院、交通、企业、通信、银行卡、专利等。

武汉长江大数据交易中心是在武汉市委市政府支持下设立的、第三方中立的、具有公信力的大数据交易中心，是武汉市政府推出的“互联网 +”产业创新工程“11711”行动计划中关于大数据产业发展的重要部署。武汉长江大数据交易中心采用市场化的运作方式，以推动政府及社会各领域数据的开放、融合为宗旨，以大数据应用为导向，汇聚数据清洗、数据加工、数据咨询、数据创意等全产业链资源，沉淀数据分析技术和供需场景，将多维度数据源与业务逻辑无缝衔接，逐步构建大数据交易生态，解决数据流通困局，让大数据真正成为推动区域经济转型升级的强大动力。

武汉东湖大数据交易中心[10] 成立于 2015 年 7 月，是经武汉市政府批准成立的华中地区首家大数据交易机构，也是国内最早探索并实施” 政务数据运营解决方案” 的服务机构，注册资金 6000 万元。东湖大数据联合武汉市互联网信息办公室和武汉市国有资产管理公司制订《武汉市政务数据资产运营中心成立方案》，参与筹建” 武汉市政务数据资产运营中心”，将成为全国首个以政务行业为主的大数据运营机构，将参与相关标准、规则的制定，来促进政府数据开放，带动整个产业链的发展。交易中心由具有政府背景的——武汉国有资产经营公司；领先的数据资产运营商——中润普达；优秀的地理空间信息公司——武大吉奥，及武汉市智慧产业投资公司、汉口银行等行业领军企业共同组建。交易中心目前整合的大数据覆盖了 200 多个行业、30 大品类，实现了数千万条数据的汇集。截止 2016 年 5 月，已经服务了近百个省市区政府、金融机构、产业集团等客户。涉及到的数据集类型有交通环境、公共服务、健康医疗、金融商贸、科研应用、社交征信、科研应用、文娱音乐、知识产权、智慧生活、产业数据、政府数据。

2.3 已有的数据定价策略

基于订阅的定价机制是一个传统的数据商品定价机制。在那些实行订阅定价机制的数据交易平台，数据买家需要根据预先设定的事务数支付访问数据的费用。Azure Data Market 就是一个很好的例子来解释这种定价机制。Azure 有两种按月订阅类型：有限型和无限型。表2–2给出了 Azure Data Market 的一个具体的价目表。比如数据集 2010 Key US Demographics 就是完全受限类型。如果一个数据买家想要每个月访问这个数据集 10 次，那么他就要支付 $9.95。然而，数据集 EU Health Data Service UK 和 Business Verification 就与 2010 Key US Demographics 稍有不同了，因为他们是部分受限类型的。数据买家可以每个月以免费的价格访问 EU Health Data Service UK 5 次或者 Business Verification 10 次。一旦数据买家访问这些数据集超过相应的免费访问次数，那么他也会被收取相应的费用。虽然，数据拥有者能比较容易地采用基于订阅制的定价机制来给商品定价，但是如果订阅

价格设计的不够精密时，就会出现套利现象，从而给数据卖家带来经济损失，同时也会对其他没有套利的买家造成不公平。

表 2–2 Azure 数据市场部分商品价目表

Table 2–2 Datasets tariff in Azure data market

Dataset[6]	Transaction Limit	Price Level
2010 Key US Demographics	10	$9.95
	50	$24.95
	150	$49.95
EU Health Data Service UK	5	$0.00
	100	$293.09
	200	$732.83
Business Verification	100	$0.00
	200	$100.00
	2,500	$1,250.00
	5,000	$2,425.00

基于查询的定价机制是来源于关系数据库中的 query。最近，一些数据交易平台开始采用这种机制来售卖他们的数据集。具体来说，数据买家为其想要的数据集向数据平台发起不同的特定的请求，而数据平台返回相应数据集的视图作为查询结果给数据买家，数据卖家根据查询的复杂度收取一定的费用。比如，CustomList[11] 以 $399 售卖其全美商业数据库。数据买家可能只在意该数据集中有邮件地址的公司的数据子集，那么买家就向交易平台发起这么一个查询，交易平台查到相应视图返回结果并向买家收取 $299。然而，现在的数据交易平台并不支持复杂的查询操作，因为目前仍然并不清楚如何给不同的查询结果定一个合适的价格。Koutris[12] 等人提出了一个基于查询的数据定价框架，该框架允许卖家给一些基本视图事先赋一些价格，然后当买家发起查询时，将查询到的基本视图的价格的和作为查询费用。但是，他们的工作仍旧没有解决如何给数据集的基本视图赋予合适价格的问题。

捆绑销售定价机制是起源于资本数据市场，它代表了一种聚合技术[13]。在资本数据市场中，卖家经常将其多种产品捆绑，并对不同的客户以不同的价格销售。因此，这就会产生价格歧视效应[14]。举例来说，Dow Jones 是一个金融信息公司，它将其信息和其他在线服务（比如新闻邮件推送服务，对特定公司的新闻实行监控和过滤服务）。Dow Jones 为订阅者免费提供部分信息服务，但如果订阅者想要检索信息的全部内容，就需要开始付费了。此外，现在一些信息公司也开始实施了根据信息内容深度进行区别定价的策略。比如，Dow Jones 对新闻的标题收费 $0.20，对新闻摘要收费 $1.00，对完整的新闻收费 $3.50。需要指出的是只有当捆绑的产品具有负相关性时，捆绑销售策略才能被市场接受。对于文本信息商品，比如新闻、文章，人们能比较弄清地分辨标题、摘要和全文的关系，然后根据内容深度的不同来给它们定价。但是，现如今，大部分的信息商品是非结构化的数据，比如音频、图像和视频。数据拥有者想要辨清这些数值数据的内在关系是困难的，因此想要根据信息深度来定价是不太可行的。

2.4 商品在线拍卖

拍卖一个古老但有效的定价机制，它最早出现在公元前 500 年。大量的商品通过拍卖这一形式被交易。William Vickrey 是第一个提出系统的拍卖理论的研究者[15]。之后，在 Vickrey 的理论基础上，涌现了更多拍卖的研究工作[16–18]。Riley 等人[18] 在拍卖者独立同分布假设下研究了最优拍卖的性质。具体来说，他们比较了不同拍卖机制下卖家的期望收益。他们发现对于多数拍卖规则，如果卖家不接受低于保留价的报价时，英式拍卖或者荷兰式拍卖能最大化卖家期望收益。

互联网的快速发展使得传统商品的在线拍卖变成了可能，最早的在线拍卖出现在 1993 年[19]。目前出现了一大批在线拍卖市场，比如国外的 eBay Live Auction[1]，国内的淘宝拍卖会[2]。大多数关于在线拍卖的理论研究都集中于 B2C (Business-to-Customer) 或者 C2C (Customer-to-Customer) 的。这很大原因是研究者本身充当的角色多数是 Customer。Beam 等人[20] 使用搜索引擎分析了 100 个 B2C 和 C2C 在线拍卖实例，Riley 等人[18] 也通过类似的方式分析了 142 个 B2C 和 C2C 拍卖实例。他们两组人的研究得到了一些相似的结论，即目前在线数据拍卖只使用了传统且有限的四组拍卖机制：英式拍卖，荷兰式拍卖，第一价格拍卖，以及 Vickrey 拍卖。在文献[19] 中，Lucking 认为传统拍卖理论的假设是不适用于新兴出现的在线拍卖的。在线拍卖是非常不同于传统拍卖的。一些因素，比如拍卖持续时长、竞拍人数、拍卖起价和保留价等在传统拍卖和在线拍卖中起的作用不尽相同。Pinker 等人[21, 22] 称延长在线拍卖时长能有助于抬高成交价。Roth 等人[23] 研究发现延迟进入拍卖 (late bidding) 和狙击拍卖 (snipe bidding) 在在线拍卖场景中时常发生。Reiley 等人[24] 发现设置公开的起拍价和保留价会减少竞拍人数且会使得竞拍物品流拍的可能性增大。此外，Ariely[25] 也发现起拍价和最终成交价有着正相关关系。

大多数传统商品都是通过标的价格进行售卖的，但越来越多的物品开始通过拍卖这种机制来进行交易。在实践中，交易参与者们越来越多地认识到这种交易机制的好处。尽管大多数消费者对标的价格交易机制很熟悉而且也默认这种机制为大多数物品的交易方式，但是从交易量上来说，这是一个错误的认知。如果物品的交易成本和复杂度越高，则该物品越有可能通过拍卖的方式进行交易[21]。此外，拍卖成交价能真实反映拍卖胜者的支付意愿，这能最大化拍卖参与者的总剩余。Lu 等人[26] 总结了这些通过拍卖交易的商品的共同特性：1、唯一性；2、不确定均衡价格。数据商品恰好符合这两个特性。在目前的在线数据市场，买家的数量远远大于卖家数量，这意味着数据交易市场是一个不完全竞争市场，近似于寡头市场。Harris 等人[27] 发现当市场需求超过供应时，拍卖也许是最好的交易机制。大量事实表明拍卖也许是一个对数据商品定价能起到最佳指导的机制。近些年来，大量关于传统商品的在线拍卖的研究[19, 21, 22, 24, 25]，例如邮票、古董等。然后传统商品的在线拍卖的经验是不能直接移植到数据商品上。另一方面，因为缺乏足够多的对于数据商品的评价指标，造成数据买卖双方不能很好地对待售商品进行准确的估价。

2.5 本章小结

[1] eBay Live Auction 官方网站: `https://www.ebay.com/rpp/live-auctions`

[2] 淘宝拍卖会官方网站: `https://paimai.taobao.com`

第三章 基于信息熵的数据定价研究

3.1 研究动机

近些年来，数据量的巨幅增长，尤其是从 2005 年到 2010 年，全球产生的数据量增长了 10 倍(从 130 艾字节增长到了 1227 艾字节)，目前仍在继续增长[1, 2]。数据交易的市场也以惊人的速度发展着，预计从目前到 2020 年，大数据和其商业分析的市场规模会从 1301 亿美元增长到 2030 亿美元[3]。如今，高质量和可信赖的数据商品及其相关的分析业务有着巨量的市场需求[1, 2, 4, 5]。

现在，数据商品和相关分析业务主要是由在线数据市场提供的。这些市场从数据发布者和全球网络收集数据，并对其进行清洗、挖掘和整理，然后出售给不同的消费者。具体来说，数据消费者主要由开发者和中小企业主构成。这些数据消费者需要在线数据市场提供的数据和相关分析业务来帮助他们做商业决策。至于在线数据市场，他们在整理过的、有价值的数据的基础上，提供分析、商业应用和算法等服务给消费者。现在，国外主要有三家数据交易平台，分别是 Microsoft Windows Azure Data Marketplace[6], Inforchimps[7] 以及 Factual[8]。而国内也主要有三家平台，分别是贵阳大数据交易所[9]，武汉长江大数据交易中心和武汉东湖大数据交易中心[10]。然而，在这些国内外数据交易平台，并没有一个统一的定价机制来指导整个市场。因此，当前数据产品的定价处于一个比较混乱的阶段。不同的数据交易平台采用的是不同的定价机制，而其中最普遍采用的有四种种机制：基于订阅的定价机制、基于查询的定价机制，捆绑销售定价机制以及私下协商定价机制。选择什么样的定价机制具体取决于消费者使用模式以及数据提供商之间的竞争差异。但是，需要指出的是，目前没有任何一种定价机制将数据本身所含有的信息量考虑为定价因素。从消费者角度考虑，目前很多数据消费者通常只对市场上数据集的某些子集感兴趣，他们并不需要购买完整的数据集，而交易平台往往给出的是完整数据集的价格。当消费者购买这些子集时，他们需要知道这些子集所含信息占完整数据集的比例从而评估交易平台给出的子集定价是否合理。从数据交易平台的角度看，如果他们能给出更多的子集数据价格以及它们之间的信息量关系的话，就能给消费者提供一个更加透明的定价关系，从而吸引更多的消费者进行消费。另一方面，目前已有的定价机制并不能产生最大交易剩余，这样会降低买卖双方的交易信心，从而使原本能发生的交易而没有发生，这将会对数据及其相关交易带来极大的经济损失。因此，我们进行了基于信息熵的数据交易的研究课题，期望以数据产品本身的信息量作为定价指标来更好地指导数据交易。

3.2 定价策略与模型

在本小节中，我们首先会给出基于信息熵数据定价的问题定义。然后针对这一问题，提出了基于信息熵数据定价的通用模型，并讨论了相关的性质和应用范围。

3.2.1 问题定义

如今，大部分工业界收集到的数据都是非结构化的。即使这些非结构化的数据能用矩阵形式表示，但是人们还是很难从这些数据结构中认识到结构化的信息和信息分布。因此，我们很难真正认

识到手中数据所蕴含的价值，所以也很难为其定一个合理的价格。这里，我们给出一个的士司机行驶数据集的例子来说明这一问题。

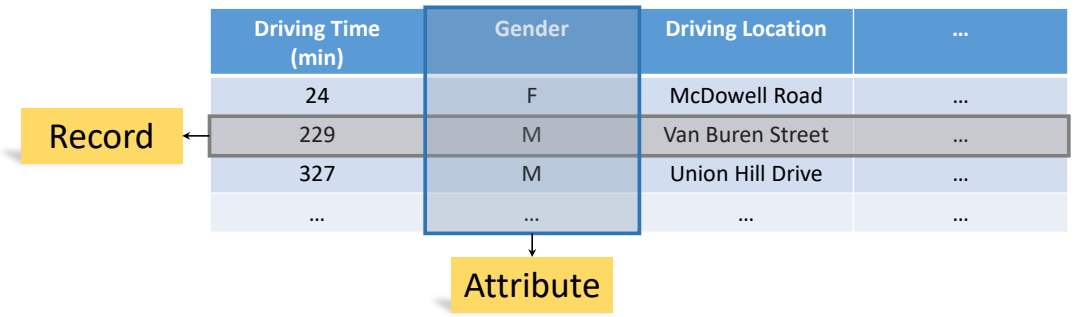


图 3–1 的士司机行车记录数据集

Fig 3–1 An illustration of taxi drivers' driving dataset

在图3–1中，我们可以看到该数据集有许多的属性，比如行车时间、司机性别以及行驶地点等等。这些属性的值可以是数值型的，也可以是文字性的。如果数据卖家想基于这些数据的信息量给这个数据集定价，那么其首要问题是弄起初这个数据集含有多少的信息量。因此，我们首要目标是去找到一个合适的方法精确度量该数据集所含有的信息量。在得到了该数据集的信息量之后，我们需要将其映射到一个合适的价格。

用更加形式化的方式叙述上述问题，对于一个有着 m 个属性和 n 条记录的数据集 $\boldsymbol{D}$，它能表示成一个矩阵 $\boldsymbol{X}$：

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}. \tag{3–1}$$

在矩阵 $\boldsymbol{X}$ 中，行向量 $\boldsymbol{r_i}$ 表示原数据集中一条记录：

$$\boldsymbol{r_i} = \begin{pmatrix} x_{i1}, & x_{i2}, & \cdots, & x_{im} \end{pmatrix}, \tag{3–2}$$

其中，$i = 1, ..., n$。令 $\boldsymbol{R}$ 表示一组记录 $\{\boldsymbol{r_{i_1}}, \boldsymbol{r_{i_2}}, ..., \boldsymbol{r_{i_k}}\}$ 的集合。类似的，列向量 $\boldsymbol{c_j^{\mathrm{T}}}$ 表示原数据集中的一列属性：

$$\boldsymbol{c_j^{\mathrm{T}}} = \begin{pmatrix} x_{1j}; & x_{2j}; & \cdots; & x_{nj} \end{pmatrix}, \tag{3–3}$$

其中，$j = 1, ..., m$。同样地，令 $\boldsymbol{C}$ 表示一组属性 $\{\boldsymbol{c_{j_1}}, \boldsymbol{c_{j_2}}, ..., \boldsymbol{c_{j_k}}\}$ 的集合。需要指出的是，$\boldsymbol{r_i}$ 中的每个元素的值可以是不同类型的，比如数值、文字、日期等。但是 $\boldsymbol{c_j}$ 的元素的值必须是同一类型的。

那么基于信息量的给数据集 $\boldsymbol{D}$ 定价的方法可以分为两步：1）量化地度量数据集 $\boldsymbol{D}$ 或其子集的信息量 H；2）基于度量结果，找到一个合适的函数 $l(\dot{})$，将信息量 H 映射到一个价格 pr，即 $pr = l(H(\boldsymbol{D}))$。

为了达到上述目标，在下文中我们首先提出一个基于信息熵的信息量检测方法。然后基于检测结果，我们给出一个定价模型。

3.2.2 数据商品信息量的测量

在本小节中，我们先定义了元组和元组集合。然后，我们基于信息熵[28] 给出了相应的信息测量方法。

定义 3.1 (元组)**.** 对于给定的一个数据集 $\boldsymbol{D}$，元组 $\boldsymbol{t}$ 被定义为 $\boldsymbol{D}$ 中一条记录 $\boldsymbol{r}$ 的非空子集，即 $\boldsymbol{t} \subseteq \boldsymbol{r}$ 且 $\boldsymbol{t} \neq \emptyset$。

定义 3.2 (元组集合)**.** 元组集合 Tup 是一系列元组 $\{\boldsymbol{t_{i_1}}, \boldsymbol{t_{i_2}}, ..., \boldsymbol{t_{i_k}}\}$ 的集合。因此，Tup 也是数据集 $\boldsymbol{D}$ 的非空子集，即 $Tup \subseteq \boldsymbol{D}$ 且 $Tup \neq \emptyset$。

对于定义3.2，元组集合 Tup 可以是数据集的子集也可以就是数据集本身。实际上，元组集合是本文提出信息测量方法的最小单元。以下四个信息熵测量指标都是基于元组的。

定义 3.3 (数据信息熵)**.** 对于一个有着 n 条元组 $\{\boldsymbol{t_i}|i = 1, ..., n\}$ 的元组集合 Tup，其数据信息熵 H_{ind} 定义为:

$$H_{ind}(Tup) = -\sum_{\boldsymbol{t_i} \in Tup} p(\boldsymbol{t_i}) \log_b p(\boldsymbol{t_i}) \tag{3–4}$$

其中，b 是公式3–4中对数的基。信息常用度量单位为比特，当对数基底 b 取为 2 时。除非特别指出，下文所有的对数都指的是以 2 为基底的对数，即 $\log_2 x$。数据信息熵是本文提出数据信息测量方法的最基础概念，它能测量出单个元组集合的信息量。

定义 3.4 (数据联合熵)**.** 对于有 n_1 条元组的元组集合 Tup_1 和有 n_2 条元组的元组集合 Tup_2，它们的数据联合熵 H_{joint} 定义为:

$$H_{joint}(Tup_1, Tup_2) = -\sum_{\boldsymbol{t_i} \in Tup_1} \sum_{\boldsymbol{t_j} \in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j}) \log p(\boldsymbol{t_i}, \boldsymbol{t_j}) \tag{3–5}$$

需要指出的是数据联合熵是可以轻易地扩展到多个元组集合的信息测量。

定义 3.5 (数据条件熵)**.** 对于有 n_1 条元组的元组集合 Tup_1 和有 n_2 条元组的元组集合 Tup_2，那么在已知 Tup_1 的条件下 Tup_2 的信息熵被定义为

$$H_{cond}(Tup_2|Tup_1) = -\sum_{\boldsymbol{t_i} \in Tup_1} \sum_{\boldsymbol{t_j} \in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j}) \log p(\boldsymbol{t_j}|\boldsymbol{t_i}), \tag{3–6}$$

其中 $H_{cond}(Tup_2|Tup_1) = 0$，当且仅当 Tup_2 被 Tup_1 完全决定。换句话说，一旦 Tup_1 已知，Tup_2 就被唯一确定了。相反的情况是，$H_{cond}(Tup_2|Tup_1) = H_{ind}(Tup_2)$ 当且仅当 $H_{ind}(Tup_1)$ 和 $H_{ind}(Tup_2)$ 是统计独立的，换句话说就是即使 Tup_1 已知，也不能获得 Tup_2 的任何信息。

定义 3.6 (数据互信息)**.** 对于有 n_1 条元组的元组集合 Tup_1 和有 n_2 条元组的元组集合 Tup_2，Tup_1 和 Tup_2 的数据互信息 I 定义为:

$$I(Tup_1; Tup_2) = \sum_{\boldsymbol{t_i} \in Tup_1} \sum_{\boldsymbol{t_j} \in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j}) \log \frac{p(\boldsymbol{t_i}, \boldsymbol{t_j})}{p(\boldsymbol{t_i}) p(\boldsymbol{t_j})} \tag{3–7}$$

相比于数据条件熵，数据互信息是被用来度量两个元组的依赖程度。

需要指出的是，元组的元素也可以是连续型数值的。然而上述定义的相关熵都是默认元组元素是离散的。以上定义可以通过将求和符号替换为积分符号后扩展到连续型元组上：

$$H_{ind}(Tup) = -\int_{\boldsymbol{t_i}\in Tup} p(\boldsymbol{t_i})\log p(\boldsymbol{t_i}), \tag{3–8}$$

$$H_{joint}(Tup_1, Tup_2) = -\int_{\boldsymbol{t_i}\in Tup_1}\int_{\boldsymbol{t_j}\in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j})\log p(\boldsymbol{t_i}, \boldsymbol{t_j}), \tag{3–9}$$

$$H_{cond}(Tup_2|Tup_1) = -\int_{\boldsymbol{t_i}\in Tup_1}\int_{\boldsymbol{t_j}\in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j})\log p(\boldsymbol{t_j}|\boldsymbol{t_i}), \tag{3–10}$$

$$I(Tup_1; Tup_2) = \int_{\boldsymbol{t_i}\in Tup_1}\int_{\boldsymbol{t_j}\in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j})\log\frac{p(\boldsymbol{t_i}, \boldsymbol{t_j})}{p(\boldsymbol{t_i})p(\boldsymbol{t_j})}. \tag{3–11}$$

然而，对于连续型变量而言，通常是很难确定其概率密度函数。另一方面，在计算机中积分的数值计算也是存在误差的。因此，计算连续型数据的数据信息熵的一般方法是将连续的输入空间切分为若干个离散的子空间，然后按照离散型变量计算其信息熵然。然而这种方法的内部误差会降低计算最后的熵的精确度，那么这就会影响定价策略。对于那些连续型、离散型数据均有的元组，其相应数据信息熵的计算方法是我们未来的工作之一。

此外，[28] 还列出一系列熵的性质，后续定价函数的讨论涉及到这些性质，因此我们简单陈述下相关性质。

性质 3.1 (数据信息熵的非负性)**.** 给定两个元组集合 Tup_1 和 Tup_2，我们有

$$H_{ind}(Tup_1) \geq 0, H_{ind}(Tup_2) \geq 0, \tag{3–12}$$

$$H_{joint}(Tup_1, Tup_2) \geq 0, \tag{3–13}$$

$$H_{cond}(Tup_1|Tup_2) \geq 0, H_{cond}(Tup_2|Tup_1) \geq 0, \tag{3–14}$$

$$I(Tup_1; Tup_2) \geq 0. \tag{3–15}$$

性质3.1的证明如下：

证明. 首先来关注数据信息熵，$H_{ind}(Tup)$，其定义如下:

$$H_{ind}(Tup) = -\sum_{\boldsymbol{t_i}\in Tup} p(\boldsymbol{t_i})\log p(\boldsymbol{t_i}) \tag{3–16}$$

由于 $0 \leq p(\boldsymbol{t_i}) \leq 1$，那么

$$\log p(\boldsymbol{t_i}) \leq 0 \tag{3–17}$$

因此，根据公式 (3–16)，$H_{ind}(Tup)$ 一定是非负的。

$H_{joint}(\cdot)$，$H_{cond}(\cdot), H_{cond}(\cdot)$ 以及 $I(\cdot)$ 的非负性同样可以采用上述类似步骤证明。

□

性质 3.2 (不同数据信息熵之间的关系)**.** 给定两个元组集合 Tup_1 和 Tup_2，我们有

$$H_{joint}(Tup_1, Tup_2) \leq H_{ind}(Tup_1) + H_{ind}(Tup_2), \tag{3–18}$$

$$H_{joint}(Tup_1, Tup_2) \geq H_{ind}(Tup_1), \tag{3–19}$$

$$H_{joint}(Tup_1, Tup_2) \geq H_{ind}(Tup_2). \tag{3–20}$$

性质3–18的证明如下:

证明. 首先从数据互信息的定义入手:

$$I(Tup_1; Tup_2) = \sum_{\boldsymbol{t_i} \in Tup_1} \sum_{\boldsymbol{t_j} \in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j}) \log \frac{p(\boldsymbol{t_i}, \boldsymbol{t_j})}{p(\boldsymbol{t_i}) p(\boldsymbol{t_j})}. \tag{3–21}$$

公式 (3–21) 能被重写为:

$$\begin{aligned} I(Tup_1; Tup_2) = & \sum_{\boldsymbol{t_i} \in Tup_1} \sum_{\boldsymbol{t_j} \in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j}) [\log p(\boldsymbol{t_i}, \boldsymbol{t_j}) \\ & - \log p(\boldsymbol{t_i}) - \log p(\boldsymbol{t_j})] \\ = & \sum_{\boldsymbol{t_i} \in Tup_1} \sum_{\boldsymbol{t_j} \in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j}) \log p(\boldsymbol{t_i}, \boldsymbol{t_j}) \\ & - \sum_{\boldsymbol{t_i} \in Tup_1} \sum_{\boldsymbol{t_j} \in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j}) \log p(\boldsymbol{t_i}) \\ & - \sum_{\boldsymbol{t_i} \in Tup_1} \sum_{\boldsymbol{t_j} \in Tup_2} p(\boldsymbol{t_i}, \boldsymbol{t_j}) \log p(\boldsymbol{t_j}). \end{aligned} \tag{3–22}$$

根据定义3.3和定义3.4，公式 (3–22) 能进一步写成:

$$\begin{aligned} H_{joint}(Tup_1, Tup_2) = & H_{ind}(Tup_1) + H_{ind}(Tup_2) \\ & - I(Tup_1; Tup_2). \end{aligned} \tag{3–23}$$

由于数据信息熵和数据互信息的非负性，我们有:

$$H_{joint}(Tup_1, Tup_2) \leq H_{ind}(Tup_1) + H_{ind}(Tup_2), \tag{3–24}$$

另一方面，根据定义3.3，定义3.4，以及定义3.5, 公式 (3–22) 能进一步写成:

$$H_{joint}(Tup_1, Tup_2) = H_{cond}(Tup_1 | Tup_2) + H_{ind}(Tup_2), \tag{3–25}$$

$$H_{joint}(Tup_1, Tup_2) = H_{cond}(Tup_2 | Tup_1) + H_{ind}(Tup_1). \tag{3–26}$$

类似地，由于数据信息熵的非负性，那么有:

$$H_{joint}(Tup_1, Tup_2) \geq H_{ind}(Tup_1), \tag{3–27}$$

$$H_{joint}(Tup_1, Tup_2) \geq H_{ind}(Tup_2). \tag{3–28}$$

□

3.2.3 通用定价模型

在本小节中，我们提出了一个基于数据信息熵的通用定价模型，而不是给出一个具体的定价函数。此外，该定价模型的相关性质和优点在本小节中也会进行讨论。

定义 3.7 (定价函数)**.** 对一个给定的数据集 $\boldsymbol{D}$，定价函数 $pr(\boldsymbol{D}) : \boldsymbol{D} \to \mathbb{R}^+$, 其中 $\mathbb{R}^+$ 非负实数。一个基于数据信息熵的定价函数是:

$$pr(\cdot) \equiv l(H(\cdot)), \tag{3–29}$$

其中 $l(\cdot)$ 是一个非递减的联系函数，它应该满足如下条件:

$$\forall x_1 \geq x_2, l(x_1) \geq l(x_2), \tag{3–30}$$

$$\forall x_1, x_2 \geq 0, l(x_1 + x_2) \leq l(x_1) + l(x_2). \tag{3–31}$$

例 3.1. 对于两个数据集 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$ 以及一个基于数据信息熵的定价函数 $pr(\cdot) \equiv l(H(\cdot))$，如果 $H(\boldsymbol{D_1}) \geq H(\boldsymbol{D_2})$，那么有 $pr(\boldsymbol{D_1}) \geq pr(\boldsymbol{D_2})$。

为了陈述简洁，上述数据信息熵的定价函数都指的是数据独立信息熵或者数据联合熵，取决于 $H(\cdot)$ 中参数个数。选择具体的定价函数要取决于具体的市场情况，这也是本文的未来工作之一。

性质 3.3 (定价函数的非负性)**.** 定价函数的输出一定是大于零的，即 $pr(\cdot) \geq 0$ 总是成立。

定价函数的非负性是显然的, 因为数据卖家在出售商品的同时还倒贴钱给数据买家。接下来, 基于数据信息熵的定价函数的*无套利*性质会被讨论，这个性质也是其最大的优点。为了详细地讨论这个性质，我们首先要介绍下一个数据集中的操作，即*连接*。

定义 3.8 (数据集连接)**.** 给定两个数据集 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$，$\boldsymbol{D_1}$ 有 n_1 条记录 $\boldsymbol{R_1}$ 和 m_1 条属性 $\boldsymbol{C_1}$，$\boldsymbol{D_2}$ 有 n_2 条记录 $\boldsymbol{R_2}$ 和 m_2 条属性 $\boldsymbol{C_2}$，$\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$ 的连接 $\boldsymbol{D_J}$ 定义为:

$$\boldsymbol{D_J} = \boldsymbol{D_1} \circledcirc \boldsymbol{D_2}. \tag{3–32}$$

这里给出数据连接操作的两个例子:

情形 1. 如果在数据集 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$ 中没有共同的属性，即 $\boldsymbol{C_1} \cap \boldsymbol{C_2} = \emptyset$，那么它们的连接 $\boldsymbol{D_J}$ 就会有 $n_1 + n_2$ 条记录和 $m_1 + m_2$ 条属性。

情形 2. 如果在数据集 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$ 中有共同的属性，即 $\boldsymbol{C_1} \cap \boldsymbol{C_2} \neq \emptyset$，那么它们的连接 $\boldsymbol{D_J}$ 就会有 $n_1 + n_2$ 条记录和 $\|\boldsymbol{C_1} \cup \boldsymbol{C_2}\|$ 条属性。

数据集*连接*的操作能被扩展到多个数据集的情况。如果有多个数据集 $\boldsymbol{D_1}, \boldsymbol{D_2}, ..., \boldsymbol{D_k}$，我们把他们的连接集合记为 $\boldsymbol{D_J} = \circledcirc_{i=1}^{k} \boldsymbol{D_i}$。值得注意的是这里的连接操作不同于 SQL 类的数据库连接操作。为了更加直观地说明这个连接操作，我们举出了如下例子:

例 3.2. 数据集 $\boldsymbol{D_1}$ 有 3 条记录和 3 个属性 $\{a, b, c\}$，数据集 $\boldsymbol{D_2}$ 有 4 条记录和 2 个属性 $\{c, d\}$。它们的连接数据集 $\boldsymbol{D_1} \circledcirc \boldsymbol{D_2}$ 将会有 7 条记录和 4 个属性 $\{a, b, c, d\}$，如图3–2所示。在连接数据集中，那些非共同属性的缺失值是用'#' 填满。

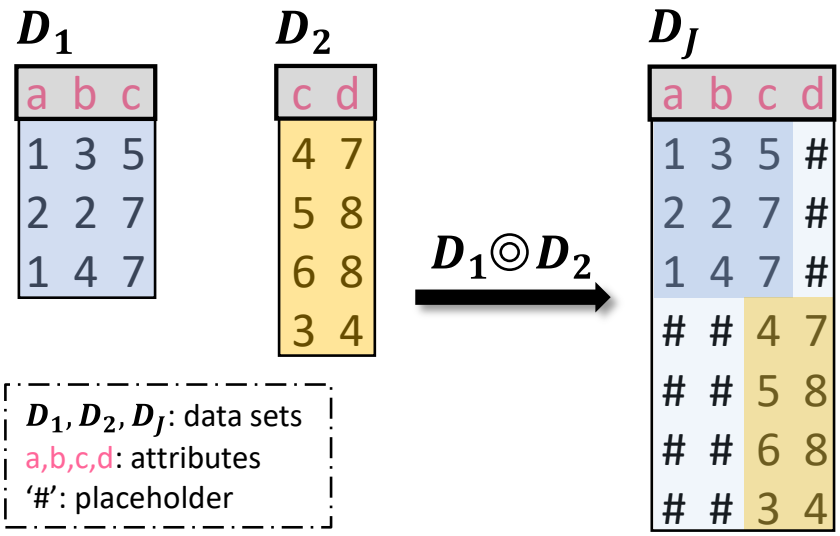


图 3–2 数据集连接操作示意图

Fig 3–2 An illustration of dataset join

性质 3.4 (定价函数的无套利性)**.** 对于数据集 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$，如果 $\boldsymbol{D_1} \circledcirc \boldsymbol{D_2}$ 的价格小于或等于 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$ 的简单求和，那么定价函数就是无套利的，即：

$$pr(\boldsymbol{D_1} \circledcirc \boldsymbol{D_2}) \leq pr(\boldsymbol{D_1}) + pr(\boldsymbol{D_2}). \tag{3–33}$$

在经济学和金融学中，套利指的是通过不同市场之间的价格差异获得利润差的行为。具体说来就是套利者可以通过在某个市场低价购入物品，然后在另一个市场高价卖出该物品。然而，这个概念在金融市场和数据市场是有些许差异的。现在假设有三个数据集 $\boldsymbol{D_1}$，$\boldsymbol{D_2}$ 和 $\boldsymbol{D_3}$，相应的价格分别为 $p_1, p_2,$ and p_3，如果数据集 $\boldsymbol{D_3}$ 可以通过连接 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$ 得到，即 $\boldsymbol{D_3} = \boldsymbol{D_1} \circledcirc \boldsymbol{D_2}$。当 $p_3 \geq p_1 + p_2$，套利的条件就产生了。一个精明的数据消 p_3 费者可以通过分别购买数据集 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$ 以一个相对原价 p_3 更低的价格 $p_1 + p_2$ 获得数据集 $\boldsymbol{D_3}$。因此，一个好的定价函数应当是能避免套利行为的。

引理 3.1. *对于定价函数 $pr(\cdot) \equiv l(H(\cdot))$，这里 $l(\cdot)$ 是一个非递减的联系函数。如果 $l(\cdot)$ 满足公式 (3–30) 和公式 (3–31)，那么 $pr(\cdot)$ 是无套利的。*

证明. 给定两个数据集 $\boldsymbol{D_1}$ 和 $\boldsymbol{D_2}$，以及一个定价函数 $pr(\cdot)$，$pr(\cdot) \equiv l(H(\cdot))$, 其中 $l(\cdot)$ 是非递减的联系函数。

根据性质 3–18, $H(\boldsymbol{D_1})$, $H(\boldsymbol{D_2})$ 和 $H(\boldsymbol{D_1}, \boldsymbol{D_2})$ 的关系如下：

$$H(\boldsymbol{D_1}, \boldsymbol{D_2}) \leq H(\boldsymbol{D_1}) + H(\boldsymbol{D_2}), \tag{3–34}$$

$$H(\boldsymbol{D_1}, \boldsymbol{D_2}) \geq H(\boldsymbol{D_1}), \tag{3–35}$$

$$H(\boldsymbol{D_1}, \boldsymbol{D_2}) \geq H(\boldsymbol{D_2}). \tag{3–36}$$

因为 $pr(\cdot) \equiv l(H(\cdot))$ 和 $l(\cdot)$ 满足公式 (3–30) 和公式 (3–31)，所以我们有：

$$pr(\boldsymbol{D_1} \odot \boldsymbol{D_2}) \leq pr(\boldsymbol{D_1}) + pr(\boldsymbol{D_2}), \tag{3–37}$$

$$pr(\boldsymbol{D_1} \odot \boldsymbol{D_2}) \geq pr(\boldsymbol{D_1}), \tag{3–38}$$

$$pr(\boldsymbol{D_1} \odot \boldsymbol{D_2}) \geq pr(\boldsymbol{D_2}). \tag{3–39}$$

因此，定价函数 $pr(\cdot)$ 是无套利的。

□

无套利的性质是基于数据信息熵的定价函数最好的一个性质，这也是之前很多研究[12, 29, 30] 所追求的目标。

3.2.4 讨论

理论上，我们提出的信息量测量方法能被应用到任何类型的数据。但是，实际中只有全离散型或者全连续型能被精确地度量。因为在这种情形下，是相对容易去找到其概率密度函数的。测量混合型数据的信息量是我们未来的工作之一。

关于本小节中提出的基于数据信息熵定价函数，其最大的优点就是其无套利性质。之前的一些基于查询的数据定价的研究[12, 29, 31] 都是致力于使他们提出的定价函数变成无套利的。论文[31] 提出了基于查询的数据定价的原型，在该文中一个关于在线数据市场如何计算查询价格的简单例子被提出来了。尽管后续的研究者基于论文[31] 提出了不同的定价函数，但是他们仍然没有解决如何为查询的最小单元（视图）定价的问题，这些最小单元的定价仍然是主观定价的。我们提出的数据定价指标——数据信息熵，也许能提供一个新的视角去解决论文[12, 29, 31] 所遗留的问题。因为数据信息熵能够为数据定价者提供不同视图之间的明确的定价关系，以避免套利行为的发生。

3.3 实验与评估

为了推动数据市场的经营者采用本文提出的数据信息熵来指导他们的商品定价，我们首先需要验证提出的数据定价指标的合理性和有效性。我们从两个方面评估数据信息熵：1）数据集数据信息熵和该数据集大小的关系。根据大量数据交易记录，一个数据集的信息量大小一般是与该数据集的大小成正比。如果数据信息熵能度量一个数据集的信息量，那么它首先应是数据集大小的非递减函数；2）数据集的数据信息熵和和在该数据集上分类器正确率的的关系。根据大量机器学习实验的经验，经过一个数据集训练的分类器的分类正确率是与该数据集的有效信息量是有关的。更加细一点地说，如果有更多的有效信息被分类器学到了，那么该分类器之后的分类正确率就会更高。

第四章 基于信息熵的数据商品在线拍卖的研究

4.1 研究动机

随着近些年人们产生的数据量的巨量增长，关于数据交易的市场雏形可见，并可期成为一个巨大规模的市场。然而，现有的关于传统商品的交易和定价策略是不适用于数据商品的，因为这些交易和定价策略均不能最大化交易双方的总剩余，从而打击交易的积极性和信心。而数据，这一新商品，目前其交易场景多为线上。从而，交易双方的信心和积极性是促成数据交易的关键因素。目前，在这个数据交易市场上，数据买家远多于卖家，使得这个市场近似于一个寡头竞争市场。基于以上两点，我们发现拍卖可能会是一个很好的数据商品定价和交易的机制。一是因为其能最大化交易双方的总剩余，二是因为拍卖机制中本来就有为这种寡头竞争市场专门设计的规则。但这不意味着我们可以直接把为传统商品设计的拍卖机制直接应用到数据商品上，因为数据商品和传统商品有很大的区别。相比于传统商品，数据商品具有唯一性和不确定均衡价格。

在众多关于拍卖的研究文献中[15–18]，拍卖物的价值 v 被认为是服从一个均匀分布，即 $v \sim U[v_l, v_h]$，其中 v_l 是估价的下限，v_h 是估价的上限。但是这些文献均没有给出如何估计上下限的指导方法，而估值上下限却是实施拍卖机制的基础。相比于传统商品，数据商品更无形，更加抽象，因此更难以被普通消费者估计它们的价值。尽管现在已经有一些关于数据商品的评价指标，比如数据集大小，数据集生成时间等，但这仍是远远不够的。我们需要找一个更合适的指标来评估数据集，从而指导拍卖参与者对数据商品的估价。目前，Li 等人[32] 提出了数据信息熵这一新的指标，用来测定给定数据集的信息量和清晰地描述数据集内的信息分布，在他们文中还提出了一系列基于该指标定价的好处。数据信息熵可能是一个非常合适的帮助数据商品估价的指标，我们期望先用数据信息熵测定给定数据商品的信息量，然后基于它给出该商品的估值上下限，从而实施拍卖机制。

4.2 影响在线拍卖的重要因素

文献[21] 提出了一些影响在线拍卖的重要因素，但是没有真实拍卖交易数据的支持。因此，我们收集了 2012 年 12 月到 2013 年 eBay Live Auction 的一些交易记录，希望以此佐证文献[21] 提出的这些影响因素。这些数据集包含了 Cartier 手表，Palm Pilot M515 掌上智能机，Xbox 游戏终端以及 Swarovski 珍珠项链在这段时期的相关拍卖信息。具体来说，它包含了相关的九个变量，即拍卖入场时间，拍卖持续时长，成交价等。在这个数据集中，总共有三种类型的拍卖，分别是时长为 3 天的拍卖，时长为 5 天的拍卖和时长为 7 天的拍卖。为了能挖掘拍卖价格，入场时间以及拍卖持续时长的关系，我们将该数据集中部分记录进行了可视化，如图4–1。这些可视化帮助我们更加清晰地认知这些因素是如何影响拍卖的。以下是我们在设计数据商品在线拍卖模型时最关心的三个因素。

4.2.1 拍卖机制

第一价格拍卖、第二价格拍卖、英式增价拍卖、荷兰式减价拍卖是目前用的最多也是被研究的最多的四种拍卖机制。其中，英式增价拍卖在在线网络拍卖中是用得最多的，因为它是顾客最熟悉、

接触最多的拍卖机制。目前，最大的在线拍卖网站，eBay Live Auction 采用的正是英式拍卖。eBay Live Auction 已成功吸引了数以百万计的人参与他们的在线拍卖，这从侧面佐证了英式增价拍卖的可用性。

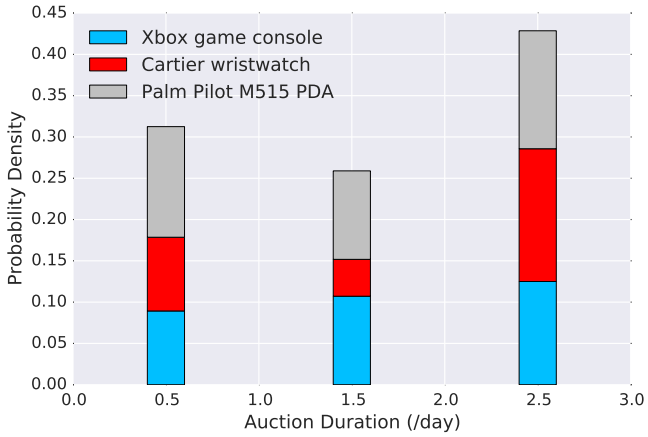
然而，Milgrom 等人[16] 提出英式拍卖机制容易受到不同欺诈行为的影响。Pinker 等人[21, 22] 认为一个好的拍卖机制必须居然三个特指：1）对不同顾客行为鲁棒的；2）分配是要高效的；3）要能有效应对欺诈行为的。设计一个最优拍卖仅仅是存在于理论中的，因为最优拍卖机制中的许多假设在现实中是难以满足的。另外，理论上的最优拍卖机制在市场中是难以实现的。Myerson[33] 在上述四个拍卖机制下提出了收益均衡定理。Lucking-Reiley[34] 通过对邮票的在线拍卖的实验，比较了英式拍卖和第二价格拍卖的收入产生。他们的目标是去验证 Myerson 的收益等价定理在英式拍卖和第二价格拍卖下是否成立。他们发现，这二者的最终收益的差别是非常小的，因此我们也认为在线英式拍卖的理论经验是可靠的。

大量关于在线拍卖的研究是集中在前向拍卖 (forward auction)，即是一个卖家向多个买家拍卖物品。以上提到的四个拍卖机制均是前向拍卖。然而，还是存在另外一大类拍卖，反向拍卖 (reverse auction)，即一个买家需要买一件物品，多个卖家以竞争的方式提供该物品，这个场景多出现在政府项目招标中。然而，对于数据的在线拍卖，是不太适合使用反向拍卖作为拍卖机制的，因为目前市场上并没有足够多的数据提供商去满足众多数据买家的需求。

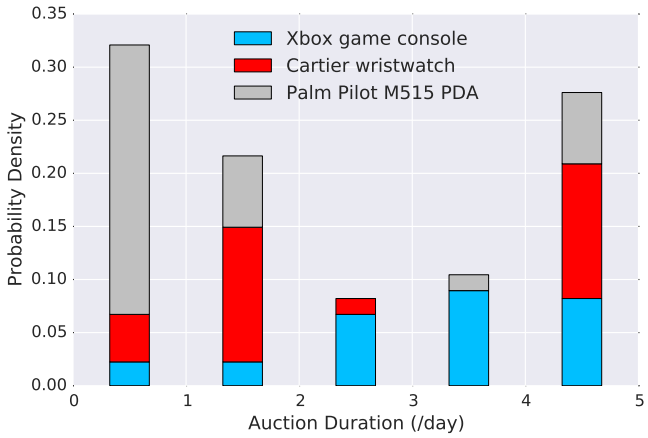
4.2.2 竞拍时长与参与人数

竞拍时长在在线拍卖中有着很大的作用。相比于传统的只持续数十分钟或数个小时的传统公开拍卖，在线拍卖往往会持续数天或数周。因为在线拍卖不会要求竞拍人到场，他们能在任何地点任何时间参与拍卖[21]。因此，持续时间很长的在线拍卖是可行的。传统拍卖往往会以一个固定的竞拍人数开始拍卖，而相应的关于期望收益和最优保留价的分析都是基于这固定的竞拍人数的。但是，同样的分析方法就不适用于在线拍卖。因为，在网络上，在线拍卖的人数是不固定的，而在线拍卖的竞拍人数往往是由竞拍时长决定的。

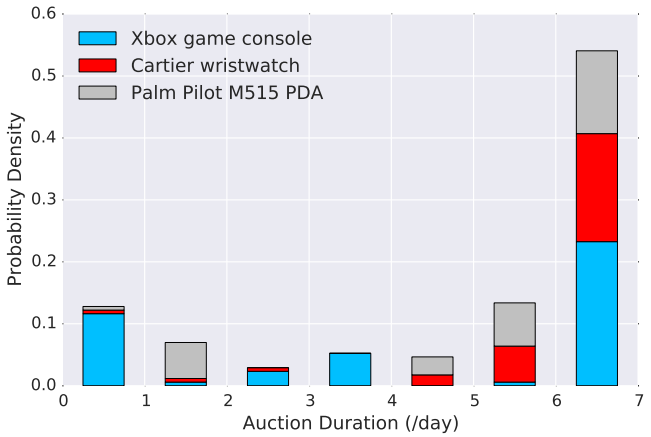
目前，在在线拍卖市场上有两大类方法终止拍卖，分别是硬结束时间 (hard deadline) 和软结束时间 (”going, going, gone”)。前一个，硬结束时间被 eBay Live Auction 采用：超过了硬结束时间的竞拍是不被接受的。后者，软结束时间被 Amazon Auction 使用[35]，在 Amazon 的在线拍卖市场上，拍卖持续时间会被自动延长只要在结束时间之前有新的竞拍。Roth 和 Ockenfels[23] 发现延迟竞拍 (late bidding) 和狙击竞拍 (snipe bidding) 时常发生在硬结束时间的在线拍卖中，比如 eBay Live Auction。我们从 eBay Live Auction 收集到的真实买拍数据中也发现了这个现象。图4–1是在三种不同持续时长下的三种不同物品的在线拍卖。可以从这三幅子图中看出，竞拍者大多数会“挤”在拍卖的最后一天参与拍卖。Pinker[21] 称当在线拍卖网站流量很低的时候，延长拍卖在一定程度上可能有助于提高成交价。然后，这并不意味着可以一味地延长拍卖时间。Ariely 等人[25] 却认为即使更短的拍卖持续时间可能会只吸引更少的竞拍者，但是这样能增加竞拍者之间的竞争激烈度。他们在自己的实验中记录到拍卖时长是与成交价呈负相关关系。在本文中，为了更好地对在线拍卖竞拍者参与过程进行动态建模，我们将竞拍人数考虑成为拍卖持续时间的分段函数。



(a) 持续时长为 3 天的拍卖



(b) 持续时长为 5 天的拍卖



(c) 持续时长为 7 天的拍卖

图 4–1 竞拍参与人数分布直方图

Fig 4–1 Number of bidders distribution histogram

4.2.3 起拍价与保留价

起拍价可以理解成为另一种形式的保留价，通常是要么公开，要么保密。起拍价迫使竞拍者需要竞拍时高于起拍价。而保留价是拍卖者为成交价设定的一个下限。如果竞拍结束时，最高竞拍价还低于保留价的话，那么卖家有权利撤销这次拍卖。

从大量的拍卖记录中，我们发现如果竞拍者人数不够的话，那么最终成交价就会比较低。而竞拍者人数不仅仅与竞拍持续时间也与起拍价相关。Vakrat 等人[22] 经过分析实例发现当没有起拍价限制时，拍卖品会以低于相应零售商给出的价格很多的成交价被拍出。当起拍价增加时，成交价很大可能会趋近于起拍价。如果起拍价进一步增加，拍卖流拍的可能性会骤增。Reiley 等人[24] 发现设置了公开的起拍价和保留价后会减少竞拍人数，并且也会增加拍卖品流拍的可能性。同样地，Ariely 等人[25] 发现了起拍价和成交价之间具有正相关关系。即使一个很低的起拍价会吸引更多的竞拍者，但是竞拍者的出价会很低，不足以在拍卖者之间产生价格竞争。此外，通过大量的实例调查，Bradlow 等人[36] 发现起拍价与竞拍时间的增幅是具有负相关关系的。这也意味着一个低的起拍价会在短时间吸引更多的竞拍者。在本文提出的在线拍卖模型中，我们将起拍价就认为是保留价，以期望去探索发现保留价和卖家期望收益的关系。

4.3 基于信息熵的在线拍卖模型

在本小节中，我将陈述基于信息熵的在线拍卖模型。首先，用 λ_t 表示在线拍卖开始时间 t 后单位时间内的竞拍者到达率。值得指出的是，λ_t 是与许多因素相关的，比如拍卖持续时间、当前参与拍卖人数、当前最高价以及竞拍规则。当然，肯定还有其他因素影响着在线拍卖。在真实情况中，在时间段 $[t, t+\delta t]$ 内到达的竞拍者人数是一个随机变量。Vakrat 和 Seidmann[37] 发现在线拍卖中新竞拍者的到达过程与指数分布非常相似。为了简化竞拍者的到达过程，同时也为了更好的抓住在线拍卖的动态特征，Chen 等人[38] 将在线拍卖的竞拍者到达过程拟合成了一个泊松分布。该泊松分布的参数设为 λ，可认为是在线拍卖中的竞拍者到达率。具体的，N 个竞拍者在时间长度 t 内到达的概率质量函数被定义为:

$$P_t(N=n) = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, n = 0, 1, 2, ... \tag{4–1}$$

然而，在实际的在线拍卖中，到达率 λ 并不会全程保持不变。我们可以从图4–1发现到达率 λ_t 在在线拍卖过程中是变化的。为了比 Chen 等人提出的模型更好地抓住在线拍卖的动态特征，我们将拍卖持续时长 $[0, T]$ 切分为一组不重叠的子区间，以使得到达率 λ_t 在不同子区间内取不同的值，然而在子区间内保持不变。

在我们的模型中，在线市场的数据卖家会列出他们要出售数据商品 D 的一些指标，去帮助买家估计该商品的价值。这些指标有 $H(D)$，数据信息熵[32]；$s(D)$, 数据集的大小；或者是数据集的生成时间等等。每一个潜在竞拍者都会在参加拍卖前浏览这些信息。利用这些信息，潜在的拍卖者可以给待拍卖的商品估计出一个价值区间 $[v_l, v_h]$。因为，本文主要关注基于信息熵的在线拍卖，我们将这一估计过程定义为如下:

$$v_l \propto H(D), \tag{4–2}$$

$$v_h \propto \alpha H(D), \alpha \geq 1, \tag{4–3}$$

其中，α 是放大系数。

在在线拍卖中，我们用元组 (a_i, b_i) 分别表示竞拍者 i 的到达时刻和出价。出价 b_i 是与拍品的真实价值 $v_i \in [v_l, v_h]$ 相关的，记为 $b_i = b(v_i)$。如果在时刻 t 有 n_t 个竞拍者参与当次拍卖，对于当前时刻 t，我们将这些竞拍者记为一个集合 B。在这个集合中，对于所有的竞拍者，其到达时间 $a_i \leq t, i = 1, 2, ..., n_t$。这些竞拍者的出价构成一个出价向量 $\vec{b}_t = (b_1, b_2, ..., b_{n_t})$。

现在我们已经为时刻 t 的在线拍卖定义了两个状态变量 n_t 和 $\vec{b}_t$。如果此次拍卖中有 k 个物品要拍卖。那么在拍卖的结尾时，那么将会有一个成交价向量 $\vec{p} = (p_1, p_2, ..., p_k)$。这个成交价向量是竞拍总人数 n_T，竞拍出价向量 $\vec{b}_T$，状态变量的最终值以及拍卖规则 $\mathbf{R}$ 的函数。对于卖家而言，他的目标是最大化拍卖物品的成交价，即:

$$\max_{\mathbf{R},T} |\vec{p}(n_T, \vec{b}_T)|_1, \tag{4–4}$$

其中 $|\cdot|_1$ 是向量的 L_1 范数。拍卖规则 $\mathbf{R}$ 规定着 1）拍卖结束时商品是如何分配的;2）成交价是如何决定的；3）出价是公开还是密封的；4）是否有保留价；5）保留价是公开的还是密封的等等[21]。假设卖家为每一个竞拍者设计一个估价区间 $V_i = [v_l, v_h]$。令 $V = \prod_{i=1}^{n} V_i$，表示所有竞拍者联合估价。除此以外卖家还需要设计分配规则 $q(v)$ 对于每一个联合估价 $v \in V$。具体来说，$q(v) = (q_1(v), q_2(v), ..., q_n(v))$ 是一个概率向量，其中 $q_i(v)$ 是第 i 个竞拍者赢得拍品的概率。此外，本文还定义了一个支付向量 $p(v) = (p_1(v), p_2(v), ..., p_n(v))$，其中 $p_i(v) \in \mathbb{R}$ 是第 i 个竞拍者期望支付价格。因为参加拍卖是自愿的，我们假设对于每个竞拍者 i，如果他的估价低于卖家的最低价，那么他会选择不参加此次竞拍，即 $np \in V_i$。具体来说，如果 $v \in V$ 并且 $v_i = np$, 那么 $p_i(v) = 0$ 且 $q_i(v) = 0$。

现如今，英式拍卖，荷兰式拍卖，第一价格拍卖以及第二价格拍卖是最普遍也是被研究地最多的四种拍卖机制。尤其是英式拍卖如今最流行的。在开始陈述本文提出的模型之前，需要作出以下假设:

假设 4.1 (个人估价私密性)**.** 对于竞拍者 i，他对拍卖品的估价 v_i 是只有他自己知道的，卖家和其他竞拍者是不清楚的。但是所有拍卖参与者都默认 v_i 是一个落在区间 $[v_l, v_h]$ 的随机变量，还知道其概率分布函数 $F_i(v_i) = \int_{v_l}^{v_i} f_i(z)dz$, 其中 f_i 是严格大于零的连续密度函数。

假设 4.2 (估价的独立性)**.** 简单来说, 每个竞拍者的估价 $v_1, v_2, ..., v_n$ 是独立的, 也就是说 $v_1, v_2, ..., v_n$ 的联合概率分布函数为

$$F(v_1, v_2, ..., v_n) = F_1(v_1)F_2(v_2)...F_n(v_n). \tag{4–5}$$

假设 4.3 (估价分布函数的对称性)**.** 这些概率分布函数是相同的，对于所有的 $i, j = 1, 2, ..., n$ 和 $v \in [v_l, v_h]$，有

$$F_i(v) = F_j(v) = F(v). \tag{4–6}$$

假设 4.4 (竞拍者风险中立)**.** 对于每个竞拍者，他的目标是最大化其期望收益。

假设 4.5 (无串通性)**.** 每一个竞拍者都是独立地决定其竞拍策略，在竞拍者之间没有串通。

不同于传统拍卖，在线拍卖并不需要召集所有的参与者到拍卖行中。因此，在线拍卖的竞拍者之间串通的可能性是远远小于传统拍卖的。

以上所有假设描述了对称独立私有价值条件 (Symmeteric Independent Private Value, SIPV)。对于卖家来说，Meyerson[33] 发现了在前面提到过的四种拍卖机制下收益是等价的。由于收益等价定理，在某种程度上来说，以上四种拍卖机制是等价的。本文选择在对称独立私有价值条件下的第二价格拍卖作为在线拍卖机制。为不失一般性，本文将第二价格拍卖扩招到适用于 k 物品的第 k 价格拍卖。本文形式化地将对于数据商品的在线第 k 价格拍卖描述在算法 xx 中。

4.3.1 带保留价的单件物品在线拍卖

在本小节中，我们先讨论设置了公开保留价的单件商品的在线拍卖，也就是这里的 $k=1$。在传统的第二价格拍卖中，Riley 和 Samuelson[18] 得出了这样的结论：当对称独立私有价值条件满足时，对于设置了合适的保留价 r 和有 n 个竞拍者的在线拍卖，卖家的期望收益为：

$$R_s = n\int_r^{v_h}(xf(x)+F(x)-1)F^{(n-1)}(x)dx. \tag{4–7}$$

在单件商品在线拍卖中，数据卖家只卖出一件数据商品。因为在在线拍卖中，竞拍者人数 N 也是一个随机变量，卖家的期望收益应当是与 N 有关的，即：

$$R_s^{online} = E_N(R_s). \tag{4–8}$$

Chen 等人[38] 将竞拍者参与在线拍卖的过程建模为一个到达率为常量 λ 的泊松过程。本文提出的在线拍卖模型也是一个泊松过程，但是该到达率是在不同时段是变化的。我之所以这样建模是基于对 eBay Live Auction 拍卖数据记录的观察，在小节4.2.2中描述了竞拍者在竞拍过程中的到达率是不一样的，一般是初期的到达率很高，中期到达率比较低，然后到竞拍快结束时达到最高峰。因此，我们将整个竞拍时长 $[0,T]$ 划分为 M 个不重叠的子区域 $\{I_i\}_{i=1}^M$，其中 $I_i=[t_{i-1},t_i]$。在不同的子区间中，到达率 λ 取不同的值 $\{\lambda_1,\lambda_2,...,\lambda_M\}$，但在子区间内到达率保持不变。因此，对于一个在时间区间 $I_i=[t_{i-1},t_i]$ 保留价为 r 的在线拍卖而言，有 n_i 个人参与的概率为：

$$P(N_i=n_i|\lambda_i)=\frac{e^{-z_i}z_i^{n_i}}{n_i!}, \tag{4–9}$$

其中，$z_i=\lambda_i(t_i-t_{i-1})(1-F(r))$，$N_i$ 是在时间区间 $I_i=[t_{i-1},t_i]$ 内参与拍卖的人数。为了易于建模，在本文中假设不同区间内的到达率 λ_i 是相互独立的。但是，在实际中，不同区间的竞拍者到达率却是与其前一个到达率和当前最高出价 b^* 相关的。对于竞拍者到达过程更加精确的建模是本文未来的工作之一。那么，总共有 N 个竞拍者参与此次拍卖的概率为：

$$P(N=n_1+n_2+...+n_M|\vec{\lambda})=\prod_{i=1}^{M}P(N_i=n_i|\lambda_i), \tag{4–10}$$

其中 $\vec{\lambda}$ 是参数向量 $(\lambda_1,\lambda_2,...,\lambda_M)$。在所有 $(n_1,n_2,...,n_M)$ 可能取值上的 $P(N=n_1+n_2+...+n_M|\vec{\lambda})$ 的和完全定义了在时间段 $[0,T]$ 上有 N 个人的概率空间。此外，参数向量 $\vec{\lambda}$ 是取决于时间段 $[0,T]$

的划分的。一旦确定了时间段的划分规则，我们就能使用极大似然估计[39, 40] 从之前收集到的拍卖数据估计出 $\vec{\lambda}$。

在数据卖家出售其数据商品之前，他还有收集和清理数据商品的固定成本。我们将这一成本定义为 c。基于公式 (4–7) (4–8) (4–9) 以及 (4–10)，本文为在线第二价格拍卖定义了以下的卖家期望收益：

定义 4.1 (卖家期望收益)**.** 假设对称独立私有价值条件被满足，且卖家设置了保留价 r，其固定成本为 c，那么对于该卖家的期望收益为：

$$\begin{aligned} R_s^{online} &= E_N(R_s|\vec{\lambda}) - c \\ &= \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}...\sum_{n_M=0}^{\infty}\prod_{i=1}^{M} P(N_i = n_i|\lambda_i)R_s - c. \end{aligned} \tag{4–11}$$

需要强调的是 R_s 中的 n 是被 $n_1 + n_2 + ... + n_M$ 替换的。

命题 4.1. *对于在线第二价格拍卖，如果对称独立私有价值条件被满足，卖家的最优保留价为：*

$$r^* = \frac{1 - F(r^*)}{f(r^*)}, \tag{4–12}$$

这样，卖家就能获得最大的期望收益。此外，观察上式可以发现最优保留价 r^* 是独立于竞拍者人数的。

根据 Vickrey 的洞见[15]，竞拍者在第二价格拍卖中的最优竞拍策略就是尽可能地出他的真实估价，即 $b_i(v_i) = v_i$。那些出价最高的竞拍者赢得商品，他们需要支付 $p_i(v)$。在单件商品第二价格拍卖中，竞拍获胜者只有一位。因此，竞拍获胜者支付价格为 $b_2^*(v)$，即该次拍卖中的第二高竞拍价。由于 $b_i(v_i) = v_i$，因此 $b_2^*(v) = v^{(2)}$。而 $v^{(2)}$ 也是一个随机变量，其概率密度函数[18] 定义为：

$$p(v^{(2)}) = n(n-1)f(v^{(2)})F^{(n-2)}(v^{(2)})[1 - F(v^{(2)})]. \tag{4–13}$$

定义 4.2 (竞拍胜者支付价格)**.** 假设对称独立私有价值条件被满足，竞拍胜者的期望支付价格 $v^{(2)}$ 定义为:

$$E(v^{(2)}) = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}...\sum_{n_M=0}^{\infty}\prod_{i=1}^{M} P(N_i = n_i|\lambda_i)\int_r^{v_h} p(x)dx. \tag{4–14}$$

$E(v^{(2)})$ 也是拍卖物品的最后成交价，它也可以被视为数据商品 D 的定价函数 $pr(\dot{})$。Li[32] 等人讨论了基于数据信息上的定价函数需要满足什么样的条件能使得该函数是无套利的。在本文中，我将 $E(v^{(2)})$ 视为与数据信息熵 H 的联系函数因为根据公式 (4–2) 和 (4–3)。

命题 4.2. *期望成交价 $E(v^{(2)})$ 是无套利的，如果 1) $E(v^{(2)})$ 是凹函数; 2) 放大系数满足如下公式：*

$$\frac{f(\alpha H)[1 - F(\alpha H)]}{f(H)[1 - F(H)]} \geq 1. \tag{4–15}$$

在实际中，放大系数 α 是取决于竞拍者对于数据商品价值的主观估计，所以 α 也是一个随机变量。该变量能帮助我们判断基于信息熵的在线拍卖无套利的概率，如何估计这一概率是我们未来的工作之一。

4.3.2 带保留价的多件物品在线拍卖

在多件物品的在线拍卖中，数据卖家会以一个公开的保留价 r 出售 k 件相同的数据商品。假设在拍卖进行的时刻 T 有 N 个风险中立的竞拍者参与进来，并且每一个竞拍者只想拍一件商品。在多件物品在线拍卖中，我们采用第 k 价格拍卖规则，其中竞拍获胜者是出价前 k 高的人，他们只需支付第 $k+1$ 高的出价来获得拍品，第 $k+1$ 高的出价记为 $b^{(k+1)}$（从另一个角度来说，这也是最低竞拍获胜价格)。而对于竞拍者来说，他们的最优竞拍策略仍是出他们的真实估价，即 $b_i(v_i) = v_i$. 因此，$b^{(k+1)} = v^{(k+1)}$。

下面，我们讨论 $v^{(k+1)}$ 的期望着。假设已有 $n > k$ 个人竞拍者参与拍卖。那么其中 k 个竞拍者出价高于或等于 $v^{(k+1)}$，剩下 $n-k$ 个竞拍者出价低于 $v^{(k+1)}$ 的概率是 $C_n^k F^{(n-k)}(v^{(k+1)})[1-F(v^{(k+1)})]^k$。那么，$v^{(k+1)}$ 的期望值为：

$$\begin{aligned} E(v^{(k+1)}) = \int_r^{v_h} & \frac{n!}{k!(n-k)!} F^{(n-k)}(v^{(k+1)}) \\ & [1-F(v^{(k+1)})]^k f(v) v dv. \end{aligned} \tag{4–16}$$

其中竞拍人数 N 是一个随机变量，因此 $v^{(k+1)}$ 对于拍卖人数 N 的期望为：

$$E_N(v^{(k+1)}|\vec{\lambda}) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} ... \sum_{n_M=0}^{\infty} \prod_{i=1}^{M} P(N_i = n_i|\lambda_i) E(v^{(k+1)}). \tag{4–17}$$

$E_N(v^{(k+1)}|\vec{\lambda})$ 为这 k 件商品的最终成交价。如果卖家的固定成本 c 被计算在内的话，那么卖家的期望收益是 $kE_N(v^{(k+1)}|\vec{\lambda}) - c$，即：

$$R_{s,k}^{online} = k(E_N(v^{(k+1)}|\vec{\lambda}) - c). \tag{4–18}$$

4.4 模型评估

在本小节中，我们为本文提出的一系列定义和命题给出可视化的数值结果。

4.4.1 卖家期望收益与保留价的关系

在前文中，我们已经定义了卖家的期望收益 R_s^{online} 和相应的最优保留价 r。我们将通过观察 R_s^{online} 随保留价 r 的变化趋势来验证我们提出的卖家期望收益 R_s^{online} 和最优保留价 r 的合理性。这里，精确的 R_s^{online} 是根据公式 (4–11) 求得。为了简化计算，我们将公式 (4–8) 中的多阶段泊松过程简化成了一阶段泊松过程以此来计算了近似的 R_s^{online}，即：

$$R_s^{online} = \lambda T \int_r^{v_h} [F(x) + xf(x) - 1] \exp[-\lambda T(1-F(r))(1-F(x))] dx. \tag{4–19}$$

这里我们将拍卖时长 T 分别设置成 $3, 5, 7$ 天，竞拍者到达率在 [1.0,4.0] 内变化。此外，我们还假设拍卖参与者的估价服从一个均匀分布，记为 $v_i \sim U[v_l, v_h]$。我们选择 $[10, 100]$ 和 $[100, 200]$ 作为分布区间。接着，根据公式 (4–19) 计算卖家的期望收益，计算结果绘制在图4–2中。

通过观察图4–2中，我们发现在每一个子图中所有曲线都呈现出相同的趋势，即卖家收益首先会随着保留价的增加而升高，随着保留价增加到最优保留价 r^* 而达到最大值，然后随着保留价的继续

增加而下降。需要指出的是，我们这里的最优保留价 r^* 是根据命题4.2中的公式 (4–12) 计算出来的。除此以外我们还观察到：1）那些有着更高到达率的曲线是能完全覆盖那些有着较低到达率的曲线；2）拍卖持续的时间越长，卖家期望收益越高；3）估价区间范围越大，卖家期望收益越高。

4.4.2 成交价 $E(v^{(2)})$ 的数值结果

定义4.2给出了第二价格在线拍卖中成交价 $E(v^{(2)})$。类似于前面卖家期望收益的计算，在这里我们也将 $E(v^{(2)})$ 的计算简化为：

$$E(v^{(2)}) = \sum_{n=0}^{\infty} \frac{e^{-z}z^n}{n!} \int_r^{v_h} n(n-1)f(x) F^{(n-2)}(x)[1-F(x)]dx, \tag{4–20}$$

其中 $z = \lambda T(1-F(r))$，T 是在线拍卖竞拍时长。

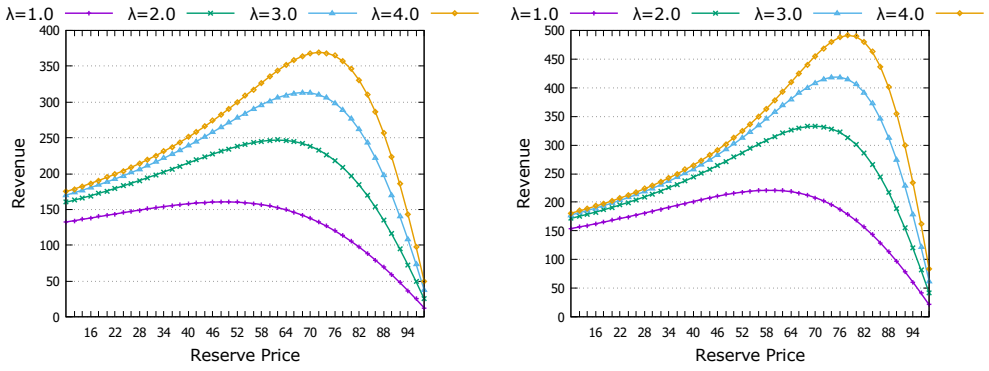
在 $E(v^{(2)})$ 的结果展示中，我们的参数设置与前一小节相同。需要指出的是，这里的保留价也设置成之前计算出来的最优保留价，因为只有当保留价达到最优值时，$E(v^{(2)})$ 才会达到最大值。相关结果呈现在图4–3。

从图4–3中可以看出：1）拍卖持续时间越长，其拍品成交价就越高；2）较高的到达率会造成较高的期望成交价，这也意味着到达率可以一定程度上表示在线拍卖市场的竞争激烈程度；3）高估值的拍品最后的成交价都很高。

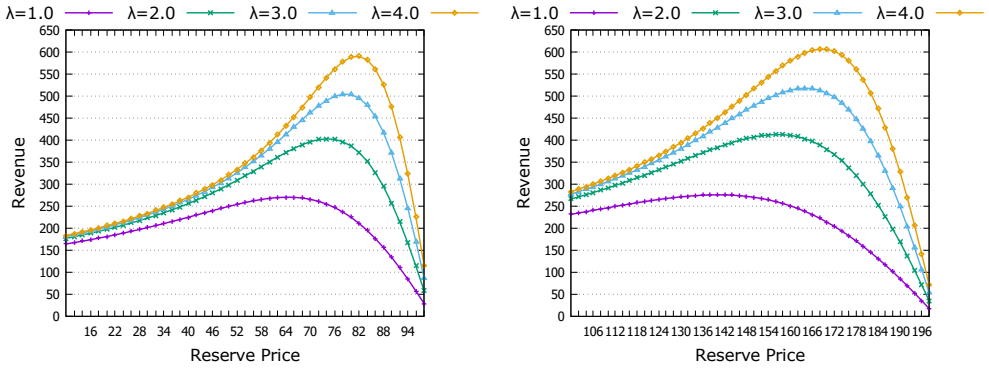
以上所有的数值结果与现有在线拍卖市场的经验调查是相符的，那么这意味着本文提出的模型是有能力模拟在线数据拍卖市场的。

4.5 本章小结

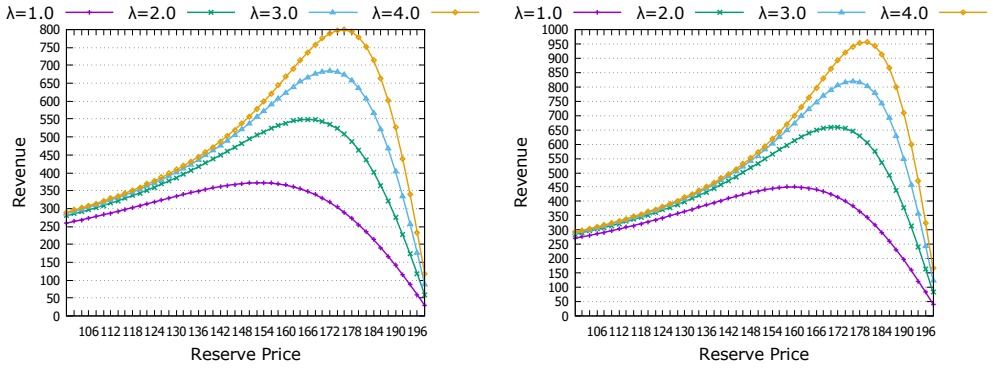
本小节提出了一个基于数据信息熵的在线数据拍卖模型。在叙述这个模型之前，我们分析了那些最影响现有传统商品在线拍卖的因素，比如拍卖机制、拍卖持续时长、拍卖参与人数以及保留价等等。在全面考虑了以上因素后，我们提出了基于数据信息熵的第 k 价格在线数据拍卖模型。在这个模型中，数据商品的数据信息熵被作为估价的参考。在这个模型的基础上，我们进一步分析了其最优保留价、期望成交价在不同出售数量下的变化趋势。此外，本文提出的一些命题和定义也在模型评估阶段进行了可视化的呈现。我们希望本文提出的模型和分析结果能给现有的在线数据拍卖市场提供些许启发，从而最大化拍卖参与者的总剩余，进一步繁荣整个数据市场。



(a) $T = 3, v_l = 10, v_h = 100$ (b) $T = 5, v_l = 10, v_h = 100$



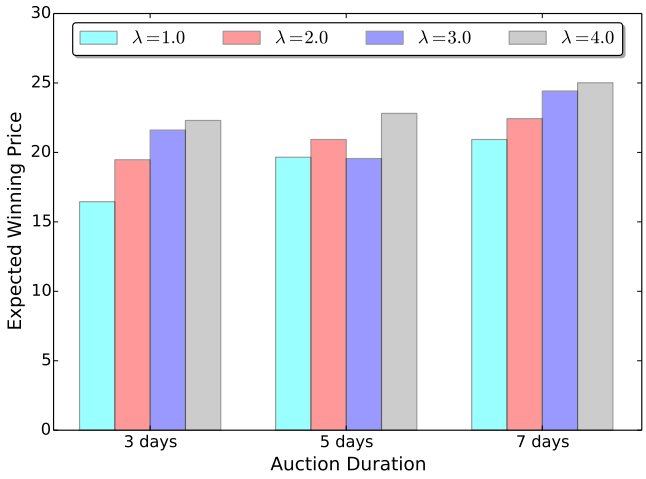
(c) $T = 10, v_l = 10, v_h = 100$ (d) $T = 3, v_l = 100, v_h = 200$



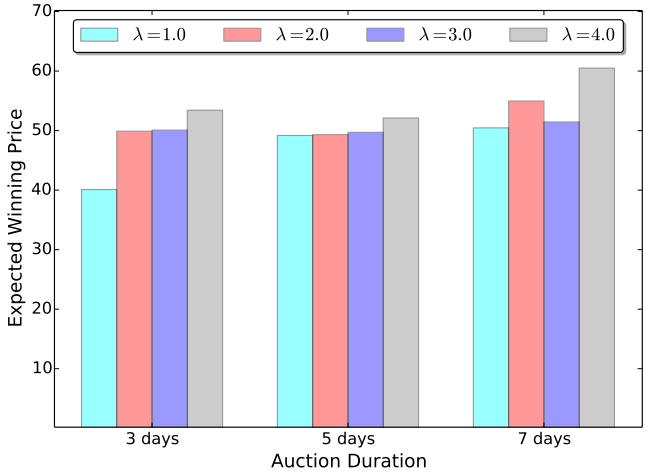
(e) $T = 5, v_l = 100, v_h = 200$ (f) $T = 7, v_l = 100, v_h = 200$

图 4–2 在不同参数设置下的卖家平均收益与保留价的变化趋势

Fig 4–2 Relationship between revenue and reserve price under different parameter configurations



(a) $v_l = 10, v_h = 100$



(b) $v_l = 100, v_h = 200$

图 4–3 在不同参数设置下的期望成交价 $E(v^{(2)})$

Fig 4–3 Numerical results of expected winning price $E(v^{(2)})$ under different parameter configurations

全文总结

这里是全文总结内容。

2015 年 2 月 28 日，中央在北京召开全国精神文明建设工作表彰暨学雷锋志愿服务大会，公布全国文明城市（区）、文明村镇、文明单位名单。上海交通大学荣获全国文明单位称号。

全国文明单位这一荣誉是对交大人始终高度重视文明文化工作的肯定，是对交大长期以来文明创建工作成绩的褒奖。在学校党委、文明委的领导下，交大坚持将文明创建工作纳入学校建设世界一流大学的工作中，全体师生医护员工群策群力、积极开拓，落实国家和上海市有关文明创建的各项要求，以改革创新、科学发展为主线，以质量提升为目标，聚焦文明创建工作出现的重点和难点，优化文明创建工作机制，传播学校良好形象，提升社会美誉度，显著增强学校软实力。2007 至 2012 年间，上海交大连续三届荣获“上海市文明单位”称号，成为创建全国文明单位的新起点。

上海交大自启动争创全国文明单位工作以来，凝魂聚气、改革创新，积极培育和践行社会主义核心价值观。坚持统筹兼顾、多措并举，将争创全国文明单位与学校各项中心工作紧密结合，着力构建学校文明创建新格局，不断提升师生医护员工文明素养，以“冲击世界一流大学汇聚强大精神动力”为指导思想，以“聚焦改革、多元推进、以评促建、丰富内涵、彰显特色”为工作原则，并由全体校领导群策领衔“党的建设深化、思想教育深入、办学成绩显著、大学文化丰富、校园环境优化、社会责任担当”六大板块共 28 项重点突破工作，全面展现近年来交大文明创建工作的全貌和成就。

进入新阶段，学校将继续开拓文明创建工作新格局，不断深化工作理念和工作实践，创新工作载体、丰富活动内涵、凸显创建成效，积极服务于学校各项中心工作和改革发展的大局面，在上级党委、文明委的关心下，在学校党委的直接领导下，与时俱进、开拓创新，为深化内涵建设、加快建成世界一流大学、推动国家进步和社会发展而努力奋斗！

上海交通大学医学院附属仁济医院也获得全国文明单位称号。

参考文献

[1] GANTZ J, REINSEL D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east[J]. IDC iView: IDC Analyze the future, 2012, 2007: 1–16.

[2] VILLARS R L, OLOFSON C W, EASTWOOD M. Big data: What it is and why you should care[J]. White Paper, IDC, 2011.

[3] Worldwide Semiannual Big Data and Analytics Spending Guide. http://www.idc.com/getdoc.jsp?containerId=IDC_P33195.

[4] MAITLAND C F, BAUER J M, WESTERVELD R. The European market for mobile data: evolving value chains and industry structures[J]. Telecommunications Policy, 2002, 26(9): 485–504.

[5] TURNER V, GANTZ J F, REINSEL D, et al. The digital universe of opportunities: rich data and the increasing value of the internet of things[J]. IDC Analyze the Future, 2014.

[6] Microsoft Windows Azure Marketplace. https://datamarket.azure.com/browse/data.

[7] Infochimps Datamarket. http://www.infochimps.com/.

[8] Factual. https://www.factual.com/.

[9] 贵阳大数据交易所. http://www.gbdex.com/website/.

[10] 东湖大数据交易中心. http://www.chinadatatrading.com.

[11] Custom Lists. http://www.customlists.net/.

[12] KOUTRIS P, UPADHYAYA P, BALAZINSKA M, et al. Query-based data pricing[J]. Journal of the ACM (JACM), 2015, 62(5): 43.

[13] BAKOS Y, BRYNJOLFSSON E. Aggregation and disaggregation of information goods: Implications for bundling, site licensing, and micropayment systems[G]//Lectures in E-Commerce. [S.l.]: Springer, 2001: 103–122.

[14] BAKOS Y, BRYNJOLFSSON E. Bundling information goods: Pricing, profits, and efficiency[J]. Management science, 1999, 45(12): 1613–1630.

[15] VICKREY W. Counterspeculation, Auctions, and Competitive Sealed Tenders[J]. The Journal of Finance, 1961, 16(1): 8–37.

[16] MILGROM P. Auctions and Bidding: A Primer[J]. Journal of Economic Perspectives, 1989, 3(3): 3–22.

[17] MILGROM P R, WEBER R J. A Theory of Auctions and Competitive Bidding[J]. Econometrica, 1982, 50(5): 1089–1122.

[18] JG RILEY W S. Optimal Auctions[J]. American Economic Review, 1981, 71(3): 381–392.

[19] LUCKING-REILEY D. Auctions on the Internet: What’s Being Auctioned, and How?[J]. Journal of Industrial Economics, 2000, 48(3): 227–252.

[20] BEAM C, SEGEV A. Auctions on the Internet: A Field Study[J]. 1998: 1032.

[21] PINKER E J, SEIDMANN A. Managing Online Auctions: Current Business and Research Issues[J]. Management Science, 2003, 49(11): 1457–1484.

[22] PINKER E J, SEIDMANN A, VAKRAT Y. Using Transaction Data for the Design of Sequential, Multi-unit, Online Auctions[J]. 2001.

[23] ROTH A E, OCKENFELS A. Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet[J]. American Economic Review, 2002, 92(4): 1093–1103.

[24] OCKENFELS A, REILEY D, SADRIEH A. Online Auctions[J]. Nber Working Papers, 2006, volume 310(3): 233–241.

[25] DAN A, SIMONSON I. Buying, Bidding, Playing, or Competing? Value Assessment and Decision Dynamics in Online Auctions[J]. Journal of Consumer Psychology, 2003, 13(1–2): 113–123.

[26] LU X, MCAFEE R P. The Evolutionary Stability of Auctions over Bargaining [J]. Games and Economic Behavior, 1996, 15(2): 228–254.

[27] HARRIS M, RAVIV A. A Theory of Monopoly Pricing Schemes with Demand Uncertainty[J]. American Economic Review, 1981, 71(3): 347–365.

[28] SHANNON C E. A mathematical theory of communication[J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1): 3–55.

[29] KOUTRIS P, UPADHYAYA P, BALAZINSKA M, et al. QueryMarket demonstration: Pricing for online data markets[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 1962–1965.

[30] LIN B.-R, KIFER D. On arbitrage-free pricing for general data queries[J]. Proceedings of the VLDB Endowment, 2014, 7(9): 757–768.

[31] BALAZINSKA M, HOWE B, SUCIU D. Data markets in the cloud: An opportunity for the database community[J]. Proc. of the VLDB Endowment, 2011, 4(12): 1482–1485.

[32] LI X, YAO J, LIU X, et al. A First Look at Information Entropy-Based Data Pricing[C]//Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on. IEEE. [S.l.]: [s.n.], 2017: 2053–2060.

[33] MYERSON R B. Optimal Auction Design[J]. Mathematics of Operations Research, 1981, 6(1): 58–73.

[34] LUCKING-REILEY D. Using Field Experiments to Test Equivalence between Auction Formats: Magic on the Internet[J]. American Economic Review, 1999, 89(5):

[35] Amazon Auction Official Website. `http://auctions.amazon.com`.

[36] BRADLOW E T, PARK Y H. Bayesian Estimation of Bid Sequences in Internet Auctions Using a Generalized Record-Breaking Model[M]. [S.l.]: INFORMS, 2007: 218–229.

[37] VAKRAT Y, SEIDMANN A. Implications of the Bidders' Arrival Process on the Design of Online Auctions[C]//Hawaii International Conference on System Sciences. [S.l.]: [s.n.], 2000: 6015.

[38] CHEN S, WU H, LUO Y. Optimal Design of Online Auction[J]. 2007: 64–70.

[39] BOCK R D, AITKIN M. Marginal maximum likelihood estimation of item parameters[J]. Psychometrika, 1982, 47(3): 369–369.

[40] ZHOU M, TIAN S, PARK T. An empirical test of Tobit model robustness in estimating online auction prices over various distributions[J]. International Journal of Mathematics in Operational Research, 2017, 10.

致 谢

感谢所有测试和使用交大学位论文 LaTeX 模板的同学!
感谢那位最先制作出博士学位论文 LaTeX 模板的交大物理系同学!
感谢 William Wang 同学对模板移植做出的巨大贡献!

攻读学位期间发表的学术论文

[1] CHEN H, CHAN C T. Acoustic cloaking in three dimensions using acoustic metamaterials[J]. Applied Physics Letters, 2007, 91:183518.

[2] CHEN H, WU B I, ZHANG B, et al. Electromagnetic Wave Interactions with a Metamaterial Cloak[J]. Physical Review Letters, 2007, 99(6):63903.

攻读学位期间参与的项目

[1] 973 项目“XXX”

[2] 自然基金项目“XXX”

[3] 国防项目“XXX”