

数字普惠金融丛书

MODELING ONLINE AUCTIONS

Statistics in Practice

Wolfgang Jank Galit Shmueli

网上 拍卖数据建模 统计方法理论与实践

(美)

沃尔夫冈·詹克

加利特·士穆里 著

曲春青 译

WILEY



东北财经大学出版社
Dongbei University of Finance & Economics Press

国家一级出版社
全国百佳图书出版单位

数字普惠金融丛书

MODELING ONLINE AUCTIONS

Statistics in Practice

Wolfgang Jark Galit Shmueli



网上 拍卖数据建模 统计方法理论与实践

(美)

沃尔夫冈·詹克

加利特·士穆里 著

曲春霞 译

WILEY

FE 东北财经大学出版社
Dongbei University of Finance & Economics Press

大连

辽宁省版权局著作权合同登记号：图字06-2015-35号

Wolfgang Jank, Galit Shmueli: Modeling Online Auctions: Statistics in Practice, original ISBN: 978-0-470-47565-2.

Copyright © 2010 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

This translation published under license.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400, fax 978-646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, 201-748-6011, fax 201-748-6008, or online at www.wiley.com/go/permissions. Copies of this book sold without a Wiley sticker on the cover are unauthorized and illegal.

本书简体中文翻译版由约翰·威立父子有限公司授权东北财经大学出版社独家出版发行。未经授权的本书出口将被视为违反版权法的行为。未经出版者预先书面许可，不得以任何方式复制或发行本书的任何部分。

图书在版编目 (CIP) 数据

网上拍卖数据建模：统计方法理论与实践 / (美) 沃尔夫冈·詹克 (Wolfgang Jank), (美) 加利特·士穆里 (Galit Shmueli) 著；曲春青译. —大连：东北财经大学出版社，2017.11
(数字普惠金融丛书)
ISBN 978-7-5654-2524-0

I. 网… II. ①沃… ②加… ③曲… III. 互联网络-应用-拍卖-统计数据-研究 IV. F713.359-39

中国版本图书馆CIP数据核字 (2016) 第 267572号

东北财经大学出版社出版发行

大连市黑石礁尖山街217号 邮政编码 116025

网 址：<http://www.dufep.cn>

读者信箱：dufep@dufe.edu.cn

大连图腾彩色印刷有限公司印刷

幅面尺寸：170mm×240mm 字数：353千字 印张：25.25

2017年11月第1版

2017年11月第1次印刷

责任编辑：刘东威

责任校对：刘 佳

封面设计：冀贵收

版式设计：钟福建

定价：66.00元

教学支持 售后服务 联系电话：(0411) 84710309

版权所有 侵权必究 举报电话：(0411) 84710523

如有印装质量问题，请联系营销部：(0411) 84710711

本书受国家自然科学基金资助，项目编号：71373038。

译者前言

网上拍卖 (auction online) 是指通过互联网实施的价格谈判交易活动, 即利用互联网在网站上公开发布将要招标的物品或者服务的信息, 通过竞争投标的方式将它出售给出价最高或最低的投标者。其实质是以竞争价格为核心, 建立生产者和消费者之间的交流与互动机制, 共同确定价格和数量, 从而达到均衡的一种市场经济过程。^①相对于传统拍卖, 网上拍卖的优势是非常明显的。从时间上看, 一场网上拍卖可以持续一天甚至几天, 竞拍者可以不分昼夜地出价; 从空间上看, 网上拍卖不受地域限制, 竞拍者可以方便地参与异地拍卖。因此, 网上拍卖可以说是一种更有效率的竞价交易方式。

网上拍卖虽然是一种较为新颖的交易方式, 但是近年来却取得了飞速发展, 比如国外的 eBay.com、uBid.com, 国内的淘宝拍卖会等网上拍卖平台的交易量迅猛增长。但是网上拍卖数据的特殊性, 导致了相关的理论和实证研究比较滞后。基于网络用户行为生成的竞价数据并不是标准的时间序列数据, 竞价事件的发生并不是均匀的, 并且表现出伴随拍卖进程变化的动态性, 同时拍卖的竞争 (既有卖家的竞争, 也有买家的竞争) 也很难捕捉。这些困难导致传统的统计方法已经不能胜任网上拍卖数据的分析和

^① 该定义引自百度百科。

研究，急需新方法的提出和指导。

沃尔夫冈·詹克、加利特·士穆里所著的《网上拍卖数据建模：统计方法理论与实践》正是针对网上拍卖数据的特殊性，创新性地提出了一些统计建模方法，并展示了这些方法在实践领域的应用，是网上拍卖数据建模领域一本不可多得的好书。本书系统地介绍了网上拍卖数据的收集、挖掘和建模的方法。首先，该书介绍了网上拍卖数据的获取方法，重点介绍了网络爬虫和网络服务两种常用方法。然后，该书又介绍了网上拍卖数据存在的一些特殊问题以及相关研究方法，比如竞拍历史数据的不均匀间隔问题及其与截面信息的结合问题、网上拍卖的并发性以及相互竞争性问题、数据半连续性问题、数据分层问题等。接下来，该书讨论了几类相关模型，包括用于捕捉拍卖数据之间相互关系的统计模型、捕捉拍卖竞争的统计模型、捕捉竞拍者参与和竞拍参与的统计模型。最后，该书进一步讨论了网上拍卖预测模型及应用问题。该书介绍和讨论的理论、模型和研究方法，不仅适用于网上拍卖数据的研究，同样适用于其他一些电子商务领域的研究。因此，该书既可以作为经济学专业、电子商务专业和统计学专业本科高年级学生、研究生学习的教材，也可以作为电子商务领域，尤其是网上拍卖领域研究人员、从业人员、交易者学习和使用的参考用书。

本书由曲春青博士翻译，在翻译过程中东北财经大学统计学院的杨仲山教授、屈超博士为本书的翻译提出了许多建设性的意见和建议，范立夫教授、郭凯副教授、刘伟副教授、张晓东副教授、康书隆副教授、周学仁副研究员、陈菁泉副研究员、解维敏副教授、霍红博士、姜美华博士及东北财经大学出版社的田玉海编辑、刘佳编辑为本书的翻译提供了许多帮助，东北财经大学的硕士研究生刘锐、白方玉、钟艳秋、李松、陈慧芳、谭洁蔓、陈丽茹、张方方、李亚云、陆明、田志信参与了本书部分章节的校对以及部分公式的录入工作。东北财经大学出版社的刘东威编辑、李季

编辑、吉扬编辑为本书的出版付出了辛勤的劳动。在此，我们表示衷心的感谢。由于译者水平所限，书中定有不妥甚至错误之处，恳请读者批评指正。

译 者

2016年12月

序

我们对于网上拍卖（online auction）的研究兴趣始于2002年。那时，有关网上拍卖数据的实证研究刚刚兴起，其研究大多聚焦于经济学和信息系统领域。起初，我们只是简单地运用描述性统计方法来研究网上拍卖数据，试图解答一些问题，比如“如何绘制单独的竞拍历史数据”“如何在避免信息过载（information overload）的前提下绘制1 000个竞拍的历史数据”“如何描绘一个正在进行的拍卖的价格走势”“如何有效地收集海量的数据”。当时，我们与一些业内人士（主要是非统计人员）展开了合作。他们向我们介绍了拍卖理论、网络环境下的局限性以及高速、有效收集海量网上拍卖数据的现代技术。我们开始利用统计学的思维和方法，以一种全新的、数据驱动的方法去研究诸如“什么因素影响最终价格”“如何量化赢者诅咒（winner curse）”等问题。在我们接触到海量的数据并深入了解了网上拍卖的运作机制以后，我们开始对一些更加复杂、更具挑战性的问题进行研究，比如“如何捕捉一个拍卖的价格动态”“如何运用动态的方法对拍卖的结果进行预测”。从那时起，我们就这些问题的研究做了大量工作。事实上，也取得了一些颇具特色的创新性成果。首先，我们不仅向自己提出这些问题，同时也向我们的博士研究生提出了这些问题，产生了一个全新的研究方向。我们的毕业生利用他们深厚的数学或统计学功底，解决了一些相关的商业和电

子商务方面的问题。正如之前 Davenport 与 Harris 合著的专著 *Competing on Analysis* 以及其他同类书籍中所述的，这种与数学和统计学的结合，对于现代商业分析的发展至关重要。

我们的研究催生了一个新的研究领域——电子商务统计学。我们与许多同事展开交流与合作，并创办了一个至今（2010 年）已经六年的同名的年度研讨会（见 <http://www.statschallenges.com>）。这一研讨会吸引了不同领域的众多研究者，他们都有这样一种信念，就是使用实证技术对电子商务数据进行研究，从而更好地理解网络世界。

关于网上拍卖方面的研究之所以吸引人，主要基于以下几个方面。认真回顾长期以来网上拍卖的数据，我们惊奇地发现这些数据的结构并不能使用标准的统计工具去表示或建模，比如事件发生的不规则性以及网上拍卖数据的“半连续”性（semicontinuous）。为了处理这些数据并从中挖掘出尽可能多的信息，我们不仅使用了现有的统计学方法，而且开发了一些新的工具和方法。我们对网上拍卖统计方法的研究，催生了大量针对拍卖数据探索、建模和预测的方法和工具。虽然本书主要讨论的是网上拍卖问题，但是本书内容的应用领域相当广泛，对那些研究其他网上问题的研究人员同样有益。

本书适用于各类对网上拍卖统计建模抱有兴趣的研究人员和学生。一方面，本书详细叙述了在网拍卖研究中统计人员可以有所作为的一些领域并提供了大量的网拍卖数据，这些可以激发统计人员在该领域开发新方法以及将研究应用到实践的兴趣。另一方面，本书还详细叙述了我们挖掘网拍卖数据的方法以及一些重要结论（可能是令人吃惊的），对于那些正在寻求新方法去深入了解网拍卖或其他网上市营销人员、经济学家以及信息系统研究人员大有益处。为了使研究人员能够复制（并改进）我们的方法和模型，我们提供了程序的代码。相关的数据和程序代码可以在我们的合作伙伴网站（<http://ModelingOnlineAuction.com>）上面找到。我们的研究始于数据，因此我们也希望大家能够直接通过网拍

卖数据进行数据探索。

沃尔夫冈·詹克 加利特·士穆里

2010年1月

目 录

第1章 引言 1

- 1.1 网上拍卖与电子商务 3
- 1.2 网上拍卖和统计学挑战 5
- 1.3 网上拍卖研究的统计方法 7
- 1.4 本书结构 8
- 1.5 数据和代码的获得 11

第2章 网上拍卖数据的获取 12

- 2.1 从网络采集数据 12
- 2.2 网络数据采集和统计抽样 23

第3章 网上拍卖数据的探索 38

- 3.1 竞拍历史：出价与“当前价格” 39
- 3.2 竞价历史数据与截面拍卖信息数据的整合 42
- 3.3 并发性拍卖（concurrent auctions）的可视化 50
- 3.4 价格演变和价格动态的探索 54
- 3.5 通过交互可视化对价格曲线和拍卖信息进行整合 70
- 3.6 探索层次信息 73
- 3.7 通过曲线聚类方法探索价格动态 77
- 3.8 探索分布假设 88

3.9 网上拍卖探索：未来研究方向 115

第4章 网上拍卖数据建模 117

4.1 建模基础（价格过程的表示） 118

4.2 价格动态和拍卖信息之间关系的建模 160

4.3 拍卖竞争的建模 189

4.4 竞价与竞拍者参与的建模 226

4.5 拍卖网络的建模 283

第5章 网上拍卖的预测 300

5.1 单个拍卖的预测 301

5.2 竞争性拍卖的预测 333

5.3 自动化出价决策 348

参考文献 361

第1章 引言

近年来，网上拍卖如井喷之势流行开来，出现了一些诸如 eBay.com、uBid.com、Swoopo.com 等允许买卖双方交易商品或信息的网站。在线拍卖平台与 Amazon.com 之类的固定价格零售平台有所不同，它允许买卖双方就交易进行协商。多方面的原因导致了网上拍卖的兴起。首先，网上拍卖网站的持续有效性使得卖家可以在任何时间摆卖物品，也使得买家可以不分昼夜地出价。商品通常展示几天，给予买家足够的时间去搜索、比较、判断和出价。其次，网上拍卖不受现实中的地域限制，买家可以参与另一地的拍卖活动。广泛的地域可达性也使得人们摆出大量的产品以供销售（无论是全新商品还是二手商品）。最后，由于网上拍卖使得参与者聚集到一个竞争性环境当中，它也有娱乐的成分。事实上，网上拍卖过程中的社会化行为有点类似于赌博，竞拍者参与竞拍并期待获胜，而且常常在最后时刻由于竞拍失败而出现情绪化反应。

网上拍卖是一种相对而言较为新颖的事物。我们所说的网上拍卖指的是基于网络的拍卖，即交易发生在互联网上。然而，早在互联网拍卖出现之前，电子拍卖就已经通过电子邮件、讨论组以及新闻组（newsgroup）等方式出现了。David Lucking-Reiley（2000）讲述了 rec.games.deckmaster 新闻组的故事。早在 1995 年，互联网用户便开始在那里交易万智牌（Magic cards）。他这样写道：

截止到1995年春季，每周的发帖量达到6 000个，这使得rec.games.tradingcards.marketplace成为互联网上流量最大的新闻组。每月的26 000个帖子当中，近90%都与万智牌的交易有关，剩下的10%是其他游戏牌的交易。

Lucking-Reiley (2000) 简要介绍了网上拍卖的发展历史，并提供了一份关于1998年网上拍卖门户网站的调查报告。Onsale和eBay是最早成立网上拍卖网站的，成立于1995年。后来，Onsale（即现在的Egghead）将拍卖业务出售给了Yahoo!，转而进入固定价格零售领域。1999年，Yahoo!和Amazon上线了它们自己的网上拍卖业务。但是在10年时间里，它们都停止了拍卖业务，只从事固定价格业务（在本书写作之时，Yahoo!只保留了中国香港、台湾地区和日本的网上拍卖业务）。从1995年起至今（2010），C2C网上拍卖市场大体经历了如下发展：在最初网上拍卖网站大量涌现之后，出现了一个剧烈的并购期，最后剩下了一些今天为我们所熟知的优秀拍卖网站，比如针对一般商品的eBay、uBid、Swoopo，针对印度艺术品的SaffromArt和针对P2P借贷的Prosper。

关于网上拍卖的实证研究出现了爆炸式增长。事实上，网上拍卖的增长速度远远超过了传统实体拍卖的增长速度。对此，我们不禁要问：“相较于传统拍卖，为何网上拍卖的数据研究更为流行？”我们认为答案其实很简单，一言以蔽之：数据。实际上，正在进行或是过去发生的网上拍卖交易数据的公开可得性，为实证研究人员研究买卖双方行为提供了新的机会。此外，基于经济学和线下实体拍卖研究所得的理论结果，已经被证实网络环境下不再成立。网上拍卖和线下拍卖出现差异的原因可能是由于以下几点，比如互联网的全球覆盖性、用户的匿名性、虚拟资源的无限性、持续有效性以及变动的连续性。

在一份关于网上拍卖的初期研究中（Lucking-Reiley等人，2000），实证经济学家发现竞价行为，尤其是在eBay上的竞价行为，通常与传统拍卖理论所预测的竞价行为有显著的差异。从那时起，在信息系统、营销

学、计算机科学、统计学和其他相关领域内，涌现出大量使用网上拍卖数据进行研究的实证研究，这些研究从不同视角去探讨网络环境下的竞价行为。一类研究主要关注新竞价行为和现象的识别与定性，如狙击出价（bid sniping）（Roth 和 Ockenfels，2002）和竞价自抬（bid shilling）（Kauffman 和 Wood，2005）；竞拍者的分类（Bapna 等人，2004）；用于捕捉竞价活动以及竞拍者行为的描述性概率模型（Borle 等人，2006；Park 和 Bradlow，2005）；用于捕捉竞价时机和竞拍价的描述性概率模型的开发（Shmueli 等人，2007；Russo 等人，2008）。另一类研究主要关注网上拍卖的价格演化趋势，比如价格动态研究（Wang 等人，2008a，b；Bapna 等人，2008b；Dass 和 Reddy，2008；Reddy 和 Dass，2006；Jank 和 Shmueli，2006；Hyde 等人，2008；Jank 等人，2008b，2009）；动态预测拍卖价格的新模型研究（Wang 等人，2008a；Jank 和 Zhang，2009a，b；Zhang 等人，2010；Jank 和 Shmueli，2010；Jank 等人，2006；Dass 等人，2009）。进一步的研究课题包括诸如 eBay 的消费者剩余（Bapna 等人，2008a）之类的经济价值的量化问题。最近，网上拍卖数据被用于网络环境下的买卖双方关系研究（Yao 和 Mela，2007；Dass 和 Reddy，2008；Jank 和 Yahav，2010），以及产品、拍卖形式、拍卖平台的竞争行为研究（Haruvy 等人，2008；Hyde 等人，2006；Jank 和 Shmueli，2007；Haruvy 和 Popkowski Leszczyc，2009）。总之，网上拍卖的实证研究正处在蓬勃发展阶段。

1.1 网上拍卖与电子商务

网上拍卖是电子商务的分支之一。电子商务通常是指任意形式的源于网络的交易活动。与十年前相比，电子商务对我们今天的生活方式产生了巨大的影响：它变革了经济、消除了障碍、为若干年前无法想象的创新敞开了大门，并产生了顾客与商人之间互动的全新方式。尽管在 20 世纪 90

年代末期，多数人曾预言电子商务会随着互联网泡沫的破裂而消失，但是电子商务却日益繁荣。电子商务包括在线购买、在线销售以及在线投资。举几个例子，比如人们在 Amazon.com 上购物或是在 eBay.com 上参与网上拍卖；在 Craigslist.com 上买卖二手物品；使用网络广告服务（比如 Google、Yahoo! 和 Microsoft 提供的广告服务）；在线预订和购买门票（比如旅行券和电影票）；发布和下载音乐、视频和其他在线内容；在 Epinions 或 Aamazon 等网站上发表产品意见或对产品评分；通过在线商店、在线拍卖购买或者提供服务（比如 Amazon Mechanical Turk 或者 eLance）等等。

电子商务的实证研究包括了从宏观到微观的诸多问题。微观问题研究的例子有：网上二手市场对 CD 和 DVD 销售的影响问题（Telang 和 Smith, 2008）、开源软件的演化问题（Stewart 等人, 2006）、软件行业网上价格偏差的最优化问题（Ghose 和 Sundararajan, 2006）、互联网广告中的资源有效配置问题（Agarwal, 2008）、广告客户的最优竞价策略问题（Matas 和 Schamroth, 2008）、互联网公司的进入与退出问题（Kauffman 和 Wang, 2008）、网上销售的地域影响问题（Jank 和 Kannan, 2008）、虚拟股票市场的效率和有效性问题（Spann 和 Skiera, 2003；Foutz 和 Jank, 2009）以及网络百科全书（维基百科）的影响问题（Warren 等人, 2008）。

宏观问题的研究包括：电子商务交易中的私密性和可信性问题（Fienberg, 2006, 2008）、互联网交易的数据探索问题（Banks 和 Said, 2006）、点击流数据的建模问题（Goldfarb 和 Lu, 2006）以及电子商务数据间关系的时变性问题（Overby 和 Konsynski, 2008）。除此之外，宏观问题的研究还涉及我们如何通过网络经验加深对于线上世界的理解问题（Forman 和 Goldfarb, 2008），用户的网络行为对经济的影响问题（Ghose, 2008），大规模电子商务数据的收集、验证和分析问题（Bapna 等人, 2006），如何评估市场干预的影响问题（Rubin 和 Waterman,

2006; Mithaset 等人, 2006), 社交网络和口头相传的影响问题 (Hill 等人, 2006; Dellarocas 和 Narayan, 2006)。

互联网广告是实证研究迅速发展的另一个领域, 但其研究更多局限于公司内部, 学术界研究较少。诸如 Google、Yahoo! 和 Microsoft 这类公司使用大规模的竞价行为数据和竞价结果数据库去研究网络广告客户的行为, 从而实现资源 (比如广告投放) 的有效分配 (Agarwal, 2008)。网络广告客户和提供服务的广告公司也会对竞价数据进行研究。他们研究竞价和利润 (或者其他代表成功的指标) 之间的关系, 以期实现广告客户竞价策略的最优化 (Matas 和 Schamroth, 2008)。

另一个实证研究迅速发展的领域与预测市场 (prediction markets) 问题有关, 也被称为“信息市场”、“想法市场”或者“博彩市场 (betting exchange)”。预测市场是在网络社区集合众人智慧 (Surowieck, 2005) 预期未来事件的结果。这类研究的应用范围很广, 从预测自然灾害造成的经济影响到电影票房的销量。尽管有一些实证研究 (Spann 和 Skiera, 2003; Forsythe 等人, 1999; Pennock 等人, 2001) 表明了最终交易价格预测的精确性, 但仍有证据表明预测市场并不是完全有效的, 这给人们提出了新的统计学挑战 (Foutz 和 Jank, 2009)。

实证分析网上拍卖和电子商务时, 人们面临的统计学挑战具有许多相似之处。接下来, 我们将集中在网上拍卖领域讨论这些挑战。更多关于电子商务方面的研究, 参见 Jank 等人 (2008a) 或者 Jank 和 Shmueli (2008a)。

1.2 网上拍卖和统计学挑战

海量数据的可得性是网上拍卖实证研究大量涌现的一个关键性因素。虽然海量数据为研究新型问题打开了一扇大门, 但是同时也带来了新的挑