

个人数据交易：从保护到定价

关键词：个人数据市场 数据服务 安全隐私 定价机制

牛超越 郑臻哲 吴帆 等
上海交通大学

个人数据流通和交易现状

在大数据驱动经济发展的今天，个人数据已经日趋商品化。国内外知名的互联网公司往往通过提供免费的在线服务以获取个人数据，许多大数据创业公司通过支付用户酬劳来获取对数据的使用权限。然而，当互联网用户逐渐意识到伴随数据分享而来的隐私泄露等风险时，将拒绝在自己的敏感信息没有得到全面保护、隐私泄露没有得到合理补偿的情况下提供个人数据。政府部门和企业作为个人数据资源的主要采集者和拥有者，也存在用户隐私保护和商业机密隐藏等安全需求，他们大都只在内部分析和使用用户数据。海量的个人数据缺乏开放和流通，形成了大量的信息孤岛，严重地抑制了市场对数据的需求，成为大数据发展的瓶颈，亟须安全可靠的数据交易平台促进个人数据资源开放、推动数据应用和释放数据价值。

为了促进个人数据流通，许多数据代理商纷纷出现，在数据贡献者和数据消费者之间搭建桥梁：一方面，通过金钱性的补偿来激励数据贡献者分享数据；另一方面，为数据消费者提供数据服务并收取一定的费用。根据美国联邦贸易委员会(Federal Trade Commission)在2014年5月发表的关于九个具有代表性数据市场的调查表明：总部位于美国阿肯萨斯州的安客诚公司(Acxiom)作为最大的数据代理商从全球约7亿用户处采集个人数据，并为全球顶尖的企业提供基于数据智能分析的商业解决方案^[1]。同年8月，美国哥伦比亚广播公司新闻部门的《60

分钟》节目(CBS News 60 Minutes)对此提出质疑：数据代理商从用户的个人数据中获取暴利，却没有对用户的隐私泄露进行补偿^[2]。围绕数据交易市场构建的相关话题也引起了我国政府相关部门的高度重视，并引发了新闻媒体的广泛讨论。2015年9月5日，《国务院关于印发促进大数据发展行动纲要的通知》正式发布。该行动纲要的核心是推动各部门、各地区、各行业、各领域的数据资源共享开放，并明确提出要引导培育大数据交易市场，建立健全数据资源交易机制和定价机制^[3]。工业和信息化部于2017年1月发布的《大数据产业发展规划(2016—2020年)》进一步指出要研制数据资源分类、开放共享、交易、标识、统计、产品评价、数据能力、数据安全等国家通用标准^[4]。在国家政策的积极鼓励以及地方政府和产业界的带动下，数据交易从概念逐步落地，贵州、武汉等地的数据交易平台先后投入运营，并在数据定价、交易模式、交易标准等方面进行了有益的探索。2018年5月，国家信息中心的大数据研究专栏发表了3篇文章评析我国大数据交易的发展现状和面临的困难，呼吁制定国家层面的法律规范^[5]。

个人数据市场框架

个人数据市场^[6~8]主要有三种类型的参与者：数据贡献者、数据代理商以及数据消费者，框架如图1所示。该框架主要包括数据采集层和数据交易层。

在数据采集层，数据代理商从数据贡献者处采

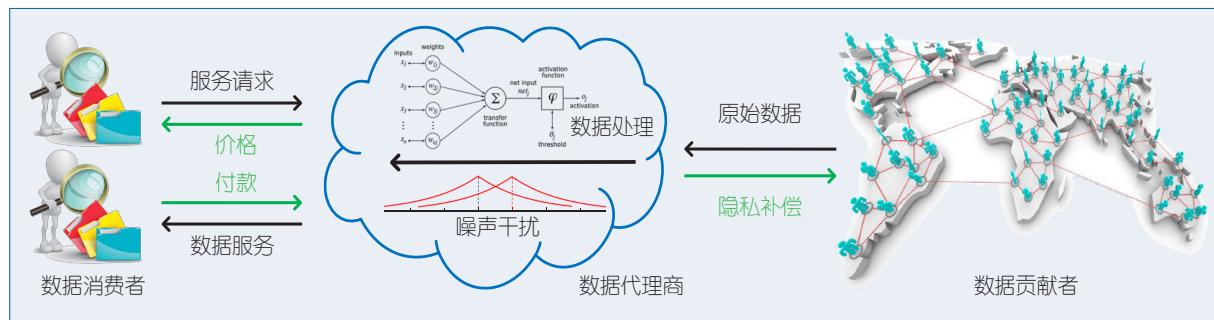


图1 基于干扰型数据服务的个人数据市场框架

集大量的个人数据，例如社交媒体数据、运动轨迹、医疗记录、住宅能源消耗量、用户评分等。由于现实生活中存在社交、行为、基因等多种多样的交互活动，实际数据之间往往存在复杂的关联性，例如，互为好友的两个用户的活动轨迹在时间和空间维度上具有很强的一致性。此外，不同类型的数据具有一些其他的特征，例如，一些用于决策的数据具有很高的时效性；时空感知数据具有不完整、不精确、易错误等特性。因此，实际可行的数据交易框架应该将数据的具体特征考虑在内。

在数据交易层，数据代理商倾向于交易数据服务而非原始数据。这里的数据服务是数据处理产生的结果。相比于交易原始数据，提供数据服务主要有以下优势：对数据贡献者来说，数据服务更能保护隐私；对数据代理商来说，原始数据的所有权和版权难以界定和管理，同时，增值的数据服务的市场需求更大，产生的收益更高；对数据消费者来说，数据服务更能直观深入地刻画整个数据集潜在的特征与规律，充分发挥大数据的实用性。此外，每个数据消费者可以自定义服务请求。值得注意的是，除了具体的数据处理方法，服务请求还包括在返回的数据处理结果中添加噪声的等级，例如噪声服从分布的方差。数据代理商采用随机化机制回答服务请求：返回结果的期望是数据处理准确的结果，而方差不能超过所指定的方差。相比于普通的数据服务，干扰型数据服务允许数据消费者选择适合自己精确度需求的数据服务并支付相应的价格。

根据具体的服务请求，数据代理商一方面向数据消费者收取特定的费用，另一方面需要补偿数据

贡献者的隐私泄露。服务请求中的噪声干扰等级越高，返回的数据服务越不准确，数据消费者需要支付的费用也越低，隐私泄露程度越低，数据贡献者的隐私补偿也应该越少。鉴于整个数据市场框架需要满足收支平衡，即数据代理商的收益必须大于等于零，这也意味着数据服务的价格要大于等于隐私补偿的总和。

关键研究问题

安全可信的数据交易环境

数据市场的安全性研究主要针对私密保护和可验证性这两个有机关联的问题。首先，数据贡献者在数据市场中不希望泄露自己显式的身份标识，例如身份证号、手机号等，即身份私密性；其次，数据贡献者和数据消费者都不希望暴露自己敏感的数据内容，例如运动轨迹、医疗记录等，即数据私密性；再次，数据代理商不希望泄露自己的数据处理模型，例如支持向量机中的支持向量、神经网络中的权值矩阵等，即模型私密性。在保护各参与方敏感信息的前提下，数据交易还有着可验证性的安全需求。数据消费者不仅需要验证数据来源的真实性，还需要验证数据处理结果的正确性与完整性。

研究现状

个人数据有广泛的应用前景，能够提供高价值、高质量的信息，但是如果直接将个人数据用来交易，将存在隐私泄露等问题。近几年，研究者逐渐将目光聚焦到数据市场的安全机制设计。在文献[9]提

出的 DataLawyer 系统中，数据代理商能够显式地制定数据使用规则，并且能够在数据消费者使用数据的时候自动检测数据使用是否满足相应的规则，以确保数据不被非法使用。典型的数据使用规则包括数据溯源、限制查询频率、禁止多元数据聚合等。

中国科学技术大学教授李向阳的团队考虑了不可信的数据消费者二次贩卖数据集的问题^[10]，并将此问题转化成集合相似度的比较问题。他们考虑了文本数据、视频图像数据和图表数据等多种类型的数据。最近，李向阳教授团队针对语音数据提出了保护隐私和可用性的数据发布方案^[11]。他们针对图片数据提出了基于群智感知的大规模、高质量的数据采集方案，并考虑了所有权和隐私等安全问题^[12]。

存在问题：缺少针对数据市场三方模型全面立体的保护机制

已有的相关研究工作主要考虑了数据市场中的单个环节，也相应采用了密码学系统中经典的两方模型，但没有全面考虑数据市场新颖的三方模型，因此很难在现实的数据市场中真正得到应用。数据市场三方模型使得设计安全保护机制有了新的挑战，主要体现在数据贡献者有保护个人隐私的需求，数据代理商有保护数据处理模型私密性的需求，数据消费者有验证数据采集与处理真实性的需求。现有的密码学工具，例如数字签名机制，是在泄露身份隐私的情况下保证数据采集的真实性验证。此外，数据处理的真实性验证与经典的外包计算场景中的可验证性计算^[13~20]有着本质区别，即数据消费者作为验证者并不知道原始数据集与外包函数。

考虑到私密保护与可验证性的内在矛盾性，如何保证数据来源的真实性且维持身份隐私与数据私密性，如何保证数据处理结果的正确性与完整性，且不破坏数据私密性与模型私密性，这两个都是极具挑战性的研究问题。目前在密码学领域出现的理论工作大都只保证其中的一个性质。此外，所假设的系统模型与新颖的数据市场三方模型有着本质区别。因此，我们需要充分考虑数据贡献者、数据代理商、数据消费者的私密保护与可验证性需求，研究实际可用的数据市场保护机制，提供全方位安全

可信的数据交易环境。

初步探索

我们从数据和模型的私密性、数据来源及数据处理结果的可验证性两方面研究具有私密可保护的可验证数据交易机制。

由于需要对数据代理商隐藏数据贡献者的原始数据或数据消费者的输入数据，同时还需要保证数据代理商能够高效地处理原始数据和服务请求，我们无法直接使用传统的对称加密算法或非对称加密算法。对于数据私密性的保护，同态加密技术是一种值得考虑的方法，包括支持在密文上同时进行加和乘运算的全同态加密系统与支持一些指定运算的部分同态加密系统。同时，同态加密技术也可以应用到数据处理模型参数的保护，而不损失数据处理模型在数据消费者这一方的可用性。

数据消费者需要验证数据来源的真实性，这就要求数据贡献者在密文上进行签名。值得注意的是，数据来源的真实性验证类似于数字签名机制中的不可抵赖性，而不是数据的真实性与完整性，因为数据消费者是作为第三方而不是数据接收方来验证数据发送方的合法性。数据消费者验证数据处理结果的真实性与完整性，最为简单直接的方式就是利用同态性质重新计算，并且检查两次结果的一致性。当然在这种朴素的设计方案中，数据消费者需要额外花费昂贵的计算开销，这正是我们的结果验证协议需要避免的地方。

待解决问题

同态加密算法与数字签名算法的结合为敏感数据保护提供了基本的技术思路，但仍存在四方面的问题。第一，数字签名算法采用的是顺序验证的方式，需要花费较大的计算开销，传输数字签名和维护数字证书也需要花费较大的通讯开销。因此，该环节可能会成为大规模数据市场的瓶颈。第二，已有的签名算法大都把签名者的身份标识当作公共参数，而在实际的数据市场中，数据贡献者作为签名者希望保护自己的身份标识。但是，如果隐藏了所有的身份标识，数据市场中的违法数据贡献者就难以被发现，即需要解决身份隐私与可追溯性之间的

矛盾。第三，完全同态加密协议虽然可以支持密文上多种类型的计算，但其较高的计算开销仍然无法很好地适应大规模数据市场的需求。第四，数据处理结果正确性与完整性的证明通常涉及数据消费者的输入数据与数据处理模型的参数，这与保护数据私密性与模型私密性的目标相违背。同时，数据服务请求的大规模性与数据处理模型的复杂性对于验证协议的设计提出很高的性能要求。

自底向上的定价机制

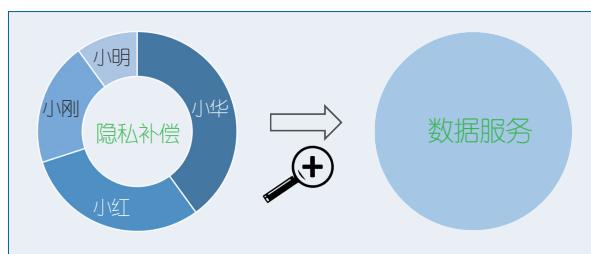


图 2 自底向上的定价机制设计

在数据交易场景下，数据消费者需要获得数据处理的结果，数据代理商则需要衡量每次所提供的数据服务对于每个数据贡献者的隐私泄露程度，并对他们进行合理的隐私补偿。从这个角度来看，隐私补偿机制可以看作传统的激励机制在个人数据采集场景下的变种。如果将隐私补偿机制放入整个数据市场的定价框架中，并采用自底向上的设计思路（图 2），即底层的隐私补偿总和决定上层的数据服务价格。隐私补偿机制是数据市场定价机制的核心与基石。

研究现状

数据市场需要高效的数据采集机制来为数据代理商不断补充优质、海量的数据资源，从而保证数据市场的持续繁荣。众包机制被认为是大规模数据采集的一种有效方法，其核心问题是如何设计激励机制以提高用户参与度。Lee 等人设计了基于动态定价的逆向拍卖机制^[21]，该机制的优化目标是使众包平台数据采集的花费最小化，并保证一定的用户参与量。Jaimes 等人考虑了数据贡献者的地理位置信息，并将数据采集建模成存在预算限制的覆盖最

大化问题^[22]。以上两个工作并没有考虑众包平台中可能存在的操纵策略。Yang 等人将用户的策略行为建模成两种不同的博弈模型：以平台为中心的模型和以采集用户为中心的模型，并分别采用斯塔克尔伯格 (Stackelberg) 博弈和逆向拍卖来设计数据采集机制^[23]。清华大学的杨铮等人提出了三种在线激励机制，以处理数据采集过程中数据贡献者随机出现的情况^[24]。针对个人数据采集的场景，亚利桑那州立大学 Wang 等人研究了在数据代理商不可信的情况下如何通过激励机制购买添加噪声干扰的隐私数据，并建立了博弈模型来衡量隐私的价值^[25]。

由于采集数据的质量参差不齐，众包平台还需要设计评估数据质量的管理方案以引导数据贡献者提供高质量的数据。Liu 等人借鉴在线学习的思想来提高众包平台采集数据的质量^[26]。Karger 等人利用推断算法来检测数据冗余^[27]。此外，众包平台中衡量数据质量最为棘手的问题是真实数据的缺失，即众包平台不仅需要衡量数据贡献者的数据质量，还需要预测真实数据。数据挖掘领域中真值发现 (truth discovery) 框架为处理该问题提供了行之有效的方案^[28]，其核心思想类似于期望最大化算法。然而这些工作并没有将数据质量衡量与酬劳机制联系起来，无法从本质上激励用户贡献高质量的数据。Peng^[29] 和 Jin^[30] 等人考虑了基于数据质量的酬劳激励机制设计。Jin 等人进一步将真值发现拓展到策略博弈环境^[31]。

上述激励机制研究工作的设计目标主要集中在社会效益最大化和数据采集酬劳开销最小化两方面。哈佛大学陈怡玲 (Yiling Chen) 研究组系统地研究了在策略博弈环境下机器学习任务导向型的数据采集方案^[32-34]。文献 [35] 介绍了如何为线性回归模型设计真实可信的数据采集机制。文献 [36] 将该采集机制拓展到更普适的回归模型。文献 [37] 将线性回归建模成非合作博弈模型，并充分考虑用户在数据采集过程中的隐私。文献 [38] 考虑了激励相容条件下的回归学习框架。

存在问题：未考虑数据的实际特征，隐私泄露的量化不准确

初步探索

设计合理的隐私补偿机制的主要挑战在于如何准确地衡量每个数据贡献者的隐私泄露程度，但当前学术界和工业界与个人数据相关工作的主要出发点与落脚点是保护隐私，例如，谷歌^[39]、苹果^[40]、微软^[41]等公司利用差分隐私框架来保护用户隐私。从统计科学角度来看，差分隐私的目标是尽可能多地挖掘关于整体数据集的规律，同时尽可能少地泄露个人的信息。我们近期的研究工作^[8]发现：量化隐私泄露本质上是保护隐私的逆过程。因此，我们可以利用隐私保护机制中的基本原理和准则实现量化隐私泄露的目标。个体隐私泄露可以定义为有无某个数据贡献者的数据对于数据处理结果分布的影响。这里的“有无某个数据贡献者的数据”代表着差分隐私框架下的一对相邻数据库。

在数据关联性方面，我们采用广义的差分隐私框架——河豚隐私 (Pufferfish privacy)^[42,43]。河豚隐私主要通过引入数据分布这一参数来更好地保护关联型数据的隐私。这里的数据分布主要用来形式化数据关联性，例如，社交网络数据中的关联性可以用贝叶斯网络来刻画，而时间序列数据中的关联性可以用马尔可夫链来表达。

待解决问题

差分隐私和河豚隐私为我们定义隐私泄露提供了基本的框架，但距离实际的隐私泄露衡量以及后续补偿机制的设计还有以下关键问题有待解决。第一，计算可行性和计算有效性。如果直接利用上述定义精确地计算隐私泄露，数据代理商需要考虑所有可能相邻数据库的实例，这在大规模数据集上是计算不可行的。第二，从隐私泄露的衡量跨越到隐私补偿机制的设计，需要考虑数据贡献者不同的隐私策略，实用可行的隐私补偿机制需要考虑补偿方案的多样性和可满足性。第三，隐私补偿方案的设计还需要考虑公平性。文献[44]提出原始的公平性是指：如果数据处理没有涉及某个数据贡献者的数据，那么该数据贡献者获得的补偿为零。我们需要论证该公平性的定义是否适用于数据之间存在关联性的情况，如果不适用，应该如何定义广义的公平

性。第四，数据关联性更新算法的高效性。随着时间或地点的变化，数据之间的关联性也会随之改变，重新计算的成本较高，因此需要提出增量式数据关联性的度量方法。第五，隐私补偿的本身也会造成数据贡献者隐私的泄露，需要设计相应的保护机制。

自顶向下的定价机制

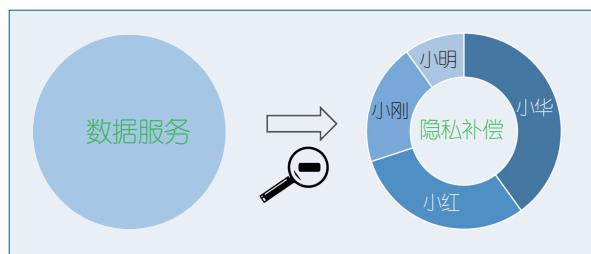


图3 自顶向下的定价机制设计

数据代理商还可以采用自顶向下的数据服务定价机制（见图3）。数据代理商匀出部分数据服务收入作为隐私补偿的预算，并按照隐私泄露的大小决定每个数据贡献者的份额，而数据贡献者不需要主动参与到隐私补偿的过程。如何对干扰型数据服务定价是自顶向下定价机制的核心。

研究现状

数据定价已经成为计算机领域和经济学领域热门的研究话题。近年来，数据库领域涌现出许多研究关系型数据的定价工作。华盛顿大学 Dan Suciu 教授领导的研究组是这个方向的开拓者。在他们最早的数据定价文章中，Balazinska 等人展望了数据交易市场的前景，并且提炼出数据交易这个方向可能的研究问题^[45]。Koutris 等人^[46,47]提出了基于查询的数据定价 (query-based data pricing) 框架，并指出数据定价中两个重要的性质：无套利性 (arbitrage freeness) 和无折扣性 (discount freeness)。文献 [48] 提出了无套利的、适用于任何查询方式的定价函数。文献 [49] 提出了针对动态数据的定价方案。最近，Deep 等人在依赖于结果 (answer dependent) 和独立于实例 (instance independent) 两种不同设定下刻画了无套利定价函数的特征^[50]。基于该理论工作，他们还实现了支持大规模关系查询定价的原型系统^[51]。

上述数据库领域的数据定价相关工作关注的大多是结构化、关系型的通用数据。个人数据已经被很多数据代理商采集和分析，并且售卖给其他数据消费者来进行精准的市场营销^[52-54]。美国纽约大学的 Laudon 教授早在 1996 年就从经济学角度构想了可以交易个人数据的全国性信息市场^[55]，但是关于个人数据定价的严格理论研究则出现在 2011 年：雅虎研究院的 Ghosh 与微软研究院的 Roth 把差分隐私作为量化隐私泄露程度的指标，并提出以拍卖的形式交易隐私数据^[56]。他们主要考虑的应用场景是单次的计数查询。Li 等人的后续研究工作，通过引入无套利的概念将应用场景拓展到多次的线性查询^[44]。Dandekar 等人讨论了隐私拍卖在推荐系统中线性预测函数以及预算限制下的应用^[57]。在近期《美国计算机学会通讯》(CACM) 的两篇文章^[58,59]中，Roth 和 Li 等人总结并展望了个人数据的定价问题。

存在问题：数据服务形式单一，缺少强健的干扰型数据服务定价机制

虽然数据定价的相关研究成果已经有很多，但大都仅适用于数据库查询，而不能直接应用于现实生活中的数据服务。相比于数据库查询，数据服务中所涉及的数据处理方法往往呈现出更加复杂多样的数学形式。例如，在聚合统计场景中，不同的统计方法涉及的算子不尽相同：加权求和涉及的算子是一次多项式，高斯分布拟合中涉及的算子是二次多项式，而度分布中涉及的算子是非线性的比较操作。因此即使为同种类型的数据服务制定统一的定价策略也颇具挑战性。此外，大部分研究工作主要关注的是一般通用数据而非个人的私密数据，因此未考虑在添加不同程度干扰噪声的情况下如何定价。

初步探索

数据服务的定价策略规避套利机会的核心挑战在于解决数据服务之间的相互决定关系，而此问题类似于数据库领域中已有的查询 / 视图的可回答性问题，例如，判断某个 SQL 查询能否通过其他的查询组合进行回答。当然，数据服务呈现出更为复杂多样的形式，之间的决定关系既可能是简单的线性关系，也可能是复杂的非线性关系。关于线性的决

定关系，已有研究工作^[54,69]发现了其与线性代数中的半范数 (semi-norm) 具有结构上的相似性以及理论上的等价性。而对于非线性的决定关系，主要有两种不同的研究思路：第一种是通过引入界面数据库的方式，简化数据处理的中间过程，只考虑数据处理的最外层的数学操作。例如，我们将常见的聚合统计建模成“点积”操作的形式，并建立线性的决定关系^[16]。第二种是通过研究数据处理的计算公式，发现与其具有相似结构的数学概念，最后建立两者之间的等价关系。

关于数据服务定价函数中涉及噪声干扰的部分，我们首先需要确定噪声所服从的概率分布，还需要准确地定义噪声干扰的等级。对于同种数据服务，噪声干扰等级越高，数据服务的精确度越低，所对应的价格也应该越低，即数据服务的价格与噪声干扰的等级之间存在负相关的关系。现有的理论推导结果表明：当噪声方差在数据服务定价函数中相对独立时，数据服务无套利的定价函数随着噪声方差的递减速度不能超过线性^[8,44,59]。

完整的干扰型数据服务的定价函数需要有机地整合数据处理模型以及噪声干扰等级两个部分，并且确定两者之间是否会相互影响，现有的研究工作都是显示地或隐式地假设两部分相对对立。

待解决问题

干扰型的聚合统计定价为我们解决一般性的数据服务定价提供了基本思路，但依然存在问题。第一，从数学上解决非线性函数的决定关系，并将其归约到现有的问题上，例如集合覆盖、网络流问题等。第二，在问题一的基础上，针对具体的、相对复杂的数据处理算法进行合理的建模并设计出鲁棒性的定价机制。第三，如果噪声干扰等级在数据服务定价函数中不是相对独立的，那么如何保证无套利的性质？最后，我们需要权衡无套利的经济学性质所带来的“利”与“弊”。其中，“利”针对贪婪的数据消费者，他们需要拥有较大的算力，花费较大的开销去发现套利的机会，并且他们的攻击开销不能超过所获取的利润。从本质上来说，无套利的定价函数意味着套利攻击的计算不可行性；“弊”针

对数据代理商，他们需要从理论上保证无套利的性质，因此数据服务的定价函数必须保有严格的数学性质，例如次可加性 (subadditivity) 等，这也意味着可供选择的定价函数的种类非常有限。

总结

个人数据交易是目前数据库、数据挖掘、机器学习、网络与信息安全、经济学等多个研究领域中热门的、学科交叉的话题，并得到了学术界、工业界以及政府相关部门的高度重视。在个人数据市场框架中起到基石作用的保护机制和定价机制的相关研究方兴未艾。我们对安全可信数据交易环境的搭建、隐私泄露量化和补偿机制的设计、干扰型数据服务定价机制的设计等三个主要方面进行了初步的探索，希望能激发相关领域研究者的探索兴趣，实现大规模个人数据交易的健康化和产业化，充分释放大数据的潜力，助力国民经济的蓬勃发展。 ■



牛超越

上海交通大学计算机科学与工程系博士生。主要研究方向为数据隐私和可验证性计算。

rvince@sjtu.edu.cn



郑臻哲

CCF 专业会员。2018 CCF 优秀博士学位论文奖获得者。上海交通大学计算机科学与工程系博士后。主要研究方向为算法博弈论、云计算、无线网络。

zhengzhenzhe220@gmail.com



吴帆

CCF 专业会员。上海交通大学计算机科学与工程系教授。主要研究方向为网络经济学、无线网络、移动计算、隐私安全。

fwu@cs.sjtu.edu.cn

其他作者：陈贵海

参考文献

- [1] Federal Trade Commission (FTC). Data brokers: A call for transparency and accountability[OL]. <https://www.ftc.gov/reports/data-brokers-calltransparency-accountability-report-federal-trade-commission-may-2014>.
- [2] The data brokers: Selling your personal information[OL]. (2014-03-09). <https://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/>.
- [3] 国务院关于印发促进大数据发展行动纲要的通知 [OL].(2015-09-05).http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.
- [4] 工业和信息化部关于印发大数据产业发展规划（2016—2020年）的通知 [OL]. (2017-01-17). <http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5464999/content.html>, 2017.
- [5] 国家信息中心大数据研究 [OL].(2018). <http://www.sic.gov.cn/Column/551/0.htm>.
- [6] Niu C, Zheng Z, Wu F, et al. Trading data in good faith: Integrating truthfulness and privacy preservation in data markets[C]//Proc. of ICDE. IEEE, 2017: 223-226.
- [7] Niu C, Zheng Z, Wu F, et al. Achieving data truthfulness and privacy preservation in data markets[J].IEEE Transactions on Knowledge and Data Engineering, 2019, 31(1): 105-119.
- [8] Niu C, Zheng Z, Wu F, et al. Unlocking the value of privacy: Trading aggregate statistics over private correlated data[C]//Proc. of KDD. ACM, 2018: 2031-2040.
- [9] Upadhyaya P, Balazinska M, Suciu D. Automatic enforcement of data use policies with datalawyer[C]// Proc. of SIGMOD. ACM, 2015: 213-225.
- [10]Jung T, Li X, Huang W, et al. Accountrade: Accountable protocols for big data trading against dishonest consumers[C]//Proc. of INFOCOM. ACM, 2017: 213-225.
- [11]Qian J, Han F, Hou J, et al. Towards Privacy-Preserving Speech Data Publishing[C]//Proc. of INFOCOM. ACM, 2018: 1079-1087.
- [12] Zhang L, Li Y, Xiao X, et al. CrowdBuy: privacy-friendly image dataset purchasing via crowdsourcing[C]// Proc. of INFOCOM. ACM, 2018:2735-2743.
- [13]Gennaro R, Gentry C, Parno B. Non-interactive verifiable computing: Outsourcing computation to untrusted workers[C]//Proc. of CRYPTO. 2010: 465-482.
- [14]S. Benabbas, R. Gennaro, and Y. Vahlis, “Verifiable

- delegation of computation over large datasets," in Proc. of CRYPTO, 2011, pp. 111–131.
- [15]D. Fiore and R. Gennaro, "Publicly verifiable delegation of large polynomials and matrix computations, with applications," in Proc. of CCS, 2012, pp. 501–512.
- [16]M. Backes, D. Fiore, and R. M. Reischuk, "Verifiable delegation of computation on outsourced data," in Proc. of CCS, 2013, pp. 863–874.
- [17]D. Catalano, D. Fiore, and B. Warinschi, "Homomorphic signatures with efficient verification for polynomial functions," in Proc. of CRYPTO, 2014, pp. 371–389.
- [18]D. Fiore, R. Gennaro, and V. Pastro, "Efficiently verifiable computation on encrypted data," in Proc. of CCS, 2014, pp. 844–855.
- [19]C. Costello, C. Fournet, J. Howell, M. Kohlweiss, B. Kreuter, M. Naehrig, B. Parno, and S. Zahur, "Geppetto: Versatile verifiable computation," in Proc. of S&P, 2015, pp. 253–270.
- [20]D. Fiore, C. Fournet, E. Ghosh, M. Kohlweiss, O. Ohremenko, and B. Parno, "Hash first, argue later: Adaptive verifiable computations on outsourced data," in Proc. of CCS, 2016, pp. 1304–1316.
- [21]J.-S. Lee and B. Hoh, "Sell your experiences: a market mechanism based incentive for participatory sensing," in Proc. of PerCom, 2010, pp. 60–68.
- [22]L. G. Jaimes, I. Vergara-Laurens, and M. A. Labrador, "A location-based incentive mechanism for participatory sensing systems with budget constraints," in Proc. of PerCom, 2012, pp. 103–108.
- [23]D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing," in Proc. of MobiCom, 2012, pp. 173–184.
- [24]X. Zhang, Z. Yang, Z. Zhou, H. Cai, L. Chen, and X.-Y. Li, "Free Market of Crowdsourcing: Incentive Mechanism Design for Mobile Sensing," IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 12, pp. 3190–3200, 2014.
- [25]W. Wang, L. Ying, and J. Zhang, "The value of privacy: Strategic data subjects, incentive mechanisms and fundamental limits," in Proc. of SIGMETRICS, 2016, pp. 249–260.
- [26]Y. Liu and M. Liu, "An online learning approach to improving the quality of crowd-sourcing," In ACM SIGMETRICS Performance Evaluation Review, vol. 43, no. 1, pp. 217–230, 2015.
- [27]D. R. Karger, S. Oh, and D. Shah, "Efficient crowdsourcing for multi-class labeling," In ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 1, pp. 81–92, 2013.
- [28]Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han, "A survey on truth discovery," ACM SIGKDD Explorations Newsletter, vol. 17, no. 2, pp. 1–16, 2016.
- [29]D. Peng, F. Wu, and G. Chen, "Pay as how well you do: A quality based incentive mechanism for crowdsensing," in Proc. of MobiHoc, 2015, pp. 177–186.
- [30]H. Jin, L. Su, D. Chen, K. Nahrstedt, and J. Xu, "Quality of information aware incentive mechanisms for mobile crowd sensing systems," in Proc. of MobiHoc, 2015, pp. 167–176.
- [31]H. Jin, L. Su, and K. Nahrstedt, "Theseus: Incentivizing truth discovery in mobile crowd sensing systems," in Proc. of MobiHoc, 2017.
- [32]J. Abernethy, Y. Chen, C. Ho, and B. Waggoner, "Low-cost learning via active data procurement," in Proc. of EC, 2015, pp. 619–636.
- [33]B. Waggoner, "Acquiring and Aggregating Information from Strategic Sources," PhD diss., 2016.
- [34]Y. Liu and Y. Chen, "Machine-learning aided peer prediction," in Proc. of EC, pp. 63–80, 2017.
- [35]R. Cummings, S. Ioannidis, and K. Ligett, "Truthful linear regression," in Proc. of COLT, 2015, pp. 448–483.
- [36]Y. Liu and Y. Chen, "A bandit framework for strategic regression," in Proc. of NIPS, 2016, pp. 1821–1829.
- [37]S. Ioannidis and P. Loiseau, "Linear regression as a non-cooperative game," in Proc. of WINE, 2013, pp. 277–290.
- [38]O. Dekel, F. Fischer, and A. D. Procaccia, "Incentive compatible regression learning," Journal of Computer and System Sciences, vol. 76, no. 8, pp. 759–777, 2010.
- [39]Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," in Proc. of CCS, 2014, pp. 1054–1067.
- [40]Apple, "Approach to Privacy," <https://www.apple.com/lae/privacy/approach-to-privacy/>.
- [41]B. Ding, J. Kulkarni, and S. Yekhanin, "Collecting Telemetry Data Privately," in Proc. of NIPS, 2017.
- [42]D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in Proc. SIGMOD, 2011, pp. 193–204.

- [43]D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in Proc. of PODS, 2012, pp. 77–88.
- [44]C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," in Proc. of ICDT, 2013, pp. 33–44.
- [45]M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," PVLDB, vol. 4, no. 12, pp. 1482–1485, 2011.
- [46]P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in Proc. of PODS, 2012, pp. 167–178.
- [47]P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with querymarket," in Proc. of SIGMOD, 2013, pp. 613–624.
- [48]B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," PVLDB, vol. 7, no. 9, pp. 757–768, 2014.
- [49]Z. Liu and H. Hacigümüs, "Online optimization and fair costing for dynamic data sharing in a cloud data market," in Proc. of SIGMOD, 2014, pp. 1359–1370.
- [50]S. Deep and P. Koutris, "The design of arbitrage-free data pricing schemes," in ICDT, 2017, pp. 12:1–12:18.
- [51]S. Deep and P. Koutris, "QIRANA: A framework for scalable query pricing," in Proc. of SIGMOD, 2017, pp. 699–713.
- [52]F. Figueiredo, B. Ribeiro, J. M. Almeida, and C. Faloutsos, "TribeFlow: Mining & predicting user trajectories," in Proc. of WWW, 2016, pp. 695–706.
- [53]J. Staiano, N. Oliver, B. Lepri, R. Oliveira, M. Caraviello, and N. Sebe, "Money walks: a human-centric study on the economics of personal mobile data," in Proc. of UbiComp, 2014, pp. 583–594.
- [54]J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. Oliveira, "Your browsing behavior for a big mac: Economics of personal information online," in Proc. of WWW, 2013, pp. 189–200.
- [55]K. C. Laudon, "Markets and Privacy," Communications of the ACM (CACM), vol. 39, no. 9, pp. 92–104, 1996.
- [56]A. Ghosh and A. Roth, "Selling Privacy at Auction," in Proc. of EC, 2011, pp. 199–208.
- [57]P. Dandekar, N. Fawaz, and S. Ioannidis, "Privacy auctions for recommender systems," in Proc. of WINE, 2012, pp. 309–322.