大数据共享及交易中的机遇和挑战

关键词：数据交易　数据共享　安全计算　隐私保护

李向阳　张　兰　韩　风 等
中国科学技术大学

大数据交易共享现状

人工智能和大数据科学技术的飞速发展在揭示数据本身的属性和规律的同时，也为自然科学和社会科学提供了新的方法，并将给数据的充分利用带来巨大价值。据统计，2015 年全球大数据产业规模达到了 1403 亿美元，预计到 2020 年，将达到 10270 亿美元[1]。然而，在看到无限机遇的同时，我们不得不指出，当前开采的只是数据资源的冰山一角，网络空间中绝大部分数据还分散在一座座属于政府、机构、企业的数据孤岛，甚至是从未开采的数据荒岛。正如李克强总理 2016 年 5 月在全国推进简政放权放管结合优化服务改革电视电话会议上提到，“目前我国信息数据资源 80% 以上掌握在各级政府部门手里，‘深藏闺中’是极大的浪费”。由于数据是非独占性资源，复制成本低，且大数据具有一种稀有的属性——协同作用，即多个数据集作为一个整体的价值要大于各个数据集价值的简单相加，因此使这些隔离封闭的数据开放流通、融合应用能极大提升数据资源的利用价值，这也是大数据时代发展的趋势。

由于数据的潜在价值未知、数据所有者的自私性及对数据隐私安全的担忧等，数据所有者大多不愿免费公开 / 提供自己的数据。为克服上述困难，一种有效途径是将数据作为商品进行交易，数据所有者通过公开 / 提供自己拥有的数据获得收益。

数据交易共享市场现状

随着数据交易和共享的重要性日益凸显，数据交易和共享平台的建设正在进入井喷期，包括 Qlik、CitizenMe、Microsoft Azure Marketplace、DataExchange 等国外平台，以及数据堂、数多多、iDataAPI 和聚合数据等国内平台。此外，我国还成立了一系列政府指导的大数据交易机构，如贵阳大数据交易所和上海数据交易中心。这些平台上的**交易内容**包括数据集、网络爬虫、API、分析报告、解决方案等，**覆盖的领域**包括金融、商业、制造业、地理、交通、天气、电子商务、娱乐、电信、医疗保健、人工智能和各种个人数据（如社交媒体、地点和信用信息），**交易形式**包括交易已有的数据或订制数据。针对不同的交易内容和交易形式，其**定价策略**也不同。通常现成的数据集以固定价格出售，而订制数据的价格由卖方和买方协商确定；API 的定价策略包括“按次支付”（例如，每次执行 0.01 元）和“批发”（例如，每千次执行 10 元）。这些平台上的**数据展示**根据交易内容而不同，包括元数据、统计信息、文本 / 视频说明、数据样本和 API 用法说明。

据不完全统计，2015 年我国大数据交易的市场规模为 33.85 亿元，预计到 2020 年将达到 545 亿元[1]。虽然现有数据交易市场已具有一定规模，但对数据

交易的探索仍处于初期阶段。现有平台通常要求数据拥有者将数据及其描述和价格提交给平台，由平台代为出售。然而数据作为一种特殊的商品具有数量增长快、易复制、质量价值难衡量、权属难确定、渠道难管控、隐私安全风险高等特点，使得当前的数据市场中还存在数据质量保证、价格管控、数据隐私和版权保护等诸多关键问题亟须深入研究和解决。建立高效、可信、公平、安全的数据交易市场仍面临巨大挑战。

数据交易和共享相关政策与法规

通常来说，只要遵守当地的隐私法，各个国家都允许在公司之间进行数据共享 / 交易。此外，当用户明确同意或者数据来自公开网站时，销售用户数据也不存在法律问题。我国政府明确鼓励数据所有者和数据消费者之间共享数据，以充分发挥大数据的潜在价值，并推动技术创新和经济增长。2015 年国务院印发了《促进大数据发展行动纲要》[2]，指出到 2018 年底前要建成国家政府数据统一开放平台，率先在信用、交通、医疗等重要领域实现公共数据资源合理适度向社会开放。2016 年底工业和信息化部发布的《大数据产业发展规划（2016 – 2020 年）》进一步分析了数据共享的问题（共享程度低，规范不健全），明确了开放共享的发展原则[3]。习近平总书记在 2017 年 12 月特别强调数据开放共享和融合作为国家大数据战略一部分的重要性，鼓励政府部门之间的数据共享，以及政府和私营公司之间的数据共享、交易。我国政府还鼓励数据交易，以扩大数字生态系统。

目前还没有专门针对数据共享和交易行为的法规，但大多数国家都有数据和隐私保护的相关法律和机构来监管公司收集、使用和销售相关消费者数据的行为。面对大数据和人工智能产业对个人数据安全带来的挑战，近年来世界各国正尝试修订或增加法律法规中关于个人信息的保护范围及强化保护力度[4]。例如欧盟在 2016 年出台的《通用数据保护条例》(General Data Protection Regulation, GDPR)，旨在加强公司对用户数据使用的管理。在我国，2017 年 6 月《网络安全法》和“侵犯公民个人信息罪司法解释”生效，前者提供了一整套数据保护的规定，后者则明确了哪些侵犯个人数据的行为会构成犯罪。对于涉及知识产权的数据，如著作和发明专利等，各国都有相应法律赋予符合条件的著作者、发明者在一定期限内享有独占权利，并对其所有权和使用权的转让做了详细规定。

数据交易模型

交易模型

数据交易模型分为两种：数据代理模型和 P2P 交易模型，如图 1 所示。

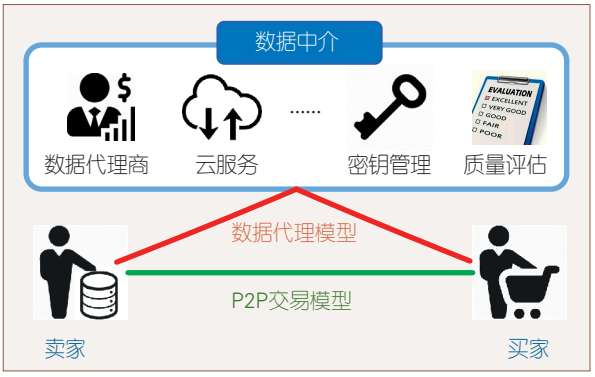

图1　数据交易模型

数据代理模型：在该模型中，数据代理商作为中间平台为买家和卖家提供交易数据的市场。交易平台由多个协作但非串通的实体组成，这些实体可以包括管理交易的数据代理商，提供存储服务的云，以及负责密钥管理、数据质量监控、异常检测、执法、税收的实体等。我们将所有这些中间机构作为一个整体称为“数据中介”。

P2P 交易模型：在该模型中，买卖双方在没有数据代理商的情况下直接交易。其典型示例包括区块链和 P2P 文件共享网络，如 Bit-torrent、eDonkey 和 Pruna 等。

当前大部分数据交易平台都采用数据代理模型，而 P2P 交易模型由于具有低效率和不透明的特性，尚未成为主流。

交易内容

数据交易的内容通常可以分为4种：(1) **数据本身**：买家拥有对数据的永久/指定期限访问权，并可以在数据上执行任意计算以尽可能多地挖掘感兴趣的信息。(2) **数据的直接功能API**：有时买家只对数据的某项简单功能感兴趣，例如搜索结果、统计信息或使用机器学习模型进行训练等。这种情况下，数据平台可以通过提供API来为买家提供相应功能，并限制其对数据的操作。(3) **数据分析结果**：是指从数据中挖掘出来的更高层次的有用信息。例如一个商家希望基于分析得到什么样的用户最可能是其潜在客户，而对原始数据并不感兴趣。(4) **数据衍生物**：与数据内容无关，而是数据的各种权利许可，例如订阅该数据的相关更新，或持续订阅不断产生的数据流，买断数据的所有权或排他的使用权，甚至一些基于区块链的证书（如基于可信飞行记录的飞行员证书）也可以进行交易。

涉及知识产权的数据交易通常是数据的各种权利许可，如作品的版权或专利的工业产权，根据法律规定其著作者或发明者在一定期限内享有独占权，并在其权利有效期内可以转让约定时间或地域范围内的所有权（购买者拥有独占权）或者使用权（分为排他和非排他两种），甚至其衍生品的商品化权（例如电影、动漫周边商品）。

交易形式

当前数据平台有两种主要的交易形式：(1) **交易现有的数据**：由卖家收集、处理数据，并向平台上传展示相关数据信息；买家在平台上检索以选择合适的数据。这种交易形式为卖家节省了订制数据的成本，但买家需要对数据进行进一步处理。(2) **订制数据**：由买家提出对数据内容和格式的要求；卖家根据买家需求整合筛选自己拥有的数据，并为买家提供符合要求的订制数据。这个过程降低了数据交易的效率，但它提高了买家对数据的满意度。

数据定价和支付

交易过程中的关键一环是确定数据的价格，通常分三种情况：(1) **固定价格**：卖家对其数据的供求关系进行市场调查，然后设置固定的出售价格。(2) **变动的价格**：卖家动态决定其出售数据的价格，可以是价格随时间而变化，也可以因买家而异，或因订单顺序而异。(3) **限制出售量和独家买断**：为了增加数据的价值，卖家可以限制出售量，而买家也会独家买断数据以防止其他人获取数据。这种情况通常由买卖双方协商价格。在难以确定数据标价的情况下，拍卖则是一种常见的数据交易形式。

数据交易的流程

数据交易的流程可分为交易前、交易中、交易后三个阶段，每个阶段包含不同的操作和问题，如图2所示。

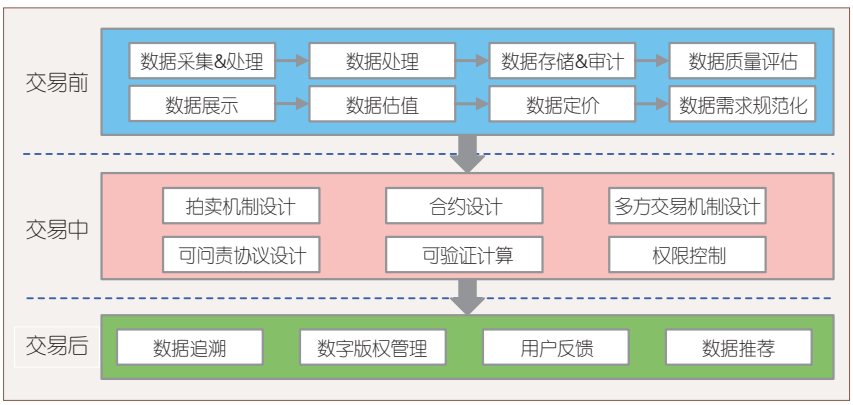

图2 数据交易的三个阶段

数据交易前的问题与挑战

在数据商品被交易前，卖家首先需要收集数据并进行清洗、标记、脱敏等处理，然后卖家需要将数据托管至云端以减少存储和通信开销，并委托代理商进行贩售。数据代理商则需要评估上传至平台的数据质量，并向买家以恰当的方式展示数据。对

于委托平台定价的数据商品，代理商还需要对数据进行估值并设计定价机制，给出合适的价格。买家为了更好地检索或定制目标数据，需要生成准确而规范的需求描述。同时买家也有评估数据质量和价值的需求，以确定自己的预算和出价。在数据交易前的准备阶段还存在以下关键问题和挑战。

数据审计

许多数据拥有者会使用云服务存储数据，但却对云存储服务不完全信任，因为云服务器可能会丢弃很少被访问的数据以降低维护成本。因此，云必须向用户证明数据是按照要求被正确存储的。这一类问题被称为可恢复性证明 (proof of retrievability)、存储证明 (proof of storage)、数据拥有证明 (provable data possession) 或存储审计 (storage auditing)。近年来也有一系列工作研究相关问题，如文献 [5, 6]。在数据交易场景中，数据代理商需要验证两方面的可恢复性：卖家是否在云端存储了声明的数据，以及数据是否可以实现声明的功能或分析结果。当数据加密存储时，代理商还须证明卖家对数据的拥有权（其可解密该数据）。由于数据代理商也并非完全可信，他无法直接访问云上的明文数据甚至是密文数据，这大大增加了可恢复性和拥有权证明的难度。此外，数据集庞大的体量使得传统的加密算法由于较高的计算和通信开销而变得不适用，研究者必须在准确度和计算复杂度之间进行权衡。

数据质量评估

为了提高用户的满意度和平台的声誉，数据代理商和买家需要对出售的数据集进行质量评估。高质量的数据本质上应该具有清晰的表示形式，能被轻松访问，并且适用于买家的任务 [7]。数据质量包含但不限于以下方面：(1) **内在质量**是指数据本身在数量、准确性、完整性、及时性、一致性、清洁度、安全性等方面的质量。(2) **表达质量**侧重于与数据格式（简洁、一致的表示形式）和意义（易于解释）相关的方面。(3) **可访问性质量**强调买家对数据易于获取或检索的程度，例如访问通信延时等。(4) **上下文质量**强调数据与应用场景的相关性。前三个指标应作为数据的一般信息由数据代理商评测并向所有用户公开，最后一个指标取决于买家，因此买家应该提供自己的评估目标场景和相应评估函数。

以前的工作提出了一系列从不同的方面和视角定义的数据质量 [8]，然而面对丰富的数据模态和复杂的数据语义，可定量的数据质量评估还有大量尚未解决的难题。其挑战一方面来自对不同数据形态的各项质量指标的高效准确量化；另一方面来自如何在允许代理商和买家评估数据的同时保护各个主体的隐私，例如应该限制在卖家数据上执行的评估函数的内容和次数，以及保护买家的上下文质量评估函数。目前已有部分工作基于同态加密等方法，为保护隐私的质量评估提供了很好的思想，但距离全面、实用的数据质量评估方案仍有很大距离。

数据展示

为了吸引数据消费者并帮助他们更好地选择产品，代理商需要展示数据的信息（如主题、优点和应用范围）。区别于实物商品，数据具有易复制的特点，因此如果将数据完全展示给潜在买家，买家则可以直接通过复制获取数据而不购买。现有部分工作主要通过数据的采样、数据的版本、元数据和数据的摘要 四种方式进行数据展示。

现有平台通常采用人工的方式生成元数据和摘要，然而面对海量的数据，数据展示的自动生成是提升效率和准确率的重要方法。由于摘要的生成需要对数据进行语义理解，因此一般比元数据生成要困难很多。已有一系列工作采用机器学习等方法为文本、音频、图像和视频生成摘要。为了更快更好地提供数据摘要，除了需要不断改进摘要模型本身，还需要充分考虑方法在大数据集上的执行效率，以及数据的安全性，即不能泄露太多有价值的信息。此外，摘要个性化（生成符合买家需求的摘要以提升销售量）的需求也提升了摘要生成的难度。

数据估值

在交易前，买家、卖家和数据代理商都需要估

算数据商品的价值，以确定自己的购买 / 销售策略。评估无形资产的方法有三种[9]。**(1) 基于成本**：数据的价值取决于收集、处理、存储的成本。由于数据通常是作为信息系统的副产品生成的，数据生产成本与其他产品共享，所以基于成本的方法难以估计数据的真实价值。**(2) 基于市场**：数据的价值取决于同一市场上可比数据的市场价格。但“可比性”的定义不明确，而且对大量数据集而言也不存在相似的数据集，故无法准确估值。**(3) 基于收益**：数据的价值取决于买家能从数据中获得的总收益。这种方法是主观的，仅在评估特定应用时有用，因此不同买方的估值可能会有很大差异。

上述三种传统方法都不能准确地评估数据的全部价值。我们也可以从买家、卖家和数据代理商的角度考虑数据的价值。**(1) 买家的估值**：买家需要评估以给定价格购买数据是否值得，或者在拍卖时应该给出什么价格。买家的估值取决于数据质量以及数据将给他带来的利润。**(2) 卖家的估值**：卖家在出售数据时需要设置最低可接受价格，为此，卖方需要根据成本、数据质量以及数据在市场中的稀缺性来估值。**(3) 数据代理商的估值**：代理商需要对数据进行估值以检验卖家设定价格的合理性。一般来说，代理商需要根据数据在市场中的稀缺性和数据质量进行评估。

数据估值的难度来源于三个方面：(1) 买家和数据代理商不能直接访问数据，且质量评估功能受限；(2) 估值依赖于买家的应用场景和需求；(3) 数据的稀缺性和重要性难以衡量。此外，数据功能和分析结果的估值同样重要且难以实现，在限量出售、独家买断、代理权和所有权转让的场景下的数据估值也有待研究。

数据定价

对于数据定价，我们需要回答三个问题，第一个问题是由谁来设定价格？第二个问题是采用三种定价策略中的哪一种？第三个问题是什么是合理的价格？

采用固定价格时，价格取决于供需关系。针对**供应估计**，由于数据是非独占资源，所以理论上它具有无限的供应；销售数据几乎没有边际成本（卖家只需向买家提供云上数据的密钥）；因为完全竞争市场，Arrow-Debreu 均衡价格几乎为零[10]。针对**需求估计**，一种思路是先在没有价格的平台上列出数据，等待客户表达他们的兴趣，然后估计需求。然而，买家可以串通起来隐藏兴趣以降低价格，或者相似类型数据集的卖家可以协商减少销售量以抬高价格。所以我们不能简单地使用供需规律来确定数据的价格。对于变动的价格，如果价格随时间变化，可以使用数学模型预测其变化。如果卖家和数据代理商想要设定针对买家的个性化价格，则需要获知买家的应用场景。一般而言买家不愿透露应用信息，但可以通过买家之前运行的上下文质量评估函数来推测其应用。针对不同的购买顺序，模拟数据价值随购买人数增多而贬值的过程。对于限量出售、独家买断、代理权和所有权转让等场景，数据是稀缺资源（供小于求），可以进行拍卖以确定其价格。在限量出售时，出售份数的选择是一个问题，同时也存在着卖家进行非法垄断（采用减少销售量的方式收取更高价格）的问题。由于数据具有非常高的固定成本和极低的边际成本，数据定价不仅取决于生产成本和市场竞争，还取决于买家对数据的价值。而如何为不同类型的数据定价，究竟什么是合理的价格仍是尚待探索的问题。

在数据交易前，数据的发现与获取也是一个重要课题。例如基于本体 (ontology) 的数据发现和信息集成，以及基于众包 (crowdsourcing) 的数据获取[17]，都能为交易带来更丰富的数据。数据隐私甄别和脱敏也是保障交易安全进行的重要前提。此外，还存在数据的收集处理、存储管理以及买方需求规范化等问题。解决好这一系列难题，才能为高效、准确、公平、安全的数据交易共享做好准备。

数据交易中的问题与挑战

在数据交易过程中，**拍卖**是一种重要的交易手段，我们需要为具有不同特性的数据设计合适的拍卖机制，并在此过程中考虑用户的隐私问题。交易

达成时，**合约**作为具有法律约束力的协议，需要被妥善设计以保护各个主体的权利。同时，除单用户购买外，数据集贩卖往往也存在**团队购买和捆绑销售**等情况。此外，对交易过程中不诚实的用户要通过设计**可问责协议** (accountable protocol) 来进行惩罚，以保证良好的数据交易生态环境。

拍卖机制设计

当需求多于供应时，可以用拍卖来决定卖给谁和收取多少费用。精心设计的拍卖机制可以将社会福利（每个实体的收益 / 效用的总和）或单边收益最大化。经典的拍卖机制包括英式拍卖、荷兰式拍卖和 Vickrey 拍卖等。数据拍卖的挑战之一是感兴趣的买家可能不会同时表现出他们的兴趣，因此需要很长时间才能将足够多的竞标者聚集在一起进行拍卖。对此，我们可以使用两种可能的拍卖模式：实时决策和延时决策。然而实时决策难以保证出售价格 / 利润最优，延时决策则面临买家可能会放弃等待，同时数据也可能贬值等。另一个挑战来自有时买方希望将出价作为隐私信息进行保护，因此保护隐私的拍卖机制也很重要。对于数据交易的拍卖机制的设计，基于机制设计的理论 [11]，我们希望其具有**激励兼容** (incentive compatibility)、**联合预防** (coalition–proofness)、**个人理性** (individual rationality)、**预算可行性** (budget feasibility) 以及**计算效率高**五个属性。为了建立更诚实公平的数据拍卖市场，我们应该针对不同的数据特性和用户需求设计相应拍卖算法，并使其尽可能满足以上性质。

合约设计

合约是明确规定了法律强制执行相关方的权利和义务的协议。在数据交易中，合约的目标可以是买家、卖家、数据代理商或他们共同的效用最大化。首先需要回答的两个问题是如何定义这些实体的效用？应优先考虑哪个实体的效用最大化？此外，合约设计的最大挑战之一是信息不对称带来的对单方利益的损害，因此需在设计合约时保证激励兼容性，并能够奖励良好的行为，并惩罚坏的行为。另一方面的挑战是我们还需考虑均匀分配风险以使得合约更公平。例如在合约签署后，买家可能重估由数据获得的利润，重估的结果可能高于或低于之前的估值，因此买家或卖家中某一方的利益很可能受到损害。一种解决方案是让买家根据交易后的实际利润而非初始估价向卖家付款。合约设计已经得到了广泛的重视。Bolton 和 Dewatripont 考虑了多种信息不对称的情况，通过将单边或联合效用最大化引入最优合约 [12]。最近，还有一系列工作利用人工智能和区块链等技术设计合约，例如基于区块链的智能合约进行设计 [13]，实现在不需要可信第三方的情况下履行合约和追踪交易。

多方交易

团购已经成为电子商务中一种流行的交易模式。在团购中，一群买家在发起者的组织下会联合起来与卖家协商并获得低于零售价的折扣价。团购同样适用于数据交易，但需要考虑几个问题，如：(1) 什么是适当的折扣，以使得品牌效应的长期收益大于折扣造成的短期利润损失？ (2) 折扣对不同应用需求的买家应该是相同的还是有差异的？ (3) 如何保证整个流程的隐私安全？由于不同买家可能对相同数据的估值不同，要求他们为此支付相同金额是不公平的，我们可以通过公平划分理论将团购的总收益分配给买家，例如 Shapley value[1]。

同理，捆绑销售也适用于数据交易，即把来自相同或不同卖家的类似或相关的数据集打包出售给买家。数据交易场景中的捆绑销售存在着以下挑战：(1) 数据集的量大且种类繁多，代理商很难提供适当的捆绑销售方案；(2) 由于数据的协同作用，组合的

[1] 是指所得与自己的贡献相等的一种分配方式。普遍用于经济活动中的利益合理分配等问题。最早由美国洛杉矶加州大学教授罗伊德·夏普利 (Lloyd Shapley) 提出。Shapley 值法的提出给合作博弈在理论上的重要突破及其以后的发展带来了重大影响。

数据可能具有更高的价值，导致数据集的估值和定价变得更加复杂；(3) 如果代理商将多个涉及重合数据对象的数据集捆绑销售，则买家可能会从中推断出较多数据对象的隐私信息；(4) 利用组合拍卖来销售多个商品的问题通常是 NP 难的。除以上两种模式外还有更复杂的交易模式，例如它们的混合模式：一群买家团购捆绑销售的商品再内部分配。

可问责协议设计

可问责性 [14] 是一种性质，强调应该指责行为不端的协议参与者。我们可以为数据交易设计可问责协议，以迫使所有参与者诚实地遵循协议。要惩罚行为不端的参与者，首先，我们应该能够检测不良行为的后果，例如一些协议的期望目标没有实现。然后，找出谁应对此负责。最后，根据签署的合约惩罚不端参与者。可问责协议的设计有两个目标：公平性和完整性。公平性要求诚实的参与者不应该受到指责，而完整性要求所有行为不端的参与者都应受到指责。在过去十年中，有一系列工作致力于不同协议的可问责性的研究。2010 年，Küsters 等人 [14] 给出了可问责性的正式定义，并展示了几个可问责协议的设计案例。Jung 等人 [15] 设计了一个简单的可问责数据交易流程。

交易后的问题与挑战

在交易完成之后，数据代理商还需进行质量认证和保护数据版权，监督买家和卖家的行为，观察他们是否履行了合约中关于数据质量的承诺和传播的限制（卖家不能卖多，买家不能转卖），但由于数据抄袭难以界定以及离线传播难以追踪，实现数据追溯仍面临着许多挑战。成熟的电子商务市场中，评价系统和推荐算法是促进商品交易进行的重要力量，在数据交易环境下，代理商也需要快速准确地为买家推荐感兴趣的数据集。

数据追溯

在一个良好的可持续发展的数据交易市场中，交易完成后数据传播的可追溯性对整个系统的可靠性至关重要，它决定了用户对系统的满意度和信任度。设计可追溯的数据交易机制是困难的。首先，难以保证数据的可追溯性，因为攻击者可能会采取任意措施来避免数据在传播中被跟踪和识别；其次，抄袭是难以检测的，因为用户可能会修改一小部分数据，然后将其列为市场上的“新”数据集；最后，数据代理商难以检测离线的非法数据交易。

一个想法是引入第三方可信机构监督市场上的所有交易，为每笔交易的数据附加水印，在准许参与者的操作之前先验证数据的现有水印，以检查是否违反了交易政策。然而，串通的参与者和离线数据流通会绕过监视。部分现有的数据交易平台声称已将区块链技术用于数据追溯，例如贵阳大数据交易所和京东万象，利用了区块链分布式数据存储的不可篡改、可追溯、可信任等特性。针对数据抄袭检测，Jung 等人 [15] 设计了相关技术，为了考量数据的原创性，他们定义了各种数据类型的原创性指数，并实验验证了该指标的有效性。还有一些可用于数据篡改检测和数据查重的工具，如默克尔树 (Merkle tree)、数字签名和局部敏感哈希 (LSH) 等。

数据版权管理

卖家可以通过数据免费试用或者限期免费退款来帮助买家买到合适的数据，并提高销售额，但如何在试用结束或退款之后确保买家删除数据呢？买家可能不仅不删除数据，而且持有该数据的拷贝或稍加改动的版本，还可能将数据转移至其他存储设备或其他人。由于数据本身易拷贝、易更改、易转移，这些侵权的风险都无法消除。这些风险同样存在于现有的数字商品中，如电子书、音乐、电影。现有的数字版权管理 (Digital Rights Management, DRM) 通过加密和开发专用的软硬件来保护数字商品的版权，比如只能用特定的软件来看电子书或听音乐并且不允许下载，通过产品密钥、限制软件安装次数、持续在线身份验证等方法自动检测盗版行为。在数据交易中，直接将明文数据发给买家会导致无法恢复的侵权损失，所以我们必须利用 DRM 技术来限

制数据的使用、下载和传播。

现有的DRM技术需要做出以下改进以支持数据商品。第一，数据的访问不像流媒体那样是连续的，可能是任意的，所以现有技术对于数据可能难以实现高速的实时在线访问(streaming)。第二，现有带版权管理的专用软件一般只允许用户浏览（如看视频、听音乐），而在数据交易中，这样的软件不仅要支持浏览，还要允许买家对数据做计算和可视化等。第三，需要禁止截屏功能并利用一些机制（如文献[16]）阻止买家对屏幕拍照或录像而间接地侵权。第四，不仅要利用产品密钥和持续在线身份验证等机制来防止侵权，还需要检测侵权是否已经发生，例如，在数据被传输时记录发送设备和接收设备等信息,结合数据的交易合约中的版权限制，软件应该自动判断买家是否已侵权，如果是，应该通过扰乱数据或完全禁止使用等方式惩罚买家。

数据推荐

许多现有的交易平台一直在使用推荐系统来帮助用户找到他们可能喜欢的新商品。流行的推荐系统可分为三类——基于内容的过滤，协同过滤，以及它们的混合体。数据交易平台同样可以通过推荐系统来推动商品交易。在买家有历史订单记录时，分析他的兴趣是较容易的。对于基于内容的过滤，我们虽难以直接衡量数据集、函数和分析结果之间的相似性，但可以通过机器学习的方式推断它们的相关性。如果买方是新客户，代理商可以应用协同过滤来找到相似用户，然后向他/她推荐他们购买过的东西。但此时需要为每个客户建立准确的资料，这可能会带来一些隐私问题，代理商需要通过隐私保护的算法匹配用户，采用安全多方计算或同态加密计算买家的相似性。此外，买家较小的购买频率和不完整的个人信息也加大了代理商做出准确推荐的难度，而数据的稀疏性是研究人员一直在努力克服的问题。

基于区块链的数据交易和追溯

近年来区块链技术的飞速发展为数据交易和追溯提供了新的思路。经典的区块链，即2009年中本聪(Satoshi Nakamoto)在比特币系统中使用的区块链，融合了非对称加密、数字签名、默克尔树、工作量证明等多种技术，为在无可信中心情况下转账信息的安全、可靠记录，提供了一套完整的解决方案。而后，研究者们对经典区块链进行了不同的修改和补充，以使其适应不同场景下的各项信息分享和记录的要求。2013年，由19岁俄罗斯少年维塔利克·布特林(Vitalik Buterin)提出的开源具有智能合约功能的公共区块链平台——以太坊(Ethereum)，在保有之前比特币区块链的支付转账功能基础上，提供了一个开放的、模块化的支持自定义高级应用的平台。以太坊支持用户编辑自己想要的应用，也就是智能合约。合约的调用过程和返回结果被记录在底层区块链中，一样的安全、可靠和不可篡改。

基于区块链及其相关技术，我们可以提出多种可能的数据交易和追溯的解决方案。例如，通过构建一个基本的区块链，即可完成分布式数据交易的基本功能；或者通过使用以太坊平台发行代币的方式，将代币和数字资产绑定，以实现数字资产的证券化和公开化；再或者通过为每一份数据建立唯一的数据档案合约，将该份数据的相关信息记入与之绑定的档案合约，对该份数据的买卖通过调用档案合约的不同功能来实现，将保有内容标签信息和数据版权的记录上链，利用链上信息的不可篡改性实现数据的防伪和版权确认。数据交易中介方还可以在本地建立合约仓库，将链上的无序合约在链下进行有序组织，通过链上链下结合的方式实现高效服务。

总结与展望

从以物易物开始，实体商品的流通及其交易市场的进化已经持续千年，并一步步演变到现今的经济全球化。可以预见数据交易也将不断发展，为国家、企业和个人带来更多的价值。对数据开放共享的急迫需求已经催生了一系列数据交易和

共享平台，但数据交易市场仍处于初级阶段，整个数据交易的流程仍然面临着许多法律及跨学科的难题和挑战，而这也代表着前所未有的机遇。随着相关研究的展开，相信总有一天我们会建立成熟的数据交易生态系统，为社会发展带来一片新的动力和繁荣。 ■

李向阳

CCF专业会员、CCCF编委。中国科学技术大学教授，国家千人计划专家，ACM中国共同主席，IEEE Fellow。主要研究方向为大数据的共享交易和隐私保护等。xiangyang.li@gmail.com

张　兰

CCF专业会员。CCF优秀博士学位论文奖获得者，阿里巴巴青橙奖获得者。中国科学技术大学特任教授。主要研究方向为跨域数据的深度理解、隐私保护和数据交易。zhanglan03@gmail.com

韩　风

中国科学技术大学硕士研究生。主要研究方向为数据理解、数据交易、安全隐私。hf1996@mail.ustc.edu.cn

其他作者：薛爽爽　钱建威　郑　翰

参考文献

[1] 2016年中国大数据交易产业白皮书 . http://www.cbdio.com/BigData/2016-06/02/content_4965656_all.htm.

[2] 国务院关于印发促进大数据发展行动纲要的通知 . http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm, 2015.

[3] 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 . http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5464999/content.html, 2017.

[4] 国家信息中心大数据研究 . http://www.sic.gov.cn/Column/551/0.htm, 2018.

[5] He K, Chen J, Du R, et al. Deypos: Deduplicatable dynamic proof of storage for multi-user environments[J]. *IEEE Transactions on Computers*, 2016, 65(12):3631-3645.

[6] Yu J, Ren K, Wang C, et al. Enabling cloud storage auditing with key-exposure resistance[J]. *IEEE Transactions on Information Forensics and Security*, 2015,10(6):1167-1179.

[7] Wang R Y, Strong D M. Beyond accuracy: What data quality means to data consumers[J]. *Journal of Management Information Systems*, 1996,12(4):5-33.

[8] Batini C, Cappiello C, Francalanci C, and Maurino A. Methodologies for data quality assessment and improvement[J]. *ACM computing surveys* (CSUR), 2009, 41(3):16.

[9] Rezaee Z. Intangible asset valuation[M]// *Financial Services Firms: Governance, Regulations, Valuations, Mergers, and Acquisitions* (Third Edition). 2012:331-344.

[10]Quah D. Digital goods and the new economy[C]// CEPR Discussion Paper No. 3846, 2003.

[11]Nisan N. Algorithmic mechanism design[M]// Young P, Zamir S, et al. *Handbook of Game Theory*. Amsterdam: North-Holland, 2014:477.

[12]Bolton P, Dewatripont M. *Contract Theory*[M]. MIT press, 2005.

[13]Kosba A, Miller A, Shi E, et al. Hawk: The blockchain model of cryptography and privacy-preserving smart contracts[C]// *IEEE Symposium on Security and Privacy*. IEEE, 2016: 839-858.

[14]Küsters Ralf, Truderung Tomasz, and Vogt Andreas. Accountability: definition and relationship to verifiability[C]// *Proceedings of the 17th ACM conference on Computer and Communications Security*. ACM, 2010: 526-535.

[15]Jung T, Li X Y, Huang W, et al. AccountTrade: Accountable protocols for big data trading against dishonest consumers[C]// *IEEE Conference on Computer*. IEEE, 2017: 1-9.

[16]Zhang L, Bo C, Hou J, et al. 2015. Kaleido: You can watch it but cannot record it[C]// *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. ACM, 2015: 372-385.

[17] Zhang L, Li Y, Xiao X, et al. CrowdBuy: Privacy-friendly Image Dataset Purchasing via Crowdsourcing. IEEE INFOCOM 2018.