数据要素流通的分账机制研究

顾 勤[1]，周 涛[2*]

(1. 成都大数据股份有限公司 成都 610095；2. 电子科技大学大数据研究中心 成都 611731)

【摘要】推动数据要素流通是进一步释放生产力的重要手段。与一般商品交易不同，数据的请求、寻址、响应和传输过程较为复杂，一次完整的请求可能需要多个数据源同时进行响应甚至竞价。因此，构建数据要素流通体系的一个基础性的问题就是如何建立一套数据要素流通的分账机制。该文分析了典型的数据请求和供给模式，建立了包括请求端节点、中介节点和响应端节点的激励网络模型，设计了几何衰减的分账机制。在上述具有普适性的框架下，该文给出了几种常见情况下如何分账的具体计算过程，并将该机制推广到了一次数据请求需要多个数据源响应且各自贡献不同的含权情境。文末讨论了如何在此框架下包容更复杂的情况，包括如何处理多数据源竞价响应的复杂情况。

关 键 词 大数据； 数据流通； 生产要素； 分账机制； 激励网络

中图分类号 TP391 **文献标志码** A doi:10.12178/1001-0548.2021005

On Credit-Splitting Mechanism in Reponses to Data Queries

GU Qin[1] and ZHOU Tao[2*]

(1. Chengdu Big Data Incorporated Company Chengdu 610095;

2. Big Data Research Center, University of Electronic Science and Technology of China Chengdu 611731)

Abstract Data circulation is a novel and important means to facilitate productivity. Different from the trade of normal products, the requesting, addressing, answering and transmitting of data involve complicated procedure, and a single requirement may lead to multiple answers or even bids. Accordingly, a fundamental issue in building up an efficient and effective system for data circulation is to design a credit allocation mechanism. This paper analyzes the typical pattern of data demand and data supply, proposes an incentive network model containing requesting node, intermediary nodes and answering nodes, and designs a geometrical decaying mechanism in credit allocation. Under the above general framework, we show some typical models and the corresponding calculation processes, and extend the single chain model to the general situation involving multiple answering nodes with different weights. Lastly, we discuss how to deal with more complicated cases under this framework, such as allowing bids in competition of multiple nodes.

Key words big data; credit allocation mechanism; data circulation; incentive networks; production factors

2020 年 4 月 9 日，中共中央、国务院印发《关于构建更加完善的要素市场化配置体制机制的意见》(以下简称《意见》)，明确了要素市场制度建设的方向及重点改革任务，并就扩大要素市场化配置范围、促进要素自主有序流动、加快要素价格市场化改革等作出了部署。《意见》首次将数据明确为与土地、劳动力、资本和技术并列的新型生产要素。数据作为生产要素参与分配具有突破性的意义，有望快速推动数据确权、数据交易和数据资本化。譬如技术作为生产要素地位的明确，就为技术的有价转让以及以知识产权作价作为股本金出资奠定了基础。如何搭建合规且高效的数据要素流通体系，是《意见》出台后亟待回答的关键问题。

数据要素的流通方式主要包括开放、共享和交易。数据开放是指向不特定主体开放的非涉密非隐私数据，一般不收取费用。某些情况下开放是面向受限主体或者有前提条件的，譬如有些科学数据的开放需要使用方提前说明使用方式并承诺不用于商

收稿日期：2021－01－05；修回日期：2021－01－09
基金项目：国家自然科学基金 (11975071)
作者简介：顾勤 (1973－)，女，主要从事数字产业和数据要素流通方面的研究和实践.
通信作者：周涛，E-mail：zhutou@ustc.edu

业目的。数据共享是指在协议或约定条件下，数据在有限主体间共享，一般也不收取费用。参与共享的主体往往同时也是数据的提供方。其他需要支付费用才能获得数据的流通方式，往往都被归为数据交易。数据交易的方式很多，包括批量下载 (大量数据一次性付费下载，如遥感数据)、权限使用 (根据权限查阅和下载数据，一般对于线程数和下载量有限制，如高校购买的电子出版物和经济社会数据集等)、API 查询 (通过接口查询，一般返回简单的是否或数值，按照查询次数付费)、API 调用 (通过接口进行下载，一般按照下载量付费)、沙箱服务 (在约定的数据环境和数据格式下进行运算并获取结果，不直接得到数据本身) 等。如果只是简单和传统的生产要素做类比，通常会认为交易才是数据作为生产要素流通的方式。但实际情况并非如此，开放的数据也可以作为重要的生产要素，如疾病致病基因的发现，需要人类表型本体 (human phenotype ontology) 数据；又如先导药物分子的发现，往往要用到大量开放的有机化学方程式库。共享的数据很多也是典型的生产要素，如多家金融机构在一定的协议约定和隐私保障下，通过数据共享可以提高风险识别的准确度，提升反欺诈、反洗钱和普惠金融服务等能力。事实上，不同于一听可乐或者一件衣服，数据很少成为最终的消费品，大部分数据的需求方都是将数据作为进一步生产的原材料，或通过对数据的利用提升决策水平、业务能力、服务效率等，这正好也是生产要素的特点。

与普通商品交易不同，随着数据需求深度和广度的增加，数据交易的结构可能非常复杂。如采集数据需求的平台可能并不具备部分或者全部的数据，数据的需求可能需要多个分布于不同位置的数据源的组合才能满足，还需要大量中介节点分解和传递数据需求、需求响应情况以及数据本身。在满足数据需求的过程中，不同数据源的数据贡献程度可能差异很大，不同数据源还可能针对同一项数据需求开展竞价。如针对罕见病的研究需要不同国家地区的多个医疗机构提供病例数据，又如对企业的深入尽调需要调取在不同地区注册的目标企业及其投资对象的多维数据。为了应对这些复杂的情境，充分发挥完成一个数据请求所涉及的多个异质主体的积极性，亟需设计一套数据要素流通的分账机制，这也是保障数据要素有效流通的基础性问题之一。

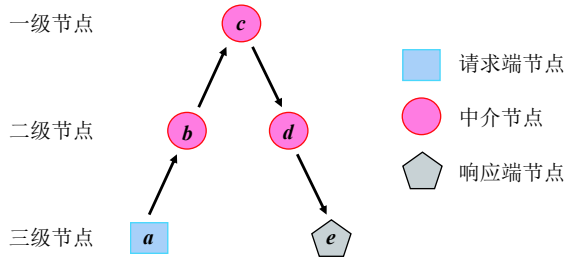
本文分析了典型的数据请求和响应模式，借鉴了 P2P 文件共享系统中请求响应的激励机制[1] 和单任务的链式衰减激励机制[2]，建立了包括请求端节点、中介节点和响应端节点的激励网络模型，设计了几何衰减的分账机制。在上述具有普适性的框架下，本文给出了几种常见情况下如何分账的具体计算过程，并将该机制推广到了数据请求需要多数据源响应且各自贡献不同的含权情境。文末讨论了如何在此框架下包容更复杂的情况，包括如何处理不同数据源针对同一数据需求进行竞价的复杂情况。

1　基本模型

一个具备数据需求分发和响应的数据要素网络至少应该包含 3 类节点：1) 请求端节点：用于采集需求方的具体需求，一般为功能性的平台，允许需求方提出数据申请，如金融机构希望获取某申请贷款企业 x 所有直接和间接控股的企业集合 $O(x)$，以及 $x\cap O(x)$ 近 3 年的纳税记录；2) 中介节点：根据协议和/或算法将未满足的数据需求转发给一个或多个其他中介节点或者响应端节点；3) 响应端节点：数据源所在地，根据数据需求提供相应的数据。注意，一个节点可能同时扮演多种角色。如请求端节点可能也拥有数据源，能够响应数据需求；而如果请求端节点不具备应对需求的完备数据，则必然也是中介节点。又如很多中介节点也是响应端节点，只是将本地无法满足的需求分发出去。

首先考虑最基本的模型，其中请求端节点收到数据需求后，通过若干中介节点的转发，最后由一个响应端节点满足其需求。在基本模型中，假设所有的数据需求一个响应端就可以全部满足，更一般化即数据需求需要多个响应端协同的情况，将在下一节讨论。因此，数据需求被满足的过程可以用一条“请求–转发–响应”链条来描述，其中需求信息从请求端到响应端所需转发的次数被称为该链条的长度。记一次成功的需求响应所有节点总的贡献为 1，每个节点分账的比例与其贡献的比例一致。如果请求端本身就有所需要的数据，自身就可以响应，则不需要任何中介节点，链条长度为 0，请求端节点完成了所有的贡献 1。一般情况下，链条的长度大于 0。譬如未来公共数据的流通体系很可能是层次架构的，某城市 a 的企业在办理业务时需要调用与城市 b 有关的数据，需求可能在城市 a 的平台提出，被转发至城市 a 所属的省级行政区节点 A，如果 A 没有相关的数据，可能要继续转发到国家中心节点 C，C 根据寻址的规则找到 b 所在省级

行政区节点 B，然后再转至城市 b 的数据中心，实现成功响应并原路回传数据。这样就形成了一个长度为 4 的链条“*a-A-C-B-b*”。图 1 给出了一个按上述层次结构组织形成的长度为 4 的“请求–转发–响应”链条示意图。注意，即便不是按照层次结构进行组织，基本模型也是完全适用的。本文给出 3 种普适性很强的简单模型。



采用模型	参数选择	$C(a)$	$C(b)$	$C(c)$	$C(d)$	$C(e)$
几何衰减	q=0.5	1/31	2/31	4/31	8/31	16/31
激励动员	N/A	1/16	1/16	1/8	1/4	1/2
固定收益	r=0.25, q=0.5	0.25	0.05	0.1	0.2	0.4

图 1 一个层次组织的长度为 4 的“请求-转发-响应”链条示意图以及在 3 种基本模型下 5 个节点贡献的比例

1) 几何衰减模型。该模型认为响应端节点的贡献最为显著，其次是将需求转发给响应端节点的中介节点，再次是将需求转发给该中介节点的中介节点，以此类推。按与响应端节点距离由近到远，贡献按照几何级数衰减，而请求端节点仅仅被看作一个普通的中介节点。记“请求–转发–响应”链条长度为 L，衰减指数为 $q(0<q\leqslant 1)$，则与响应端节点距离为 $d(0\leqslant d\leqslant L)$的节点 i 的贡献为：

$$C(i)=\frac{(1-q)q^{d}}{1-q^{L+1}}$$

如响应端节点到自身距离为 0，则其贡献为 $\frac{1-q}{1-q^{L+1}}$；而请求端节点到响应端节点的距离为 L，则其贡献为$\frac{(1-q)q^{L}}{1-q^{L+1}}$。图 1 给出了 L=4，q=0.5 的一个计算示例。

2) 激励动员模型。该模型最早是 Pentland 领衔的 MIT 团队在 2009 年 DARPA 举办的寻找美国大陆 10 个红色气象气球位置的社会动员大赛中使用的策略模型。利用该策略，MIT 团队以显著优势获得了冠军[2]。激励动员模型是一个非参模型，在该模型中，响应端节点的贡献为 1/2，将需求转发给响应端节点的中介节点的贡献为 1/4，将需求转发给该中介节点的中介节点的贡献为 1/8，依此类推。如果“请求–转发–响应”链条的长度为 L，则距离响应端节点为 $d(0\leqslant d<L)$的节点的贡献为 $(1/2)^{d+1}$。请求端节点的贡献和与其相邻的中介节点一致，为$(1/2)^{L}$。图 1 给出了 L=4 的一个计算示例，与几何衰减模型相比，激励动员模型认为请求端节点的贡献不仅仅是一个普通的中介节点，因此略微增加了分配给它的贡献。

3) 固定收益模型。上面两个模型虽然略有差异，但请求端节点分配的贡献比例都是最少的或最少的之一。然而，在互联网时代，流量的获得往往起关键性的作用。固定收益模型认为请求端节点作为流量入口，不能仅仅被看作一个中介节点，而应该享有一个固定比例的贡献值。在该模型中，其他节点的贡献值分配依然按照几何衰减模型，而请求端节点的贡献固定为 r $(0<r<1)$。依然记“请求–转发–响应”链条长度为 L，衰减指数为 $q(0<q\leqslant 1)$，则距离响应端节点距离为 $d(0\leqslant d<L)$的节点 i 的贡献为：

$$C(i)=\frac{(1-r)(1-q)q^{d}}{1-q^{L}}$$

图 1 给出了 L=4, q=0.5, r=0.25 的一个计算示例。

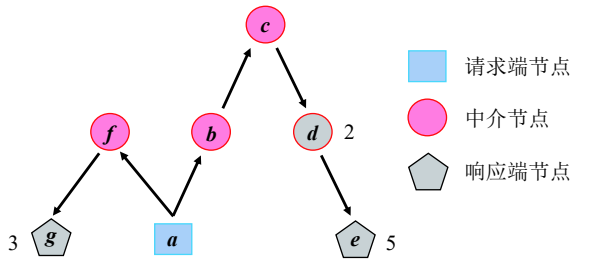
以上给出的是比较简洁，具有相当适用性的若干模型，读者在具体应用场景中还可以根据特殊需求设计更复杂的基本模型。

2 一般模型

基本模型解决的是在一条“请求–转发–响应”链条上，贡献值如何分配的问题。一般情况下，一次数据请求可能需要多个节点提供数据，且所提供的数据的价值不同。因此，对一次数据请求的响应过程可能形成多条权重不同的“请求–转发–响应”链条，这些链条两两之间可以有一个或多个除请求端节点之外的重复节点。这就要求请求端节点具备将任意在其服务范围内合法的数据请求分解成最小粒度的若干数据项需求并为每项需求赋予明确权重的能力。在此基础上，每个响应端节点根据其所满足数据需求的权重，把对应比例的贡献值在相应的“请求–转发–响应”链条上进行分配。分配的机制就是上一节所介绍的基本模型。一个节点的贡献值就是所有涉及它的链条上其贡献值的加和。

图 2 给出了一个典型的示例，其中请求端节点将收到的数据请求拆分成 10 个最小粒度的需求

项。假设这 10 个数据需求的权重相同，在转发过程中，节点 *d* 满足了其中 2 份需求，但是还不能完成所有需求，于是又继续转发给节点 *e*。节点 *e* 满足了其中 5 份需求。还有 3 份需求是节点 *g* 完成的。于是，共有 3 条“请求–转发–响应”链条参与了对该数据需求的响应，分别是“*a*-*b*-*c*-*d*”、“*a*-*b*-*c*-*d*-*e*”和“*a*-*f*-*g*”，其对应的权重分别是 0.2、0.5 和 0.3。按此权重，若采用激励动员模型，则如图 2 所示，7 个节点的贡献值分别为 $C(a)$=0.131 25、$C(b)$=0.056 25、$C(c)$=0.112 5、$C(d)$=0.225、$C(e)$=0.25、$C(f)$=0.075 和 $C(g)$=0.15。



编号	权重	*C*(*a*)	*C*(*b*)	*C*(*c*)	*C*(*d*)	*C*(*e*)	*C*(*f*)	*C*(*g*)
1	0.2	1/40	1/40	1/20	1/10	0	0	0
2	0.5	1/32	1/32	1/16	1/8	1/4	0	0
3	0.3	3/40	0	0	0	0	3/40	3/20
总计	1	21/160	9/160	9/80	9/40	1/4	3/40	3/20

图 2　一个数据请求需要多个响应端节点协同完成的示意图，其中 3 条链路分别的权重是 2、5 和 3。节点 *d* 在第一条链路中扮演了响应端节点的角色，在第二条链路中扮演了中介节点的角色。

显然，采用不同的基本模型，上述按链条进行贡献值分配并根据权重加和的框架也是适用的。

3　结 束 语

针对数据要素流通过程中如何分账的问题，本文提出了一个简单的框架，其核心组件包括：1) 流通网络由请求端节点、中介节点和响应端节点组成；2) 响应端节点贡献大于中介节点，且贡献值按照几何级数衰减；3) 一次数据请求可以由多个响应端节点满足，并根据不同权重进行贡献值的分配。尽管具体模型还可以根据不同场景的需求进行变化，但以上基本思想是具有普遍适用性的，应该能在数据要素流通体系建设中发挥重要的参考价值。

本文一个隐含的假设是中介节点知道如何找到响应端节点，或者说知道如何为一个数据需求在流通网络上寻址。对于一些简单的情况，例如一个城市 A 的数据中心就掌握该城市的所有可流通税务数据，不同数据中心按照行政所属关系形成连接，这种情况下寻址的逻辑就非常简单。然而，实际情况下数据的需求复杂多样，数据的供给方信息并不完备，此时如何给出数据线索，如何寻址，在哪些情况下要采用广播方式等等，都是值得进一步研究的问题。其中，一种更复杂的情况，就是同一个数据需求的细项，有不止一个数据源可以响应。每个得到通知的数据源原则上都可以通过网络竞价。这种情况下，如何设计竞价拍卖的机制以及在该机制下如何确定竞标价格，也是值得深究的问题。特别地，如果一个节点本身可以满足数据需求，它是否还要转发这个需求，就成了有趣的两难选择。一方面它的转发会带来新的竞争对手，造成竞价成功的可能性降低或利润空间降低；另一方面它既无法保证竞价成功，又可以寄望通过它的后继节点或后继的后继等竞价成功而获得相应分成。最近我们设计了一套机制，可以在社会化拍卖的过程中让转发拍卖信息并按照真实意愿出价恰好是纳什均衡，从而提升拍卖的效率和系统整体收益[3]。这些都可能为更好实现数据要素的流通赋能！

致谢：成都大数据产业技术研究院兰宇、清华大学廖敬仪和成都大数据股份有限公司徐忠波亦对本文有贡献，特此感谢。

参　考　文　献

[1] KLEINBERG J, RAGHAVAN P. Query incentive networks[C]//Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science. [S.l.]: ACM, 2005: 132-141.

[2] PICKARD G, PAN W, RAHWAN I, et al. Time-critical social mobilization[J]. Science, 2011, 334(6055): 509-512.

[3] LI B, HAO D, ZHAO D, et al. Mechanism design in social networks[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2017: 586-592.

编　辑　蒋　晓