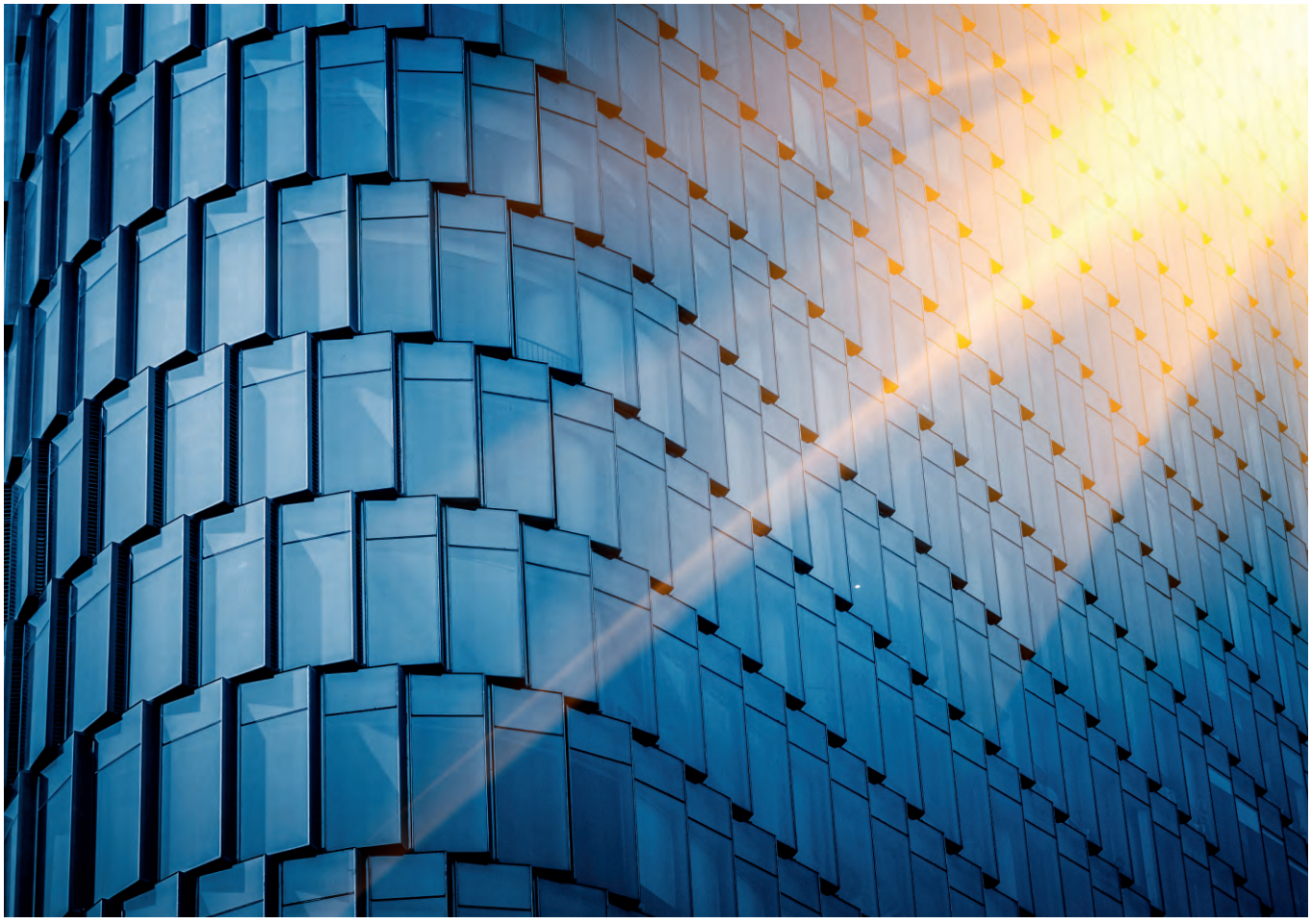
栏目编辑
丁开艳
dingky@pbcsf.tsinghua.edu.cn

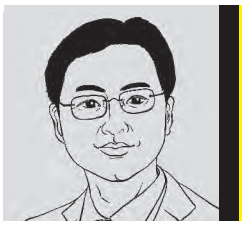
INTERNET FINANCE
互联网金融



DOI:10.19409/j.cnki.thf-review.2018.09.025

区块链技术的激励相容：基于博弈论的经济分析

如果说区块链技术的出现是各类信息技术融合带来的“化学反应”，那么，经济机理则是其中的“催化剂”。区块链技术不仅有技术逻辑层上的支撑，还有经济逻辑层上的保障。本文从经济理性人的角度出发，将“无组织”群体行动区分为三个层次，并利用博弈论，基于经济学的视角剖析了区块链技术的激励相容设计，认为通过激励相容的算法规则和相关契约安排，区块链明确各方的经济利益，充分调动各方的积极性，使分布式协同作业真正成为可能。



Introduction

姚前

密码学是区块链的根基，它集成各类密码学原语和方案，同时还应用了P2P网络协议、智能合约等技术。如果说区块链技术的出现是各类信息技术融合带来的“化学反应”，那么，经济机理则是其中的“催化剂”。区块链技术不仅需要技术逻辑层上的支撑，还需要经济逻辑层上的保障，否则难以在“去组织”化的环境下开展分布式协同作业。因此，加密技术与经济机制设计的结合使区块链成为具有巨大潜力的新兴技术，被认为是继大型机、个人电脑、互联网、移动互联网之后计算范式的第五次颠覆式创新，有望重塑人类社会活动形态从目前的信息互联网向价值互联网的转变。

“无组织”群体行动需要经济激励

从Linux的成功案例来看，基于互联网的社区形态和技术的发展，已使得松散的个体可以不通过命令式管理，无组织地自发合作，开展某一项群体行动。一是互联网信息传递的空间泛在性，消除了信息传递的物理局限，极大地拓展了潜在参与者的范围，大大降低了寻找“志同道合”者的成本，分布式协同作业成为可能；二是伴随着移动互联网的普及和大众化，信息传递的即时性越来越强，群体之间的实时沟通和协调更有效率。但是，如果要真正实现去组织化的分布式协同作业，这两点还远远不够，它们仅是技术逻辑层上的基础。既然是生产活动，就须要消耗资源，必然涉及各方的经济利益，激励相容的设计不可或缺。

根据哈维茨（Hurwiez）创立的机制设计理论，在市场经济中每个理性经济人都会有自利的一面，其个人行为会按自利的规则行为行动。如果能有一种制度安排使行为人追求个人利益的行为，正好与企业实现集体价值最大化的目标相吻合，这一制度安排就是激励相容。

企业实践表明，只有贯彻激励相容原则，才能有效地解决个人利益与企业利益之间的矛盾冲突，使个体行为符合企业价值最大化的目标。对于去组织化的分布式协同作业亦是如此。“无利不起早”，经济学的自愿原则以及经济个体对帕累托改进的动机，在加密经济的算法机制中依然成立。

当然，我们也观察到在开源经济和加密经济发展的初始，一些技术极客或创始人的付出并非完全出于经济利益的动机，而是怀着社会理想或者大爱精神，乐于无私合作，无须财务报酬。比如Linus Torvalds崇尚自由软件精神，发起Linux开源软件项目，拒绝商业利益；Nakamoto提出比特币的概念，设计和发布了相应的开源软件，初衷也并非为了追求经济利益。但应注意的是，对于大多数普通人，经济利益仍是最重要的动机。

共识算法使“无组织”群体行动的经济激励显性化

根据经济利益的重要程度，我们可以将“无组织”群体行动分为以下层次：

第一是“非经济利益驱动”的群体行动，如公益活动、慈善活动。这类活动通常不需要参与者付出太多的时间和金钱成本，因此对群体行动的激励相容要求最低。

第二是“潜在经济利益驱动”的群体行动，如社交平台上的主题分享、P2P网络资源分享。这类活动的参与者并没有明确的经济利益诉求，而是持着相互分享、共惠共利的目的，即“我可能在今天从别人的利益出发采取行动，即便此时给我带来一些风险或代价，但我期望的是，别人会记得这一切，并在未来回报我”，经济学将这样非正式的基于社会关系的互惠互利称为“社会资本”。社会资本是一种较为隐蔽的潜在经济利益。

第三是“明确经济利益驱动”的群体行动，也就是本文研究的加密经济的分布式协同作业。应该说，分布式协同作业不一定非得要选择明确经济利益驱动模式，也可以选择潜在经济利益驱动模式，比如Linux等开源项目即是如此。根据常见的自由软件授权方式（General Public License，GPL），如果使用者在免费取得自由软件的源代码后修改了源代码，那么基于公平互惠的原则，他也必须公开其修改的成果，这实质上就是为了保障潜在经济利益的分享与互惠。但需要注意的是，除了Linux等少数开源项目之外，成功的开源项目相当有限，许多开源项目往往是失败的。原因很简单，开源项目缺乏经济逻辑层上的机制设计，如果不给予活动参与者真正的经济激励，仅依靠人们对自由的热爱和追求以及业余的投入，难以激发出成员有价值的付出。比如，洪流网站（Torrent Sites）的点对点文件共享就是个失败的案例。系统旨在让每个下载者在下载的同时也保持着向网络里的其他下载者提供种子（上传已下载的数据）。这是一种基于潜在经济利益——社会资本的设计。而在没有经济激励的情况下，参与者越来越倾向于认为持续上传种子是一件对自己不利的事情，特别是当这一行为会占据电脑里更多的存储空间时，因此失去了继续参与文件共享的兴趣。

在某种意义上，开源项目是一种无明确生产目标、不讲究效率的生产过程，不是一种有效的分布式协同作业。与之

不同，区块链技术通过激励相容的算法规则和相关契约安排，明确了各方的经济利益，充分调动了各方的积极性，使有效的分布式协同作业真正成为可能。

区块链“无组织”群体行动和“拜占庭将军问题”

区块链是一种数字账本，是由一个个区块按时序组成的一串链条。一个区块包含两个部分：区块头(Block Header)和交易信息部分。区块记录的所有交易通过默克尔树(Merkle Tree)组织起来，默克尔树根(Root)的哈希值作为本区块里所有交易的信息被放入区块头。区块头还包含以下字段：前一个区块头的哈希值（或称哈希指针）、本区块的时间戳、高度(从第一个区块开始数，本区块是第几个块)以及其他信息。

在区块链系统开展的分布式协同作业，是众多互不相识的参与者一起对区块的账本信息进行验证、确认和达成共识，形成统一的交易账本。新的区块在经过系统共识验证后被添加到区块链上。由于任何输入端的细微变化都会对哈希函数的输出结果产生较大影响，再加上哈希指针的设计，区块链被认定为是难以篡改的。比如，若有人尝试改写1号区块里的数据，那么存储在2号区块里的1号区块的哈希值将会产生巨大的变化，从而导致2号区块的哈希值随之发生变化，接着又影响存储在3号区块的2号区块的哈希值，以此类推，后续的所有区块数据都会发生变化。所以说，想改动一个区块，必须同时改动该区块后面的所有区块。而对任何一个区块的改动，均须获得共识，这就使得更改一条记录的困难程度按时间的指数倍增加，时间越早的记录越难更改。因此，意图修改一整条已获得系统共识的区块链数据几乎不可能做到。

而攻击主要发生在对新增区块进行验证和共识的过程中，最典型的方式是攻击者从某个区块开始构造一条秘密的区块链，当秘密构造的区块链比当前公开的区块链更长时，将其公开，其他节点将会视其为“正确”的链条，在该链条上继续工作和延长它，使被攻击区块包含的交易被撤销，制造“双花攻击”，从而破坏系统参与者原来达成的共识。

如何在“无组织”的群体中形成共识即是经典的“拜占庭将军问题”：在一个一致意见具有绝对必要性的系统里，如何在缺乏信任机制的情况下，通过一个可信的方法，将一个一致意见同步给所有人？或者说，诚实者如何战胜破坏者，形成一个多数一致的、可信的意见？

经济激励和惩罚为何重要：一个简单的博弈逻辑

传统上，解决“拜占庭将军问题”的算法是BFT（Byzantine Fault Tolerant）算法，其中最著名的是PBFT，该算法是基于消息传递的一致性算法，在弱同步网络下，算法经过三个阶段可以达成一致性。在无法达成一致时，这些阶段会重复进行，直到超时。PBFT算法的优点是收敛速度快、节省资源、具有理论上的安全界（理论上允许不超过1/3的恶意节点存在，即总节点数为3k + 1，其中正常节点超过2k + 1个时，算法可以正常工作）。缺点是随着参与共识节点的增加，通信开销会急剧上升，达成共识的速度则快速下降，难以支撑上万节点规模的分布式系统。尤其是，PBFT假设系统的所有节点是已知的，且节点参与共识首先要获得投票权，因此需要为节点的加入和退出过程设计额外的机制，这不仅增加了协议复杂度和实现难度，还不允许节点自由加入和退出，不符合加密经济的开放性要求。

而区块链技术与BFT算法不同，它通过引入经济激励和惩罚机制，来解决“拜占庭将军问题”。下面通过简单的博弈例子来阐述其中的逻辑。

对于参与共识验证的参与者，存在两种策略是“协作”与“攻击”，选择“协作”即成为诚实者，选择“攻击”即成为攻击者。参与者权衡利弊后选择博弈策略。当参与者发现攻击的收益要高于协作时，参与者选择攻击，否则选择协

表1 无激励和惩罚的纳什均衡解

	协作	攻击
协作	(8, 8)	(-2, 6)
攻击	(6, -2)	(-1, -1)

表2 引入激励后的纳什均衡解

	协作	攻击
协作	(11, 11)	(1, 6)
攻击	(6, 1)	(-1, -1)

表3 引入惩罚后的纳什均衡解

	协作	攻击
协作	(8, 8)	(-2, 3)
攻击	(3, -2)	(-4, -4)

作。假定攻击没有成本，那么，如果大家都是攻击者，相当于“一拍而散”，双方收益均为负；如果有一方攻击，一方协作，则攻击者获利，协作者受损；如果大家都是协作者，则共赢，收益均为正。假定相应的收益矩阵为表1。对其求解，可以得到纳什均衡解：（协作，协作）和（攻击，攻击）。换言之，参与者可能协作，也可能攻击，因此系统存在安全隐患。

若在此基础上引入激励和惩罚，结果则会发生改变。激励机制是，系统给予协作者一个正向的激励，比如在表1，给予协作者三个单位的正向收益，那么表1的收益矩阵变为表2。此时求解得到的纳什均衡为（协作，协作），即参与者的最优策略均是协作，而不是选择攻击，从而消除了系统的攻击行为。惩罚机制是，系统给予供给者一个负向的惩罚，即攻击须付出一定的成本，比如在表1，对攻击者施予三个单位的惩罚，那么表1的收益矩阵变为表3，得到的纳什均衡解为（协作，协作）。可见，同激励机制一样，惩罚机制也消除了系统的攻击行为。毋庸置疑，若同时施加恰当的激励和惩罚机制，系统的安全性更能得到保障。

工作量证明机制（PoW）的激励相容

Nakamoto（中本聪）提出的工作量证明机制（PoW）同时包含了分布式共识的激励和惩罚机制。

（一）激励机制

如前述所言，区块链是一个公共可见的账本，用来记录交易的历史信息。当一笔新的资产交易被创建时，资产转出方须通过签名脚本来证明自己是资产的合法使用者，并且指定输出脚本来限制未来对本交易的使用者（资产收入方）。如果是合法创建并签名的，则该笔交易现在就是有效的，它将被广播到区块链网络并被传送，每一个收到交易的节点将会首先验证该交易，确保只有有效的交易才会在网络中传播，而无效的交易将会在第一个节点处被废弃，直至抵达挖矿节点。

挖矿节点在验证交易后，会将这些交易添加到自己的内存池中，构建新的区块。在PoW机制，矿工们接着通过反复尝试求解一种基于哈希算法的数学难题来竞争获得记账权，具体而言，矿工不断更换区块头的填充随机数并计算这个区块头信息的哈希值，看其是否小于当前目标值。如果小于，则成功“出块”，随后矿工将这个区块发给它的所有相邻节点。这些节点在接收后进行一系列的检查标准去验证区块的正确性。检查的标准包括区块的数据结构和区块包含的交易合法有效；区块头的哈希值小于目标难度（确认包含足够的工作量证明）等。一旦一个节点验证了一个新的区块，它就会将新的区块连接到累积了最大工作量证明的区块链中，矿工挖矿成功。

在上述过程中，矿工获得两方面奖励：一是代币奖励。矿工构建的新区块中的第一笔交易是一笔特殊交易，称为创币交易或者Coinbase交易。矿工挖矿成功后，将获得这笔新创造的加密代币。在比特币网络，每隔10分钟将一个新的区块添加至链上，每添加一个区块可以获得50枚比特币作为奖励（每四年减半）。二是记账决策权与交易手续费。矿工拥有记账决策权，有权决定将哪些交易添加至新

构建的区块，并对收录在区块内的所有交易收取手续费。

（二）惩罚机制

通过惩罚设计，PoW设置了两道门槛：第一道门槛设在矿工竞争记账权的时候，使得矿工不能随便“发言”（新增区块）。一方面，矿工为获得记账权，须不断求解哈希难题，因此付出“不菲”的成本，这一成本是沉没成本，只要矿工想参与“发言”，那么无论他最终能否成功“发言”，他均必须付出这一笔建言成本；另一方面，由于哈希难题的验证要比求解来得简单，对新出区块的验证成本微乎其微，因此只要矿工一错误“发言”（如交易无效、格式不符等），就会很快地被其他节点检测出来废弃掉，他之前付出的建言成本相当于对他的惩罚。

第二道门槛则设在区块被成功添加区块链后的修改，使得矿工不能随意更改区块链。在比特币网络，每2016个区块(大约两周)后，所有客户端把新区块的实际数目与目标数量相比较，并且按照差异的百分比调整目标哈希值，来增加（或减少）产生区块的难度，确保每10分钟 1 块的恒定出块速度。挖矿难度值的提高，增加了攻击的成本。攻击者如果要构造出一条比真实区块链更长的秘密区块链，需要在比特币网络产出6个区块的同时秘密产出7个区块。

截至2018年2月，专业的比特币挖矿机器（以Bitmain生产的AntMiner S9为例）价格为2700美元，这台矿机以2017年2月27日为基准可挖0.0012枚比特币。一台AntMiner S9每天耗电33度，按照居民用电价格计算，大概每天的电费是2.6美元。假定AntMiner S9的折旧年限为3年，可推算每天固定资产折旧为2700/（365*3）=2.5美元，加上耗电费用2.6美元，得到挖出一枚比特币的生产成本为(2.5+2.6)/0.0012=4250美元。那么，无论攻击成功与否，攻击者都要付出4250*7=29750美元，约3万美元的成本，而且这一成本随着挖矿难度的增加不断上升，再加上与诚实者的算力竞争，显然对算力提出了巨大的要求：只有掌握了比特币全网51%算力的攻击者，才可以用这些算力来重新计算已经确认过的区块。

上述两道门槛使得无论是新增还是更改区块，均要付出不菲的成本，尤其是对后者的要求更为苛刻，这就是Nakamoto面对“拜占庭将军问题”的全新思路。从某种意义上来说，PoW机制的“工作量”相当于现代资产交易或拍卖的保证金制度，免除了随意报价，同时还确保了比特币各区块哈希值的唯一性及难以篡改，这正是PoW这一机制设计精巧的地方。

权益证明机制（PoS）的激励相容

（一）自私挖矿攻击和P+Epsilon攻击

PoW机制存在着缺陷。首先，在性能上，PoW的挖矿要耗费算力资源，且随着算力竞争，挖矿难度值不断提高，每10分钟 1 块的恒定出块速度制约交易性能。其次，在安全性上，PoW存在自私挖矿攻击和P+Epsilon攻击的隐患。

自私挖矿攻击由美国康奈尔大学两位学者Ittay Eyal（伊泰·艾尔）和 Emin Gun Sirer（冈塞尔）提出，它是指，自私矿工不公开挖到的块，产生秘密分支，这时候诚实矿工还会基于较短的公开分支挖矿，当自私矿工选择性地公开秘密分支上的区块，将导致诚实矿工抛弃掉较短的公开分支，基于秘密分支计算最新的块，由此就浪费了诚实矿工花费在公开分支的算力，使得自私矿池获得高于全网算力比例的收益。当自私矿工所在的矿池占总网算力的1/3时，其获得的收益会大于相对算力，于是理性矿工会源源不断地加入自私矿池，最终导致矿池算力超过总网络的50%。

P+Epsilon攻击由以太坊创始人Vitalik Buterin提出，它是一种贿赂攻击者模型，即攻击者进入系统，以可信的预算贿赂其他矿工们参与攻击，但事后却无须付出任何成本。以表1为例，假定攻击者给予矿工们一个可信的贿赂预期：当其他人选择“协作”时，如果你选择“攻击”，我将给予你比选择“协作”更高的收益，

表4 P+Epsilon攻击

	协作	攻击
协作	（8，8）	（-2，8+Epsilon）
攻击	（8+Epsilon，-2）	（-1，-1）

通过这样的贿赂，其他人选择“协作”，我选择“攻击”的收益由6变为8+ε，见表4。此时的纳什均衡解就变为（攻击，攻击），即所有矿工均选择“攻击”，各自的收益均为-1，事后，攻击者没有付出成本，因为所有人都选择了攻击，攻击者不用兑现贿赂承诺。也就是说，攻击者只要以可信的预算和承诺（例如将资金锁定在智能合约），就可零成本地实现对系统的攻击。

（二）PoS机制降低了攻击的经济激励

PoS机制在理论上可以克服上述缺陷。其一，在PoS机制，节点获得区块创建权的概率取决于该节点在系统中所占有的权益比例的大小，不耗费算力，因此就不存在自私挖矿攻击的问题；其二，针对P+Epsilon攻击，Vitalik Buterin提出可以在PoS机制中引入严厉的惩罚予以预防，即要求矿工们提取一定比例的私人财富（或称权益）作为抵押物，投注于未来的区块中，随后根据投注的情况进行处罚，比如，如果事后可以明确地证明一个特定的区块是有问题的，那么就对这个区块的投注者进行最大限度的惩罚，这就改变了P+Epsilon攻击时矿工的预期收益，使他们一旦在真实链上投注了自己的权益，就会有更大的动力继续在真实链上工作，而不是参与作恶，同时又大大增加了攻击者的贿赂预算。理论上，攻击者需要拥有51%的权益（数字代币），才有可能发起成功的攻击，无疑，倘若攻击成功，攻击者自身的权益也会受到很大的损失，因此就降低了发起P+Epsilon攻击的激励。

（三）PoS机制可引入经济惩罚来解决“利益无关者”攻击

“利益无关者攻击”（Nothing at the Stake Attack）是指，由于PoS机制不需要耗费算力，因此矿工在伪造链上挖矿无须成本，但一旦伪造链被确认时，则会获得收益，这就会激励即便是诚实的矿工也可能会在伪造链上挖矿，尤其是权益越少的矿工，这种“投机”的心态越强，因为虽然他们知道这种攻击行为会造成整个系统的价值降低，但他们的权益很少，他们并不在乎，而当少数人“积少成多”时，就会对系统的整体安全性带来隐患。“利益无关者攻击”也称为平凡人悲剧（Tragedy of the Commons）。

对此，以太坊的Casper权益证明机制引入了名为Slasher的惩罚机制，即如果有人尝试了攻击，其他人发现了可以公布证据，系统将对这个人进行惩罚。由此，就抑制了矿工在伪造链上挖矿的经济激励，有效预防“利益无关者攻击”。

（四）PoS机制的实践

PeerCoin最早提出并实现了PoS共识协议。PeerCoin没有完全抛弃PoW，每个节点有自己的PoW难度值，币龄（Coinage）是该难度值的计算参数。币龄越高，则PoW难度越低，越容易计算出满足难度的哈希值。PoS一般需要用户时刻在线，对应用带来了很大的挑战。为了解决这个问题，衍生出了DPoS（Delegated Proof of Stake）共识，其核心思想是从先从全网节点中选出部分节点，保证这些节点的有效性，然后在该子节点集合内进行PoS共识。BitShares是第一个采用DPoS的区块链。在BitShares中，全网节点投票选出101名代表，来负责区块的生成。

总结与展望

基于不同的经济激励和惩罚设计，PoW机制和PoS机制构建了一个激励相容的开放式环境，让众多互不相识的参与者自愿参与，一起对区块的账本信息进行验证、确认和达成共识，形成统一的交易账本，从而可以在无需第三方机构的情况下实现资产的确权、交易和转移。目前，PoW机制与PoS机制孰优孰劣，尚未有定论。比如，有人就认为，PoW机制的资源耗费不是无意义的，恰恰是“真金白银”的投入，才凸显区块链系统的价值与可信。

在理论上，共识机制的研究成果正不断丰富，比如在PoW机制方面，康奈尔大学Rafael Pass提出Thunderella算法，使状态机与同步协议无异，不仅可以实现快速的异步处理，还可以在异常时启动回滚机制，同时实现了拜占庭容错和交易瞬间响应；在PoS机制方面，研究者们提出了各种算法，如康奈尔大学的Elaine Shi等提出的基于Sleepy Model的PoS共识，Silvio Micali等提出的Algorand协议，爱丁堡大学Aggelos Kiayias等提出的Ouroboros算法等；Krzysztof Pietrzak和Bram Cohen则提出了一种新的取代PoW的共识机制，他们称之为空间证明机制（Proof of Space）。这些共识机制和安全模型仍需在实践上进一步检验。

［姚前为中国人民银行数字货币研究所所长，本文得到国家重点研发计划（批准号：2016YFB0800600）和SFI（上海新金融研究院）资助，本文仅代表个人学术观点，与任何机构无关。本文编辑/王蕾］