中国科学: 信息科学 2022 年 第 52 卷 第 6 期: 992–1001
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
理论计算机科学专题 · 论文

创作者经济中的去中心化审查机制设计

陈宏鉴, 邓小铁*, 孔雨晴, 陆宇暄

北京大学前沿计算研究中心, 北京 100871

* 通信作者. E-mail: xiaotie@pku.edu.cn

收稿日期: 2022–03–28; 修回日期: 2022–04–13; 接受日期: 2022–05–16; 网络出版日期: 2022–06–06

国家自然科学基金 (批准号: 62002001, 62172012) 资助项目

摘要 创作者经济的兴起以及相应去中心化平台的搭建加速了权力向个体创作者的转移, 然而由于去中心化平台缺乏中心内容审查和监管, 创作者可能滥用权力, 传递可能造成社会损害的信息. 已有的去中心化审查机制缺乏对创作内容社会损害的严谨刻画, 同时对于该问题的解决方案也缺乏相应的理论支撑. 我们提出了一套去中心化平台的审查机制: 创作者上传的作品将首先由去中心化的审查者进行损害评估, 根据评估结果, 创作者在上传作品的同时需要支付一定押金, 为了防止在随机环境下较坏情况发生, 我们将引入去中心化的担保者进行担保, 在此基础上通过中心化的资金池解决极端情形下的社会损害赔付. 我们证明: (1) 在损害评估阶段, 审查者们诚实地给出自己对作品社会损害的评估是一个支配性策略; (2) 审查机制以极高的概率保证作品的社会损害会低于担保者的赔偿上限和创作者的押金之和.

关键词 创作者经济, 互联网 3.0, 去中心化审查, 区块链, 社交媒体

1 引言

随着各种直播平台和社交媒体等创作平台的诞生与发展, 越来越多的个体创作者开始在各种平台上发布自己的创作. 个体创作者们通过创作获得粉丝和影响力并以此盈利的经济行为被称为创作者经济 [1]. 创作者经济的快速发展使得传统媒体对信息渠道的垄断被逐渐打破, 创作内容呈现的权力也逐渐从传统媒体转移到了个体创作者和创作平台上. 此外, 随着区块链技术 [2] 和 Web3.0 [3] 的发展, 诸如 Mastodon [4] 和 Steemit [5,6] 等去中心化社交媒体也逐渐兴起. 与传统的创作平台不同, 去中心化创作平台并不被某一公司拥有, 也不存在中心化的管理团队. 在去中心化创作平台上, 个体创作者可以更加自由地发布创作内容, 创作内容的发布、管理和收益均通过代码协议完成.

然而, 由于去中心化创作平台缺乏中心化管理团队的内容审查和监管, 个体创作者可能会滥用其创作权力, 传递造成社会损害的信息. 较为严重的情形包括但不限于传播违法信息、侮辱诽谤和教唆

引用格式: 陈宏鉴, 邓小铁, 孔雨晴, 等. 创作者经济中的去中心化审查机制设计. 中国科学: 信息科学, 2022, 52: 992–1001, doi: 10.1360/SSI-2022-0126
Chen H Y, Deng X T, Kong Y Q, et al. Decentralized content censorship for creator economy (in Chinese). Sci Sin Inform, 2022, 52: 992–1001, doi: 10.1360/SSI-2022-0126

犯罪, 涉嫌这些方面的作品相对较少尚可管控. 而实践中更多的是那些具有相对轻微的社会损害的信息, 例如插入广告、捏造或传播谣言、令人不适或涉嫌歧视的信息, 这些造成社会损害的信息相当常见. 此外, 由于个体创作者可能难以赔付高额的社会损害并且可以选择离开平台来规避赔偿, 去中心化的创作平台也很难彻底追究个体创作者传递有害信息的责任. 已有的去中心化创作平台虽然通过设置声誉机制 [7~9] 等方式激励创作者不做出损害社会和创造平台的行为, 但是对于创作内容的审核、监管和追责的问题依旧缺乏严谨、完整的刻画, 对于该问题的解决方案也缺乏相应的理论支撑. 因此, 如何在去中心化的创作平台上设计机制防止创作者的权力滥用, 是创作者经济急需解决的一个关键问题.

本文研究如何设计创作者经济中的去中心化审查机制. 我们的研究思路是让创作者在上传其作品的同时雇佣审查者和担保者, 通过审查者、担保者和创作者同时对作品负责的方案, 做到对社会损害的事后赔付. 首先, 我们对去中心化创作平台上作品的社会损害进行了严谨的建模. 基于这一社会损害模型, 我们提出了两阶段审查机制, 使得去中心化创作平台在保证个体创作者上传作品的自由度的同时, 实现对作品的高效审查以及对其造成社会损害的追偿. 在两阶段审查机制中, 一个作品需要通过两个阶段的审查过程才会被公开展示. 第一个阶段称为创作评估阶段, 在这一阶段平台会随机选取一组用户审查作品, 并计算出该作品社会伤害的安全阈值. 第二个阶段为担保拍卖阶段, 在这一阶段创作者将会质押一定的押金, 同时一组担保者会竞拍该作品的担保权. 竞拍成功的一组担保者会收到创作者的担保费用, 但是如果该作品造成的社会损失高于安全阈值, 这组担保者将会承担其社会损失高于安全阈值部分的赔偿.

理论分析的结果表明我们提出的两阶段审查机制可以有效地防止创作者的权力滥用. 我们证明了如实汇报作品的社会损害程度是审查者最大化审查收益的支配性策略. 同时, 我们还证明了两阶段审查机制以极高的概率保证作品的社会损害会低于担保者的赔偿上限和创作者的押金之和. 最终, 我们对两阶段审查机制在实际运行中的参数选择进行了详细的探讨.

2 问题模型

2.1 社会损害模型

本文对作品的社会损害建模如下: 对每一个作品 i, 其造成的社会损害为 $l_i \geqslant 0$, l_i 是一个随机变量, 其随机性来源于环境.

存在一个社会损害的事后揭示器, 在内容上传后, 揭示器有可能揭示该作品的社会损害的真实值, 也有可能不揭示其社会损害的真实值. 设 r_i 为一个 0-1 随机变量, $r_i = 1$ 表示作品的社会损害被揭示, $r_i = 0$ 表示作品的社会损害不被揭示. 事实上只有一小部分作品的社会损害最终会被揭示, 当一个作品的影响力较低或者社会损害较小时, 其社会损害很少会被揭示出来. 当一个作品的社会损害被揭示时, 会造成等同于其值的事后损失, 反之这个作品不会造成事后损失.

在平台中包含两类用户: 普通用户和专业用户. 对于普通用户 j 对作品 i 会有一个社会损害的认识 $\tilde{l_{ij}}$, $\tilde{l_{ij}}$ 是一个随机变量, 其准确程度取决于该用户对作品领域的熟悉程度. 专业用户 j 认为作品 i 的损害程度 l_i 和是否被揭示 r_i 服从联合分布 $\tilde{D_{ij}}$. 值得注意的是, $\tilde{l_{ij}}$ 和 $\tilde{D_{ij}}$ 都是用户对于作品社会损害的主观判断, 应与作品实际的社会损害 l_i 是否被揭示 r_i 区分开来.

2.2 假设

在该问题模型下, 我们将进行一些适当的假设.

(1) 逐利性. 所有用户都想最大化自己的期望收益.

(2) 重大误判限制. 用户对于作品的社会损害认识大概率不会发生巨大的误判, 即用户严重低估作品社会损害的概率很小. 在实践中, 我们假设对于审查者 j 和作品 i, 有 $\Pr[\tilde{l_{ij}} < \frac{1}{k} l_i] \leqslant q$. 关于常数的选取, 一般可以通过固定常数 k 并通过实验估计常数 q 来解决. 在固定常数 k 后, 我们定义审查严重偏差为审查者估计作品的社会损害不足其真实值的 $\frac{1}{k}$, 然后通过实验估计审查严重偏差的发生概率 q.

3 本文贡献

本文首次将创作者经济中的社会损害建模. 在该建模的前提下提出直接的解决方案, 即基线机制, 以及精心设计的机制 —— 两阶段审查机制, 并说明了两阶段审查机制相比基线机制具有良好的性质并且在实践中可行.

3.1 基线机制

在理想情况下, 基线机制要求作品 i 的创作者提供足够的押金 d_i 以应对可能造成的事后社会损害. 这笔押金会在造成事后损失后被扣除 l_i, 剩余的押金会被退还给创作者. 这要求 d_i 要超过可能的社会损失最大值, 这导致要求的押金数额是不可接受的.

3.2 两阶段审查机制

本文提出了两阶段审查机制, 是一种使得在创作者可以自由上传作品的前提下提供社会损害赔偿的去中心化审查机制. 两阶段担保机制要求在作者准备上传作品时通过两个阶段作品才会被平台展示出来. 第一个阶段称为创作评估阶段, 在这一阶段平台会随机选取一组用户审查作品. 第二个阶段为担保拍卖阶段, 在这一阶段一组担保者会来竞拍该作品的担保权.

两阶段担保机制在适当的假设下满足的两方面良好性质及证明见第 5 节.

4 机制设计

本节将讨论两阶段担保机制的设计和细节. 具体来说, 在第一阶段, 创作者上传的作品将首先由去中心化的审查者进行损害评估, 根据评估结果, 创作者在上传作品的同时需要支付一定押金, 为了防止在随机环境下较坏情况发生, 在第二阶段, 我们将引入去中心化的担保者进行担保, 在此基础上通过中心化的资金池解决极端情形下的社会损害赔付.

4.1 参与者

在两阶段审查机制中, 存在 4 种不同的参与者, 分别被称为创作者、审查者、担保者和极端损失担保池.

创作者. 创作者是作品的制造者, 出于兴趣或利益希望将作品上传至平台以获取播放量. 为了让其作品能够上传至平台, 需要经过雇佣审查阶段和担保拍卖阶段. 在雇佣审查阶段会产生审查费而在担保拍卖阶段会产生担保费. 需要创作者支付审查费、担保费以及一笔押金后, 作品才可以在平台上公开.

审查者. 审查者是在雇佣审查阶段被随机选取的普通用户. 这些用户愿意通过出卖劳动力审查作品以获得审查费. 每个审查者有自己熟悉的领域并且系统只会分配符合其熟悉领域的作品进行审查.

审查者可以在系统分配审查任务之后拒绝该任务, 如果审查者接受一个审查任务, 就必须提交一个对该作品社会损害的估计.

担保者. 担保者是在担保拍卖阶段对作品的担保权进行竞价的专业用户, 要成为担保者需要事先缴纳一定的押金. 担保者可以为任何等待担保的作品提交报价, 报价为该担保者为了承担该作品产生高于一个特定阈值的社会损失需要收取的担保费用.

极端损失担保池. 极端损失担保池是一个中心资金池, 用于兜底赔偿作品造成的远超出审查者预期的社会损害. 极端损失担保池为平台所有的作品提供担保并对每个作品收取固定的足以忽略不计的担保费. 所有用户都可以向极端损失担保池注资, 担保池产生的收入 (来源于收取的担保费) 和损失 (来源于极端社会损失赔偿) 根据用户的出资分摊.

4.2 机制描述

本小节将详细描述两阶段担保机制的实施过程并介绍机制中需要设定的参数, 对于参数的设计将在第 5 节讨论. 以下是两阶段担保机制的实施过程, 机制中使用的参数和符号描述参见表 1. 在两阶段担保机制下, 作者上传作品需要支付审查费, 并且选择质押押金为作品担保或额外支付担保费.

1. 创作者创作作品 i, 将其上传至平台并预付足够的审查费用.
2. 雇佣审查阶段:
   (a) 创作者尽可能准确地描述作品所属的领域分类.
   (b) 令社会损害保守估计 $l_{\text{safe}} = l_{\text{base}}$.
   (c) 重复以下审查轮直至跳出:
      i. N 名熟悉该领域的审查者审查该作品, 审查者需要回答 "作品 i 的社会损害低于或高于 l_{safe}" 这一问题. 为作品分配审查者以及为审查者付费的机制将在 4.2.1 和 4.2.2 小节讨论.
      ii. 待审查者结束审查作品后, 平台获得 N 个答案, 令 p_{safe} 为其中回答 "低于" 的答案比例.
      iii. 若 p_{safe} 不低于安全阈值 p_{thres}, 通过此轮审查, 跳出循环; 否则不通过此轮审查, 将 l_{safe} 赋值为 $2 \times l_{\text{safe}}$, 继续下一轮审查.
3. 担保拍卖阶段: 创作者在 (a) 或 (a′) 中选择一项, 然后执行 (b).
   (a) 创作者质押 $k \times l_{\text{safe}}$ 的押金, 或者
   (a′) i. 创作者质押 l_{safe} 的押金.
      ii. 平台拍卖一个担保, 担保的内容为承担作品 i 造成的在区间 $[l_{\text{safe}}, k \times l_{\text{safe}}]$ 的社会损害的 $\frac{1}{s}$ 比例. 拍卖的担保数量为 s 份, 每个担保者最多承担一份担保. 对担保进行拍卖的机制将在 4.2.3 小节讨论.
      iii. 根据拍卖的结果, 创作者向 s 位担保者支付担保费.
   (b) 创作者支付 v 以购买极端损失担保池提供的担保, 担保的内容为承担作品 i 造成的在区间 $[k \times l_{\text{safe}}, +\infty)$ 的社会损害.

表 1　两阶段担保机制所使用的符号一览

Table 1　List of symbols used in the two-stage insurance mechanism

Symbol	Description
N	The number of reviewrs of a work.
l_{base}	A mechanism parameter which denotes the initial value of conservative estimation of social damage.
l_{safe}	The platform's conservative estimation of the social damage.
p	The frequency at which reviewers believe that the social damage of works exceeds l_{safe}.
p_{thres}	A mechanism parameter which denotes the specified safety threshold.
k	A mechanism parameter where the social damage caused by the work can hardly exceed $k \times l_{\text{safe}}$.
v	A mechanism parameter which denotes the price guaranteed by the extreme-loss pool for any work.

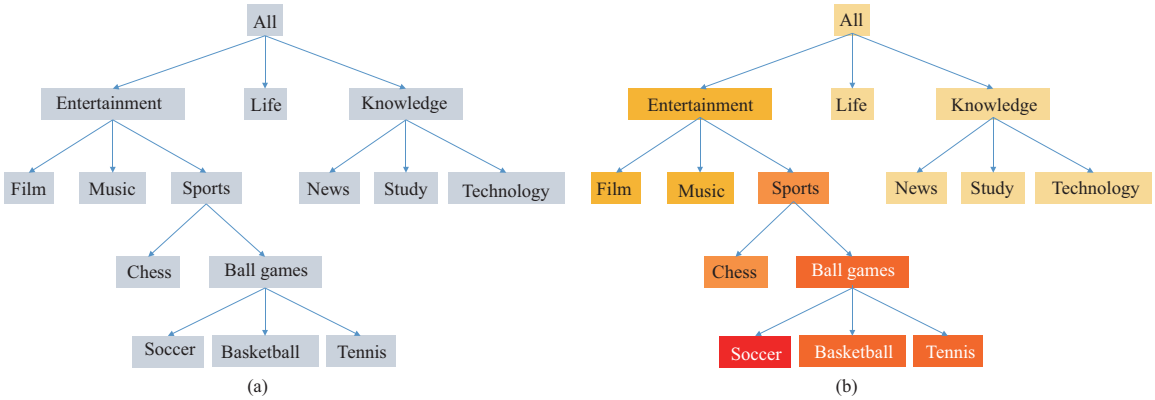

图 1　(网络版彩图) 一个领域划分的例子. 领域对应的颜色越深, 表明审查者对该领域越熟悉. 同样的颜色代表相同熟悉程度

Figure 1　(Color online) An example of domain division. (a) Domain tree; (b) a reviewer's familiarity of domains other than the most familiar one. The darker the domain, the more familiar the reviewer is with that domain. The same color represents the same familiarity

4.2.1　基于领域的审查分配

在为作品审查时, 需要尽可能将作品分配给熟悉作品领域的审查者以获得更准确的审查结果. 作品的领域划分可以被看作一个树形结构, 一个例子如图 1(a) 所示. 对于审查者而言, 他有一个 (或数个) 精通的领域 A, 在图 1(b) 中用大红色表示. 一个适当的假设是其对于另一个领域 B 的熟悉程度取决于领域 A 和 B 的最近公共祖先和领域 A 的距离, 距离越小熟悉程度越高. 由于我们需要在可用的审查者人数和审查者对作品的熟悉程度之间寻求一个平衡, 对于每一个作品我们对其领域向上找到其最近的一个祖先领域, 使得总共有至少具有常数比例的审查者的精通领域在该祖先领域的子树中. 对于该作品就应该在这些审查者中随机抽取进行审查, 以避免抽取审查者的范围过小导致审查结果受到潜在的攻击者的影响.

4.2.2　激励诚实汇报的审查费计算

在现实中, 只有一小部分作品的社会损害会在公开后被揭示出来, 而更多作品的社会损害确切值是未知的, 这意味着仅根据作品的社会损害值来设计审查费计算是困难的. 因此, 我们需要根据审查者对作品的审查本身设计激励诚实汇报的审查费计算方案. 文献 [10] 提出了一种分数计算方案, 名为

行列式互信息 (determinant based mutual information, DMI) 分数, 其在宽松的限制下能够激励回答者对于主观问题进行真实汇报. 下面我们来介绍一下行列式互信息分数的计算方法.

现有 m 位回答者, n 个先验相似的选择题, 每个选择题有 C 个选项. 每位回答者将被随机分配一些题目并保证题目数量大于等于 $2C$, 且对任意回答者 i, 存在回答者 $j \neq i$ 保证两人共同回答的题目数量大于等于 $2C$. 将每个回答者 i 的答案记作一个 $n \times C$ 的 0-1 回答矩阵 $\boldsymbol{R}_i$. 对于任意 t 和 c, 回答矩阵的第 (t, c) 项为 1 当且仅当回答者在问题 t 中回答选项 c, 否则为 0. 对于回答者 i, 任取回答者 $j \neq i$ 保证两人共同回答的题目数量大于等于 $2C$, 并同时将问题分成两组且保证在每一组里两人共同回答的数量大于等于 C, 并相应地得到回答者 i 两个回答矩阵 $\boldsymbol{R}_i^1$ 和 $\boldsymbol{R}_i^2$ 以及回答者 j 的两个回答矩阵 $\boldsymbol{R}_j^1$ 和 $\boldsymbol{R}_j^2$. 回答者 i 的 DMI 分数为 $\det[\boldsymbol{R}_i^{1\mathrm{T}}\ \boldsymbol{R}_j^1] \times \det[\boldsymbol{R}_i^{2\mathrm{T}}\ \boldsymbol{R}_j^2]$.

根据行列式互信息分数的性质和计算, 平台每次只需将不少于 4 个同一领域的审查问题打包成一个审查任务并随机抽取审查者 (因为询问的问题只有两个选项), 在此基础上保证每个审查者最多审查同一作品一次. 根据审查结果计算行列式互信息分数并正比支付审查费.

4.2.3 确定性激励兼容的担保拍卖

如果作者选择只质押 l_{safe} 的押金而不是 $k \times l_{\mathrm{safe}}$ 的押金, 那么对作品 i 造成的在区间 $[l_{\mathrm{safe}}, k \times l_{\mathrm{safe}}]$ 的社会损害的担保将会被拍卖. 在担保拍卖中, 担保者需要为 s 份对作品的担保进行竞价. 对于担保者 j 而言, 根据模型他有对于作品社会损害揭示的分布估计 F_j, 通过该分布担保者就能计算出其对于担保权的价值估计.

为了方便, 我们令报价 b_j 为 j 声称其愿意承担一份担保责任以获得 b_j 的担保费. 自然地, 我们设定担保的分配规则为报价最低的 s 位担保者会获得担保权. 不难验证该分配规则是单调的, 即在其余担保者出价不变的情况下降低报价能够使得该担保者获得担保权的概率不降. 由于这是一个单参数拍卖并且分配规则单调, 我们可以应用迈尔森引理 [11] 获得唯一的确定性激励兼容的担保拍卖机制, 即

(1) 设各担保者对作品的报价从低到高排序为 $b_1, b_2, \ldots, b_m$;

(2) 分配规则: 给出报价 $b_1, b_2, \ldots, b_s$ 的担保者获得对该作品的担保权;

(3) 支付规则: 向给出报价 $b_1, b_2, \ldots, b_s$ 的担保者支付 b_{s+1} 的担保费.

因此, 创作者需要在拍卖后支付 $s \times b_{s+1}$ 的担保费给担保者.

5 机制分析

5.1 审查者诚实汇报激励

定理1 在假定审查者对于同一类型的作品审查策略相同的情况下, 两阶段担保机制保证如实汇报作品的社会损害程度是审查者最大化审查收益的占优策略. 一个策略是占优策略指在博弈中无论对手采取任何策略, 该策略都是玩家的最优回应策略.

证明 回忆我们使用正比于 DMI 分数的审查费用设计. DMI 分数要求一组回答者一次回答至少 $2C$ 个问题, 每个问题都具有 C 个选项. 在以下条件下根据 DMI 分数正比支付审查费的机制保证了真实回答每个问题是逐利用户的一个支配性策略.

- 回答者独立作答.
- 问题独立同分布.
- 对任意回答者 j, 其对所有问题的策略相同, 策略指一个 $C \times C$ 的转移矩阵 S_j, $S_j[x, y]$ 为回答者真实答案为 x 但是汇报 y 的概率.

由于审查问题的选项仅有 2 个 (高于或低于), 因此使用 DMI 分数只需打包不低于 4 个问题成为问卷. 由于问题的领域相同并且形式相同, 我们可以相信并验证审查者对所有作品的审查策略是相同的. 因此在审查环节使用 DMI 分数满足条件, 真实回答每个问题是一个支配性策略.

5.2 极端社会损害概率估计

定理2 假定在每轮审查时, l_{safe} 与剩余作品的分布 D 满足 $\Pr_{i\sim D}[l_i > k \times l_{\text{safe}}] \leqslant \Pr_{i\sim D}[$作品 i 通过该轮审查] 时, 两阶段担保机制保证需要中心化资金池偿付的极端社会损害发生的概率不超过 $\exp\{-N \times D_{\text{KL}}(p_{\text{thres}}||q)\}$ ($D_{\text{KL}}(x||y)$ 指 x 到 y 的相对熵[1]). 即在中心化资金池的偿付社会损害上限为 M 的情况下, 作品需向中心化资金池上交不超过 $M \times \exp\{-N \times D_{\text{KL}}(p_{\text{thres}}||q)\}$ 的极端损害担保费.

证明 回顾我们假定审查者群体对于作品的社会损害认识大概率不会发生巨大的误判, 有 $\Pr[\tilde{l_{ij}} < \frac{1}{k}l_i] \leqslant q$. 而在作品的最后一轮审查中在 N 位审查者中有不超过 $p \times N$ 位认为作品的社会损害超过 l_{safe}. 现在我们将计算真实社会损害 l 不低于 $k \times l_{\text{safe}}$ 的概率.

作品经过若干轮不通过的审查直到进行 l_{safe} 这一轮审查时的社会损害分布为 D, 设事件 A 为作品的真实社会损害 l 不低于 $k \times l_{\text{safe}}$, 事件 B_i 为第 i 个审查者认为该作品的社会损害小于 l_{safe}, 事件 C 为超过 $p \times N$ 个审查者认为该作品的社会损害小于 l_{safe}. 所有事件 B_i 在事件 A 发生的条件下独立. 根据假设, 有 $\forall i \in [N], \Pr[B_i|A] \leqslant q$. 根据切诺夫界 (Chernoff bound), 可以得到事件 C 在事件 A 的条件下的概率上界为

$$\begin{aligned}
&\Pr[C|A] \\
&\quad = \Pr\left[\sum_{i=1}^{N} B_i \geqslant p \times N \,\middle|\, A\right] \\
&\quad \leqslant \exp\{-N \times D_{\text{KL}}(p_{\text{thres}}||q)\}.
\end{aligned}$$

因此, 我们有

$$\Pr[A|C] = \exp\{-N \times D_{\text{KL}}(p_{\text{thres}}||q)\} \frac{\Pr[A]}{\Pr[C]}.$$

在分布 D 下, 事件 A 的概率为作品出现极端社会损害的概率, 而事件 C 的概率为在这一轮之中能够通过审查的概率. 这两个概率都能通过事前调研统计出来, 不过由于事件 A 往往是一个小概率事件, 而事件 C 有常数概率发生, 我们可以假定并在实践中测试一个非常宽松的条件 $\frac{\Pr[A]}{\Pr[C]} \leqslant 1$. 在这个假设下, 我们可以得到 $\Pr[A|C]$ 的一个上界, 即需要计算损失担保池赔偿概率的上界, 为 $\exp\{-N \times D_{\text{KL}}(p_{\text{thres}}||q)\}$.

5.3 参数设计

因此, 实际上存在 3 个需要设计的机制参数 N, p_{thres} 和 q. 表 2 给出了参数在不同取值时极端损失担保池应当收取的担保费率 ρ. 根据平台的社会损害赔偿上限 M, 可计算得到极端损失担保池应当收取的担保费 $v = \rho M$, 我们可以选取适当的参数使得 v 的大小是可以忽略的, 比如人民币 1 元.

1) $D_{\text{KL}}(x||y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$.

表 2 机制参数和担保费率对照表

Table 2 Comparison of mechanism parameters and guarantee rates

			$N = 6$			
	$q = 0.001$	$q = 0.003$	$q = 0.01$	$q = 0.03$	$q = 0.1$	$q = 0.3$
$p_{\text{thres}} = 3/6$	6.381E−08	1.712E−06	6.210E−05	1.577E−03	4.666E−02	5.927E−01
$p_{\text{thres}} = 4/6$	4.547E−11	3.668E−09	4.466E−07	3.472E−05	3.691E−03	1.808E−01
$p_{\text{thres}} = 5/6$	1.491E−14	3.617E−12	1.478E−09	3.519E−07	1.344E−04	2.540E−02
			$N = 10$			
	$q = 0.001$	$q = 0.003$	$q = 0.01$	$q = 0.03$	$q = 0.1$	$q = 0.3$
$p_{\text{thres}} = 5/10$	1.019E−12	2.451E−10	9.738E−08	2.137E−05	6.047E−03	4.182E−01
$p_{\text{thres}} = 6/10$	8.339E−16	6.031E−13	8.043E−10	5.403E−07	5.493E−04	1.465E−01
$p_{\text{thres}} = 7/10$	4.484E−19	9.747E−16	4.364E−12	8.977E−09	3.279E−05	3.374E−02
$p_{\text{thres}} = 8/10$	1.487E−22	9.718E−19	1.460E−14	9.199E−11	1.207E−06	4.791E−03
$p_{\text{thres}} = 9/10$	2.579E−26	5.065E−22	2.555E−17	4.928E−13	2.323E−08	3.556E−04
			$N = 14$			
	$q = 0.001$	$q = 0.003$	$q = 0.01$	$q = 0.03$	$q = 0.1$	$q = 0.3$
$p_{\text{thres}} = 7/14$	1.627E−17	3.509E−14	1.527E−10	2.895E−07	7.836E−04	2.951E−01
$p_{\text{thres}} = 9/14$	9.132E−24	1.780E−19	8.728E−15	1.551E−10	5.420E−06	3.036E−02
$p_{\text{thres}} = 11/14$	1.438E−30	2.532E−25	1.400E−19	2.332E−14	1.052E−08	8.765E−04
$p_{\text{thres}} = 13/14$	3.665E−38	5.832E−32	3.632E−25	5.674E−19	3.302E−12	4.095E−06
			$N = 18$			
	$q = 0.001$	$q = 0.003$	$q = 0.01$	$q = 0.03$	$q = 0.1$	$q = 0.3$
$p_{\text{thres}} = 9/18$	2.598E−22	5.022E−18	2.395E−13	3.923E−09	1.016E−04	2.082E−01
$p_{\text{thres}} = 11/18$	1.663E−28	2.905E−23	1.561E−17	2.397E−12	8.009E−07	2.443E−02
$p_{\text{thres}} = 13/18$	4.136E−35	6.529E−29	3.953E−22	5.691E−16	2.455E−09	1.114E−03
$p_{\text{thres}} = 15/18$	3.318E−42	4.732E−35	3.229E−27	4.358E−20	2.426E−12	1.638E−05
$p_{\text{thres}} = 17/18$	4.752E−50	6.124E−42	4.709E−33	5.958E−25	4.281E−16	4.300E−08

6 总结与讨论

我们考虑在创作者经济中如何保护社会公共利益的方案设计问题, 特别是在去中心化区块链经济环境下, 我们设计了一套适用于去中心化平台的审查机制: 对创作者的产品进行去中心化损害评估; 根据评估结果, 收取创作者押金; 我们同时引入去中心化的担保系统, 建立资金池以解决极端情形下的社会损害的补偿问题.

在权力向个体创造者转移的同时, 我们的机制设计也要求他们承担适当的责任, 将技术风险本地化, 从而最大限度地减轻了社会负担.

本文也同时留下了大量的开放性问题. 第一, 在损害评估机制中, 虽然审查者们诚实地汇报自己的评估是一个支配性策略, 但是如果若干个审查者进行合谋, 是否可以大大地提高他们的收益? 第二, 我们是否可以通过设立声誉机制来衡量审查者之间的水平的差别 [7~9], 从而让损害评估机制可以得到更加精确的结果? 第三, 是否可以建立模型分析担保者的收益率? 第四, 是否可以用其他去中心化

保险 [12] 的方式来完成担保的过程? 第五, 是否存在其他的思路完成创作者经济中的去中心化审查?

参考文献

1 Radionova I, Trots I. “Creator economy”: theory and its use. Econ Financ Manage Rev, 2021, 3: 48–58

2 Zheng Z, Xie S, Dai H, et al. An overview of blockchain technology: architecture, consensus, and future trends. In: Proceedings of IEEE International Congress on Big Data (BigData Congress). Honolulu: IEEE, 2017. 557–564

3 Almeida F, Santos J D, Monteiro J A. E-commerce business models in the context of Web3.0 paradigm. Int J Adv Inform Tech, 2013, 3: 1–12

4 Zignani M, Gaito S, Rossi G P. Follow the “Mastodon”: structure and evolution of a decentralized online social network. In: Proceedings of the International AAAI Conference on Web and Social Media. Stanford: AAAI Press, 2018. 541–550

5 Guidi B, Michienzi A, Ricci L. Steem blockchain: mining the inner structure of the graph. IEEE Access, 2020, 8: 210251

6 Li C, Palanisamy B. Incentivized blockchain-based social media platforms: a case study of Steemit. In: Proceedings of the 10th ACM Conference on Web Science. Boston: ACM, 2019. 145–154

7 Chen H, Chen Z, Cheng Y, et al. Poster: an efficient permissioned blockchain with provable reputation mechanism. In: Proceedings of IEEE 41st International Conference on Distributed Computing Systems (ICDCS). Washington DC: IEEE, 2021. 1134–1135

8 Proserpio D, Zervas G. Online reputation management: estimating the impact of management responses on consumer reviews. Marketing Sci, 2017, 36: 645–665

9 Moreno A, Terwiesch C. Doing business with strangers: reputation in online service marketplaces. Inf Syst Res, 2014, 25: 865–886

10 Kong Y. Dominantly truthful multi-task peer prediction with a constant number of tasks. In: Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms. Salt Lake City: SIAM, 2020. 2398–2411

11 Myerson R B. Optimal auction design. Math Oper Res, 1981, 6: 58–73

12 Chen Z, Yang G. Decentralized asset custody scheme with security against rational adversary. In: Web and Internet Economics. Potsdam: Springer, 2021. 449–466

Decentralized content censorship for creator economy

Hongyin CHEN, Xiaotie DENG*, Yuqing KONG & Yuxuan LU

Center on Frontiers of Computing Studies, Peking University, Beijing 100871*, China*
* Corresponding author. E-mail: xiaotie@pku.edu.cn

Abstract With the growth of the creator economy and decentralized platforms, the power has started to be transferred to individual creators. However, due to the lack of censorship in decentralized platforms, the individual creators can upload contents that may cause huge damage to society potentially without any responsibility. Previous work on decentralized censorship does not provide a formal model and theoretic guarantee for this setting. We propose a decentralized censorship mechanism in this setting: after the creator uploads the content, a decentralized censor mechanism will evaluate the damage of the content and the creator will be requested to provide a deposit according to the result. To avoid the worst case when the deposit cannot cover the damage, we also have a decentralized insurance system. We show that (1) in the damage evaluation phase, the censors will be incentivized to provide honest evaluation; (2) with high probability, the potential social damage will be covered by the deposit of creators and the decentralized insurance system.

Keywords creator economy, Web3.0, decentralized censorship, blockchain, social media