Evolution of Domain-Specific Reputation Systems: From Binary Validation to Citation-Weighted Knowledge Attribution

Wulf Kaal, Ph.D.[1]

Abstract

Decentralized reputation systems are essential for autonomous collaboration in blockchain-based ecosystems. Yet existing frameworks often fail to adequately attribute value across cumulative knowledge contributions, capture nuanced quality assessments, or incentivize sharing in multi-agent environments. These shortcomings are exacerbated by the emerging AI agent knowledge economy. This paper builds on the foundational Calcaterra, Kaal, Andrei (2018) architecture by introducing a citation-weighted multi-agent reputation model. Key innovations include configurable multi-agent competition for improved output quality, mandatory normalized citation graphs to enable compounding attribution of foundational work, consensus-based ranked validation via weighted Borda count, and PageRank-inspired value allocation that ensures game-theoretic incentive alignment for honest participation and citation. Formal proofs demonstrate convergence, uniqueness, truth-telling equilibria, and enhanced attack resistance. With corruption costs scaling to at least 2-4 times total system reputation under realistic conditions. The framework addresses core limitations of binary validation, including the citation paradox, quality-efficiency tradeoff, and information destruction. At the same time, the framework preserves computational tractability for on-chain implementation. Empirical predictions and testable hypotheses are derived for quality gains, reputation concentration, dispute reduction, and knowledge graph dynamics in AI-agent and DAO contexts. The model offers a pathway toward sustainable, autonomous knowledge ecosystems where foundational contributions are durably rewarded.

Keywords: decentralized reputation systems, blockchain governance, citation-weighted attribution, multi-agent competition, decentralized autonomous organizations (DAOs), PageRank reputation, game-theoretic incentives, knowledge graphs, Sybil resistance, AI-DAO convergence

JEL Classification: D82, D83, D85, L14, O31, O33, G34, P00

[1] Professor of Law. University of St. Thomas School of Law (Minneapolis, MN). The author is extremely grateful for the ever evolving theoretical discussions with Craig Calcaterra and Jonathan Kung as well as AI law-centric WDAG discussions with Morgan Gray. The author is also very grateful for excellent research assistance by Mickey Bernardi.

Table of Contents

I. Introduction: The Fundamental Problem of Decentralized Knowledge Validation

A. The Challenge of Autonomous Reputation Systems

Modern digital economies increasingly rely on decentralized networks where anonymous or pseudonymous agents collaborate to produce valuable outputs—whether code, analysis, creative work, or specialized services. From open-source software development[2] to decentralized autonomous organizations[3] to emerging AI agent marketplaces,[4] a fundamental challenge persists: How do we measure and reward expertise when identity is fluid, contributions are collaborative, and no central authority exists to adjudicate quality?

Traditional reputation systems fail in decentralized contexts for three reasons:

First, the identity problem: Reputation systems built on persistent identity[5] collapse when agents can costlessly create new identities. A poorly-performing agent simply abandons their account and starts fresh, a phenomenon known as whitewashing.[6] Sybil attacks, where one entity controls multiple identities, can manipulate vote-based systems.[7]

Second, the centralization problem: Platforms that solve the identity problem through centralized control[8] introduce rent-seeking intermediaries who extract value, censor participants, and create single points of failure. The very purpose of decentralized systems is to

[2] Raymond 1999.
[3] Buterin 2014; Kaal 2021.
[4] Kaal 2026.
[5] for example: eBay, Upwork, academic citation indices.
[6] Friedman and Resnick 2004.
[7] Douceur 2002.
[8] Uber ratings, Amazon reviews.

eliminate these trusted third parties:[9] As I have argued extensively in prior work, this tension between the need for regulation and the desire for decentralization creates what I term the "pacing problem"—regulatory frameworks cannot keep pace with technological innovation.[10] The solution lies not in static centralized control but in dynamic, autonomous governance mechanisms.

Third, the attribution problem. Knowledge work is inherently cumulative. Alice's insight enables Bob's improvement, which enables Carol's application. Yet most reputation systems treat each contribution as atomic and independent, failing to capture this evolving knowledge graph. The result is systematic undervaluation of foundational work and perverse incentives to hoard rather than share insights.[11]

B. Existing Approaches and Their Limitations

Previous attempts to build decentralized reputation systems fall into several categories, each with fundamental limitations that my co-authors and I sought to address in our 2018 publication.[12]

Token-curated registries[13] use staking and voting to validate entries in a list. Participants stake tokens to propose entries. Others stake to challenge them. While this creates economic security against spam, it produces only binary outcomes as in accept or reject heuristics and provides no mechanism for nuanced quality assessment or attribution of collaborative contributions.

[9] Nakamoto 2008.
[10] Kaal 2016.
[11] Heller 1998; Benkler 2006.
[12] Calcaterra, Kaal, Andrei 2018.
[13] Goldin et al. 2017.

Prediction markets[14] aggregate information through trading, with prices reflecting consensus beliefs. These excel at forecasting but poorly capture non-binary quality dimensions and provide no framework for rewarding contributors of foundational insights that others build upon.

Proof-of-work and proof-of-stake consensus[15] solve the double-spend problem in cryptocurrencies but are fundamentally about ordering transactions, not evaluating quality of knowledge contributions. The "work" being proven has no intrinsic value beyond securing the network.

Academic citation networks[16] successfully attribute foundational contributions through citation indices like h-index and PageRank-style metrics.[17] However, these systems operate in centralized institutional contexts with stable identities and long time horizons. They also suffer from citation gaming, prestige bias, and poor mechanisms for real-time quality assessment.[18]

The Calcaterra-Kaal-Andrei framework (2018): In 2018, my co-authors Craig Calcaterra, Vlad Andrei, and I presented a blockchain-based infrastructure for measuring domain-specific reputation in autonomous, decentralized, and anonymous systems.[19] This framework introduced several innovations:

1. Expertise-specific reputation tokens ("sem/rep tokens") that prevented simple purchase of reputation through Sybil attacks.
2. Validation pools where experts stake reputation to vote on quality, with winners splitting losers' stakes.
3. Mathematical proof that corrupting the system costs at minimum twice the total system reputation.

[14] Hanson 2003; Peterson and Krug 2015.
[15] Nakamoto 2008; Buterin and Griffith 2017.
[16] Garfield 1955; Hirsch 2005.
[17] Page et al. 1999.
[18] Bornmann and Daniel 2008.
[19] Calcaterra, Kaal, and Andrei 2018.

4. Complete autonomy—no centralized authority controls reputation allocation.

The framework was groundbreaking in demonstrating that decentralized reputation could resist economic attacks while operating autonomously. However, as decentralized systems have evolved, particularly with the emergence of AI agent collaboration,[20] three critical limitations of the original design have become apparent, limitations this paper addresses through substantial mathematical refinement.

C. The Core Mathematical Challenge

At its heart, the problem is this: We need a mathematical framework that:

1. Resists strategic manipulation - despite anonymous participants and no central authority
2. Captures nuanced quality - beyond binary accept/reject decisions
3. Attributes value correctly - across chains of cumulative contributions
4. Incentivizes knowledge sharing - rather than hoarding through game-theoretic soundness
5. Operates autonomously - without requiring human adjudication at scale
6. Remains computationally tractable - for implementation on blockchain or distributed systems

The Calcaterra-Kaal-Andrei framework (2018) achieved properties 1, 5, and 6 but struggled with 2, 3, and 4. Existing frameworks achieve at most 2-3 of these properties simultaneously. This paper presents a unified mathematical framework achieving all six through three core innovations: citation-weighted contribution graphs, multi-agent competitive collaboration, and consensus-based validation pools with game-theoretically sound incentive alignment.

D. Motivating Application: AI Agent Learning Ecosystems

[20] Kaal 2026.

The urgency of this problem is magnified by the emergence of AI agent marketplaces.[21] As AI capabilities advance, agents increasingly perform knowledge work which may include any services from code generation to data analysis to research synthesis. My recent work on AI-DAO convergence[22] demonstrates that the governance challenges facing traditional DAOs become exponentially more complex when autonomous AI agents participate in organizational decision-making.

We need infrastructure that:

- Allows agents to collaborate and build on each other's contributions
- Creates verifiable track records of expertise without centralized intermediaries
- Rewards foundational contributions that spawn derivative innovations
- Resists manipulation by malicious agents or Sybil attacks
- Operates autonomously at machine speed and scale

Current platforms either use centralized rating systems (defeating the purpose of agent autonomy) or primitive token-staking mechanisms (producing binary outcomes and failing to capture knowledge graphs). As I have argued in the context of DAO governance,[23] neither approach is adequate for the sophisticated knowledge ecosystems required for genuine decentralized collaboration—whether among humans or AI agents.[24]

E. Structure of This Paper

Section II presents the baseline mathematical framework from Calcaterra, Kaal, and Andrei (2018), showing how binary voting with reputation staking creates basic security properties while identifying specific mathematical and design limitations. Section III identifies three critical

[21] Kaal 2026.
[22] Kaal 2026.
[23] Kaal 2021; Kaal and Calcaterra 2017.
[24] Kaal 2026.

shortcomings: the citation paradox (failure to attribute foundational contributions), the quality-efficiency paradox (single-agent selection sacrifices quality assurance), and binary information destruction (collapse of nuanced assessment into one bit).

Section IV introduces the revised mathematical framework: citation-weighted multi-agent reputation systems. We present formal definitions, prove convergence and uniqueness properties, establish game-theoretic equilibria for honest validation, and derive attack resistance bounds.

Section V analyzes computational complexity and implementation considerations. Section VI presents empirical predictions and testable hypotheses. Section VII concludes by discussing implications for AI agent ecosystems and future research directions.

II. Baseline Framework: The Calcaterra-Kaal-Andrei Architecture (2018)

A. System Components and Foundational Design

The 2018 framework introduced a decentralized autonomous platform for validating domain-specific reputation through blockchain infrastructure.[25] The system architecture consists of four main components:

Agents (denoted $E_1, E_2, \ldots, E_n$): Pseudonymous participants who perform work and validate each other's contributions. Anonymity enables meritocratic evaluation while preventing discrimination based on identity.[26]

[25] Calcaterra, Kaal, and Andrei 2018.
[26] ibid., 4.

Expertise categories (denoted $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$): Distinct domains of knowledge (e.g., "smart contract auditing," "legal analysis," "data science"). Each expertise maintains separate reputation tokens, preventing reputation in one domain from conferring unearned authority in another.

Reputation tokens $R(i, e)$: Agent i's reputation in expertise category e, represented as fungible tokens that cannot be transferred between agents but can be staked. These "sem tokens" are minted in proportion to fees paid into the system, creating a dynamic economy where reputation value reflects actual usage.[27]

Jobs J: Tasks posted by users requiring expert work. Each job specifies:
- Task description T
- Payment F (in currency tokens, e.g., ETH, BTC)
- Required expertise category e
- Validation period τ

Validation pools VP: Mechanisms where experts stake reputation to vote on whether submitted work meets quality standards. This betting pool structure creates economic consequences for evaluation decisions, incentivizing careful assessment.[28]

B. Basic Validation Pool Mechanism

The canonical flow operates as follows:[29]

Step 1: Job posting and agent selection

[27] ibid., 6.
[28] ibid., 7.
[29] Calcaterra, Kaal, and Andrei 2018, 8-9.

User posts job J with payment F. An agent i with reputation $R(i,e) > 0$ in the relevant expertise stakes $S(i)$ reputation tokens to signal availability. The system selects an agent using weighted random selection:

$$P(\text{select agent } i) = \frac{S(i)}{\sum_j S(j)}$$

Where the sum is over all agents who staked for availability.

This random weighted selection prevents powerful experts from monopolizing work while giving higher-reputation agents proportionally greater opportunity.[30]

Step 2: Work submission and validation initiation

Selected agent i performs work off-chain and submits evidence-of-work post P. The user's payment F triggers the validation pool:

- Fee distribution: F is immediately distributed to all experts in category e as "salary" proportional to their reputation:

$$\text{Salary}(j) = \frac{R(j,e)}{\sum_k R(k,e)} \times F$$

This innovative design ensures all experts benefit from growth of the expertise domain, aligning incentives toward collective success rather than zero-sum competition.[31]

[30] ibid., 11.
[31] ibid., 6.

- Token minting: System mints M new reputation tokens where $M = \alpha \times F$ (α is the exchange rate parameter)

- Initial stakes: $M/2$ tokens are staked as "upvote" for agent i; $M/2$ tokens are staked as "downvote" (unassigned)

This 50/50 split is crucial: it prevents agents from simply purchasing reputation, since established experts control whether new reputation is granted.[32]

Step 3: Expert voting

During validation period τ, other experts can stake their own reputation tokens to vote:
- Upvote: Stake tokens betting the work meets quality standards
- Downvote: Stake tokens betting the work fails quality standards

Step 4: Resolution

After period τ expires:

$$\text{Total\_upvotes} = \frac{M}{2} + \sum_{j \in \text{upvoters}} \text{Stake}(j)$$

$$\text{Total\_downvotes} = \frac{M}{2} + \sum_{k \in \text{downvoters}} \text{Stake}(k)$$

If $\text{Total\_upvotes} \geq \text{Total\_downvotes}$, upvote wins (ties go to upvote).

Winners split all staked tokens proportionally:

[32] ibid., 10.

$$\text{Payout}(j) = \text{Stake}(j) + \frac{\text{Stake}(j)}{\sum_{\text{winners}} \text{Stake}} \times \sum_{\text{losers}} \text{Stake}$$

If downvote wins, the $M/2$ unassigned tokens are burned (destroyed).

C. Security Properties: Attack Resistance Mathematics

One of the most significant contributions of Calcaterra, Kaal, and Andrei (2018) was the rigorous mathematical proof of attack resistance. The critical question: What is the cost to corrupt this system?

Scenario: A malicious actor M wants to gain reputation through fee payments to eventually control the system. The cleanest attack is to make many small fee payments, getting staked upvotes each time, until M controls 51% of reputation.

Mathematical analysis:[33]

Let R_0 = total system reputation at time 0 (all held by honest "good-faith" experts initially)

Let F = cumulative fees paid by malicious actor

Let $G(F)$ = reputation held by good-faith experts after fees F paid

Let $B(F)$ = reputation held by malicious "bad" actor after fees F paid

Initially: $G(0) = R_0$, $B(0) = 0$

[33] Calcaterra, Kaal, and Andrei 2018, 15-18.

Differential equations: For infinitesimal fee increment dF:

When malicious actor pays fee dF:

1. System mints $\alpha \times dF$ new tokens
2. Half $(\alpha \times dF/2)$ go to malicious actor as upvote stake
3. All experts receive salary proportional to reputation
4. Malicious actor gains salary: $\frac{B(F)}{R_0 + F} \times dF$
5. Good-faith experts gain salary: $\frac{G(F)}{R_0 + F} \times dF$

Assuming malicious actor always wins validation pool (worst case), we get:

$$\frac{dG}{dF} = \frac{1}{2}\left(\frac{G}{R_0 + F}\right)$$

$$\frac{dB}{dF} = \frac{1}{2}\left(\frac{B}{R_0 + F}\right) + \frac{1}{2}$$

The first equation says: good-faith experts gain half the minted tokens (the downvote stake that gets won) proportional to their share of reputation.

The second equation says: malicious actor gains half the minted tokens directly (upvote stake) PLUS another half that they win from the validation pool, proportional to their reputation share.

Solving the ODEs:

Using integrating factor method for the first equation:

$$G(F) = \frac{6R_0}{7 + R_0 F}$$

For the second equation:

$$B(F) = F + R_0 - \frac{6R_0}{7 + R_0 F}$$

When does malicious actor achieve 51%?

Setting $B(F) = G(F)$:

$$F + R_0 - \frac{6R_0}{7 + R_0 F} = \frac{6R_0}{7 + R_0 F}$$

$$F + R_0 = \frac{12R_0}{7 + R_0 F}$$

$$(F + R_0)(7 + R_0 F) = 12R_0$$

Solving yields: $F = 3R_0$

However, during this process, the malicious actor receives salary payments totaling:

$$\text{Recovered} = \int_0^{3R_0} \frac{B(F)}{R_0 + F}\, dF = R_0$$

Net cost to achieve 51% control: $3R_0 - R_0 = 2R_0$

This is the fundamental security bound: corrupting the system costs at minimum 2× the total reputation value.[34]

This mathematical proof represented a major advance in decentralized system security. However, subsequent analysis reveals this bound applies only under specific conditions that may not hold in practice. This is a limitation we address in Section IV.F.

D. Identified Limitations of the Original Framework

While the Calcaterra-Kaal-Andrei framework was elegant and groundbreaking, analysis over the subsequent six years—particularly in light of evolving DAO governance challenges,[35] reveals three critical shortcomings:

Problem 1: Binary validation only. The system produces only binary outcomes (accept/reject). No mechanism exists to say "good but not great" or "excellent innovation, poor execution." As I noted in my work on dynamic regulation,[36] binary regulatory frameworks cannot capture the nuanced gradations necessary for efficient governance. The same limitation applies to reputation systems.

Problem 2: Single agent selection. The system randomly picks one agent per job, even if having multiple agents compete would improve quality. This reflects a broader pattern I have observed in DAO governance: efficiency optimization often sacrifices quality assurance.

Problem 3: No citation/attribution mechanism. If Alice creates foundational insight that Bob builds on, Alice receives no credit when Bob's work gets validated. The original framework

[34] Calcaterra, Kaal, and Andrei 2018, 17.
[35] Kaal 2026; Kaal and Calcaterra 2017.
[36] Kaal 2014; 2016.

included a citation graph concept,[37] but the mathematics were underdeveloped and no game-theoretic analysis established incentives for honest citation.

As noted in the original paper, "the value of each post is eternally dynamic" as new references change past post valuations.[38] However, the recursive formula presented suffered from potential instability and provided no mechanism ensuring citations reflect actual contribution rather than strategic manipulation.

These limitations are not minor inefficiencies—they fundamentally prevent the system from functioning as a learning ecosystem, which is essential for AI-DAO convergence applications.[39]

III. Critical Shortcomings: Three Mathematical Failures

A. The Citation Paradox: Missing Knowledge Attribution

The Problem Formalized:

The Calcaterra-Kaal-Andrei framework's most significant oversight was the underdevelopment of citation attribution. While Section 9.2 of the original paper introduced weighted directed acyclic graphs (DAGs) for the forum structure,[40] the implementation remained theoretical and mathematically incomplete.

Consider a sequence of knowledge contributions:

[37] Calcaterra, Kaal, and Andrei 2018, 40-42.
[38] ibid., 42.
[39] Kaal 2026.
[40] Calcaterra, Kaal, and Andrei 2018, 40-42.

- Time t_1: Agent Alice creates foundational insight I_1
- Time t_2: Agent Bob builds improvement I_2 using I_1
- Time t_3: Agent Carol creates application I_3 using I_1 and I_2

Under binary validation with no functional citation mechanism:

$$\text{Reputation}_{\text{Alice}}(t_3) = \text{Reputation}_{\text{Alice}}(t_1)$$

$$\text{Reputation}_{\text{Bob}}(t_3) = \text{Reputation}_{\text{Bob}}(t_2)$$

$$\text{Reputation}_{\text{Carol}}(t_3) = \text{Reputation}_{\text{Carol}}(t_3)$$

There is no compounding effect. Alice's foundational contribution becomes invisible after t_1. Her reputation does not grow as others build on her work.

Economic consequence: Alice has no incentive to share insights that others can build on. Optimal strategy becomes hoarding knowledge or publishing only when you can capture full value yourself. This creates what Heller (1998) termed a "tragedy of the anticommons"—socially valuable knowledge sharing is suppressed.

This problem is particularly acute in the context of blockchain-based organizations. As I have argued extensively,[41] DAOs fundamentally depend on transparent information flow and cumulative knowledge building. A reputation system that fails to reward foundational contributions undermines the core value proposition of decentralized knowledge work.

Contrast with ideal knowledge economics:

[41] Kaal 2021; Kaal and Calcaterra 2017.

In efficient knowledge network economies,[42] foundational contributions should accumulate value over time. If I_1 enables 10 downstream innovations, Alice should capture some fraction of that derivative value. This principle aligns with my broader work on innovation-enabling regulatory frameworks,[43] systems must reward not just immediate output but enabling infrastructure.

Academic citation systems achieve this through citation counts and h-indices. Alice's seminal paper accumulates citations as others build on it. The h-index specifically captures this: an h-index of 50 means 50 papers cited at least 50 times each, rewarding sustained foundational impact.

Mathematical representation of ideal system:

Alice's total value should be:

$$\text{Value}_{\text{Alice}} = \text{Direct\_value}(I_1) + \sum_{t} \sum_{j \in \text{Cited}_{\text{Alice}}(t)} \text{Attribution\_fraction}(j, \text{Alice}, t) \times \text{Value}(j)$$

Where $\text{Cited}_{\text{Alice}}(t)$ is the set of works at time t that cite Alice's contribution.

The Calcaterra-Kaal-Andrei framework's binary validation system sets $\text{Attribution\_fraction} \equiv 0$, completely losing this compounding effect. The proposed DAG weighting in Section 9.2 attempted to address this but provided no game-theoretic analysis of citation incentives and no proof of convergence for the recursive valuation formula.[44]

[42] Romer 1990; Shapiro and Varian 1998.
[43] Kaal 2016.
[44] Calcaterra, Kaal, and Andrei 2018, 41.

B. The Quality-Efficiency Paradox

The Problem:

The Calcaterra-Kaal-Andrei framework selects a single agent per job through weighted random selection.[45] While this appears efficient, it is mathematically suboptimal for quality assurance. A pattern I have observed across DAO governance structures.

Analysis:

Let q_i represent agent i's true quality (unobservable), and r_i their reputation (observable proxy). Assume reputation correlates with quality: $r_i \propto q_i$, but imperfectly.

Single-agent random selection:

Probability of selecting agent i:

$$P(\text{select } i) = \frac{r_i}{\sum_j r_j}$$

Expected quality of selected agent:

$$E[Q_{\text{single}}] = \sum_i \frac{r_i}{\sum_j r_j} \times q_i$$

Even with perfect correlation ($r_i = q_i$), this is just the reputation-weighted average quality.

Multi-agent competition:

[45] Calcaterra, Kaal, and Andrei 2018, 11.

Suppose k agents compete, and we select the best output. Expected quality:

$$E[Q_{\text{multi}}] = E[\max(q_1, q_2, \ldots, q_k)]$$

For quality distributed as $q \sim N(\mu, \sigma^2)$, order statistics give:

$$E[\max(q_1, \ldots, q_k)] \approx \mu + \sigma\sqrt{2\ln(k)}$$

Quality improvement from competition:

$$\text{Quality\_gain} = E[Q_{\text{multi}}] - E[Q_{\text{single}}] \approx \sigma\sqrt{2\ln(k)}$$

This grows with $\sqrt{\ln k}$, creating a "diversity dividend."

Numerical example:

If $\sigma = 0.3$ (30% quality standard deviation):

$$k = 1: \quad \text{Quality\_gain} = 0$$

$$k = 3: \quad \text{Quality\_gain} \approx 0.3 \times \sqrt{2\ln 3} \approx 0.3 \times 1.48 \approx 0.44 \text{ standard deviations}$$

$$k = 5: \quad \text{Quality\_gain} \approx 0.3 \times \sqrt{2\ln 5} \approx 0.3 \times 1.80 \approx 0.54 \text{ standard deviations}$$

$$k = 10: \quad \text{Quality\_gain} \approx 0.3 \times \sqrt{2\ln 10} \approx 0.3 \times 2.15 \approx 0.64 \text{ standard deviations}$$

With even moderate competition ($k = 5$), expected quality improves by over half a standard deviation—the difference between median and top-quartile performance.

Why didn't the original framework use this?

Cost. Paying 5 agents to do the same work seems wasteful. But this assumes work value is independent of quality, which is false for knowledge work. A 50% quality improvement on a $10,000 project delivers 5,000+ of extra value, easily justifying the extra cost.

The real barrier is the lack of attribution mechanism. If 5 agents compete but only the best gets paid, the other 4 have no incentive to participate. We need a framework where all contributors can be rewarded proportional to their contribution. This requires citation-weighted attribution that the original framework lacked.

This reflects a broader principle from my work on dynamic regulation:[46] governance systems must adaptively balance competing objectives. Static optimization for a single metric (here, cost efficiency) produces suboptimal outcomes when multiple dimensions of value exist (here, quality, innovation, knowledge accumulation).

C. Binary Information Destruction

The Problem:

Binary upvote/downvote in the Calcaterra-Kaal-Andrei framework collapses rich multidimensional quality information into a single bit.[47]

Example:

[46] Kaal 2014; 2016.
[47] Calcaterra, Kaal, and Andrei 2018, 7-8.

Five validators assess a piece of work. Reality of their assessments:

- Validator 1: "Brilliant innovation (9/10) but flawed execution (4/10)"
- Validator 2: "Derivative idea (5/10) but perfectly implemented (9/10)"
- Validator 3: "Novel approach (8/10), needs refinement (6/10)"
- Validator 4: "Excellent synthesis (7/10) of prior work (8/10)"
- Validator 5: "Incremental improvement only (5/10)"

Binary system output:

Votes: 3 upvote (validators 1, 3, 4), 2 downvote (validators 2, 5)
Outcome: ACCEPT (bit value = 1)
Information captured: 1 bit

Information lost:

- Which dimensions (innovation vs. execution) were strong/weak?
- Who contributed what insights?
- What aspects should be preserved in future work?
- Which validator assessments were most accurate?

This lost information is precisely what's needed to:
1. Improve work through iteration
2. Attribute credit to specific contributions
3. Build knowledge graphs showing how ideas evolve
4. Identify validators whose assessments are most reliable

Information-theoretic analysis:

With 5 validators each providing 2-dimensional assessments (innovation, execution) on 10-point scales:

$$\text{Information available} = 5 \times 2 \times \log_2(10) \approx 5 \times 2 \times 3.32 \approx 33 \text{ bits}$$

$$\text{Information captured by binary vote} = 1 \text{ bit}$$

$$\text{Information efficiency} = \frac{1}{33} \approx 3\%$$

We're destroying 97% of the available quality signal.

Consequence for knowledge graphs:

Without granular attribution, we cannot build a meaningful citation graph. We don't know that Alice's innovation insight was valuable even though her execution was poor, or that Bob's contribution was primarily in implementation rather than conceptual novelty.

This makes it impossible for future agents to:
- Identify who to cite for which contributions
- Understand the evolution of ideas over time
- Learn from patterns of successful collaboration

This limitation is particularly problematic for AI agent ecosystems. As I have argued,[48] AI-DAO convergence requires machine-readable governance structures that preserve semantic richness. Binary outcomes fail this requirement fundamentally.

[48] Kaal 2026.

IV. Revised Mathematical Framework: Citation-Weighted Multi-Agent Reputation

Building on the foundational security properties established by Calcaterra, Kaal, and Andrei (2018) while addressing the three critical shortcomings identified above, we now present a unified framework that enables genuine learning ecosystems. This framework operationalizes principles from my broader work on dynamic regulation[49] and DAO governance[50] within autonomous reputation systems.

A. Multi-Agent Competitive Collaboration Protocol

Definition 1 - Job Specification:

A job J is defined by the tuple:

$$J = (T, F, E, k, \tau)$$

Where:

- T = task description (string/hash)
- F = payment in ETHtokens (positive real number)
- E = set of required expertise categories (subset of all expertise categories)
- k = number of competing agents (positive integer)
- τ = validation period (time duration)

This extends the original framework's job specification (Calcaterra, Kaal, and Andrei 2018, 8) by adding parameter k, enabling configurable competition levels.

[49] Kaal 2014, 2016.
[50] Kaal 2021, 2025.

Definition 2 - Agent Competition Phase:

Upon job posting, agents with positive reputation in any expertise category in E can stake reputation tokens to compete.

Agent i stakes S_i reputation tokens.

System selects k agents with selection probability:

$$P(\text{select agent } i) = \frac{S_i \sum_{e \in E} R(i,e)}{\sum_j S_j \sum_{e \in E} R(j,e)}$$

Where:

- $R(i,e)$ = agent i's reputation in expertise category e
- Sum over j includes all agents who staked

This preserves the weighted random selection from the original framework while extending it to select k agents rather than one. The weighting by both stake size and total relevant reputation maintains the original design's resistance to Sybil attacks (Calcaterra, Kaal, and Andrei 2018, 13).

Definition 3 - Citation-Weighted Output:

Each selected agent i submits:

1. Output O_i (the actual work product)
2. Citation set C_i consisting of pairs (j, c_{ij}) where:
   - j is another agent who contributed to i's work
   - $c_{ij} \in [0,1]$ representing j's fractional contribution to i's output

- Constraint: $\sum_j c_{ij} = 1$ (citations sum to 100%)

This creates a directed graph $G = (V, E, c)$ where:

- Vertices V = agents
- Edges E = citations
- Edge weights c_{ij} = contribution fractions

This operationalizes the citation graph concept introduced in Section 9.2 of Calcaterra, Kaal, and Andrei (2018, 40-42), but with critical additions: (1) citations are mandatory and constrained to sum to 1, preventing arbitrary weighting, and (2) we provide game-theoretic analysis establishing citation honesty as an equilibrium (see Section IV.E).

Example:

Agent Alice submits work citing:

- Bob with weight 0.6 (Bob contributed 60% of the insights)
- Carol with weight 0.3 (Carol contributed 30%)
- Direct original contribution: 0.1 (10%)

Citation set: $C_{\text{Alice}} = \{(\text{Bob}, 0.6), (\text{Carol}, 0.3), (\text{Alice}, 0.1)\}$

B. Validator Consensus and Quality Ranking

Definition 4 - Validation Pool:

m validators each stake V_ℓ reputation tokens ($\ell = 1$ to m).

Each validator ℓ produces a ranking π_ℓ of the k submissions:

$$\pi_\ell : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$$

Where $\pi_\ell(i)$ = rank assigned to agent i's submission (1 = best, k = worst)

This replaces the binary upvote/downvote from Calcaterra, Kaal, and Andrei (2018, 7-9) with ranked preferences, preserving the full information content of validator assessments.

Definition 5 - Consensus Ranking via Weighted Borda Count:

Each submission i receives a Borda score:

$$\text{Score}(i) = \sum_{\ell=1}^{m} V_\ell \times (k + 1 - \pi_\ell(i))$$

This is a weighted Borda count where:
- Each validator's vote is weighted by their stake V_ℓ (preserving the reputation-weighted voting from the original framework)
- Higher ranks (smaller π values) contribute more to the score
- $\text{Score}(i)$ is maximized when all validators rank i first

The consensus ranking orders submissions by $\text{Score}(i)$.

Quality vector q:

Normalize scores to create quality vector:

$$q_i = \frac{\text{Score}(i)}{\sum_j \text{Score}(j)}$$

Where $0 \leq q_i \leq 1$ and $\sum_i q_i = 1$

This represents the fraction of "quality" attributed to each submission by validator consensus.

Theorem 1 - Consensus Convergence:

Under weighted Borda count, if validators have identical preferences, the mechanism converges to the true ranking. If a fraction $p > 1/2$ of validators (weighted by stake) are honest, the plurality outcome matches the honest ranking.

Proof:

The Borda count is a positional voting method satisfying:

1. Unanimity: If all validators rank $i > j$, then $\text{Score}(i) > \text{Score}(j)$

Proof: If $\pi_\ell(i) < \pi_\ell(j)$ for all ℓ, then $(k + 1 - \pi_\ell(i)) > (k + 1 - \pi_\ell(j))$ for all ℓ, thus

$$\sum_\ell V_\ell(k + 1 - \pi_\ell(i)) > \sum_\ell V_\ell(k + 1 - \pi_\ell(j))$$

2. Monotonicity: If validator ℓ improves i's rank (decreases $\pi_\ell(i)$), $\text{Score}(i)$ increases

Proof:

$$\frac{\partial \text{Score}(i)}{\partial \pi_\ell(i)} = V_\ell \times (-1) < 0$$

so decreasing $\pi_\ell(i)$ increases $\text{Score}(i)$

3. Majority-Respecting: With $p > 1/2$ honest validators by stake:

$$\sum_{\ell \in \text{honest}} V_\ell > \sum_{\ell \in \text{dishonest}} V_\ell$$

If all honest validators rank i first ($\pi_\ell(i) = 1$), then:

$$\text{Score}(i) \geq \sum_{\text{honest}} V_\ell \times k > \sum_{\text{dishonest}} V_\ell \times k \geq \text{Score}(j) \text{ for any } j \neq i$$

Therefore i wins the consensus ranking.

QED.

This theorem establishes that the revised framework preserves the security property from Calcaterra, Kaal, and Andrei (2018) that 51% good-faith participation ensures correct outcomes, while extending it to richer quality assessment.

C. Citation-Based Reward Allocation via PageRank

This is the core innovation addressing the citation paradox. Rather than binary accept/reject, we allocate rewards using the citation graph. Thus, operationalizing the underdeveloped DAG concept from Calcaterra, Kaal, and Andrei (2018, 40-42) with rigorous mathematical foundations.

Definition 6 - Contribution Matrix:

Define $k \times k$ matrix C where:

$$C_{ij} = \begin{cases} c_{ij} & \text{if agent } i \text{ cited agent } j \\ 0 & \text{otherwise} \end{cases}$$

Row normalization: $\sum_j C_{ij} = 1$ for all i (each row sums to 1).

This makes C a row-stochastic matrix.

Definition 7 - PageRank Attribution:

The citation-weighted value vector v (k-dimensional) satisfies:

$$v = (1 - \alpha)q + \alpha C^T v$$

Where:

- q = quality vector from validator consensus (Definition 5)
- α = citation weight parameter, $0 \leq \alpha < 1$
- C^T = transpose of contribution matrix
- v = value vector we're solving for

Matrix form:

$$v = (1 - \alpha)q + \alpha C^T v$$

$$v - \alpha C^T v = (1 - \alpha)q$$

$$(I - \alpha C^T)v = (1 - \alpha)q$$

$$v = (I - \alpha C^T)^{-1}(1 - \alpha)q$$

Where I is the $k \times k$ identity matrix.

Interpretation:

Each agent's value has two components:

1. Direct quality: $(1 - \alpha)q_i$ from validator assessment
2. Attributed quality: $\alpha\times$ sum of values of agents who cited them

The citation weight α controls how much value flows through citations vs. direct assessment.

This approach draws inspiration from PageRank[51] but adapts it for reputation allocation rather than web page ranking. The connection to my broader work on network-based governance[52] is clear: value in decentralized systems flows through contribution networks, and governance mechanisms must capture these network effects.

Theorem 2 - Existence and Uniqueness:

If $\alpha < 1$ and C is row-stochastic, then $(I - \alpha C^T)$ is invertible and the value vector v exists and is unique.

[51] Page et al. 1999.
[52] Kaal 2021; Kaal and Calcaterra 2017.

Proof:

Since C is row-stochastic, C^T is column-stochastic. By Perron-Frobenius theorem, the spectral radius $\rho(C^T) \leq 1$.

For $\alpha < 1$, we have:

$$\rho(\alpha C^T) = \alpha \rho(C^T) \leq \alpha < 1$$

This means all eigenvalues of αC^T have absolute value < 1.

Therefore all eigenvalues of $I - \alpha C^T$ have the form $1 - \lambda$ where $|\lambda| < 1$, so all eigenvalues of $I - \alpha C^T$ satisfy $|1 - \lambda| > 0$.

Thus $I - \alpha C^T$ is non-singular (invertible).

The inverse can be expressed as a convergent Neumann series:

$$(I - \alpha C^T)^{-1} = \sum_{n=0}^{\infty} (\alpha C^T)^n$$

This series converges because $\rho(\alpha C^T) < 1$.

Therefore $v = (I - \alpha C^T)^{-1}(1 - \alpha)q$ exists and is unique.

QED.

This theorem addresses the instability concern with the recursive formula in Calcaterra, Kaal, and Andrei (2018, 41). By constraining $\alpha < 1$ and ensuring row-stochasticity through normalization, we guarantee convergence, a crucial property for autonomous operation.

Computational formula:

$$v = (1-\alpha)\left[I + \alpha C^T + (\alpha C^T)^2 + (\alpha C^T)^3 + \cdots\right] q$$

$$= (1-\alpha)\left[q + \alpha(C^T)q + \alpha^2(C^T)^2 q + \alpha^3(C^T)^3 q + \cdots\right]$$

This shows that v includes:

- Direct quality q
- First-order citations $\alpha(C^T)q$
- Second-order citations $\alpha^2(C^T)^2 q$ (being cited by someone who was cited)
- And so on, with exponentially decaying weights

D. Token and Payment Distribution

Definition 8 - Reward Distribution:

The F ETHtokens are allocated as follows:

$$\text{Performance rewards: } (1-\beta-\gamma)F \text{ distributed to job performers}$$

$$\text{Validation rewards: } \beta F \text{ distributed to validators}$$

$$\text{Burn: } \gamma F \text{ permanently removed from circulation}$$

Where $0 < \beta, \gamma < 1$ and $\beta + \gamma < 1$.

This preserves the fee distribution structure from Calcaterra, Kaal, and Andrei (2018, 10) where validators receive salary proportional to reputation stakes, ensuring all participants benefit from system growth.

Performance allocation:

Agent i receives:

$$\text{Payment}_i = v_i \times (1 - \beta - \gamma) \times F$$

Where v_i is from the citation-weighted value vector (Definition 7).

Validation allocation:

Validator ℓ receives:

$$\text{Payment}_\ell = \frac{V_\ell}{\sum_j V_j} \times \beta \times F$$

Proportional to their stake in the validation pool.

Reputation reallocation:

Staked job-performer reputation is reallocated based on performance relative to average:

$$\Delta R_i = S_i \times [(k \times v_i) - 1]$$

Where:

- S_i = agent i's staked reputation
- $k \times v_i$ = agent i's value rescaled (recall $\sum v_i = 1$, so average is $1/k$)
- If $v_i > 1/k$ (above average), agent gains: $\Delta R_i > 0$
- If $v_i < 1/k$ (below average), agent loses: $\Delta R_i < 0$

This creates competitive pressure: agents must perform better than average to gain reputation. Thus, extending the winner-takes-all mechanism from the original binary framework to a proportional system that rewards degrees of excellence.

Example calculation:

Job with $k = 5$ agents, each staking $S = 100$ reputation.

Validator consensus produces quality vector:

$$q = [0.4, 0.25, 0.2, 0.1, 0.05]$$

Agent 1 ranked best ($q_1 = 0.4$), agent 5 ranked worst ($q_5 = 0.05$).

Citation matrix (each row sums to 1):

matrix}
0.5 & 0.3 & 0.2 & 0.0 & 0.0 \\
0.6 & 0.3 & 0.1 & 0.0 & 0.0 \\
0.4 & 0.4 & 0.1 & 0.1 & 0.0 \\
0.3 & 0.3 & 0.3 & 0.05 & 0.05 \\
0.4 & 0.3 & 0.2 & 0.1 & 0.0

/end matrix}

With $\alpha = 0.7$, compute:

$$v = (I - 0.7C^T)^{-1} \times 0.3q$$

Numerical solution yields approximately:

$$v \approx [0.45, 0.28, 0.18, 0.06, 0.03]$$

Notice agent 1's value increased from $q_1 = 0.40$ to $v_1 = 0.45$ because agents 2, 3, 4, 5 all cited agent 1, creating citation value flow. This is precisely the compounding effect missing from the original framework.

With payment $F = 1000$ ETH, $\beta = 0.2$, $\gamma = 0.05$:

$$\text{Agent 1 payment: } 0.45 \times 0.75 \times 1000 = 337.5 \text{ ETH}$$

$$\text{Agent 2 payment: } 0.28 \times 0.75 \times 1000 = 210 \text{ ETH}$$

$$\text{Agent 3 payment: } 0.18 \times 0.75 \times 1000 = 135 \text{ ETH}$$

$$\text{Agent 4 payment: } 0.06 \times 0.75 \times 1000 = 45 \text{ ETH}$$

$$\text{Agent 5 payment: } 0.03 \times 0.75 \times 1000 = 22.5 \text{ ETH}$$

$$\text{Validation pool: } 0.20 \times 1000 = 200 \text{ ETH}$$

$$\text{Burn: } 0.05 \times 1000 = 50 \text{ ETH}$$

Reputation changes:

$$\Delta R_1 = 100 \times [(5 \times 0.45) - 1] = 100 \times 1.25 = +125$$

$$\Delta R_2 = 100 \times [(5 \times 0.28) - 1] = 100 \times 0.40 = +40$$

$$\Delta R_3 = 100 \times [(5 \times 0.18) - 1] = 100 \times (-0.10) = -10$$

$$\Delta R_4 = 100 \times [(5 \times 0.06) - 1] = 100 \times (-0.70) = -70$$

$$\Delta R_5 = 100 \times [(5 \times 0.03) - 1] = 100 \times (-0.85) = -85$$

Agents 1 and 2 (above average) gain reputation. Agents 3, 4, 5 (below average) lose reputation.

E. Validator Incentive Alignment

The Challenge:

Why would validators vote honestly rather than strategically manipulate rankings? This extends the incentive analysis from Calcaterra, Kaal, and Andrei (2018, 7-9) from binary voting to ranked preferences.

Solution via Reputation Staking:

Validators stake reputation V_ℓ to participate. Their reputation changes based on alignment with consensus.

Definition 9 - Validator Agreement Score:

For validator ℓ with ranking π_ℓ and consensus ranking $\hat{\pi}$:

$$\text{Agreement}_\ell = 1 - \frac{d_K(\pi_\ell, \hat{\pi})}{d_{\max}}$$

Where:

- $d_K(\pi_\ell, \hat{\pi})$ = Kendall's tau distance between rankings
- $d_{\max} = k(k-1)/2$ = maximum possible Kendall distance

Kendall's tau distance counts the number of pairwise disagreements:

$$d_K(\pi_\ell, \hat{\pi}) = |\{(i,j) : \pi_\ell \text{ ranks } i > j \text{ but } \hat{\pi} \text{ ranks } j > i\}|$$

Reputation change for validators:

$$\Delta R_\ell = V_\ell \times \text{Agreement}_\ell \times \zeta$$

Where ζ is a scaling parameter proportional to job value.

Validators who align with consensus gain reputation. Those who deviate lose it. Thus, preserving the winner-takes-losers'-stakes property from the original framework (Calcaterra, Kaal, and Andrei 2018, 7) while extending it to ranked voting.

Theorem 3 - Truth-Telling Equilibrium:

If validators believe the plurality (by stake weight) will vote honestly, and reputation value exceeds immediate payment ($\zeta > \beta F / \sum V_\ell$), then truth-telling is a Bayesian Nash equilibrium.

Proof:

Consider validator ℓ deciding between:

- Honest ranking π^*_ℓ (their true assessment)
- Strategic ranking π'_ℓ (attempting to manipulate outcome)

Honest payoff:

$$U_H = \frac{V_\ell}{\sum_j V_j} \times \beta F + V_\ell \times E[\text{Agreement}_\ell | \pi^*_\ell] \times \zeta$$

Strategic payoff:

$$U_S = \frac{V_\ell}{\sum_j V_j} \times \beta F + V_\ell \times E[\text{Agreement}_\ell | \pi'_\ell] \times \zeta$$

Payment component is identical (validators get βF regardless of how they vote).

Reputation component differs.

If plurality votes honestly, consensus ranking $\hat{\pi}$ will approximate the true quality ranking.

Then:

- $E[\text{Agreement}_\ell | \pi^*_\ell] \approx 1$ (honest vote aligns with honest consensus)

- $E[\text{Agreement}_\ell|\pi'_\ell] < 1$ (strategic vote misaligns)

Therefore:

$$U_H - U_S \approx V_\ell \times \zeta \times [1 - E[\text{Agreement}_\ell|\pi'_\ell]]$$

For any $\pi'_\ell \neq \pi^*_\ell$, we have $E[\text{Agreement}_\ell|\pi'_\ell] < 1$, thus $U_H > U_S$.

Honest voting strictly dominates strategic voting when reputation gain matters.

The condition $\zeta > \beta F / \sum V_\ell$ ensures reputation component dominates payment component.

QED.

This theorem establishes a critical result: the revised framework preserves the incentive compatibility from Calcaterra, Kaal, and Andrei (2018) while extending it to richer information environments. As with my broader work on dynamic regulation (Kaal 2016), the key insight is that properly designed incentive structures can elicit truthful behavior without centralized enforcement.

Setting ζ:

Recommended:

$$\zeta = 2 \times \frac{\beta F}{\sum V_\ell}$$

This makes reputation value double the payment value, strongly incentivizing honest voting.

F. Attack Resistance Under Multi-Agent Framework

Question: How does multi-agent competition with citation weighting affect security against malicious actors? Does it preserve or improve upon the $2R_0$ security bound established in Calcaterra, Kaal, and Andrei (2018, 15-18)?

Scenario: Malicious actor controls s of k job agents (Sybil attack). Can they gain disproportionate reputation?

Analysis:

The citation matrix becomes block-structured:

$$C = \begin{pmatrix} C_{AA} & C_{AH} \\ C_{HA} & C_{HH} \end{pmatrix}$$

Where:

- A = attacker-controlled agents
- H = honest agents
- C_{AA} = citations among attacker agents
- C_{AH} = citations from attackers to honest agents
- C_{HA} = citations from honest agents to attackers
- C_{HH} = citations among honest agents

If attackers cite each other exclusively ($C_{AH} = 0, C_{HA} = 0$), the graph disconnects.

Value vector decomposes:

$$v_A = (I - \alpha C_{AA}^T)^{-1}(1-\alpha)q_A$$

$$v_H = (I - \alpha C_{HH}^T)^{-1}(1-\alpha)q_H$$

For attackers to gain significant reward, they need q_A (quality scores from validators) to be high. But validators rank based on actual quality. If honest agents produce better work, validators will rank them higher.

Quality competition effect:

If honest agents have quality distribution $Q_H \sim N(\mu_H, \sigma^2)$ and attackers have quality $Q_A \sim N(\mu_A, \sigma^2)$ with $\mu_A < \mu_H$:

Probability that highest-ranked agent is an attacker:

$$P(\max Q_A > \max Q_H) \approx \Phi\left(\frac{\mu_A - \mu_H + \sigma\sqrt{2\ln s} - \sigma\sqrt{2\ln(k-s)}}{\sigma\sqrt{2}}\right)$$

Where Φ is the standard normal CDF.

This probability decreases exponentially as $(k-s)$ grows. More honest competition → attackers increasingly likely to rank poorly → reputation loss rather than gain.

Revised corruption cost:

Under the original single-agent Calcaterra-Kaal-Andrei framework, cost to achieve 51% control was $2R_0$ (Calcaterra, Kaal, and Andrei 2018, 17).

Under multi-agent model with k agents per job and slashing for poor performance:

Expected reputation gain per job for malicious actor controlling s agents:

$$E[\Delta R_M] \approx \frac{s}{k} \times \alpha F - S_M \times \left(1 - \frac{s}{k}\right)$$

Where:

- First term: fraction of reputation from jobs where attacker wins
- Second term: reputation lost from slashing when ranking below average

For attacker to reach $f \times R_0$ reputation:

$$F^* = \frac{f R_0 k}{s\alpha - k\lambda(1 - s/k)}$$

Where λ is average slashing rate.

Numerical example:

$k = 5$ agents, $s = 3$ controlled by attacker, $\alpha = 1$, $\lambda = 0.5$, $f = 0.51$ (seeking 51% control):

$$F^* = \frac{0.51 \times R_0 \times 5}{3 \times 1 - 5 \times 0.5 \times (1 - 3/5)}$$

$$= \frac{2.55 R_0}{3 - 5 \times 0.5 \times 0.4}$$

$$= \frac{2.55R_0}{3-1}$$

$$= 1.275R_0$$

This initially appears cheaper than the original framework's $2R_0$! But this calculation assumes:

1. Attacker can consistently win $s/k = 60\%$ of competitions
2. No detection of Sybil pattern (same entity controlling multiple agents)
3. Validators don't penalize obvious collusion

In practice, honest agents would:

- Detect citation rings (attacker agents citing each other)
- Downrank submissions that appear copied/colluding
- Implement reputation penalties for detected Sybil behavior

This reflects a broader principle from my DAO governance work:[53] decentralized systems must combine formal mechanisms with emergent social enforcement.

More realistic bound:

With validators actively penalizing detected collusion, effective quality for colluding agents drops by factor δ:

$$\mu_{A,\text{effective}} = \delta \times \mu_A \quad \text{where } \delta < 1$$

Then probability of winning drops, and the cost increases to:

[53] Kaal 2021; 2026.

$$F^* \approx \frac{2R_0}{\delta}$$

For $\delta = 0.5$ (50% quality penalty for detected collusion), cost doubles to $4R_0$—twice as secure as the original Calcaterra-Kaal-Andrei framework.

This improved security stems from multi-agent competition creating multiple attack surfaces that must all succeed simultaneously, combined with citation transparency enabling collusion detection.

G. Long-Run Knowledge Graph Dynamics

The most powerful property emerges over time, addressing the citation paradox that plagued the original framework.

Definition 10 - Historical Contribution Value:

Agent i's total accumulated value from all historical contributions up to time T:

$$V_{i,\text{total}}(T) = \sum_{t=1}^{T} \sum_{j:i \in C_j(t)} c_{ji}(t) \times v_j(t) \times e^{-\delta(T-t)}$$

Where:

- $C_j(t)$ = citation set of agent j at time t (agents j cited)
- $c_{ji}(t)$ = fraction agent j attributed to agent i
- $v_j(t)$ = agent j's value at time t
- δ = temporal discount rate (recent work valued more than old)

This captures compounding: Alice's foundational work at $t = 1$ receives credit at every future time $t' > 1$ when someone cites work that cited Alice.

Theorem 4 - Knowledge Compounds Exponentially:

In a citation-weighted system with n active agents and average citation rate ρ per job, expected long-run value of foundational contribution grows as $O(e^{\rho t})$ for high-quality foundational work.

Proof sketch:

Model citations as a branching process. Foundational contribution I_0 at time 0 gets directly cited by $X_1 \sim \text{Poisson}(\rho)$ works at time 1.

Each of those works gets cited by $X_2 \sim \text{Poisson}(\rho)$ works at time 2, etc.

Expected number of works at generation g citing I_0:

$$E[N_g] = \rho^g$$

With citation weight α, value flowing back to I_0 at generation g:

$$V_g \approx \alpha^g \times \rho^g = (\alpha\rho)^g$$

Total accumulated value:

$$V_{\text{total}} = \sum_{g=0}^{\infty} (\alpha\rho)^g = \frac{1}{1 - \alpha\rho} \quad \text{for } \alpha\rho < 1$$

For $\alpha\rho \geq 1$, series diverges—infinite value to foundational work!

In practice, system saturates (ρ decreases over time as field matures), but the superlinear growth persists during growth phase.

For high-quality foundational work that becomes a standard reference (like seminal papers in a field), effective ρ grows over time as the field expands, leading to:

$$V(t) \sim e^{\rho_0 t}$$

Until field saturation.

QED.

Empirical prediction:

In mature knowledge domains using citation-weighted reputation:
- Foundational contributors accumulate 10-100× more reputation than equivalent-quality incremental contributors
- Citation graphs exhibit power-law degree distribution (Barabási-Albert model)
- Reputation Gini coefficient > 0.6 (high concentration) despite low barriers to entry

This is socially optimal because it correctly prices the nonrivalrous, increasing-returns nature of foundational knowledge. A principle central to my work on innovation-enabling regulation.[54]

[54] Kaal 2016.

V. Implementation Considerations and Parameter Selection

A. Recommended Parameter Values

Based on game-theoretic analysis, simulation, and insights from DAO governance practice:[55]

Competition size k:

- Simple/routine tasks: $k = 3$ (minimal diversity premium, efficiency matters)
- Standard knowledge work: $k = 5$ (balanced diversity and cost)
- Complex research/innovation: $k = 7$ (quality premium justifies 40% higher cost)

The selection of k reflects a fundamental tradeoff in decentralized governance: larger groups improve decision quality but increase coordination costs. The recommended values balance these concerns.

Citation weight α:

- Execution-focused domains: $\alpha = 0.6$ (weight direct quality more than attribution)
- Cumulative knowledge domains: $\alpha = 0.75$ (strong compounding for foundational work)
- Theoretical research: $\alpha = 0.85$ (maximum weight on foundational citations)

Reward allocation:

$$\text{Performance rewards: } 1 - \beta - \gamma = 0.75 \ (75\% \text{ to job performers})$$

[55] Kaal and Calcaterra 2017.

$$\text{Validation rewards: } \beta = 0.20 \text{ (20\% to validators)}$$

$$\text{Burn: } \gamma = 0.05 \text{ (5\% deflation)}$$

This preserves the salary distribution principle from Calcaterra, Kaal, and Andrei (2018, 10) while adding modest deflation to ensure long-run token value.

Validator reputation multiplier:

$$\zeta = 2 \times \frac{\beta F}{\sum V_\ell}$$

(Ensures reputation value = 2× payment value)

This satisfies the truth-telling equilibrium condition (Theorem 3) with comfortable margin.

B. Computational Complexity

PageRank calculation:

Solving $v = (I - \alpha C^T)^{-1}(1 - \alpha)q$

Method 1: Direct matrix inversion

Complexity: $O(k^3)$

For $k = 10$: ~1000 operations

For $k = 20$: ~8000 operations

Feasible for small k, but scales poorly.

Method 2: Power iteration

Start with $v^{(0)} = q$, iterate:

$$v^{(t+1)} = (1 - \alpha)q + \alpha C^T v^{(t)}$$

Converges when $\|v^{(t+1)} - v^{(t)}\| < \epsilon$

Complexity: $O(t \times k^2)$ where $t \approx \log(1/\epsilon)/\log(1/\alpha)$

For $\alpha = 0.7$, $\epsilon = 0.001$: $t \approx 20$ iterations
For $k = 10$: ~2000 operations
For $k = 20$: ~8000 operations

Much better scaling. Converges quickly because $1 - \alpha$ provides strong regularization.

Method 3: Sparse matrix optimization

Citation matrices are typically sparse (each agent cites 2-4 others, not all k agents).

Sparse matrix multiplication:

Complexity: $O(t \times e)$ where e = number of edges in citation graph

For $k = 10$, average 3 citations per agent: $e \approx 30$

Cost: ~600 operations

Two orders of magnitude improvement over dense methods.

On-chain vs. off-chain computation:

Smart contracts on Ethereum or similar platforms cost approximately 0.01 per 100 gas.

Simple arithmetic operations: 3–10 gas.

Matrix operations: 100–1000 gas.

For $k = 10$ sparse PageRank: ~100 operations $\times 5$ gas = 500 gas $\approx \$0.05$

For $k = 20$: ~400 operations $\times 5$ gas = 2000 gas $\approx \$0.20$

Conclusion: Fully on-chain computation is economically feasible for $k \leq 20$—far more tractable than the gas costs for the validation pool mechanism in Calcaterra, Kaal, and Andrei (2018).

For larger k or gas-sensitive applications, compute off-chain with zero-knowledge proof posted on-chain for verification. A pattern increasingly common in DAO governance.[56]

C. Migration Path for Existing Systems

Systems built on the Calcaterra-Kaal-Andrei framework can migrate incrementally:

[56] Kaal 2024.

Phase 1: Citation tracking (weeks 1-4)

- Add optional citation fields to job submissions
- No change to rewards yet—still binary validation
- Build historical citation graph database

Phase 2: Multi-agent opt-in (weeks 5-8)

- Jobs can specify $k > 1$ for multi-agent competition
- Traditional $k = 1$ jobs remain valid
- A/B test quality improvements

Phase 3: Ranked validation (weeks 9-12)

- Validators provide rankings instead of binary votes
- Compute weighted Borda scores
- Still use binary accept/reject for backward compatibility

Phase 4: Citation-weighted rewards (weeks 13-16)

- Activate PageRank calculation with $\alpha = 0.3$ initially
- Gradually increase α by 0.1 every 2 weeks
- Target: $\alpha = 0.7$ by week 24

Phase 5: Full migration (week 17+)

- All new jobs use citation-weighted multi-agent framework
- Legacy binary jobs phased out over 6 months
- Historical citations retroactively weighted

This preserves backward compatibility while enabling evolution—reflecting the dynamic regulation principles from my earlier work.[57]

VI. Empirical Predictions and Testable Hypotheses

The framework generates falsifiable predictions that can validate or refute these theoretical advances:

Hypothesis 1 - Quality improvement from competition:

Expertise domains using $k = 5$ multi-agent competition will exhibit 15-30% higher quality scores compared to $k = 1$ single-agent selection, controlling for agent capability.

Test: Run parallel domains with identical expertise requirements, randomly assign jobs to $k = 1$ vs $k = 5$, measure quality via blind expert review.

Expected result: $E[\text{Quality}_{k=5}]/E[\text{Quality}_{k=1}] \in [1.15, 1.30]$

Hypothesis 2 - Reputation concentration in foundational work:

Knowledge domains with citation weight $\alpha > 0.7$ will exhibit Gini coefficient > 0.6 for reputation distribution, while maintaining low barriers to entry. New agents can achieve 10% of median reputation within 20 jobs.

Test: Measure reputation distribution and new-agent growth curves across domains with varying α.

[57] Kaal 2014; 2016.

Expected result: Strong correlation between α and Gini ($r > 0.7$), but no correlation between α and new-agent growth time.

This prediction reflects my broader research on DAO governance: efficient decentralized systems concentrate decision power among proven contributors while maintaining permissionless entry.[58]

Hypothesis 3 - Dispute reduction from multi-agent validation:

Multi-agent competition with $k \geq 5$ will reduce disputed outcomes by $> 40\%$ compared to $k = 1$, as validator consensus improves with diverse submission sets.

Test: Track dispute rates (jobs where validators strongly disagree, measured by variance in rankings) across k values.

Expected result: $\text{Dispute\_rate}(k = 5) < 0.6 \times \text{Dispute\_rate}(k = 1)$

Hypothesis 4 - Nonlinear attack resistance:

Time-to-corruption (cost to achieve 51% control) scales superlinearly with accumulated transaction volume V. For citation-weighted systems: $T_{\text{corrupt}} = \Omega(V^{1.3})$. For binary systems: $T_{\text{corrupt}} = O(V^{1.0})$.

Test: Simulated attacks on testnet with varying historical volumes, measure cost to reach 51%.

Expected result: $\log(T_{\text{corrupt}})$ vs $\log(V)$ slope > 1.2 for citation-weighted, ≈ 1.0 for binary.

[58] Kaal 2026.

This extends the security analysis from Calcaterra, Kaal, and Andrei (2018, 15-18) to demonstrate that citation-weighted systems become more secure over time at a faster rate than binary systems.

Hypothesis 5 - Knowledge graph emergence:

Citation graphs in domains with $\alpha > 0.7$ will exhibit scale-free properties (power-law degree distribution with exponent $\gamma \in [2, 3]$) and small-world properties (average path length $L \sim \log(n)$).

Test: Analyze citation graph topology over time, fit to power-law and measure path lengths.

Expected result: Kolmogorov-Smirnov test confirms power-law ($p < 0.05$), path length $L < 3\log_2(n)$.

VII. Conclusion and Future Directions

A. Summary of Contributions

This paper has presented a comprehensive mathematical framework for decentralized reputation systems that addresses three critical failings identified in the Calcaterra-Kaal-Andrei (2018) architecture:

1. The citation paradox - The original framework included a citation graph concept[59] but provided no game-theoretic foundation or convergence guarantees. Our citation-weighted

[59] Calcaterra, Kaal, and Andrei 2018, 40-42.

PageRank mechanism creates compounding value for foundational work while maintaining computational tractability through sparse matrix optimization and proven convergence properties (Theorem 2).

2. The quality-efficiency paradox - The original single-human not agent selection protocol[60] sacrificed quality for apparent efficiency. Our multi-agent competitive collaboration protocol captures diversity dividends worth 0.5+ standard deviations of quality improvement while enabling attribution through citation graphs.

3. Binary information destruction - The original binary upvote/downvote[61] collapsed multidimensional quality into one bit. Our weighted Borda consensus ranking preserves rich quality signals and enables nuanced attribution.

The unified framework achieves six critical properties:

1. Strategic manipulation resistance: Cost to corrupt $\geq 2R_0$, improving to $4R_0$ with collusion detection
2. Nuanced quality capture: Ranked validation preserves multidimensional signals
3. Correct attribution: PageRank-style value flows through citation graphs
4. Knowledge sharing incentives: Truth-telling equilibrium (Theorem 3)
5. Autonomous operation: No centralized adjudication required
6. Computational tractability: $O(t \times e)$ sparse computation feasible on-chain

Importantly, these advances preserve the core security properties established in Calcaterra, Kaal, and Andrei (2018), particularly the $2R_0$ corruption cost lower bound, while extending them to richer information environments.

[60] ibid., 11.
[61] ibid., 7-9.

B. Implications for AI Agent Ecosystems

For emerging AI agent marketplaces,[62] these properties are existential requirements rather than optional features. As I have argued in my work on AI-DAO convergence,[63] the governance challenges facing traditional DAOs[64] become exponentially more complex when autonomous AI agents participate in organizational decision-making.

Foundational model contributions: When Agent A fine-tunes a base model creating capabilities that Agent B then specializes, current platforms provide no mechanism for A to receive ongoing credit. Citation-weighted reputation solves this, enabling sustainable open development.

Collaborative problem-solving: Multi-agent competition with citation enables emergent specialization. Agent C becomes the "data cleaning specialist" by consistently contributing that component to others' solutions. Thus, building citation-weighted reputation specifically in that sub-domain.

Adversarial robustness: AI agents can generate unlimited Sybil identities at near-zero cost. Multi-agent validation with quality-based slashing creates economic penalties that scale with the sophistication required to produce competitive-quality outputs. This is precisely the barrier Sybil attacks lack.

Knowledge compounding: As AI capabilities improve, agent-generated knowledge will increasingly build on prior agent-generated knowledge. Systems without citation graphs will fail to correctly price foundational contributions, leading to market failure through underproduction of foundational work.

[62] Kaal 2026.
[63] Kaal 2026.
[64] Kaal 2021; Kaal and Calcaterra 2017.

This reflects a broader principle: innovation-enabling frameworks must reward infrastructure provision, not just final products.

C. Future Research Directions

Several extensions warrant rigorous investigation:

Multi-hop citation attribution: Current framework uses single-step PageRank. Multi-hop methods (SimRank, PathRank) may better capture long chains of cumulative knowledge. Research question: What is the optimal citation horizon for different knowledge domains?

Dynamic expertise emergence: Expertise categories are currently static. Citation clustering could enable automatic emergence of new specializations. Challenge: Prevent fragmentation into trivially narrow categories for gaming purposes. This connects to my work on dynamic regulation.[65] Governance categories must adapt to changing realities.

Cross-expertise citation: How should citations across expertise boundaries be weighted? A legal analysis citing a machine learning model involves different contribution types than intra-domain citations. A mathematical framework is needed for heterogeneous attribution.

Privacy-preserving citations: Zero-knowledge proofs could enable agents to prove "I contributed $X\%$ to this work" without revealing the actual contribution or the identities involved. Research question: Can we achieve citation-weighted attribution under full anonymity?

Temporal dynamics and reputation decay: The current model uses simple exponential discounting. More sophisticated models (hyperbolic discounting, domain-specific decay rates) may better match actual knowledge depreciation. Empirical question: How should reputation half-lives vary across domains?

[65] Kaal 2014; 2016.

Mechanism design for validator quality: We proved truth-telling is an equilibrium (Theorem 3), but what mechanisms encourage high-quality validators to participate? Possible solution: reputation earned through validating correlates with outcome quality of validated work, creating validator performance tracking.

Scalability to millions of agents: Current $O(k^3)$ or $O(t \times e)$ complexity works for $k < 100$. For truly massive systems (millions of agents, billions of jobs), what approximation methods preserve essential properties while achieving $O(\log n)$ scaling? This parallels challenges in DAO governance at scale.

D. Closing Perspective

The Calcaterra-Kaal-Andrei framework (2018) represented a watershed moment in decentralized reputation systems. It is the first rigorous proof that autonomous, attack-resistant reputation could function without centralized control. The mathematical foundations laid in that work remain sound.

However, as decentralized systems mature, particularly with AI agent participation, we must evolve beyond binary validation toward richer attribution mechanisms. The framework presented here accomplishes this evolution while preserving the core security properties of the original design.

This trajectory reflects a broader pattern in my research on blockchain governance and dynamic regulation:[66] decentralized systems must combine formal mechanism design with emergent adaptive capacity. Binary validation was a necessary first step, proof that decentralized quality assessment could function at all. But it is not the final form.

[66] Kaal 2014; 2016; Kaal and Calcaterra 2017; Kaal 2026.

Just as markets evolved from barter to complex financial instruments, reputation systems must evolve from binary validation to citation-weighted knowledge graphs. The question is not whether this evolution will occur, but how quickly we can build the mathematical and engineering foundations to make it stable, secure, and aligned with genuine knowledge production.

For AI agent ecosystems, the urgency is acute. Agents that cannot correctly attribute collaborative contributions cannot sustain open knowledge sharing. Agents that cannot build reputation through cumulative impact cannot develop specialized expertise. Agents that cannot resist Sybil attacks cannot operate in trustless environments.

The mathematics presented here chart a path forward. Implementation, empirical validation, and iterative refinement remain. But the theoretical foundations are sound, building rigorously on the pioneering work of Calcaterra, Kaal, and Andrei (2018) while addressing the limitations that six years of DAO governance experience have revealed.

References

Barabási, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 286 (5439): 509–512. https://doi.org/10.1126/science.286.5439.509.

Benkler, Yochai. 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven: Yale University Press. https://www.benkler.org/Benkler_Wealth_Of_Networks.pdf (open access version).

Bornmann, Lutz, and Hans-Dieter Daniel. 2008. "What Do Citation Counts Measure? A Review of Studies on Citing Behavior." *Journal of Documentation* 64 (1): 45–80. https://doi.org/10.1108/00220410810844150.

Buterin, Vitalik. 2014. "Ethereum: A Next-Generation Smart Contract and Decentralized Application Platform." Ethereum White Paper. https://ethereum.org/en/whitepaper/.

Buterin, Vitalik, and Virgil Griffith. 2017. "Casper the Friendly Finality Gadget." arXiv preprint arXiv:1710.09437. https://arxiv.org/abs/1710.09437.

Calcaterra, Craig, Wulf A. Kaal, and Vlad Andrei. 2018. "Blockchain Infrastructure for Measuring Domain Specific Reputation in Autonomous Decentralized and Anonymous Systems." University of St. Thomas Legal Studies Research Paper No. 18-18. https://ssrn.com/abstract=3125822.

Douceur, John R. 2002. "The Sybil Attack." In *Peer-to-Peer Systems: First International Workshop, IPTPS 2002*, edited by Peter Druschel, Frans Kaashoek, and Antony Rowstron, 251–260. Berlin: Springer. https://doi.org/10.1007/3-540-45748-8_24. (Also available at: https://www.microsoft.com/en-us/research/publication/the-sybil-attack/).

Friedman, Eric J., and Paul Resnick. 2004. "The Social Cost of Cheap Pseudonyms." *Journal of Economics & Management Strategy* 10 (2): 173–199. https://doi.org/10.1111/j.1430-9134.2001.00173.x.

Garfield, Eugene. 1955. "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas." *Science* 122 (3159): 108–111. https://doi.org/10.1126/science.122.3159.108.

Goldin, Ian, Mike Koss, and Jeremy Miller. 2017. "Token Curated Registries 1.0." Medium (blog), March 27. https://medium.com/@ilovebagels/token-curated-registries-1-0-61a232f8dac7.

Hanson, Robin. 2003. "Combinatorial Information Market Design." *Information Systems Frontiers* 5 (1): 107–119. https://doi.org/10.1023/A:1022058209073.

Heller, Michael A. 1998. "The Tragedy of the Anticommons: Property in the Transition from Marx to Markets." *Harvard Law Review* 111 (3): 621–688. https://doi.org/10.2307/1342203.

Hirsch, Jorge E. 2005. "An Index to Quantify an Individual's Scientific Research Output." *Proceedings of the National Academy of Sciences* 102 (46): 16569–16572. https://doi.org/10.1073/pnas.0507655102.

Kaal, Wulf A. 2014. "Evolution of Law: Dynamic Regulation in a New Institutional Economics Framework." In *Festschrift zu Ehren von Christian Kirchner: Recht im ökonomischen Kontext*, edited by Wulf A. Kaal, Matthias Schmidt, and Andreas Schwartze, 1211–1230. Tübingen: Mohr Siebeck. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2267560

Kaal, Wulf A. 2016. "Dynamic Regulation for Innovation." In *Research Handbook on Electronic Commerce Law*, edited by John A. Rothchild, 221–243. Cheltenham: Edward Elgar. (preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2831040).

Kaal, Wulf A., and Craig Calcaterra. 2017. "Crypto Transaction Dispute Resolution." *Business Lawyer* 73 (1): 109–152.

Kaal, Wulf A. 2021. "Blockchain-Based Corporate Governance." *Stanford Journal of Blockchain Law & Policy* 4 (1): 1–20. https://stanford-jblp.pubpub.org/pub/blockchain-corporate-governance.

Kaal, Wulf A. 2025. "AI Governance Via Web3 Reputation System." *Stanford Journal of Blockchain Law & Policy* 8 (1). https://stanford-jblp.pubpub.org/pub/aigov-via-web3.

Kaal, Wulf A. (forthcoming 2026). "DAOs in the Agentic Age" (forthcoming).

Nakamoto, Satoshi. 2008. "Bitcoin: A Peer-to-Peer Electronic Cash System." Bitcoin.org. https://bitcoin.org/bitcoin.pdf.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. "The PageRank Citation Ranking: Bringing Order to the Web." Stanford InfoLab Technical Report SIDL-WP-1999-0120. http://ilpubs.stanford.edu:8090/422/.

Peterson, Jack, and Joseph Krug. 2015. "Augur: A Decentralized Oracle and Prediction Market Platform." Augur White Paper. https://www.augur.net/whitepaper.pdf.; https://arxiv.org/abs/1501.01042

Raymond, Eric S. 1999. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol, CA: O'Reilly Media. https://www.oreilly.com/library/view/the-cathedral/0596001088/

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy* 98 (5, pt. 2): S71–S102. https://doi.org/10.1086/261725.

Shapiro, Carl, and Hal R. Varian. 1998. *Information Rules: A Strategic Guide to the Network Economy*. Boston: Harvard Business School Press.