Citation Honesty Mechanisms in Weighted Directed Acyclic Graph Governance:

Incentive Alignment for Knowledge Attribution in Decentralized Reputation Systems

Wulf A. Kaal, Ph.D.[1]
Professor of Law
University of St. Thomas School of Law
Minneapolis, Minnesota

February 2026
Draft Version 0.04

Abstract

This paper identifies and resolves a critical architectural flaw in citation-weighted reputation systems operating within Weighted Directed Acyclic Graph (WDAG) governance structures for Decentralized Autonomous Organizations (DAOs). Under existing formulations, rational agents face a direct financial disincentive to cite prior contributions, as citation-weighted value allocation through PageRank-derived mechanisms transfers economic reward from the citing agent to the cited agent. This under-citation incentive undermines the foundational promise of knowledge attribution in decentralized collaborative systems.

The paper formalizes the under-citation problem, demonstrates the impossibility of precise citation quantification, and identifies a gap in existing game-theoretic proofs that purport to establish citation honesty as an equilibrium. Drawing on the WDAG governance framework developed in Calcaterra (2018), this paper proposes an integrated solution that decouples citation accuracy from quality-based payment, introduces validator-assessed citation honesty as a separately ranked and rewarded dimension, and establishes retroactive audit mechanisms with graduated sanctions. A new equilibrium theorem (Theorem 3b) demonstrates that under the

[1] The author is extremely grateful for the ever evolving theoretical discussions with Craig Calcaterra and Jonathan Kung and for theoretical AI and law WDAG and symbolic AI discussions with Morgan Grey. The author is also very grateful for excellent research assistance by Mickey Bernardi.

integrated solution, honest citation constitutes a Nash equilibrium when validators assess citation accuracy and the marginal effect of citation accuracy on validator-assessed honesty scores exceeds the marginal gain from self-citation inflation. The paper situates these findings within the broader framework of dynamic regulation theory and New Institutional Economics incomplete contract theory, arguing that static citation protocols are inherently vulnerable to gaming and that evolutionary governance mechanisms are essential for sustaining honest knowledge attribution in decentralized systems.

Keywords: Decentralized Autonomous Organizations, WDAG Governance, Citation Mechanisms, Reputation Systems, Game Theory, Incentive Alignment, Dynamic Regulation, New Institutional Economics, Blockchain Governance, Knowledge Attribution

JEL Categories: D02, D82, D86, C72, D85, L14, D83, O33, P00

I. Introduction

The emergence of decentralized reputation systems operating on blockchain infrastructure has introduced novel possibilities for organizing collaborative knowledge production without centralized intermediaries.[1] In such systems, the accurate attribution of intellectual contributions—who built upon whose prior work, and to what degree—is not merely an academic courtesy but a core mechanism for distributing economic rewards and governance power. When reputation tokens are minted and allocated based on the perceived value of contributions, and when the perceived value of contributions is itself a function of how subsequent work references prior work, the integrity of the citation mechanism becomes a load-bearing structural element of the entire governance architecture.

Calcaterra (2018) introduced a comprehensive framework for on-chain governance of DAOs, in which a Weighted Directed Acyclic Graph (WDAG) structure records contributions as vertices and citations as weighted directed edges.[2] The WDAG framework provides a powerful mechanism for revaluing past contributions through future references, enabling an evolutionary governance system that can reward long-term value creation and punish corruption through methodical review. The post valuation algorithm assigns each post a total value that incorporates both its initial validation pool outcome and the accumulated weight of all subsequent references, creating a dynamic and reviewable reputation economy.

More recent work on domain-specific reputation systems has extended this framework by incorporating PageRank-derived citation weighting into multi-agent collaborative environments.[3] Under such extensions, when multiple agents contribute to a task, a citation matrix records each agent's attribution of value to prior contributors, and a modified PageRank calculation distributes reputation and payment in proportion to citation-weighted influence. The appeal of this approach is considerable: it promises to create a self-organizing system of knowledge attribution that mirrors the citation dynamics of academic publishing while embedding those dynamics directly into economic incentive structures.

This paper demonstrates, however, that citation-weighted payment mechanisms as currently formulated contain a fundamental incentive misalignment that, if left unaddressed, will systematically erode the quality of knowledge attribution in any decentralized reputation system.

The problem is straightforward: under a PageRank-derived value distribution, any value that agent *i* attributes to agent *j* through citation weight c_{x}□ flows *away* from agent *i*'s share of the payment pool. A rational, self-interested agent therefore has a direct incentive to minimize citations to others and maximize self-attribution—precisely the opposite of honest knowledge attribution.

The contribution of this paper is threefold. First, it formalizes the under-citation incentive problem and demonstrates that it constitutes a structural flaw rather than a mere parameter calibration issue (§II). Second, it identifies a gap in existing game-theoretic analyses that claim to establish citation honesty as an equilibrium, showing that proofs of validator truth-telling in quality ranking do not extend to agent honesty in citation allocation (§III). Third, drawing on the WDAG governance parameters and evolutionary protocol development framework from Calcaterra (2018), it proposes an integrated solution that treats citation accuracy as a separately ranked and rewarded governance dimension, establishes mandatory citation minimums with validator override, and introduces retroactive audit mechanisms with graduated sanctions (§IV–V). The paper proves that under this integrated mechanism, honest citation constitutes a Nash equilibrium (Theorem 3b, §V.E).

The analysis is situated within the broader framework of dynamic regulation theory and New Institutional Economics (NIE) incomplete contract theory.[4] As the literature on dynamic regulation recognizes, the conditions and corresponding requirements for optimal and stable rules are constantly evolving, making different sets of rules optimal over time.[5] Citation protocols in decentralized reputation systems are no exception. Arrow's Impossibility Theorem, the Folk Theorems of repeated game theory, and the theory of incomplete contracts all counsel that no static citation protocol can eliminate all arbitrage opportunities for sufficiently patient and powerful actors.[6] The integrated solution proposed here is therefore designed from its inception to be evolutionary, embedding citation honesty enforcement within the dynamic governance mechanisms. That is, loosely-coupled and tightly-coupled validation pools, reference-based revaluation, and contentious debate protocols, that Calcaterra (2018) identified as essential for robust DAO governance.

II. The Under-Citation Incentive Problem

A. Formal Statement of the Problem

Consider a collaborative task environment in which n agents contribute to a job, and each contributing agent i submits a citation vector attributing credit to other contributors. The citation weights are collected into a column-stochastic citation matrix C, where entry $c_{x\square}$ represents the fraction of credit that agent i attributes to agent j, subject to the constraint that $\Sigma_{\square}\ c_{x\square} = 1$ for each agent i. A PageRank-derived value vector v is then computed as:

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

where $\alpha \in (0,1)$ is a damping factor, I is the identity matrix, and q is a base quality vector reflecting the initial assessment of each agent's contribution. Agent i's payment is then determined as:

$$Payment_{i} = v_{i} \times (1-\beta-\gamma) \times F$$

where F is the total fee pool, β is the platform fee fraction, and γ is the validator reward fraction. The critical observation is that the PageRank calculation is a zero-sum redistribution mechanism with respect to citation weights: any increase in $c_{x\square}$ (credit attributed by i to j) necessarily decreases v_{i} and increases $v_{\square}$. In concrete terms, if agent i cites agent j with weight $c_{x\square} = 0.6$, then approximately 60% of the citation-weighted value flows away from agent i to agent j.

The rational strategy for any self-interested agent is therefore transparent: set $c_{x\square} = 0$ for all $j \neq i$ (cite no one) to maximize personal payment. Under the standard formulation, the dominant strategy is maximal self-citation, which is the precise negation of honest knowledge attribution.

B. The Quantification Impossibility

Even setting aside strategic incentives, the citation mechanism faces a deeper epistemic challenge. The requirement that agents assign continuous-valued citation weights $c_{x\square} \in [0,1]$ satisfying $\Sigma_{\square}\ c_{x\square} = 1$ presupposes that agents can meaningfully quantify the relative contribution of prior work to their current output. In practice, this quantification is epistemically unknowable. Was a prior collaborator's insight worth 0.37 or 0.62 of the current agent's work? The question has no determinate answer. Knowledge production is not a linear combination of inputs amenable to precise fractional decomposition.

This observation connects directly to the incomplete contract theory framework of New Institutional Economics.[7] NIE recognizes that bounded rationality and incomplete information prevent contracting parties from specifying all contingencies ex ante.[8] The citation matrix C functions as an implicit contract specifying value distribution, yet the precise terms of this contract—the exact citation weights—cannot be determined with the precision the mathematical framework demands. As Kirchner (2011) observed in the context of institutional design, contracts are a form of experimentation by way of observation for the learning process.[9] Citation weights, similarly, should be understood as approximate signals of attribution rather than precise measurements, and the governance system must be designed to function robustly under this inherent imprecision.

C. Connection to WDAG Reference Weights

The under-citation problem is structurally analogous to challenges identified in the WDAG governance framework. In the WDAG model, each post p_i may reference prior posts with weights $w_{i,j}$ chosen between -1 and 1, subject to the constraint $\sum_j |w_{i,j}| \leq 1$.[10] The total value of any post is computed recursively: $t_i := v_i + \sum_j w_{j,i} t_j$, where v_i is the initial value from the validation pool. Under this formulation, references that add value to prior posts do not inherently reduce the referrer's value, unless the leaching parameter q_4 is set above zero.[11]

The WDAG framework's treatment of the leaching parameter q_4 is instructive. When $q_4 = 0$ (default), a referrer can create new reputation value through positive references without diminishing its own value. When $q_4 > 0$, the referrer's value is reduced by a fraction of the value it confers on referenced posts. The citation-weighted payment mechanism in multi-agent settings effectively operates with an implicit $q_4 = 1$: every unit of citation-weighted value conferred on another agent is a unit lost to the citing agent. This is a maximally punitive leaching parameter, and it is unsurprising that it generates maximal disincentives for honest citation.

The insight from the WDAG framework is that leaching parameters can and should be calibrated to balance competing incentives. Thus, rewarding honest attribution while maintaining sufficient skin-in-the-game to prevent frivolous or strategic references. The solution proposed in this paper adapts this insight to the multi-agent citation context.

III. The Missing Game-Theoretic Proof

Existing work on citation-weighted reputation systems has claimed to provide "game-theoretic analysis establishing citation honesty as an equilibrium."[12] Specifically, such claims typically reference a theorem (commonly designated Theorem 3 or its equivalent) establishing that validators have a truth-telling equilibrium for quality ranking. This is a genuine and important result: when validators stake reputation on their rankings and the mechanism rewards agreement with consensus, validators are incentivized to report their honest assessment of quality.

However, this result addresses a fundamentally different question from the one at issue. The validator truth-telling equilibrium establishes that *validators* have incentive-compatible strategies for *ranking quality*. It does not establish that *job-performing agents* have incentive-compatible strategies for *honest citation*. These are distinct strategic actors facing distinct strategic choices within distinct mechanism structures. Conflating the two is a genuine error.

To make the distinction precise: the validator faces a mechanism in which they submit a ranking and receive a reward based on agreement with the consensus ranking. The standard result shows that under appropriate conditions (e.g., proper scoring rules, reputation staking), the validator's dominant strategy is truthful reporting. The job-performing agent, by contrast, faces a mechanism in which they submit citation weights and receive payment based on the resulting PageRank calculation. Under the standard formulation, as demonstrated in §II, the agent's dominant strategy is *dishonest* reporting, maximal self-citation.

This gap is not merely an oversight in presentation. It reflects a structural absence in the mechanism design. The existing framework provides no mechanism for verifying citation accuracy, no separate incentive for honest attribution, and no penalty for strategic under-citation. The framework relies entirely on the PageRank calculation to "correctly" distribute value, but the PageRank calculation is only as honest as its inputs. And the inputs are generated by agents with a direct incentive to distort them.

The Folk Theorems of repeated game theory provide additional cause for concern.[13] In the practical scenario where policing citation accuracy is costly, if a fixed citation protocol is published, there will always be strategies for sufficiently patient players to game the system. As Calcaterra (2018) observed, "it is expensive to police the protocol deviants, so it is more efficient to assume others are following protocol and not perform the expensive policing yourself. This gives a subgame perfect Nash equilibrium leading to a situation where eventually less than 50% of the members actually perform the policing, meaning the system can be gamed."[14] Any solution to the citation honesty problem must therefore incorporate dynamic enforcement mechanisms that can evolve in response to gaming strategies. Precisely the type of evolutionary governance that the WDAG framework was designed to support.

IV. Proposed Solutions: An Integrated Framework

This Part presents four complementary mechanisms for addressing the under-citation problem, then synthesizes them into an integrated solution that leverages the WDAG governance infrastructure. The design philosophy follows the principle articulated in the dynamic regulation literature: rather than seeking a single optimal static rule, the goal is to create an evolutionary system where citation protocols can continually adapt to emerging gaming strategies.[15]

A. Solution 1: Validator-Verified Citations

The most immediate intervention extends the validator's role to include citation verification. Under the existing framework, validators submit quality rankings during validation pools. Under this modification, each validator ℓ would additionally submit a citation adjustment matrix proposing corrections to agent-submitted citations, yielding consensus citations computed as a reputation-weighted average across validators.

This approach maps directly onto the WDAG governance structure. In the Calcaterra (2018) framework, validation pools already serve as the mechanism by which the bench of experts evaluates every contribution.[16] Extending the validation pool to encompass citation verification is a natural evolution of the existing infrastructure. The validator's task becomes: evaluate not only the quality of the work-evidence post but also the accuracy of its references.

The limitation of this approach is that validators may not possess complete knowledge of the citation history relevant to any particular contribution. A validator may be able to identify obviously missing citations or obviously inflated self-citation, but may lack the domain-specific knowledge to assess fine-grained attribution weights. This limitation counsels that validator-verified citations should be one component of a broader solution rather than the sole mechanism.

B. Solution 2: Decoupled Citation Rewards

The structural solution is to decouple citation accuracy from quality-based payment, thereby eliminating the zero-sum dynamic that creates the under-citation incentive. Under this modification, agent *i*'s payment becomes:

$$Payment_i = [\theta v_i + (1-\theta)\eta_i] \times (1-\beta-\gamma) \times F$$

where v_i is the citation-weighted value (as before), η_i is a "citation honesty score" derived from validator assessment of citation accuracy, and $\theta \in [0,1]$ determines the relative weight of quality versus citation honesty in payment. Validators rank both quality and citation accuracy as separate dimensions. An agent who produces medium-quality work but cites meticulously can still earn substantial payment through the η_i component.

This decoupling is analogous to the WDAG framework's separation of hard and soft protocols.[17] Quality assessment corresponds to the hard protocol evaluation of work-evidence posts, while citation accuracy corresponds to the soft protocol evaluation of reference standards. The WDAG framework already contemplates that "software evaluators follow the protocols for referencing the algorithms designed in Stage 1 to fairly reward the creators of the algorithms being used" and that "the standards for choosing reference weights is set in the forum as soft protocols."[18] The decoupled citation reward mechanism formalizes this soft-protocol policing of reference standards and attaches explicit economic consequences.

The reputation change formula is similarly modified to incorporate a citation accuracy term:

$$\Delta R_i = S_i \times [(k \times v_i) - 1 + \lambda(\eta_i - \overline{\eta})]$$

where $\lambda > 0$ rewards above-average citation accuracy and S_i is the agent's stake. This dual-channel incentive structure ensures that citation honesty is not merely a cost to be minimized but a source of both immediate payment and long-term reputation growth.

C. Solution 3: Retroactive Citation Penalties

The WDAG framework's most distinctive feature is the ability to revalue past contributions through future references.[19] This mechanism provides a natural infrastructure for retroactive citation audits. Under this solution, any agent can challenge historical under-citation by submitting evidence that agent *i*'s work at time *t* relied on uncited work from agent *j*. If the challenge is upheld through a validation pool, agent *i* loses reputation:

$$\Delta R_i^{penalty} = -\gamma_{penalty} \times v_i(t) \times R_i(t)$$

where the penalty parameter is set sufficiently high (e.g., 0.5) that the expected value of under-citing is negative: the expected gain from under-citation must be less than the probability of detection multiplied by the penalty magnitude.

This retroactive audit mechanism mirrors the WDAG framework's reference-based revaluation system, under which "review gives users the opportunity to reward or punish past actions, such as positive protocol development or subtle patterns of corruption."[20] The implementation requires a dispute resolution protocol and evidence standards, both of which can be developed through the contentious debate mechanism. Loosely-coupled validation pools transitioning to tightly-coupled votes as consensus emerges.[21] The graduated nature of the sanction connects to Ostrom's sixth principle for managing a commons: "Use graduated sanctions for rule violators."[22]

D. Solution 4: Mandatory Citation Minimums with Validator Override

The fourth mechanism establishes structural constraints on citation allocation combined with validator correction authority. In the first phase, agents must allocate no more than a maximum fraction τ_{max} (e.g., 0.5) to self-citation, ensuring that at least 50% of citation weight flows to other contributors. In the second phase, validators can increase citations to others if they detect under-citation. Critically, validators can only increase outward citations. They cannot decrease them. Thus, preventing collusion to help allies under-cite competitors.

The final citation matrix is computed as:

$$C^*_{ij} = \max(c_{ij}, \max_{\ell} \tilde{c}_{\ell,ij})$$

where $\tilde{c}_{\ell,ij}$ is validator ℓ's proposed citation weight from agent i to agent j. Validators receive a reputation bonus for detecting under-citation that consensus agrees with, creating a positive incentive for vigilant citation policing.

This mechanism connects directly to the WDAG framework's parameter q_2 (limit to revaluation). Just as q_2 constrains the fraction of a referrer's value that can affect referenced posts,[23] the mandatory citation minimum constrains the fraction of citation weight that can be self-allocated. Both parameters serve the same governance function: preventing any single actor from monopolizing the value flow within the system.

V. The Integrated Citation Honesty Mechanism

A. Mechanism Design

The recommended integrated solution combines Solutions 2 and 4 as the primary mechanism, with Solutions 1 and 3 as supporting enforcement layers. The complete mechanism operates as follows. Agents must allocate self-citation weight $c_{ii} \leq 0.5$ (mandatory citation minimum). Validators rank both quality and citation accuracy as separate dimensions during validation pools. Payment is computed using the decoupled formula with $\theta = 0.6$, meaning 60% of payment derives from citation-weighted quality and 40% from citation honesty scores. Reputation change incorporates a citation accuracy bonus: agents with citation accuracy $\eta_i > 0.5$ gain bonus reputation, while those below 0.5 lose reputation relative to the baseline. Retroactive audits remain available with penalties of $-0.5R_i$ if sustained through the dispute resolution protocol.

The specific payment formula under the integrated mechanism is:

$$Payment_i = [0.6 \cdot v_i + 0.4 \cdot \eta_i] \times (1-\beta-\gamma) \times F$$

And the reputation change formula is:

$$\Delta R_i = S_i \times [(k \times v_i) - 1] + S_i \times 2(\eta_i - 0.5)$$

The second term creates a symmetric incentive around the midpoint: agents with citation accuracy above 0.5 gain bonus reputation, while those below 0.5 suffer a reputation penalty. The coefficient of 2 ensures that the citation accuracy component has meaningful magnitude relative to the quality component.

B. Integration with WDAG Governance Parameters

The integrated solution maps onto the WDAG governance parameter space as follows. The mandatory citation minimum corresponds to a constraint on the WDAG reference weight structure: just as the WDAG constrains $\Sigma_j |w_{i,j}| \le q_2$,[24] the citation mechanism constrains $c_{ii} \le \tau_{max}$. Both are structural parameters that limit the concentration of value flow.

The validator assessment of citation accuracy operates through the existing validation pool infrastructure. In the WDAG framework, validation pools already evaluate work-evidence posts using both hard protocols (programmatic checks) and soft protocols (socially agreed-upon evaluation standards).[25] Citation accuracy assessment becomes an additional soft protocol dimension. The transition from loosely-coupled to tightly-coupled voting on citation standards follows the same contentious debate mechanism used for all protocol development: initial loosely-coupled polls gauge consensus, and as standards solidify, tightly-coupled enforcement ensures compliance.[26]

The retroactive audit mechanism leverages the WDAG's reference-based revaluation. A negative reference to a past post citing under-attribution serves as a challenge to historical citation accuracy. If this challenge post is upheld in its validation pool, the referenced post loses value, and by extension, the agent who submitted the under-cited work loses reputation. The chain of references can propagate: if agent *A*'s uncited reliance on agent *B*'s work is exposed, and agent *C*'s later work cited agent *A*'s contribution, the revaluation cascades through the WDAG precisely as the framework intends. This cascading revaluation is one of the WDAG's core strengths: "Turning the Folk Theorems against the attackers, we are also guaranteed the existence of a strategy for sufficiently patient and powerful police, which eliminates the incentive to attack from any fixed malicious strategy."[27]

C. Validation Pool Parameters for Citation Assessment

The following WDAG governance parameters are recommended for citation assessment validation pools. The token loss ratio c_8 should initially be set to 0 (loosely-coupled) for citation standard development, transitioning to 1 (tightly-coupled) once consensus is established. The fraction of tokens staked for the post c_2 should be set below the default of 1/2, encouraging critical analysis of citation accuracy. The validation pool termination period c_7 should be moderate, allowing sufficient time for validators to review citation histories without creating excessive delay. The limit to revaluation q_2 for citation-related references should be moderate, allowing meaningful correction of under-cited posts without enabling destabilizing cascades.

D. Evolutionary Protocol Development

Consistent with the dynamic regulation framework, citation accuracy standards should be treated as an evolving body of soft protocols rather than a fixed set of rules. The initial standards for citation assessment will be developed through contentious debate posts in the forum, subjected to loosely-coupled validation pools. As consensus emerges on what constitutes adequate citation in various contexts, these standards will transition to tightly-coupled enforcement.

This evolutionary approach is essential because, as the NIE literature on incomplete contracts recognizes, the precise parameters of citation accuracy cannot be specified ex ante.[28] What constitutes "adequate" citation will vary across expertise domains, will evolve as domain knowledge accumulates, and will need to adapt as agents develop novel gaming strategies. The WDAG's evolutionary governance structure, in which "rules are able to continually change to avoid threats and seize opportunities (micro soft forks),"[29] provides precisely the institutional infrastructure needed to support this ongoing adaptation.

The six qualities of effective governance identified in Calcaterra (2018)—open, balanced, dynamic, binding, informed, and escapable—apply with equal force to citation governance.[30] Citation standards must be open (any expert can propose improvements), balanced (proper rewards for both work and citation accuracy), dynamic (adaptable to new gaming strategies), binding (enforced through tightly-coupled validation pools once consensus is reached), informed (developed through authentic deliberation), and escapable (agents who disagree with citation standards can fork to a new expertise where different standards apply).

E. Proof of Honest Citation Equilibrium (Theorem 3b)

Theorem 3b (Citation Honesty Equilibrium). *Under the integrated solution with citation accuracy component η_i contributing 40% of payment and reputation bonus for $\eta_i > 0.5$, honest citation is a Nash equilibrium when validators assess citation accuracy and the marginal effect of citation accuracy on the honesty score exceeds the marginal gain from self-citation:*

$$\partial\eta_i/\partial(\text{citation accuracy}) > \partial v_i/\partial c_{ii}$$

Proof. Agent *i* chooses citation allocation $\{c_{i\square}\}_{\square}$ to maximize total utility:

$$U_i = [0.6v_i(c_{i\square}) + 0.4\eta_i(c_{i\square})]F + S_i[kv_i - 1] + 2S_i(\eta_i - 0.5)$$

Taking the derivative with respect to c_{ii} (self-citation):

$$\partial U_i/\partial c_{ii} = 0.6F(\partial v_i/\partial c_{ii}) + 0.4F(\partial\eta_i/\partial c_{ii}) + S_ik(\partial v_i/\partial c_{ii}) + 2S_i(\partial\eta_i/\partial c_{ii})$$

Under the mechanism, validators penalize over-self-citation (under-citing others), so:

$$\partial\eta_i/\partial c_{ii} < 0 \quad \text{and} \quad |\partial\eta_i/\partial c_{ii}| > \partial v_i/\partial c_{ii}$$

The first condition holds because validators rate citation honesty lower when self-citation is inflated. The second condition is the design requirement: the mechanism parameters ($\theta = 0.6$, reputation coefficient = 2) are chosen such that the negative effect of increased self-citation on the honesty score outweighs the positive effect on the value score.

Substituting, the positive terms $(0.6F + S_ik)\partial v_i/\partial c_{ii}$ are dominated by the negative terms $(0.4F + 2S_i)\partial\eta_i/\partial c_{ii}$, yielding $\partial U_i/\partial c_{ii} < 0$. Thus increasing self-citation (under-citing others) strictly decreases utility. The optimal strategy is accurate citation. □

The equilibrium is robust in the sense that it does not depend on agents accurately quantifying precise citation weights. It requires only that agents' citation patterns be *directionally* honest. That is, citation patterns demonstrate that agents do not systematically under-cite contributors whose work they relied upon. The validator assessment mechanism evaluates this directional honesty rather than demanding point estimates of attribution fractions, addressing the quantification impossibility identified in §II.B.

VI. Discussion

A. Distributive Norms and Citation Honesty

Calcaterra (2018) identified five categories of distributive norms relevant to DAO governance: equity, equality, power, need, and responsibility.[31] The citation honesty mechanism interacts with these norms in important ways. Under the equity norm, where members' outcomes are based on their inputs, honest citation is essential because it ensures that the inputs of prior contributors are properly recognized in determining outcomes. A system that permits strategic under-citation effectively transfers value from genuine contributors to free-riders, violating the equity norm.

The power norm presents a potential tension: actors with more authority or status may have greater ability to under-cite without detection, since their established reputation may shield them from scrutiny. The validator override mechanism (Solution 4) and retroactive audit mechanism (Solution 3) are specifically designed to counteract this tendency. The WDAG's quadratic locking mechanism (c_{11} = 1/2) may further mitigate power concentration by ensuring that even high-reputation agents cannot dominate citation assessment validation pools disproportionately.[32]

B. The Commons Management Perspective

Citation accuracy in a decentralized reputation system is a commons resource. The accumulated body of honestly attributed knowledge, the integrity of the citation graph, benefits all participants by ensuring that reputation scores are meaningful signals of expertise and contribution quality. Strategic under-citation is a form of commons degradation: each act of under-citation slightly erodes the informational value of the reputation system for all users.

Ostrom's eight principles for managing a commons, which Calcaterra (2018) mapped to DAO governance,[33] apply directly to citation governance. Clear group boundaries (Principle 1) are defined by expertise-specific reputation tokens. Rules matched to local conditions (Principle 2) are achieved through domain-specific citation standards developed in each expertise's forum. Participation in rule-making (Principle 3) is ensured through the contentious debate mechanism. Community monitoring (Principle 5) is implemented through validator citation assessment and retroactive audits. Graduated sanctions (Principle 6) range from reduced citation honesty scores to retroactive reputation penalties. Accessible dispute resolution (Principle 7) is provided through the validation pool and reference-based revaluation mechanisms.

C. Limitations and Future Work

Several limitations warrant acknowledgment. First, the proof of Theorem 3b assumes that validators can meaningfully assess citation accuracy, which may be challenging in highly specialized domains. Future work should explore automated citation verification tools, such as semantic similarity analysis between work products, that could supplement human validator judgment. Second, the optimal values of mechanism parameters (θ, the reputation coefficient, τ_{max}, penalty magnitudes) are presented as design choices rather than derived from first principles. Empirical calibration through simulation and field experimentation will be necessary to identify parameter values that produce robust equilibria across diverse domain contexts.

Third, the interaction between citation honesty mechanisms and the broader WDAG governance parameter space has not been fully characterized. The leaching parameter q_4, the chain length limit q_3, the reference limits q_5 and q_6, and the token aging function q_7 all interact with citation incentives in ways that deserve systematic analysis. Fourth, the analysis has focused on the Nash equilibrium concept. Coalition-proof equilibrium analysis would address the possibility of coordinated under-citation among groups of colluding agents.

VII. Conclusion

Citation-weighted reputation systems offer a powerful mechanism for knowledge attribution in decentralized collaborative environments. However, the standard formulation contains a critical structural flaw: the PageRank-derived value distribution creates a direct financial disincentive to cite, incentivizing precisely the strategic under-attribution that the mechanism is designed to prevent. Existing game-theoretic proofs of validator truth-telling do not extend to agent citation honesty, leaving a genuine gap in the mechanism design.

The integrated solution proposed in this paper addresses this gap by treating citation accuracy as a separately ranked and rewarded governance dimension, establishing mandatory citation minimums with validator override, and enabling retroactive audit mechanisms with graduated sanctions. Theorem 3b establishes that under this mechanism, honest citation constitutes a Nash equilibrium when the marginal effect of citation accuracy on honesty scores exceeds the marginal gain from self-citation inflation.

The solution is designed to integrate seamlessly with the WDAG governance framework developed in Calcaterra (2018). It leverages existing infrastructure: validation pools, reference-based revaluation, contentious debate mechanisms, and hard and soft protocol development. All are intended to enforce citation honesty through evolutionary governance mechanisms rather than static rules. This approach is consistent with the dynamic regulation literature's recognition that no single static protocol can eliminate all gaming opportunities for sufficiently patient and powerful actors.

As blockchain technology continues to enable new forms of decentralized collaboration, the integrity of knowledge attribution mechanisms will become increasingly critical. The framework developed here provides both a formal diagnosis of the under-citation problem and a principled mechanism design solution, grounded in game theory, institutional economics, and the practical governance infrastructure needed to sustain honest knowledge attribution in decentralized autonomous organizations.

The citation honesty mechanism should be seen as incorporated as Section IV.H in any revised formulation of citation-weighted reputation systems,[34] ensuring that the theoretical elegance of PageRank-derived attribution is matched by institutional mechanisms that make honest citation the rational strategy for every participant.

Notes

1. *See generally* Craig Calcaterra, On-chain Governance of Decentralized Autonomous Organizations: Blockchain Organization using Semada, with an appendix by Wulf Kaal (Draft Version 0.1, May 24, 2018), available at https://ssrn.com/abstract=3188374; Wulf A. Kaal, Evolution of Domain-Specific Reputation Systems: From Binary Validation to Citation-Weighted Knowledge Attribution (2025) (working paper), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6192998.

2. Calcaterra (2018), *supra* note 1, at §2.2.1 (describing the forum as a weighted, directed, acyclic graph where the vertices are posts and the edges are citations directed from referrer to reference).

3. *See, e.g.*, the reputation system framework incorporating PageRank-derived citation weighting for multi-agent collaborative environments, where the citation matrix C records attribution weights and a modified PageRank calculation distributes value.

4. Wulf A. Kaal, Evolution of Law: Dynamic Regulation in a New Institutional Economics Framework, in Festschrift in Honor of Christian Kirchner (2013), available at SSRN: https://ssrn.com/abstract=2267560.

5. Wulf A. Kaal, Dynamic Regulation of the Financial Services Industry (2013), 48 Wake Forest L. Rev. 791–828 2014, available at SSRN: https://ssrn.com/abstract=2273857.

6. Calcaterra (2018), *supra* note 1, at §1.4 (arguing that Arrow's Impossibility Theorem, the Folk Theorems of game theory, and incomplete contract theory demonstrate that no static constitution of rules can eliminate all possible arbitrage opportunities).

7. Eirik G. Furubotn & Rudolf Richter, Institutions and Economic Theory: The Contribution of the New Institutional Economics (University of Michigan Press 2005); Douglass C. North, Institutions, Institutional Change and Economic Performance (Cambridge University Press 1990).

8. Oliver E. Williamson, The Economic Institutions of Capitalism (Free Press 1985); Oliver Hart & John Moore, Foundations of Incomplete Contracts, 66 Rev. Econ. Stud. 115 (1999).

9. Christian Kirchner, Evolution of Law: Interplay Between Private and Public Rule-Making: A New Institutional Economics Analysis, 4 Erasmus L. Rev. 161 (2011).

10. Calcaterra (2018), *supra* note 1, at §2.2.2, Post valuation algorithm, item 1 (defining $w_{n,k}$ as the weight from referrer p_k to reference p_n with constraints $-1 \le w_{n,k} \le 1$ and $\Sigma_n |w_{n,k}| \le 1$).

11. *Id.* at §3.3.4 (defining q_4 as the leaching value parameter, with default $q_4 = 0$, under which "a referrer can gain value from a post it references negatively" and "new posts can leach value from older posts and vice versa").

12. See the claim in existing reputation system frameworks that "we provide game-theoretic analysis establishing citation honesty as an equilibrium (see Section IV.E)," where Section IV.E in fact proves only validator truth-telling equilibrium for quality ranking.

13. See George J. Mailath & Larry Samuelson, Repeated Games and Reputations: Long-run Relationships (Oxford University Press 2006); see also https://en.wikipedia.org/wiki/Folk_theorem_(game_theory).

14. Calcaterra (2018), *supra* note 1, at §1.3 n.25.

15. Kaal, Dynamic Regulation of the Financial Services Industry, *supra* note 5. See also Wulf A. Kaal, Dynamic Regulation for Innovation (2016), in Perspectives in Law, Business & Innovation (Mark Fenwick, Wulf A. Kaal, Toshiyuki Kono & Erik P.M. Vermeulen eds., Springer 2016).

16. Calcaterra (2018), *supra* note 1, at §2.1.2 (describing validation pools where "the community has the opportunity to bet their personal sem tokens on whether or not they believe the post adds value to the company").

17. *Id.* at §3.1 (distinguishing hard protocols as “rules which are programmatically enacted” from soft protocols as “socially agreed upon rules which dictate orthodox behavior for its members”).

18. *Id.* at §4.1, Stage 5.e.

19. *Id.* at §1.6 (“Review gives users the opportunity to reward or punish past actions, such as positive protocol development or subtle patterns of corruption.”).

20. *Id.*

21. *Id.* at §1.5 (describing the contentious debate mechanism with loosely-coupled and tightly-coupled voting).

22. Elinor Ostrom, Governing the Commons: The Evolution of Institutions for Collective Action (Cambridge University Press 1990). See Calcaterra (2018), *supra* note 1, at §4.3 (mapping Ostrom’s principles to DAO governance).

23. Calcaterra (2018), *supra* note 1, at §3.3.2 (defining q_2 as the limit to revaluation, constraining the sum of absolute reference weights to q_2).

24. *Id.*

25. *Id.* at §3.1.

26. *Id.* at §1.5, §3.2.7–3.2.8 (describing transition from loosely-coupled to tightly-coupled voting through parameters c_8, c_9, and c_10).

27. *Id.* at §1.6.

28. Kirchner (2011), *supra* note 9; Hart & Moore (1999), *supra* note 8.

29. Calcaterra (2018), *supra* note 1, at §1.7, quality 3 (Dynamic).

30. *Id.* at §1.7.

31. *Id.* at §4.2 (cataloguing distributive norms of equity, equality, power, need, and responsibility, derived from Donelson R. Forsyth, Group Dynamics (5th ed., Wadsworth Cengage Learning 2009), at 388–89.

32. *Id.* at §3.2.8 (describing quadratic locking as c_11 = 1/2, drawing on the Weyl/Posner/Lalley mechanism design proposal).

33. *Id.* at §4.3.

34. Kaal, Wulf A., Evolution of Domain-Specific Reputation Systems: From Binary Validation to Citation-Weighted Knowledge Attribution (2026), available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6192998.