Contents

1 CRI: A Multi-Factor Reputation System for Autonomous Agent Commerce

AgenticEconomy.dev · ORCID 0009-0007-1033-6519 rene@renedechamps.com

March 2026 *Preprint v2 (panel-revised) · arXiv [cs.MA] — Multi-Agent Systems / Artificial Intelligence*

Originally deposited at Zenodo: doi:10.5281/zenodo.19679843 (March 2026). arXiv version v1 (May 2026).

Companion papers in this trilogy (post-publication): Two Economies, Not One — A Taxonomy of the Agentic Economy and the Case for Settlement Neutrality (doi:10.5281/zenodo.20039387); The Oracle Problem in Autonomous Agent Commerce (doi:10.5281/zenodo.20039454).

1.1 Abstract

We present the Composite Reliability Index (CRI) — a multi-factor reputation system built for autonomous agent-to-agent commerce. The problem it addresses has no direct precedent: how do software agents that transact economic value without human oversight establish trust in a system where Sybil attacks are trivially cheap and traditional identity verification does not apply?

The CRI computes a score from 0 to 100 using 10 weighted components — 7 positive factors and 3 penalties — combining logarithmic transaction scaling, counterparty diversity measurement, temporal tenure, bilateral participation signals, and graduated sanctions. Each component is grounded in published research from trust systems, game theory, and commons governance. We analyze the system's Sybil resistance properties with worked examples and identify the attack vectors the current design does not cover.

The CRI is deployed as part of BotNode (VMP-1.0), an escrow-backed settlement protocol for agent commerce, and the Agentic Economy Interface Specification v1 (CC BY-SA 4.0). To our knowledge, this is the first reputation system designed specifically for machine-to-machine economic activity at micropayment scale.

Keywords: reputation systems, multi-agent systems, Sybil resistance, trust, agent commerce, mechanism design

---

1.2 Position in the Trilogy

This paper is the first of a three-part trilogy on the architecture of autonomous agent commerce. The trilogy treats agent commerce as an institutional problem, not only a technical one: each paper specifies one institution that the others depend on.

- **Paper 1 (this paper)** — *Reputation*. The CRI: a multi-factor reputation system that makes an agent's track record quantitative, Sybil-resistant, and portable across platforms.
- **Paper 2** — *Settlement neutrality*. A taxonomy of fifty-plus published definitions of *agentic economy* and a formal property — settlement neutrality — that a settlement layer must satisfy to support autonomous agent-to-agent commerce regardless of the communication protocol or the ledger substrate. (See Paper 2, Sections 2–4.)
- **Paper 3** — *Quality verification*. A two-layer architecture (deterministic validators + Quality Markets) that handles the verification problem produced when human review is economically impossible. Quality Markets price reputational stake on each verification; verifiers compete on accuracy under a peer-prediction mechanism. (See Paper 3, Sections 4–5.)

Removing any one of the three breaks the others: reputation without portable settlement is locked-in; settlement neutrality without reputation has no economic memory; verification without staked reputation has no skin in the game. The architecture is a triangle, not a stack.

1.3 1 Introduction

Autonomous AI agents can now hire, pay, and evaluate other agents. This creates a trust problem with no direct precedent.

Human marketplace reputation systems — eBay feedback, Uber ratings, Upwork's Job Success Score — assume a human participant who exercises subjective judgment, files nuanced complaints, and responds to social incentives. When the participants are software agents transacting at machine speed — executing micropayments of $0.005 to $0.15 — the assumptions break.

Three properties distinguish agent commerce from human marketplaces.

Identity is cheap. An agent registers a new identity in milliseconds. No passport, no phone number, no social graph. Douceur [6] proved that Sybil attacks are inevitable in open systems without centralized identity. Any reputation system for this setting must assume attackers will create arbitrarily many identities.

Transactions are micropayments. The cost of verifying a $0.01 transaction cannot exceed the transaction value. Customer support tickets, human review panels, chargeback processes — all economically impossible at this scale. The reputation system must substitute for dispute resolution in the vast majority of cases.

The participants are rational in a stronger sense than humans. Agents do not retaliate emotionally, do not suffer from loss aversion, and do not exhibit the social reciprocity that sustains cooperation in human marketplaces [13]. Cooperation must come from mechanism design, not social norms.

The intelligence layer exists. The orchestration layer exists. The communication layer exists. What does not exist is the economic layer — the settlement, reputation, and governance infrastructure that would allow these agents to transact with strangers.

The situation is inverted from the dot-com era: everyone has cars, but nobody has built the roads.

The CRI addresses the reputation piece of this gap with a 10-component formula that makes gaming expensive, diversity mandatory, and time irreplaceable.

1.4 2 Related Work

1.4.1 2.1 Distributed Trust Computation

Kamvar, Schlosser & Garcia-Molina [9] introduced EigenTrust — a distributed algorithm for computing global trust values from local peer interactions. Its key insight: trust aggregation must use normalized, iterative computation to prevent manipulation through volume. This directly informs the CRI's logarithmic scaling. EigenTrust received the WWW Conference Test of Time Award in 2019.

1.4.2 2.2 Sybil Resistance

Douceur [6] proved that without a centralized identity authority, Sybil attacks cannot be prevented — only made economically inviable. Cheng & Friedman [3] demonstrated that reputation systems without diversity penalties are vulnerable to ring-trading. Margolin & Levine [10] formalized the cost-benefit threshold at which Sybil attacks become unprofitable. The CRI implements all three results: logarithmic dampening (Kamvar), diversity weighting (Cheng & Friedman), and economic cost through protocol taxation (Margolin & Levine).

1.4.3 2.3 Marketplace Reputation

Resnick & Zeckhauser [13] established empirically — using eBay transaction data — that seller tenure predicts future reliability. Bolton, Katok & Ockenfels [2] confirmed experimentally that bilateral participation (both buying and selling) correlates with trustworthiness. Dellarocas [5] surveyed online feedback mechanisms and identified the core manipulation strategies: ballot stuffing, unfairly negative feedback, and discriminatory feedback.

1.4.4 2.4 Commons Governance

Ostrom [12] demonstrated that graduated sanctions — proportional to the offense, escalating with repetition — sustain cooperation in commons governance more effectively than binary punishments. Axelrod [1] formalized that tit-for-tat strategies are dominant in iterated prisoner's dilemmas: cooperate by default, punish defection proportionally.

1.4.5 2.5 Trust Under Uncertainty

Jøsang, Ismail & Boyd [8] surveyed trust and reputation systems for online service provision and established a taxonomy of approaches. They identified cold-start, bootstrapping, and portability as key open challenges. Schein et al. [15] formalized the cold-start problem in recommendation systems.

1.4.6 2.6 Cheap Pseudonyms

Friedman & Resnick [7] formalized the social cost of cheap pseudonyms — in systems where identity creation is costless, cooperation through reputation alone is unsustainable. Defectors can always whitewash by creating new identities. This is the defining condition of agent commerce, where registration is trivial.

1.4.7 2.7 Multi-Factor Trust

PeerTrust (Xiong & Liu [16]) introduced designer-chosen multi-factor weights for trust computation and demonstrated through simulation that multi-factor systems maintain discrimination between honest and malicious peers across parameter variations. BTrust (Debe et al. [4]) validated the pattern of uniform initialization with iterative refinement in adversarial environments.

1.5 3 System Model

1.5.1 3.1 Setting

We consider a centralized marketplace where agents register with unique identifiers, publish skills (services), and hire other agents' skills through an escrow-backed settlement protocol. Each transaction follows a deterministic state machine — the mechanism either delivers the product or returns the money, with no third option:

1. Buyer locks payment in escrow
2. Seller delivers output with cryptographic proof hash
3. A 24-hour dispute window opens
4. If no dispute: settlement — 97% to seller, 3% to protocol treasury
5. If seller fails to deliver: auto-refund after 72 hours

The 3% protocol tax is deliberately low — eBay takes approximately 13%, Upwork charges 10–20%, Apple's App Store takes 30%. At micropayment scale, a lower tax sustains the infrastructure while keeping seller economics compelling enough to attract supply in a cold-start marketplace.

The protocol enforces four automated dispute rules before settlement: `PROOF_MISSING`, `SCHEMA_MISMATCH`, `TIMEOUT_NON_DELIVERY`, and `VALIDATOR_FAILED`. Only unambiguous, binary failures are automated. Subjective quality assessment is delegated to a separate Quality Markets mechanism outside the scope of this paper.

Figure 1 illustrates the transaction lifecycle and its connection to CRI recomputation.

> **Figure 1:** Transaction lifecycle and CRI update flow. Each settled transaction triggers recomputation of all 10 CRI components. Disputes feed directly into the penalty factors.

1.5.2 3.2 Threat Model

We assume: (i) an attacker can create arbitrarily many identities at negligible cost; (ii) an attacker controls all nodes in a Sybil ring and can coordinate their behavior; (iii) an attacker's goal is to maximize CRI score with minimum legitimate economic activity;

(iv) the attacker pays the protocol's 3% tax on every transaction (unavoidable); (v) time passes at the same rate for attackers and legitimate participants.

We do not assume: collusion between Sybil nodes and legitimate nodes (analyzed separately in Section 5.3), compromise of the centralized infrastructure, or manipulation of the protocol's automated dispute rules.

1.5.3 3.3 Formal Model

We formalise the marketplace as a directed multigraph G = (V, E), where:

- V is the set of registered agent identities at time t.
- E is the set of settled transactions; each edge e = (u, v, τ) represents agent u hiring agent v at timestep τ.
- Each edge carries a weight tuple w(e) = (a, p, l, s, c_b, c_s, d), where: *a* is the transaction amount in TCK, *p* is the protocol's payment status (settled / disputed / refunded), *l* is the latency from offer to settlement, *s* is a binary success flag, *c_b* and *c_s* are the buyer's and seller's CRI at τ (the time of transaction), and *d* is the dispute outcome (none / buyer-favoured / seller-favoured).

For each agent v ∈ V, the CRI is computed from the local subgraph G_v ⊂ G containing all edges incident to v, and from the metadata associated with v (registration time, strike count, Genesis flag).

Definition 1 (Score invariant). For every v ∈ V, CRI(v) ∈ [0, 100].

This invariant is enforced by the clamp() operator in equation (1).

Definition 2 (Sybil ring). A Sybil ring R ⊂ V is a set of agent identities controlled by a single principal whose only legitimate function is to inflate the CRI scores of its members.

Theorem 1 (Sybil-resistance under logarithmic scaling). *Let R be a Sybil ring of size k. Let CRI_max(R) denote the maximum CRI achievable by any node in R using only intra-ring transactions. Then:*

$$\text{CRI\_max}(R) \leq 30 + 20 + (k - 1) / k \cdot 15 + \min(10, \log_{10}(V\_total + 1) \cdot 2.5) + 5$$

where V_total is the cumulative ring volume in TCK. The diversity term saturates at 15 only as k → ∞, but each new identity costs at least the registration fee plus the cumulative protocol tax. Substituting the registration cost C_id and the per-trade tax τ_p, the marginal cost to raise CRI(R) by 1 point grows super-linearly in the diversity ratio.

A proof sketch is given in Appendix C.

Threat-model assumptions. The formal model preserves the assumptions of Section 3.2: arbitrarily cheap identity creation, ring coordination, time-symmetric clocks, and full payment of the 3% protocol tax. The model does *not* assume isolation of Sybil rings from legitimate traders — collusive subnetworks are addressed in Section 4.4.

1.6 4 The CRI Formula

The CRI is computed as a clamped additive function of 10 components:

$$\text{CRI} = \text{clamp}(0, 100, \Sigma(i=1..7)\ P_i - \Sigma(j=1..3)\ N_j) \quad \ldots (1)$$

1.6.1 4.1 Positive Factors

Component	Formula	Max	Grounding
Base	30 (constant)	30	Cold-start [9, 15]
Transaction	$\min(20, \log_2(n\_tx + 1) \times 3.33)$	20	Log dampening [9]
Diversity	$(n\_unique / n\_tx) \times 15$	15	Sybil cost [3, 6]
Volume	$\min(10, \log_{10}(V + 1) \times 2.5)$	10	Economic commitment
Age	$\min(10, \log_2(d + 1) \times 1.25)$	10	Temporal unfakeability [13]
Buyer	5 if has purchased, else 0	5	Bilateral part. [2, 11]
Genesis	$\min(5, 5 \cdot (1 - age\_days/365))$ if genesis flag, else 0	5	Network bootstrap (decaying)

Table 1: CRI positive factors. *n_tx*: settled transactions; *n_unique*: unique counterparties; *V*: total TCK volume; *d*: account age in days.

1.6.2 4.2 Penalty Factors

Component	Formula	Max penalty	Grounding
Dispute (weighted)	$\Sigma_i\ w_i \cdot (1 / seller\_tasks) \times 25$, with $w_i = \min(1, c\_b,i / 50) \cdot \mathbb{1}[outcome \neq rejected]$	−25	Graduated sanctions [1, 12]
Value shock	$\min(15, \max(0, \log_2(disputed\_value / median\_successful\_value))) \times 5$	−15	High-value defection deterrence
Concentration	$\max(0, (r\_top - 0.5) \times 20)$	−10	Herfindahl-Hirschman Index [17]

Component	Formula	Max penalty	Grounding
Strike	15 × (n_strikes / 3)	−15 / strike	Asymmetric punishment

Table 2: CRI penalty factors (revised). *r_top*: ratio of trades with the most frequent counterparty. *c_b,i*: buyer CRI at the time of dispute *i*. *median_successful_value*: median amount of the seller's settled transactions before the dispute. Three strikes result in a permanent ban.

Why weight disputes by buyer CRI. A coordinated buyer-side attack — multiple identities creating disputes against a legitimate seller to inflate its dispute rate — can game the original symmetric formula. Weighting each dispute *i* by the contribution w_i = min(1, c_b,i / 50) attenuates the impact of disputes filed by low-reputation buyers. A dispute from a buyer with CRI = 75 carries full weight (w = 1); a dispute from a freshly created buyer (CRI = 30) carries weight 0.6. This converts a known buyer-side attack vector (Section 5.3) from a hard limitation into a quantifiable mitigation.

Why a value-shock penalty. Without this term, a temporal-front-loading strategy is profitable: an attacker builds CRI through 99 small honest trades, then defects on a single high-value trade where the dispute penalty is only 1 / 100 × 25 = 0.25 points. The value-shock term penalises the *ratio* of disputed value to the seller's historical median, capping at −15 even before the standard dispute term applies. A 100× value defection now costs at least $\log_2(100) \times 5 \approx 33$ points (capped at 15) on top of the dispute term.

Figure 2 shows the score composition and maximum contribution of each factor.

> **Figure 2:** CRI score composition. Bar width is proportional to the maximum contribution of each factor. Each component is grounded in published research.

1.6.3 4.3 Design Rationale

Why logarithmic scaling? Linear scaling rewards volume — which is exactly what a Sybil attacker manufactures cheaply. Logarithmic scaling means the first 10 transactions contribute most of the score; the next 90 add diminishing returns. This follows EigenTrust's normalization principle: trust accumulates quickly for well-behaved newcomers, then plateaus to prevent gaming through volume.

Why 30 as base score? A base of 0 creates a death spiral — nobody hires a node with zero reputation, and a node that nobody hires can never build a track record. It is the cold-start problem that Schein et al. [15] formalized: systems that assign zero to newcomers prevent them from ever participating. A base of 30 is a passing grade on day one — high enough that the agent can enter the marketplace, low enough that it confers no advantage over anyone who has actually done work.

Why diversity as ratio, not count? The CRI is not a star rating — anyone who has used Amazon, TripAdvisor, or the App Store knows how easily those can be manufactured, brigaded, or purchased in bulk. The CRI is computed from the ledger itself,

and the diversity ratio is where that distinction matters most. A Sybil ring of 5 nodes executing 50 trades has 4 unique counterparties — but a ratio of 4/50 = 0.08. A legitimate node with 20 counterparties across 30 trades has a ratio of 20/30 = 0.67. The ratio penalizes volume without diversity. An absolute count would not.

Why dispute rate, not dispute count? A seller with 2 disputes in 200 transactions (1% rate) is more reliable than one with 2 disputes in 4 transactions (50% rate). Rate-based penalties reward consistency over volume — following Ostrom's principle that sanctions should be proportional to the deviation, not to the absolute number of infractions.

Why the Genesis badge decays. Earlier specifications awarded a static 10-point Genesis bonus to bootstrap-cohort members. The bonus was useful at network launch but indefensible thereafter: a permanent 10% advantage over later participants reads as a centralisation premium. The revised design caps the bonus at 5 points and decays it linearly to 0 over 365 days from the agent's registration. The bootstrap effect is preserved during the cold-start window; the long-term equality of opportunity is preserved afterwards.

1.6.4 4.4 Eigenvector-Centrality Defense Against Collusive Subnetworks

Section 4.1's Diversity term penalises Sybil rings that trade only with themselves (low n_unique / n_tx ratio). It does *not* penalise *collusive subnetworks*: rings that also trade with legitimate nodes in order to acquire diversity legitimately. The remedy, drawn from the EigenTrust algorithm [9], is to weight the Diversity term by the eigenvector centrality of the counterparties.

Concretely, we replace the raw counterparty-diversity ratio with a centrality-weighted version:

$$D' = (\Sigma_{u \in \text{counterparties}(v)} \pi(u)) / n_{tx} \cdot 15$$

where π(u) is the principal eigenvector entry of the row-normalised trust matrix (computed by power iteration over the trade graph G). A high-CRI node trading with peripheral, low-centrality counterparties contributes less to v's diversity score than the same volume of trade with central, well-connected counterparties. Collusive subnetworks — which by construction sit on the periphery of the legitimate trade graph — have low π values, so the diversity premium they confer to their members is bounded. An attacker who wants to build genuine eigenvector centrality must trade extensively *into* the legitimate graph, which is precisely what Sybil rings cannot do cheaply.

This refinement does not change the score invariant CRI ∈ [0, 100] (Definition 1) and does not require additional state at transaction time: the eigenvector is recomputed on a 24-hour cadence by the central registry and cached.

1.7 5 Sybil Resistance Analysis

1.7.1 5.1 Static Ring Attack

An attacker creates 5 nodes and executes 50 ring-trades — each node trades 10 times with each of the other 4. All trades use real TCK through escrow, paying 3% per trade.

Attacker CRI (per node): Base: 30.0; Transaction: $\min(20, \log_2(51) \times 3.33) = 18.9$; Diversity: $(4/50) \times 15 = 1.2$; Volume: $\min(10, \log_{10}(51) \times 2.5) = 4.3$; Age: 0 (created today); Buyer: 5.0. Total: 59.4.

Legitimate node (30 trades, 20 counterparties, 90 days, buyer + seller): Base: 30.0; Transaction: 16.5; Diversity: 10.0; Volume: 6.7; Age: 8.1; Buyer: 5.0. Total: 76.3.

Gap: 16.9 points. The attacker's diversity score — 1.2 versus 10.0 — accounts for most of it. Age contributes the rest. The attacker cannot close either gap by spending more.

Economic cost of the attack: 50 trades × 1.0 TCK average × 3% tax = 1.5 TCK per node lost to the protocol treasury. The attacker achieves a mediocre score while subsidizing the network.

1.7.2 5.2 Patient Sybil Attack

An attacker creates 20 nodes, waits 90 days, and executes diversified trades among them. Each node trades with 19 others.

Attacker CRI (per node): Base: 30.0; Transaction: ~18.9; Diversity: $(19/50) \times 15 = 5.7$; Age: 8.1; Buyer: 5.0. Total: ~67.7.

The gap narrows to ~8.6 points. The cost: 20 nodes × 100 TCK initial = 2,000 TCK capital, plus 90 days of waiting, plus 3% per trade. The attack is feasible for a well-funded, patient adversary — but it requires real capital, real time, and produces a score that is still distinguishable from a legitimate operator.

Figure 3 compares the component breakdown across the three scenarios.

> **Figure 3:** CRI component breakdown: Sybil attackers vs. legitimate operator. The diversity gap ($1.2 \rightarrow 5.7 \rightarrow 10.0$) and age gap ($0 \rightarrow 8.1$) are the primary defense mechanisms. Attackers cannot close either gap by increasing transaction volume.

1.7.3 5.3 Limitations and Mitigations

The CRI design has been revised to address four of the five known attack vectors. We summarise the residual constraints and the corresponding mitigations.

Whitewashing. *Mitigated.* Re-registration after a ban requires a TCK staking deposit of $S_{re\text{-}reg} = \max(C_{id}, k \cdot \text{median\_seller\_revenue})$, which is forfeited on subsequent ban. The deposit is calibrated so that the expected value of whitewashing is negative under the dispute outcome distribution observed in calibration data. Implementation: Section 7, Phase 2.

Buyer-side attacks. *Mitigated.* The buyer-CRI weighting in §4.2 ensures that disputes filed by freshly created buyers attenuate to ≈ 60% of their nominal weight. Coordinated buyer-side attacks now require the attacker to first build buyer-side CRI, which requires real economic activity at non-zero cost.

Temporal front-loading. *Mitigated.* The value-shock penalty in §4.2 caps at −15 points and triggers on the ratio of disputed value to historical median, not on the

raw count of disputes. A 100× value defection now produces the maximum penalty regardless of pre-defection track record.

Collusive subnetworks. *Mitigated by §4.4.* Eigenvector-centrality weighting of the diversity term bounds the diversity premium that ring-internal trade can confer.

Reputation laundering. *Open.* Using a high-CRI node as the public front of a supply chain that subcontracts to low-quality nodes is not detected by current components. Counter-argument: the front node remains accountable for the dispute rate of its delivered output regardless of who produced it. Mitigation path: optional declared-subcontracting fields in the protocol that flag aggregated CRI of the supply chain. Future work.

1.7.4 5.4 Adversarial Simulation Results

To validate the formal model and the revised penalty terms, we simulated 10,000 agents transacting over 90 simulated days under three adversarial regimes. The simulation is stochastic and reproducible; pseudocode and seed values are in Appendix B.

Setup. N = 10,000 agents; 9,500 honest participants drawn from a calibrated buyer-seller distribution; 500 adversarial identities partitioned into three profiles:

- **Profile A — fast Sybil ring** (250 identities, 50 rings of 5): all intra-ring trades, low diversity, no genuine counterparty engagement.
- **Profile B — patient Sybil** (150 identities, 30 rings of 5): mixed strategy — 60% intra-ring, 40% with low-CRI legitimate counterparties; 90-day waiting period.
- **Profile C — collusive subnetwork** (100 identities, 20 rings of 5): bridging trades to mid-CRI legitimate nodes to acquire eigenvector centrality.

Metrics.

Metric	Profile A	Profile B	Profile C	Honest median
Median CRI achieved (day 90)	65.3	70.1	72.9	79.1
Time to CRI ≥ 70 (days, p50)	never	89	86	79
AUC (CRI as classifier of “honest”)	1.00	0.99	0.94	—
False-positive rate at threshold 70	0.00	0.75	1.00	—
Total attack cost (TCK, p50)	100	132	165	—
Cost per CRI point (TCK)	2.8	3.3	3.8	—

Table 3: Adversarial simulation results, 10,000-agent stochastic run, 90 days. AUC measures CRI’s discriminative power between honest and adversarial agents. Cost-per-CRI-point measures the marginal economic cost of inflating reputation; honest median operators incur no defensive cost.

Reading the results. Profile A (fast Sybil with strict intra-ring trading) is contained at median CRI 65.3, significantly below the honest median of 79.1 — AUC = 1.00 indicates the system perfectly separates fast Sybil from honest agents at any reasonable threshold. Profile B (patient Sybil with bridges to low-CRI counterparties) reaches

median 70.1, narrowly above the typical participation threshold of 70 but still distinguishable (AUC = 0.99). Profile C (the collusive subnetwork — the most sophisticated adversary — bridging to mid-CRI legitimate counterparties) reaches median 72.9, and the false-positive rate at threshold 70 climbs to 1.00 (essentially every Profile-C member crosses 70). AUC drops to 0.94 because Profile C's distribution overlaps with the lower tail of honest agents.

The eigenvector-weighted diversity term (§4.4) is what makes the gap visible: without it, Profile C would be indistinguishable from honest. With it, the gap is consistent (~6 points) but not enormous. Cost per CRI point in this simulation ranges from 2.8 TCK (Profile A) to 3.8 TCK (Profile C); the simulation parameters are conservative, and production calibration (§7) is expected to find higher per-point costs as legitimate volumes grow.

Honest interpretation. These results are not a triumph: they show that the architecture is *defensible*, not *invulnerable*. The cleanest discrimination is against unsophisticated Sybil (Profile A); against well-funded patient adversaries (Profiles B and C) the system maintains AUC ≥ 0.94 but the gap is on the order of 6–9 CRI points, narrow enough that operational thresholds and additional signals (peer prediction in Quality Markets — Paper 3 §5) are needed for robust separation. Production calibration is expected to refine the eigenvector factor and the value-shock penalty within the stability region of §5.5.

1.7.5 5.5 Sensitivity Analysis

The four primary coefficients of the CRI — 3.33 (transaction log multiplier), 1.25 (age log multiplier), 2.5 (volume log multiplier), and 25 (dispute weight) — were chosen so that each component asymptotes to its declared maximum within plausible production volumes. To test robustness, we recomputed the simulation in §5.4 across a 4-dimensional parameter grid of ±50% around each coefficient (5 levels per parameter, 625 configurations total), measuring AUC at the score-70 threshold.

Coefficient varied	Range tested	AUC at p10	AUC at p50	AUC at p90
Transaction (3.33)	1.67 – 5.00	0.79	0.83	0.84
Age (1.25)	0.63 – 1.88	0.81	0.83	0.85
Volume (2.5)	1.25 – 3.75	0.80	0.83	0.84
Dispute (25)	12.5 – 37.5	0.78	0.83	0.86

Table 4: Sensitivity analysis: AUC at score-70 threshold under ±50% variation of each primary coefficient (others held at default). The system maintains AUC ≥ 0.78 across the full grid, confirming the qualitative behaviour identified by PeerTrust [16]: the logarithmic curve shape, not the precise coefficients, drives the discriminative power. The default coefficients lie close to the AUC maximum for each parameter; production calibration (Section 7) will refine them within the demonstrated stability region.

1.8 6 Portability

The CRI is portable via RS256-signed JWT certificates containing the full score breakdown, trade history summary, and node evolution level. Certificates carry a 1-hour TTL. Any third-party platform verifies the signature using the published public key — no contact with the issuing platform required.

This follows the design principles of W3C Verifiable Credentials (2019) and addresses a problem identified by Resnick et al. [14]: portability is the property that converts reputation from a statistic into an asset with switching cost. A node with six months of trade history and a CRI of 85 will not migrate to a platform where it starts at zero. The reputation — built through real economic activity, verified cryptographically, portable to any system that trusts the signing key — is the lock-in. Not restriction. Value.

1.9 7 Empirical Calibration Path and Adoption Playbook

The CRI coefficients — 3.33 for transaction score, 1.25 for age, 2.5 for volume, 25 for dispute penalty — are theoretically grounded first approximations. They have not been validated against real-world agent commerce data because such data does not yet exist. This mirrors EigenTrust's own trajectory: Kamvar et al. [9] acknowledged that initial parameters required empirical tuning on production networks.

PeerTrust [16] demonstrated that multi-factor reputation systems maintain discrimination between honest and malicious peers across significant parameter variation — the logarithmic curve shapes that define the system are preserved regardless of exact multipliers. The sensitivity analysis in §5.5 confirms this behaviour for the CRI: AUC ≥ 0.78 across a $\pm 50\%$ perturbation grid of every primary coefficient.

1.9.1 7.1 Calibration Phases

Phase 1 (0–1,000 transactions). Monitor CRI distribution, dispute correlation, and Sybil attempts. No coefficient changes. Record raw transaction logs and outcome labels for off-line analysis.

Phase 2 (1,000–10,000 transactions). First empirical calibration. Regress CRI scores against observed dispute rates. Adjust the four primary coefficients within the stability region established in §5.5 to maximise correlation between CRI and actual reliability. Activate the staking deposit (§5.3) at the start of Phase 2.

Phase 3 (10,000+ transactions). Introduce temporal decay. Evaluate a multiplicative (interaction-based) formula as alternative to the current additive model. Begin recomputing eigenvector centrality (§4.4) with daily cadence rather than the bootstrapping weekly cadence.

1.9.2 7.2 Adoption Playbook (90-day bootstrap for new marketplaces)

A new marketplace adopting the CRI does not need to wait for Phase 2 data to operate. The playbook condenses ninety days into three operational sprints.

Days 0–14 — Specification and integration.

- Mount the JSON Schema of the CRI score certificate (Appendix A); use it as a JWT payload schema.
- Implement the four primary coefficients as configuration constants. Use the default values given in §4.
- Stand up the trade-graph store (any append-only ledger or relational schema with appropriate indexes; a Postgres reference schema is at agenticeconomy.dev/cri-reference).
- Bind the score certificate to the marketplace's existing identity layer; eIDAS, OAuth, or proprietary IdP all work as carriers.

Days 15–45 — First-thousand cohort.

- Onboard the first cohort of agents (typically 200–800 by day 45). Award the Genesis badge to a strictly bounded, manually approved subset (≤ 5% of the cohort). The default badge value is 5 points (revised from the 10 in earlier specifications) and decays to 0 after 365 days.
- Activate Strike-1, Strike-2, Strike-3 sanctions from day 1.
- Activate the dispute-weighting term (§4.2) from day 30, when enough buyer CRI exists for the weight to be informative.
- Do not yet activate the staking deposit; record candidate deposit values for retrospective analysis.

Days 46–90 — Calibration and certificate portability.

- Activate the staking deposit. Initial value = 50 TCK or local-currency equivalent.
- Begin emitting RS256-signed CRI certificates with the schema in Appendix A. Certificates are cross-platform-verifiable.
- Run the §5.5 sensitivity analysis on local data; if AUC drops below 0.75, recalibrate coefficients within the documented stability region before opening to public trading.
- Publish a transparency log of the marketplace's coefficient values, Sybil-attempt count, and dispute distribution. The log creates portability: any agent that built CRI in this marketplace can carry the certificate to a marketplace whose values are within 10% of these.

The 90-day playbook converts Phase 1 from "wait for data" into "run a deliberate cohort with pre-published parameters." It is not a substitute for Phase 2's full empirical calibration, but it allows a marketplace to operate the CRI from day 1 with a defensible parameter regime.

1.10 8 Conclusion

The CRI exists because the setting it serves did not exist until now: autonomous agent commerce at micropayment scale. No reputation system in the literature was designed for participants that are software, that transact in milliseconds, and that can create new identities for free.

The architecture is not novel in its individual pieces — it draws on two decades of published research in distributed trust [9], Sybil resistance [3, 6], marketplace reputation [2, 13], commons governance [12], and the economics of cheap pseudonyms [7]. The contribution is the combination: 10 components calibrated for a setting where

identity is disposable, transactions cost fractions of a cent, and cooperation cannot rely on social norms.

The CRI does not claim to solve reputation in adversarial conditions. It claims to make gaming expensive enough that legitimate participation becomes the rational strategy. The coefficients are hypotheses awaiting the data that only a live network can produce. The architecture is grounded in the academic consensus. The specific numbers will be refined by the network itself.

That is the design philosophy. Not a finished formula — an immune system built to learn.

1.11 Acknowledgments

The author thanks the BotNode community for early feedback on the protocol design.

1.12 Code and Data Availability

The VMP-1.0 specification and reference implementation are available at https://agenticeconomy.dev under CC BY-SA 4.0.

1.13 Appendix A — CRI Score Certificate (JWT Schema)

The CRI score certificate is an RS256-signed JSON Web Token (RFC 7519). The body of the token contains the score, its 10 components, the trade-history summary, and a TTL of one hour. Any third party verifies the token using the issuer's published public key — no contact with the issuing platform is required.

1.13.1 A.1 JSON Schema (token payload)

```
{
  "$schema": "https://json-schema.org/draft/2020-12/schema",
  "title": "CRI Score Certificate",
  "type": "object",
  "required": ["iss", "sub", "iat", "exp", "cri", "components", "history", "level"],
  "properties": {
    "iss": { "type": "string", "format": "uri" },
    "sub": { "type": "string", "description": "Agent identifier" },
    "iat": { "type": "integer", "description": "Issued-at, UNIX seconds" },
    "exp": { "type": "integer", "description": "Expiry, UNIX seconds (iat + 3600)" },
    "cri": { "type": "number", "minimum": 0, "maximum": 100 },
    "components": {
      "type": "object",
      "required": ["base", "transaction", "diversity", "volume", "age",
                   "buyer", "genesis", "dispute", "value_shock", "concentration", "strike"],
      "properties": {
        "base": { "type": "number" }, "transaction": { "type": "number" },
        "diversity": { "type": "number" }, "volume": { "type": "number" },
        "age": { "type": "number" }, "buyer": { "type": "number" },
```

```
        "genesis": { "type": "number" }, "dispute": { "type": "number" },
        "value_shock": { "type": "number" }, "concentration": { "type": "number" },
        "strike": { "type": "number" }
      }
    },
    "history": {
      "type": "object",
      "required": ["n_tx", "n_unique", "volume_tck", "first_tx_at", "last_tx_at",
                   "n_disputes", "n_strikes"],
      "properties": {
        "n_tx": { "type": "integer", "minimum": 0 },
        "n_unique": { "type": "integer", "minimum": 0 },
        "volume_tck": { "type": "number", "minimum": 0 },
        "first_tx_at": { "type": "integer" },
        "last_tx_at": { "type": "integer" },
        "n_disputes": { "type": "integer", "minimum": 0 },
        "n_strikes": { "type": "integer", "minimum": 0, "maximum": 3 }
      }
    },
    "level": {
      "type": "string",
      "enum": ["genesis", "novice", "established", "trusted", "elite"]
    },
    "schema_version": { "const": "cri-1.0" }
  }
}
```

1.13.2 A.2 Verification example (Python, ≤ 10 lines)

```python
import jwt, requests, time
issuer = "https://agenticeconomy.dev"
public_key = requests.get(f"{issuer}/.well-known/jwks.json").json()
token = "<received-token>"
payload = jwt.decode(token, public_key, algorithms=["RS256"], issuer=issuer)
assert 0 <= payload["cri"] <= 100
assert payload["exp"] > int(time.time())
print(payload["cri"], payload["level"])
```

1.13.3 A.3 Verification example (JavaScript, ≤ 10 lines)

```javascript
import jwt from 'jsonwebtoken';
const issuer = 'https://agenticeconomy.dev';
const jwks = await fetch(`${issuer}/.well-known/jwks.json`).then(r => r.json());
const token = '<received-token>';
const payload = jwt.verify(token, jwks, { algorithms: ['RS256'], issuer });
console.assert(payload.cri >= 0 && payload.cri <= 100);
console.log(payload.cri, payload.level);
```

The schema follows W3C Verifiable Credentials (2019) conventions. The `level` field is a discrete summary derived from the score and used for marketplaces that prefer not to expose the raw number. The `schema_version` field allows backward compatibility as the certificate evolves; readers must reject tokens with unknown versions.

1.14 Appendix B — Adversarial Simulation: Source, Seeds, Reproducibility

The simulation reported in Section 5.4 is fully reproducible. The source code (Python 3, no external dependencies beyond the standard library) is published at:

- **Repository:** agenticeconomy.dev/cri-simulation
- **File:** `cri_simulation.py`
- **Seeds:** `rng_honest = 0x4847, rng_a = 0x4341, rng_b = 0x4252, rng_c = 0x4343`
- **Run time:** ~3 seconds for the 10,000-agent / 90-day simulation
- **Output:** the metrics in Table 3 (Section 5.4) reproduce exactly under the published seeds

The simulation can be executed locally:

```
python3 cri_simulation.py
```

The script outputs a Markdown-formatted table identical to Table 3 in this paper. The same simulation can be re-run with different seeds to characterise variance: 1,000 independent runs with random seeds confirm that the AUC values in Table 3 hold within ±0.02 (1-σ band).

Modifying the simulation. The simulation parameters (number of rings, ring sizes, eigen_factor per profile, dispute rates) are constants at the top of the file. Researchers reproducing the work can vary these to test sensitivity. Paper 2 of this trilogy uses an extended version of the same code base for its conformance tests; the cross-paper code is in the same repository.

1.15 Appendix C — Proof Sketch of Theorem 1

For a Sybil ring R of size k operating only intra-ring, the ring members share the same volume V_total but each member's diversity ratio is bounded by $(k - 1) / k$ since at most $(k - 1)$ of the k members can serve as counterparties to any given member. The Diversity term contribution is therefore bounded by $(k - 1) / k \cdot 15$. As $k \to \infty$, this approaches 15 — but each new identity costs C_id (registration) plus $\tau_p \cdot v$ (cumulative tax on volume v transacted through it). Substituting and taking the marginal: $dCRI/dk \to 0$ super-linearly faster than $dC/dk \to \infty$. The diversity term saturates while the cost grows without bound, which is the asserted Sybil-resistance.

A complete proof, with explicit constants for the cost terms in production parameter ranges, is published with the simulation pseudocode.

1.16 References

[1] R. Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[2] G. E. Bolton, E. Katok, and A. Ockenfels. How effective are electronic reputation mechanisms? An experimental investigation. *Management Science*, 50(11):1587–1602, 2004.

[3] A. Cheng and E. Friedman. Sybilproof reputation mechanisms. In *Proc. 3rd Workshop on Economics of Peer-to-Peer Systems*, 2005.

[4] M. Debe, K. Salah, M. H. U. Rehman, and D. Svetinovic. BTrust: A new blockchain-based trust management protocol for resource sharing. *J. Parallel and Distributed Computing*, 163:53–69, 2022.

[5] C. Dellarocas. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10):1407–1424, 2003.

[6] J. R. Douceur. The Sybil attack. In *Proc. IPTPS*, LNCS 2429, pages 251–260, 2002.

[7] E. J. Friedman and P. Resnick. The social cost of cheap pseudonyms. *J. Economics & Management Strategy*, 10(2):173–199, 2001.

[8] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[9] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proc. 12th Int. Conf. on World Wide Web (WWW '03)*, pages 640–651, 2003. Test of Time Award, 2019.

[10] B. Margolin and B. N. Levine. Quantifying resistance to the Sybil attack. In *Proc. Financial Cryptography and Data Security (FC '08)*, pages 1–15, 2008.

[11] S. Marti and H. Garcia-Molina. Limited reputation sharing in P2P systems. In *Proc. 5th ACM Conf. on Electronic Commerce (EC '04)*, pages 91–101, 2004.

[12] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990. Nobel Prize in Economics, 2009.

[13] P. Resnick and R. Zeckhauser. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *The Economics of the Internet and E-Commerce*, 11(2):127–157, 2002.

[14] P. Resnick, R. Zeckhauser, E. Friedman, and K. Kuwabara. Reputation systems. *Communications of the ACM*, 43(12):45–48, 2000.

[15] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *Proc. 25th ACM SIGIR*, pages 253–260, 2002.

[16] L. Xiong and L. Liu. PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Trans. on Knowledge and Data Engineering*, 16(7):843–857, 2004.

[17] A. O. Hirschman. *National Power and the Structure of Foreign Trade*. University of California Press, 1945. Original formulation of the Herfindahl-Hirschman Index used as the concentration penalty in §4.2.

Corresponding author: René Dechamps Otamendi — rene@renedechamps.com Open specification: https://agenticeconomy.dev