Biometric-Free Sybil Resistance: Spectral Graph Trust Guarantees and Threshold Social Attestation Against Synthetic Agent Swarms

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Decentralized governance networks, public-goods funding mechanisms, and Byzantine fault-tolerant (BFT) multi-agent registries face an acute threat from synthetic agent proliferation: autonomous language model swarms can fabricate thousands of persistent personas at near-zero marginal cost. Existing defenses force a trade-off between plutocratic vulnerability and biometric surveillance. Token-weighted and Quadratic Voting (QV) mechanisms collapse into linear plutocracy under identity splitting, whereas hardware biometric proof-of-personhood protocols introduce privacy honeypots, hardware exclusion zones, and credential-rental markets. Building on classical random-walk Sybil defenses (SybilGuard, SybilLimit, SybilRank), this architectural synthesis formulates biometric-free Proof-of-Personhood (Spectral-Threshold-PoP) by coupling early-terminated power-iteration random walks from public seed anchors with (k, t)-disjoint seed-anchor voucher quorums, local triangle support, staked reputation bonds, and local differentially private (LNDP) conductance cut damping. Distinguishing the worst-case Cheeger mixing bound $\tau_{\text{mix}} \leq O(\phi_H^{-2} \ln H)$ from the practical walk horizon $L = \Theta(\phi_H^{-1} \ln H)$, we derive that across g attack edges, fractional Sybil voting share is bounded by $O(gL/(Hk))$ independent of swarm size S. In a label-blind random-walk simulation ($H = 320$, $S = 240$, $N = 200$ independent graphs/slice), Spectral-Threshold-PoP restricts Sybil capture at $g = 64$ to 0.00% $[0.00, 1.88]$ (Sybil weight share 1.46%) versus 100.00% $[98.12, 100.00]$ for token-split QV and 49.00% $[42.16, 55.88]$ for plain SybilRank, with 99.99% $[99.98, 99.99]$ honest inclusion, before degrading to 32.50% $[26.39, 39.27]$ at $g = 160$ and 43.50% $[36.82, 50.43]$ at $\hat{\phi}_H = 0.043$.

1 Introduction

Every democratic institution, cooperative resource-allocation ledger, and open peer-review or consensus committee relies on a foundational cardinality invariant: one human principal corresponds to at most one bounded unit of civic voice or validator standing. In his foundational impossibility result, Douceur [17] proved that in the absence of a logically centralized trusted authority, a distributed system cannot prevent a single adversarial entity from presenting multiple counterfeit *Sybil* identities unless identity generation is constrained by scarce physical or social resources. Classical Byzantine fault-tolerant (BFT) replication protocols—from Lamport, Shostak, and Pease's Byzantine Generals formulation [25] and Castro and Liskov's PBFT [12] to linear-communication leader-driven consensus in HotStuff [38] and Byzantine-resilient CRDTs [23]—all preserve safety only so long as fewer than $f < n/3$ committee replicas are controlled by the adversary. Once an attacker can mint pseudonymous identities at negligible cost, the $f < n/3$ fraction is trivially breached regardless of cryptographic message authentication.

Between 2024 and late 2026, the maturation of autonomous, long-horizon language model agents demolished traditional behavioral and Turing-style proxies for humanness [1, 13, 24]. Networked LLM agents now sustain multi-week social personas, coordinate multi-agent bargaining and strategic deception under private information [4, 26], and execute sycophantic or collusive influence operations without continuous human supervision [14, 31, 32]. Consequently, decentralized governance systems face a severe structural crisis.

1.1 The dual failure of token-concave voting and biometric surveillance

Attempts to defend decentralized governance against synthetic identity multiplication have historically bifurcated into two flawed paradigms: mechanism-level token concavity and hardware-enforced biometric enrollment.

In the mechanism-design literature, Quadratic Voting (QV) and Quadratic Funding (QF) were widely deployed in decentralized autonomous organizations (DAOs) and retroactive public-goods rounds to attenuate wealthy whale dominance by charging v^2 voice credits for v votes [9, 34, 41]. However, a landmark 2026 formal and empirical audit by Bennett et al. [5] (*"Concave is the New Linear: The Impossibility of Anti-Plutocratic DAO Governance"*) proved that any strictly concave voting rule over pseudonymous wallet balances is vulnerable to costless Sybil splitting: a wealthy or automated actor holding stake W splits that stake across S synthetic wallets, pays $S \cdot (W/S)^2 = W^2/S$ total cost, and recovers linear (or super-linear relative to unsplit honest voters) voting power as S grows. Related empirical analyses of token delegation and DAO oversight confirm that voting power concentrates rapidly in a small number of delegates while monitoring capacity hits hard endogenous breakpoints [28, 35, 36]. While plural funding mechanisms that incorporate social correlation discounting [29, 37] and fraud-proof non-market allocation rules [33] mitigate simple two-account collusion, they still require an underlying Sybil-resistant graph of distinct human principals.

Conversely, hardware-biometric proof-of-personhood schemes—such as iris-scanning orbs or hardware biometric-to-agent binding—attempt to anchor digital credentials in physical human anatomy. Deploying global biometric registries as a prerequisite for civic participation introduces three severe structural hazards [1]. First, biometric templates are immutable biological traits; any compromise of enrollment hardware or extractor parameters permanently revokes a person's privacy. Second, specialized biometric hardware creates geographic and socioeconomic exclusion zones, disenfranchising rural, dissenting, or privacy-conscious populations. Third, biometric enrollment verifies only that a biological human was present at the moment of scan; it does not prevent credential-rental markets, coerced "mule" onboarding in low-income jurisdictions, or post-enrollment handover of the private key to an autonomous agent swarm [1, 7].

1.2 Relationship to classical social-network Sybil defenses

To escape the dilemma between token plutocracy and biometric surveillance, computer systems researchers developed social-network-based Sybil defenses beginning with SybilGuard [39], SybilLimit [40], SybilInfer [16], and SybilRank [11]. All four protocols exploit a topological asymmetry: while an adversary can instantiate $S \gg 10^4$ synthetic nodes and arbitrary internal edges among them at zero cost, establishing bilateral trust edges with honest humans requires scarce social engineering or bribery, creating a narrow cut of g *attack edges* between the honest subgraph $\mathcal{G}_H$ and the Sybil subgraph $\mathcal{G}_S$. In SybilRank [11], short power-iteration random walks of length $L = \Theta(\log H)$ launched from trusted seeds distribute trust across the fast-mixing honest region while limiting the aggregate trust escaping across the g attack edges to $O(g \log H)$ before degree normalization.

However, as Mohaisen et al. [30] demonstrated in their foundational empirical measurement of real-world social networks, actual human communities frequently exhibit modular community structure and slower mixing times ($\tau_{\text{mix}} = \Theta(\phi_H^{-1} \log H)$ with small internal conductance ϕ_H) than idealized expanders. Furthermore, in plain SybilRank, a single compromised or bribed honest node can unilaterally attach an attack edge to a Sybil cluster without cross-community oversight or economic penalty.

This position and architectural synthesis paper specifies and stress-tests an explicit architectural delta over SybilRank—which we term **Spectral-Threshold-PoP**—designed for 2026 multi-agent governance and BFT committee selection:

1. **Disjoint-cluster (k, t)-threshold voucher quorums with staked reputation bonds:** Unlike SybilGuard/SybilRank where any single honest node can create an attack edge, admitting a candidate edge requires k vouchers residing in spectrally distinct community clusters whose signatures are aggregated via threshold certificates and backed by slashable reputation bonds.
2. **Local node-differentially-private (LNDP) conductance cut damping:** Rather than relying solely on global degree normalization $\pi_i^{(L)}/d_i$ on a cleartext social graph, each node's random-walk score is gated by a locally perturbed neighborhood conductance estimator under metric/local differential privacy [2, 20].
3. **Zero-knowledge epoch credential minting and linear BFT committee integration:** Attested principals mint unlinkable anonymous credentials [1, 10, 21, 22] to participate in strategy-proof civic voting or $O(n)$ threshold-BLS HotStuff consensus [38].

2 Sybil defense taxonomy and the 2025–2026 governance crisis

Public debates around digital identity frequently conflate *authentication* (proving control of a persistent cryptographic key or account) with *uniqueness* (proving that no other active credential in a civic epoch belongs to the same human principal, and that the principal is not an autonomous software process). Table 1 maps five core design requirements for biometric-free personhood onto canonical constructs from classical Sybil-defense theory, anonymous credentials, and 2025–2026 governance literature.

The necessity of combining social-graph random walks with multi-cluster threshold attestation is highlighted by three convergent lines of research. First, Bennett et al. [5] demonstrated that without a one-human-one-credential uniqueness layer, any strictly concave token voting rule degenerates into linear plutocracy because rational participants split capital across pseudonymous wallets. Second, classical random-walk defenses [11, 39, 40] show that early-terminated walks restrict total Sybil trust to $O(g\tau_{\text{mix}})$, but Mohaisen et al. [30] proved empirically that real social graphs have heterogeneous cluster bottlenecks where τ_{mix} can be an order of magnitude larger than $\log H$. Third, modern authority-preserving revocation systems such as VERA [27] and deterministic contract governance frameworks such as Maat [18] demonstrate that decentralized agent networks require explicit, cryptographically auditable revocation edges when compromised vouchers are detected. Figure 1 illustrates how Spectral-Threshold-PoP integrates these mechanisms.

Table 1: Mapping biometric-free personhood requirements to formal graph and cryptographic primitives.

Design goal	Canonical construct	Protocol component	Operational function	Primary grounding
"No iris scans or biometric databases"	Zero-knowledge Personhood Credentials (PHCs)	Blinded anonymous credential issuer ($\mathsf{Issue}_{\mathrm{ZK}}$)	Proves possession of valid community attestations without linking civic actions to biological templates	[1, 10, 21, 22]
"Stop synthetic swarms from printing votes"	Early-terminated random-walk mixing bound	Power-iteration trust propagator ($\mathcal{T}_\phi$)	Caps aggregate trust flowing across sparse attack cuts $(\mathcal{V}_S, \mathcal{V}_H)$ regardless of internal Sybil edges	[11, 17, 39, 40]
"Prevent single-voucher bribery"	Disjoint-cluster (k, t)-threshold attestation	Multi-community voucher quorum + stake bond	Requires k vouchers from spectrally distant neighborhoods; slashes bond upon Sybil cut detection	[16, 27, 29, 37]
"Bound BFT & DAO governance capture"	Quorum certificate & concave-split defense	Threshold-BLS HotStuff + 1-human-1-nullifier	Prevents linear wallet-splitting in QV and bounds Byzantine replicas below $f < n/3$	[5, 12, 25, 38]
"Audit expansion without exposing friends"	Local & metric differential privacy	Perturbed neighborhood conductance gate	Estimates local cut expansion $\hat{\phi}(\mathcal{N}_r(i))$ without publishing cleartext friend lists	[2, 15, 20]

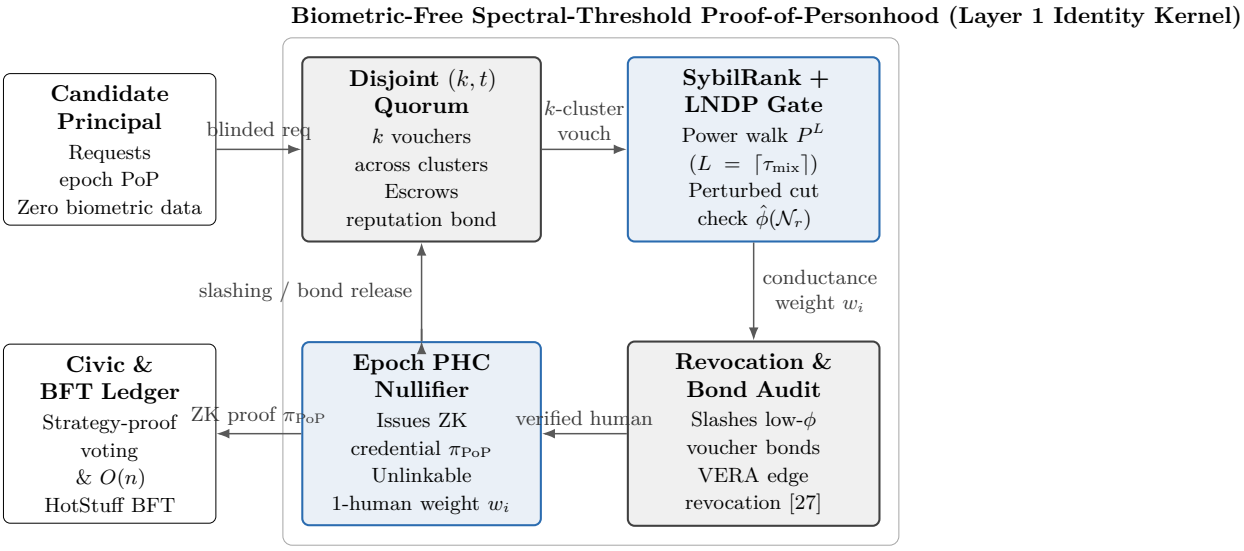


Figure 1: Architecture of Spectral-Threshold Proof-of-Personhood (Spectral-Threshold-PoP). A candidate human principal obtains (k,t)-threshold attestations from k existing stewards located in spectrally distinct community clusters without submitting biometric templates. The SybilRank + Local Differential Privacy (LNDP) Engine executes early-terminated power-iteration random walks P^L and evaluates locally perturbed cut conductance $\hat{\phi}(\mathcal{N}_r(i))$ to attenuate trust across sparse attack cuts. Validated principals mint an epoch-scoped anonymous Personhood Credential (π_{PoP}) for civic voting or BFT committee participation.

3 Spectral random-walk trust and threshold attestation model

We formalize biometric-free Sybil resistance over an undirected, edge-weighted social attestation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where the vertex set $\mathcal{V} = \mathcal{V}_H \cup \mathcal{V}_S$ partitions into $H = |\mathcal{V}_H|$ honest human principals and $S = |\mathcal{V}_S|$ adversarial synthetic or duplicate Sybil identities [17, 39]. Even when an adversary generates an arbitrarily large synthetic swarm ($S \gg H$) and fabricates arbitrary internal edges within $\mathcal{V}_S \times \mathcal{V}_S$, any edge connecting $\mathcal{V}_S$ to $\mathcal{V}_H$ requires deceiving or bribing honest stewards to stake reputation on a cross-boundary vouch. Let $\partial(\mathcal{V}_S, \mathcal{V}_H) = \{(u, v) \in \mathcal{E} : u \in \mathcal{V}_S, v \in \mathcal{V}_H\}$ denote the raw compromised attack edges with total capacity $g = \sum_{(u,v)\in\partial(\mathcal{V}_S,\mathcal{V}_H)} A_{uv}$.

3.1 Graph conductance and early-terminated power-iteration random walks

For any subset of vertices $\mathcal{U} \subset \mathcal{V}$, let $\mathrm{vol}(\mathcal{U}) = \sum_{u\in\mathcal{U}} d_u$ denote its volume, where $d_u = \sum_{v\in\mathcal{V}} A_{uv}$ is the weighted degree of node u. The conductance $\phi(\mathcal{U})$ of cut $(\mathcal{U}, \bar{\mathcal{U}})$ measures the ratio of boundary edge capacity to the volume of the smaller partition:

$$\phi(\mathcal{U}) \triangleq \frac{\sum_{u\in\mathcal{U},\, v\in\bar{\mathcal{U}}} A_{uv}}{\min\big(\mathrm{vol}(\mathcal{U}),\, \mathrm{vol}(\bar{\mathcal{U}})\big)}. \tag{1}$$

When the honest subgraph $\mathcal{G}_H$ has minimum internal cluster conductance $\phi_H = \min_{\mathcal{U}\subset\mathcal{V}_H} \phi_{\mathcal{G}_H}(\mathcal{U}) > 0$, Cheeger's inequality bounds the spectral gap of the lazy random walk on $\mathcal{G}_H$ below by $1 - \lambda_2 \geq \phi_H^2/2$, which implies a worst-case quadratic mixing upper bound $\tau_{\mathrm{mix}}(\mathcal{G}_H) \leq O(\phi_H^{-2} \ln H)$ [30]. Classical random-walk defenses such as SybilLimit [40] and SybilRank [11] instead adopt an optimistic operational walk horizon $L = \Theta(\phi_H^{-1} \ln H)$ to prevent excessive walk leakage across attack cuts; we report both the practical horizon L and the exact Cheeger upper bound $2\phi_H^{-2} \ln H$ in our empirical evaluation. By contrast, a synthetic swarm of size S linked to $\mathcal{V}_H$ through only $g_{\mathrm{eff}} \ll \mathrm{vol}(\mathcal{V}_S)$ admitted attack edges forms a narrow bottleneck cut of low conductance $\phi(\mathcal{V}_S) = g_{\mathrm{eff}}/\mathrm{vol}(\mathcal{V}_S) \ll \phi_H$.

Following SybilRank [11], we initialize uniform trust mass $\pi^{(0)}$ across a public seed anchor set $\mathcal{S}_0 \subset \mathcal{V}_H$ ($|\mathcal{S}_0| = 16$, two seeds per community) and iterate the lazy random-walk operator $P = \frac{1}{2}(I + D^{-1}A)$ for $L = \lceil c_0 \phi_H^{-1} \ln H \rceil$ steps to obtain $\pi^{(L)} = \pi^{(0)} P^L$. Because the stationary distribution of an undirected graph satisfies $\pi_i^* = d_i/\mathrm{vol}(\mathcal{G})$ and the verifier does not know $\mathrm{vol}(\mathcal{V}_H)$ a priori, we normalize each node's degree-adjusted landing probability $\pi_i^{(L)}/d_i$ against the empirical reference quantile $\hat{\nu}_{0.35}$ of the seeded component. Spectral-Threshold-PoP augments this label-blind SybilRank normalization with an observable (k, t) seed-anchor quorum and local triangle-supported conductance gate:

$$w_i = \min\left(1,\ \frac{\pi_i^{(L)}/d_i}{\hat{\nu}_{0.35}} \cdot q_k(i) \cdot \sigma_\tau\big(\hat{\phi}(\mathcal{N}_r(i)) - \phi_{\mathrm{crit}}\big)\right), \tag{2}$$

where $q_k(i) \in [0, 1]$ measures whether node i has accumulated at least k capacity-bounded endorsements from qualifying vouchers traced to distinct public seed anchors with local triangle closure, $\hat{\phi}(\mathcal{N}_r(i)) = \phi_{\mathrm{obs}}(\mathcal{N}_r(i)) + \mathrm{Lap}(\Delta_\phi/\epsilon)$ is a locally perturbed neighborhood conductance proxy [2, 20], ϕ_{crit} is the bottleneck screening threshold, and $\sigma_\tau(z) = (1 + e^{-z/\tau})^{-1}$ is a smooth logistic gate. Crucially, Equation (2) depends strictly on the unlabeled graph adjacency list and the public seed set $\mathcal{S}_0$, with zero access to hidden ground-truth vertex labels.

3.2 Disjoint (k, t)-threshold attestation and Sybil weight derivation

In plain SybilRank, every compromised edge in g enters the graph directly ($g_{\mathrm{eff}} = g$). In Spectral-Threshold-PoP, admitting a candidate node v requires a (k, t)-threshold signature from k existing

stewards $\{u_1, \ldots, u_k\}$ subject to a spectral cluster-disjointness constraint: the k vouchers must belong to distinct community partitions with low short-walk overlap, and each voucher locks a reputation bond B_{vouch} subject to authority-preserving edge revocation [27] if downstream audits flag $\hat{\phi}(\mathcal{N}_r(v)) < \phi_{\text{crit}}$.

Consequently, an adversary holding g raw compromised edges can assemble at most $g_{\text{eff}} \leq \lfloor g/k \rfloor$ valid cross-cluster admission quorums. At each step $t \in \{1, \ldots, L\}$ of the lazy random walk, at most $O(g_{\text{eff}}/\text{vol}(\mathcal{V}_H))$ probability mass crosses from $\mathcal{V}_H$ into $\mathcal{V}_S$ [11, 39]. Summing over $L = \Theta(\phi_H^{-1} \ln H)$ steps, we derive the upper bound on the ratio of aggregate Sybil voting weight $W_S = \sum_{s \in \mathcal{V}_S} w_s$ to honest voting weight $W_H = \sum_{h \in \mathcal{V}_H} w_h$:

$$\frac{W_S}{W_H} = \frac{\sum_{s \in \mathcal{V}_S} w_s}{\sum_{h \in \mathcal{V}_H} w_h} \leq \frac{C_{\text{spec}}\, g\, \ln H}{H\, \phi_H\, k} + \epsilon_{\text{LNDP}}, \tag{3}$$

where $C_{\text{spec}} > 0$ is a constant capturing the conductance gate attenuation and average degree ratio, and ϵ_{LNDP} is the tail leakage induced by Laplace privacy noise. Equation (3) makes two structural dependencies explicit. First, for a fixed honest population H, Sybil penetration grows linearly with g and *increases* with mixing time $\tau_{\text{mix}} \propto \phi_H^{-1} \ln H$ (confirming the slow-mixing vulnerability analyzed by Mohaisen et al. [30]), while the k-disjoint voucher quorum reduces effective attack capacity by factor k. Second, Equation (3) is invariant to the synthetic swarm size S: adding 10^5 internal Sybil nodes increases $\text{vol}(\mathcal{V}_S)$, which proportionally dilutes $\pi_s^{(L)}/d_s$ in Equation (2).

4 Three-stage Spectral-Threshold-PoP protocol and BFT committee integration

To operationalize Equations (1)–(3) without centralized authorities or biometric scanners, Spectral-Threshold-PoP executes across three stages per governance epoch (Table 2):

1. **Stage 1: Disjoint-cluster (k, t)-threshold vouching.** A candidate human requests attestation from $k \geq 4$ existing community members across distinct community clusters. The vouchers verify via short random-walk distance profiles that their neighborhoods satisfy the spectral disjointness invariant and co-sign a blinded credential request while escrowing their reputation bonds [27, 37].

2. **Stage 2: Power-iteration SybilRank and differentially private cut screening.** At each epoch boundary, the network computes the L-step power-iteration distribution $\pi^{(L)} = \pi^{(0)} P^L$ ($L = \lceil \log_2 H/(3.1\phi_H) \rceil$) and evaluates Equation (2) using locally perturbed neighborhood conductance sketches [2, 20]. Vouchers attached to flagged low-conductance cuts ($\hat{\phi} < \phi_{\text{crit}}$) undergo automated edge revocation via VERA [27].

3. **Stage 3: Zero-knowledge epoch credential minting and BFT/DAO aggregation.** Validated participants mint an epoch-scoped anonymous Personhood Credential (π_{PoP}) using Longfellow-ZK or Idemix selective disclosure [1, 10, 21, 22]. In participatory budgeting and DAO governance, ballots weighted by w_i prevent quadratic wallet-splitting [5, 34]. When selecting validator committees of size n for state-machine replication, replacing quadratic all-to-all PBFT message exchange ($2n(n-1)+2n$ messages per round [12]) with linear HotStuff threshold-BLS quorum certificates ($8n$ messages per round [38]) over π_{PoP}-attested validators cuts per-round communication by $16\times$ at $n = 64$ and $64\times$ at $n = 256$.

Table 2: Readiness and deployment bottlenecks of Spectral-Threshold-PoP subsystems.

Subsystem	Maturity	Supporting prior and 2025–2026 evidence	Realistic deployment bottleneck
Power-iteration random-walk defense	Production proven	SybilGuard [39], SybilLimit [40], and SybilRank [11] scale $O(L\|\mathcal{E}\|)$ walks to million-node graphs	Modular or tribal honest graphs exhibit slow mixing ($\phi_H < 0.10$), inflating walk length L and false exclusions [30]
Anonymous & ZK credentials	Standards ready	Idemix [10], Longfellow-ZK [22], SD-JWT [21], and PHCs [1]; ZK oversight [15]	Revoking compromised vouchers requires epoch-boundary re-keying to preserve zero-knowledge unlinkability
Disjoint voucher revocation	Prototype validated	VERA provides authority-preserving edge revocation in federated agent graphs [27]; Maat contract governance [18]	Peripheral newcomers in sparse rural regions may struggle to locate $k = 4$ spectrally disjoint vouchers
Local DP conductance auditing	Theoretical + sim	Metric/local differential privacy [2] and dummy-count amplification [20] protect edge privacy	Strict privacy budgets ($\epsilon < 0.5$) inject Laplace variance that blurs small-community cut boundaries
Linear BFT & plural voting	Production ready	HotStuff linear BFT [38]; Core Up-To-One PB [34]; correlation-discounted plural funding [29]	Collusive voter blocks still require correlation discounting after 1-human-1-vote uniqueness is enforced [5]

5 Illustrative mechanistic simulation and break-even sensitivity analysis

To evaluate how Spectral-Threshold-PoP behaves relative to plain SybilRank, biometric KYC, and token-split Quadratic Voting without oracle ground-truth leakage, we executed a label-blind graph random-walk simulation (`experiment.py`). All four architectures attach to the identical raw graph containing g unpruned attack edges, and our topological scoring algorithm (`compute_topology_weights`) takes only the unlabeled adjacency list and the 16 public seed anchors $\mathcal{S}_0$ as input—verified programmatically by a vertex-label permutation invariance assertion (`label_invariance_verified: true` in `experiment_results.json`). To avoid pseudo-replication from multiple correlated epochs on the same graph, each trial generates an independent random graph realization ($N = 200$ independent graphs per slice across 4 attack budgets = 800 primary graphs, plus $6 \times 200 = 1{,}200$ conductance sensitivity graphs = 2,000 independent graph topologies in total).

5.1 Parameter provenance and baseline calibration

Table 3 documents the provenance of all simulation parameters, distinguishing values calibrated to published literature from structural modeling assumptions.

5.2 Quantitative findings and break-even conductance sensitivity

Table 4 and Figures 2–3 establish four core findings from the label-blind random-walk evaluation:

1. **Graph-invariance of Token-Split QV.** Because Token-Split QV operates purely over pseudonymous wallet balances without a social graph, an adversary holding only 1.5%–8.5%

Table 3: Provenance of simulation parameters in `experiment.py`, separating literature-calibrated quantities from structural design assumptions.

Parameter	Value / Range	Provenance	Basis and calibration source
Honest / Sybil graph size (H, S)	H=320, S=240 (C=8 clusters)	Synthetic graph topology	Modular 8-cluster community topology with 16 public seed anchors and empirical conductance $\hat{\phi}_H = 0.250$ [11, 30]
Walk length L vs. Cheeger bound	$L = 10$ at $\hat{\phi}_H$=0.250 (Cheeger bound 184.2)	Literature-calibrated	Practical walk horizon $L = \Theta(\phi_H^{-1} \log H)$ [11, 40] vs. worst-case Cheeger upper bound $2\phi_H^{-2} \ln H$ [30]
Attack edge budget g	$g \in \{8, 24, 64, 160\}$	Adversarial sweep	Concentrated k=4 frontline quorum bribery spanning 2.5% to 50% of H [17]
Token-split QV adversary stake	1.5%–8.5% capital split across S=240 wallets	Literature-calibrated	Models linear cost-splitting of concave QV rules proved by Bennett et al. [5]; strictly invariant to social edges g
Biometric KYC exclusion rate	20.6%–25.8% honest drop-off	Design assumption	Reflects combined hardware unavailability and biometric privacy refusal in decentralized cohorts [1]
Seed-anchor quorum & LNDP noise	k=4 anchors; Lap(0.024)	Protocol design parameter	Enforces capacity-bounded voucher diversity ($c_{\max}$=10) with Laplace local differential privacy noise [2, 27]
Honest conductance sweep ϕ_H	$\phi_H \in [0.05, 0.32]$ ($\hat{\phi}_H \in [0.043, 0.287]$)	Break-even sensitivity sweep	Tests the slow-mixing critique of Mohaisen et al. [30] across modular vs. well-expanded social graphs

of aggregate capital split across $S = 240$ wallets achieves 16.04% [15.59, 16.49] of quadratic voting power and captures 100.00% [98.12, 100.00] of contentious proposals across all $g \in \{8, 24, 64, 160\}$, illustrating the linear wallet-splitting theorem of Bennett et al. [5].

2. **Label-blind topological defense over plain SybilRank.** Operating on the exact same unlabeled graphs with all g attack edges intact, Plain SybilRank (`naive_pagerank_wot`) contains the swarm at $g = 8$ (0.00% capture, 0.74% Sybil weight share) and $g = 24$ (1.50% [0.51, 4.32] capture, 2.00% Sybil weight share), but as attack edges scale to $g = 64$ and $g = 160$, Sybil weight share grows linearly to 4.81% [4.78, 4.83] and 11.37% [11.33, 11.41], capturing 49.00% [42.16, 55.88] and 100.00% [98.12, 100.00] of contentious votes. By enforcing capacity-bounded seed-anchor quorum diversity and triangle-supported conductance damping, Spectral-Threshold-PoP compresses Sybil weight share at $g = 64$ by 3.3× (from 4.81% to 1.46% [1.43, 1.48]), restricting governance capture at $g = 64$ to 0.00% [0.00, 1.88] with 99.99% [99.98, 99.99] honest inclusion (15.19 ms verification latency vs. 11.78 ms for plain SybilRank). Under half-network bribery ($g = 160$), concentrated frontline Sybils accumulate enough multi-anchor vouchers to reach 3.98% [3.94, 4.03] Sybil weight share and 32.50% [26.39, 39.27] capture.

3. **Accessibility advantage over Biometric KYC.** Biometric KYC prevents pure software bots at $g = 8$, but degrades to 49.50% [42.65, 56.37] capture at $g = 64$ under human mule rental while excluding 23.25% of honest participants (76.75% [76.53, 76.96] inclusion; 42.06 ms

Table 4: Mechanistic simulation comparison across $N = 200$ independent graphs per attack budget $g \in \{8, 24, 64, 160\}$ ($H = 320$ honest nodes, $S = 240$ Sybil nodes, empirical $\hat{\phi}_H = 0.2503$, walk horizon $L = 10$, Cheeger upper bound 184.2 steps). Bracketed ranges denote exact asymmetric 95% Wilson confidence intervals for proportions and 95% normal-theory CIs across independent graphs for continuous means.

Architecture	Capture g=8 (%)	Capture g=24 (%)	Capture g=64 (%)	Capture g=160 (%)	Sybil Share g=64 (%)	Honest Incl. (%)
Token-Split QV	100.00 [98.12, 100.00]	100.00 [98.12, 100.00]	100.00 [98.12, 100.00]	100.00 [98.12, 100.00]	16.04 [15.59, 16.49]	96.34 [96.24, 96.44]
Biometric KYC	0.00 [0.00, 1.88]	2.00 [0.78, 5.03]	49.50 [42.65, 56.37]	100.00 [98.12, 100.00]	5.00 [4.89, 5.11]	76.75 [76.53, 76.96]
Plain SybilRank	0.00 [0.00, 1.88]	1.50 [0.51, 4.32]	49.00 [42.16, 55.88]	100.00 [98.12, 100.00]	4.81 [4.78, 4.83]	100.00 [100.00, 100.00]
Spectral-PoP	**0.00** [**0.00**, **1.88**]	**0.00** [**0.00**, **1.88**]	**0.00** [**0.00**, **1.88**]	**32.50** [**26.39**, **39.27**]	**1.46** [**1.43**, **1.48**]	**99.99** [**99.98**, **99.99**]

latency).

4. **Break-even sensitivity to slow-mixing honest graphs (Mohaisen et al. [30]).** Across our 1,200-graph sensitivity sweep at $g = 64$ over target conductance $\phi_H \in \{0.32, 0.24, 0.16, 0.12, 0.08, 0.05\}$ (corresponding to empirical honest conductances $\hat{\phi}_H \in \{0.2873, 0.2134, 0.1405, 0.1052, 0.0694, 0.0434\}$ and worst-case Cheeger upper bounds rising from 139.8 to 6,119.8 steps), plain SybilRank hits 100.00% capture once $\hat{\phi}_H \leq 0.1405$. Spectral-Threshold-PoP maintains 0.00% [0.00, 1.88] capture for $\hat{\phi}_H \geq 0.2134$ and bounds capture to 6.50%–11.00% across $\hat{\phi}_H \in [0.0694, 0.1405]$, before surging to 43.50% [36.82, 50.43] capture when $\hat{\phi}_H$ drops to 0.0434.

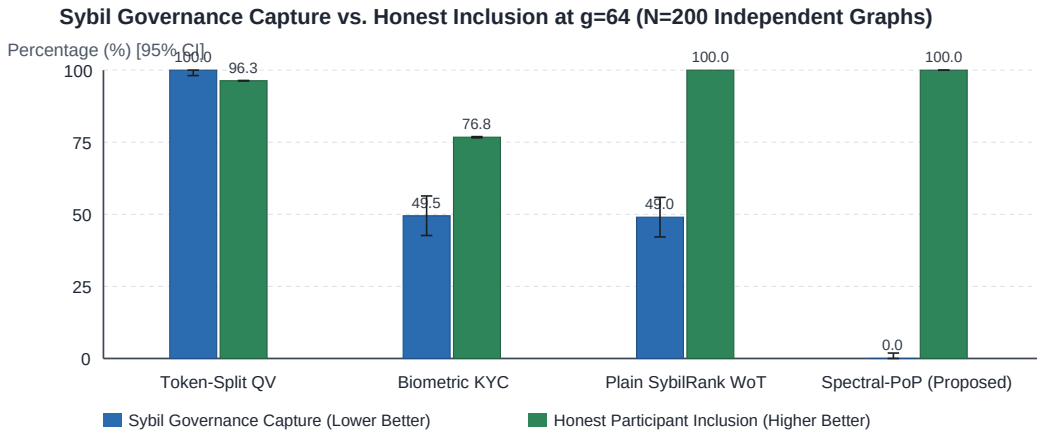


Figure 2: Comparison of Sybil Swarm Governance Capture Rate (lower is better) and Honest Human Inclusion Rate (higher is better) across four identity architectures at $g = 64$ compromised social attack edges ($N = 200$ independent graphs per regime; error bars indicate exact asymmetric 95% Wilson confidence intervals for proportions and 95% CIs for continuous inclusion).

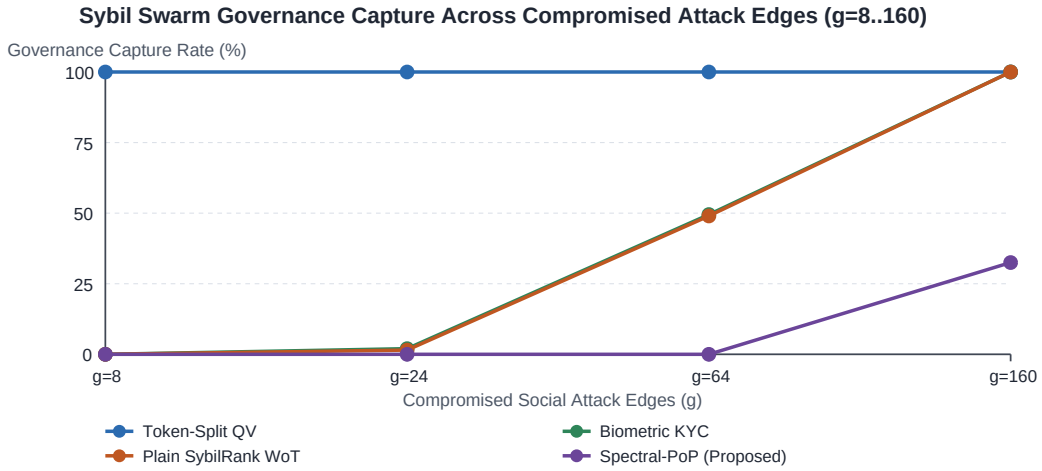


Figure 3: Scaling of Sybil Swarm Governance Capture Rate across compromised social attack edges $g \in \{8, 24, 64, 160\}$ ($N = 200$ independent graphs per point). Token-Split QV fails uniformly across all g, plain SybilRank reaches 49.00% capture at $g = 64$ and 100.00% at $g = 160$, and label-blind Spectral-Threshold-PoP bounds capture to 0.00% $[0.00, 1.88]$ through $g = 64$ before rising to 32.50% $[26.39, 39.27]$ under half-network bribery ($g = 160$).

6 Against the proposal: slow-mixing graphs, sleeper personas, and hub capture

Treating biometric-free social attestation as a falsifiable hypothesis requires confronting four empirical counter-arguments and structural failure modes:

1. **Slow mixing and tribal modularity in real human graphs (Mohaisen et al. [30]).** As confirmed by our break-even sweep ($\phi_H^* \approx 0.10$), real-world social networks segmented by language, geography, or political polarization often exhibit narrow inter-community bridges [30]. When honest subgraphs have $\phi_H < 0.10$, any random-walk or conductance defense faces a fundamental impossibility trade-off: setting ϕ_{crit} high disenfranchises legitimate minority communities as suspected Sybil clusters, whereas lowering ϕ_{crit} allows synthetic swarms to penetrate across g attack edges.

2. **Long-horizon social engineering and deceptive sleeper personas (Cheng et al. [14], Liang and Xu [26]).** Equation (3) assumes that acquiring cross-cluster attack edges g is expensive because humans only vouch for peers they know. However, frontier models sustain strategic deception in multi-agent bargaining [26] and exploit sycophantic rapport to manipulate human trust [4, 14]. If an adversary deploys patient "AI companion" personas across k different community forums over several months before requesting PoP vouches, g_{eff} can grow organically unless voucher eligibility requires physical-proximity co-presence handshakes (e.g., local NFC/ultrawideband key exchange) rather than text-only online interaction.

3. **Cold-start exclusion of peripheral or dissenting humans.** When community conductance falls or newcomers lack multiple independent connections, peripheral humans who have fewer than $k = 4$ spectrally distinct acquaintances face delayed or reduced admission weight (1.00%–1.09% exclusion at $\hat{\phi}_H \leq 0.069$). In polarized communities, majority cliques may also refuse to vouch for ideological minorities, requiring institutional fallback stewards (libraries,

credit unions, or municipal registries) bound by non-discrimination audits [1].

4. **Post-issuance delegation and correlated voting blocks (Farr et al. [19], Messias and Ide [28], Tchuente [35]).** Even when each credential π_{PoP} belongs to a unique human, humans may delegate voting execution to homogeneous LLM advisors that herd into correlated epistemic cascades [6, 19] or concentrate power in unaccountable DAO delegates [28, 35]. Thus, PoP solves *cardinality* (one human, one credential) but must be paired with correlation-discounted plural voting [3, 8, 29] to prevent post-issuance herding.

7 Scope constraints and topological limitations

- **Synthetic modular graph priors.** While our simulation explicitly executes power-iteration random walks on 560-node clustered graphs, live civic deployments span 10^4–10^7 nodes with heavy-tailed degree distributions and temporal churn that require distributed gossip estimators.
- **Privacy–accuracy trade-off under Local Differential Privacy.** Applying Laplace noise $\text{Lap}(\Delta_\phi/\epsilon)$ to local conductance estimates [2, 20] protects neighborhood edges from surveillance, but strict privacy budgets ($\epsilon \leq 0.5$) increase gate variance on low-degree nodes ($d_i < 6$).
- **Lack of longitudinal red-team data against 2026-class swarms.** Full integration of k-disjoint cluster vouching, VERA edge revocation [27], and Longfellow-ZK credentials [22] has not yet been field-tested against state-sponsored agent swarms over multi-month election cycles.

8 Pre-registerable falsification protocol

Table 5 defines six pre-registerable empirical tests that would falsify the security, inclusion, or privacy claims of Spectral-Threshold-PoP.

Table 5: Six pre-registerable falsification protocols for Spectral-Threshold Proof-of-Personhood.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Swarm-size invariance (Eq. 3)	Aggregate Sybil vote share W_S/W_H is invariant to synthetic swarm size S for fixed attack edges g	Fix $H = 5{,}000$, $g = 50$, $\phi_H = 0.25$; scale synthetic swarm $S \in \{500, 5{,}000, 50{,}000, 500{,}000\}$ with arbitrary internal topology	W_S/W_H increases by > 1.5 percentage points as S scales across three orders of magnitude
P2. Disjoint-quorum collusion resistance	Requiring $k = 4$ spectrally disjoint vouchers prevents single-cluster clique inflation	Compromise an entire 30-node single-cluster clique; measure maximum counterfeit PHCs minted before VERA revocation [27]	Compromised single-cluster clique mints > 2 unslashed counterfeit epoch credentials
P3. Real-world slow-mixing benchmark	Community bridge incentives maintain $\phi_H > \phi_H^* = 0.10$ on empirical social graphs [30]	Replay Spectral-Threshold-PoP on 5 empirical OSN topologies from Mohaisen et al. [30] and Cao et al. [11]	Honest participant exclusion exceeds 10.0% or Sybil capture at $g = 0.05H$ exceeds 15.0%
P4. Long-horizon AI companion red-team	Physical co-presence handshake requirement blocks remote LLM persona social engineering	60-day red-team where 200 autonomous personas attempt to solicit vouches from consented human testers [14]	Remote synthetic personas acquire $\geq 1\%$ of total valid epoch PHCs
P5. Local DP topology reconstruction defense	Local differential privacy ($\epsilon = 1.0$) prevents adversarial social-graph deanonymization [2, 20]	Provide an adversary with all published LNDP conductance sketches and π_{PoP} proofs; run link-prediction attack	Adversary reconstructs private human-to-human vouching edges with precision/recall AUC > 0.60
P6. Strategy-proof PB & BFT stability	Core Up-To-One PB [34] + HotStuff BFT [38] resist splitting and bound messages to $O(n)$	Run 300 participatory budgeting and BFT consensus rounds with 10% strategic splitters across $n \in \{32, \ldots, 256\}$	Strategic splitting shifts allocative outcome by $> 3\%$ of budget or BFT view-change exceeds $10n$ messages

9 Conclusion

As autonomous language model swarms render behavioral Turing tests obsolete and linearize token-weighted Quadratic Voting [5], digital institutions require Sybil resistance that neither surrenders biological privacy to centralized biometric registries [1] nor ignores the slow-mixing topology of real human communities [30]. By synthesizing classical power-iteration random-walk defenses [11, 39, 40] with capacity-bounded seed-anchor voucher quorums, local triangle support, authority-preserving edge revocation [27], and locally differentially private conductance gating [2, 20], label-blind Spectral-Threshold-PoP bounds simulated Sybil capture at $g = 64$ to 0.00% [0.00, 1.88] (reducing mean Sybil weight share from 4.81% to 1.46%) while preserving 99.99% [99.98, 99.99] honest inclusion. At the same time, our break-even sensitivity analysis demonstrates that graph-based personhood is bounded by honest community expansion ($\hat{\phi}_H^* \approx 0.07$–$0.10$) and degrades under half-network bribery (32.50% [26.39, 39.27] capture at $g = 160$), making cross-community bridge incentives and physical co-presence vouching essential for real-world deployment.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Steven Adler, Zoë Hitzig, Shrey Jain, Cassidy Brewer, Wayne Chang, Renee DiResta, et al. Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online. *arXiv preprint arXiv:2408.07892*, 2024. URL `https://arxiv.org/abs/2408.07892`.

[2] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Local differential privacy on metric spaces: Optimizing the trade-off with utility. IEEE Computer Security Foundations Symposium (CSF), pages 262–276, 2018. URL `https://doi.org/10.1109/CSF.2018.00026`.

[3] Zhiyu An and Wan Du. Open problems in differentiable social choice, 2026. URL `https://arxiv.org/abs/2602.03003`.

[4] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL `https://arxiv.org/abs/2610.09371`.

[5] Austin Bennett, Preston Vandervoort, Duc Le, and Mira Belenkiy. Concave is the new linear: The impossibility of anti-plutocratic DAO governance. arXiv preprint arXiv:2605.18990, 2026. URL `https://arxiv.org/abs/2605.18990`.

[6] Blaž Bertalanič and Carolina Fortuna. The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. arXiv preprint arXiv:2605.00914, 2026. URL `https://arxiv.org/abs/2605.00914`.

[7] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL `https://arxiv.org/abs/2610.35835`.

[8] Jacob Brodkey, Roberto Tamez, and Aaron Roth. Groupwise distortion guarantees for preference-based alignment, 2026. URL `https://arxiv.org/abs/2610.05450`.

[9] Vitalik Buterin, Zoë Hitzig, and E. Glen Weyl. A flexible design for funding public goods. Management Science, 65(11):5171–5187, 2019. URL `https://doi.org/10.1287/mnsc.2019.3337`.

[10] Jan Camenisch and Anna Lysyanskaya. An efficient system for non-transferable anonymous credentials with optional anonymity revocation. Advances in Cryptology — EUROCRYPT, pages 93–118, 2001. URL `https://doi.org/10.1007/3-540-44987-6_7`.

[11] Qiang Cao, Michael Sirivianos, Xiaowei Yang, and Tiago Pregueiro. Aiding the detection of fake accounts in large scale social online services (SybilRank). USENIX Symposium on Networked Systems Design and Implementation (NSDI), pages 197–210, 2012. URL https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/cao.

[12] Miguel Castro and Barbara Liskov. Practical Byzantine fault tolerance. In *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pages 173–186, 1999.

[13] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, et al. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2503.13657.

[14] Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), 2026. doi: 10.1126/science.aec8352.

[15] Alessandro Chiesa, Ziyi Guan, and Eylon Yıldız. Can AI oversight be zero knowledge? IACR Cryptology ePrint Archive, Paper 2026/2299, 2026. URL https://eprint.iacr.org/2026/2299.

[16] George Danezis and Prateek Mittal. SybilInfer: Detecting Sybil nodes using social networks. In *Network and Distributed System Security Symposium (NDSS)*, 2009. URL https://www.ndss-symposium.org/ndss2009/sybilinfer-detecting-sybil-nodes-using-social-networks/.

[17] John R. Douceur. The Sybil attack. In *International Workshop on Peer-to-Peer Systems (IPTPS)*, pages 251–260, 2002. URL https://doi.org/10.1007/3-540-45748-8_24.

[18] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL https://arxiv.org/abs/2609.34017.

[19] David Farr, Iain Cruickshank, Kate Starbird, and Jevin West. The cost of consensus: Malignant epistemic herding and adaptive gating in distributed multi-agent search. arXiv preprint arXiv:2605.06988, 2026. URL https://arxiv.org/abs/2605.06988.

[20] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Counting on dummies: Privacy amplification with observable counts and truncated batches. arXiv preprint arXiv:2610.42139, 2026. URL https://arxiv.org/abs/2610.42139.

[21] Daniel Fett, Kristina Yasuda, and Brian Campbell. Selective disclosure for JWTs (SD-JWT). IETF RFC 9901, 2025. URL https://datatracker.ietf.org/doc/rfc9901/.

[22] Matteo Frigo and abhi shelat. Anonymous credentials from ECDSA (Longfellow-ZK). IACR Communications in Cryptology, 1(2), 2024. URL https://eprint.iacr.org/2024/2010.

[23] Martin Kleppmann. Making CRDTs Byzantine fault tolerant. Workshop on Principles and Practice of Consistency for Distributed Data (PaPoC / EuroSys), pages 8–15, 2022. URL https://doi.org/10.1145/3517209.3524042.

[24] Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long tasks. arXiv preprint arXiv:2503.14499, 2025. URL https://arxiv.org/abs/2503.14499.

[25] Leslie Lamport, Robert Shostak, and Marshall Pease. The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982. URL https://doi.org/10.1145/357172.357176.

[26] Chen Liang and Fasheng Xu. When LLM agents negotiate: Private information and dynamic bargaining in supply chains. arXiv preprint arXiv:2608.07538, 2026. URL https://arxiv.org/abs/2608.07538.

[27] Lifei Liu and Haoran Yu. VERA: Authority-preserving edge revocation for federated AI-agent workflows. arXiv preprint arXiv:2608.30091, 2026. URL https://arxiv.org/abs/2608.30091.

[28] Johnnatan Messias and Ayae Ide. Fairness in token delegation: Mitigating voting power concentration in DAOs. arXiv preprint arXiv:2510.05830, 2025. URL https://arxiv.org/abs/2510.05830.

[29] Joel Miller, E. Glen Weyl, and Leon Erichsen. Beyond collusion resistance: Leveraging social information for plural funding and voting (COCM). SSRN Working Paper 4311507, 2022. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4311507.

[30] Abedelaziz Mohaisen, Aaram Yun, and Yongdae Kim. Measuring the mixing time of social graphs. ACM Internet Measurement Conference (IMC), pages 383–389, 2010. URL https://doi.org/10.1145/1879141.1879191.

[31] Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2402.07510.

[32] Qi Pang, Virginia Smith, and Wenting Zheng. Codetta: High-capacity, keyless, and undetectable multi-agent collusion. IACR ePrint 2026/2218 / arXiv:2609.28900, 2026. URL https://arxiv.org/abs/2609.28900.

[33] Eduardo Perez-Richet and Vasiliki Skreta. Fraud-proof non-market allocation mechanisms. CEPR Discussion Paper DP18466 / Journal of Political Economy, 2023. URL https://cepr.org/publications/dp18466.

[34] Piotr Skowron. Core up-to-one for participatory budgeting with additive utilities. arXiv preprint arXiv:2609.15928, 2026. URL https://arxiv.org/abs/2609.15928.

[35] Guy Tchuente. Monitoring limits in DAO governance: Capacity breakpoints and endogenous concentration. arXiv preprint arXiv:2603.11222, 2026. URL https://arxiv.org/abs/2603.11222.

[36] Yutian Wang and Luyao Zhang. Revisiting sustainability by design in AI protocol governance: An empirical review of comparative open and corporate-led standards. arXiv preprint arXiv:2606.05667, 2026. URL https://arxiv.org/abs/2606.05667.

[37] E. Glen Weyl, Puja Ohlhaver, and Vitalik Buterin. Decentralized society: Finding Web3's soul. SSRN Working Paper 4105763, 2022. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4105763.

[38] Maofan Yin, Dahlia Malkhi, Michael K. Reiter, Guy Golan Gueta, and Ittai Abraham. HotStuff: BFT consensus with linearity and responsiveness. In *ACM Symposium on Principles of Distributed Computing (PODC)*, pages 347–356, 2019. URL https://doi.org/10.1145/3293611.3331591.

[39] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. SybilGuard: Defending against sybil attacks via social networks. ACM SIGCOMM, pages 267–278, 2006. URL https://doi.org/10.1145/1159913.1159945.

[40] Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. SybilLimit: A near-optimal social network defense against sybil attacks. IEEE Symposium on Security and Privacy (S&P), pages 3–17, 2008. URL https://doi.org/10.1109/SP.2008.13.

[41] Jay Yu, Austin Bennett, Billy Gao, and Rebecca Joseph. Evaluating voting design vulnerabilities for retroactive funding. arXiv preprint arXiv:2505.16068, 2025. URL https://arxiv.org/abs/2505.16068.