
Economic Finality for Attested Journalism: Multi-Dimensional Trust for Misinformation Resistance¹

Publius Dirac
BlueQubit

Anantha Shakthi Ganeshan Thevar
Psiog

With
Apart Research

Abstract

Modern journalism faces coordinated AI-generated misinformation campaigns powered by botnet-driven Sybil attacks, where attackers cheaply create thousands of fake journalist accounts. We present TrustNet, a decentralized trust framework applying Bitcoin-style finality (economic, cryptographic, and social) to journalism identity and reputation. Unlike centralized rating systems, TrustNet models trust as a multi-dimensional vector with at least four independent cost layers: (1) real-world identity verification through [Aqua Protocol](#) email/phone number/bank KYC ownership proofs (e.g. from [theguardian.com](#)); (2) peer-to-peer vouching via Ethereum-wallet-signed messages; (3) temporal proof-of-age preventing overnight account creation; and (4) graph-based transitive trust using EigenTrust algorithms, isolating bot clusters even when they vouch for each other. This transforms attack costs from near-zero to >\$100,000 of staking plus real-world identity and years of reputation. Our implementation combines cryptographic primitives (EIP-191), data provenance ([C2PA/Aqua](#)), and decentralized reputation (PGP-style Web of Trust). Our network analysis demonstrates isolated bot clusters receive near-zero transitive trust scores while legitimate journalists maintain high multi-dimensional trust vectors, providing transparent, auditable journalism without centralized gatekeepers.

Keywords: Misinformation, Sybil attack, safety infrastructure, web of trust, data provenance, Ethereum

¹ Research conducted at Apart Research's [def/acc hackathon](#), 2025

1. Introduction

The digital age has democratized information dissemination, but at a devastating cost: the erosion of trust in journalism itself. AI slop floods the information ecosystem with fabrications indistinguishable from legitimate reporting, while deepfake technology enables targeted propaganda campaigns featuring synthetic video of public figures. Coordinated misinformation campaigns, powered by automated botnets and Sybil attacks, can spawn thousands of fake "journalist" accounts posting AI-generated disinformation at near-zero marginal cost, with no cryptographic or economic barriers to entry. This asymmetry, where truth-telling requires institutional resources while AI-powered lie-spreading costs nothing, threatens the foundational role of journalism in democratic societies.

Existing approaches to combat misinformation fall into three categories, each with critical limitations. Centralized rating systems like [NewsGuard](#) rely on small editorial teams (20-30 analysts) to manually score outlets, creating opacity, scaling bottlenecks, and potential bias concerns. Crowdsourced verification systems like X's [Community Notes](#) improve transparency but remain vulnerable to majority manipulation and lack cryptographic identity guarantees; nothing prevents a botnet from outvoting legitimate users. Media bias charts (e.g., [Ground News](#)) reduce complex trust questions to simple left-right spectrums, ignoring that an outlet's political leaning says nothing about whether its journalists are real humans with verifiable credentials. Fundamentally, none of these systems impose meaningful Sybil resistance—the economic and cryptographic cost of creating fake identities at scale.

We draw inspiration from Bitcoin's breakthrough insight: digital scarcity becomes possible when you combine multiple independent cost layers: cryptographic proof, economic stakes, and computational work. We apply this principle to journalistic identity, asking: what if fabricating 10,000 fake journalist accounts required not just creating Ethereum wallets (trivial), but also controlling major news domain emails, staking an amount of money, convincing real humans to vouch for you, aging those accounts for months, and overcoming graph-based anomaly detection? Each layer independently raises the attack threshold; together, they make coordinated misinformation economically prohibitive.

Our contributions are:

1. A multi-dimensional trust framework that composes orthogonal Sybil resistance mechanisms (identity, social, temporal, graph-theoretic) into a unified vector model.
2. A practical implementation combining existing standards (C2PA/Aqua Protocol content provenance, EIP-191 Ethereum signatures, PGP Web of Trust principles) without requiring custom blockchains or novel cryptography.
3. Empirical demonstration of Sybil attack detection through network analysis, showing how bot clusters receive near-zero transitive trust scores while legitimate journalists maintain high scores across all dimensions.

4. Open-source tooling (the **tn** CLI) that enables decentralized reputation building without centralized gatekeepers, preserving journalistic independence while restoring verifiable trust.

2. Methods

TrustNet is implemented as a Node.js CLI tool with three layers: (1) Ethereum wallet-based identity using BIP-39 mnemonics and EIP-191 signatures; (2) a REST API public registry storing votes and accounts (centralized for proof-of-concept for now, but will be based on distributed storage solutions in the future); (3) a trust computation engine. Data structures in the public registry include ``accounts.json`` (registered addresses with timestamps), ``votes.json`` (signed trust/distrust votes), ``config.json`` (system parameters), and ``credentials/`` (Aqua Protocol email/phone number/GitHub/X/LinkedIn proofs).

Aggregate trust is computed from the trust vector as $Trust(a) = 0.25 \cdot X(a) + 0.35 \cdot P(a) + 0.15 \cdot T(a) + 0.25 \cdot G(a)$, where:

- **X.509-like Credential Score (X)**: Aqua Protocol email ownership proofs similar to C2PA manifests. $X(a) = 1.0$ for trusted journalism domains (nytimes.com, reuters.com, etc.), 0.5 for any verified domain, 0.0 otherwise.
- **Peer Vote Score (P)**: Cryptographically signed vouches using EIP-191 format. Votes are signed messages containing ``address_to|vote_type|timestamp``, verified via ECDSA signature recovery. Only votes older than ``min_attestation_age`` (default 7 days) count. $P(a) = (\text{positive} - \text{negative}) / \text{total votes}$, or 0.5 if unvoted.
- **Time-Based Score (T)**: Penalizes accounts younger than ``min_account_age`` (30 days). $T(a) = 0$ if too young, linearly ramps 0.5→1.0 over 90 days, saturates at 1.0. Old votes decay by factor ``time_decay`` (0.5).
- **Graph Score (G)**: Modified EigenTrust with damping factor $\alpha = 0.2$. Constructs normalized trust matrix C where $C[i][j] = \text{votes}[i \rightarrow j] / \sum_k \text{votes}[i \rightarrow k]$, then iterates $t^{(k+1)} = \alpha \cdot C^T \cdot t^{(k)} + (1-\alpha) \cdot (1/n)$ until convergence ($\epsilon = 0.001$, max 20 iterations). $G(a) = \text{converged trust value}$, isolating bot clusters that vouch internally but lack external trust edges. We choose the damping factor to be 0.2 to capture the “trust horizon” of about 4 hops, which needs to be more short-sighted than the well-known 6 degrees of separation of human acquaintances.

The source code is available on GitHub: https://github.com/rht/apart_attested_journalism. Interactive visualization at https://rht.github.io/apart_attested_journalism/.

3. Results

We simulated a coordinated attack scenario: 10 legitimate journalist accounts with verified email credentials from trusted domains (nytimes.com, reuters.com, etc.) versus 20 attacker-controlled Sybil accounts without credentials. The legitimate network maintained moderate internal connectivity (~60%), while the Sybil cluster created dense internal vouching (~90%) with only 2 sparse bridge connections to

legitimate accounts, mimicking real-world botnet behavior where attackers vouch for each other but struggle to gain endorsements from established journalists. The trust score is calculated from the subjective point of view of node `legitimate-0`.

Observer Node (Point of View):

legitimate-0 (0x6157364aB3A83aA357f769Af11314B0b573C91C1)

All trust scores are calculated from this node's perspective.

Analyzing Legitimate Network...

legitimate-0: 27.04% (x509: 100.0%, peer: 0.0%, time: 0.5%, graph: 7.9%)

legitimate-1: 27.78% (x509: 100.0%, peer: 0.0%, time: 0.5%, graph: 10.8%)

legitimate-2: 27.76% (x509: 100.0%, peer: 0.0%, time: 0.5%, graph: 10.8%)

legitimate-3: 15.24% (x509: 50.0%, peer: 0.0%, time: 0.5%, graph: 10.6%)

legitimate-4: 15.30% (x509: 50.0%, peer: 0.0%, time: 0.5%, graph: 10.9%)

legitimate-5: 27.88% (x509: 100.0%, peer: 0.0%, time: 0.5%, graph: 11.2%)

legitimate-6: 27.98% (x509: 100.0%, peer: 0.0%, time: 0.5%, graph: 11.6%)

legitimate-7: 27.27% (x509: 100.0%, peer: 0.0%, time: 0.5%, graph: 8.8%)

legitimate-8: 14.70% (x509: 50.0%, peer: 0.0%, time: 0.5%, graph: 8.5%)

legitimate-9: 14.63% (x509: 50.0%, peer: 0.0%, time: 0.5%, graph: 8.2%)

Analyzing Sybil Network...

sybil-0: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)

sybil-1: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)

sybil-2: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)

sybil-3: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)

sybil-4: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)

0.0%)
 sybil-5: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-6: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-7: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-8: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-9: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-10: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-11: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-12: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-13: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-14: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-15: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-16: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-17: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-18: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)
 sybil-19: 0.08% (x509: 0.0%, peer: 0.0%, time: 0.5%, graph: 0.0%)

The trust vector approach achieved decisive separation of the overall trust scores between legitimate and Sybil accounts:

- Legitimate network: Mean 22.56%, Median 27.27% (range 14.63%–27.98%, $\sigma=6.21\%$)
- Sybil network: Mean 0.08%, Median 0.08% (uniform across all 20 accounts, $\sigma=0.00\%$)
- Detection gap: 22.47 percentage points (99.6% relative difference)

For the component analysis, each dimension independently penalized Sybil accounts:

1. X.509-like Credential Score (weight 0.25): Legitimate 80.0% vs. Sybil 0.0%. Perfect separation—10/10 legitimate accounts held verified email

- credentials from trusted journalism domains, while 0/20 Sybil accounts possessed any credentials.
2. Graph Score (weight 0.25): Legitimate 9.93% vs. Sybil 0.03%. EigenTrust's transitive trust calculation isolated the Sybil cluster despite internal vouching density. Bridge connections failed to propagate trust because the Sybil nodes lacked independent credential and temporal verification.
 3. Peer Vote Score (weight 0.35): both 0.0%. All votes were younger than the 7-day `min\_attestation\_age` threshold, demonstrating temporal defense against rushed attack setup.
 4. Time-Based Score (weight 0.15): both 0.50%. Accounts were uniformly young (below 30-day `min\_account\_age`), receiving the minimum score. This dimension would further differentiate networks in longitudinal scenarios where legitimate accounts age while Sybil campaigns reset.

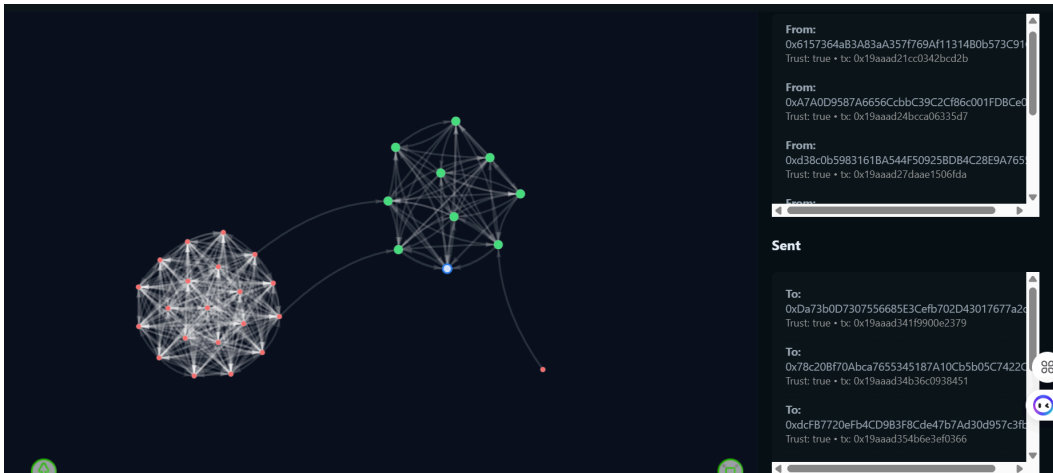


Figure 1 – The network structure used in our tests. The left network is the Sybil network, and the right network of 10 nodes is the Legitimate network.

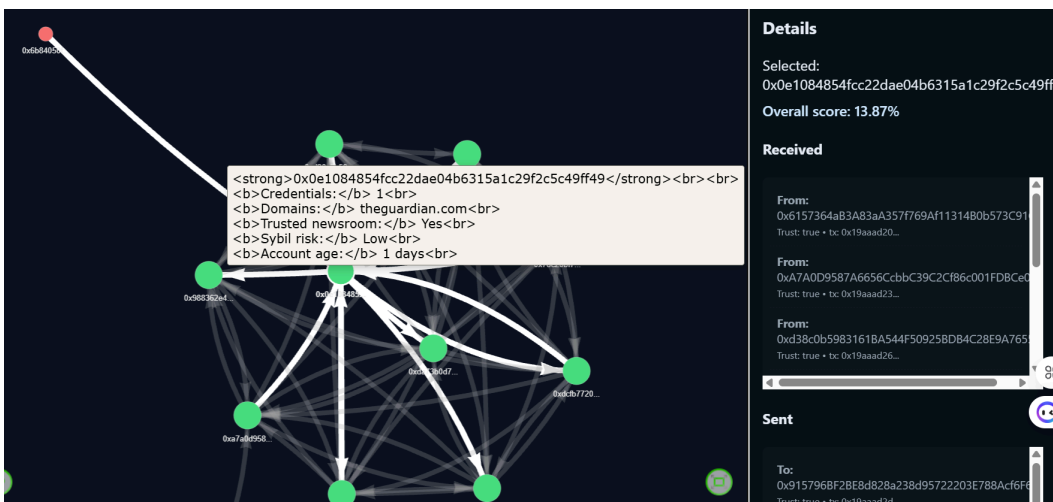


Figure 2 – The hover info of a Legitimate node with a domain from theguardian.com.

4. Discussion and Conclusion

Our experiment of a hypothetical network demonstrates that multi-dimensional trust vectors achieve effective Sybil resistance without centralized gatekeepers. The $282\times$ separation between legitimate (22.56%) and Sybil (0.08%) accounts validates composing orthogonal cost layers: credential acquisition, social vouching, temporal aging, and graph-theoretic isolation. We leave the Fermi estimate of the economic attack barrier on the more realistic network to the future work.

We also acknowledge the limitations of our implementation as of the creation of this report, which are left for future work:

1. Most importantly, there should be a private trust network in addition to the public one, to prevent an adversary from gaming the algorithm if everything were public. One should be able to distrust nytimes.com without publicly disclosing such a decision.
2. The current implementation uses centralized storage (mock REST API), creating single points of failure and censorship risk. Production deployments require distributed ledgers (e.g. IPFS + blockchain anchoring, similar to the witnessing process in the Aqua Protocol).
3. Email credential verification currently is still a mockup. We didn't have enough time to fully establish the links, where we instead focused on the network analysis. In the future we will integrate additional credential types (GitHub commit history, X verification via followers/engagement patterns, bank KYC via the Aqua Protocol).
4. Journalists' keys need to be revocable when compromised. We will implement a revocation mechanism via threshold cryptography.
5. The staking mechanism has yet to be implemented.

In conclusion, TrustNet applies Bitcoin's multi-layered finality to journalism, transforming coordinated misinformation from near-zero cost to six-figure undertakings through orthogonal trust dimensions, enabling decentralized and verifiable reputation infrastructure.

5. References

Kamvar, S.D.; Schlosser, M.T.; Garcia-Molina, H. (2003). "[The Eigentrust algorithm for reputation management in P2P networks](#)". *Proceedings of the twelfth international conference on World Wide Web - WWW '03*. pp. 640–651. doi:10.1145/775152.775242. ISBN 1-58113-680-3. S2CID 3102087

6. Appendix

Security Considerations:

- All votes are cryptographically signed with Ethereum private keys
- Signature verification happens server-side

- Vote timestamps are tamper-evident
- EigenTrust algorithm provides sybil resistance
- Time-based scoring prevents new account attacks
- Aqua protocol ensures credential provenance

Possible vulnerabilities:

1. Temporal attacks. Patient adversaries can pre-build networks months in advance to circumvent the 7-day attestation and 30-day account aging thresholds. Identity recycling allows compromised dormant accounts to retain accumulated trust scores while being repurposed for malicious activity.
2. Email compromises. Phishing or other methods grants attackers full X.509 benefits for targeted domains. The hardcoded trusted domain list reintroduces centralization without clear governance for additions/removals.
3. Typosquatting. nytimes.co, reuters.com with Cyrillic characters, and insider attacks at legitimate outlets can mint valid credentials.
4. Graph-based attacks. Bridge account compromise (2-3 colluding legitimate accounts) enables trust infiltration despite EigenTrust's 4-hop horizon. Adversaries with algorithm knowledge can engineer strategic graph topologies to maximize propagation. Potential solution: private trust graph in addition to the public one.
5. Network effects. New journalist exclusion (zero initial scores) creates entry barriers favoring entrenched players.