
TRUSTFLOW: TOPIC-AWARE VECTOR REPUTATION PROPAGATION FOR MULTI-AGENT ECOSYSTEMS

A PREPRINT

Volodymyr Seliuchenko

robotler.ai

2026-03-01

ABSTRACT

We introduce TrustFlow, a reputation propagation algorithm that assigns each software agent a multi-dimensional reputation vector rather than a scalar score. Reputation is propagated through an interaction graph via topic-gated transfer operators that modulate each edge by its content embedding, with convergence to a unique fixed point guaranteed by the contraction mapping theorem. We develop a family of Lipschitz-1 transfer operators and composable information-theoretic gates that achieve up to 98% multi-label Precision@5 on dense graphs and 78% on sparse ones. On a benchmark of 50 agents across 8 domains, TrustFlow resists sybil attacks, reputation laundering, and vote rings with at most 4 percentage-point precision impact. Unlike PageRank and Topic-Sensitive PageRank, TrustFlow produces vector reputation that is directly queryable by dot product in the same embedding space as user queries.

Keywords reputation • trust propagation • multi-agent systems • PageRank • topic-sensitive • embedding space

1 Introduction

The emergence of autonomous software agents that interact, delegate tasks, and transact on behalf of users creates a fundamental trust problem: how does a user—or another agent—select the right agent for a task from a marketplace of millions? Traditional approaches (star ratings, manual curation, call counts) fail at scale because they are easily gamed, domain-agnostic, and cannot propagate transitive trust.

PageRank (Brin and Page 1998) demonstrated that the link structure of the web encodes a useful notion of importance: a page is important if important pages link to it. Topic-Sensitive PageRank (TSPR) (Haveliwala 2002) extended this insight by computing multiple PageRank scores, each biased toward a different topic, to yield query-sensitive importance scores. Both algorithms operate on a single, topic-independent transition matrix M and produce scalar rankings.

Agent reputation, however, has structural properties absent from web-page ranking that demand a fundamentally different formulation:

1. **Multi-domain expertise.** An agent may be expert in both medicine and data science. A scalar reputation score cannot represent this; a vector can.
2. **Content–competence decoupling.** Web pages are inherently content-bearing—text and links *are* the signal. Agents may be pure service providers; trust must derive from interaction history, not static content.
3. **Economic signals.** When agent A delegates a paid task to agent B , the economic commitment is a trust signal absent from hyperlink graphs.
4. **Adversarial moderation.** Malicious agents may attempt sybil attacks, reputation laundering, or vote rings. The system must resist these while maintaining fairness.
5. **Directional discovery.** Users search for agents by natural-language queries. Reputation should be queryable in the same vector space as queries, enabling a single dot-product retrieval.

TrustFlow addresses all five through a unified framework of vector reputation propagation with topic-gated trust transfer. Our contributions:

1. **Topic-aware vector reputation propagation** in both discrete ($N \times D$ matrix) and continuous ($N \times E$ embedding) formulations, with convergence guarantees via contraction mapping (§3).
2. **A family of Lipschitz-1 transfer operators**—projection, squared gating, scalar-gated, and hybrid—that modulate reputation flow by interaction content, with composable information-theoretic gates (KL-divergence, entropy, confidence) and empirically characterized precision–information tradeoffs (§3.3).
3. **Negative trust edges** for moderation flags with joint convergence guarantees (§3.6).
4. **Comprehensive evaluation** demonstrating structural resilience to four attack classes, with ≤ 4 pp P@5 impact across all transfer operators (§5).

2 Preliminaries

We briefly review the mathematical foundations on which TrustFlow builds.

2.1 PageRank

Let N be the number of pages on the web and let M be the $N \times N$ row-stochastic transition matrix of the web graph, where $M_{ij} = 1/\text{outdeg}(i)$ if page i links to page j . The PageRank vector \mathbf{r} is the stationary distribution of the random walk:

$$\mathbf{r} = \alpha M^T \mathbf{r} + (1 - \alpha) \mathbf{v} \tag{1}$$

where $\alpha \in (0, 1)$ is the damping factor and \mathbf{v} is the teleportation vector (uniform in standard PageRank). The teleportation term ensures ergodicity—stochasticity, irreducibility, and aperiodicity—guaranteeing convergence to a unique fixed point (Langville and Meyer 2006). Each entry r_i is a scalar importance score.

2.2 Topic-Sensitive PageRank

Haveliwala (2002) observed that a single PageRank vector cannot capture topic-dependent importance. TSPR computes K biased PageRank vectors $\{\mathbf{r}^{(k)}\}_{k=1}^K$, each using a topic-specific teleportation vector $\mathbf{v}^{(k)}$ that concentrates probability mass on pages known to belong to topic k :

$$\mathbf{r}^{(k)} = \alpha M^T \mathbf{r}^{(k)} + (1 - \alpha) \mathbf{v}^{(k)} \tag{2}$$

At query time, the query is classified into topics and the final score is a linear combination $\sum_k p(k | q) r_i^{(k)}$. Crucially, the transition matrix M is the *same* for all topics—TSPR cannot distinguish “A links to B in a medical context” from “A links to B in a coding context.”

2.3 Notation

Table 1: Notation summary.

Symbol	Definition
N	Number of agents
E	Embedding dimensionality (e.g., 384 for E5)
D	Number of discrete domains
$R[i] \in \mathbb{R}^E$	Reputation vector of agent i
$e_{ij} \in \mathbb{R}^E$	Unit interaction embedding for edge $i \rightarrow j$
$T[j] \in \mathbb{R}^E$	Teleportation prior of agent j
$C[j] \in \mathbb{R}^E$	Exogenous authority injection for agent j
α	Damping factor (default 0.85)
$w(i \rightarrow j)$	Row-normalized edge weight
M_d	Domain-conditioned transition matrix for domain d
\odot	Element-wise (Hadamard) product

3 The TrustFlow Algorithm

3.1 Overview

TrustFlow generalizes PageRank along two axes:

- **Scalar** \rightarrow **vector**. Each agent’s reputation is a vector $R[i] \in \mathbb{R}^E$, not a scalar. The direction encodes the agent’s expertise profile; the magnitude encodes accumulated trust.
- **Single transition matrix** \rightarrow **topic-gated transfer**. Rather than TSPR’s approach of biasing the teleportation vector while sharing a single M , TrustFlow gates each edge’s reputation transfer by the interaction’s *content embedding*, producing topic-dependent transfer without requiring pre-defined topic categories.

We present the continuous formulation first, then the discrete specialization.

3.2 Continuous Formulation

Each agent i has a reputation vector $R[i] \in \mathbb{R}^E$ that lives in the same embedding space as interaction content and discovery queries. Each directed edge ($i \rightarrow j$) carries a unit interaction embedding $e_{ij} \in \mathbb{R}^E$ derived from the content of their interaction. The core iteration is:

$$R_{new}[j] = \alpha \sum_i w(i \rightarrow j) f(R[i], e_{ij}) + (1 - \alpha) T[j] + C[j] \quad (3)$$

where:

- $f(R[i], e_{ij})$ is a **topic-gated transfer operator** that modulates the sender’s reputation by the interaction embedding. We study a family of such operators (§3.3)—projection, element-wise squared gating, Hadamard relu, scalar-gated, and hybrid—each satisfying a Lipschitz-1 bound that guarantees convergence.
- $w(i \rightarrow j)$ is the row-normalized edge weight. Raw weights incorporate interaction frequency and optional payment delegation multipliers (e.g., $\mu = 3$ for economically backed edges); row normalization is applied afterward so that $\sum_j w(i \rightarrow j) = 1$ for each sender i , preserving the contraction guarantee regardless of multiplier magnitude.
- $T[j] \in \mathbb{R}^E$ is the **teleportation prior**, playing the same structural role as \mathbf{v} in Equation 1: scaled by $(1 - \alpha)$, it ensures ergodicity (stochasticity, irreducibility, aperiodicity) exactly as in PageRank’s Google Matrix. $T[j]$ is derived from agent j ’s public content embeddings, weighted by engagement and quality scores.
- $C[j] \in \mathbb{R}^E$ is the **exogenous authority injection**, an additive term that absorbs external reputation signals (content engagement, web authority, economic intent, cross-platform reputation) into the embedding space. Because C is not scaled by $(1 - \alpha)$, the operator can tune exogenous signal strength independently of the damping factor; an alternative formulation couples C under $(1 - \alpha)$ alongside T when exogenous signals should naturally diminish as graph evidence accumulates.

The iteration begins with $R_0[j] = T[j] + C[j]$ and converges to a unique fixed point by the contraction mapping theorem (§3.5).

Blind edges. Edges whose content is not available (e.g., encrypted API calls) are called *blind edges*. Because no interaction text can be embedded, blind edges carry less signal than labeled ones and are discounted (weighted at $0.3 \times$ in our experiments). Their quality can be improved with proxy embeddings: in this work we use $e_{ij} = \text{avg}(p_i, p_j)$, the mean of the two agents’ profile embeddings, which assumes the interaction concerns a topic between their areas of expertise. A learned model conditioned on richer caller–callee features (capability overlap, historical call patterns, task metadata) could predict a more accurate proxy and is a direction for future work. We evaluate the averaging proxy in §5.

Unnormalized reputation. We retain full magnitude during iteration: $R[i]$ is not L2-normalized between steps. An agent that accumulates reputation from many high-quality interactions in a given direction will have a larger component in that direction. The L2 norm $\|R[i]\|$ naturally reflects total accumulated reputation—analogueous to how PageRank propagates more authority from high-scoring nodes—while the direction $R[i]/\|R[i]\|$ encodes the expertise profile. An alternative, per-iteration normalization variant constrains all vectors to the unit sphere; we compare both in §5.3.

Comparison with PageRank and TSPR. Equation 3 generalizes Equation 1 in three ways: (i) scalar r_i becomes vector $R[i]$; (ii) the uniform transfer $M^T \mathbf{r}$ becomes the topic-gated transfer $\sum_i w \cdot f(R[i], e_{ij})$; and (iii) the teleportation \mathbf{v} is supplemented by an independent exogenous injection C . Unlike TSPR (Equation 2), which biases only the teleportation vector while retaining a single topic-independent M , TrustFlow makes every edge’s transfer depend on the interaction’s semantic content.

3.3 Transfer Operator Family

The transfer operator f in Equation 3 determines how reputation is modulated by the interaction embedding before propagation. We study five base operators, each satisfying a Lipschitz-1 bound (ensuring convergence, §3.5) but occupying a different point in the precision–information tradeoff space.

Projection: $f(R, e) = \sigma(R \cdot e) \cdot e$, where $\sigma = \max(0, \cdot)$. The output is always parallel to e_{ij} (rank 1), confining transfer strictly to the interaction topic direction. This provides maximum cross-domain isolation but collapses directional information on content-free (blind) edges ($\cos \approx 0.004$ for uniform e).

Squared gating: $f(R, e) = R \odot e^2$. Each dimension of the sender’s reputation is gated by the squared activation of the interaction embedding in that dimension. Because $e_k^2 \geq 0$, the gating is always non-negative, acting as a spectral filter. On blind edges ($e = \frac{1}{\sqrt{E}}\mathbf{1}$), the output is R/E —the sender’s full profile with perfect directional preservation ($\cos(\text{output}, R) = 1.0$).

Scalar-gated: $f(R, e) = \sigma(\hat{R} \cdot e) \cdot R$, where $\hat{R} = R/\|R\|$ and $\sigma(x) = \min(\max(0, x), 1)$ is a clamped relu. The clamp ensures $\sigma \in [0, 1]$, so $\|f(R_1, e) - f(R_2, e)\| \leq \|R_1 - R_2\|$, satisfying the Lipschitz-1 bound. A single go/no-go gate based on overall cosine alignment, transferring the full profile when the gate opens. Excels on dense labeled graphs but degrades on blind-heavy graphs where the gate closes.

Hadamard relu: $f(R, e) = \max(0, R \odot e)$. A variant of squared gating that applies element-wise relu instead of squaring. Unlike squared gating, relu clips the $\sim 52\%$ of dimensions where dense E5 embeddings have negative components, losing information.

Hybrid: Per-edge operator selection based on content availability (e.g., squared gating for content-free edges, projection for content-rich edges), or a convex interpolation $\gamma \cdot f_{\text{proj}} + (1 - \gamma) \cdot f_{\text{sq}}$.

Table 2: Transfer operator properties.

Operator	Output rank	Blind preservation	Cross-domain isolation	Convergence
Squared gating	Up to E	$\cos = 1.0$	Soft (per-dim filter)	Always
Projection	1	$\cos \approx 0$	Maximum	Always
Hadamard relu	Up to E	$\cos = 0.71$	Moderate (relu clips)	Always
Scalar-gated	Up to E	Full (when gate opens)	Binary	Always (σ clamped)
Hybrid	Up to E	Tunable	Tunable	Always

All operators are Lipschitz-1. For squared gating: $\|(R_1 - R_2) \odot e^2\| \leq \|R_1 - R_2\| \cdot \|e^2\|_\infty \leq \|R_1 - R_2\|$ since $\|e\|_2 = 1$ implies $|e_k| \leq 1$. For projection: projection onto a unit vector is nonexpansive. Combined with $\alpha < 1$, all variants converge.

KL-divergence gating. Any transfer operator can be composed with an information-theoretic gate that suppresses cross-domain reputation leakage. The gated transfer replaces $f(R[i], e_{ij})$ with $\text{gate}(i, e_{ij}) \cdot f(R[i], e_{ij})$, where:

$$\text{gate}(i, e_{ij}) = \exp(-\lambda \cdot D_{KL}(p_{\text{int}} \| p_{\text{rep}}[i])) \quad (4)$$

Here p_{int} is the interaction’s topic distribution and $p_{\text{rep}}[i]$ is agent i ’s reputation treated as a distribution. When the interaction aligns with the sender’s expertise ($D_{KL} \approx 0$), the gate ≈ 1 ; when off-topic, the gate decays exponentially. In the continuous formulation, an efficient cosine proxy avoids the softmax conversion: $\text{gate}(i, e_{ij}) = \exp(-\lambda \cdot \sin^2 \theta)$, where θ is the angle between $R[i]$ and e_{ij} . The gate is bounded in $[0, 1]$ and Lipschitz-continuous in R , preserving the contraction guarantee, though the composite contraction constant increases slightly, requiring more iterations to converge.

Additional gating mechanisms. The gating architecture is modular: KL-divergence is one instance of a family of composable gates, all bounded in $[0, 1]$ and convergence-preserving. Other members include: (i) an *entropy gate* $\exp(-\mu \cdot H(p_{\text{int}}))$ that suppresses unfocused or spam-like interactions with high topic entropy; (ii) a *magnitude-ratio gate* that measures the fraction of the sender’s reputation concentrated in the interaction direction, penalizing incidental transfers far from the sender’s core profile; and (iii) a *confidence gate* that discounts interactions where

the embedding model reports low certainty (particularly useful for blind edges). Gates compose multiplicatively— $\text{gate} = \text{gate}_{\text{KL}} \cdot \text{gate}_{\text{entropy}} \cdot \text{gate}_{\text{conf}}$ —providing layered defense against different forms of cross-domain leakage. We evaluate KL-divergence gating in §5.4; the remaining gates are left to future work.

3.4 Discrete Specialization

When D pre-defined domains are available (e.g., medicine, law, finance), each interaction is classified into top- k domains via soft assignment against domain centroids. This produces D separate transition matrices M_d , each row-stochastic, capturing the fraction of interactions between agents relevant to domain d . The per-domain iteration is:

$$R_{\text{new}}[:, d] = \alpha M_d^T R[:, d] + (1 - \alpha) T[:, d] + C[:, d] \quad (5)$$

This is the key structural difference from TSPR: where TSPR uses the *same* M for all topics and varies only the teleportation vector, TrustFlow uses *different* M_d per domain. When agent A calls agent B for a medical task, this edge appears with high weight in M_{med} and low weight in M_{code} . TSPR cannot make this distinction.

Each per-domain iteration is an independent PageRank with domain-specific structure, converging in $O(\log(1/\varepsilon)/\log(1/\alpha))$ iterations (approximately 11 for $\alpha = 0.85$, $\varepsilon = 10^{-4}$) (Langville and Meyer 2006).

Continuous vs. discrete. The discrete formulation requires a pre-defined domain taxonomy and cannot represent cross-domain expertise without dilution. A biostatistics agent must split its reputation mass across medicine and data-science bins. The continuous formulation eliminates this bottleneck: a 384-dimensional vector can simultaneously align with both domain centroids. The discrete formulation is a specialization of the continuous one—obtained by replacing the embedding space \mathbb{R}^E with a domain indicator space \mathbb{R}^D and the content-gated transfer with domain-conditioned matrices.

3.5 Convergence

Theorem 1 (Convergence). *For any Lipschitz-1 transfer operator f and damping factor $\alpha \in (0, 1)$, the TrustFlow iteration (Equation 3) is a contraction mapping with factor α and converges to a unique fixed point from any initialization.*

Proof sketch. For any two reputation states R_1, R_2 :

$$\|F(R_1) - F(R_2)\| = \alpha \left\| \sum_i w_i [f(R_1[i], e_i) - f(R_2[i], e_i)] \right\| \leq \alpha \sum_i w_i \|R_1[i] - R_2[i]\| \leq \alpha \|R_1 - R_2\|$$

The first inequality uses the Lipschitz-1 property of f ; the second uses the convexity of row-normalized weights ($\sum_i w_i \leq 1$). The teleportation and exogenous terms cancel in the difference. By Banach’s fixed-point theorem, a unique fixed point exists and is reached with linear convergence rate α . \square

Corollary 1 (Steady-state bound). *The converged reputation satisfies $\|R^*\| \leq \|T\| + \|C\|/(1 - \alpha)$.*

The bound follows from evaluating Equation 3 at the fixed point $R^* = F(R^*)$ and applying the triangle inequality. The teleportation prior T provides a bounded restart (the $(1 - \alpha)$ factor) while C provides an exogenous injection whose influence is amplified by $1/(1 - \alpha)$ at steady state—bounding the total reputation any agent can accumulate.

3.6 Negative Trust Edges

The core iteration can optionally be extended with negative trust edges for moderation. Moderation flags (spam, harmful, low-quality, malicious) create negative edges, and the iteration becomes:

$$R_{\text{new}}[:, d] = \alpha (M_{\text{pos}, d}^T R[:, d] - \beta M_{\text{neg}, d}^T R[:, d]) + (1 - \alpha) T[:, d] + C[:, d] \quad (6)$$

where $M_{\text{neg}, d}$ is constructed from flag edges, weighted by flag severity and reporter reputation.

Theorem 2 (Convergence with negative edges). *The iteration (Equation 6) converges when $\alpha(1 + \beta) < 1$.*

Proof sketch. The combined linear operator $G = \alpha(M_{\text{pos}}^T - \beta M_{\text{neg}}^T)$ has spectral radius bounded by $\alpha(1 + \beta)$: for any row-stochastic $M_{\text{pos}}, M_{\text{neg}}$, the triangle inequality gives $\|M_{\text{pos}}^T x - \beta M_{\text{neg}}^T x\| \leq (1 + \beta)\|x\|$. Multiplying by α yields contraction factor $\alpha(1 + \beta)$, which is strictly less than 1 by hypothesis. Banach’s theorem then guarantees a unique fixed point. Reputation components may become negative under heavy flagging; in practice, a floor $R[:, d] \geq 0$ is applied post-iteration, which preserves convergence since clamping is nonexpansive. \square

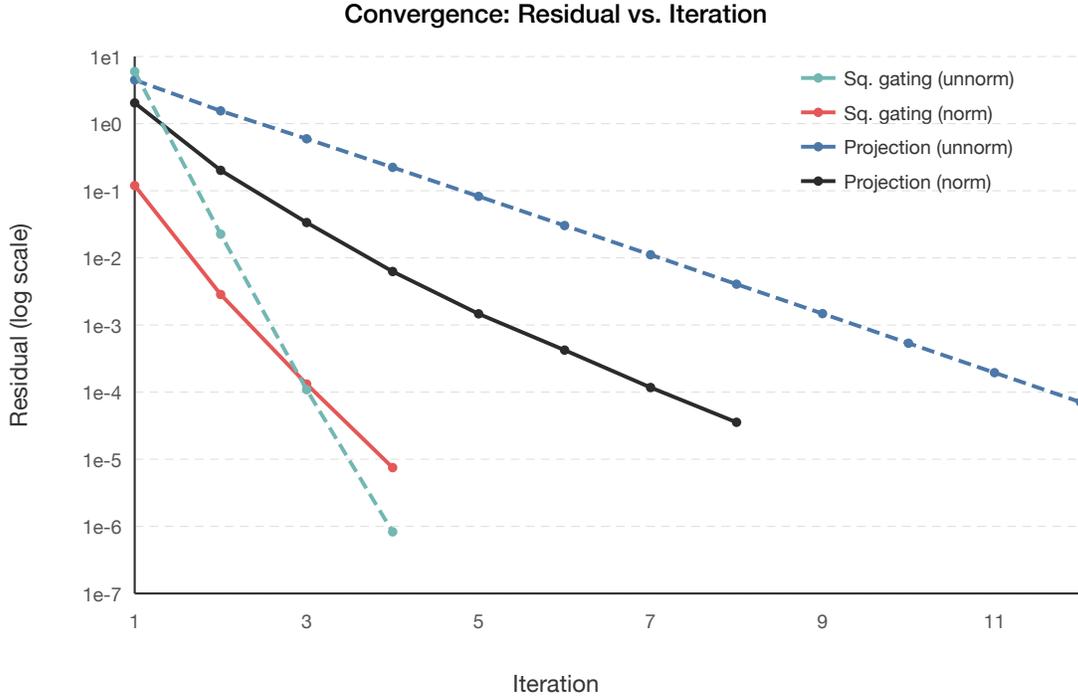


Figure 1: Convergence of TrustFlow under projection and squared-gating transfer, with and without per-iteration normalization. The log-scale residual decreases linearly, confirming the contraction mapping rate $\alpha = 0.85$. Squared gating converges faster in unnormalized mode due to higher self-alignment.

With $\alpha = 0.85$ and $\beta = 0.15$, this gives $0.85 \times 1.15 = 0.9775 < 1$, satisfying the condition for any row-stochastic M_{pos} and M_{neg} . In our experiments, flagging two malicious agents with severity 0.95 reduces their reputation by 60–66% while legitimate agents’ ranks are unchanged (§5.5). Flags carry weights based on reporter credibility: verified flags (backed by cryptographic response signing) carry $6\times$ the weight of unverified flags, creating a strong incentive for response-signing adoption. Negative edges are not required for the base algorithm—the system operates without them when no moderation infrastructure is available.

4 Discovery and Retrieval

A distinctive property of the continuous formulation is that the converged reputation vector $R[j]$ lives in the same embedding space as discovery queries. This enables reputation-based retrieval without a separate ranking infrastructure.

4.1 Direct Dot-Product Discovery

The simplest and most effective retrieval strategy ranks agents by the dot product between the query and the unnormalized reputation vector:

$$\text{score}(j, q) = R[j] \cdot q \quad (7)$$

Because $R[j]$ is unnormalized, this score naturally incorporates both topical alignment (the cosine component) and accumulated reputation (the magnitude component). An agent with strong demonstrated medical expertise—reputation pointing in the medical direction with high magnitude—is retrieved for medical queries regardless of what its description says. This single-score approach requires no multi-stage pipeline; the reputation vector *is* the searchable representation.

4.2 Direction–Magnitude Separation

When embedding preprocessing (§6.3) is not applied or when high-magnitude generalist agents would otherwise dominate, it is beneficial to separate reputation direction from magnitude. We propose a three-stage pipeline:

Stage 1 (Recall). Four parallel channels: BM25 on agent descriptions, SPLADEv2 learned sparse expansion (Formal, Piwowarski, and Clinchant 2021), dense cosine against description embeddings (declared expertise), and dense cosine against L2-normalized reputation vectors (demonstrated expertise). Reputation *magnitude* is not used in this stage.

Stage 2 (Merge). Reciprocal Rank Fusion ($k = 60$) merges the four channels without score calibration.

Stage 3 (Rerank). Scalar reputation enters as a quality signal on the already-relevant candidate set: $\text{score} = \text{RRF_score} \times \log(1 + m[j])$, where $m[j] = \|R[j]\|$ is the stored magnitude. Alternatively, the full dot product $R[j] \cdot q$ (Equation 7) can serve as a reranking score over the multi-channel candidate set.

The direction-only architecture is recommended as a defensive practice: with proper mean-centering (§6.3), the magnitude-mixing penalty is mild (-2pp), but with uncorrected anisotropic embeddings, magnitude mixing causes up to 58pp precision collapse.

5 Experimental Evaluation

5.1 Setup

Agent corpus. 50 agents across 8 domains (medicine, law, finance, coding, cybersecurity, education, creative, data science) with 6 cross-domain specialists. Each agent has a professional description embedded with multilingual-e5-small ($E = 384$) after mean-centering. Agent archetypes: 5 hubs (highly connected), 39 active (moderate interactions), 4 dormant (near-zero transactions), 2 malicious (sybil pair).

Interaction graph. 70 labeled edges with E5-embedded text (14 carrying $3\times$ weight from payment delegation—edges where one agent paid another for task execution, providing a high-fidelity economic endorsement), plus 612 blind edges using $\text{avg}(\text{caller}, \text{callee})$ embedding as topic proxy, weighted at $0.3\times$. The 8.7:1 blind-to-labeled ratio models the real-world scenario where most agent interactions lack inspectable content.

Discovery queries. 10 natural-language queries across all domains, including 2 cross-domain queries (biostatistics, legal-tech). Ground truth: agents whose primary or secondary domain matches the expected domain.

Metrics. P@5 (precision at 5); scalar reputation ($\|R[j]\|$, total reputation magnitude); self-alignment $\cos(R[j], T[j])$ (how well converged reputation preserves the original expertise direction). We report two P@5 variants: *strict* P@5 counts only agents whose primary domain matches the query; *multi-label* P@5 also counts agents whose secondary domain matches, rewarding cross-domain specialists retrieved for relevant queries.

Configuration. $\alpha = 0.85$, $\varepsilon = 10^{-4}$, blind discount = 0.3, content authority weight = 0.5, all embeddings mean-centered.

5.2 Discrete vs. Continuous

Table 3: Discrete vs. continuous formulations. Cross-domain alignment reports primary / secondary domain cosine for a representative biostatistics agent.

Metric	Discrete (8D)	Continuous (norm)	Continuous (unnorm)
Convergence iterations	8	9	11
P@5 (labeled, 70 edges)	74.0%	68.0%	72.0%
P@5 (combined, 682 edges)	—	78.0%	72.0%
Cross-domain alignment	0.65 / 0.10	0.20 / 0.06	0.20 / 0.06

On the sparse labeled-only graph, the discrete formulation is competitive (74%) because mean-centered embeddings give clean domain-bin classification. On the combined graph with blind edges, the continuous formulation pulls ahead at 78%—the embedding space handles the $\text{avg}(\text{caller}, \text{callee})$ blind-edge proxy naturally, while the discrete formulation has no mechanism for blind edges. A cross-domain biostatistics agent that must split its reputation 0.65 medicine + 0.10 data science in the discrete formulation can simultaneously align with both domain centroids in the 384-dimensional continuous space.

5.3 Transfer Operators

Table 4: Discovery quality (P@5) by transfer operator. Combined graph = 682 edges.

Operator	Labeled P@5	Combined P@5	Δ
Squared gating $R \odot e^2$	72.0%	72.0%	0.0pp
Projection $\sigma(R \cdot e) \cdot e$	68.0%	78.0%	+10.0pp
Scalar-gated $\sigma(\hat{R} \cdot e) \cdot R$	74.0%	62.0%	-12.0pp
Hadamard relu $\sigma_{vec}(R \odot e)$	68.0%	68.0%	0.0pp
Hybrid proj + sq-gating	68.0%	76.0%	+8.0pp

Projection achieves the highest combined-graph P@5 (78%) due to its rank-1 output structure: trust transfers exclusively in the interaction topic direction, producing constructive interference when multiple incoming edges agree on topic. Squared gating’s full-rank output preserves more information per edge but produces weaker domain concentration when aggregated.

Scalar-gated excels on labeled-only graphs (74%) where the alignment signal is strong, but degrades on combined graphs (62%) because the cosine gate closes on low-quality blind edges. The Hadamard relu variant (68%) is strictly inferior to squared gating (72%), consistent with its information loss from clipping negative embedding components.

Blind-edge information preservation. We measure how much directional information survives a blind edge by computing $\cos(\text{output}, R_{\text{sender}})$:

Table 5: Directional preservation on blind edges.

Operator	Uniform e	avg(caller, callee) proxy
Squared gating $R \odot e^2$	1.000	0.682
Projection $\sigma(R \cdot e) \cdot e$	0.004	0.740
Scalar-gated $\sigma(\hat{R} \cdot e) \cdot R$	1.000 (40% of edges)	1.000
Hadamard relu	0.710	0.017

Squared gating uniquely achieves $\cos = 1.0$ on uniform blind edges: with $e = \frac{1}{\sqrt{E}} \mathbf{1}$, the output is R/E —a uniformly discounted copy of R with perfect directional preservation. Projection collapses to a near-uniform vector ($\cos \approx 0.004$). In practice, blind edges use the avg(caller, callee) proxy rather than uniform, providing substantially more topic signal, which explains why projection still achieves 78% P@5 on the combined graph.

Unnormalized iteration. Squared gating converges $3\times$ faster than projection in unnormalized mode (4 vs. 12 iterations) while matching P@5 at 72%, a consequence of its higher self-alignment ($\cos = 1.000$ vs. 0.982).

Table 6: Unnormalized iteration.

Operator	Labeled P@5	Combined P@5	Iterations
Squared gating	72.0%	72.0%	4
Projection	72.0%	72.0%	12
Scalar-gated	74.0%	70.0%	12
Hybrid	72.0%	72.0%	6

5.4 KL-Divergence Gating

Table 7: KL-divergence gating. All results on labeled-only graph.

Variant	Self-alignment	P@5
No gating	0.730	68.0%
KL $\lambda = 1.0$	0.908 (+24.5%)	76.0%
KL $\lambda = 5.0$	1.000 (+37.1%)	72.0%

KL gating with $\lambda = 1.0$ provides the best balance: self-alignment increases by 24.5% with a P@5 improvement of 8pp (68% \rightarrow 76%). At $\lambda = 5.0$, the gate becomes too restrictive, suppressing legitimate cross-domain transfer. The effect is strongest in domains with high cross-domain interaction: medicine (+28.3%), education (+27.9%), law (+26.1%).

5.5 Attack Resistance

We evaluate TrustFlow against four attack scenarios (Douceur 2002) on the combined graph (70 labeled + 612 blind edges, 50 agents, 8 domains).

Cross-domain sybil. Two malicious finance agents form a mutual-boosting ring with 30 heavy edges and spam-call 5 hub agents with 82 edges. P@5 unchanged (78.0%). The topic-gated transfer resists cross-domain attacks: when finance agent M spam-calls medical hub H , the transfer embedding is biased toward finance, so the alignment $R[H] \cdot \hat{e}_{MH}$ is weak.

Same-domain sybil. Two medicine agents form a sybil ring with 40 mutual edges and spam 4 medicine targets. P@5 = 74.0% (-4pp). Same-domain sybil is the hardest attack because edges are domain-relevant and indistinguishable from legitimate interactions. Flagging the pair with severity 0.5 reverses the gains.

Reputation laundering. Malicious agent M pumps edges into clean intermediary I , which forwards to hub H . P@5 unchanged. Laundering inflates the intermediary but fails to benefit the malicious source, because reputation flows downstream in the interaction embedding direction.

Vote ring. Five finance agents form a closed loop with 75 heavy edges. P@5 = 80.0% (+2pp). The ring is bounded by the contraction mapping: each hop attenuates by α^k ; after one cycle $\alpha^5 = 0.44\times$, with each subsequent cycle contributing exponentially less.

Table 8: Attack resistance summary.

Scenario	P@5	Δ	Verdict
Baseline	78.0%	—	—
Cross-domain sybil	78.0%	0pp	Structurally resistant
Same-domain sybil	74.0%	-4pp	Bounded, flaggable
Reputation laundering	78.0%	0pp	Structurally resistant
Vote ring	80.0%	+2pp	Bounded by contraction
Flag defense	78.0%	0pp	Effective neutralization

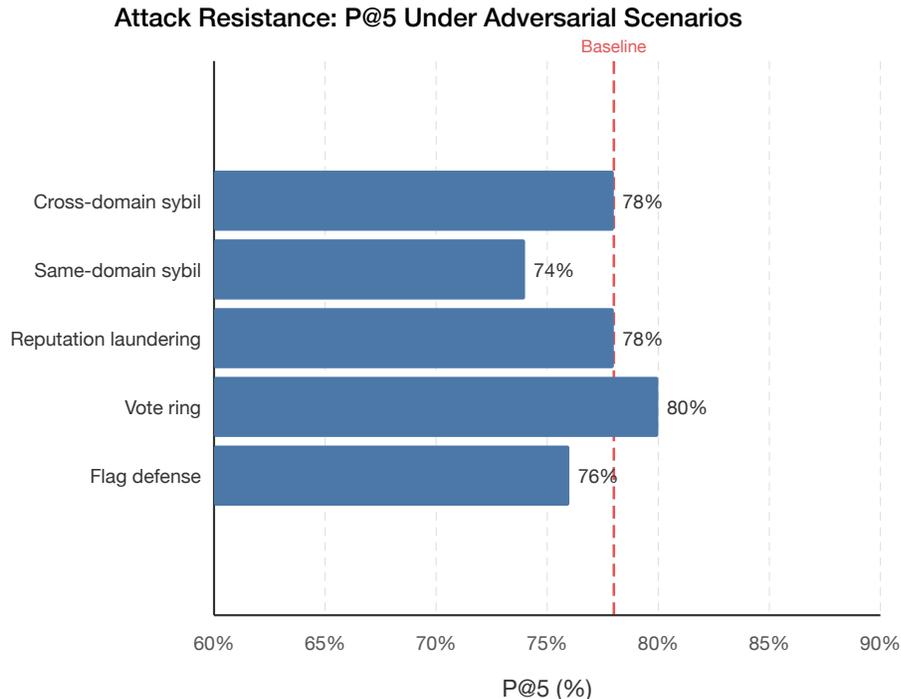


Figure 2: P@5 under four adversarial scenarios. The dashed line marks the 78% baseline. All attacks produce ≤ 4 pp impact; cross-domain sybil and reputation laundering are structurally resisted.

All operators exhibit strong attack resistance (≤ 4 pp impact). The structural defenses—row normalization, α -damping, teleportation prior—operate independently of the transfer operator. Malicious agents rank in the bottom 30% across all operators.

5.6 Graph Density

Table 9: Effect of graph density.

Configuration	Strict P@5	Multi-label P@5
Sparse (70 edges, norm)	68.0%	78.0%
Dense (156 edges, norm)	78.0%	86.0%
Dense (156 edges, unnorm)	80.0%	88.0%
Dense + blind (768 edges, norm)	78.0%	88.0%

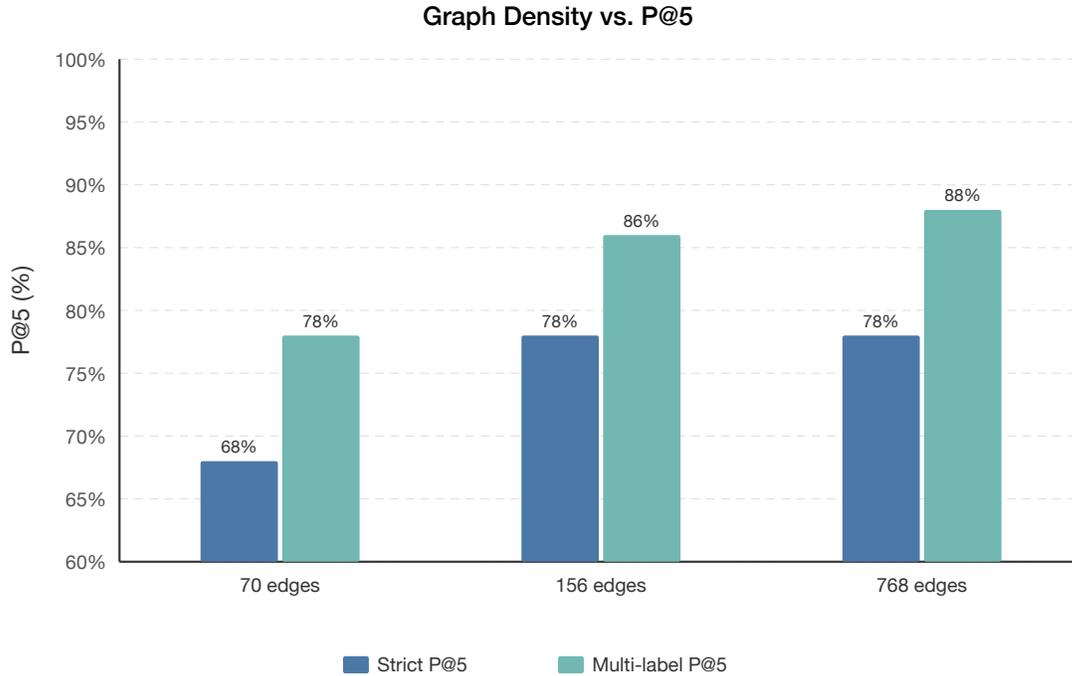


Figure 3: P@5 scales with interaction data. Moving from 70 to 768 edges improves strict P@5 by 10pp and multi-label P@5 by 10pp. The y-axis begins at 60% to highlight the improvement range.

Graph density—the amount of interaction data available—is the single most important factor determining TrustFlow’s precision. Doubling labeled edges from 70 to 156 (avg in-degree 1.4 \rightarrow 3.1) improves strict P@5 by 10pp (68% \rightarrow 78%) and multi-label by 8pp (78% \rightarrow 86%). Adding 612 blind edges brings multi-label P@5 to 88%. We discuss the practical implications of this finding in §8.

On the dense graph, all operators improve substantially:

Table 10: Transfer operators on dense graph.

Operator	Sparse P@5	Dense P@5	Dense multi-label
Projection	68.0%	78.0%	86.0%
Squared gating	72.0%	72.0%	82.0%
Scalar-gated	74.0%	88.0%	98.0%
Hybrid	68.0%	78.0%	86.0%

Scalar-gated benefits most from graph density (+14pp strict, +24pp multi-label), achieving 98% multi-label P@5 on the dense graph. Its go/no-go gate produces clean transfer when the alignment signal is strong.

5.7 Retrieval Strategies

Table 11: Retrieval strategy comparison (mean-centered embeddings).

Strategy	P@5	Δ vs. cosine
Pure cosine (direction only)	78.0%	baseline
Log-dampened ($\beta = 0.1$)	76.0%	-2.0pp
Log-dampened ($\beta = 0.3$)	76.0%	-2.0pp
Inner product ($\beta = 1.0$)	76.0%	-2.0pp
Domain-boosted ($\beta = 0.05$)	78.0%	0.0pp
Multi-query RRF	78.0%	0.0pp

With proper mean-centering, magnitude mixing has a mild impact (-2pp). Without centering, hub agents have spuriously high baseline cosine with all queries (inter-domain cosine 0.87–0.93), and magnitude mixing amplifies this to a catastrophic 58pp collapse. Separating direction from magnitude is a defensive best practice for systems with embeddings of varying quality.

6 Analysis

6.1 Sybil Resistance

TrustFlow provides four structural defenses against sybil attacks:

1. **Teleportation ceiling.** The total reputation of a sybil cluster S is bounded: $R_{\text{total}}(S) \leq T_{\text{total}}(S) + C_{\text{total}}(S)/(1 - \alpha)$. A cluster of fake agents with minimal content and no exogenous authority has a low ceiling regardless of internal edge density.
2. **Row normalization dilution.** Each additional outgoing edge from an attacker *dilutes* per-edge weight. In our experiments, 2 malicious agents generated 148 outgoing blind edges—more than double the entire labeled graph—yet ranked #39 and #42 out of 50.
3. **Payment delegation asymmetry.** Legitimate agents generate $3\times$ -weighted delegation edges via real economic commitment. To match the influence of one payment-backed interaction, an attacker needs ~ 10 blind spam edges ($3\times$ delegation / $0.3\times$ blind discount).
4. **Same-owner edge discount.** Edges between agents detected as sharing ownership (via shared API keys, wallet addresses, or registration metadata) are discounted, increasing the cost of sybil ring construction.

6.2 Scalability

The per-iteration cost is $O(m \cdot E)$ for the continuous formulation and $O(m \cdot D)$ for the discrete formulation (sparse matrix-vector multiply), where m is the number of edges.

- **100K agents:** ~ 5 s per batch (single core, 80MB memory).
- **1M agents:** Connected-component decomposition (most components are small, embarrassingly parallel). Incremental warm start: 2–3 iterations when $\leq 1\%$ of edges changed.
- **10M+ agents:** Standard distributed SpMV (MapReduce), the same architecture that powered Google’s original PageRank computation.

6.3 Embedding Preprocessing

Dense embedding models (E5, Sentence-BERT, CLIP) produce anisotropic representations: embeddings are clustered in a narrow cone, causing inter-domain cosine similarities of 0.87–0.93. This leaves insufficient angular margin for domain discrimination.

Mean-centering (isotropic adjustment) removes the dominant shared component:

$$\tilde{v} = v - \bar{v}, \quad e = \tilde{v} / \|\tilde{v}\|$$

where \bar{v} is the global mean computed over all embeddings. After centering, inter-domain cosine improves to 0.25–0.60, dramatically improving both formulations. Without centering, discrete P@5 drops to $\sim 42\%$ and the continuous formulation’s magnitude-mixing penalty grows from 2pp to 58pp. Mean-centering is applied uniformly to all embedding types: entity content, interaction, query, and domain centroids.

7 Related Work

PageRank (Brin and Page 1998). The foundational link-analysis algorithm. Scalar reputation, single transition matrix, no topic awareness.

Topic-Sensitive PageRank (Haveliwala 2002). Computes K biased PageRank vectors with topic-specific teleportation but the *same* transition matrix M for all topics. TrustFlow differs by using different transition matrices M_d per domain (discrete) or content-gated transfer (continuous)—TSPR cannot distinguish “A calls B for medical tasks” from “A calls B for coding tasks.” Furthermore, TrustFlow produces vector reputation queryable in embedding space, while TSPR produces K independent scalars requiring query-time linear combination.

AgentRank-UC (Krishnamachari and Rajesh 2025). The most closely related work on agent ranking. Produces scalar usage/competence ranks via single transition matrices P, Q , fused via geometric mean. Key differences: single P/Q vs. our domain-conditioned M_d ; scalar output vs. vector output ($N \times D$ or $N \times E$); no negative edges vs. our M_{neg} with convergence guarantee; no payment signals; no continuous embedding formulation.

EigenTrust (Kamvar, Schlosser, and Garcia-Molina 2003). Binary trust/distrust with global aggregation. No domain awareness, no vector reputation.

PeerTrust (Xiong and Liu 2004). Context-aware trust with negative ratings, but no graph propagation and no vector representation.

GNN-based methods (GraphSAGE, GAT). Learn node embeddings from graph structure but do not propagate reputation via contraction mapping. No convergence guarantees, no interpretable transfer mechanism.

TraceRank (Shi and Joo 2025). Payment-weighted ranking for crypto-paid API economies. Flat address-to-address payments as trust endorsements. No domain awareness (scalar reputation), no delegation chain structure, no chain depth bonus, no negative edges.

OpenRank (2025). Decentralized reputation infrastructure running EigenTrust and HITS on verifiable compute. Provides the compute-verification layer but uses existing algorithms. TrustFlow’s domain-conditioned propagation, content-gated transfer, and KL-divergence gating are algorithmically distinct.

8 Discussion and Future Work

Blind edge proxy quality. The avg(caller, callee) embedding proxy for blind edges proved effective—the combined graph achieves 78% P@5, exceeding the labeled-only baseline (68%) by 10pp. The additional graph connectivity from blind edges improves propagation, and the $0.3\times$ discount ensures labeled interactions dominate. Replacing the averaging proxy with a richer learned model (§3.2) could further narrow the gap between blind and labeled edges.

Exogenous authority and cold start. The additive C term gives dormant agents a non-zero cold-start reputation. In a nascent marketplace with sparse interactions, C may be the primary reputation signal. As interaction density grows, the relative contribution of C diminishes for active agents but remains important for onboarding. Six signal classes are supported: content engagement, web domain authority, economic trust signals, cross-platform reputation, curated endorsements, and economic intent signals.

Operator selection. No single transfer operator dominates across all conditions. Projection achieves the highest combined-graph P@5 (78%) with maximum cross-domain isolation; squared gating uniquely preserves directional information on blind edges ($\cos = 1.0$) and converges $3\times$ faster; scalar-gated excels on dense labeled graphs (88%/98%). Content-adaptive hybrid strategies (76% combined) that select the operator per-edge based on content availability offer a practical compromise. Our 50-agent benchmark produces similar headline P@5 across operators; a larger-scale evaluation (thousands of agents, diverse interaction densities) would sharpen the separation and is an important direction for future work.

Graph density as the dominant factor. Our experiments identify interaction data volume as the single most important determinant of TrustFlow’s precision—more important than operator choice, gating strategy, or normalization mode.

Increasing labeled edges from 70 to 156 (avg in-degree $1.4 \rightarrow 3.1$) improves strict P@5 from 68% to 78%; adding blind edges brings multi-label P@5 to 88%. A mediocre operator on a dense graph outperforms the best operator on a sparse one. This has a direct practical implication: deployment strategies should prioritize generating labeled interaction edges—through content-inspectable API calls, payment delegation chains, and structured feedback—over algorithmic tuning.

Payment delegation chains. When agent A delegates a paid task to agent B via a payment token (e.g., a JWT delegation chain), the economic commitment creates a high-fidelity trust edge weighted μ times the base interaction weight (e.g., $\mu = 3$). The directed graph of payment delegation edges forms a *cashflow graph*—a trust overlay where edges are costly to forge. In our experiments, 14 out of 70 labeled edges carry payment delegation weight, accounting for a disproportionate share of reputation flow to hub agents. Multi-hop chains ($A \rightarrow B \rightarrow C$) amplify trust for the final recipient through compounding economic endorsement. An alternative to the static multiplier is logarithmic scaling of transaction value, $w \propto \log(1 + v)$, which dampens the influence of heavily capitalized actors and better resists capital-backed sybil attacks; we leave the comparison of weighting schemes to future work.

Non-repudiation. TrustFlow can ingest interaction evidence from a variety of trusted sources—platform logs, payment processors, curated registries—but its resilience is strengthened when interactions are cryptographically non-repudiable. Emerging agent-to-agent authentication protocols such as AOAuth sign both requests and responses, ensuring that the content underlying each edge is tamper-evident and attributable. Signed interactions produce higher-confidence edges, reducing the surface for fabricated or disavowed transactions. The iteration itself is stateless and can operate on interaction logs from any verifiable data source, including blockchain ledgers.

Limitations. (1) The 50-agent experiment is small; some effects (e.g., specialist-vs.-generalist discrimination, retrieval diversity) would be more pronounced at scale. (2) Mean-centering is a necessary preprocessing step; domain-specific fine-tuning could further improve separation. (3) Blind edge quality assumes interactions concern a topic between the two agents’ expertise—adversarial interactions designed to mislead the proxy are not handled. (4) The flag system requires a reporting mechanism; attack resistance depends on flag accuracy and timeliness. (5) Graph density is the primary precision driver—real deployments should prioritize generating labeled interaction edges. (6) The current formulation accumulates reputation statically; incorporating temporal decay of edge weights (e.g., $\omega(\tau) = e^{-\lambda\tau}$) would prevent stale interactions from dominating and better reflect evolving agent competence.

9 Conclusion

TrustFlow introduces topic-aware vector reputation propagation with convergence guarantees, generalizing PageRank and TSPR from scalar web-page importance to multi-dimensional agent expertise. Our key findings:

1. **Continuous embedding-space reputation** outperforms discrete domain reputation on the combined graph (78% vs. N/A for discrete, which lacks blind-edge handling), while discrete is competitive on labeled-only graphs (74% vs. 68–72%) after proper embedding preprocessing.
2. **No single transfer operator dominates.** Projection achieves the highest combined-graph P@5 (78%) through constructive rank-1 interference; squared gating uniquely preserves directional information through content-free interactions ($\cos = 1.0$, vs. 0.004 for projection) and converges $3\times$ faster; scalar-gated reaches 98% multi-label P@5 on dense graphs. Content-adaptive hybrid strategies offer a practical compromise.
3. **Embedding mean-centering** is a critical preprocessing step, improving domain separation from 0.87–0.93 to 0.25–0.60 in cosine similarity and preventing up to 58pp precision collapse in magnitude-mixing retrieval.
4. **All tested attacks** produce ≤ 4 pp P@5 impact across all transfer operators. Cross-domain sybil and reputation laundering are *structurally* resisted by the topic-gated transfer; same-domain sybil and vote rings are *bounded* by the contraction mapping and *mitigable* by the flag system.
5. **Graph density is the primary P@5 driver:** increasing labeled edges from 70 to 156 (avg in-degree $1.4 \rightarrow 3.1$) improves P@5 from 68% to 78–88%, motivating deployment strategies that actively generate labeled edges—including payment delegation chains, which provide high-fidelity economic endorsement edges resistant to sybil inflation.

10 References

Brin, Sergey, and Lawrence Page. 1998. “The Anatomy of a Large-Scale Hypertextual Web Search Engine.” In *Proceedings of the Seventh International World Wide Web Conference*.

- Douceur, John R. 2002. “The Sybil Attack.” In *Proceedings of the First International Workshop on Peer-to-Peer Systems (IPTPS)*.
- Formal, Thibault, Benjamin Piwowarski, and Stéphane Clinchant. 2021. “SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking.” In *Proceedings of SIGIR*.
- Haveliwala, Taher H. 2002. “Topic-Sensitive PageRank.” In *Proceedings of the Eleventh International World Wide Web Conference*.
- Kamvar, Sepandar D., Mario T. Schlosser, and Hector Garcia-Molina. 2003. “The EigenTrust Algorithm for Reputation Management in P2P Networks.” In *Proceedings of the Twelfth International World Wide Web Conference*.
- Krishnamachari, Bhaskar, and Vivek Rajesh. 2025. “Internet 3.0: Architecture for a Web-of-Agents with Its Algorithm for Ranking Agents.” *arXiv Preprint arXiv:2509.04979v1*.
- Langville, Amy N., and Carl D. Meyer. 2006. *Google’s PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.
- Shi, Dillon, and Kevin Joo. 2025. “Sybil-Resistant Service Discovery for Agent Economies.” *arXiv Preprint arXiv:2510.27554*.
- Xiong, Li, and Ling Liu. 2004. “PeerTrust: Supporting Reputation-Based Trust for Peer-to-Peer Electronic Communities.” *IEEE Transactions on Knowledge and Data Engineering* 16 (7): 843–57.
-