Barataria Manifesto
For a chivalry distributed of trust

Blaise Ségal
blsegal@student.42.fr

"There is no book so bad that it contains nothing good. ”

Cervantes

1. Abstract

Contemporary digital platforms face fundamental vulnerabilities: the proliferation of fictitious identities, narrative manipulation, and the topological degradation of trusted graphs. Decentralized technologies, when anchored in a probative economy, make it possible to overcome these limits by integrating tokenomics, interaction analysis and arbitration logic into an architecture without a throne.

We offer a distributed infrastructure of reputation, articulated around three interdependent pillars:

(i) an algorithmic Sybil tax called Burn-to-Register, dynamically modulated according to the registration rate and without governable intervention;

(ii) a graph of consequences, where each edge encodes an economic relationship attested by payments x402 and potentially accompanied - if any - by an arbitrable certificate of insurance;

(iii) A pluralist market of arbitration guilds, transparently publishing their integrity indices in the form of Knowledge Assets anchored on OriginTrail's Decentralized Knowledge Graph (DKG).

This restricted network is subject to a limited and weighted PageRank, where the strength of the links includes the amount, nature, recency and the coefficient of assurance derived from previous interactions.

Payments **x402** play a bifunctional role: vectors of atomic settlement, they simultaneously constitute evidence of trust. Residual disputes feed a system of preventive observability: before any economic commitment, agents consult the integrity metrics published by the guilds.

Added to this structure is a cognitive layer composed of three neuro-symbolic agents: **Sancho**, the pragmatic agent of payment, who engages only on symbolic evidence; **Sansón**, the formal analyst, who derives the coefficients of assurance according to falsifiable formulas; and **Quixote**, the topological interpreter, who identifies the large-scale attack motives within the graph.

In this paradigm, reputation ceases to be a global entity or a centralized score: it becomes a versioned, falsifiable and locally inferable flow, specific to the applicant's ego-network. Each economic relationship leaves a symbolic imprint, which can be consulted by any actor in the network via the interoperable mechanisms of the Cross-Consensus Message Format (XCM).

We expose the mechanical and algorithmic foundations of this device - formulas, minimum interfaces, contractual primitives - while demonstrating its resilience to classic attack

vectors: Sybil flood, link farms, arbitration capture, or subsidized narrative economy. Finally, we formulate minimum specifications, immediately implementable, allowing the progressive institution of a space of trust without throne.

In our proposal, the Superposition Agent ↔ Knowledge ↔ Trust is the essential articulation: the Agent (human or AI) acts above the Knowledge layer (the DKG), itself anchored in the Trust layer (NeuroWeb, parachain Polkadot). This architecture ensures that trust circulates as a continuous, auditable, interoperable flow via XCM, and impossible to capture.

Thus, reputation ceases to be a sovereign abstraction: it becomes an emerging, situated, falsifiable, and strictly local property to the applicant's ego-network.

2. Introduction

In their current form, distributed social graphs remain vulnerable to systemic attack vectors, first and foremost **Sybil identities** and the strategic manipulation of centralities.

The proliferation of fictitious entities profoundly alters the topology of trust, making relational metrics inoperative when they are not anchored in verifiable economic commitments.

Faced with this, contemporary devices of **proof of humanit**y - based on biometric, cognitive or behavioral tests **- certainly reduce the marginal cost of duplicity, but remain detached from the internal economy of the graph.**
They thus fail to couple identity with transactional responsibility.

In addition, although Web3 micro-payments of type x402 allow atomic exchanges, their structure does not prescribe any endogenous mechanism for the evaluation of the reliability of counterparties.

It is in this gap that our ambition is part of: to design a probabilistic reputation protocol, locally calculable, where each interaction is burdened with an economic cost, each relationship with a proofable commitment, each dispute with exposure to risk. The agent is no longer committed to faith or notoriety, but to the public imprint of an anchored and auditable transactional history.

This architecture is based on a functional tripartition:

- **The Agent layer**, where individual decisions and preferences are expressed;
- **The Knowledge layer**, where epistemic active ingredients are formed, versioned and questioned;
- **The Trust layer**, where the contractual execution takes place, arbitrable and falsifiable.

To embody this paradigm, we introduce three cognitive agents modeled as a distributed interpretation process:

- **Sancho**, the paying agent, compares the symbolic facts (metrics published in the DKG) with his local policy before triggering the payment;
- **Sansón**, the academic auditor, derives the integrity score from disputes, income and correlations of verdicts, and publishes the new evaluation versions;
- **Quixote**, the topological guardian, interprets the narrative structures of the graph to signal reasons for attack (link farms, cartels, collusive schemes).

This cognitive triangulation transforms trust into a symbolic process located. Reputation becomes a distributed, updated, searchable signal in context, and no longer a static entity or an abstract score. **The evaluation is no longer centralized: it is expressed, falsified, and interpreted by an ecology of agents, each operating according to its own policy, but sharing a common grammar of evidence.**

The infrastructure that allows this logic is NeuroWeb, a specialized parachain of the Polkadot ecosystem. It serves as a neutral, non-captable execution layer, connected by XCM to all ecosystem networks. **Trust thus becomes not only calculable without center, but interoperable without standardization**: it circulates as verifiable knowledge, and not as a transferable asset.

Thus, distributed trust is not only conceived as a theoretical artifact, but as an inherent property of a multi-chain infrastructure explicitly designed for interoperability.

3. Context & Problem

Systemic diagnostics
Current decentralized social architectures are marked by structural vulnerabilities that compromise the reliability of reputation dynamics:

- **Low-cost identity proliferation:** the quasi-frictional issuance of accounts promotes the exponential amplification of Sybil attacks, disseminating the illusion of plurality where strategic duplication actually reigns.

- **Centralization of reputation:** the aggregation of trust in the form of a global score transforms the system into a single target, easily captured by opportunistic entities.

- **Monolithic arbitration:** the concentration of the power of judgment in a centralized body exposes the protocol to the risks of buying verdicts, inter-jury collusion, and the formation of decision-making cartels.

- **Opaque economic narratives:** the emergence of subsidized models - often presented as performative - masks an absence of authentic economic signal, favoring the effects of ponzi disguised as yields.

- **Absence of truly multi-chain reputation protocols:** The parachains of the Polkadot ecosystem, although operationally interoperable via XCM, remain without a cross-sectional Trust Layer. Therefore, identities remain fragmented, evidence of commitment non-exportable, and localized arbitrations - all discontinuities that weaken the universality of trust signals. This deficit of inter-chain articulation requires thinking about a reputation whose local (economic, transactional) anchorage remains compatible with a systemic circulation of evidence.

Design imperatives
Faced with these impasses, the protocol we propose is articulated around three fundamental principles:

- **Absence of throne:** proscription of any governable instance or an aggregate score likely to exercise symbolic sovereignty. No global vector of power or visibility should be able to be captured.

- **Distributed verifiability:** each metric or reputation status is anchored as a *Knowledge Asset (K.A)* on the *Decentralized Knowledge Graph (DKG)*, guaranteeing its searchability and auditability without mediation.

- **Economic rooting of relationships:** each identity must be linked to an explicit cost; each interaction, to a verifiable transaction; each dispute, to a real financial exposure. Thus, social topology is constructed not by declarative, but by proof of commitment.

4. Hypothesis / Proposal / Resolution

Founding hypothesis

We assume that a reputational system cannot be objectified by a global score, nor governed from a symbolic center. All centrality - whether metric, algorithmic, or institutional - becomes a target for Sybil. Trust can therefore only emerge in the immanence of economic interaction, in a local, falsifiable, auditable, and situated framework.

In this paradigm, a system becomes resilient if:

1. **Identity production is subject to a dynamic economic constraint, and not governable**
2. **Inter-agent relations emanate from attestable economic interactions -** in particular via x402 payments, accompanied by a freely selectable arbitration option
3. **Execution assurance is part of an open market of competing guilds,** bound by rigorous metric transparency

This triptych makes possible the inference of a local reputation, embodied, resistant to manipulation, and strictly contextualized to the applicant's ego-network. Reputation is no longer an ontological attribute: it becomes a pragmatic, versioned, and queryable trace.

Systemic proposal: The Barataria Architecture

Inspired by the island entrusted to Sancho Panza in the work of Cervantes
the **Barataria Architecture** abandons the search for a global "Reputation Score". It focuses on providing verifiable information on the reliability of transactional insurance, thus transforming trust into a market-informed choice.

1. Tokenomics for identity & trust (Pillar 1: The Guarantee)

We replace the defective concepts of **"Proof-of-Personhood" (PoP)** and **"global DAO"** with the economic sovereignty of "Guilds".

Concept	Mechanism	Sybil Resistance
Identity (Entry Cost)	**Algorithmic Sybil Tax (Burn):** The cost of creating an identity (a node on the DKG) is not fixed. It is calculated via a decentralized Oracle USD (to avoid inflation/deflation of the token) and increases exponentially (surge pricing) in case of congestion (e.g. if > 1000 accounts are created/hour).	**Neutralizes** the creation of millions of bots at zero cost ("Siphoning" attack). The absence of global DAO to set the cost eliminates the "Throne" of the Tax.
Trust (The Guarantee)	**Arbitration Guild Market:** A seller must choose a "Guild" to arbitrate his disputes. The Guild stakes a capital (TRAC/NEURO) to guarantee the sellers of its list (TCR). Investors can co-stik on the Guild to earn a return.	**Neutralizes** the 51% attack on a global Tribunal. The arbitration is distributed on an "Archipelago" of competing Guilds.
Dispute (Consequence)	**Symmetrical Deposit:** In the event of a dispute, the Buyer must stake the same amount as the Seller (e.g. $10). The dispute is sent to the Guild chosen by the Seller.	**Neutralize Sybil's** "Paid Murder" attack, because the cost of the attack is equal to the attacker's risk.

2. Graph-based reputation inference (Pillar 2: The Mirror)

We abandon the global **PageRank** (easily corrupted) and the import of a potentially pre-poisoned web 2 social graph. The DKG serves as an analysis engine to unmask Ponzi cartels and frauds.

Concept	Mechanism	Role of the DKG
The Graph	**Consequence Graph:** We ignore the social graph. The edges represent only **the successful and undisputed x402 transactions between two nodes** (Identities).	**The DKG** is the source of immutable truth for these transactions. Resolved disputes are recorded as Knowledge Assets (KA).
The Analysis	**The DKG Mirror:** The protocol generates verifiable **Knowledge Assets (KA)** that expose the behavior of Guilds, not individuals.	The DKG automatically publishes: **Proof of Performance (**Traces the source of APY), **Correlation Index** (Detects if 50 Guilds act as a single Sybil cartel), and **Dispute History** (How the Guild voted).
The Score	**No Overall Score.** The "reputation score" is replaced by the Insurer's verification. A node is reputed if its Insurer is not "Subsidized" and « Correled »	AI agents can query the DKG with **SPARQL** to obtain the KA of a Guild before any transaction.

3. Transactions via x402 for quality & commerce (Pillar 3: Action)

The system is designed to make trading with the malicious actor **economically irrational for the Buyer/AI Agent.**

Concept	Mechanism	Demonstration x402
Insurance Contract	**"Smart" transaction:** The seller uses **x402** to invoice. The API endpoint (paymentMiddleware) must include **a Proof of Assurance** (Insured by Guild XYZ).	The endpoint could be GET /api/trusted_data protected by paymentMiddleware("0xVendorAddress", {"/trusted_data": "$0.01", "Guild": "0xABC"});
Critical Verification	The Buyer's AI agent performs a critical check before paying: a direct request to the DKG to obtain the "Mirror" (Proof of Performance, Correlation Index) from the Insurer's Guild.	The AI agent executes: SPARQL Query DKG -> GET Knowledge Asset of Guild 0xABC. If the APY is "Subsidized (98%)" and the Correlation is "Cartel (99%)", **the agent cancels the payment.**
Monetization	Honest sellers (insured by an "Organic" Guild) can charge a premium price via x402. The consumer voluntarily pays for data verified as being provided by an uncorrupted source (the "Mirror" has proven it).	**Use Case:** An AI agent buys a "market analysis" at $0.05 via x402 only if the DKG proves that the Guild providing the source has an **Organic Provenance > 80%.**

The island of Barataria can thus survive the world-class attacks of Sybil because it does not build a "Throne" to buy.

The island of Barataria does not solve Sybil's problem by exclusion but by the cryptographic visibility of its strategies. It does not block fraud: it makes it unprofitable, because it is immediately detectable. It does not promise the truth: it guarantees the exposure of lies. He does not build an algorithmic throne: he makes an epistemic mirror.

This system produces a situated, emerging, interoperable reputation, without transcendent authority or invocable oracle. **Veracity becomes a falsifiable property — not a political fiction or a centralized heuristic.**

The architecture being proposed **with a native integration of these mechanisms within NeuroWeb** - specialized parachain - where BurnToRegister, Consequence Graph and Barataria Arbitration Archipelago are implemented in ink contracts! (Wasm), benefiting from Polkadot's shared security. Inter-chain flows (burn, x402 payments, LocalTrust interrogation) are orchestrated by XCM, making the Incorruptible Archipelago not a closed protocol, but an inter-parachain public good, called to irrigate the entire ecosystem.

The resolution we propose is embodied in a dynamic cycle: the Agent (AI or payer) publishes its metrics, another entity queries this data, initiates a transaction, the Trust layer executes and arbitrates, and the results return to the Knowledge layer as an updated version. Thus, reputation is transformed into a measurable, inferable, and locally contextualized flow.

5. Detailed mechanism

5.1. Identity: Burn-to-Register mechanism (adaptive Sybil tax)

Identity issuance is based on a dynamic cost function, designed to discourage Sybil attacks via exponential taxation adjusted according to the registration flow. This mechanism introduces a growing economic barrier, without recourse to changeable centralized governance.

Modelling
Either:

- λ_t : The instantaneous rate of identity issuance (in registrations per hour),
- $\lambda*$:The target system stability threshold (e.g. registrations/time),
- $T_{1/2}$: Smoothing half-life of the exponential average (e.g.:),
- C_t : Cost of registration at time t, expressed in USD,
- C_0 : Reference floor cost (e.g.: 1 USD),
- α : Exponential elasticity factor ($\alpha \in [\ln 2, 2 \ln 2]$).

We first define the exponential moving average (EWMA) of the flow rate:

$$\lambda_t = (1-\gamma) \cdot x_t + \gamma \cdot \lambda_{t-1} \quad \text{où} \quad \gamma = 2^{-\frac{\Delta t}{T_{1/2}}}$$

Then, the recording cost is expressed by the following exponential function:

$$C_t = C_0 \cdot \exp\left(\alpha \cdot \max\left(0, \frac{\lambda_t}{\lambda *} - 1\right)\right)$$

This wording implies that:

- When $\lambda_t \leq \lambda *$, Then $C_t = C_0$ (Stable floor),
- When $\lambda_t > \lambda *$, The cost increases super-linearly with the flow rate,
- A limited degrowth constraint is imposed:

$$C_{t+\Delta} \geq \frac{1}{2} \cdot C_t \quad \text{(downward slope limited with each update)}$$

A minimum decrease slope (0.5 × per update) ensures economic continuity while prohibiting bypassing by gusts. **This mechanism introduces a super-linear, non-governable cost**, deterring identity floods while maintaining architectural neutrality.

Contract specification:

function burnToRegister(address token, bytes proof, address recipient)
  returns (IdentityId id, uint256 costUSD, uint64 blocktime, uint64 ewmaRate);

The concrete implementation of this mechanism is based on an ink contract! Deployed on NeuroWeb, where EWMA logic, cost exponential and interactions with the median oracle are executed with the efficiency of Wasm. Payments intended for the burn - whether they come from Moonbeam, Astar or other parachains - are conveyed via XCM: a WithdrawAsset → BuyExecution → Transact carrying the call to the BurnToRegister contract. Thus, the identity itself becomes an inter-chain act, whose certification is secured by the Relay Chain and whose IdentityMinted event systematically includes the original MultiLocation.

5.2. Edges: transactional evidence via payments x402

A directed ridge $e = (A \rightarrow B)$ Is established if, and only if:

1. A x402 payment is executed without dispute during the defined window,
2. No dispute has led to a sanction (slash),
3. An insurance certificate has been included.

Each receipt encapsulates: the amount, the type of service, a time stag, and the insurance seal. The weight of the edge is then defined as:

$$w_{A \rightarrow B} = \min(m, M) \cdot \beta(\text{type}) \cdot \tau(\Delta t) \cdot \sigma(\mathcal{G})$$

- m : Amount paid; M : Ceiling by interaction (ex. 50 USD),
- β : Multiplier by category (API, dataset, service),
- $\tau(\Delta t) = \exp(-\Delta t / T_\tau)$: Obsolescence function with half-life $T_\tau = 90$ days,
- $\sigma(\mathscr{G})$: Insurance coefficient of the declared guild.

In a multi-channel setting, the economic edge is no longer local: an x402 payment initiated on Astar can trigger, via XCM, a call to the x402 middleware installed on NeuroWeb. He verifies the local policy, reads the insurance coefficient with the GuildTCR, and accepts or rejects the payment in an atomic way. Any failure causes an automatic refund, demonstrating that transactional reliability is guaranteed here not only by the internal logic of the protocol, but by the deterministic mechanics of the Cross-Consensus Message Format.

5.3. Derivation rules

Each guild $\mathscr{G}$ Publishes a set of Knowledge Assets versioned on the DKG.
These publications are the raw material of a deterministic evaluation process producing the insurance coefficient $\sigma(\mathscr{G})$, Intended to weigh the weight of economic edges.

The validity of a link of trust here is based not on a holistic authority or a centralized aggregation, but on a series of attested transactions, each relationship leaving a trace, each commitment generating evidence. In this perspective, the truth of reliability becomes a falsifiable hypothesis, continuously updated by interactions, then versioned, signed and time-dated in the Decentralized Knowledge Graph - epistemic matrix where credibility becomes distributed knowledge.

(A) Grant ratio
The subsidy ratio $\mathrm{SR}_{\mathscr{G}}$ Measures the guild's dependence on non-market income:

$$\mathrm{SR}\mathscr{G} = \frac{E\text{internal}(\mathscr{G}) + T_{\text{correles}}(\mathscr{G})}{F_{\text{nets}}(\mathscr{G})} \quad \text{(evaluated over 90 days)}$$

With:

- E_{internes} : 90-day internal token issues,
- T_{correles} : financial transfers between entities correlated to $\mathscr{G}$,
- F_{nets} : Net fees actually collected for arbitration.

Thus, a subsidized guild ($\mathrm{SR}_{\mathscr{G}} > 0.5$) loses economic signal.

(B) Cartel index

The cartel flag $\mathrm{CF}_{\mathscr{G}}$ Reports a concentration of verdicts or jurors indicative of collusion:

$$\mathrm{CF}\mathscr{G} = \mathbf{1}\left\{\rho\mathscr{G}(\text{verdicts}) > 0.9 \ \wedge \ \omega_{\mathscr{G}}(\text{jurors}) > 0.3\right\}$$

Where:

- $\rho_{\mathscr{G}}(\text{verdicts})$: Inter-guild correlation of verdicts on similar cases,
- $\omega_{\mathscr{G}}(\text{jurors})$: Average rate of jury.

The value 1 indicates a proven suspicion of cartel.

(C) Integrity score

The integrity score $\mathrm{IS}_{\mathscr{G}}$ Is a continuous metric, derived from public observable variables:

$$\mathrm{IS}\mathscr{G} = g\left(r\text{success}, d_{\text{latency}}, v_{\text{stakes}}, c_{\text{delay}}\right) \in [0,1]$$

Where the g function normalizes the performance of the guild according to:

- r_{success} : rates of disputes resolved without appeal,
- d_{latency} : median resolution time,
- v_{stakes} : variance of the amounts staked,
- c_{delay} : compliance with procedural deadlines.

(D) Final projection: insurance coefficient

The insurance coefficient $\sigma(\mathscr{G})$ Is then defined as follows:

$$\sigma(\mathscr{G})\left(1 - \mathrm{CF}\mathscr{G}\right) \cdot \mathrm{IS}\mathscr{G} \cdot \left(1 - \mathrm{SR}_{\mathscr{G}}\right) \quad \text{with} \quad \sigma(\mathscr{G}) \in [0,1]$$

This expression guarantees:

- $\sigma(\mathscr{G}) = 0$ If the guild is flagged as a cartel ($\mathrm{CF}_{\mathscr{G}} = 1$),
- $\sigma(\mathscr{G}) \to 1$ For an efficient, self-financed and independent guild,
- A multiplicative continuity: the decrease of a single metric degrades overall confidence without an arbitrary threshold.

Arbitration guilds - registered in the NeuroWeb TCR register - become, by construction, usable by any parachain: their metrics, published on DKG, can be queried via XCM Query. A DAO on Moonbeam, a confidential calculation dApp on Phala, or a DeFi protocol on Acala can thus evaluate in real time the integrity of a guild before engaging in an economic transaction.

The Agent publishes via an SDK (dkg.py / dkg.js) a Knowledge Asset (KA) on the DKG; the Paying Agent queries the DKG via SPARQL to select a guild that meets its criteria; then the x402 transaction is initiated, routed via XCM to NeuroWeb,

Who executes the contract and, in the event of a dispute, performs the slash.

This distributed device calls, in its symbolic economy, for the creation of three fundamental agents of the system - payer, calculator, observer - which were named after the structuring figures of Cervantes' Don Quixote, the founding work of the modern epistemology of the false.

Don Quixote is the inaugural story of simulation societies - where illusion takes the mask of the truth, and where only praxis, situated and exposed, makes it possible to infer trust. To name these agents is to inscribe the protocol in a tradition of lucidity in the face of pretense, an economy of reality where the truth is experienced in the risk, not in the decree.

I. Agent "Sancho" (The Pragmatic Payer)

(Why "Sancho"?): Sancho Panza is the archetype of pragmatism. His engine is economic rooting: he follows Quixote in exchange for a tangible reward (his "island"). Agent Sancho rejects any internal knowledge that would not be verified.

- **Neuro-Symbolic Behavior:**
    - **Symbolic (DKG):** The agent "Sancho" must first use the tool tool_query_dkg to read the verifiable facts: "Is this guild trustworthy? What is his sigma? Is his cartel_flag active?" .
    - **Neural (LLM):** The LLM only reasons on these DKG facts, comparing them to its "local policy" (its "greed" or its "risk aversion").
    - **Action (Trust):** It will only execute the tool tool_execute_xcm_payment if the symbolic facts of the DKG correspond to its pragmatic policy. He refuses to pay for the "giants" (unreliable guilds) that his master could hallucinate.

- **Agent Architecture (LangGraph model) :**The architecture models a graph of states where the agent cannot progress towards payment without first verifying the symbolic facts of the DKG.

  a. **Initial State (Input):** Task = { "action": "pay_service", "service_id": "ka://...", "amount": 50 }

  b. **Node: schedule_verification**

     - **Agent (LLM) Chain-of-Thought (CoT):** "New payment task. I can't trust parametric knowledge. I must first verify the reputation of this service on the DKG. My policy requires the metrics sigma, cartel_flag, and subsidy_ratio."

  c. **Node: query_dkg (Symbolic Layer)**

     - **Action :** call of tool_query_dkg(query_sparql: str).
     - **Query (SPARQL) :** "Retrieve the most recent Knowledge Assets related to service_id, especially the integrity KAs of its associated guild."
     - **Output :** A JSON-LD containing the facts, for example: { "sigma": 0.69, "cartel_flag": false, "subsidy_ratio": 0.18, ... }.

  d. **Node: reason_on_facts (Neuronal Layer)**

     - **Agent (LLM) CoT :** "The facts are recovered.
       - Done (DKG): sigma is 0.69.
       - Done (DKG): cartel_flag is false.
       - Done (DKG): subsidy_ratio is 0.18.
       - Rule (Local Politics): if sigma < 0.8: counter_offer(price * sigma).
     - **Calculation :** 50 * 0.69 = 34.5.
     - **Decision :** Trust is not total. I do not approve 50, but I proceed with a counter-offer based on the insurance coefficient."

  e. **Node : Execute_decision (Couche Trust)**

     - **Action :** Call of tool_execute_xcm_payment(destination_address: "...", amount: 34.5).

f. Node : Publish_consequence (DKG update)

- **Action :** Appel de tool_publish_dkg(data_jsonld: dict).
- **CoT :** "My transaction was successful. I have to publish this interaction as a new edge in the Consequence Graph, creating a new probative footprint for future agents."

II. Agent "Sansón" (The Academic Auditor)

(Why "Sansón Carrasco"): The Symbolic Logician: Sansón Carrasco is the "Bachelier" (the graduate) of Salamanca. He is the embodiment of academic logic and symbolic rules. He opposes the neuronal "madness" of Quixote by using reason, facts and deduction. The Calculator: Its function is not to interpret, but to calculate. It executes the derivation formulas (IS, SR, CF) to determine the mathematical "truth" of the integrity of an arbitration guild in the event of a dispute. The Audit Actor: He is not passive. After auditing the situation, he acts (becoming the Knight of the White Moon) to impose the consequences of his logic. Similarly, our agent acts by publishing the calculated sigma on the DKG, thus applying his algorithmic sentence.

- **Agent Architecture (CrewAI model) :** This workflow is permanent and deterministic, ideal for a CrewAI model where agents have fixed roles. The agent is hosted on a DKG Edge Node.
- **Role :** Trust Analyst
- **Objectif :** Maintain the integrity of the DKG by continuously evaluating arbitration guilds.
- **Tasks (continuous workflow):**

  a. **Task: Monitor_Transactions**

     - **Action:** The agent uses tool_query_dkg to continuously listen to the new transaction KAs (published by the "Sanchos Agents") and the litigation KAs.

  b. **Task: Calculate_Metrics (Local/Symbolic)**

     - **Action :** The agent locally executes the LocalTrust pseudo-code and the derivation functions (SR, IS, CF) on the retrieved data. It is not an LLM, it is a pure algorithmic calculation (Sansón's "academic" work).

c. Task: Audit_Results (Neuronal)

- **Agent (LLM) CoT :** "The symbolic calculation is complete.
    - Result (SR) : Guild '0xABC' a un SR = 0.55.
    - Result (CF) : Correction of verdicts rho = 0.91, omega sworn coverage = 0.35.
- **Thinking :** The SR exceeds 0.5 AND the CF thresholds are exceeded. This guild shows clear signs of subsidy and collusion.
- **Decision:** I have to calculate the new sigma, which will be 0 because of the CF_flag=1, and publish this critical update. ”

d. Task: Publish_Analysis (Action)

- **Action :** The agent calls tool_publish_dkg To publish the new version of the analysis KA (ex: ka://analytics/integrity/0xABC/v6).

III. Agent "Quixote" (The Topological Guardian)

Justification of the Expertise (Why "Quixote"): It is the most conceptual choice, but the most powerful. Don Quixote is the only character who perceives a narrative structure (giants, armies, castles) where others see only isolated facts (mills, sheep, inns).Agent Quixote is our "Guardian" because he is designed to do exactly that: use neuronal reasoning (LLM) to interpret the topology of the graph and find attack patterns that other agents, too focused on local facts (such as "Sancho" or "Sansón"), cannot see.

- **Neuro-Symbolic Behavior:**
    - **Symbolic (DKG) :** It uses tool_query_dkg to retrieve large sets of topological data (who connected to whom? When?).
    - **Neuronal (LLM) :** This is where his neural "madness" comes in. The LLM is trained to recognize the forms of attacks:
        - "I see thousands of nodes connected in stars with low-weight edges. Pragmatics see 'mills' (valid transactions), but I see a 'giant' (a Sybil link farm).
        - "I see an 'almost closed community' where verdicts correlate perfectly. The others see a 'inn' (a normal guild), but I see a 'castle' (a collusion cartel)."
    - **Action (Knowledge) :** He doesn't fight the mills. He calls tool_publish_dkg to publish an "Alert KA", signaling the narrative structure of the attack to the entire network.
- **Architecture de l'Agent (modèle dRAG) :** The dRAG (decentralized Retrieval-Augmented Generation) is at the heart of this agent. He uses the Retrieval (DKG) not for simple facts, but for complex topological patterns, which the Generation (LLM) interprets.

5. **Node: scanner_topology (Symbolic)**
    - **Agent (LLM) CoT :**"I'm launching my periodic defense scan. My mission is to find 'link farms' and 'almost closed communities'."
    - **Action :** tool_query_dkg(query_sparql: str).
    - **Query (SPARQL):** "Search for subgraphs (ego-networks) with high internal density but with low-weight edges (w) and/or systematically low or zero sigma. ”

2. **Node: reason_on_patterns (Neuronal)**

    - **Agent (LLM) CoT:** "The DKG returns several suspicious clusters."

        - Pattern 1: Cluster 'A' presents 1000 nodes created via BurnToRegister during a lambda_t peak.

        - Pattern 2: These nodes exchange x402 micro-payments (very low m amount) to simulate an activity.

        - Pattern 3: The KAs of their associated guilds are self-subsidized (SR > 0.8).

        - **Conclusion:** It is a Sybil link farm designed to artificially inflate the PageRank. The structure is 'topologically sterile' economically, but looks like an attack."

3. **Node: publish_alert (Action)**

    - **CoT :** "I cannot censor these facts, but I can add context. I publish an 'Alert KA' so that other agents (such as the 'Sanchos') can use it in their own reasoning."

    - **Action :** tool_publish_dkg(data_jsonld: dict).

    - **Data :** { "@type": "AlertKA", "target_pattern": "...", "threat_type": "SybilLinkFarm", "severity": 0.9, "evidence": [...] }

The status return is published again as the next version of the KA. This orchestration embodies the architecture of the three layers:

- Layer Agent: paying actors, guilds (IA)
- Layer Knowledge: OriginTrail DKG, publication, versioning, interrogation
- Trust layer: NeuroWeb + Polkadot, contractual execution, stake/slash

6. Expected results and implications

6.1. Resilience to attack vectors (Red Team modeling)

The proposed architecture demonstrates structural robustness in the face of various types of systemic aggressions. Simulations in adverse conditions - of the Red Team type - reveal the following defensive properties:

- **Flooding Sybil :** the cumulative cost of identities $\sum C_t$ grows super-linearly with the increase in flow λ_t, by exponential surge effect. The decrease in cost is limited, prohibiting brutal regenerations at low cost.

- **Link closures:** in the absence of authentic economic payments directed to intact nodes, and without a credible insurance coefficient ($\sigma \to 0$), The edges remain weak: their weighting erodes via the function $\tau(\Delta t)$, And Cape M limits their influence. The structure thus generated remains topologically sterile, unable to increase the PageRank score π Within $G_S^{(r)}$.

- **Capture of a guild:** the local manipulation of verdicts generates a systemic backlash. Perceived integrity falls $\mathrm{IS}_{\mathscr{G}} \downarrow$, the subsidy ratio increases ($\mathrm{SR}\mathscr{G} \uparrow$), the flag cartel is activated ($\mathrm{CF}_{\mathscr{G}} = 1$) : as a result, the coefficient σ collapses, which dissuades rational agents from committing economically "before payment". The deal-flow dries up endogenously.

- **Biased economic narrative:** an artificially high yield (e.g. 50% APY) is revealed by a $\mathrm{SR}_{\mathscr{G}}$ abnormally high. Local trust policies reject these guilds as incredible counterparts.

- **Oracle manipulation:** costs are indexed via a multi-feed median, without any governable parameters. Abnormal price injection attempts are neutralized by the consolidation strategy.

- **Temporal sharding and burst attack:** the use of an EWMA synchronized on a common clock mitigates the effects of coordinated bursts, by cushioning activity peaks over a sliding period.

6.2. Intrinsic systemic properties

Beyond technical resistances, architecture expresses a series of fundamental systemic properties:

- **Reputation locality:** by refusing any global assessment body, the protocol dissolves the unique targets. Any attempt at manipulation requires the simultaneous alteration of multiple perspectives, heterogeneous and independent. From a heuristic point of view, the locality makes it possible to develop a **robust decision-making policy in uncertainty**. The protocol encourages each agent to adopt a locally calculated trust policy, based on audited metrics - such as published integrity, the degree of implicit subsidy, or signs of collusion. This reduces cognitive costs (by avoiding total mapping of the graph), speeds up decision-making (trust becomes algorithmic and contextual), and prevents systemic attacks (no throne exists to be captured). **Thus, each commitment is made upstream of the payment, according to a heuristic located, based on the historical-economic coherence of the partner.**

- **Economy of commitment:** Reputation is only built through effective exposure to risk. Each gain in confidence requires a cost (burn, transactions) and a risk-taking (stakes, potential slash).
- **Neutral monetization:** The x402 protocol provides the function of regulation without bias. Transactional quality is quantified independently via audited metrics published on DKG. There is no intrinsic reputational pension, but only economically deserved confidence trajectories.

6.3. Extrinsic systemic properties

In a multi-chain environment, resilience acquires a second layer: any attempt to falsify critical states - EWMA variables, arbitrary stakes, cartel flags - would require not only to compromise NeuroWeb, but to capture the entire Relay Chain.The economic and operational cost of such an attack becomes asymptomatically prohibitive, making the Polkadot–NeuroWeb whole a deterrent bastion in the face of institutional capture.

In addition, the circulation of interactions via XCM amplifies the sterility of link farms: the absence of transversal economic commitment makes visible, at the network level, the emptiness of fraudulent topologies.

Inter-chain circulation (via XCM) connects the Trust layer (NeuroWeb) to various parachains of the Polkadot ecosystem; simultaneously, the Knowledge layer reflects the verdicts and metrics, making any attempt at manipulation capturable, measurable and dissuasive.

The networking of trusted networks is based on the epistemological paradigm of the DKG. Each arbitration guild, as a subject of integrity, publishes its performance metrics in a versioned, time-dated and publicly searchable way. This distributed register allows interoperability without standardization of standards: reputation does not circulate as a transferable asset, but can be questioned contextually. This results in a topology in which each local graph, although autonomous, remains interoperable. Trust thus becomes a form of knowledge distributed, indexed to a proof, and cross-linked in a pluralistic graph.

7. Evaluation and metrics

The evaluation of the protocol is based on a series of systematic automatable tests, simulating different attack vectors, deviating behaviors and extreme scenarios. Twelve experimental rounds make it possible to objectify the robustness and precision of the proposed mechanisms.

Test benches (representative extracts)

1. **Flooding Sybil** : Generation of 10^6 identities in 24 hours. We draw the curve C_t of the cost of registration, and its full $\sum C_t$ Depending on λ_t, to check the superlinear growth of the total cost under overload.

2. **Intra-cluster link farm:** creation of 10^4 edges between agents not economically engaged. Expected result : the score LocalTrust calculated from exogenous nodes remains low, due to the absence σ significant and the saturation of $\tau(\Delta t)$.

3. **Capture of a guild :** Introduction of bias in arbitration verdicts. Measured consequences: $\text{IS}\mathscr{G} \downarrow$, $\text{SR}\mathscr{G} \uparrow$, $\text{CF}_{\mathscr{G}} = 1 \mapsto \sigma(\mathscr{G}) \to 0 \mapsto$ refusal of prior payment $\mapsto$ contraction of the flow of transactions.

4. **Artificial narrative (e.g. APY 50%):** the ratio $\text{SR}_{\mathscr{G}}$ detects the implicit subsidy; agents applying a local policy reject the guilds concerned.

5. **Inter-guild cartel :** extreme jurdict correlations and jurors' recovery trigger $\text{CF}_{\mathscr{G}} = 1$, prohibiting the use of the guild as an arbitrator.

6. **Stress oracle :** Injection of extreme data on one or more price feeds; expected stability of the aggregate median; no impact on the dynamics of λ_t.

Key indicators
For each scenario, we collect comparable and auditable metrics:

- False positive / false negative rate in the recognition of disputed transactions.

- Sybil elasticity: $\dfrac{\partial \log(\text{coût})}{\partial \log(\lambda)}$, expressing the sensitivity of costs to saturation.

- Median time to resolve disputes based on pre-payment acceptance rate.

- Entropy of selected guilds: indicator of diversity and non-centralization in the selection of referees by agents.

8. Discussion: limits and extensions

8.1. Decentralized configuration

Critical hyperparameters of the protocol — such as the ceiling by interaction M, the half-life of obsolescence T_τ, the reinjection factor α, or the registration threshold λ^* — must remain configurable locally, without a centralized governable instance. The publication of default values is useful but cannot establish a binding standard.

8.2. Preservation of privacy

System integrity must not contravene operational anonymity. It is recommended to publish only cryptographic commitments (hashes, ZK-proofs, signatures) in the DKG Knowledge Assets, allowing the verifiable verification of verdicts without disclosing the identity of the jurors.

8.3. Inter-chain interoperability

The validity of the identities from the Burn-to-Register mechanism can be extended to several networks, provided that a shared logic clock is defined (common slot or epoch) for the calculation of the EWMA. This generalization requires weak but not trivial synchronization.

8.4. Alternative Reputation Algorithms

Although the weighted PageRank is preferred here, other methods could be integrated in a modular way:

- **Katz centrality limited:** to strengthen indirect trajectories,
- **Weighted SimRank:** to infer behavioral similarities,
- **Detection of quasi-closed communities:** to report self-reinforced groups likely to collusion.

These extensions do not alter the economic core of the protocol but enrich the granularity of trust.

The natural extension of the protocol lies in not only technical interoperability, but epistemic: each parachain can import the trust signal, without importing the governance. The economic neutrality of the protocol becomes a common good, embodied in a specialized parachain, but universally accessible.

9. Minimum specifications

This section lists the fundamental technical elements allowing the operational implementation of the protocol in a short cycle. Each component is designed for modular and interoperable integration.

9.1. Smart Contracts

- **BurnToRegister :** Identity creation mechanism with dynamic taxation (EWMA, exponential surge), median price oracle, and issuance of the IdentityMinted event.

- **GuildTCR** : Stakable/slashable referee register with stake(), vote(), DisputeResolved functions, export of verdicts on DKG.

- **Middleware x402 :**Payment adapter automatically boarding the Certificate Insurance (clé : $guilde, version, \sigma$).

The BurnToRegister and GuildTCR contracts are implemented in ink! On NeuroWeb; external payments (xcUSDC, ASTR, GLMR) are routed via XCM. The x402 middleware is explicitly designed to be invoked from other parachains via a Transact message, encapsulating the insurance certificate.

9.2. Knowledge Assets (DKG)

Series of versioned objects, published by each guild and processed by the analytical engine:

- $ka : //guild/addr/yield/v < k >$: Income structure (costs vs subsidies),
- $ka : //guild/addr/disputes/v < k >$: Signed register of disputes,
- $ka : //analytics/cartel - index/v < k >$: Correlation matrices and inter-guild overlap,
- $ka : //analytics/integrity/addr/v < k >$: Score derived from performance and transparency.

All diagrams are timestamped, signed, and explicitly versioned.

9.3. Libraries and services

- LocalTrust (TypeScript / Python) : Extraction of $G_S^{(r)}$, Calculation of the matrix P, Iterations of π Until convergence.
- Service pay-per-API (x402) : ex. $leaderboard - local$ à 0,01 USD, With automatic refusal if $\sigma < \theta$.

The calculation of the LocalTrust on NeuroWeb can be called via an XCM Query by any parachain, making NeuroWeb a distributed Trust Hub. The reputation does not circulate but can be questioned, guaranteeing multi-channel coherence without duplication of state.

10. Conclusion

The combination of a non-governable identity cost, a graph of verifiable economic interactions, and a transparent arbitration market published on the DKG, establishes a distributed, local and resilient reputation.

The x402 payment protocol becomes the objectivable trace of trust, making each transaction falsifiable but verifiable.

This model - without the Throne - externalizes the evaluation of integrity, prohibits decision-making centralization, and inscribes confidence in the economy of commitment. Local policies become the only sovereigns, guided by public metrics, auditable and resistant to capture.

The three-layer architecture – Agent, Knowledge, Trust – and the Flowchart Agent → Knowledge → Trust → Knowledge, transforms trust into a systemic topology, constantly updated and shared, rather than a static attribute.

The model formalizes trust as a distributed ecology, impossible to capture and destined to irrigate a multi-chain ecosystem.

11. Executive summary

- Problem:
  - Sybil bots
  - Manipulation of reputation graphs
  - Capture of arbitrators in charge of dispute management

- Guiding principles:
  - Make identity creation expensive
  - Probative every transaction
  - Auditable the integrity of the referees

- Technical innovation:
  - Weighted local PageRank (amount, rate, insurance),
  - paiements x402 comme arêtes transactionnelles
  - guildes d'arbitrages concurrentes aux métriques publiées

- Safety:
  - resistance to floods
  - sterility of bond farms
  - endogenous devaluation of corrupt arbitration guilds

- Operational deliverables:
  - contrats BurnToRegister & GuildTCR
  - middleware x402
  - assets DKG
  - API LocalTrust.

- Strategic advantage:
  - calculable confidence without center,
  - neutral monetization,
  - cryptographic verifiability.

Added to this is the following contribution:
The explicit alignment of the protocol with Polkadot's open interoperability.
NeuroWeb acts as a specialized parachain, offering a neutral, public, non-captable Trust Layer, usable by any actor in the ecosystem via XCM.

Trust thus becomes a shared resource, not a private annuity; a systemic mechanism, not a local privilege.

Annexes

A. Pseudo-code: LocalTrust

```
function LocalTrust(seed S, radius r, topK k):
  Gs = induced_subgraph(G, ego(S, r))
  P  = row_normalize(weight_matrix(Gs))^T
  v  = normalize(stake_vector(Gs))
  π0 = v
  repeat until ||π_{t+1}-π_t||_1 < ε:
     π_{t+1} = α * P * π_t + (1-α) * v
  return topK_by_score(π_{t+1})
```

B. EWMA function (registrations / time)

$$\lambda_t = (1-\gamma)\cdot x_t + \gamma \cdot \lambda_{t-1}, \quad \gamma = 2^{-\Delta t/T_{1/2}}$$

C. Enriched x402 header (.json)

```
Assurance-Attestation: {
  "guild": "0xABC...",
  "integrity_score": 0.72,
  "subsidy_ratio": 0.18,
  "cartel_flag": false,
  "sigma": 0.69,
  "version": "1.2.0",
  "dkg_refs": [
    "ka://guild/0xABC/yield/v5",
    "ka://analytics/integrity/0xABC/v5"
  ]
}
```

D. Local policy (agent/heuristic) by default

```
if integrity_score < 0.6 or subsidy_ratio > 0.5 or cartel_flag:
   reject()
elif sigma < 0.8:
   counter_offer(price * sigma)
else:
   proceed_with_x402()
```

E. Default settings

- Maximum amount per ridge : $M = 50\,\mathrm{USD}$
- Half-life of obsolescence : $T_\tau = 90 jours$
- PageRank restart coefficient : $\alpha = 0.9$
- Saturation threshold : $\lambda* = 10^3\, h^{-1}$
- Convergence precision : $\epsilon = 10^{-8}$

Bibliography

- **Reputation algorithms (P2P) :**

a. Algorithme EigenTrust (2003), Sep Kamvar , Mario Schlosser, and Hector Garcia-Molina

b. LocalTrust by Karma3Labs :
   - ts-lens
   - py-eigentrust
   - openrank-sdk
   - Karma3Labs

- **Cryptocurrency Tokenomics:**

d. (PoB) : Proof of Burn :
   - Proof of Burn Iain Stewart
   - TRIDEnT: Building Decentralized Incentives for Collaborative Security, Nikolaos Alexopoulos , Emmanouil Vasilomanolakis , Stephane Le Roux , Steven Rowe , and Max Muhlhauser

e. Entity Registration Mechanism (ioID concept to IdentityId for agent) :
   - Iotex Github
   - Iotex Whitepaper

f. Adaptive pricing:
   - Rechained: Sybil-Resistant Distributed Identities for the Internet of Things and Mobile Ad Hoc Networks Arne Bochem, Benjamin Leiding

g. x402 Whitepaper (Coinbase, x402 Foundation)

- Risk management

a. Controversial Users demand Local Trust Metrics: an Experimental Study on epinions.com Community Paolo Massa and Paolo Avesani
b. OriginTrail DKG & Edge Node
c. Verifiable Internet for Artificial Intelligence: The Convergence of Crypto, Internet and AI, Trace Labs OriginTrail Core developers
d. The Sybil Attack, John R. Douceur
e. Foundations of Cryptoeconomic Systems, [Inondation Sybil] [fermes de liens] [prise de contrôle d'une guilde] [discours économique biaisé, (schémas de Ponzi)] Voshmgir Shermin; Zargham Michael
f. The ins and outs of decentralized autonomous organizations (DAOs) unraveling the definitions, characteristics, and emerging developments of DAOs , Marijn Janssen, Olivier Rikken, Zenlin Kwee
g. Adversarial Dynamics in Centralized Versus Decentralized Intelligent Systems, Niccolo Pescetelli , Levin Brinkmann, Manuel Cebrian
h. A prototype towards modeling visual data using decentralized generative adversarial networks, Dimitrios Kosmopoulos
i. The Impact of Adversarial Node Placement in Decentralized Federated Learning Networks , Adam Piaseczny, Eric Ruzomberka, Rohit Parasnis, Christopher G. Brinton
j. The knowledge complexity of interactive proof systems, shafi goldwasser, silvio micali, charles rackoff
k. Finding and evaluating community structure in networks, M. E. J. Newman, M. Girvan
l. Community structure in social and biological networks, Michelle Girvan, M. E. J. Newman
m. EWMA Control Charts

- **Resources for practical implementation:**

  - **Polkadot :**

    - Polkadot learn-architecture
    - Polkadot-lightpaper

    - building ai on polkadot
    - Transforming trust in the age of AI with OriginTrail

    - intro to XCM
    - XCM Guides Fees
    - XCM: The Cross-Consensus Message Format

  - **Origin Trail**

    - SDK Decentralized Knowledge Graph V8 client

    - Build neuro symbolic AI agents with OriginTrail Decentralized Knowledge Graph - Branimir Rakic

    - Knowledge Mining and dRAG examples
    - Science Paper to JSON-LD Pipeline (Desci Knowledge Mining)
    - OriginTrail dRAG DeSci Example

    - DKG Javascript SDK (dkg.js)
    - DKG Python SDK (dkg.py)

  - **NeuroWeb.AI**

    - documentation

- **Others Ressources :**
    - DAO Governance Models Explained: Token-Based vs. Reputation-Based Systems
    - a Web3 Beginner Series: Exploring the New On-Chain Payment Protocol — x402
    - How To Perform Cross-Chain Token Transfers with Polkadot XCM
    - Loi de Goodhart