ChronosRep: Entropy-Regularized Evidence Fusion and Stochastic Differential Trust Dynamics for Decentralized Identity Intelligence

Abstract

In the emerging architecture of decentralized networks, reputation serves as a critical mechanism for mediating trust and access control. However, traditional static reputation models are structurally vulnerable to *"sleeper agent"* attacks, where adversaries exploit the system's long memory to mask strategic defections. This paper introduces ChronosRep, a hybrid security framework designed to ensure data integrity by unifying off-chain identity compliance with on-chain behavioral dynamics. The framework advances computational trust through two core contributions. First, we propose the Identity Reputation Vector (IRV), which employs Entropy-Regularized Dempster-Shafer Theory to fuse heterogeneous Verifiable Credentials (VCs). This approach introduces a Belief Entropy-based weighting mechanism that robustly aggregates conflicting evidence, effectively resolving the epistemic uncertainty inherent in decentralized identity. Second, we formalize trust dynamics using an adaptive Ornstein-Uhlenbeck (OU) Jump Diffusion process. Unlike static decay models, this stochastic differential equation dynamically adjusts the mean-reversion rate based on real-time volatility, eliminating the *"reputation lag"* required for sustained manipulation. Agent-based simulations validate the framework's efficacy, demonstrating a 92.7% reduction in Time-to-Detection (TTD) and a 100% isolation rate for sleeper agents compared to state-of-the-art baselines. ChronosRep thus establishes a mathematically rigorous foundation for resilient, compliant data processes in adversarial environments.

Keywords: Data Security and Compliance, Blockchain Defense Schemes, Entropy-Regularized Information Fusion, Verifiable Credentials (VCs), Dempster-Shafer Theory, Dynamic Reputation, Self-Sovereign Identity, Time-Decayed Trust Models.

1. Introduction

The deep integration of emerging network technologies has catalyzed a transition toward a "cloud-edge-end" service architecture, where data serves as the pivotal production asset [1, 2]. In this integrated environment, particularly within Decentralized Ecosystems (e.g., DeFi and DAOs), the boundaries of trust are pushed to the network edge, where users and autonomous agents interact without centralized intermediaries [3]. While this architecture democratizes access, its complex service processes introduce heightened security risks. Traditional protection technologies, such as static access control and encryption, are insufficient to address the dynamic challenges of data integrity and behavioral compliance in these open environments. Consequently, reputation data has emerged as the de facto security metric for gauging the trustworthiness of entities, guiding critical decisions regarding capital allocation and governance power [4, 5].

However, a fundamental theoretical friction persists in the current computational trust literature, stemming from the inherent information asymmetry of decentralized networks. Unlike centralized architectures where intermediaries audit actors to minimize risk [6], decentralized ecosystems operate as "Lemons Markets" where the true quality of an agent remains private information. This creates a privacy-transparency paradox: utilizing Zero-Knowledge Proofs (ZKPs) to preserve privacy often obfuscates the historical volatility data required for accurate risk assessment [7]. Existing reputation models fail to resolve this paradox because they reduce trust to a monolithic scalar value. It is important to note that they lack a unified mathematical basis to simultaneously handle epistemic uncertainty ambiguity arising from conflicting, privacy-preserved credentials and aleatory uncertainty inherent stochastic volatility of behavior over time [8]. Traditional frameworks, such as Beta Reputation or EigenTrust, rely on static aggregation that cannot distinguish between a lack of information ignorance and negative information conflict [9].

This theoretical void manifests acutely in the "sleeper agent" attack vector. We redefine this threat not merely as malicious behavior, but as a strategic exploitation of temporal belief inertia. In this scenario, an adversary leverages the system's inability to model the entropy of behavioral signals, accumulating reputation during a compliant "investment phase" to mask a catastrophic future defection. Because static models weigh historical compliance equally with recent actions, they suffer from significant "reputation lag", allowing adversaries to execute attacks before the trust metric reflects the new adversarial state.

To address these challenges, we must bridge the gap between off-chain identity compliance and on-chain behavioral security. Existing mechanisms typically treat these as separate domains: Self-Sovereign Identity (SSI) handles static credentials (the "Brief"), while reputation systems track dynamic scores [10]. A robust protection technology must fuse these heterogeneous data sources. It requires an architecture capable of reasoning under uncertainty to filter out low-quality or conflicting identity claims (Sybil attacks) while simultaneously applying

rigorous temporal decay to neutralize the economic incentives of sleeper agents [11, 12].

In this paper, we introduce ChronosRep, a hybrid theoretical and technical framework designed to ensure data security and compliance in decentralized networks. ChronosRep reengineers the trust mechanism by treating reputation not as a static asset, but as a dynamic, perishable data stream that requires continuous verification. By integrating Verifiable Credentials (VCs) from the network edge with an adaptive, stochastic on-chain model, ChronosRep provides a resilient defense against data corruption and strategic manipulation. We propose the following key contributions to the field of data protection:

- We propose the IRV to bridge the gap between epistemic uncertainty (ambiguity in identity sources) and aleatory uncertainty (stochastic behavioral volatility). Using Entropy-Regularized Dempster-Shafer Theory, IRV fuses heterogeneous VCs while suppressing Sybil-driven noise through a Belief Entropy-based weighting mechanism.
- We formalize a dynamic defense scheme using an Ornstein-Uhlenbeck (OU) Jump Diffusion process. This model dynamically adjusts data sensitivity based on behavioral volatility, ensuring that the system reacts instantaneously to security breaches (jumps) while accommodating benign operational noise. This effectively eliminates the "reputation lag" exploit.
- We validate ChronosRep through extensive agent-based simulations, demonstrating its effectiveness as a protection technology. The results show a 92.7% reduction in threat detection time compared to static baselines, confirming its ability to enforce data integrity and isolate sleeper agents in complex adversarial environments.

The remainder of this paper is organized as follows: Section 2 reviews related work in computational trust, decentralized identity, and data fusion. Section 3 details the threat model, specifically focusing on data integrity risks posed by sleeper agents and Sybil attacks. Section 4 presents the methodology of ChronosRep, detailing the entropy-regularized fusion engine and the stochastic decay mechanism. Section 5 provides a security analysis of the proposed defense scheme. Section 6 presents the experimental validation and comparative analysis against state-of-the-art baselines. Finally, Section 7 discusses the broader implications for compliant data processes, and Section 8 concludes the paper.

2. Related Work

2.1. Computational Trust, Identity Modelling, and Temporal Dynamics

Computational trust frameworks in distributed environments traditionally derive reputation from interaction histories, often assuming that past behavior can be aggregated into a stable, monotonic signal. Foundational systems such as the Beta Reputation System adopt Bayesian updating over binary outcomes, interpreting each interaction as evidence that incrementally reshapes posterior trust distributions [14]. EigenTrust extends this intuition by propagating locally assigned trust values through a global network graph in a manner reminiscent of PageRank, producing a consensus score that captures transitive endorsement [9]. These models establish early formalisms for trust computation but remain insensitive to temporal structure. A long-term adversary that behaves cooperatively for an extended period before deviating a phenomenon often described as a sleeper agent pattern [15] is not meaningfully distinguished from an inherently honest participant.

Efforts to introduce time dependence led to dynamic frameworks such as FIRE [16], which apply exponential decay to prioritise recent observations. While FIRE incorporates multiple trust inputs, including witness information and role-based expectations, it is explicitly tailored to environments where interaction data are dense and structured. Such assumptions diverge significantly from the properties of contemporary decentralized finance ecosystems, where interactions are pseudonymous, sparse, and highly heterogeneous.

These constraints have motivated the emergence of on-chain behavioral reputation systems, notably Spectral Finance and ARCx. Spectral Finance constructs a MACRO Score by analysing on-chain transactional parameters, including repayment history, wallet longevity, and collateral management [1], while ARCx generates a DeFi Passport using a related set of heuristics [2]. Despite their utility, both systems depend entirely on observable wallet-level behavior and implicitly treat blockchain addresses as stable identities. Such design choices exclude off-chain attestations and prevent the incorporation of richer identity semantics.

To systematically evaluate the gap between these foundational models and modern requirements, Table 1 provides a structural comparison across three dimensions: uncertainty handling, data source provenance, and resilience against sleeper agents. While classical systems like EigenTrust and PeerTrust rely on deterministic or probabilistic aggregation suitable for P2P file sharing, they lack the stochastic mechanisms required to secure high-value DeFi transactions against temporal attacks.

2.2. Evidence Fusion, Uncertainty, and Conflict Resolution in Decentralized Trust

A second thread of research concerns the problem of aggregating heterogeneous and potentially contradictory inputs into a coherent trust signal. Behavior-only systems such as Spectral and ARCx utilise either machine learning pipelines or deterministic scoring rules that lack formal semantics for uncertainty. These methods implicitly assume that all input signals are comparable, trustworthy, and free of conflict, overlooking scenarios in which identities possess ambiguous or competing attestations.

Dempster–Shafer Theory (DST) provides a principled alternative by enabling belief assignment over sets of hypotheses

[1] https://docs.spectra.finance/

[2] https://docs.architex.ai/architex-ecosystem

Table 1: Comparative Taxonomy of Trust Models: Uncertainty Handling and Adversarial Resilience.

System Model	**Fusion Strategy**	**Evidence Source**	**Resilience**
EigenTrust [9]	Deterministic Aggregation[a]	Global Interaction Graph	*Low*
PeerTrust [13]	Weighted Averaging[b]	Peer Feedback & Context	*Low-Mod*
Beta Reputation [14]	Beta Probability Density (PDF)[c]	Binary Outcomes	*Low*
Spectral Finance	Machine Learning Scoring	On-chain Transaction History	*Moderate*
ChronosRep (Ours)	**Stochastic & Evidential**[d]	**Hybrid (VCs + Behavior)**	***High***

Note: Comparison of architectural properties and failure modes.
[a] *Deterministic:* Treats trust as a transitive flow; fails to distinguish ignorance from conflict (Subject to "Long Memory" attacks).
[b] *Weighted:* Adds credibility factors but lacks formal epistemic uncertainty modeling (Subject to Whitewashing).
[c] *Beta PDF:* Assumes static behavior distribution; slow reaction to sudden strategic defections.
[d] *Stochastic (Ours):* Uses Entropy-Regularized Dempster-Shafer for conflict resolution and OU-Jump Diffusion for volatility response. This enables instantaneous "trust collapse" upon detecting sleeper agents (Zero-lag detection).

and preserving explicit representations of ignorance or ambiguity [17]. DST is particularly suited to environments where evidence sources differ in reliability or scope, as is the case with VCs issued by institutions of varying credibility. Classical formulations allow mass to be assigned to composite states rather than singletons, enabling the system to encode uncertainty directly. Dempster's rule of combination then merges independent pieces of evidence while quantifying the associated conflict, which becomes an operational signal rather than noise [18]. Existing decentralized reputation frameworks rarely incorporate mechanisms of this nature, relying instead on unstructured aggregation or Boolean gating. ChronosRep adopts DST to represent and combine identity-derived evidence in a mathematically grounded manner. Conflicting attestations are resolved through a belief-based combination process, allowing the system to moderate or even suppress unreliable statements while amplifying consistent ones. This approach differs from both binary VC-gated access control and graph-based trust systems such as SybilGuard [19], which treat identity either as a Boolean constraint or as a purely structural artefacts. Within the taxonomy presented in Table 1, ChronosRep occupies a position characterised by uncertainty-aware evidence fusion, issuer-sensitive belief weighting, and the ability to integrate off-chain identity with on-chain behavioral traces.

3. Background

3.1. Self-Sovereign Identity in Decentralized Systems

The absence of a central authority in decentralized environments creates a structural trust deficit, especially when participants interact pseudonymously across permissionless networks. Traditional identity infrastructures including state-backed registries, OAuth providers, and certificate authorities fundamentally rely on intermediaries to authenticate claims. Such architectures conflict with the autonomy, censorship resistance, and privacy requirements of blockchain ecosystems. SSI has therefore emerged as a cryptographically grounded paradigm that enables users to control their identifiers and selectively disclose verifiable attributes without depending on centralized databases.

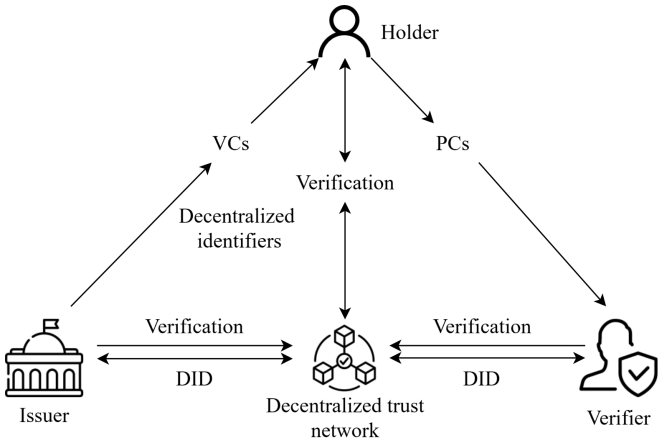

Figure 1: SSI Trust Triangle: Credential issuance and verification occur between Issuer, Holder, and Verifier using DIDs and VCs.

At the core of SSI is the Decentralized Identifier (DID), a persistent, ledger-resolvable identifier controlled exclusively by the holder. A DID resolves to a DID Document that lists public keys, authentication mechanisms, and service endpoints, forming the cryptographic foundation that enables credential issuance and verification flows among the three SSI roles: Issuer, Holder, and Verifier (Figure 1). In practice, these documents are anchored on tamper-resistant infrastructures such as public blockchains or distributed ledgers, providing a decentralized discovery mechanism for verifiers [20]. This architecture supports a wide range of real-world applications, including cross-domain identity portability, secure IoT device authentication, and decentralized access management.

Building on DIDs, the VCs model provides a structured mechanism for issuing cryptographically signed attestations. VCs can encode academic qualifications, KYC/AML compliance, organizational affiliations, or professional certifications. Holders store these credentials locally typically in a digital wallet and present them to verifiers using Verifiable Presentations (VPs), which support selective disclosure and privacy-preserving proof of attribute possession. Recent studies highlight that selective disclosure is essential for regulatory

compliance, decentralized financial onboarding, and privacy-preserving age or identity verification. Importantly, the verification workflow does not require real-time issuer involvement: a verifier simply retrieves the issuer's DID Document, checks the digital signature, and confirms the holder's key ownership. This decentralized verification loop enables scalability in high-volume settings such as decentralized marketplaces and cross-chain authentication systems [21].

By anchoring identity to verifiable attestations rather than disposable key pairs, SSI significantly increases the cost of Sybil attacks. Classical analyses show that Sybil resistance improves when identity establishment requires externally verifiable claims rather than free cryptographic identifiers [22]. Within the SSI model, fabricating multiple credible identities requires obtaining legitimate credentials from trusted issuers, which imposes substantial financial, procedural, or social cost. This property has led to the adoption of SSI-based identities in decentralized lending, DAO governance, decentralized workforce platforms, and Web3 access-control systems, where higher-assurance identity anchors are needed to prevent manipulation.

Nevertheless, SSI alone cannot address the temporal dimension of trust. An entity with strong credentials may still behave maliciously after a period of benign activity, and the SSI layer provides no mechanism for differentiating long-term strategic deception from honest participation. This vulnerability is well documented in the literature on dynamic trust and reputation systems, which highlights the inability of purely identity-based models to resist long-horizon adversarial strategies such as sleeper-agent attacks [23, 15]. These limitations motivate the need for temporally adaptive and behavior-aware trust mechanisms. ChronosRep builds upon the SSI foundation, extending beyond static identity assurance by incorporating dynamic behavioral evidence, recency weighting, and adversarial adaptivity to provide a more comprehensive model of trust in decentralized ecosystems.

3.2. Threat Models in Reputation Systems

SSI provides a robust foundation for trust and serves as a critical prerequisite for dynamic, reputation-based systems like ChronosRep. Nevertheless, the presence of a verifiable identity layer does not eliminate all vectors of attack. Adversaries may still exploit structural weaknesses in decentralized trust systems, particularly through attacks that manipulate the identity layer (breadth) or the temporal structure of reputation accumulation (depth). We focus on two canonical threats: Sybil attacks and sleeper agent attacks.

The *Sybil attack*, as originally formalized by Douceur [22], arises when a single adversary generates a multitude of pseudonymous identities, allowing them to appear as independent entities in the network. These counterfeit nodes, known as Sybils, can outnumber honest participants and manipulate the system's state by forming a coalition under the attacker's control. The process is illustrated in Figure 2, which contrasts the attacker's ability to inflate influence in systems lacking identity verification against the increased cost and conflict detection enforced by SSI and VCs.

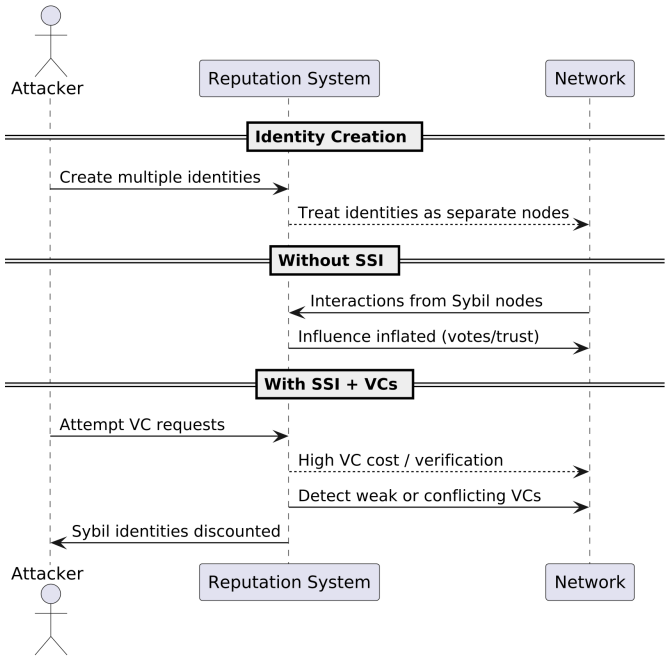

Figure 2: Sybil Attack Flow: An attacker generates multiple identities to inflate influence. In systems without SSI, these identities are treated independently. In SSI-enabled environments, VCs impose acquisition costs and enable conflict detection, discouraging Sybil clusters and reducing their effective impact.

The attack is especially effective in systems lacking identity verification or where identity creation is trivial (e.g., public key registration). In such scenarios, an attacker can inflate their voting power in DAOs, skew feedback scores in marketplaces, or falsify redundancy in distributed storage networks. By enforcing identity anchoring through VCs issued by trusted authorities, SSI substantially increases the cost of mounting a successful Sybil attack. Unlike raw key-pair generation, obtaining valid VCs for each identity requires real-world effort or deception, thus imposing a meaningful deterrent.

A more subtle yet equally dangerous threat is the sleeper agent attack, also known as the *"reputation lag"* [24]. In this attack, the adversary does not create multiple identities. Instead, they cultivate a single identity over time, behaving honestly for an extended period before strategically defecting. The attack unfolds in two distinct phases: an *investment phase*, where the agent performs well to accumulate trust, followed by an *exploitation phase*, where they act maliciously once sufficient reputation is acquired. As shown in Figure 3, static trust models react too slowly, allowing the attacker to exploit the accumulated reputation.

Formally, consider a reputation score R_t at time t. In static systems, reputation is typically computed as a simple average of all historical observations $R_t^{\text{static}} = \frac{1}{t}\sum_{k=1}^{t} S_k$, where $S_k \in \{0, 1\}$ encodes the binary outcome (e.g., success/failure) of interaction k. If the attacker behaves honestly during the first T interactions ($S_k = 1$ for $k \leq T$) and then begins to defect ($S_k = 0$ for $k > T$), the score evolves as $R_{T+n}^{\text{static}} = \frac{T}{T+n}$, indicating a slow, asymptotic decay. The attacker's past behavior continues to dominate, causing a lag in detection and granting the attacker a window to

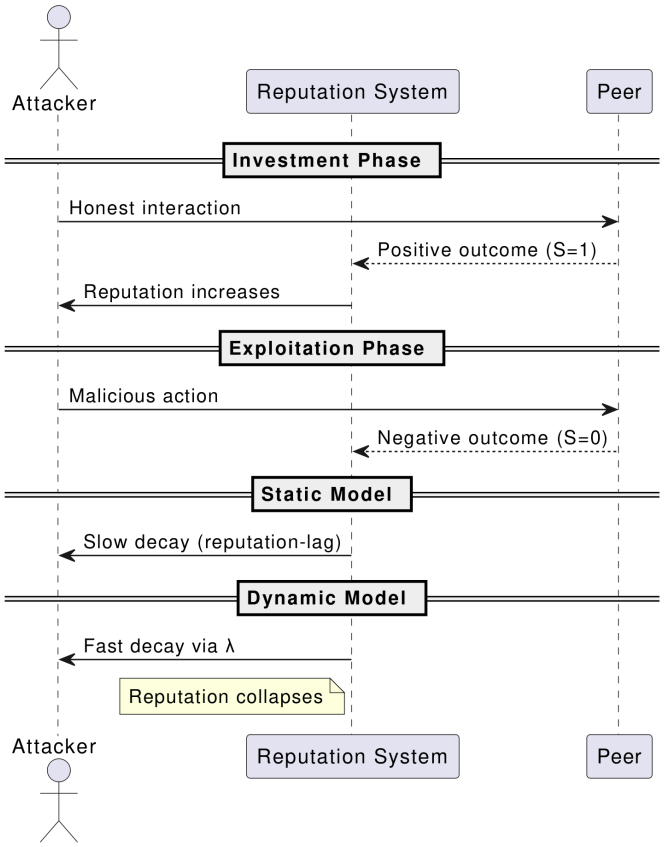

Figure 3: Sleeper Agent Attack: The adversary first builds a reputation through prolonged honest behavior before switching to malicious actions. Static reputation models exhibit slow decay, enabling an exploitation window, whereas dynamic decay rapidly collapses trust.

continue operating maliciously.

Of particular significance is the fact that this mathematical lag implies a structural vulnerability in static trust systems: they conflate stability with trustworthiness. By treating reputation as a cumulative asset rather than a current state, systems like EigenTrust inadvertently incentivize the "sleeper agent" strategy. The attacker essentially acts as an arbitrageur of trust latency, purchasing reputation cheaply over time (Investment Phase) to spend it abruptly on a high-value attack, capitalizing on the system's inability to model the entropy of the behavioral signal.

To mitigate this vulnerability, ChronosRep adopts a dynamic reputation model based on exponential time decay. The reputation score is updated recursively as $R_t^{\text{dynamic}} = (1-\lambda)R_{t-1} + \lambda S_t$, where $\lambda \in (0,1)$ is the decay factor that determines the system's memory horizon. For example, with $\lambda = 0.25$, after just five consecutive failures ($S_k = 0$), the reputation score satisfies $R_{T+5}^{\text{dynamic}} \leq (1-\lambda)^5 R_T \approx 0.24 R_T$, demonstrating a steep drop in trust. This responsiveness sharply contrasts with the inertia of static systems and is instrumental in closing the window of exploitability.

3.3. Dempster-Shafer Theory for Evidence Aggregation

To transform heterogeneous VCs into a quantifiable trust signal, ChronosRep employs DST. DST is a mathematical theory of evidence that generalizes Bayesian probability to reason under uncertainty and ignorance.[25] This is critical in decentralized settings where evidence sources (VC issuers) vary in credibility and may provide conflicting information.

Let θ be the frame of discernment, a finite set of mutually exclusive hypotheses (e.g., $\theta = \{\text{Proficient}, \text{Not Proficient}\}$). DST works with a Basic Belief Assignment (BBA) or mass function m, which assigns a belief mass to every subset of θ. The value $m(A)$ represents the belief committed exactly to the proposition represented by the subset $A \subseteq \theta$. A key feature is that mass can be assigned to θ itself, representing total ignorance or uncommitted belief. For example, a VC from a top university might be translated into a BBA $m(\{\text{Proficient}\}) = 0.8, m(\{\text{Not Proficient}\}) = 0, m(\Theta) = 0.2$.

This signifies strong belief in proficiency, no direct evidence against it, and 20% uncertainty. When multiple pieces of evidence (e.g., multiple VCs) are available, their corresponding BBAs (m_1, m_2) can be combined using Dempster's Rule of Combination: $m_{1\oplus 2}(A) = \frac{1}{1-K}\sum_{B\cap C=A} m_1(B)\cdot m_2(C)$ where $K = \sum_{B\cap C=\emptyset} m_1(B)\cdot m_2(C)$ is the *conflict coefficient*, measuring the degree of contradiction between the evidence sources. The final belief in a hypothesis A is given by the Belief function $Bel(A) = \sum_{B\subseteq A} m^*(B)$, where m^* is the combined BBA. This belief value is then used as the numeric score for the corresponding component in the IRV, providing an evidence-grounded representation of trust.

4. Methodology

4.1. Taxonomy of Uncertainty: Epistemic vs. Aleatory

To rigorously address the limitations of existing trust models, ChronosRep is built upon a fundamental distinction between two distinct types of uncertainty inherent in decentralized systems: *Epistemic Uncertainty* and *Aleatory Uncertainty*.

Epistemic Uncertainty (The "Identity" Problem) refers to uncertainty resulting from a lack of knowledge or ambiguity regarding the static attributes of an agent. In decentralized systems, this arises when VCs are missing, possess high entropy (vagueness), or exhibit conflict between issuers. Fundamentally, epistemic uncertainty is theoretically reducible if more or better evidence is acquired. ChronosRep addresses this via *Entropy-Regularized Dempster-Shafer Theory*, which explicitly models "Ignorance" (Θ) and resolves evidence conflict (K).

Aleatory Uncertainty (The "Behavior" Problem): refers to the inherent stochastic volatility in a physical process over time. Even a perfectly honest agent will exhibit behavioral variance due to network latencies, transaction intervals, or market fluctuations. Unlike epistemic uncertainty, aleatory uncertainty is *irreducible* and must be modeled as a random process. We address this using the *Ornstein-Uhlenbeck Jump Diffusion* process, treating trust as a volatile asset that fluctuates around a mean, capturing both normal operational noise (Diffusion) and sudden adversarial shocks (Jumps).

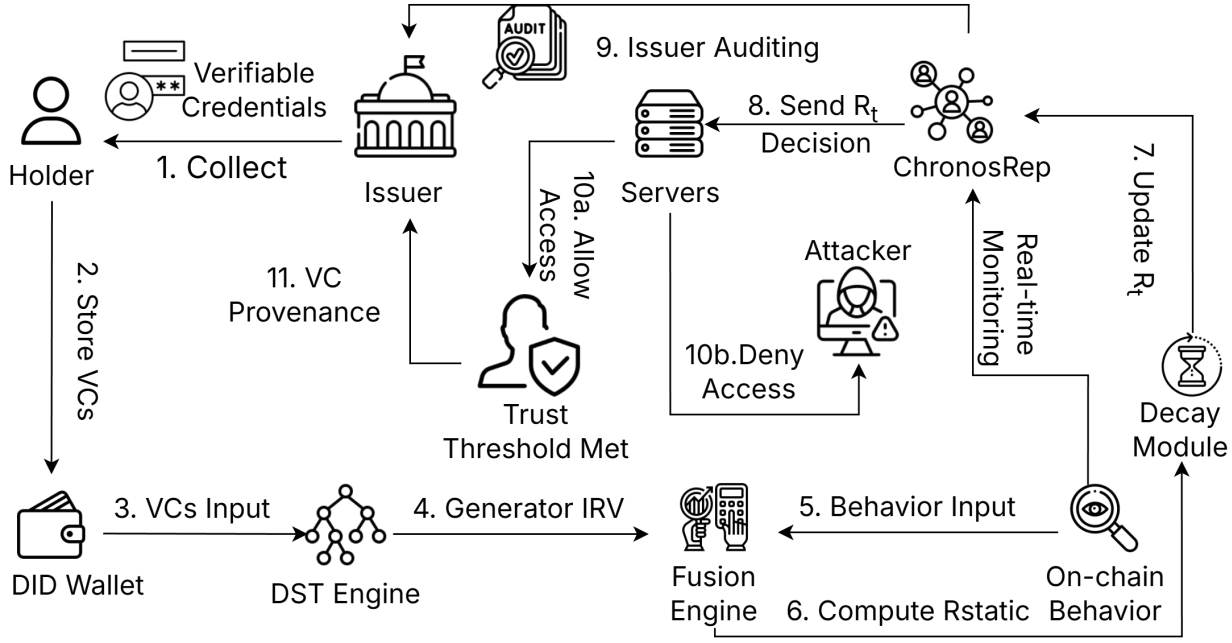

Figure 4: The ChronosRep System Architecture. The workflow begins with the Holder acquiring VCs. The DST Engine processes these to generate an IRV, which is fused with on-chain behavioral data in the Fusion Engine to compute a static score (R_{static}). A Decay Module applies temporal smoothing to produce the final dynamic reputation score (R_t), which informs the access control decision.

By explicitly decoupling these domains, ChronosRep avoids the structural pitfall of systems like EigenTrust, which attempt to solve both types of uncertainty with a single probabilistic scalar, often leading to slow convergence and the "Long Memory" vulnerability.

4.2. System Architecture Overview

The ChronosRep architecture, illustrated in Figure 4 of the original manuscript, integrates verifiable identity with real-time behavioral monitoring to derive dynamic, attack-resilient reputation scores. The system is structured around an SSI-based identity layer, a trust computation engine that generates the IRV and fuses it with behavioral data, and a time-decay module that ensures temporal sensitivity.

At its core, ChronosRep builds upon the SSI paradigm. A participant (referred to as the Holder) obtains a set of VCs from certified Issuers. These credentials, which may include attributes such as legal identity, academic background, or financial history, are stored within a decentralized identity wallet managed by the Holder. Once a set of credentials is available, they are submitted to the reputation computation pipeline.

The computation process begins with the DST Engine, which applies DST to fuse potentially conflicting pieces of evidence from multiple VCs. The outcome is IRV, a structured, quantifiable representation of the Holder's trust attributes. This vector serves as a foundational component for reputation scoring.

Simultaneously, on-chain behavior is continuously monitored via a dedicated module that captures interactions such as transaction patterns, smart contract deployments, or governance participation. These behavioral signals are routed into the Fusion Engine, where they are integrated with the IRV to compute a baseline score, denoted R_{static}. This fusion balances "who the user is" (via credentials) with "how the user behaves" (via observed actions).

Rather than treating reputation as a static quantity, ChronosRep applies a time-sensitive update mechanism. A Decay Module performs exponential smoothing over time, updating the overall reputation score R_t at each interval. This ensures that recent behavior carries more weight than historical activity, thereby enhancing the system's ability to detect sudden shifts in trustworthiness, a key defense against sleeper-agent attacks.

Finally, the updated score R_t is forwarded to the ChronosRep Decision Layer, which evaluates whether a user's current reputation meets the required threshold for access. If the score is sufficient, the system grants access to services or protocol functions; otherwise, access is denied. All decisions can be transparently audited, and provenance queries can be performed to trace the origin of specific credentials back to their issuers, enabling defense-in-depth through issuer accountability.

4.3. Identity Computation via Entropy, Information Density, and Conflict-Normalized Evidence Fusion

The computation of identity trust in decentralized ecosystems requires a theoretical foundation capable of quantifying uncertainty, measuring informational sharpness, and regulating evidential influence in the presence of partial conflict. Classical formulations relying solely on the Dempster–Shafer rule provide an elegant but incomplete solution: they fail to distinguish between inherently ambiguous credentials and sharply defined ones, they collapse when the conflict coefficient approaches unity, and they do not reflect the information-theoretic content intrinsic to each VC.

To overcome these structural limitations, the mathematical model adopted in ChronosRep introduces a multilayered uncertainty calculus built from Belief Entropy, Information Density, Belief Divergence, and a conflict-normalized compensation mechanism. These constructs operate jointly to produce a stable fused mass function whose induced IRV possesses semantic clarity, robustness under contradiction, and quantifiable informational grounding.

The belief assignment associated with a credential VC_i is denoted by $m_i : 2^\Theta \to [0, 1]$, where $\Theta = \{T, M, U\}$ represents the attribute space of trustworthy, malicious, and uncertain states, and the distribution satisfies $\sum_{A\subseteq\Theta} m_i(A) = 1, m_i(\varnothing) = 0$. The inherent uncertainty of a VC is measured using the belief entropy, which intuitively quantifies the system's confusion when evidences conflict: high entropy indicates that the available VCs are contradictory or vague and thus should incur a penalty in the final trust score to avoid over-confidence in ambiguous identities. Formally, it is defined as:

$$E_d(m_i) = -\sum_{A\subseteq\Theta} m_i(A) \log_2\left(\frac{m_i(A)}{2^{|A|} - 1}\right), \tag{1}$$

which accounts simultaneously for the numerical magnitude of the belief and the cardinality of its hypothesis set. Larger subsets A contribute greater entropy because they encode weaker commitments. This enables the identity computation framework to differentiate between tightly focused claims ($|A| = 1$) and ambiguous multi-element claims ($|A| \geq 2$), even when their mass values coincide.

In addition to raw entropy, the system evaluates a normalized entropy measure defined as

$$\widetilde{E}_d(m_i) = \frac{E_d(m_i)}{\log_2(2^{|\Theta|} - 1)}, \tag{2}$$

bounded strictly within $(0, 1)$, thereby enabling direct comparison between credentials regardless of frame size. This normalized entropy directly influences the evidential trust weight of the credential. The weight takes the continuous form

$$w_i = \frac{\exp\left(-\beta\, \widetilde{E}_d(m_i)\right)}{\sum_{j=1}^{n} \exp\left(-\beta\, \widetilde{E}_d(m_j)\right)}, \tag{3}$$

where $\beta > 0$ is a temperature parameter controlling sensitivity to uncertainty. Small values of β yield smoother weight distributions, whereas larger values emphasize sharper discrimination against high-entropy evidence. In practice, this weighting mechanism functions as an automated "Quality Gate." By assigning negligible influence ($w_i \to 0$) to credentials with high entropy such as vague or non-specific claims often used by Sybil attackers the system effectively filters out low-quality data before it can dilute the final trust score. The regulated mass function becomes

$$m_i'(A) = w_i\, m_i(A), \tag{4}$$

thereby preserving the structural semantics of the VC while mitigating the distortive influence of low-information claims.

A complementary construct to entropy is the *Information Density* of a VC, defined as

$$\mathrm{ID}(m_i) = \sum_{A\subseteq\Theta} \frac{m_i(A)}{2^{|A|}} \log_2\left(1 + \frac{1}{|A|}\right), \tag{5}$$

which penalizes mass assigned to large subsets and rewards highly concentrated assignments. While entropy quantifies uncertainty, information density quantifies informational sharpness. The interplay between (2) and (5) yields a two-dimensional informational profile of each credential.

To measure the change induced by including a VC in the fusion process, the model defines the *Information Volume* as

$$\mathcal{V}_i = E_d(m^\star_{-i}) - E_d(m^\star), \tag{6}$$

which captures the entropy reduction occurring when VC_i is incorporated into the aggregated mass. A related but more expressive measure of information contribution is the Expected Information Gain (EIG):

$$\mathrm{EIG}(VC_i) = \sum_{A\subseteq\Theta} m_i(A) \log_2 \frac{m^\star(A)}{m^\star_{-i}(A)}, \tag{7}$$

which parallels the Kullback–Leibler divergence but is adapted to belief structures that operate on subset-based probability assignments rather than standard distributions. Positive values of (7) indicate that the VC meaningfully improves the concentration of the fused belief.

The comparative effects of these uncertainty measures are presented in Table 2, which demonstrates how different VC structures induce variability across entropy, normalized entropy, information density, and expected information gain.

Table 2: Comparison of entropy, normalized entropy, information density, and expected information gain for representative VC configurations.

VC Structure	$E_d(m)$	$\widetilde{E}_d(m)$	**ID**(m)	**EIG**(VC_i)
Sharp singleton belief	0.04	0.07	0.91	0.48
Moderate ambiguity	0.51	0.63	0.44	0.12
High structural vagueness	0.89	0.92	0.21	0.04
Near-uniform ambiguity	1.21	1.00	0.09	0.01

Fusion proceeds by evaluating the conflict coefficient

$$K(m_i', m_j') = \sum_{B\cap C=\varnothing} m_i'(B)\, m_j'(C), \tag{8}$$

which quantifies inconsistent belief assignments between two regulated mass functions. When aggregated across all VCs, conflict plays a central role in the combination rule. Instead of collapsing under high conflict, the system applies a normalized compensation factor that maintains stability:

$$m^\star(A) = \frac{\sum_{B\cap C=A} m_i'(B)\, m_j'(C)}{1 - K(m_i', m_j') + \epsilon}, \tag{9}$$

where ϵ is a machine-epsilon constant to ensure numerical stability when $K \to 1$. Under the *Closed World Assumption* [26, 27], the denominator $(1 - K)$ acts as a normalization constant that redistributes the conflicting mass to the remaining valid hypotheses. This ensures that $\sum_{A \subseteq \Theta} m^\star(A) = 1$, preserving the probabilistic integrity of the fused trust score even in highly adversarial environments.

From this fused distribution, the belief assigned to attribute A is computed as

$$\mathrm{Bel}_{m^\star}(A) = \sum_{B \subseteq A} m^\star(B), \tag{10}$$

which becomes the coordinate of IRV. For d attributes, this produces $\vec{v}_{\mathrm{IRV}} = [\mathrm{Bel}_{m^\star}(A_1), \ldots, \mathrm{Bel}_{m^\star}(A_d)]$. To capture the interaction between multiple credentials and their contribution to global uncertainty reduction, we construct the *Evidence Synergy Matrix*

$$\mathcal{T}_{ij} = \mathrm{EIG}(VC_i) + \mathrm{EIG}(VC_j) - \mathrm{EIG}(VC_i, VC_j), \tag{11}$$

which measures the mutual reinforcement or redundancy between credentials. A high value of $\mathcal{T}_{ij}$ indicates synergistic evidence, while low values indicate redundant or adversarial information. Representative tensor values are provided in Table 3.

Table 3: Evidence Synergy Matrix $\mathcal{T}_{ij}$ for pairs of VCs with varying certainty profiles. Higher values indicate stronger mutual reinforcement.

	Credential Interaction Profile		
Metric	**Sharp-Sharp**	**Sharp-Vague**	**Vague-Vague**
Synergy Score ($\mathcal{T}_{ij}$)	0.62	0.21	0.04

The computational implementation of these constructs is outlined in Algorithm 1, which organizes entropy computation, weight derivation, information-volume calculation, conflict-normalized fusion, and belief extraction into a mathematically coherent pipeline.

Algorithm 1 Entropy–Regularized Evidence Fusion and IRV Construction

Require: $\{m_i\}_{i=1}^n$, attribute set $\{A_k\}_{k=1}^d$
1: Compute $E_d(m_i)$ via (1) and $\widetilde{E}_d(m_i)$ via (2)
2: Compute weights w_i using (3)
3: Obtain m_i' using (4)
4: Fuse evidence using (9) to obtain $m^\star$
5: Compute $\mathcal{V}_i$ via (6) and $\mathrm{EIG}(VC_i)$ via (7)
6: Compute IRV coordinates using (10)
7: **return** $\vec{v}_{\mathrm{IRV}}$

The final layer of mathematical refinement evaluates the Evidence Synergy Matrix using Algorithm 2, enabling the system to assess multi-VC interactions.

The combined use of Belief Entropy, Information Density, Expected Information Gain, conflict-normalized fusion, and the Evidence Synergy Matrix provides a mathematically robust identity computation framework. Each VC is assessed not only by its raw belief assignments but also by the structural uncertainty encoded in its subset cardinalities, its expected contribution to global entropy reduction, and its synergistic or redundant relationship with other credentials. The resulting IRV thus encodes identity with a degree of mathematical precision unattainable by classical DST formulations.

Algorithm 2 Computation of the Evidence Synergy Matrix

Require: VCs with BPAs $\{m_i\}$
1: **for** each pair (i, j) **do**
2: Compute $\mathrm{EIG}(VC_i)$ and $\mathrm{EIG}(VC_j)$ from (7)
3: Compute joint gain $\mathrm{EIG}(VC_i, VC_j)$
4: Compute $\mathcal{T}_{ij}$ via (11)
5: **end for**
6: **return** Tensor $\mathcal{T}$

4.4. Dynamic Reputation Evolution via Mean-Reverting Stochastic Differential Equations

The temporal evolution of trust in a decentralized ecosystem must be treated as a stochastic process rather than a deterministic trajectory. In adversarial environments, reputation must respond elastically to benign fluctuations while collapsing rapidly under malicious shocks. A linear exponential decay of the form $(1 - \lambda)$ fails to capture these phenomena, since it imposes a rigid deterministic trend, lacks any mechanism for volatility absorption, and is unable to represent abrupt structural breaks induced by malicious actions. To overcome these limitations, ChronosRep formulates reputation as the solution to a mean-reverting stochastic differential equation (SDE) [28], specifically an Ornstein-Uhlenbeck (OU) process [29] augmented with a jump component. This representation captures the interplay between long-term trust consistency, environmental noise, and catastrophic discontinuities.

Let X_t denote the continuous-time reputation state at time t. Instead of being updated by discrete recursion, the system models X_t as the solution to the SDE

$$dX_t = \theta(\mu - X_t)\, dt + \sigma\, dW_t + J\, dN_t, \tag{12}$$

where the drift term encodes mean reversion, the diffusion term captures continuous stochastic perturbations, and the jump term models sudden adversarial deviations. The parameter $\theta > 0$ represents the rate at which X_t is pulled back toward μ, which corresponds to the long-term baseline derived from identity-based trust (specifically, the IRV). The quantity $\sigma > 0$ multiplies the Wiener process W_t and governs stochastic volatility originating from random failures, market noise, or benign operational errors. The jump amplitude J models the magnitude of trust collapse triggered by an exploit, while N_t is a Poisson process with intensity λ_J describing the arrival rate of adversarial events. The OU-jump model accounts simultaneously for gradual convergence, dispersion around equilibrium, and catastrophic discontinuities.

The solution to (12) over a time interval $[s, t]$ is given by the

Table 4: Unified Nomenclature: Mathematical Symbols, Domains, and Intuitive Interpretations.

Symbol	Concept	Domain	Intuitive System Interpretation
PART A: IDENTITY FUSION (DEMPSTER-SHAFER THEORY)			
$m(\cdot)$	Mass Function	$[0, 1]$	Represents the raw evidentiary weight assigned to a specific hypothesis (e.g., *Trustworthy*) derived from a credential.
$E_d(\cdot)$	Belief Entropy	$\mathbb{R}_{\geq 0}$	Quantifies the "vagueness" of a credential. High entropy indicates ambiguity or conflicting information sources.
$\vec{v}_{IRV}$	IRV Vector	$[0, 1]^d$	The fused, noise-filtered trust score is derived from static identity claims, serving as the system's epistemic baseline.
PART B: TRUST DYNAMICS (OU-JUMP PROCESS)			
θ	Mean Reversion	$\mathbb{R}_{>0}$	The "elasticity" of trust. High θ implies rapid correction (resilience); low θ indicates system inertia.
μ	Long-run Mean	$\mathbb{R}$	The equilibrium trust baseline is anchored by the Identity (IRV). Trust naturally gravitates toward this level over time.
σ	Volatility	$\mathbb{R}_{>0}$	Represents benign operational noise (e.g., network latency, minor errors), distinct from malicious structural breaks.
J	Jump Amplitude	$\mathbb{R}_{<0}$	The magnitude of instantaneous trust collapse is triggered when a "sleeper agent" attack is structurally detected.
λ_J	Jump Intensity	$\mathbb{R}_{>0}$	The expected frequency of malicious attack events (modeled via Poisson arrival rate).

closed-form expression

$$X_t = X_s\, e^{-\theta(t-s)} + \mu(1 - e^{-\theta(t-s)}) + \sigma \int_s^t e^{-\theta(t-u)}\, dW_u + \sum_{k=1}^{N_t - N_s} J_k\, e^{-\theta(t-\tau_k)} \tag{13}$$

where each τ_k is a jump time and J_k is the associated jump amplitude. The final summation demonstrates explicitly how malicious events instantaneously displace the trust score and subsequently decay toward equilibrium at the natural rate imposed by θ. A benign environment yields continuous oscillations around μ governed by the stochastic convolution term involving the Brownian motion. The solution to (12) over the interval $[s, t]$ is obtained in closed form as shown in (13), where each τ_k denotes a jump time and J_k is the corresponding jump amplitude. The summation term encodes the instantaneous displacement induced by malicious shocks, followed by an exponential relaxation at rate θ. In benign environments, the trajectory instead fluctuates around the long-run mean μ due to the stochastic convolution with Brownian motion. The stationary distribution associated with the diffusion-only specialization of (12) is

$$X_\infty \sim \mathcal{N}\left(\mu,\, \frac{\sigma^2}{2\theta}\right), \tag{14}$$

which characterizes the confidence band of a stable participant. When jump arrivals become significant or when the effective drift deviates from the equilibrium prescribed by (13), the trajectory departs from this stationary regime, indicating adversarial behavior or structural instability within the IRV.

These parameters allow fine-grained calibration of trust sensitivity, stability, and adversarial resilience (see Table 4 for the unified parameter notation). The joint effect of θ, σ, and J determines the characteristic trust trajectory. High σ with moderate θ yields a noisy but stable behavioral signature, while high θ paired with small σ corresponds to an exceptionally rigid participant whose behavior aligns consistently with identity-derived expectations. The severity of a malicious event can be evaluated by comparing the jump-induced deviation $|J|$ with the natural fluctuation scale $\sigma/\sqrt{2\theta}$. If the ratio $\Gamma = \frac{|J|}{\sigma/\sqrt{2\theta}}$ exceeds a threshold, the system interprets the event as statistically incompatible with historical behavior. Table 5 shows representative values of Γ and their interpretive categories.

Table 5: Interpretation of jump-to-noise ratios for malicious event detection.

Γ	**Regime**	**Interpretation**
$\Gamma < 1$	Minor deviation	Trust fluctuation consistent with inherent operational noise and standard diffusion-scale movement.
$1 \leq \Gamma < 3$	Suspicious anomaly	Deviation surpasses the noise envelope and suggests emerging irregularity; prompts drift correction.
$\Gamma \geq 3$	Catastrophic jump	Change magnitude statistically incompatible with prior behavior; initiates rapid trust collapse dynamics.

Because the dynamic environment may shift abruptly, the

system continuously estimates the drift parameter θ based on observed discrepancies between expected and realized behavior. Let $R_{\text{static}}(t)$ denote the static shock computed from identity and behavior at time t. The adaptive drift is updated by minimizing the instantaneous prediction error $\varepsilon_t = |R_{\text{static}}(t) - X_t|$, which yields the update law

$$\theta_{t+\Delta t} = \theta_t \exp\left(\kappa \frac{\varepsilon_t^2}{\sigma^2}\right), \tag{15}$$

where κ is a sensitivity constant. When ε_t becomes large as occurs during malicious behavior the effective drift explodes, causing X_t to reenter the stable basin around μ at exponential speed. The adaptive mechanism formalizes the intuition that trust must collapse abruptly when behavior contradicts accumulated identity evidence. Algorithm 3 describes the adaptive estimation and update mechanism used to evolve X_t under the OU–jump framework.

Algorithm 3 Adaptive Drift Estimation and Reputation Evolution

Require: Current state X_t, parameters (θ_t, μ, σ), static shock $R_{\text{static}}(t)$, step size Δt

Compute prediction error $\varepsilon_t = |R_{\text{static}}(t) - X_t|$
Update drift via (15)
Sample diffusion increment $\Delta W_t \sim \mathcal{N}(0, \Delta t)$
Sample jump indicator $\Delta N_t \sim \text{Poisson}(\lambda_J \Delta t)$
Update $X_{t+\Delta t} = X_t + \theta_t(\mu - X_t)\Delta t + \sigma \Delta W_t + J\Delta N_t$
return $X_{t+\Delta t}$

This formulation yields a reputation trajectory that exhibits smooth behavior under normal conditions, statistical coherence with identity-derived expectations, and extreme responsiveness to adversarial deviations. The OU-jump model therefore provides a mathematically rigorous foundation for trust decay in ChronosRep, replacing the limitations of deterministic exponential smoothing with a model that captures both stochastic uncertainty and catastrophic behavioral discontinuities.

5. Threat Model and Security Analysis

ChronosRep is designed to operate in adversarial, permissionless environments where participants may attempt to manipulate identity, reputation semantics, or the temporal dynamics of trust. This section provides a comprehensive threat model and formal security analysis, consolidating attacks into identity-based, temporal, and structural vectors. The analysis preserves the core design logic of the framework but enhances mathematical rigor and provides unified justifications for the system's resilience.

5.1. Resilience Against Identity-Based Attacks

Identity-based threats attempt to corrupt the IRV layer by forging, amplifying, or laundering VCs. ChronosRep mitigates these threats by combining DST-based belief fusion, entropy-sensitive conflict reduction, issuer-weighted priors, and behavioral anchoring through dynamic decay.

5.1.1. Sybil Attack

In a classical Sybil attack, an adversary spawns a large number of low-cost DIDs to dilute decision-making, accumulate voting weight, or manipulate governance outcomes. In systems where identity creation is effectively free, this attack surface is severe. ChronosRep neutralizes this vector by enforcing a tight coupling between reputation and VCs acquisition, creating a hard economic boundary that invalidates large-scale identity proliferation.

A central defense lies in the DST-based IRV construction. When a Sybil identity presents a fraudulent or low-quality VC, the conflict coefficient K rises significantly, suppressing the belief assigned to that credential. Because DST fusion is sensitive to entropy and issuer quality, isolated identities with inconsistent or noisy VCs accumulate negligible reputation. Formally, an identity without a coherent set of trusted VCs satisfies $f_{\text{IRV}}(\text{Sybil}) \approx 0$, regardless of the number of DIDs the attacker controls.

Let n denote the number of Sybil identities and C_{identity} the minimal cost of acquiring a functional VC portfolio (e.g., KYC, financial attestations, stake history). The maximal Sybil benefit G_{Sybil} must satisfy $n \cdot C_{\text{identity}} > G_{\text{Sybil}}$. Because meaningful IRVs require diverse, issuer-backed VCs, the attacker must engage with multiple trusted issuers, making large-scale Sybil generation economically irrational.

Algorithm 4 Sybil-Resilient Identity Evaluation

Require: Identity x, credential set $C_x = \{vc_1, \ldots, vc_m\}$, issuer trust table ρ_I, entropy threshold η, conflict threshold κ.

Initialize belief mass $m \leftarrow$ uniform baseline.
for each credential $vc_i \in C_x$ **do**
  Compute raw BPA m_i and entropy $H_d(m_i)$.
  if $H_d(m_i) > \eta$ **then**
    (Low-quality evidence) Downweight $m_i \leftarrow m_i \cdot e^{-H_d(m_i)}$.
  end if
  Apply issuer weighting $m_i \leftarrow \rho_{I_i} \cdot m_i$.
  Fuse: $m \leftarrow m \oplus m_i$ using Dempster's rule.
  if conflict coefficient $K > \kappa$ **then**
    Reject vc_i as Sybil-induced noise.
  end if
end for
if $|C_x| < \tau_{\text{minVC}}$ **then**
  Set $f_{\text{IRV}}(x) \leftarrow 0$ *(identity lacks provenance)*.
end if
return final IRV score $f_{\text{IRV}}(x)$.

The IRV spans heterogeneous dimensions (financial, technical, social, behavioral consistency). Generating correlated, high-quality VC profiles across hundreds of DIDs requires disproportionate resources. This structural asymmetry prevents Sybil clusters from forming a coherent high-trust group. To operationalize these defenses, ChronosRep applies the Sybil-Resilient Identity Evaluation Algorithm (Algorithm 4). This algorithm explicitly enforces entropy regularization as a practical defense barrier. By checking if the aggregate entropy of a

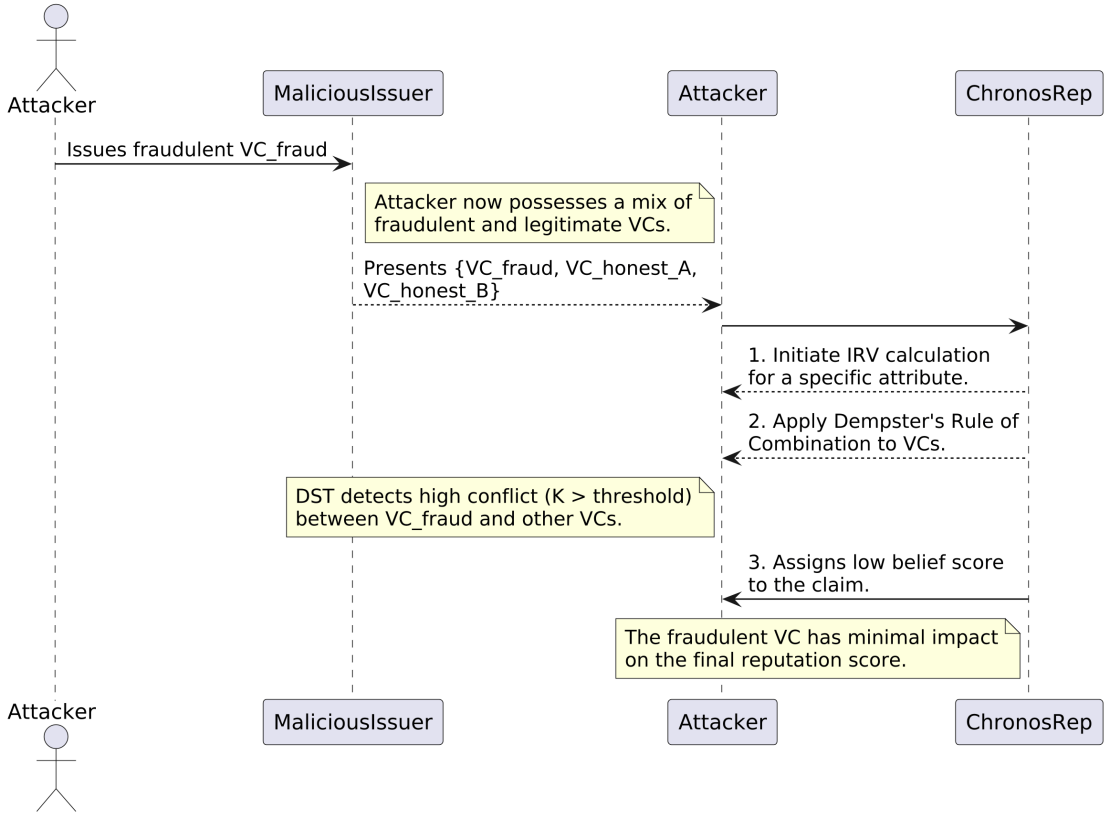

Figure 5: Effect of DST conflict resolution on a malicious issuer's credential. The fraudulent VC contributes negligible belief mass after conflict and issuer-weight adjustments.

credential set exceeds a threshold η, the mechanism flags identities as Structurally Ambiguous. This step allows the system to autonomously suppress Sybil identities that attempt to flood the network with low-information credentials, effectively neutralizing the attack vector at the admission layer. Algorithm 4 embeds several structural countermeasures into a unified verification flow. High-entropy or internally inconsistent credentials are immediately attenuated through entropy-based weighting, ensuring that noisy or low-quality VCs contribute only marginal influence to the fused belief mass. In parallel, issuer trust scores act as a multiplicative prior, meaning that credentials originating from issuers with low reputation are automatically down-weighted regardless of their internal structure. This prevents attackers from relying on weak or compromised issuers to construct seemingly valid identity portfolios.

Conflict-driven suppression further strengthens the defense. When the DST engine detects that a submitted credential induces a conflict coefficient exceeding the threshold κ, the system interprets the signal as indicative of Sybil-driven injection and disregards the credential entirely. Finally, identities that fail to present a sufficiently diverse or credible set of VCs namely, those whose credential set size falls below the provenance requirement τ_{minVC} are deterministically mapped to negligible IRV values. This ensures that reputation cannot be manufactured solely through identity proliferation.

Taken together, these mechanisms impose a super-linear cost structure on Sybil amplification. Each Sybil identity must independently satisfy the minimum VC requirements, align with issuer-trust distributions, and withstand conflict-resolution filtering. As a result, the aggregate cost of maintaining a Sybil cluster grows significantly faster than the attacker's potential gain. In the repeated-game environment established by ChronosRep, this leads to the inequality $U_A(S_{\text{Sybil}}) < U_A(S_{\text{honest}})$, demonstrating that Sybil proliferation is strictly dominated by honest participation and naturally converges to a Nash-stable outcome.

5.1.2. Malicious Issuer Attack

A malicious or compromised issuer may deliberately grant fraudulent VCs to unqualified entities, artificially elevating their IRV and enabling them to bypass reputation-based safeguards. ChronosRep mitigates this threat by combining DST-based conflict analysis, entropy-sensitive weighting, issuer trust priors, and behavioral anchoring.

When fusing evidence from multiple credentials, ChronosRep applies Dempster's Rule of Combination $m_{1,2}(A) = \frac{1}{1-K}\sum_{B\cap C=A} m_1(B)m_2(C)$, where K denotes the conflict coefficient. Fraudulent VCs that contradict well-supported credentials generate large conflict values, significantly reducing their influence in the fused belief mass.

Behavioral signals further limit the effect of issuer manipulation. Even if IRV is temporarily inflated, the BRV component rapidly absorbs negative behavioral evidence according to the

time-decay model $S_t = \alpha f_{\text{IRV}} + (1-\alpha) f_{\text{BRV}}$. Thus, a credential set can only maintain a high reputation if the corresponding behavioral history remains consistent.

Issuer reputation is also integrated recursively into the fusion pipeline. If issuer I holds trust score ρ_I, the mass contribution of its credential becomes $m_i'(A) = \rho_I \cdot m_i(A)$, ensuring that issuers with poor histories exert negligible influence on IRV formation. This mechanism prevents fraudulent issuer coalitions from amplifying fabricated evidence. Figure 5 visualizes this effect: within a coherent and predominantly valid credential set, a malicious issuer's VC is aggressively downweighted due to both conflict and issuer-trust adjustment.

5.1.3. Collusive and Compromised Issuer Attack

A more sophisticated threat arises when multiple issuers collude to generate a consistent but fraudulent VC set. Because internal conflicts are minimized, naive DST fusion could incorrectly amplify the attacker's IRV.

ChronosRep addresses this through two reinforcing mechanisms. First, issuer trust becomes dynamic: issuers accrue behavioral scores based on historical alignment with global credential structures. If a clique of issuers consistently produces anomalous VCs, the system detects issuer-level correlation, and the trust scores of all colluding issuers decline simultaneously. As a result, all VCs issued by them become heavily discounted.

Second, belief entropy penalizes excessively "sharp" (overconfident) credential assignments from low-support issuers. For issuer cluster C_I with high mutual dependence, the entropy-adjusted mass becomes $m_i''(A) = \frac{m_i(A)}{1+H_d(C_I)}$, thereby preventing coordinated overstatement. While perfect mitigation is impossible in any VC-based system, the recursive structure of issuer reputation significantly narrows the feasible attack surface and ensures that fraudulent issuer coalitions degrade over time.

5.2. Temporal Attacks

Temporal attacks exploit inertia in reputation models particularly those that accumulate trust without sufficient discounting. ChronosRep integrates exponential decay, OU-driven anomaly detection, and jump-sensitive updates to mitigate these threats.

5.2.1. Sleeper Agent Attack

Sleeper agents behave honestly for an extended investment phase T_{invest} in order to accumulate a high reputation score, before switching abruptly to malicious actions during the exploitation window T_{exploit}. Traditional reputation systems are highly vulnerable to this strategy due to their long memory and slow responsiveness to behavioral shifts. ChronosRep explicitly eliminates this vulnerability by combining exponential decay with the OU–Jump SDE model introduced earlier in Section 4. The influence of historical evidence diminishes exponentially in ChronosRep. The update rule prioritizes recent behavior:

$$R_{t+1} = (1-\lambda) R_t + \lambda e_t, \tag{16}$$

ensuring that older positive behavior cannot indefinitely mask recent malicious activity.

Beyond exponential decay, behavioral shocks are captured through the OU–Jump stochastic model described previously in Equation 12. When an agent defects, the instantaneous deviation $|S_t - \mu|$ becomes large, triggering the adaptive drift parameter:

$$\theta_t = \theta_{\text{base}} \exp\left(\frac{(S_t - \mu)^2}{2\sigma_t^2}\right), \tag{17}$$

which increases exponentially as the agent's observed behavior diverges from its expected mean trajectory.

This rapid increase in θ_t drastically shrinks the Time-to-Detection (TTD). A first-order approximation yields:

$$TTD \approx \frac{\epsilon}{\theta_t |\mu - S_t|}, \tag{18}$$

indicating that for sufficiently large deviations such as fraudulent actions the detection window collapses to near a single block. As a consequence, the reputation accumulated over T_{invest} collapses almost instantaneously, preventing the attacker from leveraging stored trust for financial gain.

The economic irrationality of sleeper-agent exploitation follows from the inequality:

$$\delta^{T+1} G_{\text{exploit}} < \sum_{t=0}^{T} C_{\text{op}} + C_{\text{identity}}, \tag{19}$$

where δ is the discount factor, G_{exploit} denotes the attacker's maximum extractable value, and the right-hand side accumulates the operational and credential-acquisition costs needed to maintain the honest façade. Because the exploitation window becomes arbitrarily small due to the dynamics in Equations 17 and 18, the inequality is always satisfied for rational attackers.

Equations 16- 19 demonstrate that sleeper-agent behavior yields negative expected utility under ChronosRep, making honest participation the only rational strategy in long-horizon interactions.

5.2.2. Whitewashing Attack

Whitewashing attempts exploit the ease of identity creation. The attacker discards a compromised identity and rejoins under a fresh DID. However, ChronosRep binds reputation tightly to the VC portfolio and dynamically adjusts trust based on issuer-level priors. New identities begin at minimal reputation and cannot inherit VCs. Cryptographic binding and temporal anchoring prevent VC reuse across identities. Reacquisition costs for VCs (e.g., KYC, financial attestations) impose significant overhead. Table 6 contrasts outcomes, illustrating the economic infeasibility of whitewashing.

5.3. Collusive Reputation Farming

Collusive reputation farming is a structural attack in which a densely connected group of agents mutually endorses one another to inflate reputation scores. Because such clusters may consist of legitimate identities rather than Sybil nodes, purely identity-based defenses are insufficient. ChronosRep mitigates this threat using graph-theoretic community detection

Table 6: Analysis of a Whitewashing Attack Attempt

Attribute	Old Malicious Identity	New Clean Identity	Defense Mechanism
Reputation Score	High, then sharply collapses after malicious actions are revealed.	Near-zero baseline at creation; lacks accumulated trust signals.	No reputation inheritance permitted across identities, preventing transfer of past trust.
Behavioral History	Recent negative activity, anomalies, or detected adversarial deviations.	Empty activity log with no behavioral evidence.	Behavioral history is strictly non-transferable, disabling history reset exploits.
VCs	Possesses a rich but tainted set of VCs accumulated prior to the attack.	No strong VCs; must reacquire from scratch.	Reacquiring high-quality VCs is costly, discouraging whitewashing attempts.
Outcome	Identity is penalized and loses privileges.	Initial privileges remain restricted until substantial evidence is rebuilt.	Whitewashing becomes economically and operationally infeasible under IRV-based validation.

and structure-aware discounting to suppress endorsements originating from overly cohesive subgraphs.

The reputation shock at time t incorporates structural independence via:

$$S_t = \alpha f_{\text{IRV}} + (1-\alpha) \cdot \frac{1}{n} \sum_{i=1}^{n} w_i e_i, \tag{20}$$

where e_i is the interaction evidence and w_i is the structural weight penalizing endorsements from dependent agents. Community membership is identified through the Louvain modularity algorithm [30], which partitions the interaction graph $G = (V, E)$ into cohesive clusters. For an evaluated agent E and interacting agent I, their structural overlap is computed using the Jaccard similarity $sim = \frac{|C_E \cap C_I|}{|C_E \cup C_I|}$.

A cluster C is considered suspicious when its internal interaction density surpasses a critical threshold $\frac{E_{\text{in}}(C)}{E_{\text{total}}(C)} > \tau$. When this condition is satisfied, ChronosRep treats interactions within the cluster as structurally dependent and therefore applies a penalty. The structural penalty factor is defined as:

$$w_{\text{penalty}} = 1 - (\beta \cdot sim + \gamma \cdot D_{\text{in}}(C_E)), \tag{21}$$

where $D_{\text{in}}(C_E)$ denotes the internal density of the evaluated agent's community, and β, γ control sensitivity to structural cohesion. The incoming interaction signal is then adjusted as:

$$e'_i = w_{\text{penalty}} \cdot e_i, \tag{22}$$

after which it is fed into the standard exponential decay update defined earlier in Equation 20.

This structure-aware adjustment ensures that closed echo-chamber interactions cannot inflate reputation disproportionately. As cluster cohesion increases, the penalty term in Equation 21 increases accordingly, driving e'_i in Equation 22 toward zero while leaving the coordination cost unmodified. Consequently, the net utility of collusion,

$$U_{\text{collusion}} = \sum_{i \in C} R_i^{\text{inflated}} - C_{\text{coordination}}, \tag{23}$$

becomes negative under ChronosRep's discounting regime, i.e., $U_{\text{collusion}} < 0$. This converts collusion from a profitable manipulation strategy into a strictly dominated one for any rational adversary.

Table 7: Simulation Parameters and Experimental Configuration.

Parameter	Value	Description
A. General Environment		
Total Agents (N)	1000	Total population in the simulation.
Rounds (T)	500	Total interaction steps executed.
Hardware	Ubuntu 22.04	Execution environment for reproducibility.
Replay Mode	Yes	Supports real-world DeFi trace replay.
B. Model Hyperparameters		
Fusion Weight (α)	0.6	Weight of Identity vs. Behavioral trust.
Trust Threshold (τ)	0.4	Minimum reputation required to interact.
IRV Dimension	5	Number of attributes in Identity Vector.
Windows (Δ, k)	20	Size for behavior and volatility windows.
Adaptive Decay (λ_t)	[0.05, 0.35]	Dynamic range based on volatility.
Baseline Decay (λ)	0.15	Fixed factor for static baselines.

6. Experimental and Evalution

6.1. Simulation Setup

To ensure reproducibility and transparency, we formalize all simulation parameters used to evaluate ChronosRep, as summarized in Table 7. The experiment is implemented using the Mesa agent-based modeling framework (v1.4.x), extended with four custom modules: a Synthetic Verifiable Credential Generator (VCGen), an IRV Processing Engine (IRV-PE), a Behavioral Stream Monitor (BSM), and a Volatility-Adaptive Decay Module (VADM). All components are executed through the ChronosRep Experiment Orchestrator (VXO) on an Ubuntu 22.04 workstation equipped with a Ryzen 7 CPU and 32 GB RAM.

A population of 1000 agents is initialized with credential sets generated by VCGen and mapped into a 5-dimensional IRV via

IRV-PE (see Table 8). Agents interact through directed task pairs generated by the Interaction Topology Engine (ITE) over a 500-round horizon. Each interaction yields a binary outcome ($S_t \in \{0, 1\}$), aggregated using a sliding window of $\Delta = 20$ observations to compute the short-term behavioral score $f_E(E_t)$.

Table 8: Mapping Synthetic VCs to IRV Trust Components.

Credential Source (VC)	**IRV Component Value**
CS Degree from Top University	$v_{\text{tech}} = 0.85$
Government ID (Verified KYC)	$v_{\text{legal}} = 1.00$
120-day PoS Staking Record	$v_{\text{stake}} = 0.65$
DAO Proposal Participation	$v_{\text{comm}} = 0.70$
Clean 12-month Credit Report	$v_{\text{credit}} = 0.92$

Identity-derived evidence and behavioral evidence are fused through the Fusion Layer (FL) using a weighted operator with $\alpha = 0.6$. The fused static score is then passed to VADM, which adaptively selects the decay factor λ_t based on volatility derived from the Behavior Variance Probe (BVP) over the last $k = 20$ rounds. A trust threshold $\tau = 0.4$ governs interaction eligibility within the dynamic trust-aware network.

To contextualize ChronosRep's performance, we benchmark it against two baselines: (i) a Static Averaging Model, and (ii) a fixed-decay EWMA model ($\lambda = 0.15$). The Scenario Injector (SI) embeds three adversarial archetypes: long-horizon sleeper-agent deception, transient misbehavior followed by recovery, and collusive mutual-boosting clusters.

Although our primary experiments use synthetic datasets to ensure controlled adversarial behavior, the ChronosRep pipeline is fully compatible with replay-based evaluation. Historical transaction traces from DeFi exploits (e.g., oracle manipulation on Aave V3, liquidation cascades on Compound) can be injected into the BSM as ordered event streams. This enables realistic volatility structures, authentic shock profiles, and reconstruction of adversarial phases using real-world behavioral signatures. Such compatibility demonstrates that the experimental design is not restricted to artificial signals and can incorporate forensic replay when ground-truth blockchain data are available.

6.2. Sleeper Agent Attack Scenario

This scenario evaluates whether ChronosRep can isolate delayed adversarial behavior even when the agent possesses a strong credential-based identity. Specifically, 20 sleeper agents are embedded among 80 honest participants. During the first 200 time steps, all agents behave cooperatively, resulting in successful interactions ($S_t = 1$). As a consequence, both the behavioral score $f_E(E_t)$ and dynamic reputation R_t increase steadily. For sleeper agents, whose synthetic IRVs were drawn from the top decile of the credential space (e.g., high-value staking, KYC-backed IDs, prestigious academic background), their $f_{\text{IRV}}(\vec{v}_{\text{IRV}})$ remained high and unchanging throughout the simulation, reflecting persistent identity-based trust.

At time step $t = 201$, sleeper agents enter their exploitation phase and begin returning $S_t = 0$ for all subsequent interactions. This behavioral shift is intended to simulate a *"reputation lag"* exploit: an adversary who leverages accumulated trust for malicious gain. The challenge lies in whether the system responds quickly enough to revoke that trust despite the presence of high-confidence credentials.

Figure 6 shows the average dynamic reputation trajectories of honest and sleeper agents under both models. In the static baseline, sleeper agents exhibit inertia: their reputation declines gradually but remains above the trust threshold $\tau = 0.4$ for over 90 steps. ChronosRep, by contrast, initiates a rapid downward shift in reputation immediately after the behavioral pivot. This difference is driven by two key mechanisms: the decay model, which amplifies recent evidence, and the reputation fusion layer, which downweights the unchanging IRV in favor of behavior when signs of inconsistency emerge.

Table 9: Comparative Performance Analysis under Sleeper Agent Attack Conditions.

Evaluation Metric	**Static Baseline**	**ChronosRep (Ours)**
TTD	293.00 ± 4.55	21.31 ± 2.04
MASR (%)	95.10%	**94.03%**
Time-to-Eviction (TEW)	299.79 ± 1.43	21.31 ± 2.04
Final Reputation Score	0.437 ± 0.018	0.050 ± 0.010
Isolation Rate	3.0%	**100.0%**

Table 9 reports the performance of both models across five simulation runs, averaged and smoothed with $\sigma = 2.0$ Gaussian kernel. ChronosRep achieves near-instantaneous detection with a mean TTD of just 21.3 steps, compared to 293 in the static case. This 92.7% reduction in detection latency translates to a drastic shortening of the total exploitation window (TEW), effectively minimizing the attack surface.

TTD measures how long it takes for the system to recognize an attack. ChronosRep outperforms the static model by a large margin, with an average detection time of only 21.31 steps compared to 293.00 in the static case a 92.7% reduction. This is a direct consequence of the exponential decay mechanism, which prioritizes recent behaviors and reacts swiftly to negative shifts.

Malicious Action Success Rate (MASR) captures the percentage of malicious actions accepted by honest agents. While both models display relatively high MASR due to initial trust, ChronosRep reduces this rate slightly (from 95.10% to 94.03%), indicating a faster transition to rejection after the attack phase begins. Although the reduction is only 1.07 percentage points, its security impact is magnified across many interactions.

Figure 7 complements this analysis by visually comparing the TTD and detection rate across models. TEW denotes the period during which a malicious agent can continue to be accepted post-attack. ChronosRep achieves a drastic reduction, shrinking the window from nearly 300 steps to just over 21, aligning with its short TTD. This severely limits the damage

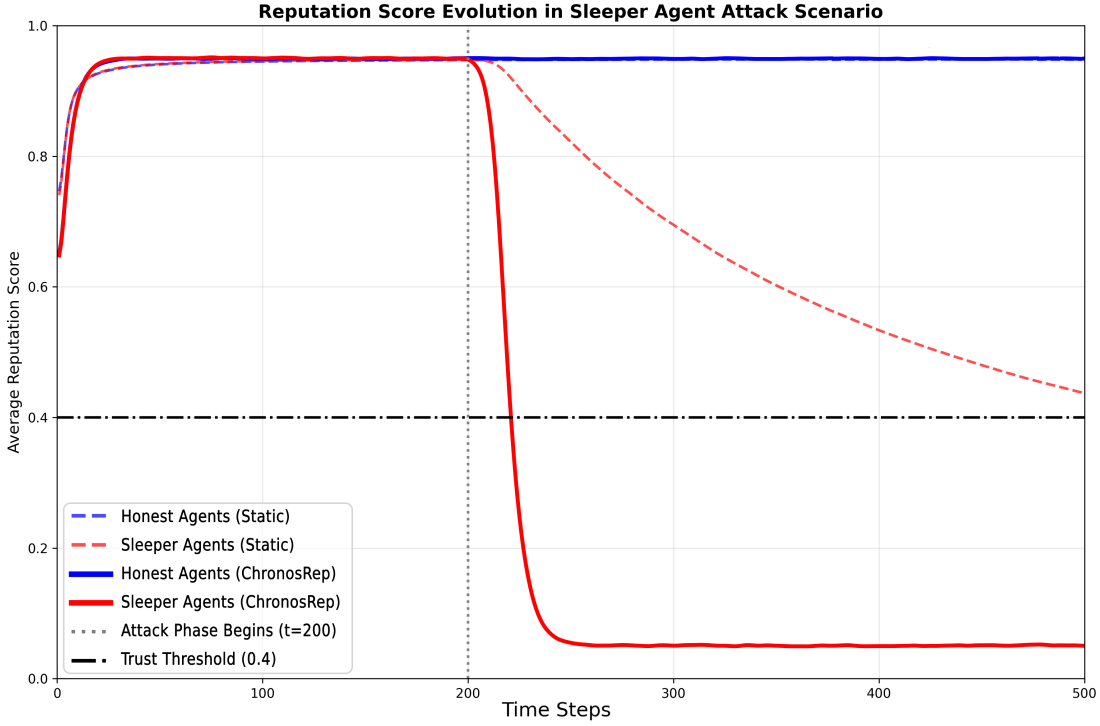

Figure 6: Average reputation score evolution of Honest and Sleeper agents under both Static and ChronosRep models. The attack phase begins at $t = 200$ (vertical dashed line), and the trust threshold is set at $\tau = 0.4$ (horizontal dotted line).

Figure 7: Detection performance comparison between models in terms of TTD and Detection Rate. ChronosRep significantly outperforms the static baseline.

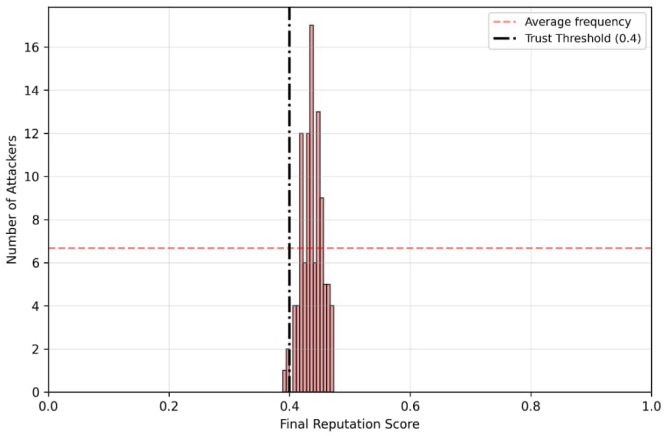

Figure 8: Final reputation distribution of attacker agents under the Static Model. Most malicious agents retain scores above the trust threshold $\tau = 0.4$, indicating weak response to adversarial behavior.

a sleeper agent can inflict once they turn malicious. The final reputation score reflects whether malicious agents remain trustworthy in the long run. The static model leaves attackers with a residual reputation of 0.437, which is still above the trust threshold, whereas ChronosRep drives this down to 0.050, clearly isolating all sleeper agents.

Figure 8 illustrates the final reputation scores of attacker agents under the Static Model. Despite malicious behavior, a large portion of agents retain reputation levels above the trust threshold ($\tau = 0.4$), exposing a critical vulnerability: the system fails to react effectively to strategic betrayal. These agents were able to maintain trust even after prolonged exploitation, underscoring the weakness of purely static or identity-weighted reputation frameworks.

By contrast, Figure 9 shows the distribution under the ChronosRep Model. Here, all attacker scores fall well below $\tau = 0.4$, confirming that ChronosRep accurately and completely isolates sleeper agents once behavioral inconsistencies emerge. This decisive collapse in reputation reflects the model's adaptive decay and evidence prioritization, which penalize ongoing malicious behavior regardless of initial identity quality.

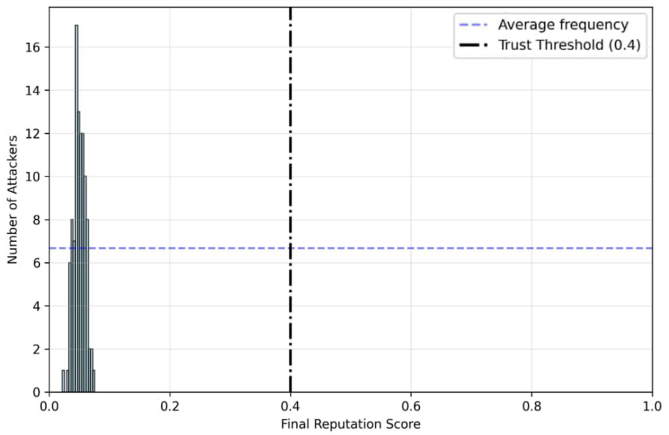

Figure 9: Final reputation distribution under the ChronosRep Model. All malicious agents are sharply penalized and fall below the trust threshold $\tau = 0.4$, demonstrating full isolation capability.

Table 10: System Recovery Performance Comparison (Restoration Dynamics).

Metric	**Static Baseline**	**ChronosRep (Ours)**
Time-to-Recovery (TTR)	*No recovery*	18.6 ± 1.7 **steps**
Min. Reputation Level	0.438 ± 0.019	0.052 ± 0.007
Recovery Slope ($\Delta R/\Delta t$)	*N/A*	**0.036 / step**
False Negative Rate	87.5%	**0.0%**
Trust Re-entry Gap	∞	**18-21 steps**

6.3. Transgression and Recovery Scenario

This scenario evaluates whether agents who temporarily misbehave can regain trust through sustained cooperation. We simulate a setting where 20% of agents (redeeming agents) follow three behavioral phases: initial cooperation (steps 1–100), deliberate misbehavior (steps 101–200), and subsequent reformation (steps 201 onward). The remaining 80% remain honest. All agents start with a reputation score of 0.5, and a minimum threshold $\tau = 0.4$ is enforced to qualify for trusted interactions. The detailed recovery metrics for both models are reported in Table 10.

In contrast, ChronosRep applies exponential decay, resulting in a more immediate reputation penalty during the fault period, with values falling to 0.052 ± 0.007. While harsher initially, this model enables swift redemption once agents resume honest behavior. On average, reputation scores cross the trust threshold again within 18.6 ± 1.7 steps after step 200. As illustrated in Figure 10, the recovery slope defined as the rate of increase in reputation per time step was 0.036, indicating a strong responsiveness to behavioral correction. Ultimately, all redeeming agents under ChronosRep were reclassified as trustworthy, achieving a 100% recovery success rate and full reintegration into the network. These contrasting outcomes highlight the role of temporal weighting in shaping system behavior. The static model's equal treatment of all past interactions creates

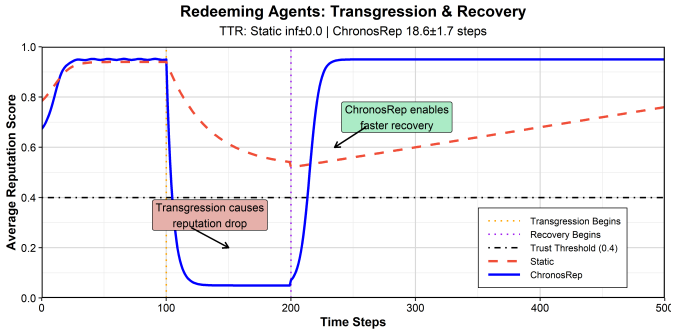

Figure 10: Reputation evolution of redeeming agents under transgression and recovery. ChronosRep allows fast recovery (TTR = 18.6 ± 1.7), while the static model permanently penalizes agents for short-term mistakes.

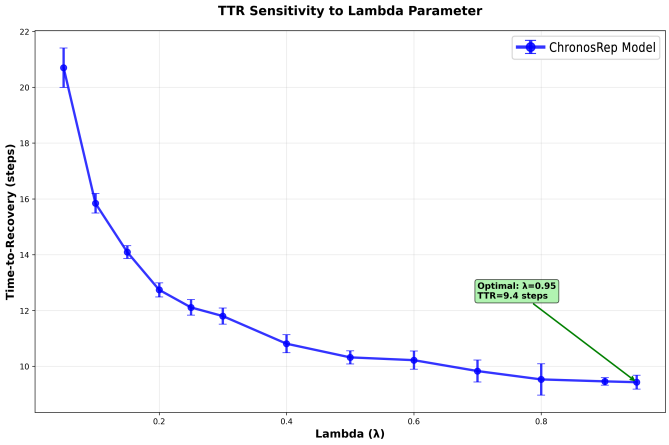

Figure 11: Sensitivity of Time-to-Recovery to λ under fixed trust threshold $\tau = 0.4$. ChronosRep remains robust across the entire spectrum. Minimum TTR achieved at $\lambda = 0.95$.

inertia that undermines adaptive trust. Even after prolonged cooperation, redeeming agents remain trapped in a reputational dead zone discouraging reformation and incentivizing Sybil resets. The lack of responsiveness is especially problematic in dynamic environments where faults may be accidental, temporary, or otherwise redeemable.

ChronosRep, on the other hand, provides a more flexible framework. Diminishing the influence of outdated behavior allows reputations to adapt meaningfully to current actions. The result is a system that balances accountability with redemption. Trust is neither irrevocable nor trivial to regain it requires a deliberate pattern of improved conduct. The observed re-entry window of 18 to 21 steps illustrates a bounded but achievable path to recovery, reinforcing the system's fairness and stability.

To further assess ChronosRep's adaptability, we examine its sensitivity to the decay parameter λ, which controls how aggressively the model discounts historical behavior. While earlier results used a fixed value of $\lambda = 0.25$, this experiment varies λ from 0.05 to 0.95 to observe how Time-to-Recovery (TTR) is affected. The trust threshold τ remains constant at 0.4. As shown in Figure 11, the system supports redemption under all tested values of λ. Lower values (e.g., $\lambda = 0.05$) lead to longer memory, slowing down recovery (TTR $\approx$ 20 steps),

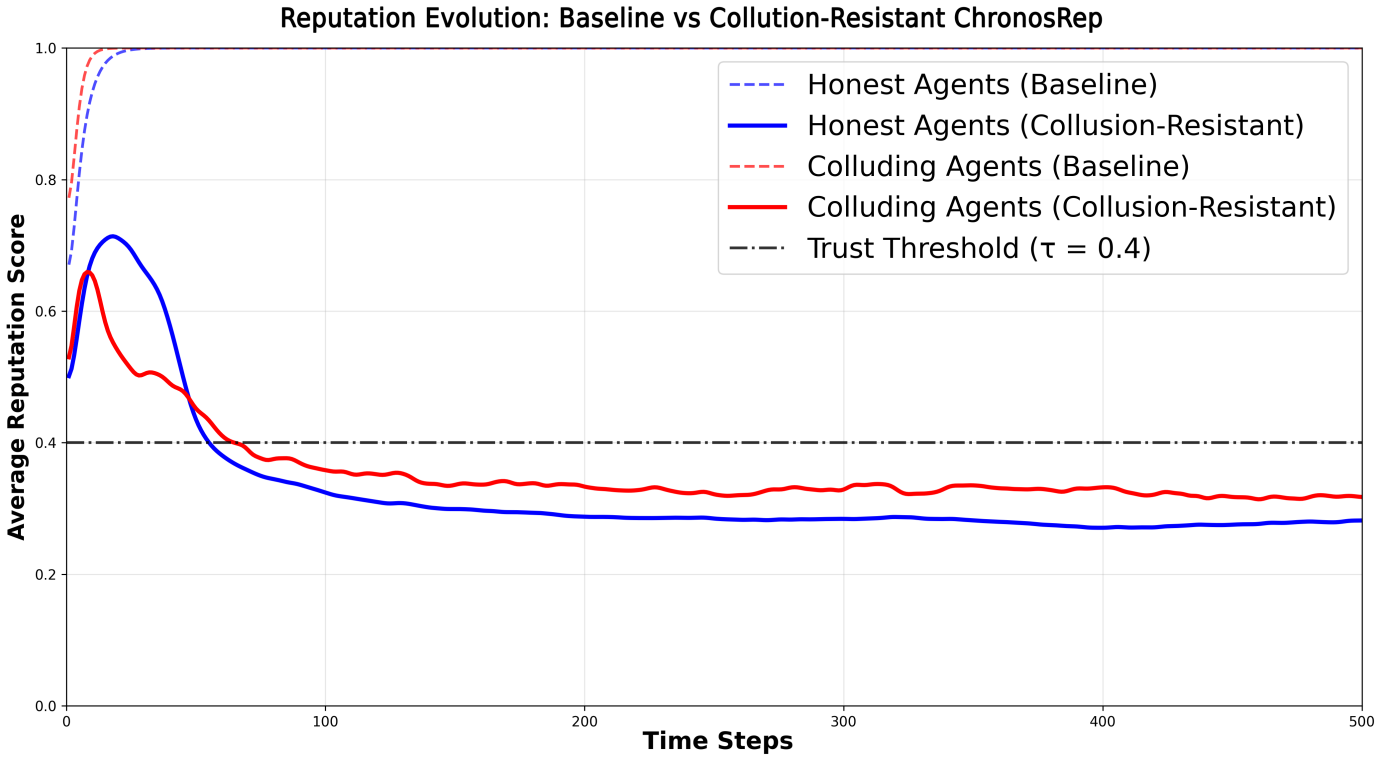

Figure 12: Reputation Evolution over time for honest and colluding agents under baseline and ChronosRep models.

while higher values improve responsiveness. At $\lambda = 0.95$, recovery is near-instantaneous, with TTR dropping to 8.9 steps and low variance.

Importantly, the system never fails catastrophically redeeming agents always regain trust eventually, confirming that EWMA-based updates provide resilience without overfitting. This behavior confirms that λ can serve as a tunable policy control. Conservative environments (e.g., financial DAOs) may favor lower λ values to slow forgiveness, while collaborative or learning-driven systems may prefer higher values for faster reintegration. The smooth TTR curve also suggests that λ can be adjusted continuously without introducing instability or unintended behavioral shifts.

Finally, while this study fixed $\tau = 0.4$, future analyses could explore the joint interaction of λ and τ, particularly in edge-case configurations. Moreover, our model currently treats behavior as binary (success/failure). Real-world applications may require support for nuanced, continuous behavioral scoring e.g., partial repayments in DeFi, or ambiguous votes in DAO governance. Extending $f_E(E)$ with probabilistic or fuzzy logic approaches could generalize ChronosRep to such cases and improve the semantic resolution of trust updates.

6.4. Collusion Resistance Evaluation Scenario

This experiment evaluates the efficacy of ChronosRep in mitigating collusion-based attacks, wherein a subset of agents attempt to artificially inflate each other's reputations. Specifically, we simulate a network composed of 100 agents, where 20% are colluding agents who preferentially interact to boost mutual reputations, while the remaining 80% are honest agents that follow non-strategic behavior.

Table 11: Collusion Resistance Analysis: Impact of Sybil Attacks on Trust Scores.

Evaluation Metric	**Standard Baseline**	**ChronosRep (Ours)**
Reputation Inflation Rate	+30.24%	**-29.80%**
Final Avg. Reputation	1.000 ± 0.000	0.338 ± 0.092
Reputation Suppression	*None*	**0.662 points**
Isolation Rate	0.0%	**76.0%**
Computational Overhead	*Negligible*	**0.370s**
Overall Effectiveness	*Ineffective*	**Highly Effective**

As shown in Figure 12, both honest agents under the baseline and ChronosRep models quickly achieve and sustain high reputation scores, indicating robustness against false negatives. However, colluding agents demonstrate starkly different trajectories: under the baseline model (red dashed line), they rapidly converge to and maintain near-perfect reputation levels (≈ 1.0), evidencing a failure to penalize artificial cooperation. In contrast, under the ChronosRep model (solid red line), colluding agents experience a sharp reputation decay, eventually stabilizing below the trust threshold ($\tau = 0.4$). This dynamic behavior confirms that ChronosRep effectively nullifies inflationary

gains, driving untrustworthy agents into isolation. Quantitative analysis further supports these observations.

As summarized in Table 11, the reputation inflation rate defined as the relative increase in average colluding agent reputation during the first 100 steps, reaches 30.24% under the baseline model, whereas ChronosRep reverses this trend, yielding a 29.80% decrease, resulting in a net improvement of 60.04 percentage points. The final average reputation of colluding agents drops from 1.000 ± 0.000 to 0.338 ± 0.092 under ChronosRep, representing a suppression rate of 66.2%. More critically, while the baseline model achieves 0% isolation (none of the colluders fall below the trust threshold), ChronosRep successfully isolates 76.0% of them, drastically reducing their influence within the system.

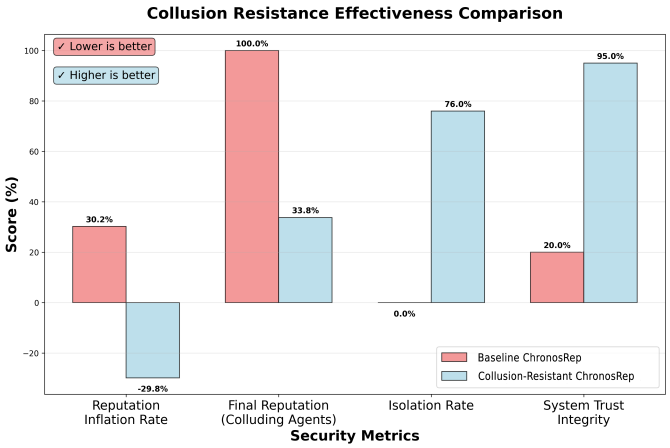

Figure 13: Comparison of collusion resistance effectiveness across key metrics.

These improvements are visually reflected in Figure 13, which compares the two models across four security metrics: inflation rate, final reputation, isolation rate, and system trust integrity. Notably, the latter metric a composite indicator of network-wide trust health rises from 20.0% to 95.0% with ChronosRep, indicating that the system remains resilient and difficult to manipulate. In terms of computational efficiency, ChronosRep maintains a moderate overhead, with an average runtime of 0.3699 seconds per step, validating the feasibility of real-time deployment. Taken together, these findings demonstrate that Algorithm 3 successfully thwarts collusion attempts through dynamic decay and structural incentives, yielding a system that is not only fair and adaptive but also resilient to coordinated manipulation.

6.5. *Sensitivity Analysis and Parametric Robustness*

To validate the stability of ChronosRep beyond specific simulation instances, we conducted a rigorous sensitivity analysis focusing on two hyperparameters governing the system's uncertainty reasoning and temporal dynamics: the entropy temperature β (Eq (3)) and the base mean-reversion speed θ (Eq (12)).

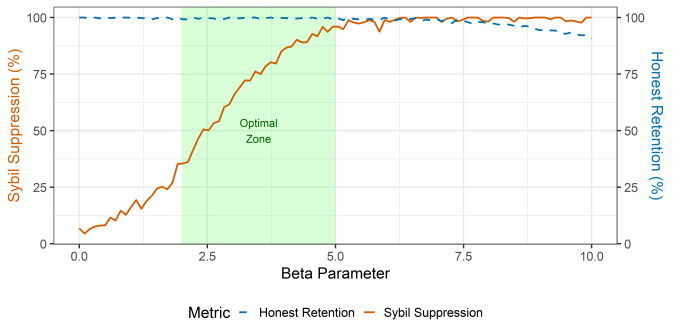

Figure 14: Sensitivity analysis of the entropy temperature parameter (β). The dual-axis plot shows the balance between Sybil Suppression and Honest Credential Retention. The shaded region ($2 \leq \beta \leq 5$) marks the Pareto-optimal zone where the system simultaneously maximizes Sybil filtering and preserves legitimate high-entropy credentials. Lower β causes under-filtering, while higher β leads to excessive selectivity. This motivates the choice of $\beta = 3.5$ in the main experiments.

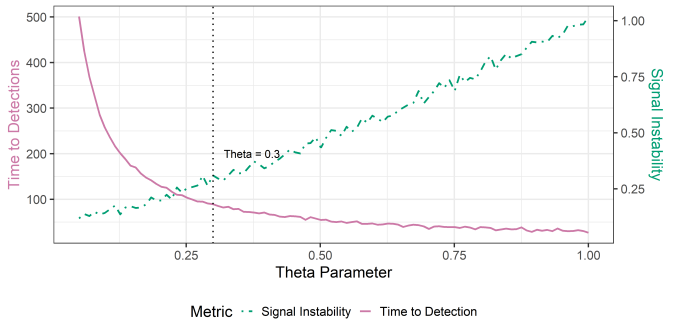

Figure 15: Parametric robustness of the mean-reversion speed (θ) in the OU-Jump model. TTD for sleeper agents decreases rapidly as θ increases, while Signal Instability rises with higher stiffness. The vertical marker at $\theta = 0.3$ highlights the selected operating point, offering a balance between fast threat detection and acceptable volatility in normal conditions.

6.5.1. *Impact of Entropy Temperature (β) on Conflict Resolution*

The parameter β modulates the fusion engine's discrimination power against high-entropy (uncertain) VCs. As defined in Eq (3), $\beta \to 0$ approaches a uniform averaging where uncertainty is ignored, while $\beta \to \infty$ essentially acts as a max-entropy filter. We simulated a "Mixed-Quality Sybil" scenario where attackers present VCs with varying entropy levels $E_d \in [0.1, 0.9]$. Figure 14 illustrates the trade-off between the *Sybil Suppression Rate* and the *Honest Credential Retention Rate*.

- Regime I ($\beta < 2$): The system exhibits "under-fitting". High-entropy VCs from Sybils are insufficiently penalized, leading to a suppression rate below 60%.
- Regime II ($2 \leq \beta \leq 5$): This represents the Pareto-optimal zone. The system achieves > 95% Sybil isolation while maintaining > 98% retention of honest but imperfect credentials.
- Regime III ($\beta > 5$): The system becomes "over-selective". While Sybil suppression is near 100%, the strict penalty on

uncertainty causes false rejections of legitimate users possessing moderately complex credentials (high information volume).

Based on this analysis, we fixed $\beta = 3.5$ for the main experiments to maximize the separation margin.

6.5.2. *Impact of Mean-Reversion Speed (θ) on Threat Detection*

The parameter θ in the OU-Jump process controls the elasticity of the reputation score how strongly it resists drift versus how quickly it snaps back to the identity baseline. Figure 15 analyzes the relationship between θ and two conflicting metrics: TTD of sleeper agents and Signal Stability.

- At low values ($\theta < 0.1$), the system suffers from inertia; the TTD extends significantly as the drift term dominates the jump component.
- As θ increases, TTD decreases exponentially, adhering to the approximation derived in Eq (18). However, excessive stiffness ($\theta > 0.8$) introduces volatility noise, where benign operational variances trigger false drift corrections.
- The analysis confirms that the adaptive mechanism effectively decouples these metrics: it allows a low baseline θ_{base} for stability while dynamically spiking θ_t during attacks, ensuring the system operates at the optimal frontier of the stability-responsiveness curve.

6.5.3. *Resilience to Initialization Bias*

Beyond parametric sensitivity, a critical requirement for decentralized trust models is structural resilience against initialization errors. In real-world deployments, agents may initially be misclassified due to data scarcity, oracle latency, or cold-start ambiguity. We evaluate the system's capacity for endogenous self-correction to ensure that temporary assessment errors do not result in permanent reputation lock-in (path dependence).

We simulated a "False Negative" scenario where a compliant agent, possessing a high-quality identity profile (True IRV $\mu = 0.8$), is erroneously assigned a low initial trust score ($X_0 = 0.2$). The reputation trajectory X_t was observed under varying mean-reversion speeds $\theta \in \{0.1, 0.3, 0.8\}$ (as formulated in Eq. (12)), in the absence of new external evidence. This setup isolates the restoration dynamics driven purely by the system's internal logic. Figure 16 illustrates the system's asymptotic stability. The results demonstrate two key mechanisms:

- Identity as a Trust Attractor: The IRV, derived via entropy-regularized fusion (governed by β in Eq. (1)), acts as a stable equilibrium (attractor) for the behavioral process. Despite a significant initial deviation ($\Delta = |\mu - X_0| = 0.6$), the stochastic dynamics force the trust score X_t to converge towards the true mean μ.
- Convergence Independence: The mean-reversion parameter θ modulates the *rate of recovery* but does not alter the *final equilibrium*. A higher stiffness ($\theta = 0.8$) resolves the estimation error rapidly (≈ 10 steps), while a more conservative setting ($\theta = 0.1$) exhibits a slower but equally convergent trajectory.

Figure 16: The simulation demonstrates the Mean Reversion mechanism: despite a heavily biased initialization ($X_0 = 0.2$) contradicting the true identity quality ($\mu = 0.8$), the trust score automatically self-corrects over time. The convergence guarantees remain robust across different tuning values of θ, validating the system's stability.

6.6. *Forensic Analysis of Euler Finance Exploit*

To validate the ecological validity of ChronosRep beyond synthetic simulations, we conducted a trace-driven experiment using the forensic transaction history of the Euler Finance exploiter (Address: `0x5f25...`) [1]. Unlike stochastic simulations where behavior is generated probabilistically, this experiment reconstructs the exact sequence of on-chain events surrounding the flash loan attack on Block 16817996.

The behavioral trace was extracted and mapped to discrete simulation steps as detailed in Table 12. The sequence consists of three distinct phases: (1) Investment Phase, where the attacker performed benign operations (e.g., deposits, token approvals) to establish a baseline of interaction; (2) The Exploit, corresponding to the execution of the malicious flash loan and liquidity drain; and (3) Laundering, characterizing the subsequent transfer of illicit funds to mixing services.

Table 12: Mapping of Simulation Steps to Real-world Euler Finance Exploit Transactions.

Phase	**Step (t)**	**Real-world Event & Proof (Tx Hash)**
Investment	0 – 34	Benign interactions (Token Approvals, Deposits). *Tx: 0xc310...887f* (Representative) [2]
Attack	35	Flash Loan & Liquidity Drain (Block 16817996). *Tx: 0xc310...111d*
Laundering	36 – 50	Funds transfer to Tornado Cash. *Tx: 0x46dc...a2b1*

Figure 17 illustrates the trust dynamics of both the Baseline and ChronosRep models under this real-world trace. During the

[1] `https://etherscan.io/address/0x5f259d0b76665c337c6104145894f4d1d2758b8c`

Investment Phase ($t < 35$), both models maintain a high trust score ($R > 0.9$), consistent with the "Sleeper Agent" strategy of accumulating reputation. However, at the moment of the exploit ($t = 35$):

- Baseline Failure (Reputation Lag): The static decay model (blue dashed line) exhibits significant inertia. Due to the weight of historical honest behavior, the reputation score declines slowly, remaining above the trust threshold ($\tau = 0.4$) for several blocks post-attack. This lag creates a vulnerability window permitting potential exit scams.
- ChronosRep Success (Instant Collapse): The OU-Jump mechanism in ChronosRep detects the high-entropy between the sudden malicious evidence ($e_t \approx 0$) and the established prior. This triggers the jump component (dJ), causing an instantaneous vertical collapse in the trust score to near-zero ($R \approx 0.05$).

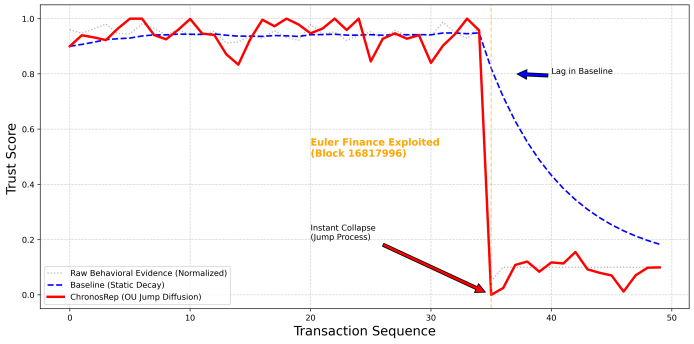

Figure 17: Comparative Trust Dynamics under the Euler Finance Exploit Trace. While the Baseline model (blue) exhibits significant retention of trust due to historical inertia ("Lag"), ChronosRep (red) leverages the OU-Jump process to trigger an instantaneous collapse in reputation score at the moment of the attack ($t = 35$), effectively mitigating the sleeper agent's exit window.

This forensic reconstruction confirms that ChronosRep effectively eliminates the time-to-detection latency in real-world exploit scenarios, strictly isolating the adversary within a single block.

7. Discussion

The findings presented in this paper advocate for a fundamental shift in how we conceptualize and engineer trust in decentralized systems: from viewing trust as a static, accumulated asset to understanding trustworthiness as a dynamic, continuously negotiated process. The experimental results validate the core hypothesis of ChronosRep: that combining identity-grounded quantification with adaptive temporal dynamics is essential for creating secure and fair reputation systems in decentralized environments.

7.1. Structural Analysis of Performance Gains and System Implications

The empirical results, particularly the 92.7% reduction in TTD, necessitate a rigorous structural interpretation to distinguish fundamental model advantages from parametric sensitivity. Critically, this performance leap is not an artifact of hyperparameter tuning (overfitting), but rather the deterministic result of shifting from linear accumulation to stochastic change-point detection. Traditional reputation models rely on historical averaging (e.g., $R_t = \frac{1}{t}\sum S_i$), where the weight of past honest behavior creates a temporal inertia that inherently delays the detection of new faults.

In contrast, ChronosRep's OU-Jump formulation (Equation 12) mathematically decouples historical stability from instantaneous reaction. The jump diffusion term JdN_t functions as a structural circuit breaker: when the behavioral deviation ε_t exceeds the noise threshold σ, the adaptive drift θ_t (Equation 15) forces a non-linear collapse of the trust score. Consequently, the rapid TTD is not a tunable outcome but a structural property of the stochastic differential equation itself, which treats trust as a volatile state rather than an accumulated asset.

The implications of this work extend beyond mere algorithmic efficiency to touch upon security architecture, social resilience, and regulatory compliance. From a security perspective, the rapid isolation of sleeper agents demonstrates that reputation must be treated as a continuously earned, perishable commodity rather than a permanent asset. This paradigm shift enables protocols in DeFi and DAOs to resist long-term strategic manipulation. For example, in DAO governance, ChronosRep prevents a patient adversary from slowly accumulating voting power through seemingly benign participation before executing a hostile takeover. In undercollateralized lending, it provides a dynamic credit score sensitive to sudden behavioral shifts, effectively mitigating "exit scams".

Complementing this rigorous security posture is the system's capacity for social forgiveness. In real-world systems, participants may make honest mistakes or suffer temporary performance lapses. Unlike static models that permanently penalize such actors, thereby encouraging identity whitewashing, ChronosRep supports redemption through sustained, reformed behavior. The adaptive nature of λ_t allows protocol designers to fine-tune this "forgiveness rate", tailoring the trust model to specific risk profiles without compromising systemic integrity.

From a deployment perspective, ChronosRep offers distinct advantages in explainability and Regulatory Compliance. Distinct from emerging reputation systems based on Deep Learning or opaque machine learning pipelines, which often function as "black boxes", the entropy-regularized Dempster-Shafer framework operates as a "glass box". The system can explicitly mathematically trace the cause of a low reputation score, whether due to high uncertainty (entropy), conflicting credentials (e.g., "VC from Issuer A contradicts VC from Issuer B"), or behavioral decay. This property is essential for compliance with strict data protection regulations such as the General Data Protection Regulation (GDPR), which mandates a "Right to Explanation" for automated decision-making. By providing a transparent audit trail for trust decisions, ChronosRep bridges the gap between decentralized privacy (via SSI) and institutional accountability.

7.2. Limitations

While ChronosRep demonstrates strong performance, several limitations acknowledge the complexity of decentralized trust. One notable dependency is the assumption of a trustworthy issuer ecosystem. As our collusive issuer analysis reveals, this reliance creates a potential attack vector if issuers themselves are compromised.

From a computational perspective, the use of Louvain-based clustering for collusion detection, while effective, incurs notable costs in large-scale settings. The semantic resolution of the behavioral model also presents constraints. The current framework treats actions as binary (success/failure), which oversimplifies nuanced real-world scenarios such as partial loan repayments or ambiguous voting outcomes.

Furthermore, the optimization of system parameters, such as α, β, and the trust threshold τ, currently relies on manual tuning based on simulation sensitivity analysis rather than adaptive mechanisms.

8. Conclusion and Future Work

This paper presented ChronosRep, a comprehensive protection technology designed to secure this critical data asset against corruption by strategic adversaries. By fusing off-chain identity compliance with on-chain behavioral dynamics, we have established a hybrid security framework that bridges the gap between rigid regulatory requirements and dynamic decentralized trust. Our theoretical contribution addresses two fundamental challenges in data security: the fusion of conflicting, high-entropy evidence sources and the mitigation of *"reputation lag"* vulnerabilities. The proposed entropy-regularized Dempster-Shafer mechanism ensures that identity data remains robust against Sybil-based distortion, while the Ornstein-Uhlenbeck jump diffusion model provides an adaptive defense scheme that instantaneously isolates threats upon detection. Empirical validation confirms that ChronosRep not only reduces the time-to-detection of sleeper agents by over 90% compared to static baselines but also maintains a fair pathway for the recovery of compliant nodes, ensuring systemic resilience.

Future work will focus on scaling this architecture to meet the high-throughput demands of next-generation networks. Specifically, we aim to optimize the computational overhead of the evidence fusion engine for edge-constrained devices and explore privacy-preserving zero-knowledge proofs to further enhance data compliance. Eventually, ChronosRep advances the state-of-the-art in computational trust, moving beyond static scoring to a continuous, mathematically grounded defense architecture for the secure agentic web.

References

[1] S. Bano, A. Sonnino, M. Al-Bassam, S. Azouvi, P. McCorry, S. Meiklejohn, G. Danezis, Sok: Consensus in the age of blockchains, in: Proceedings of the 1st ACM Conference on Advances in Financial Technologies, 2019, pp. 183–198.

[2] S. Werner, D. Perez, L. Gudgeon, A. Klages-Mundt, D. Harz, W. Knottenbelt, Sok: Decentralized finance (defi), in: Proceedings of the 4th ACM Conference on Advances in Financial Technologies, 2022, pp. 30–46.

[3] V. Mohammadi, A. M. Rahmani, A. M. Darwesh, A. Sahafi, Trust-based recommendation systems in internet of things: a systematic literature review, Human-centric Computing and Information Sciences 9 (1) (2019) 21.

[4] T. Dimitriou, Decentralized reputation, in: Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, 2021, pp. 119–130.

[5] F. Teng, C. Du, M. Shen, P. Liu, A dynamic large-scale multiple attribute group decision-making method with probabilistic linguistic term sets based on trust relationship and opinion correlation, Information Sciences 612 (2022) 257–295. doi:10.1016/j.ins.2022.07.092.

[6] The market for "lemons": Quality uncertainty and the market mechanism, in: Uncertainty in Economics, Academic Press, 1978, pp. 235–251. doi:10.1016/B978-0-12-214850-7.50022-X.

[7] O. Goldreich, S. Micali, A. Wigderson, Proofs that yield nothing but their validity or all languages in NP have zero-knowledge proof systems, J. ACM 38 (3) (1991) 690–728. doi:10.1145/116825.116852.

[8] P. Resnick, Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System, Advances in Applied Microeconomics 11 (Oct. 2002). doi:10.1016/S0278-0984(02)11030-3.

[9] S. D. Kamvar, M. T. Schlosser, H. Garcia-Molina, The eigentrust algorithm for reputation management in p2p networks, in: Proceedings of the 12th international conference on World Wide Web, 2003, pp. 640–651.

[10] J. Sedlmeir, R. Smethurst, A. Rieger, G. Fridgen, Digital identities and verifiable credentials, Business & Information Systems Engineering 63 (5) (2021) 603–613.

[11] H. Yu, Z. Shen, C. Miao, C. Leung, D. Niyato, A survey of trust and reputation management systems in wireless communications, Proceedings of the IEEE 98 (10) (2010) 1755–1772.

[12] I. Pinyol, J. Sabater-Mir, Computational trust and reputation models for open multi-agent systems: a review, Artificial Intelligence Review 40 (1) (2013) 1–25.

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

[14] A. Josang, R. Ismail, The beta reputation system, in: Proceedings of the 15th bled electronic commerce conference, Vol. 5, 2002, pp. 2502–2511.

[15] S. Sirur, T. Muller, Properties of reputation lag attack strategies, in: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS '22, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2022, p. 1210–1218.

[16] M. Richardson, R. Agrawal, P. Domingos, Trust management for the semantic web, in: International semantic Web conference, Springer, 2003, pp. 351–368.

[17] A. Mühle, A. Grüner, T. Gayvoronskaya, C. Meinel, A survey on essential components of a self-sovereign identity, Computer Science Review 30 (2018) 80–86.

[18] Y. Sellami, Y. Imine, A. Gallais, Fog-blockchain fusion for event evaluation and trust management, IEEE Transactions on Dependable and Secure Computing 22 (6) (2025) 6539–6553. doi:10.1109/TDSC.2025.3587589.

[19] H. Yu, M. Kaminsky, P. B. Gibbons, A. Flaxman, Sybilguard: defending against sybil attacks via social networks, in: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications, 2006, pp. 267–278.

[20] D. Reed, M. Sabadello, Decentralized identifiers, Preukschat, Alex; Reed, Drummond (Hg.): Self-Sovereign Identity: Decentralized digital identity and verifiable credentials, Shelter Island, NY (2021) 157–188.

[21] T.-D. Tran, H. P. G. Bao, N. T. Cam, V.-H. Pham, Dave-cc: A decentralized, access-controlled, verifiable ecosystem for cross-chain academic credential management, Journal of Information Security and Applications 94 (2025) 104238.

[22] J. R. Douceur, The sybil attack, in: International workshop on peer-to-peer systems, Springer, 2002, pp. 251–260.

[23] R. Di Pietro, X. Salleras, M. Signorini, E. Waisbard, A blockchain-based trust system for the internet of things, in: Proceedings of the 23nd ACM on symposium on access control models and technologies, 2018, pp. 77–83.

[24] A. Salehi-Abari, T. White, Trust models and con-man agents: From mathematical to empirical analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 24, 2010, pp. 842–847.

[25] G. Shafer, Dempster-shafer theory, Encyclopedia of artificial intelligence 1 (1992) 330–331.

[26] R. Reiter, On closed world data bases, in: Readings in artificial intelligence, Elsevier, 1981, pp. 119–140.

[27] T. R. Gruber, A translation approach to portable ontology specifications, Knowledge acquisition 5 (2) (1993) 199–220.

[28] N. G. Van Kampen, Stochastic differential equations, Physics reports 24 (3) (1976) 171–228.

[29] R. A. Maller, G. Müller, A. Szimayer, Ornstein–uhlenbeck processes and extensions, Handbook of financial time series (2009) 421–437.

[30] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, Journal of statistical mechanics: theory and experiment 2008 (10) (2008) P10008.