Confianza Tejida: consenso y gobernanza anclados en reputación verificable

Whitepaper técnico, v3 — edición en español

Índice

Resumen

Presentamos Confianza Tejida (*Woven Trust*), la arquitectura de una red descentralizada en la que la autoridad — la elegibilidad para el comité de finalidad, los jurados de disputas y la gobernanza — se deriva de una métrica de reputación calculada de forma determinista sobre evidencia registrada en cadena, en lugar de derivarse de trabajo computacional o de capital en depósito. La reputación se computa por dominios separados a partir de dos fuentes objetivas: atestaciones bilaterales de transacciones liquidadas y un grafo de avales con fianza, propagadas por un algoritmo de la familia EigenTrust en aritmética de punto fijo (i128), de modo que todos los nodos obtienen valores idénticos bit a bit. Los poderes que exigen integridad transversal usan un agregado cóncavo de los cuatro dominios que veta a los perfiles sin cobertura mínima. La resistencia Sybil y a la colusión se obtiene por defensa en profundidad: descuento por proximidad de grafo, amortiguación de comunidades densas, fianzas de aval con ventanas de prueba y un estrato judicial por sorteo ponderado. La finalidad la otorga un comité renovado por épocas,

extraído por sorteo sembrado con un faro de compromiso-y-revelado endurecido contra *grinding*, y protegido por una guardia de concentración que congela la finalidad, de forma reversible, si el peso de una sola entidad cruza un techo. Documentamos el modelo económico de doble activo con emisión ligada a servicio medido, el procedimiento de arranque con dilución fundacional forzada por protocolo, y una evaluación adversarial con libro de validación público: 17/17 escenarios de ataque resistidos por el motor de reputación, captura de jurado por colusión a dos saltos cuantificada y cerrada, y supervivencia del consenso a la pérdida y reincorporación de validadores a través de fronteras de época en dos entornos de pruebas independientes. Cerramos con las limitaciones abiertas: prueba de persona única, metadatos de transporte y parámetros aún en calibración.

1. Introducción

Los sistemas de consenso abiertos anclan su seguridad en un recurso que el adversario debe adquirir. La prueba de trabajo lo ancla en energía [8]; la prueba de participación, en capital. Ambos recursos son comprables en mercado, de modo que la seguridad de la red es, en el límite, una función del presupuesto del atacante. Este trabajo explora un anclaje distinto: **tiempo-reputación** — historial verificable de conducta dentro del propio sistema, que se acumula despacio, decae con el desuso y no admite transferencia. Un adversario no puede alquilar años de historial atestado por contrapartes independientes; tiene que producirlos bajo las mismas reglas que vigilan al resto de participantes.

La dificultad clásica de las métricas de reputación es doble. Primero, el ataque Sybil [2]: si crear identidades es gratis, cualquier métrica agregable se infla con identidades fabricadas. Segundo, la colusión: subconjuntos coordinados que se atestan mutuamente pueden fabricar señal sin actividad económica real. Un tercer problema, menos tratado en la literatura, es institucional: en los sistemas de reputación de plataforma [7], el operador custodia la base de datos y puede editarla, con lo que la métrica hereda la confianza — y la corruptibilidad — de su custodio.

Contribuciones. Este paper describe un sistema completo, implementado y validado adversarialmente, con las siguientes aportaciones:

1. Un **esquema de reputación multidominio** computado exclusivamente desde evidencia objetiva en cadena (atestaciones bilaterales de operaciones liquidadas; avales con fianza), sin puntuaciones subjetivas, con decaimiento temporal por dominio y agregación cóncava anti-especialista (§4).
2. Una **batería anti-colusión en profundidad** — descuento por proximidad, amortiguación de comunidades, fianzas con ventana de prueba y cascada de responsabilidad acotada — evaluada contra un catálogo de ataques (§5, §6).
3. Un **mecanismo de finalidad por comité reputacional**: sorteo ponderado sobre un faro de compromiso-y-revelado con guardia de peso reputable (anti-*grinding* y anti-congelación), votos firmados con separación de dominio, y una guardia de concentración con parada reversible (§7).
4. Un **modelo económico de doble activo** en el que la emisión paga servicio medido algorítmicamente y ninguna autoridad discrecional acuña ni asigna (§9).
5. Una **metodología de validación** de evidencia-sobre-afirmación con libro público que liga cada propiedad declarada a una ejecución reproducible y al hash del binario que la produjo (§11).

Implementación. El sistema está escrito en **Rust** sobre un framework de runtime de cadena de bloques de propósito general (familia Substrate): los mecanismos de este paper (reputación, justicia, faro, tokens) son módulos de runtime independientes, y los motores de cómputo (propagación de reputación, elección de comité, faro) viven en *crates* núcleo sin dependencias externas, compilables a `no_std`/wasm. Todos los caminos de cómputo relevantes para el consenso corren en aritmética entera de punto fijo (i128, escala fija) para garantizar determinismo entre arquitecturas. El análisis adversarial previo al cableado se hizo sobre prototipos en Python de biblioteca estándar con generación aleatoria sembrada, contrastados después contra los núcleos de Rust mediante tests-oráculo de paridad exacta (Apéndice A).

2. Trabajo relacionado

Propagación de confianza. EigenTrust [1] calcula una confianza global como punto fijo de una cadena de Markov sobre el grafo de interacciones, con un vector de pre-confianza que ancla el cálculo. Nuestro motor pertenece a esta familia (§4.3), con dos diferencias: la pre-confianza se deriva de evidencia objetiva y no de un conjunto fiado a priori, y el cálculo corre por dominios separados en punto fijo determinista. La estructura de teletransporte es análoga a la de PageRank [6].

Defensas Sybil estructurales. SybilGuard [3] y sucesores explotan que las identidades fabricadas se conectan al grafo honesto por pocos cortes. Nuestro descuento por proximidad y la amortiguación de comunidades densas (§5) siguen el mismo principio estructural, aplicado al peso de la evidencia en lugar de a la admisión de identidades.

Finalidad por comité. La separación entre producción de bloques y finalidad por votación de un comité es análoga a los *gadgets* de finalidad de tipo BFT (p. ej. GRANDPA [9]). La diferencia central es la fuente del peso de voto: reputación no transferible en lugar de participación económica en depósito.

Funciones de retardo verificable. El endurecimiento propuesto del faro contra el sesgo del último revelador usa VDFs [4].

Reputación de plataforma. Los sistemas de reseñas centralizados [7] agregan opiniones subjetivas bajo un custodio editable; el presente diseño invierte las tres decisiones (hechos en lugar de opiniones, multidominio en lugar de escalar, recomputación universal en lugar de custodia).

Precedentes institucionales. La viabilidad de órdenes de intercambio sostenidos por reputación sin ejecución estatal está documentada históricamente; véase el análisis del derecho mercantil medieval de Milgrom, North y Weingast [5] y, sobre autogobierno de recursos comunes, Ostrom [10].

3. Modelo del sistema y modelo de amenaza

3.1 Actores y supuestos

Los participantes se identifican con seudónimos criptográficos: pares de claves sr25519; la dirección derivada de la clave pública es la identidad completa del miembro dentro del sistema. No existe vínculo con identidad legal alguna, por diseño (la capa de red se trata en §12). Los nodos replican

la cadena y recalculan todo el estado de reputación de forma determinista; no existe ningún nodo con estado privilegiado.

Se asume el modelo habitual de red parcialmente síncrona, criptografía estándar (firmas sr25519, hashes resistentes a colisión) y un adversario adaptativo que puede: (a) crear identidades ilimitadas a coste marginal nulo, (b) coordinar subconjuntos de identidades propias o sobornadas, (c) retener o reordenar sus propios mensajes de protocolo, y (d) participar honestamente durante periodos largos antes de desviarse.

3.2 Objetivos de seguridad

- **S1 (anti-Sybil):** identidades sin evidencia independiente no acumulan peso de consenso, con independencia de su número.
- **S2 (anti-colusión):** el peso obtenible por un subconjunto coordinado crece con su evidencia externa al subconjunto, no con su actividad interna.
- **S3 (no-transferibilidad):** el peso no puede comprarse, venderse ni heredarse; decae sin actividad.
- **S4 (determinismo):** todos los nodos computan pesos idénticos bit a bit desde el mismo prefijo de cadena.
- **S5 (seguridad de finalidad):** ningún bloque queda finalizado sin supermayoría del peso reputacional del comité vigente; bajo concentración excesiva de peso el sistema prefiere detener la finalidad (reversiblemente) a finalizar bajo captura.
- **S6 (no-discrecionalidad):** ninguna cuenta posee capacidad de interpretación, edición o asignación discrecional; donde hace falta juicio humano, se sortea (§8).

3.3 Catálogo de ataques considerado

Anillos Sybil; clústeres de colusión (admiración mutua); *whitewashing* (abandono y reentrada); *slashing* dirigido a un dominio de la víctima; aval-kamikaze y bomba-aval (§6); *grinding* del faro y retención de revelados (§7.3); congelación de finalidad por participante sin peso (§7.3); captura de jurado por colusión a dos saltos (§8); acaparamiento del comité durante la ventana de arranque (§10).

4. El sistema de reputación

4.1 Evidencia admisible

Solo dos tipos de eventos generan señal de reputación, ambos objetivos y verificables en cadena:

1. **Atestación bilateral.** Una contraparte registra el hecho y la beneficiaria lo reclama — dos firmas, dos voluntades. No existen puntuaciones graduadas ni reseñas de terceros: cuenta solo el testimonio de una contraparte que empeña su propio nombre, y cuatro frenos deterministas le ponen precio a la fabricación de señal: un presupuesto de atestaciones por época, un suelo de reputación exigido al atestante, un tope de magnitud (nadie da fe de más peso del que él mismo posee) y un desgaste geométrico por pareja que satura la señal repetida. La liga dura entre atestación y operación liquidada en cadena — de modo que solo un trato liquidado pueda atestarse — es un endurecimiento declarado, pendiente de la capa de liquidación (§12).
2. **Aval con fianza** (§4.2).

Qué tipos de evento son admisibles lo fijan reglas públicas y recomputables, sin autoridad de admisión discrecional (objetivo S6). La clave de arranque que siembra los tipos iniciales expira automáticamente a una altura de bloque fijada (§10).

4.2 El aval como fianza

Un miembro puede avalar la entrada de otro. El aval: (i) bloquea una fracción de la reputación del avalista, en el mismo dominio, como prenda; (ii) la prenda se libera gradualmente a medida que el avalado acumula reputación **independiente**, definida como evidencia con contrapartes cuya proximidad de grafo al avalista queda por debajo de un umbral; (iii) el aval no transfiere peso: un miembro cuya única arista es su avalista tiene peso estructural nulo; (iv) el fraude del avalado propaga pérdida hacia la cadena de avales con topes por incidente (§6). La estructura de avales forma un grafo dirigido con raíz en el génesis, público y analizable: los anillos Sybil presentan firmas topológicas detectables.

4.3 Motor de propagación

Sea G_d el grafo de evidencia del dominio d (aristas ponderadas por atestaciones y avales, con descuentos de §5). El peso s_d de cada identidad en el dominio d se computa como el punto fijo de una iteración de propagación de la familia EigenTrust [1]:

- **Pre-confianza:** el vector de anclaje no es un conjunto fiado a priori sino la distribución de evidencia objetiva (operaciones liquidadas, obra verificada) de cada identidad.
- **Teletransporte:** con probabilidad fija, la iteración salta a la distribución de pre-confianza, lo que acota la influencia obtenible por subgrafos densos sin evidencia externa (análogo al factor de amortiguación de PageRank [6]).
- **Punto fijo determinista:** toda la aritmética es entera (i128, escala fija). No hay coma flotante en ningún camino de consenso; el resultado es idéntico bit a bit en toda arquitectura (objetivo S4).

4.4 Cuatro dominios y agregación cóncava

La reputación se computa en cuatro dominios independientes: intercambio, obra, juicio y gobernanza (los cuatro palos de la baraja, en la nomenclatura del proyecto). Cada dominio admite únicamente evidencia de su especie. Para los poderes que exigen integridad transversal (comité de finalidad, jurados, potestad de avalar), los cuatro valores se agregan:

$$\mathrm{agg}(s) = (1-\lambda)\cdot \min_d(s_d) + \lambda \cdot \mathrm{med}_d(s_d)$$

con $\lambda \in [0,1]$ un dial de mezcla (provisionalmente $\lambda = 0.5$; parámetro abierto en calibración final, §12). La forma cóncava conserva la propiedad anti-especialista — cubrir menos de la mitad de los dominios da mediana nula y agregado casi nulo para todo λ, de modo que λ **no puede** rescatar a un perfil sin cobertura — mientras suaviza el mínimo puro, que el análisis adversarial mostró convertible en arma: con agg = min, un golpe dirigido a un solo dominio anula a la víctima ($-100\,\%$); con $\lambda = 0.5$ la herida se acota a $\approx -50\,\%$ en el caso simétrico ideal ($-47\,\%$ medido sobre perfiles reales en hardware, libro fila 16). El valor $\lambda = 0.5$ es provisional y está pendiente de re-validación contra la batería de ataques completa (§12).

4.5 Decaimiento

El peso decae multiplicativamente por época en ausencia de evidencia nueva, por dominio, con semivida de ≈ 24 meses (retención por época $\rho = 0.999052$). El decaimiento implementa la no-transferibilidad temporal (S3): no existen cuentas envejecidas con valor de reventa, y la ventaja de los participantes tempranos se disuelve sin intervención (§10).

5. Defensas anti-colusión

Ninguna defensa aislada satisface S2; el sistema apila cuatro, todas deterministas:

1. **Descuento por proximidad.** El peso de toda atestación se descuenta como función de la distancia de grafo entre atestante y atestado. La señal fabricada dentro de un clúster es estructuralmente barata.
2. **Amortiguación de comunidades.** Las comunidades densas cuya suma de aristas internas excede en proporción configurada su evidencia externa se amortiguan como grupo (detección e imposición en punto fijo, deterministas).
3. **Fianzas y cupos de aval** (§4.2) acotan el abanico de identidades fabricadas financiables por un atacante con presupuesto de reputación dado.
4. **Estrato judicial.** La colusión probada ante jurado (§8) se sanciona en los cuatro dominios simultáneamente, como único caso de sanción transversal (§6).

6. Sanción y cascada de responsabilidad

El fallo de confianza propaga responsabilidad con tres propiedades, cada una validada frente a su ataque específico:

- **Proporcionalidad con tope.** La pérdida del avalista por fraude del avalado se acota por incidente. Cierra el **aval-kamikaze**: usar un fraude enorme de una identidad sacrificable para aniquilar a un avalista honesto.
- **Ventana de prueba.** El fraude cometido inmediatamente después de recibir un aval no propaga cascada. Cierra la **bomba-aval**: solicitar un aval y detonarlo en la misma ventana.
- **Localidad de dominio.** La sanción cae en el dominio de la falta; la única excepción transversal es la traición estructural (operación Sybil probada, colusión probada, ataque a la red).

Las sanciones son graduales (marca pública, suspensión, expulsión), priorizan la restitución y son reversibles por el mismo proceso que las impone. El mecanismo completo de expulsión-sin-confiscación — cómo la expulsión acarrea consecuencias reales mientras los saldos permanecen estructuralmente intocables — se desarrolla en el Apéndice B.

7. Consenso

7.1 Arquitectura

La producción de bloques se desacopla de la finalidad. La producción usa un mecanismo convencional de ronda; la **finalidad** la otorga un comité extraído por sorteo ponderado por $\mathrm{agg}(s)$ (§4.4), renovado cada época. Un bloque es final cuando acumula votos firmados que representan una supermayoría del peso reputacional del comité vigente.

7.2 Votos

Los votos son firmas sr25519 con separación de dominio sobre (altura, hash). En la difusión, la firma se verifica antes de retransmitir cada voto, de modo que un par que inunde la red con votos falsificados no se amplifica a través de los nodos honestos. En el conteo, cada voto pasa por la comprobación de pertenencia al comité y la verificación de firma, y se deduplica por (altura, votante). La renovación de época poda y refresca el estado del comité en la transición; una condición de carrera en la frontera (miembros del comité saliente votando a través del borde), descubierta en pruebas multi-nodo de caída y reincorporación, fue reproducida y cerrada (libro de validación, filas 4 y 33).

7.3 El faro de aleatoriedad

El sorteo del comité y de los jurados se siembra con un faro de **compromiso-y-revelado**: los participantes comprometen un valor, lo revelan en la ventana siguiente, y la entropía combinada de los revelados siembra la selección. Dos endurecimientos:

1. **Sesgo del último revelador.** Se propone superponer una función de retardo verificable [4] a la salida del faro, eliminando la ventaja de decidir revelar o no una vez vistos los demás revelados.
2. **Guardia de peso reputable.** El diseño original congelaba la selección si el conjunto de revelados era débil, lo que permitía a un solo participante sin reputación — comprometiendo y no revelando jamás — detener la finalidad de toda la red indefinidamente. La versión corregida solo delega en el faro cuando el conjunto comprometido acumula peso reputable suficiente; en caso contrario la selección cae a un mecanismo determinista de respaldo. El ataque fue reproducido en test antes de la corrección y fracasa después de ella.

7.4 Guardia de concentración

Una guardia en cadena monitoriza la concentración del poder de finalidad. Si el peso de una sola entidad supera un techo configurado, la finalidad se **congela de forma reversible** y está diseñada para recuperarse sola cuando la concentración se re-dispersa (el estado actual de la ruta de recuperación automática se detalla en §12). La justificación es de teoría de la seguridad: detener la finalidad es un fallo de vivacidad recuperable; finalizar historia bajo comité capturado es un fallo de seguridad irreversible. La guardia prefiere el primero.

8. Resolución de disputas

Las disputas entre miembros se resuelven por jurados sorteados, con el sorteo ponderado por el dominio de juicio y tres restricciones estructurales:

1. **Exclusión de interesados:** las partes y toda identidad con proximidad de grafo medible a la disputa quedan fuera de la urna.
2. **Independencia por pares:** el jurado debe satisfacer una distancia de grafo mínima entre cada par de jurados — una disputa no puede ser juzgada por un clúster.
3. **Cierre por supermayoría:** el veredicto lo decide una supermayoría del jurado sorteado; las dos protecciones anteriores garantizan que esa mayoría no es la de un clan interesado. (Una capa adicional que puntúe a cada jurado por coherencia con la evidencia — recompensando

o penalizando el voto según se alinee con el expediente, en la línea de un punto de Schelling [11] — es trabajo futuro, §12; hoy no está implementada.)

La captura de jurado por colusión a dos saltos se cuantificó en simulación (8–10 $\times 10^3$ ensayos por configuración) y se cerró con penalizaciones de peso a profundidad 2, verificadas en unitarios y en devnet. El veredicto opera únicamente sobre el seudónimo y el hecho probado; el sistema no posee información para des-seudonimizar, y ningún procedimiento la genera (§12). El jurado *certifica* el hecho en lugar de decretar un castigo; las consecuencias de solo-estatus que se siguen, y su fundamento doctrinal y empírico, se exponen en el Apéndice B.

9. Modelo económico

El dinero no es poder. Antes que cualquier mecanismo, una regla gobierna la economía y la ata a la constitución (§6, «el origen del poder»): la moneda no compra autoridad. Ni HLQ ni SOV pueden comprar reputación, peso de consenso, un asiento de jurado ni voz alguna en las decisiones comunes; no existe *staking*-por-poder en ninguna parte del protocolo. La riqueza y la influencia son *ejes ortogonales* — se puede tener mucho de una y nada de la otra. Esto es lo que hace el modelo anarcocapitalista y no plutocrático, y es la propiedad que cada decisión de diseño de abajo está construida para proteger: la capa de retribución nunca debe volverse una puerta trasera del saldo a la autoridad.

Doble activo. Dos monedas, cada una con un rol distinto y un tope duro que nunca se supera — sin inflación artificial, sin banco central, sin acuñación discrecional. **HLQ** (tope duro 540×10^6, finamente divisible) es el medio de intercambio — la moneda del trato diario, las tasas y los pagos. **SOV** (tope duro 54×10^6, indivisible) es la reserva — escasa, lenta, para guardar más que para gastar. Las tasas de transacción se pagan en HLQ y se validan contra el mismo libro contable que las cobra — un descuadre entre el libro de validación de tasas y el de cobro, hallado en pruebas de estrés adversarial, habría impedido a toda cuenta nueva pagar su primera tasa; está corregido y cubierto por regresión.

Emisión por servicio medido — el mecanismo. No existe emisión por calendario. En cada frontera de época el runtime lee, de un feed de participación no falsificable, el servicio verificado que rindió cada nodo esa época — bloques que autoreó en sus turnos elegidos más votos de finalidad emitidos como miembro del comité, ambos re-comprobados contra firmas sr25519 y pertenencia al comité antes de contar. Esto da un vector de pesos sobre los nodos que sirvieron. Antes de repartir moneda, el vector pasa un **tope anti-sybil por nodo**, $w_i \leftarrow \min(w_i,\ \lceil \text{mediana}(w) \cdot k \rceil)$ con $k \approx 2$: el tope aplana la cola alta para que ningún operador — ni una granja de máscaras fingiendo serlo — pueda llevarse una porción desproporcionada por pesar más que la mediana. El fondo programado de la época se divide entonces **en proporción a los pesos capados** (`split_service_reward`) y se acredita. El orden importa: los receptores se verifican *antes* de acuñar, así que una época sin servicio verificado no acuña nada ni consume nada contra el tope — cerrando el defecto de acuñación fantasma (una versión previa reservaba el tope primero y vaporizaba la emisión no reclamada en oferta fantasma permanente).

Por qué servicio y no stake. Ligar la emisión a servicio medido y firmado — y no a un calendario ni a tenencias — es lo que mantiene la entrada de moneda alineada con la seguridad de la red: la única forma de recibir moneda nueva es haber hecho trabajo de consenso que la cadena pudo verificar y que el tope anti-sybil no pudo inflar. El peso comprable (los ataques de flash-loan y delegación pagada que se repiten en la gobernanza on-chain) no tiene aquí de dónde

agarrarse, porque la acuñación lee *servicio*, no saldo.

SOV por aguante. Los dos activos se ganan distinto a propósito. HLQ, el medio de intercambio, se paga por servicio puntual cada época. SOV, la reserva, se paga por **constancia**: el peso SOV de un nodo en una época es su servicio multiplicado por su racha de épocas consecutivas (con techo k; la racha se reinicia a cero la primera época que falla). Dos nodos que hacen el mismo trabajo en una época ganan el mismo HLQ, pero el que ha servido sin interrupción gana mucho más SOV. Así la reserva se acumula por aguante — el único recurso que un sybil no puede fabricar, porque es tiempo realmente invertido, no capital ni reputación multiplicables entre máscaras — y derrota directamente al cártel de entrar-y-salir que cosecha recompensas puntuales sin sostener jamás la red.

Desinflación — la curva. La emisión sigue una curva geométrica, siempre acotada por el tope duro que nunca puede superar: procede por eras, y cada era la recompensa programada se escala por una razón fija y pública — 15/16 para HLQ, un decaimiento deliberadamente plano que reparte la emisión entre generaciones de operadores en lugar de concentrarla al principio, y 3/4 para SOV, una senda más pronunciada para la moneda más escasa. El decaimiento reputacional (§4.5) corre con su propia constante, sin relación ($\rho = 0.999052$ por época, semivida ≈ 24 meses). Las dos capas implementan la misma *propiedad* — la ventaja temprana se diluye sola, por aritmética, sin que nadie lo decida — pero con constantes separadas, cada una afinada a su capa, no con un parámetro compartido.

Comisiones y la quema — una oferta en dos sentidos. La emisión por servicio es unidireccional: solo puede añadir moneda. Para dar a la oferta una vía de *contraerse* con la demanda — la disciplina que la redención aporta en los sistemas de banca libre, y que a una moneda sin respaldo le falta —, una fracción fija de cada tasa de transacción HLQ se **quema** (se destruye, no se recicla a una tesorería), pagando el resto al nodo procesador. La moneda entra así solo por la curva capada y sale solo por el fuego, de modo que el flotante responde en ambos sentidos a cuánto se usa de verdad la red.

Separación de retribuciones. El principio es doble: cobrar no da voz (el pago no confiere peso de gobernanza) y el asiento no cobra por serlo (la emisión paga servicio medido, no rango). El servicio que la cadena mide y paga hoy es el de comité (autoría de bloques y votos de finalidad, con tope anti-acaparador); el pago del servicio de infraestructura pura (alojamiento, custodia, disponibilidad) mediante pruebas auto-verificables está diseñado y aún no cableado (§12). Las identidades fundacionales arrancan sin saldo alguno — el génesis no asigna moneda a nadie — y el compromiso de diseño es que tampoco perciban emisión por su servicio de arranque; el mecanismo que lo fuerza (exclusión en protocolo o separación operativa verificable) está en decisión final y se documentará con su validación.

Sin premine, sin venta, sin acuñación discrecional. No hay ICO, ni asignación a inversores, ni premine: no se crea moneda para iniciados antes de que la red corra, ni se reparte por persona ni por cuenta. La moneda entra *solo* pagando servicio medido, por el calendario decreciente publicado, y circula después por intercambio voluntario — jamás por decreto, jamás por una acuñación que alguien controle. Las identidades fundacionales arrancan con saldo cero y no perciben emisión por su servicio de arranque (§10). Los instrumentos colectivos (fondos de bienvenida, ayuda mutua, casas de cambio) pertenecen a la capa de libre asociación: los construye y los posee quien los levanta, fuera del protocolo, que no los prohíbe ni los opera.

Parámetros abiertos: forma exacta de la curva de emisión; λ; calendario de prueba-de-servicio; reparto interno de retribuciones. Siguen en calibración; los valores con los que la red corre hoy son provisionales, ajustables por el proceso de actualización, y cada uno se publicará con su evidencia a medida que se fije (§12).

10. Procedimiento de arranque

El arranque de una economía de reputación es su problema de seguridad más delicado: toda asimetría inicial es una semilla de captura. Reglas del génesis:

1. **Siembra mínima, pública y caduca.** El génesis siembra una cohorte fundacional etiquetada como andamiaje, con decaimiento activo desde el bloque 0 y emisión cero. Regla de ceremonia verificable: toda identidad sembrada corresponde a un nodo operativo en el génesis.
2. **Dilución fundacional por protocolo.** La ventaja de los nodos fundacionales se disuelve por el decaimiento de reputación (§4.5): sin actividad atestada, su peso cae bajo el del ciudadano medio en poco más de una semivida, sin intervención humana. La clave de arranque que admite los tipos de evidencia iniciales expira automáticamente a una altura de bloque fijada — por protocolo, no por buena voluntad — condición implementada y probada en hierro (libro, fila 14). La autoridad de actualización fundacional (multifirma, vigente solo durante la construcción) porta además expiración en el propio pallet de actualización, con una holgura de convergencia todavía en revisión (§12). Un ocaso explícito del *rol de validador* fundador ligado a una métrica de dispersión en cadena es trabajo futuro (§12), no un mecanismo cableado hoy: la dispersión efectiva la produce, por ahora, el decaimiento.
3. **Techo de concentración** (§7.4) activo desde el bloque 0, cuando la red es más pequeña y vulnerable.
4. **Entropía externa.** El protocolo de génesis ancla la semilla a un bloque público de Bitcoin fijado en la ceremonia, eliminando la capacidad de los propios fundadores de moler su génesis. La ceremonia de lanzamiento se ejecutó el 18 de julio de 2026: la semilla queda anclada al bloque de Bitcoin 958536, y la red corre desde el hash de génesis `0xa1ab5e0b...da1fed6`.
5. La captura y la expiración de la ventana de arranque se barrieron adversarialmente en simulación antes de la existencia de valor (libro, fila 10), y el arranque sin captura se validó además en hardware (fila 13).
6. **Primer upgrade constitucional, ejecutado en vivo (24 de julio de 2026).** Seis días después del génesis, la red ejerció la autoridad fundacional de actualización de punta a punta sobre la cadena viva: un upgrade del runtime se propuso en el bloque 26798 desde un asiento fundador, se co-firmó hasta el quórum de 3-de-5 (el asiento de custodia no puede firmar upgrades), se expuso a la ventana de objeción completa de 14.400 bloques y se aplicó en el bloque 41367. Dos propiedades del mecanismo quedaron confirmadas en producción y no solo sobre el papel. Primera: la aplicación es deliberadamente sin permiso y exenta de tasas una vez vencida la ventana — la autoridad vive en el quórum sellado y en el código anclado por hash (el `blake2_256` del código enviado debe igualar el hash anunciado), no en quien lo envía; de hecho el upgrade se aplicó desde una cuenta recién nacida y sin saldo. Segunda: la primera rotación de comité bajo el runtime nuevo (bloque 43200, la frontera de época que había detenido a una compilación incoherente durante la validación previa al lanzamiento) se completó con la finalidad ininterrumpida. Libro, fila 35.

11. Evaluación

Metodología. El proyecto opera bajo un protocolo de evidencia-sobre-afirmación: ninguna propiedad se declara sin una ejecución que la exhiba, y las auditorías preceden a la existencia de valor. El comportamiento multi-nodo se valida en devnets de dos entornos independientes operados por separado; los hallazgos se contrastan con paneles de revisión adversarial independientes bajo un protocolo anti-alucinación (toda afirmación se verifica contra ejecución, nunca se acepta de prosa). Un **libro de validación público** (`VALIDATION-LEDGER.md` en el repositorio público del proyecto) liga cada afirmación a la ejecución que la sostiene, el hash del binario que la produjo y las salidas en crudo (35 filas al cierre de esta versión). El **Apéndice A** enumera la batería adversarial completa y traza cada resultado de la tabla siguiente a su fila del libro, su comando de reproducción y el entorno donde corrió.

Resultados seleccionados.

Propiedad	Método	Resultado
Resistencia del motor de reputación (anillos Sybil, clústeres de colusión, slash dirigido, whitewashing)	batería adversarial	17/17 escenarios resistidos
Seguridad y vivacidad del consenso con caída y reincorporación de validadores a través de fronteras de época	devnet multi-nodo, dos entornos	supervivencia verificada; carrera de frontera cerrada
Captura de jurado por colusión a dos saltos	simulación, $8\text{–}10 \times 10^3$ ensayos/configuración	cuantificada; cerrada con penalización a profundidad 2
Congelación de finalidad vía faro (retención de revelado)	reproducción en test	reproducida; cerrada con guardia de peso reputable
Aval-kamikaze y bomba-aval	batería adversarial	cerrados (topes por incidente; ventana de prueba)
Determinismo del cómputo de reputación	recomputación cruzada	idéntico bit a bit entre nodos y arquitecturas
Descuadre validación/cobro de tasas (cuenta nueva impagable)	estrés adversarial	hallado; corregido; regresión activa

Capa normativa. El documento constitucional del sistema fue sometido al mismo proceso adversarial (12 fallas halladas y resueltas antes del sellado), y cada mecanismo de este paper está trazado al artículo normativo al que sirve.

12. Limitaciones

Un estudio se gana la confianza nombrando lo que *no* ha resuelto. Lo que sigue son los residuos honestos del sistema tal como se evalúa en este paper — enunciados como límites actuales, no

como una hoja de ruta:

- **Prueba de persona única.** La puerta de unicidad de persona no tiene aún implementación en cadena. El muro de reputación (§4) la hace no crítica para la seguridad del consenso; sigue siendo relevante como cota del arranque en frío (§10) y para la unidad de ciudadanía base. Mecanismo objetivo y sustituible, en evaluación.
- **Liquidación y atestación.** No existe aún capa de liquidación de operaciones en cadena; la atestación se sostiene hoy en los cuatro frenos deterministas de §4.1. Ligar cada atestación a una operación liquidada es el endurecimiento previsto.
- **Servicio de infraestructura.** El servicio medido que la emisión paga hoy es el de comité (autoría y finalidad); el pago del alojamiento puro mediante pruebas auto-verificables está diseñado y no cableado.
- **Metadatos de transporte.** El contenido es ciego de extremo a extremo, pero la correlación en la capa de red (tiempos, tamaños, direcciones) es la frontera real de la privacidad. En construcción: capa de transporte de enrutado en cebolla. Hasta entonces, la afirmación honesta es privacidad de contenido, no de tráfico.
- **Separación de claves.** Autenticación y gasto comparten actualmente una clave con separación de dominio criptográfica (v1); la separación en subclaves está diseñada y pendiente.
- **Parámetros en calibración.** λ (barrido en curso con la batería adversarial completa), forma de la curva de emisión y calendario de prueba-de-servicio: los valores de lanzamiento son provisionales, y cada uno se fijará con evidencia publicada a través del proceso de actualización, nunca por decreto.
- **VDF del faro.** Propuesta, no implementada; el faro opera hoy con la guardia de peso reputable y respaldo determinista.
- **Difusión de votos de finalidad.** Una auditoría interna halló que la capa de difusión (gossip) retransmitía votos de finalidad sin verificar su firma sr25519 antes de propagarlos, lo que permitía a un par externo amplificar votos falsificados a través de la red (la validez seguía comprobándose en el conteo, de modo que no afectaba a la seguridad de la finalidad, solo abría un vector de amplificación). Corregido: la firma se verifica antes de retransmitir (§7.2), validado en hardware y vivo desde el binario de lanzamiento.
- **Guardia de concentración del comité.** El umbral de la guardia usa un piso absoluto de cuentas reputables, no una fracción del conjunto reputable; deja una ventana estrecha y autosanable durante el arranque, cuando el conjunto es pequeño. Residual reconocido, de baja severidad.
- **Clustering de entidad partida.** La métrica de concentración agrega máscaras conectadas por avales. Una entidad que parte su peso entre unas pocas máscaras y deliberadamente nunca avala entre ellas (tres máscaras a un tercio cada una, en el barrido adversarial) es invisible al clustering por aristas y puede quedar bajo la barra de la guardia. El clustering conductual (voto correlacionado, tiempos, topología) es la respuesta diseñada, aún no cableada; hasta entonces, el límite honesto es que la guardia mide concentración *conectada*, no concentración *coordinada*.
- **Recuperación automática tras una parada por concentración.** La parada reversible de §7.4 tiene dos mitades: la *congelación* bajo concentración excesiva y el *deshielo automático* una vez que la concentración se re-dispersa. La congelación — la mitad crítica para la seguridad — se sostiene desde el principio: la cadena se niega a finalizar bajo captura y prefiere detenerse, verificado en hardware. El deshielo regresó en el binario de lanzamiento

(un endurecimiento posterior del re-anclaje del comité, hecho determinista para cerrar un vector de división ajeno, rompió la ruta de recuperación del lado del nodo; se detectó en pruebas adversariales). La recuperación reparada pasó el hierro adversarial (34/34 escenarios) y se desplegó en el primer upgrade constitucional de la red (24 de julio de 2026, §10.6). El límite honesto que queda: el deshielo reparado aún no lo ha ejercitado una parada real por concentración en producción — solo el hierro.

- **Dependencias.** El framework de cadena subyacente arrastra avisos conocidos en dependencias transitivas, rastreados; el análisis de cuáles alcanzan la superficie del nodo vivo está en curso.

13. Conclusión

Todo sistema de consenso abierto responde a una pregunta: ¿qué tiene que adquirir un atacante para apoderarse de él? La prueba de trabajo responde *energía*; la prueba de participación, *capital*. Ambas respuestas son precios, y un precio se puede pagar — así que la seguridad de estas redes es, en el límite, el tamaño del presupuesto de un adversario. Este trabajo se propuso dar una respuesta distinta, una que no se compra de golpe: **tiempo comportándose bien, bajo reglas públicas, dentro del propio sistema**.

Confianza Tejida hace esa respuesta concreta. La autoridad — elegibilidad para el comité de finalidad, para los jurados de disputas, para la gobernanza — se deriva de una métrica de reputación computada de forma determinista desde evidencia objetiva en cadena: atestaciones bilaterales de tratos liquidados y un grafo de fianzas de aval, jamás puntuación subjetiva, jamás un custodio que pueda editar el libro. Esa métrica no es transferible y decae, de modo que una ventaja ni se vende ni se guarda. En torno a ella el paper compone una defensa en profundidad contra los dos fallos clásicos de los sistemas de reputación — el ataque Sybil y la colusión: agregación cóncava multidominio que veta perfiles estrechos, propagación con descuento por proximidad y amortiguación de comunidad, fianzas de aval con cascadas de responsabilidad acotadas y ventanas de prueba, sorteo sembrado por un faro de compromiso-revelación endurecido contra el *grinding*, y una guardia de atrincheramiento que congela la finalidad *de forma reversible* en cuanto el peso de una sola entidad cruza un techo. El efecto neto es invertir el problema del atacante: para capturar la finalidad hay que gastar primero tiempo real ganando reputación verificada dentro del propio orden que se pretende subvertir — y la reputación ganada así es, por construcción, la conducta que el orden quiere.

Nada de esto se afirma desde la prosa. Cada propiedad del paper se ata, a través de un libro de validación público (Apéndice A), a una ejecución que la exhibe, el hash del binario que la produjo y su salida cruda — reproducible por cualquiera: diecisiete de diecisiete escenarios adversariales resistidos por el motor de reputación, captura de jurado a dos saltos cuantificada en simulación y cerrada, finalidad sobreviviendo a la caída y reingreso de validadores a través de fronteras de época en dos entornos operados de forma independiente, y reputación bit-idéntica entre nodos y arquitecturas. El texto constitucional que gobierna el sistema pasó por el mismo proceso adversarial antes de sellarse. Declaramos también, en el mismo registro, lo que *no* está hecho aún (§12) — porque un estudio que esconde sus residuos gana menos confianza, no más.

La afirmación que creemos que la evidencia sostiene es estrecha y comprobable: **una red donde la autoridad no está en venta se puede construir con herramientas de hoy, y su seguridad no es un eslogan sino una cantidad medida.** La razón mayor para construirla

es más vieja que cualquiera de estos mecanismos — que la gente puede cooperar, mantener su palabra y dirimir sus disputas por un nombre ganado con obras y no por la fuerza ni por la riqueza. Confianza Tejida es un intento de darle a esa posibilidad antigua una máquina que no se pueda capturar en silencio. En qué se convierte es ya asunto de quienes la usen.

Referencias

[1] S. D. Kamvar, M. T. Schlosser, H. Garcia-Molina. *The EigenTrust Algorithm for Reputation Management in P2P Networks.* Proc. WWW 2003.

[2] J. R. Douceur. *The Sybil Attack.* Proc. IPTPS 2002.

[3] H. Yu, M. Kaminsky, P. B. Gibbons, A. Flaxman. *SybilGuard: Defending Against Sybil Attacks via Social Networks.* Proc. SIGCOMM 2006.

[4] D. Boneh, J. Bonneau, B. Bünz, B. Fisch. *Verifiable Delay Functions.* Proc. CRYPTO 2018.

[5] P. R. Milgrom, D. C. North, B. R. Weingast. *The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs.* Economics & Politics 2(1), 1990.

[6] L. Page, S. Brin, R. Motwani, T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web.* Stanford InfoLab, 1999.

[7] P. Resnick, K. Kuwabara, R. Zeckhauser, E. Friedman. *Reputation Systems.* Communications of the ACM 43(12), 2000.

[8] S. Nakamoto. *Bitcoin: A Peer-to-Peer Electronic Cash System.* 2008.

[9] A. Stewart, E. Kokoris-Kogias. *GRANDPA: A Byzantine Finality Gadget.* arXiv:2007.01560, 2020.

[10] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press, 1990.

[11] T. C. Schelling. *The Strategy of Conflict.* Harvard University Press, 1960.

Apéndice A: mapa de evidencia

Este apéndice sustituye la cifra por la lista: qué tests son, dónde viven y cómo re-ejecutarlos. La convención de entornos es la del libro de validación: **local** = estación de construcción; **hardware** = nodos reales en infraestructura separada (máquina y red distintas), la evidencia más fuerte.

A.1 La batería adversarial del motor de reputación (los 17/17 del Resumen)

Suite `tests/test_engine.py` del prototipo de reputación (`reputation-engine/` en el repositorio público, Python de biblioteca estándar, PRNG sembrado — reproducible por cualquiera). Los 17 tests, por nombre exacto y propiedad que fijan:

1. `test_sybil_has_no_power` — 200 identidades fabricadas (40 % de la red) capturan 0,00 % del comité (S1).
2. `test_independence_penalises_inbreeding` — el descuento por proximidad penaliza la señal endogámica (§5, defensa 1).

3. `test_damping_reduces_laundering` — la amortiguación corta 8,7× el lavado de reputación hacia un anillo denso (§5, defensa 2).
4. `test_community_reduces_scattered_ring` — la detección de comunidades corta el anillo disperso que evade la defensa local.
5. `test_adaptive_community_does_not_worsen` — fragmentar el anillo para evadir la etiqueta no aumenta lo lavado bajo defensa.
6. `test_adaptive_has_no_consensus_power` — a cualquier fragmentación, el poder de consenso de los títeres es ≈ 0 (§4.4).
7. `test_asymmetric_evades_damping_but_no_power` — el embudo dirigido (muchos avalan a uno) evade la amortiguación local pero no compra poder estructural.
8. `test_in_concentration_closes_funnel` — la señal de concentración de entrada corta el embudo ~70–80 % a toda diversificación.
9. `test_in_concentration_closes_cross_dim_funnel` — el embudo entre dominios también cae: el déficit de evidencia es por dominio (§4.4).
10. `test_whitewashing_loses_everything` — abandonar el seudónimo y reentrar limpio cuesta toda la reputación ganada (S3).
11. `test_cascade_slashing` — el fraude propaga pérdida por la cadena de avales con atenuación por salto; el no-avalista queda intacto (§6).
12. `test_sublinear_quota` — el cupo de avales vivos crece sublinealmente con la reputación (§5, defensa 3).
13. `test_graduation_frees_mentor_quota` — el avalado que acumula evidencia independiente se gradúa y libera el cupo del avalista (§4.2).
14. `test_temporal_decays_inactive` — la reputación del inactivo decae época a época; la del activo se sostiene (§4.5).
15. `test_edge_aging_freshness_premium` — los propios avales envejecen: renovar la confianza da prima frente a no renovarla.
16. `test_conservative_aggregate` — el agregado anti-especialista anula al perfil sin cobertura de dominios (§4.4).
17. `test_mass_conservation` — invariante interno del motor: la propagación conserva la masa de reputación.

Reproducción: `cd reputation-engine && python3 tests/test_engine.py` (suite) y `python3 run_all.py` (regenera el informe `RESULTS.md` con las cifras citadas). Última re-ejecución verificada: 2026-07-14, 17/17.

A.2 Trazabilidad del resto de resultados (§11)

Los números de *fila del libro* de abajo son las entradas del libro de validación público (`VALIDATION-LEDGER.md`), que registra más ejecuciones de las que se resumen aquí; cada resultado seleccionado se mapea a su propia fila, así que la numeración es la del libro y no es consecutiva.

Resultado (§11)	Fila del libro	Tests / método	Entorno
Batería del motor de reputación	1	17/17 (A.1)	local

Resultado (§11)	Fila del libro	Tests / método	Entorno
Simulador de consenso (seguridad/vivacidad bajo ataque, partición, pérdida)	2	11/11 testrig + 13/13 propiedad	local
Paridad Rust $\leftrightarrow$ prototipo (punto fijo determinista, `no_std`)	3	35/35 tests-oráculo	local
Gadget de finalidad multi-nodo (votos firmados, prueba en importación, rotación)	4	6/6 + 6/6 + 8/8; devnets 3/6/7 nodos	local + hardware
Decaimiento / anti-atrincheramiento	5	barrido sim + cuenta real 0,90/época exacto	sim + hardware
Justicia: jurado por sorteo con exclusión de interesados	6	8/8 + 14/14 unit; 9/9 en cadena	local + hardware
Recuento sin disparador privilegiado (época automática)	7	15/15; extrinsic ausente de la metadata	local + hardware
Parámetros de red de producción (cadencias, α, τ)	8	simulador re-ejecutado + metadata verificada	local + hardware
Barrido de captura en arranque frío	10	3 tandas adversariales	sim
Captura de jurado a dos saltos + penalización profundidad-2	11, 11b	8–10 $\times 10^3$ ensayos; 44 unit + 8/8 en cadena	sim + hardware
Faro compromiso-revelado anti-grinding (jurado + comité)	12	70 unit; 7/7 en cadena + conmutación verificada	local + hardware
Arranque con siembra declarada (sin halt, guardias sanas)	13	finalización sostenida; fundadores 20 % exactos	hardware

Resultado (§11)	Fila del libro	Tests / método	Entorno
Agregado cóncavo (golpe dirigido −47 % frente a −100 % con MIN)	16	20/20 unit + verificación en cadena	local + hardware
Topes de cascada y ventana de prueba (anti-kamikaze, anti-bomba)	17, 18	21/21, 22/22 unit + verificación en cadena	local + hardware
Descuadre validación/cobro de tasas (cuenta nueva impagable)	32	21/21 unit + gasto real de cuenta fresca	local + hardware
Carrera de frontera de época (comité obsoleto votando a través del borde)	33	reproducción capturada + KILL/REJOIN tras el fix	hardware

Los paquetes de evidencia en crudo (salidas de ejecución, scripts de validación) están inventariados en el propio libro de validación con sus hashes sha256, de modo que cualquier copia publicada es verificable contra lo que el libro compromete.

A.3 Ejecuciones reproducibles en hierro

Las siguientes ejecuciones ejercitan el *binario de lanzamiento* (sha256 `fa79dda9...6666a6`, la compilación que ahora corre en producción) sobre nodos reales, y cada una se reproduce byte a byte en cualquier máquina: las claves de nodo usadas son las conocidas claves deterministas de desarrollo (`0x00...01`...`0x00...06`), de modo que los PeerID y el génesis resultantes son idénticos en todas partes, y ninguna dirección o clave de nodo de producción aparece en el procedimiento. Cada entrada dice qué prueba, cómo ejecutarla y qué se observó.

E1 — El binario que corre es el auditado. *Prueba:* que no hubo sustitución entre auditoría y ejecución. *Ejecutar:* `sha256sum harlequin-node` y comparar con el publicado `fa79dda9...6666a6`, luego `harlequin-node --version`. *Observado:* el hash coincide; versión `0.1.0`, rol `AUTHORITY`.

E2 — La superficie de pallets coincide con la constitución ratificada. *Prueba:* que cada decisión constitucional está cableada en el runtime y, sobre todo, que ningún camino de llamada deja a la justicia tocar un saldo (Artículo II). *Ejecutar:* un solo nodo con `--consensus manual-seal-1000` (producción de bloques determinista sin comité); leer los metadatos del runtime con un cliente SCALE. *Observado:* la llave de actualización expone exactamente el diseño incorruptible (caducidad automática `LifeEnd`, renovación co-firmada por la red `PendingRenewal`/`apply_renewal` con la muerte geométrica retirada, quema irreversible, veto costoso, modo desastre, los cinco asientos, y un asiento `Custody` solo-llave con cero llamadas); la justicia expone únicamente `open_case`, `open_incapacity_case`, `cast_vote`, `close_case`,

y ninguna de ellas toma ni mueve un saldo — un veredicto culpable escribe una marca de estatus (`GuiltyMarks`), nunca moneda; participación expone los acumuladores de aguante (`ConsecutiveEras`).

E3 — La finalidad está viva y se niega a finalizar por debajo del quórum. *Prueba:* que el gadget finaliza bajo un comité sano y *se detiene en vez de bifurcar* cuando los votos caen bajo el quórum — seguridad sobre vivacidad. *Ejecutar:* seis nodos, `--consensus woven-trust-<ms> --validator --vote-as //Alice ... //Ferdie`, claves de nodo deterministas, nodo 1 como bootnode. *Observado:* con cinco votantes vivos (comité 6, quórum $\alpha = 5$) la cadena reportó *finalidad bajo quórum — esperando, sin finalizar* y **no** finalizó; al levantar el sexto nodo se produjo *finalizado #15 (comité 6, α 5/6)*, tras lo cual la finalidad avanzó y se sostuvo. Corriendo el mismo escenario a `woven-trust-2000` y a la cadencia de ceremonia `woven-trust-12000` se produjo el **hash finalizado idéntico** — el comportamiento del gadget es determinista e independiente de la cadencia de bloques.

E4 — Las guardias de la justicia rechazan casos inseguros en la puerta. *Prueba:* que el pallet de justicia se niega a abrir un caso sobre evidencia ajena o sin jurado válido — falla ruidosamente al abrir, jamás poniendo algo a cero en silencio en el veredicto. *Ejecutar:* en una cadena `--dev` con evidencia sembrada, enviar `open_case` como extrínseco firmado. *Observado:* la evidencia que no pertenece al demandado se rechaza al abrir con `ForeignEvidence`; evidencia válida con reputación génesis cero se rechaza con `EmptyJury` — el comportamiento correcto de arranque en frío, ya que los fundadores tienen reputación cero en el génesis (se gana por servicio) y no hay jurado elegible hasta que la reputación se acumula.

E5 — «La justicia toca estatus, nunca saldo», por tres vías independientes. *Prueba:* que el Artículo II es un límite estructural, no una política que un jurado pudiera doblar. *Las tres vías:* (i) *código* — el gancho de culpabilidad enruta solo al pallet de reputación, y un barrido del pallet de justicia encuentra cero referencias a token, moneda o saldo; (ii) *superficie* (E2) — ninguna llamada de justicia mueve un saldo; (iii) *en vivo* (E4) — las guardias hacen inalcanzable un camino de veredicto inseguro. La transición veredicto-culpable-a-`GuiltyMarks` está además cubierta por un test unitario determinista en el que la pérdida cae en la dimensión de reputación, nunca en el monedero.

Alcance honesto. El ciclo completo de veredicto-culpable-hasta-readmisión en un devnet vivo necesita primero que la reputación se acumule a lo largo de varias épocas (el arranque en frío de E4); hoy está cubierto de forma determinista en la capa de tests unitarios, y la ejecución multi-época en vivo es un siguiente paso declarado y reproducible — no un hueco en el mecanismo.

El código fuente, el libro de validación y los paquetes de evidencia son públicos y verificables de forma independiente.

Apéndice B: La expulsión sin confiscación: un mecanismo de sanción de solo-estatus

Este apéndice es autocontenido: sus referencias [1]–[34] se numeran de forma independiente de la bibliografía principal del paper.

B.1 Planteamiento del problema

Una sociedad construida sobre la asociación voluntaria, pseudónima y no coercitiva se enfrenta a una contradicción aparente. Si la pertenencia es genuinamente voluntaria y la propiedad es genuinamente inviolable, ¿cómo puede la sociedad responder a un agresor dentro de ella —un estafador, un ladrón, un vendedor de contrabando que perjudica a terceros no consintientes— sin (a) tolerar la agresión indefinidamente en nombre de la no coerción, o (b) recurrir a la confiscación, que exige exactamente el poder coercitivo y discrecional sobre los saldos ajenos que la sociedad existe para clausurar? Un mecanismo que resuelva esto debe superar dos exigencias independientes. Primero, debe funcionar realmente: la expulsión tiene que acarrear consecuencias reales, no un mero gesto simbólico que deje al agresor libre para seguir operando bajo la misma identidad. Segundo, el mecanismo mismo —quienquiera que ostente el poder de expulsar— debe resistir la captura; una herramienta construida para castigar una agresión consensuada es, por construcción, también una herramienta que una facción puede señalar contra una minoría impopular. El compromiso del Artículo II de Harlequin con saldos de moneda incondicionalmente inviolables convierte el primer cuerno del dilema en algo estructural y no negociable, lo cual obliga a la pregunta de diseño que aborda esta sección: ¿pueden separarse el estatus y la propiedad como palancas de sanción, y puede evitarse que la palanca que queda —el estatus— colapse de nuevo en un mecanismo capturable, cuasi-propietario, por derecho propio?

B.2 Fundamento doctrinal e histórico

La afirmación de que expulsión y confiscación son palancas separables no es un argumento aislado, sino una convergencia alcanzada de forma independiente por al menos nueve tradiciones que no se citan entre sí: el derecho consuetudinario medieval islandés, la democracia ateniense, el derecho comercial privado de las ferias medievales (lex mercatoria), la disciplina eclesial cuáquera, el derecho canónico católico, la teoría anarcocapitalista rothbardiana/hoppeana, la economía de la gobernanza privada de Stringham y Benson, la jurisprudencia analítica (la distinción hohfeldiana entre un poder sobre el estatus y un derecho de reclamación sobre la propiedad) y el marco de la meta-utopía de Nozick [1–11]. El ostracismo ateniense desterraba a un ciudadano hasta por diez años mediante voto de quórum, pero dejaba intactos sus bienes, rentas y familia [1]. El *fjörbaugsgarðr* islandés, la «proscripción menor», imponía un exilio de tres años y la retirada de la protección legal sin tocar la propiedad —reservando la confiscación efectiva para el *skóggangr*, la proscripción total, la pena más dura y más rara, exactamente el modelo contrario [2]. Tanto el repudio cuáquero como la excomunión católica retiran únicamente lo que la propia institución había conferido —voz en una reunión, cargo sacramental— nunca la propiedad secular que el miembro poseía de forma independiente [4, 5]. La *Meidung* amish se aplica solo a quienes aceptaron libremente el bautismo adulto, un pacto explícito ex ante [3]. El diseño contemporáneo de la gobernanza cripto reproduce el mismo patrón mediante una ingeniería convergente y no coordinada: el mecanismo guildKick de MolochDAO no quema las participaciones de un miembro expulsado, sino que las convierte en tokens LOOT sin derecho a voto que siguen siendo redimibles a prorrata contra la tesorería mediante ragequit, de modo que el derecho económico sobrevive aunque se revoque la voz [12].

La teoría hoppeana de la comunidad de pacto (covenant community) ofrece el anclaje doctrinal más claro: la expulsión opera dentro de un marco propietario contractual y es «no violenta, pero coherente con los derechos de propiedad», una lectura confirmada de forma independiente por varias fuentes y aclarada por el propio Hoppe en una entrevista posterior en el sentido de

que se refiere a la distancia física lograda por medios no violentos —ostracismo, ridiculización, denuncia colectiva, nunca la fuerza— [6]. La derivación rothbardiana del castigo a partir de la autodefensa de la víctima no excluye, en una lectura atenta, el boicot colectivo o el ostracismo no agonístico, que Rothbard trata como no invasivos [7]. El caso negativo es The DAO (2016): cuando la comunidad de Ethereum empleó el poder colectivo para revertir un robo moviendo saldos, el resultado fue una escisión permanente de la cadena, con críticos que sostuvieron que la propia «corrección» fue la violación de propiedad más grave de todo el episodio [13]. La lección es arquitectónica, no moral: en cuanto un jurado puede mover un saldo por una razón legítima, no existe una línea nítida que separe ese primer uso justificado de un uso posterior corrompido.

Este fundamento no elimina una tensión doctrinal genuina. La teoría del estoppel de Kinsella sostiene que restringir el remedio de la víctima a la sola expulsión la revictimiza —podría legítimamente optar en su lugar por la confiscación proporcional—, lo cual contradice directamente un techo del tipo «el saldo es intocable» [6, 8]. Harlequin resuelve esto no invocando un falso consenso libertario, sino convirtiendo el propio techo en una cuestión de consentimiento ex ante: el límite del remedio se acepta en el momento de crear una máscara, de modo que la futura víctima también lo ha consentido de antemano, y no se le está quitando nada que poseyera de forma incondicional. El precedente de The DAO (§ anterior) aporta el argumento estructural de por qué este techo debe ser un límite constitucional duro y no una discrecionalidad del jurado caso por caso [13].

B.3 Fundamento empírico

Tres resultados producidos de forma independiente cierran la cuestión de la plutocracia con una convergencia inusual: la medición sobre 21 sistemas de gobernanza on-chain muestra coeficientes de Gini del poder de voto «cercanos a 1,0» (por ejemplo, 0,992 para los titulares de tokens de Uniswap y 0,996 para los de Compound, Tabla 2), una mayoría del poder de voto controlada por menos de 10 participantes en 17 de los 21 sistemas, y coeficientes de Nakamoto de tres o menos en la mitad de la muestra [14]; una revisión de diseños de gobernanza blockchain no encuentra ninguno que satisfaga simultáneamente la mayoría de los criterios deseables [15]; y un resultado formal de imposibilidad de 2026 demuestra que *todo* esquema de voto cóncavo (cuadrático, logarítmico o de otro tipo) colapsa en resultados lineales y plutocráticos bajo una división sybil sin restricciones, con una amplificación medida de 1,172×–4,039× en cinco DAO reales [16]. Este último resultado es la advertencia estructural que sostiene todo el mecanismo expuesto más abajo: la reputación es exactamente tan capturable como el capital en cuanto las identidades pueden multiplicarse libremente. El ataque de gobernanza mediante flash-loan a Beanstalk en 2022 (182 millones de dólares) demostró que la compra instantánea de votos funciona exactamente como fue diseñada [17]. La Propuesta 289 de Compound muestra la misma dinámica mediante una acumulación paciente en lugar de un flash-loan: la propuesta, aprobada por 682.191 votos frente a 633.636 en julio de 2024, dirigió aproximadamente 24 millones de dólares (499.000 COMP) de la tesorería a una bóveda vinculada a un gran titular, empleando un poder de voto construido a partir de aproximadamente 12 millones de dólares en tokens delegados y no la cifra completa de 24 millones —una captura real pero más modesta de lo que sugirieron los titulares [18]. Un relato de primera mano en un foro de Polkadot describe una única cartera de 9 millones de DOT que hizo virar un referéndum de la aprobación mayoritaria a un rechazo del 96% sin debate alguno; esto se recoge aquí como una anécdota comunitaria no verificada, no como un hecho on-chain auditado [19]. El análisis académico de Ferreira de

2026, publicado en *Corporate Governance: An International Review* de Wiley por un profesor de finanzas de la LSE (no, como a veces se cita, un paper de Oxford —el Oxford Business Law Blog se limitó a resumirlo—), llega a la misma conclusión desde el lado de la literatura: la formalidad procedimental no confiere resistencia a la captura [20].

Los datos de mercados de la darknet aportan una vertiente empírica distinta: el aval como puerta de entrada (vouching-as-entry-gate), puesto a prueba en el caso mejor documentado (Darkode), fracasó —se aceptó el 94,5% de las solicitudes, y los propios administradores de la plataforma generaron el 38% del reclutamiento captando novatos, lo que colapsó la selectividad que el mecanismo pretendía ofrecer [31]. En cambio, la autorregulación comunitaria en torno a un consenso moral casi universal —la prohibición del material de abuso sexual infantil (CSAM)— se mantuvo de forma consistente en distintos mercados sin coordinación central, mientras que la supresión de un bien de consumo (los opioides) simplemente desplazó los anuncios en lugar de reducirlos, y la supresión específica del fentanilo combinada con presión externa sí mostró un efecto sostenido [33, 34]. Esta asimetría es relevante para la capa de mercado: los filtros de entrada no suprimen de forma fiable a los malos actores, pero una prohibición dura y de alcance estrecho sobre la agresión contra terceros no consintientes sí puede hacerlo.

B.4 El mecanismo

Harlequin compone estos hallazgos en una arquitectura de sanción de solo-estatus, estratificada por capas. **M2, la disociación individual**, es el suelo inalienable: cualquier máscara puede en todo momento dejar de tratar con cualquier otra. No tiene fuerza por sí sola, pero no puede ser capturada, porque no existe una palanca central de la que apoderarse. **M5, jurados por sorteo con sanciones de solo-estatus**, forma la capa de proceso: los jurados se seleccionan al azar y rotan —la defensa antisoborno redescubierta de forma independiente por Atenas y reimplementada por Kleros [1, 30]—, los veredictos son públicos, y la única sanción disponible es la reputación, el vínculo de aval (vouch-edge) o la pérdida de elegibilidad, nunca un saldo; este techo está escrito en el Artículo II como un límite estructural, no como discrecionalidad del jurado, precisamente a causa de la tensión de Kinsella y el precedente de The DAO (§2) [8, 13]. **M3, una escalera graduada** (advertencia → penalización reputacional → suspensión de privilegios → expulsión temporal con vía de readmisión → expulsión total) sigue el principio de diseño nº 5 de Ostrom, las sanciones graduadas —distinto de su principio nº 3, los mecanismos de elección colectiva, que la legitimidad compositiva del jurado satisface por separado [21]. La readmisión se tasa en aguante (un requisito de racha continuada, streak), lo que hace que el coste de la reentrada sea tiempo y no capital ni reputación, la única divisa que el resultado de colapso sybil no puede multiplicar barato [16]. **M4, la revocación de aval**, opera como *consecuencia* del veredicto, nunca como puerta de entrada —el resultado de Darkode descarta específicamente el aval-como-puerta, no el aval-como-consecuencia [31]. **M6, las fianzas específicas de privilegio** (para jurados, creadores de mercado, operadores de nodo) pueden perderse, pero solo ese depósito, separado arquitectónicamente de la cartera base, lo que evita la deriva al estilo comunidad de vecinos (HOA) por la que una fianza de estatus acaba convirtiéndose gradualmente en garantía sobre el propio saldo. Por último, la brecha sin resolver de Hoppe —una parte boicoteada que queda formalmente propietaria pero prácticamente indefensa— se cierra de forma explícita y por escrito: la protección de la propiedad es universal e independiente del estatus, de modo que un jurado debe juzgar una denuncia de robo presentada por una máscara expulsada exactamente igual que la de cualquier otra; y la cláusula de sanción se limita a los actos, nunca a la creencia

o la afiliación, lo cual clausura el riesgo de captura ideológica visible incluso en los escritos posteriores del propio Hoppe.

Una última restricción de encuadre proviene de la propia constitución: la justicia de Harlequin no impone castigo; hace saber la verdad. El jurado, por tanto, no decreta la expulsión: certifica un acto probado y publica el veredicto. Toda consecuencia que se derive (el registro reputacional, la pérdida de elegibilidad para funciones comunes, la pérdida de una fianza específica de privilegio, la retirada de cobijo y herramientas comunes) es o bien un efecto que el miembro consintió ex ante al crear la máscara, o bien el agregado de decisiones libres de otros miembros de disociarse. La sanción es emergente, no impuesta: el jurado es el órgano epistémico; la disociación individual es el brazo de ejecución. De aquí se siguen dos corolarios. Primero, la expulsión nunca bloquea el intercambio entre pares a nivel de protocolo —la orden ni privilegia ni excluye a ningún comerciante; retira privilegios comunes y deja constancia en el registro público, y cada contraparte decide por sí misma. Segundo, no se requiere ninguna enmienda constitucional: el mecanismo vive por entero en la especificación mutable, subordinada a los principios sellados de los que se deriva.

B.5 Limitaciones honestas

Deben declararse tres limitaciones en lugar de disimularlas. Primera, toda la arquitectura descansa en el coste anti-sybil: si las máscaras se vuelven baratas de multiplicar, el resultado formal de imposibilidad se aplica directamente y el sistema colapsa en una plutocracia de sybils, por muy cuidadosamente que se hayan separado el estatus y la propiedad [16]. Segunda, el sorteo tiene un historial de despliegue real en disputas acotadas (Kleros), pero ninguno a la escala de una actualización constitucional; extrapolar sus propiedades anti-captura al sistema de jurados de Harlequin es una hipótesis declarada, no una demostrada [30]. Tercera, el debate entre el libertarianismo «delgado» (thin) y «grueso» (thick) sobre hasta dónde puede llegar un orden voluntario en la vigilancia del «vicio» consensuado frente a la no-agresión estrictamente definida no ha sido abordado en esta ronda de investigación y sigue siendo una cuestión abierta para trabajo futuro.

B-Referencias

1. Wikipedia, “Ostracism”; World History Encyclopedia, “Ostracism”; “Pruning of the People,” *Philosophies* 8(5):81 (MDPI, peer-reviewed).
2. Hurstwic, “Viking-age Laws and Legal Procedures” (hurstwic.org); *Norse Mythology for Smart People*, “Outlawry in the Viking Age.”
3. Amish Studies, Elizabethtown College, “Church Discipline” (groups.etown.edu/amishstudies).
4. Quaker.org, “Our Understanding of Disownment.”
5. Wikipedia, “Excommunication in the Catholic Church.”
6. Stephan Kinsella, “Hoppe on Covenant Communities and Advocates of Alternative Lifestyles” (stephankinsella.com, 2010); aclaración de Hoppe en su entrevista con Michael Malice («YOUR WELCOME», Ep. 018, septiembre de 2018) de que «físicamente removido» significa *distancia* física lograda por medios no violentos — ostracismo, burla, queja colectiva — nunca fuerza.

7. Murray Rothbard, *The Ethics of Liberty*, Ch. "The Ethics of Boycotts"; Hoppe, introduction to *The Ethics of Liberty* (mises.org).
8. Stephan Kinsella, "A Libertarian Theory of Punishment and Rights" (1997/2021 repost, stephankinsella.com); *Libertarian Papers* 1:17.
9. Murray Rothbard, *The Ethics of Liberty*, Ch. 13 "Punishment and Proportionality" ("two teeth for a tooth"); further developed by Walter Block.
10. Robert Nozick, *Anarchy, State, and Utopia* (1974), meta-utopia framework.
11. Stanford Encyclopedia of Philosophy, "Freedom of Association" (Hohfeldian analysis).
12. MolochDAO documentation, guildKick/ragequit mechanism (LOOT conversion, not burn).
13. Ethereum Classic Cooperative blog, "Code is Law and the Quest for Justice"; CoinDesk, "Ethereum Executes Blockchain Hard Fork to Return DAO Funds" (2016).
14. arXiv:2302.12125, governance concentration across 21 on-chain systems.
15. Kiayias & Lazos, "SoK: Blockchain Governance," arXiv:2201.07188 — survey finding no analyzed system satisfies a majority of desirable governance properties.
16. Bennett, Vander Vos, Le, Belenkiy, "Concave is the New Linear: The Impossibility of Anti-Plutocratic DAO Governance," arXiv:2605.18990 (AFT 2026).
17. Bloomberg; CoinDesk; Cointelegraph; The Register (April 17–18, 2022), "DeFi Project Beanstalk Loses $182 Million in Flash Loan Attack."
18. Compound governance forum, Proposal 289 (passed 682,191–633,636, July 28, 2024).
19. Polkadot governance forum, community report on referendum swing by a single 9M-DOT wallet (unverified, first-hand anecdote); Messari, "Polkadot OpenGov Report."
20. Daniel Ferreira, "The Myths of Blockchain Governance," *Corporate Governance: An International Review* (Wiley, 2026); London School of Economics.
21. Elinor Ostrom, *Governing the Commons* (1990), design principles #3 (collective-choice arrangements) and #5 (graduated sanctions).
22. OpenZeppelin, Governor/Timelock documentation (GovernorTimelockCompound.sol).
23. Tezos documentation, self-amendment process (docs.tezos.com).
24. MakerDAO documentation, Governance Security Module (GSM) pause delay (docs.makerdao.com).
25. Vitalik Buterin, "Credible Neutrality as a Guiding Principle" (2020); governance-minimization writings (vitalik.eth.limo).
26. Nick Szabo, "Trusted Third Parties Are Security Holes" (2001), nakamotoinstitute.org.
27. Arbitrum governance documentation, Security Council (2 staggered cohorts of 6 members, 9-of-12 approval threshold).
28. Wikipedia, "Liberum veto"; Encyclopaedia Britannica, "Polish–Lithuanian Commonwealth."
29. BIP-148 text; uasf.org (Bitcoin UASF, 2017).
30. Kleros.io; Stanford Law School, "Kleros: A Socio-Legal Case Study of Decentralized Justice & Blockchain Arbitration."
31. ResearchGate, "Darkode Recruitment Patterns."
32. Morselli, Décary-Hétu, Paquet-Clouston & Aldridge, "Conflict Management in Illicit Drug Cryptomarkets," *International Criminal Justice Review* 27(4), 237–254 (SAGE, 2017). DOI 10.1177/1057567717709498.
33. Australian Institute of Criminology, "Impact of Darknet Market Seizures on Opioid Availability" (rr18).
34. ResearchGate, "Crime Displacement in Digital Drug Markets."