Conditional Disclosure as a Coordination Device

By Matthew Cashman*

July 16, 2026 (latest version here)

People with controversial views may stay silent rather than speak out alone, though in a large enough group they could speak safely together. A social assurance contract makes disclosure conditional: signed commitments of support stay private until enough people have signed for it to be safe to publish. In contrast, a survey can measure hidden support but cannot protect anyone who acts on it. I derive the contract's exact advantage over speaking openly - and when that advantage disappears: it can turn a few people who would speak alone into a larger coalition.
JEL: C72, D82, D83
Keywords: Assurance contracts, disclosure design, equilibrium selection, global games, coordination, failure privacy

Public silence need not reveal agreement. It may instead reflect a coordination problem: agents who would join a sufficiently large public coalition do not want to be identified in a small one. This problem can persist even when agents know that others privately share their view. It is distinct from pluralistic ignorance, which concerns mistaken beliefs about others' private views (Prentice and Miller, 1993, 1996). Preference falsification supplies a related account of why private views and public statements may diverge (Kuran, 1997). The relevant institutional question is then not only how to measure private support, but how to make public identification contingent on enough others being identified at the same time.

I study a *social assurance contract*. Individuals privately commit to endorse a statement, and the mechanism releases the complete named coalition only if a preset threshold is reached. Otherwise it releases no names. Unlike a provision-point contract, the object held in escrow is identity disclosure rather than a financial contribution. At a university, the contract might read:

> We, the undersigned, believe [controversial belief] and think the University should take the following steps [1, 2, 3, …]. Signatures on this letter will become public *only when at least [200] faculty have signed.* Until then, no one can see who has signed. If the threshold is not met, the letter and its signatures are destroyed.

The contract differs from both anonymous measurement and open expression.

* Cashman: MIT Sloan School of Management, cashman@mit.edu.

A survey or secret ballot can reveal an aggregate without naming respondents. That is sufficient when an institution can act on the aggregate. It does not produce a coalition whose members are prepared to be publicly associated with the statement. An open letter does produce such a coalition, but it names participants even when final participation is low. The contract instead stores commitments and conditions the entire list's release on final coalition size. The paper studies this final-outcome risk: exposure after a low-participation or failed-coordination outcome.

Social assurance contracts combine the threshold logic of step-level public goods with identity disclosure. Provision-point mechanisms refund contributions when a funding threshold fails (Bagnoli and Lipman, 1989); dominant assurance contracts add refund bonuses (Tabarrok, 1998); and Strausz (2017) studies provision-point design in crowdfunding. Here the item held conditionally is not money but the complete signer list. The closest institutional antecedent is the information escrow of Ayres and Unkovic (2012), which stores allegations and releases them under specified corroboration conditions. Braghieri, Bursztyn and Fasnacht (2026) also study threshold identity disclosure, but allow each voter to choose when her own vote–identity pair becomes visible. The present mechanism instead commits to release one named coalition as a unit.

The formal comparison isolates material retaliation: the cost an employer, professional group, or other audience can impose on a named participant. Both institutions have the same committed seed, one type among the agents outside that seed, private signals about a common state, and a common audience index and exposure technology. I call the agents outside the seed *marginal agents* because their participation can change with the institution. Punishment attaches to membership in a released list and is diluted by coalition size, consistent with evidence that reducing retaliation costs raises employee complaints (Heese and Pérez-Cavazos, 2021). Holding the audience environment fixed isolates the disclosure rule and contract-specific signing friction; the comparison does not assume that the two audiences begin with different beliefs. Social-image models offer a related reason why visible isolation is costly (Bernheim, 1994; Braghieri, 2024); the fixed-index benchmark separates that inference channel from the mechanism's payoff effect. Garbling protects individual whistleblowers in Chassang and Padró i Miquel (2019); conditional disclosure instead delays identification until a coalition forms.

The seed is a committed or protected core that participates under either institution and can clear the publication threshold in sufficiently favorable states. The seed creates the upper dominance region: when conditions are favorable enough, joining is worthwhile even if no marginal agent expects help from other marginal agents. Conditional disclosure is therefore an amplifier that converts an existing core into a larger coalition. Seed credibility and threshold reachability are central design variables.

The main result applies the same small-private-noise selection to open expres-

sion and the contract. To understand the resulting comparison, imagine varying marginal participation from none to all. A *rank* records a point along this conceptual path, or equivalently the fraction of marginal agents who participate. It is not an order of arrival in a dynamic campaign. The *rank-indifference value* is the expressive benefit at which participation pays zero on average across those ranks. Open expression and conditional disclosure produce different rank-indifference values because the contract keeps the signer private at every rank below the threshold.

Their difference is an *accounting gap*: the exposure attached to ranks between the seed and threshold, which private failure removes, net of signing friction. The *selected gap* adds a second requirement: a coalition that has just reached the threshold must be large enough to make public identification worthwhile for the marginal signer. It is the smaller of the accounting gap and this recruitment-coverage margin. Measured in units of the value a person places on public expression, it tells us how far the contract lowers the private value required for participation. The theorem gives three exact conditions for a positive gap. The exposure tail must exceed the friction haircut, meaning the participation advantage lost to signing costs; the threshold must be protective and reachable; and the open-expression cutoff must strictly exceed the value needed for protection at the threshold.

The selection argument uses the monotone global-games program (Carlsson and van Damme, 1993; Morris and Shin, 2003; Frankel, Morris and Pauzner, 2003): as private noise becomes small, iterated deletion of dominated actions selects an essentially unique cutoff from a complete-information game that had several equilibria. I apply exactly the same procedure to open expression and conditional disclosure.

Global-games applications study bank-run selection (Goldstein and Pauzner, 2005); principals can alter coordination through targeted subsidies (Sákovics and Steiner, 2012) or discriminatory success-contingent rewards (Winter, 2004; Halac, Kremer and Winter, 2020); and information manipulation changes signals and beliefs (Edmond, 2013). Conditional disclosure uses a different channel: it changes whether a failed participant is identified. The exposure-tail decomposition quantifies that channel. The argument requires strategic complementarity; turnout can instead be substitutable, as in the Hong Kong evidence of Davide Cantoni, David Y Yang, Noam Yuchtman and Y Jane Zhang (2019).

Anonymous measurement provides the key comparison. Randomized response can measure sensitive attitudes without naming respondents (Warner, 1965), and a secret ballot is sufficient when an institution maps an aggregate into action. Information about social norms can also change behavior (Bursztyn, González and Yanagizawa-Drott, 2020). But measurement does not insure a participant against identification after a small coalition forms. At a fixed state and audience index where either silence or full expression can sustain itself, the open-expression game retains both seed-only and full-expression equilibria. The selected comparison

holds public information fixed and measures the payoff change created by private failure. Whether publication creates common knowledge depends on who observes it and on the communication structure (Chwe, 1999); the fixed-index comparison isolates the exposure channel from those belief effects.

The design results identify the institutional source of the benefit. Consider rules that make a yes-or-no release decision based only on coalition size and then publish either the full list or nothing. Within this class, a threshold is the most permissive rule that never exposes an unprotected signer. Threshold rules also preserve strategic complementarity among the people the mechanism seeks to recruit. Failure privacy must be credible: a trusted independent custodian or a protocol that releases names only after verifiable threshold clearance must implement it. At the sharp opposite endpoint, releasing every failed list with the same exposure and expressive value as open expression reproduces open expression plus contract-specific signing friction.

The contribution is the exposure-tail decomposition and its design implications. It gives an exact formula for the value of failed-list privacy in the seeded one-type benchmark and separates that value from signing friction. The Supplemental Appendix studies a literal finite group in which agents begin with a prior belief about the state, receive private signals, and understand that one person's signature can change whether the threshold is met. It shows how signing friction and the chance of publication shape participation around a core that can clear the threshold in favorable states. Potential applications include collective whistleblowing, named organizing committees, and public letters, provided that they have a protected or committed core, a reachable threshold, and exposure that falls as the final coalition grows.

The companion paper allows potential signers to differ in their protection and other characteristics. It studies composition-dependent safety, a necessity theorem for zero-risk coalition assembly, and the risk and robust frontiers created by positive risk tolerance and leakage (Cashman, 2026). The present paper supplies the homogeneous foundation: marginal agents have the same payoff type but receive different private information, and the comparison with open expression holds the audience technology fixed.

Section I defines the institution and the seeded benchmark. Section II proves selection, derives the exposure-tail gap, and gives the design results. Section III develops applications and empirical implications, and Section IV concludes. The Appendix contains proofs. The separate Supplemental Appendix gives the finite committed-core benchmark, clarifies failure privacy and exchangeability, and develops an implementation architecture using existing tools.

I. Conditional Public Coalition Formation

The literal mechanism has a finite number of participants. The article approximates a large population by a continuum, so no single marginal agent changes coalition size but the group of marginal agents does. This approximation pro-

duces a transparent exact comparison. The Supplemental Appendix returns to the finite institution, where one signature can be pivotal. The analytical benchmark has one marginal type and holds the audience's exposure environment fixed across institutions. These choices isolate the direct safety-in-numbers channel: how a larger final coalition changes the cost of being publicly identified.

A. The Finite Institution

Let $\mathcal{I}_N = \{1, \ldots, N\}$ be the eligible set, let K_N be an integer publication threshold, and let each eligible agent choose $a_i \in \{0, 1\}$. Define the number and identities of signers by

$$Q_N = \sum_{i=1}^{N} a_i, \qquad \mathcal{C}_N = \{i : a_i = 1\}. \tag{1}$$

Here Q_N is the number of signers, while $\mathcal{C}_N$ records exactly who they are—the roster, in the companion paper's terminology. The binary release rule publishes the complete list $\mathcal{C}_N$ if and only if $Q_N \geq K_N$. If $Q_N < K_N$, it publishes neither names nor a count. Success therefore creates a named coalition, while failure reveals no signer list.

Under ideal failure privacy, the public observation is

$$o_N = \begin{cases} \mathcal{C}_N, & Q_N \geq K_N, \\ \varnothing, & Q_N < K_N, \end{cases} \tag{2}$$

where $\varnothing$ means that the exposure-cost audience observes no names, count, or failure announcement. The ideal benchmark also keeps the campaign attempt itself from that audience; the Supplemental Appendix treats failed-list privacy and attempt observability separately.

A committed or protected core $B_N(\theta)$ signs independently of the remaining agents' decisions and supplies the base that the contract amplifies. The Supplemental Appendix shows that all marginal agents can use the same signal cutoff: each signs when her private signal is above one common boundary. Because each person contributes one signature, a single decision may determine whether the list is published. The result shows how signing friction moves that boundary and changes the probability of publication. The comparison between the contract and open expression instead comes from the continuum model below.

B. The Seeded Analytical Benchmark

The state $\theta \in \mathbb{R}$ summarizes underlying conditions that make a public coalition more or less viable, such as latent support or the availability of protected participants. The committed seed has mass $b(\theta)$ and participates under every institution.

The total mass that could participate is $Y(\theta)$, so $Y(\theta) - b(\theta)$ is the pool whose decision the mechanism may change. The normalization $0 \leq b(\theta) < Y(\theta) \leq 1$ expresses all coalition sizes as shares of the relevant population. Both b and Y are continuously differentiable and weakly increasing: more favorable states never shrink either the seed or the available pool.

The remaining agents form an atomless population $\mathcal{I} = [0, 1]$ of normalized mass one. There is one marginal type: all marginal agents have the same payoff parameters. Their expressive benefit e is the value of being publicly associated with the statement, and $\alpha \geq 0$ measures sensitivity to exposure. If a fraction $w \in [0, 1]$ of the marginal population participates, named coalition mass is

$$y(w; \theta) = b(\theta) + \big(Y(\theta) - b(\theta)\big)w. \tag{3}$$

When $w = 0$, the coalition consists only of the seed and has mass $b(\theta)$. When $w = 1$, every available participant joins and the coalition has mass $Y(\theta)$. Values between zero and one trace a conceptual participation path between those endpoints, not a sequence in time. The notation Y is reserved for participating capacity; y denotes realized coalition mass, and $\mathcal{C}$ denotes the corresponding named list.

The audience index μ summarizes the public environment facing a signer. It may incorporate prior public information or a survey posterior, such as how sympathetic an employer or professional audience is likely to be. It is a compact description of that environment, not another choice made by the signer. The function $\ell(y, \mu)$ gives the baseline exposure cost of appearing in a coalition of mass y in that environment. A signer with sensitivity α therefore pays $\alpha\ell(y, \mu)$. The function is bounded, continuous, nonnegative, and strictly decreasing in y. This last property is safety in numbers: adding names to the released list lowers the cost borne by each signer. A more favorable audience index also lowers exposure. The comparison holds μ fixed across institutions, so coalition size does not itself change what the audience believes. The Supplemental Appendix uses $\mu(y)$ to describe that separate possibility, but the theorem here does not cover it. In the material-retaliation interpretation, a larger list dilutes the punishment that can be directed at any one member.

Marginal agents share the type (e, α) but have different information. Agent i observes a private signal

$$m_i = \theta + \sigma\varepsilon_i. \tag{4}$$

The noise term ε_i creates idiosyncratic disagreement, and $\sigma > 0$ controls how noisy the signals are. A higher m_i is evidence of a more favorable state. Conditional on θ, the shocks are independent across agents and identically distributed with a continuous, strictly increasing distribution F and positive density $f = F'$. The signal family satisfies the monotone likelihood-ratio property: higher signals make

TABLE 1—MECHANISM CLASSIFICATION

Mechanism	Public output	Identity disclosure	Conditional holding
Open expression	Named endorsers	Every endorser is named	No
Anonymous measurement	Aggregate statistic	No respondent is named	No
Social assurance contract	Named signer list	Complete list on success; no failed list	Yes

higher states relatively more likely.

A pure strategy is a jointly measurable map $a : \mathcal{I} \times \mathbb{R} \to \{0, 1\}$, where $a_i(m) = 1$ means that agent i joins after observing m. Strategy profiles are ordered pointwise. The benchmark maintains *conditional exact aggregation*, the continuum analogue of a law of large numbers: conditional on the state, the realized fraction of marginal participants equals the fraction predicted by the signal distribution,

$$w_a(\theta) = \int_{\mathcal{I}} \int_{\mathbb{R}} a_i(\theta + \sigma\varepsilon)\, dF(\varepsilon)\, di. \tag{5}$$

Under a symmetric cutoff strategy $a_i(m) = \mathbf{1}\{m \geq c\}$, exact aggregation becomes

$$w_c(\theta) = 1 - F\left(\frac{c - \theta}{\sigma}\right). \tag{6}$$

The expression is the probability that a signal drawn in state θ exceeds the cutoff. It is also the realized participation fraction in the continuum. This exact-aggregation convention is the idealization used for selection; the Supplemental Appendix treats a literal finite population. Public information, including the fixed index μ, is held fixed while private noise σ tends to zero.

The exchangeable-identities restriction says that two coalitions of the same size create the same exposure cost, regardless of which particular people fill the positions. Coalition mass is therefore sufficient for calculating exposure. Identities still matter because publication reveals exactly who joined and hence who bears that exposure. When seniority, subgroup, or network position changes protection, coalition composition becomes relevant; the companion paper studies that heterogeneous case.

C. *Benchmark Mechanisms*

Table 1 separates the public outputs of the three benchmark technologies. Anonymous measurement may change the common audience index before either participation game, but it does not itself store conditional commitments or release a named coalition.

Open expression names every marginal participant. A social assurance contract privately records commitments, publishes the complete named list when its mass reaches τ, and otherwise keeps the list private. Signing incurs friction $k \geq 0$. Credible custody and threshold release are maintained features of the mechanism rather than continuously varying primitives.

II. Mechanism Comparison

The comparison holds the seed, audience index, exposure technology, private signals, and selection criterion common across institutions. I refer below to open expression and the contract as the two *arms* of the comparison. Keeping everything else common measures the direct payoff effect of making identification contingent on a sufficiently large named coalition. A survey can affect behavior by changing the common index μ or agents' beliefs about the state; conditional disclosure changes what happens to a participant after coordination fails.

A. Selection, Not Information

The complete-information benchmark first identifies the coordination problem that the private-signal model will resolve. Even with the state and audience environment commonly observed, behavior can remain uncertain because each person's best action depends on how many others participate.

THEOREM 1 (Open-expression multiplicity): *Fix a state θ and any audience belief μ. In the simultaneous open-expression game among marginal agents with a participating seed of mass $b(\theta)$, the gain from expression rises with the expressing mass. If the marginal type satisfies*

$$\alpha\,\ell\big(Y(\theta),\mu\big) \;\leq\; e \;<\; \alpha\,\ell\big(b(\theta),\mu\big), \tag{7}$$

then both the seed-only profile, in which no marginal agent expresses, and the full-expression profile are equilibria. The band is nondegenerate whenever $\alpha > 0$, $b(\theta) < Y(\theta)$, and ℓ is strictly decreasing in y. An informational intervention can move μ and move a type into or out of this band. At every fixed state and index inside it, the open-expression game remains multiple.

The two equilibria are sustained by different self-confirming expectations. If every marginal agent stays silent, a person who deviates is exposed beside only the seed and prefers not to deviate. If every marginal agent speaks, the large coalition lowers exposure enough that speaking is worthwhile. The state and audience are the same in both cases; what differs is what each person expects the others to do.

I select between such outcomes by adding the small private-signal differences described above. The selection is applied to two participation payoffs. Under open

expression, a participant always receives expressive benefit e and always bears exposure. Under the contract, both the expressive benefit and exposure arrive only when the coalition reaches threshold τ, and the signer pays the contract friction k in either event:

$$u_{open}(\theta, w) = e - \alpha\,\ell\big(y(w;\theta), \mu\big),$$
$$u_{SAC}(\theta, w) = \mathbf{1}\{y(w;\theta) \geq \tau\}\big(e - \alpha\,\ell(y(w;\theta), \mu)\big) - k. \tag{8}$$

The baseline normalizes open expression to zero private action friction and assigns the escrow-specific friction k only to the contract. This asymmetry captures the costs of verification, custody, and platform trust rather than declaring public expression costless in every application.

The selection argument needs firm starting points. At sufficiently low signals, staying out must be best even if everyone else joins. At sufficiently high signals, joining must be best even if no other marginal agent joins. Payoffs must change regularly between those extremes. The first six clauses below give the economic content of these requirements. They ensure that failure is possible in bad states, the seed can produce success in good states, a coalition at the threshold protects the recruited type, and signing a contract that certainly fails is strictly costly.

ASSUMPTION 1 (Seeded selection domain): *There are states $\theta_L < \theta_H$ and an interior evaluation state $\theta \in (\theta_L, \theta_H)$ with $b(\theta) < \tau < Y(\theta)$ such that:*

1) τ-protection: $e \geq \alpha\,\ell(\tau, \mu)$ — *a just-successful coalition protects the marginal type. The threshold must protect whom it recruits.*

2) Low-state exposure: $e < \alpha\,\ell(Y(\theta_L), \mu)$ — *at low states even full participation does not protect, so the type genuinely faces a coordination problem.*

3) Genuine failure risk: $\tau > Y(\theta_L)$ — *the threshold is unreachable at low states.*

4) Seed clearance: $b(\theta_H) \geq \tau$ — *at high states the threshold is reachable by the seed alone.*

5) Seed protectability with friction: $e > \alpha\,\ell(b(\theta_H), \mu) + k$ — *at high states the seed alone makes participation worthwhile.*

6) Strict contract friction: $k > 0$, *which makes nonparticipation strictly dominant when contract failure is certain.*

For each arm $r \in \{open, SAC\}$, the rank integral

$$R_r(s) = \int_0^1 u_r(s, w)\,dw \tag{9}$$

averages the payoff from participation over every possible marginal participation rank $w \in [0, 1]$ at state s. It answers whether participation pays on average

when the conceptual recruitment path is treated as running from no marginal participation to full marginal participation. The integral is continuous and strictly single crossing on the relevant state range, with a unique zero $\theta_r^* \in (\theta_L, \theta_H)$.

The signal family satisfies the monotone-likelihood-ratio property. The proof allows either of two descriptions of beliefs after a signal. The first is a diffuse calculation: after signal m*, the agent treats the state as* $\theta = m - \sigma\varepsilon$*, with* $\varepsilon \sim F$*; this posterior kernel is associated with an improper flat prior. The second begins with a proper common prior: the state has bounded support and a continuously differentiable density* g*. The two specifications lead to the same small-noise rank calculation. Under either one, a compact interval* $\mathcal{J}$ *contains the dominance states, the crossing range, and every cutoff boundary generated below for all sufficiently small* σ*. This interval keeps the proof in a bounded region. In the proper-prior branch, it also stays away from the endpoints of the prior's support, and* $\inf_{s\in\mathcal{J}} g(s) > 0$ *ensures that the prior assigns positive density throughout it.*

Under either prior specification, the dominance claims do not depend on guessing what unresolved opponents will do. For all sufficiently small σ*, there are signals* $m_L < m_H$ *in* $\mathcal{J}$ *such that nonparticipation is a strict best response below* m_L *and participation is a strict best response above* m_H*, regardless of the strategies still available to other agents. These are the low- and high-signal footholds from which iterative reasoning begins. Every boundary generated by the two extreme cutoff iterations lies in* $\mathcal{J}$*. Public information is held fixed while private signal noise* σ *shrinks toward zero.*

The type-dependent clauses can all hold only if the signing cost is smaller than the improvement in exposure between the unfavorable and favorable benchmark states:

$$k \;<\; \alpha\big(\ell(Y(\theta_L),\mu) - \ell(b(\theta_H),\mu)\big) \tag{10}$$

with the other displayed inequalities satisfied. In practical terms, favorable conditions must improve safety enough to overcome the cost of using the contract. This is a restriction on each one-type economy, not a condition supplied by unmodeled differences among agents.

REMARK 1 (A primitive regularity example): *The following example shows that all parts of the selection domain can hold at once. Let the state lie in* $[0,1]$ *with a continuously differentiable positive prior density, let the signal errors be standard normal, and set*

$$\begin{aligned} b(s) = \frac{1}{10} + \frac{3}{5}s, \qquad Y(s) = \frac{1}{5} + \frac{4}{5}s, \qquad \ell(y,\mu) = 1 - y, \\ \tau = \frac{3}{5}, \qquad \alpha = 1, \qquad e = \frac{2}{5}, \qquad k = \frac{1}{50}. \end{aligned} \tag{11}$$

With these choices, direct integration gives

$$R_{open}(s) = -\frac{9}{20} + \frac{7}{10}s \tag{12}$$

and

$$R_{SAC}(s) = \begin{cases} -\dfrac{1}{50}, & 0 \le s \le \dfrac{1}{2}, \\ \dfrac{4(2s-1)^2}{5(2s+1)} - \dfrac{1}{50}, & \dfrac{1}{2} < s < \dfrac{5}{6}, \\ -\dfrac{47}{100} + \dfrac{7}{10}s, & \dfrac{5}{6} \le s \le 1. \end{cases} \tag{13}$$

When $s \le 1/2$, even the largest available coalition is smaller than the threshold, so the contract fails at every rank. When $1/2 < s < 5/6$, it succeeds only after enough marginal agents join. Once the state reaches $s = 5/6$, the seed alone meets the threshold, and the contract succeeds at every rank thereafter. Under open expression, the average participation payoff changes from negative to positive once, at $9/14$. Under the contract, it changes sign once, at $(81+\sqrt{321})/160$. The contract integral is continuous, and its zero lies strictly between $1/2$ and $5/6$ and below $9/14$.

The example also provides the extreme signals from which the selection argument begins. At state $s = 1/10$, even if every marginal agent joins, open participation pays $2/5 - \ell(Y(1/10), \mu) = -8/25 < 0$. The contract cannot reach the threshold and pays $-1/50$. At state $s = 9/10$, the seed is $16/25 > 3/5$, so participation pays at least $1/25$ under open expression and $1/50$ under the contract even if no other marginal agent joins. These strict inequalities continue to hold for states near those two points. Participation payoffs rise with the state. In this signal model, which adds normal noise to the state, a higher signal also points to a higher state. With sufficiently little noise, agents with low enough signals therefore prefer not to join whatever the others do, while agents with high enough signals prefer to join whatever the others do. The example thus derives the required dominance regions and shows that each average payoff changes sign only once. Outside this class, the general result remains conditional on Assumption 1.

Seed clearance creates the contract's upper dominance region. If $b(s) < \tau$ at every state, one infinitesimal signer cannot trigger publication and still pays k. Even the most favorable signal would then fail to make signing unquestionably worthwhile. The selected contract therefore expands participation around an existing seed.

The formal selection procedure is iterated deletion of interim strictly dominated actions, starting from the dominance regions. If its lower and upper boundaries meet as noise vanishes, the game has an essentially unique selected cutoff even if the original complete-information game had several equilibria.

The following definition makes that reasoning valid for arbitrary measurable strategies, not only symmetric cutoff strategies. It treats every agent at every signal separately, including exceptional signals that occur with probability zero. The diffuse branch uses the posterior kernel just stated. In the proper-prior branch, the regular conditional density is proportional to $g(\theta)f((m-\theta)/\sigma)$. Let $P_\sigma(\cdot \mid m)$ denote whichever posterior belief applies. Initially both actions are available, so $\mathcal{A}_i^0(m) = \{0,1\}$. Let $\mathcal{S}^\gamma$ collect the jointly measurable strategy profiles that choose only actions remaining after round γ. Against a profile $a \in \mathcal{S}^\gamma$, the expected, or *interim*, payoff difference between participation and nonparticipation is

$$\Delta_i(m;a) = \int u\big(\theta, w_a(\theta)\big) P_\sigma(d\theta \mid m). \tag{14}$$

An action already absent remains absent. At the next, or successor, round, action 1 is deleted exactly when $\sup_{a\in\mathcal{S}^\gamma} \Delta_i(m;a) < 0$: even the most favorable remaining behavior by others cannot make joining worthwhile. Action 0 is deleted exactly when $\inf_{a\in\mathcal{S}^\gamma} \Delta_i(m;a) > 0$: even the least favorable remaining behavior by others makes joining worthwhile.

The notation also permits the process to run for infinitely many rounds. At a limit ordinal, set $\mathcal{A}_i^\gamma(m) = \bigcap_{\beta<\gamma} \mathcal{A}_i^\beta(m)$ and again restrict attention to jointly measurable selectors. In ordinary terms, keep only actions that survived every earlier round, then continue deleting. The process must eventually stabilize: every nonstable successor round removes at least one previously available agent-signal-action triple (i,m,d), and the collection of such triples is a set. A profile survives iterated deletion if it chooses from the actions left at this stable point. The ordinal terminology supplies bookkeeping for arbitrarily many rounds; it adds no behavioral or economic assumption. This pointwise definition lets the theorem make a claim at every signal rather than treating two strategies as identical whenever they differ only on events that had zero probability before signals were observed.

LEMMA 1 (Noise-limit selection): *Under Assumption 1, fix either arm $r \in \{open, SAC\}$. As $\sigma \downarrow 0$, the lower and upper signal boundaries of the strategies surviving iterated deletion of interim strictly dominated actions both converge to θ_r^*, the unique zero of R_r. The two boundaries bracket the signals at which rationalizable strategies may still disagree. Equivalently, for every $\delta > 0$, all sufficiently small σ force participation at every signal $m > \theta_r^* + \delta$ and nonparticipation at every signal $m < \theta_r^* - \delta$. The unresolved interval can therefore be made arbitrarily narrow. This collapse of the surviving-strategy disagreement band is the meaning of* essential uniqueness.

At the fixed evaluation state θ, define an arm's rank-indifference value $\bar{e}_{arm}(\theta)$

by

$$\int_0^1 u_{arm}(\theta, w)\, dw = 0. \tag{15}$$

This value answers a simple question: how much private value from public expression would make the marginal type exactly indifferent, after averaging over the participation ranks generated by the global game? To interpret it as a cutoff, compare hypothetical populations that are identical except for their common expressive benefit e. The comparison does not place several payoff types in one population. When the one-type economy at the boundary satisfies Assumption 1, the rank-indifference value is that arm's selected expressive-benefit cutoff. The marginal type participates in otherwise identical one-type economies with e above this cutoff and does not participate in those with e below it. Appendix A proves the result over arbitrary strategies surviving iterated deletion.

The lemma turns a difficult equilibrium-selection problem into a comparison of two average participation payoffs. The next theorem computes those averages for open expression and the contract, then subtracts one from the other.

THEOREM 2 (The selection gap): *At an interior evaluation state, where the seed is below the threshold but full participating capacity is above it, let*

$$w_\tau = \frac{\tau - b(\theta)}{Y(\theta) - b(\theta)} \in (0, 1) \tag{16}$$

be the fraction of the marginal pool needed to move the coalition from seed mass $b(\theta)$ to threshold τ. Ranks below w_τ produce contract failure, while ranks at or above w_τ produce publication. Define the two rank-indifference values

$$\begin{aligned} \bar{e}_{open}(\theta) &= \alpha \int_0^1 \ell\big(y(w;\theta), \mu\big) dw, \\ \bar{e}_{SAC}(\theta) &= \alpha\, \frac{\int_{w_\tau}^1 \ell\big(y(w;\theta), \mu\big) dw}{1 - w_\tau} + \frac{k}{1 - w_\tau}, \end{aligned} \tag{17}$$

The open-expression cutoff averages exposure over the entire recruitment path because every participant is named at every rank. The contract cutoff averages exposure only over successful ranks: below the threshold, the signer is neither named nor given the value of public expression. It then adds $k/(1 - w_\tau)$ because the signer pays friction at every rank but receives expressive value only on the successful share $1 - w_\tau$ of the path. Dividing by that successful share converts the always-paid signing cost into the amount of expressive value needed on success to cover it.

Their accounting gap *is*

$$G^{acct}(\theta) \equiv \bar{e}_{open}(\theta) - \bar{e}_{SAC}(\theta) = \alpha\, w_\tau \operatorname{avg}_{[0,w_\tau]} \ell\big(y(w;\theta),\mu\big) - \alpha\, w_\tau \operatorname{avg}_{[w_\tau,1]} \ell\big(y(w;\theta),\mu\big) - \frac{k}{1-w_\tau}. \tag{18}$$

The first two lines compare exposure below and above the threshold. Because exposure falls as the coalition grows, the below-threshold average is larger. Their weighted difference is the exposure-tail advantage*: the low-coalition exposure between the seed and the threshold that private failure removes. The final line subtracts the friction haircut, which converts the signing cost into the same expressive-benefit units as the exposure terms. The accounting gap is not yet a statement about which outcome is selected; it first records the payoff advantage created by withholding failed lists. A positive value, $G^{acct} > 0$, means that the avoided exposure is worth more than the cost of using the contract.*

Accounting is not enough by itself. A contract cannot recruit this type into a just-successful coalition if being named at the threshold gives negative payoff. The monotone selection domain therefore requires $e \geq \alpha\ell(\tau,\mu)$. Combining this protection requirement with rank indifference gives the contract's expressive-benefit requirement

$$e^{sel}_{SAC}(\theta) = \max\{\bar{e}_{SAC}(\theta), \alpha\ell(\tau,\mu)\}. \tag{19}$$

Whenever the separate one-type economies invoked in the comparison satisfy the remaining clauses of Assumption 1, the selected participation-cutoff gap is

$$G^{sel}(\theta) = \bar{e}_{open}(\theta) - e^{sel}_{SAC}(\theta) = \min\{G^{acct}(\theta), \bar{e}_{open}(\theta) - \alpha\ell(\tau,\mu)\}. \tag{20}$$

The minimum appears because the contract must pass two tests: it must beat open expression in the rank accounting, and its threshold must protect the people it recruits. Thus a favorable selected comparison requires both a positive accounting gap and strict recruitment coverage. If $\bar{e}_{SAC}(\theta) \geq \alpha\ell(\tau,\mu)$, protection does not bind and $G^{sel} = G^{acct}$, so the exposure-tail formula is the selected gap. A positive G^{sel} means that there is an interval of expressive-benefit values for which the marginal type participates under the failure-private contract but not under open expression. Across the one-type economies considered here, the theorem measures the width of that interval in units of e.

The displayed rank-indifference formulas treat verification and escrow costs as specific to the contract. If public expression requires the same costly action in

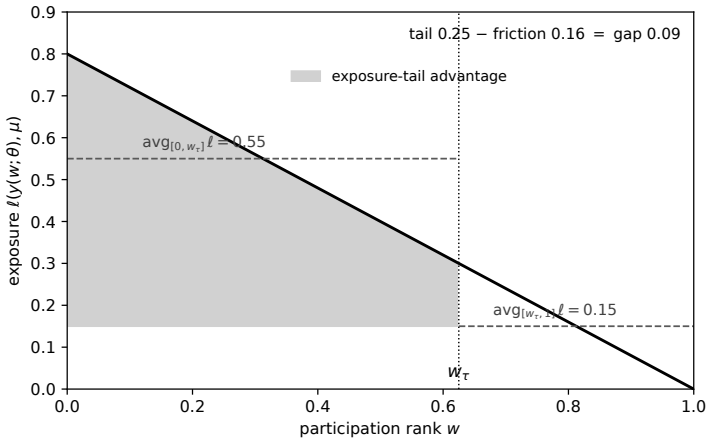

FIGURE 1. BASELINE EXPOSURE-TAIL ACCOUNTING. EXPOSURE $\ell(y(w;\theta),\mu)$ FOLLOWS THE RECRUITMENT PATH FOR $\ell(y) = 1 - y$, $b(\theta) = 0.2$, $Y(\theta) = 1$, $\tau = 0.7$, $\alpha = 1$, AND $k = 0.06$, SO $w_\tau = 0.625$. THE DASHED LINES MARK THE BELOW- AND ABOVE-THRESHOLD RANK AVERAGES, 0.55 AND 0.15. THE SHADED EXPOSURE-TAIL ADVANTAGE IS THE SCALED DIFFERENCE $\alpha w_\tau(0.55 - 0.15) = 0.25$. THE FRICTION HAIRCUT IS $k/(1 - w_\tau) = 0.16$, LEAVING $G^{acct} = 0.09$. BECAUSE $\bar{e}_{SAC} = 0.31 \geq \alpha\ell(\tau,\mu) = 0.30$, THRESHOLD PROTECTION DOES NOT BIND AND $G^{sel} = G^{acct}$.

both arms, then

$$\bar{e}^{sym}_{open}(\theta) = \bar{e}_{open}(\theta) + k, \tag{21}$$

while $\bar{e}^{sym}_{SAC}(\theta) = \bar{e}_{SAC}(\theta)$, and the friction haircut in the accounting gap is $kw_\tau/(1 - w_\tau)$ rather than $k/(1 - w_\tau)$. The open arm's cutoff rises because open expression now also costs k. The protection cap defining the selected gap applies under either normalization. The baseline is appropriate when verification and escrow custody generate the incremental friction; the symmetric formula applies when the same action cost accompanies either form of expression.

Figure 1 illustrates the baseline decomposition. The shaded area is the exposure difference between the below-threshold ranks and the above-threshold average; signing friction is then subtracted separately.

The next corollary translates the gap formula into economically distinct cases. As in the theorem, it compares separate populations whose marginal agents all have the same type; it does not combine different types in one population.

COROLLARY 1 (Conditional comparison across one-type economies): *Fix the common structural objects and consider separate one-marginal-type economies indexed by* (e,α). *Every statement about selected behavior below requires the economy in question to satisfy Assumption 1, including its type-dependent*

protection, dominance, seed-clearance, and friction clauses. Parts 2 and 3 show what happens at two useful boundaries: when coalition size no longer lowers exposure, and when the seed already reaches the threshold at the evaluation state.

1) Intermediate region. *In any such economy whose marginal type satisfies*

$$\max\{\bar{e}_{SAC}(\theta),\ \alpha\ell(\tau,\mu)\} \ < \ e \ < \ \bar{e}_{open}(\theta) \tag{22}$$

the marginal type participates under the selected contract and not under selected open expression. The comparison is across one-type economies, and the strict interval is exactly $G^{sel}(\theta) > 0$. Its weak recruitment-coverage boundary is

$$\bar{e}_{open}(\theta) \ \geq \ \alpha\,\ell(\tau,\mu) \qquad \textit{(equivalently } \int_0^1 \ell(y(w;\theta),\mu)\,dw \ \geq \ \ell(\tau,\mu)\textit{)} \tag{23}$$

with strict inequality for a nonempty strict-type interval. Monotone contract selection requires $e \geq \alpha\ell(\tau,\mu)$ because publication otherwise makes the type's payoff jump downward.

2) Boundary without safety in numbers. *If the strict-decrease assumption is relaxed and ℓ does not depend on coalition mass on the relevant range, the exposure difference between failed and successful ranks vanishes. The accounting gap then equals $-k/(1-w_\tau) < 0$: the contract pays signing friction and concentrates that cost on successful ranks without receiving any exposure benefit.*

3) Seed-at-threshold null. *If $b(\theta) \geq \tau$ at the evaluation state — a boundary case of the theorem's interior condition, read as the limit $w_\tau \downarrow 0$ with $w_\tau = \max\{0,(\tau-b)/(Y-b)\}$ — the tail term vanishes and the accounting gap equals $-k$ exactly. When the core alone already publishes, the contract does not protect any additional below-threshold recruitment ranks and therefore adds only friction. The contract's value is precisely the stretch between what the seed delivers and what the threshold guarantees.*

4) Fixed-index scope. *At any common fixed μ, the exposure-tail advantage is strictly positive when the below-threshold rank average of ℓ strictly exceeds the above-threshold average. The accounting gap is positive if and only if that tail advantage exceeds $k/(1-w_\tau)$; the selected gap is positive if and only if the accounting gap and recruitment-coverage margin are both positive. Information can change μ and thereby change both roots and gaps. Holding μ common isolates conditional disclosure's distinct accounting effect: below-threshold ranks do not expose signers when failure is private.*

COROLLARY 2 (Common-rank accounting robustness): *At an interior evaluation state, let W have any distribution H on $[0,1]$ that is applied identically to both institutions. This permits some parts of the conceptual recruitment path to*

receive more weight than others, while requiring the same weighting in both arms. Define

$$L(W) = \ell\big(y(W;\theta),\mu\big), \qquad p_H = \Pr_H(W \geq w_\tau) \in (0,1). \tag{24}$$

The corresponding H-weighted accounting values are

$$\bar{e}^H_{open} = \alpha\, \mathrm{E}_H[L(W)], \qquad \bar{e}^H_{SAC} = \alpha\, \mathrm{E}_H[L(W) \mid W \geq w_\tau] + \frac{k}{p_H}, \tag{25}$$

and their difference is

$$\begin{aligned} \bar{e}^H_{open} - \bar{e}^H_{SAC} = \alpha(1-p_H)\,(\mathrm{E}_H[L(W) \mid W < w_\tau] - \mathrm{E}_H[L(W) \mid W \geq w_\tau]) \\ - \frac{k}{p_H}. \end{aligned} \tag{26}$$

Under strict safety in numbers, the difference between the two conditional exposure averages is positive. Thus every common rank weighting retains the same avoided-tail-minus-friction accounting structure.

In ordinary language, Corollary 1 says that the contract helps a marginal type only in the middle: the type must value public expression enough to join a protected successful coalition, but not enough to accept the exposure of open expression. If coalition size offers no protection, or if the seed already reaches the threshold, conditional disclosure has no exposure advantage to trade against its signing cost. Corollary 2 separates this accounting logic from equilibrium selection. Taking H to be uniform gives $p_H = 1 - w_\tau$ and reproduces G^{acct} exactly. The corollary does not claim that every distribution H describes equilibrium beliefs. Lemma 1 selects the uniform rank; applying any other common weighting to both institutions merely preserves the form of the decomposition.

B. *Designing the Contract: Thresholds and Failure Privacy*

The selection result also guides the release rule itself. The next results ask three questions: below what coalition size must a safe mechanism keep names private, what shape of rule preserves the incentive to join as participation rises, and what remains if failed lists are disclosed? Fix the marginal type (e,α) and audience index μ, with ℓ strictly decreasing in coalition mass, and define the *protection point*

$$y^*(e,\alpha,\mu) = \inf\{y \in [0,1] :\ e \geq \alpha\ell(y,\mu)\}, \tag{27}$$

the smallest public coalition mass at which the type is willing to be named. If no coalition size makes identification worthwhile, set $y^* = +\infty$.

A safe release rule never names the signer below this protection point. Safety here is evaluated after the final coalition size is known; the companion paper's regret-free safety requirement is the composition-general version of this standard (Cashman, 2026). Among safe rules, a *pointwise maximal* rule releases the list at every coalition size where doing so is safe; it does not withhold at one size merely because it must withhold at another.

PROPOSITION 1 (Safe disclosure representation): *Restrict attention to deterministic anonymous binary rules $R : [0,1] \to \{0,1\}$: the release decision depends only on coalition size, with $R(y) = 1$ releasing the complete realized signer list and $R(y) = 0$ releasing no names. A rule is* ex post safe *if $R(y) = 1 \Rightarrow y \geq y^*$. Within this full-list-or-nothing class, the unique pointwise maximal safe rule is*

$$R^*(y) = \mathbf{1}\{y \geq y^*\}. \tag{28}$$

Thus every safe binary rule withholds the full list below y^*, and the maximal one releases it at every safe mass.

In plain terms, safety forces privacy below y^*, and maximality calls for publication everywhere above it. The proposition characterizes binary full-list-or-nothing rules. Partial lists, composition-dependent exposure, and random release define different mechanism classes.

COROLLARY 3 (Committed thresholds): *A designer may commit before anyone signs to withhold release below a chosen mass satisfying finite $y^* \leq \tau \leq 1$. Within the same binary class and subject to that prior commitment, the unique pointwise maximal safe rule is $R(y) = \mathbf{1}\{y \geq \tau\}$. Below y^* withholding is forced by safety; on $[y^*, \tau)$ it is chosen by the stronger commitment.*

Theorem 2 quantifies this stronger commitment. Raising τ moves more ranks into private failure and lowers $\ell(\tau, \mu)$, so recruitment coverage becomes easier. At the same time, it raises w_τ and the friction haircut $k/(1 - w_\tau)$ and tightens evaluation-state reachability and favorable-state seed clearance. The threshold is governed by a protection–feasibility tradeoff.

COROLLARY 4 (Limits of belief-only interventions): *Call the marginal type* exposure-dominant *if $e < \alpha\ell(b(\theta), \mu')$ for every audience index μ' attainable by a belief-only intervention — one that transmits information but releases no named coalition, so an agent acting alone is publicly associated only with the seed mass $b(\theta)$. Then no belief-only intervention makes seed-level public expression by that type ex post safe. At any induced index μ', a coalition-mass intervention that leaves the maintained payoff technology otherwise unchanged can make association safe only by supplying mass at least $y^*(e, \alpha, \mu')$.*

The corollary separates two routes to safe expression. A sufficiently favorable attainable index can make seed-level expression safe; otherwise safety requires a larger public coalition.

The next result asks which release rules preserve strategic complementarity. An irregular rule that publishes at a small coalition size but withholds at a larger one can make an agent less willing to join when participation rises. The richness condition rules out irrelevant irregularities: it requires that near every size at which release is possible, at least one recruited type would value publication. The *closure* of the release set includes both release sizes and boundary sizes that can be approached by release sizes.

LEMMA 2 (Threshold rules preserve complementarity): *Consider deterministic anonymous failure-private release rules* $R : [0,1] \to \{0,1\}$ *mapping the realized aggregate mass to a release decision, with participation payoff* $R(y)\big(e - \alpha\ell(y,\mu)\big) - k$ *and non-release paying* $-k$*. Say the recruited class is* rich *if for every mass* y *in the* closure *of the release set there is a recruited type with* $e > \alpha\ell(y,\mu)$*. Under richness, the participation payoff is nondecreasing in the aggregate for every recruited type whose success payoff is nonnegative at every released mass if and only if* R *is an upper set — once it releases at one mass, it also releases at every larger mass — and hence is a threshold rule, up to boundary conventions at the cutoff.*

The lemma gives a second reason for thresholds. They do more than protect signers below a chosen mass: they also ensure that additional participation never turns publication off and thereby never reverses the incentive to join.

PROPOSITION 2 (Complete failed-list disclosure): *Suppose every signer's identity is released whether or not the threshold is reached, and failed release gives the same expressive value and exposure cost as open expression at the realized coalition mass. This is the opposite of failure privacy: the platform still collects signatures and checks a threshold, but failure no longer protects anyone. Under the baseline normalization in which only the contract incurs signing friction, the contract payoff at every state and rank is*

$$u^{FD}_{SAC}(\theta, w) = e - \alpha\ell\big(y(w;\theta),\mu\big) - k = u_{open}(\theta, w) - k. \tag{29}$$

Its rank-indifference value is therefore $\bar{e}^{FD}_{SAC}(\theta) = \bar{e}_{open}(\theta) + k$*. Complete failed-list disclosure eliminates the exposure-tail advantage and leaves the contract worse by exactly* k *in expressive-benefit units. If the same friction is paid in both arms, their rank-indifference values coincide.*

This endpoint isolates failure privacy's role. Once failure reproduces open identification and expression, the threshold removes no exposure ranks and the contract retains only its friction disadvantage.

Three margins govern the contract's advantage: the exposure tail it removes, the friction it adds, and the recruitment coverage permitted by a protective and reachable threshold. Credible failed-list privacy preserves the first margin.

Publication also carries information. Under a proper prior with finite mean, a nondegenerate upper-state publication event raises the expected state. When the state measures support for the focal statement, successful conditional disclosure is therefore good news about that statement.

III. Applications

The model applies to settings with five features: a defined eligible population, an audience able to impose material costs, a committed seed, a protective and reachable threshold, and credible custody of failed lists. Exposure must depend primarily on final coalition size, so that adding signers protects those already named. These criteria provide a practical checklist for deciding whether the model fits an application; they are not automatic features of every collective-action setting.

Four practical relationships determine whether those criteria are met. Recruitment ties determine who hears the invitation and therefore help determine the committed seed $b(\theta)$ and the total participating capacity $Y(\theta)$. Trust in the organizers and custody arrangement determines whether people believe the invitation and the promised release rule; distrust can shrink the effective seed or capacity and raise signing friction k. Exposure relationships identify the employer, peers, professional body, or public able to punish signers and therefore determine the audience index μ and exposure function ℓ. Observation channels determine whether that audience sees a failed attempt even when it sees no names. Private recruitment can therefore coexist with failed-list privacy, while public advertising or a visible progress counter changes the information regime.

Safety in numbers is plausible when retaliation uses scarce attention, managerial time, investigative resources, or political capital. Punishing a larger group may also be more visible and institutionally costly, while members of a larger coalition can document, contest, and publicize retaliation together. In those environments, a signer expects a lower material cost in a larger released coalition, which is the maintained inequality $\ell(y', \mu) < \ell(y, \mu)$ for $y' > y$.

That inequality is an empirical scope condition, not a universal property of collective action. If an audience can impose the same sanction on each additional signer at no increasing difficulty, exposure is flat and conditional disclosure has no exposure-tail advantage. If a larger coalition provokes stronger sanctions, exposure can rise instead. A different failure occurs when only one report is needed or only the first reporter receives a reward: one person's participation then reduces, rather than raises, another person's reason to join. Participation may be strategically substitutable even if a larger coalition is safer. Such environments fall outside the monotone coordination comparison in this paper.

Collective whistleblowing is a candidate application when these conditions hold. Employees in a defined organizational unit form the eligible population, and an employer or professional audience can impose material retaliation. Wells Fargo's independent investigation documents sales pressure, internal complaints, and em-

ployee terminations (Independent Directors of the Board of Wells Fargo & Company, 2017); CFPB and DOJ records document the sales-practice enforcement that followed (U.S. Consumer Financial Protection Bureau, 2016; U.S. Department of Justice, 2020). These records show that retaliation can matter. They do not show that each reporter's risk falls as more reporters join or that a protected core can clear a threshold. Where those additional conditions hold, protected or already-disclosed reporters can form the seed. A regulator, independent counsel, or trustee quorum can authenticate employee status, hold submissions, and release a named collective report after the certified count reaches the threshold. The mechanism would keep names private if the complaint remained below the threshold. In exchange, each reporter would bear signing friction k and would be named if the threshold were met.

Pre-petition labor organizing can also fit the model when known organizers or openly pro-union workers supply a credible seed and a larger committee reduces each member's exposure. The prospective bargaining unit is the eligible population, the employer is the exposure audience, and the threshold is the committee's target size before its members go public. This stage precedes a certification election, where the NLRB maps a majority vote into certification and formal representation (U.S. National Labor Relations Board, 2026*a*,*b*). Under these conditions, conditional disclosure lets organizers recruit toward a named committee without revealing the names of workers in a committee that never grew large enough to go public.

Public letters and threshold commitments can fit the same structure. Senior or already-public signers may form the seed, and an independent committee or platform can hold the list. The application fits the model only when that seed is credible and reachable and when appearing on a short list is especially costly to each signer, with the cost falling as the list grows. The same conditions can be applied to named organizing committees and public statements backed by a fixed constituency.

Custody is part of the economic mechanism. A sympathetic independent holder can enforce the release rule directly. A distributed design instead combines authenticated encrypted submissions, an auditable threshold decision, and a quorum—a required minimum number—of independent trustees, so that no single administrator can read or release a failed list. Secure allegation escrows demonstrate closely related functions (Arun et al., 2020); Easwar Vivek Mangipudi, Donghang Lu, Alexandros Psomas and Aniket Kate (2023) develop a collusion-deterrent threshold escrow. Supplemental Appendix D combines these components into a reference architecture and states the security and governance assumptions it would require. The cited systems do not implement or audit that exact design. Counsel, a regulator, a union, or a platform could operate the escrow without belonging to the coalition itself.

A. Testable Implications

The model yields four sharp predictions. First, conditional disclosure lowers the participation cutoff when the removed exposure tail exceeds signing friction and the threshold protects the recruited type. Second, the gain disappears when the seed cannot support high-state threshold clearance or the threshold exceeds feasible capacity. Third, a higher threshold is not always better: it provides more protection if publication occurs, but makes publication harder and magnifies the friction haircut. Fourth, complete failed-list disclosure eliminates the exposure-tail advantage. Laboratory and field experiments can vary seed size, threshold, custody, and failure disclosure independently to test each margin.

These results compare institutions by the minimum private value needed to induce someone to join a named coalition. They do not by themselves rank the coalitions' social value. That welfare judgment depends on the statement's content, the action the coalition produces, and its effects on non-signers and targets.

IV. Concluding Remarks

Conditional disclosure changes what happens after failed coordination. In the seeded one-type benchmark, it removes the exposure attached to ranks between the committed core and the publication threshold. The exposure-tail decomposition measures that benefit against signing friction at a fixed state and common audience index. The selected cutoff advantage is the smaller of the accounting gap and the recruitment-coverage margin. A protective and reachable threshold, high-state seed clearance, and credible failure privacy define the region in which the contract expands willingness to participate.

Information and conditional disclosure act on different objects. A survey can change beliefs and behavior; conditional disclosure changes whether a participant is identified after a small coalition forms. Measurement and commitment therefore perform complementary functions. A survey can reveal latent support, while the contract converts that support into a named coalition by removing failed-path exposure. Complete failed-list disclosure destroys this advantage and leaves only the contract's signing friction.

A committed core supplies the high-state anchor, the threshold determines the protection offered to marginal participants, and failure privacy makes that protection credible. The finite Bayesian model in the Supplemental Appendix asks how a person's chance of affecting publication, the probability of success, and signing friction shape participation in a fixed population. It does not compare the contract with open expression; that comparison comes from the continuum benchmark and its coalition-size exposure payoff. The supplement also explains how existing institutional and cryptographic components could be combined. Whether a particular system protects signers depends on the attacks it must withstand,

how its components are integrated, and the quality of its auditing, governance, and user interface.

The empirical agenda follows directly from the theory. Experiments should vary seed size, threshold feasibility, signing friction, and failure disclosure rather than treat conditional commitment as a single intervention. The exposure-tail formula predicts which contracts should expand public coalitions, which should merely add friction, and which institutional feature accounts for the difference.

Appendix A. Proofs

This appendix collects the proofs of the results stated in the body, in order of appearance. Most use direct algebra or the fact that exposure falls as coalition size rises. The noise-limit selection proof is longer because it does not assume in advance that the strategies surviving rationality are symmetric cutoffs. Its logic has four steps. First, identify signals at which one action is best whatever others do. Second, construct lower and upper cutoff bounds. Third, show that every strategy surviving repeated deletion lies between those bounds. Finally, show that both bounds converge to the unique zero of the rank integral as private noise vanishes.

PROOF OF THEOREM 1:

Because ℓ is strictly decreasing in coalition mass, the gain from expressing rises with the expressing mass. At the seed-only profile a marginal expresser faces aggregate $b(\theta)$ and payoff $e-\alpha\ell(b(\theta),\mu)<0$ under the right inequality, so silence is a strict best response. At the full-expression profile she faces $Y(\theta)$ and payoff $e-\alpha\ell(Y(\theta),\mu)\geq 0$ under the left inequality, so expressing is a best response. Thus both profiles are equilibria. An informational intervention matters here only through any change it induces in μ or the state being evaluated; the equilibrium multiplicity follows at each fixed pair inside the displayed band.

PROOF OF LEMMA 1:

The proof follows the four-step roadmap above. It first shows that better states and greater participation make joining more attractive and identifies the signals at which one action is dominant. It then builds two cutoff sequences that approach the set of surviving strategies from opposite sides. Finally, it shows why an agent exactly at the cutoff evaluates all ranks uniformly when signal noise becomes small. That step converts the boundary agent's payoff into the rank average R.

Fix one arm. To keep the notation readable, omit the arm's subscript and write u, R, and θ^* for its payoff, rank integral, and unique crossing. For the open arm, $y(w;\theta)$ rises with both the state and marginal participation, while ℓ falls as coalition mass rises. For the contract, the payoff is $-k$ below the threshold, jumps weakly upward when publication begins because of τ-protection, and rises thereafter. Hence both payoff functions are bounded and nondecreasing in the state and in participation by others. At θ_L, open expression pays at most $e-\alpha\ell(Y(\theta_L),\mu)<0$, while the unreachable contract pays $-k<0$. At θ_H, the seed clears the threshold and clause 5 makes participation strictly profitable in both

arms. Thus low states make staying out best even under maximal participation, while high states make joining best even under minimal marginal participation. These are the state dominance regions used below.

Cutoff bounds for pointwise deletion. Against a symmetric cutoff c, define

$$D_\sigma(m;c) = \mathrm{E}\big[u\big(\theta, w_c(\theta)\big) \mid m_i = m\big], \qquad w_c(\theta) = 1 - F\left(\frac{c-\theta}{\sigma}\right), \tag{A1}$$

with $w_{-\infty} = 1$ and $w_{+\infty} = 0$. The function $D_\sigma(m;c)$ is the expected gain from participating for an agent with signal m when every other marginal agent is conjectured to use cutoff c. A higher own signal makes favorable states more likely. The monotone-likelihood-ratio property and the fact that payoffs improve with the state therefore make D_σ nondecreasing in m. A higher conjectured cutoff means that fewer other agents participate. Strategic complementarity therefore makes D_σ nonincreasing in c.

The posterior kernels give an explicit continuity argument. In the diffuse branch,

$$D_\sigma(m;c) = \int u\left(m - \sigma\varepsilon, 1 - F\left(\varepsilon + \frac{c-m}{\sigma}\right)\right) dF(\varepsilon). \tag{A2}$$

In the proper-prior branch, the same expression is

$$\frac{\int u\big(m - \sigma\varepsilon, 1 - F\big(\varepsilon + \frac{c-m}{\sigma}\big)\big)\, g(m - \sigma\varepsilon) f(\varepsilon)\, d\varepsilon}{\int g(m - \sigma\varepsilon) f(\varepsilon)\, d\varepsilon}. \tag{A3}$$

The denominator is positive on $\mathcal{J}$. For any convergent sequence (m_j, c_j), the integrands converge at each error realization except possibly where the limiting coalition mass equals τ. As ε rises, both the state and opponents' participation fall, so that equality holds at at most one value of ε. A continuous error distribution assigns probability zero to that single value. The dominated-convergence theorem then permits the limit to pass through the integral because payoffs are bounded and the other primitives are continuous. This proves that D_σ is jointly continuous in the agent's signal and the conjectured cutoff on the confined region in either branch. Extending g by zero creates possible discontinuities only at the two endpoints of the prior's support, which also receive probability zero under the continuous error density.

The cutoff $-\infty$ represents the most optimistic conjecture about others, in which every marginal agent participates; $+\infty$ represents the most pessimistic conjecture, in which none does. Starting from these extremes, set $\underline{c}_0 = -\infty$ and $\bar{c}_0 = +\infty$. Each new lower cutoff is the first signal at which joining pays under the previous optimistic bound. Each new upper cutoff is the last signal at which joining does

not pay under the previous pessimistic bound:

$$\begin{aligned} \underline{c}_{n+1} &= \inf\{m : D_\sigma(m; \underline{c}_n) \geq 0\}, \\ \bar{c}_{n+1} &= \sup\{m : D_\sigma(m; \bar{c}_n) \leq 0\}. \end{aligned} \tag{A4}$$

Uniform signal dominance makes the boundaries finite after the first round and confines them to $\mathcal{J}$. As reasoning proceeds, the optimistic lower boundary can only move upward and the pessimistic upper boundary can only move downward:

$$\underline{c}_n \uparrow \underline{c}_\sigma, \qquad \bar{c}_n \downarrow \bar{c}_\sigma, \qquad \underline{c}_\sigma \leq \bar{c}_\sigma. \tag{A5}$$

For completeness, define $\beta^-(c) = \inf\{m : D_\sigma(m; c) \geq 0\}$ and $\beta^+(c) = \sup\{m : D_\sigma(m; c) \leq 0\}$. These are the lower and upper best-response boundaries to a conjectured cutoff c. Both maps are nondecreasing in c, and continuity gives $\beta^-(c) \leq \beta^+(c)$. Since $\underline{c}_{n+1} = \beta^-(\underline{c}_n)$ and $\bar{c}_{n+1} = \beta^+(\bar{c}_n)$, induction from $-\infty \leq +\infty$ proves the two monotone directions and

$$\underline{c}_{n+1} = \beta^-(\underline{c}_n) \leq \beta^-(\bar{c}_n) \leq \beta^+(\bar{c}_n) = \bar{c}_{n+1}. \tag{A6}$$

The next step connects these symmetric cutoff calculations to all surviving strategies, including asymmetric and nonmonotone ones. Inductively, every strategy surviving n rounds obeys the pointwise outer bounds

$$\mathbf{1}\{m > \bar{c}_n\} \leq a_i(m) \leq \mathbf{1}\{m \geq \underline{c}_n\}. \tag{A7}$$

The inequalities say that every surviving strategy must prescribe participation above the upper boundary and nonparticipation below the lower boundary; only signals between the two remain unresolved. The base case is $0 \leq a_i(m) \leq 1$. If the bounds hold at round n, exact aggregation and continuity of F imply, for every measurable surviving profile a,

$$w_{\bar{c}_n}(\theta) \leq w_a(\theta) \leq w_{\underline{c}_n}(\theta) \qquad \text{for every } \theta. \tag{A8}$$

If $m > \bar{c}_{n+1}$, then $D_\sigma(m; \bar{c}_n) > 0$. Participation is therefore strictly profitable even against the lowest participation consistent with the previous bounds, and hence against every profile in $\mathcal{S}^n$, so action 0 is deleted. If $m < \underline{c}_{n+1}$, then $D_\sigma(m; \underline{c}_n) < 0$. Participation is strictly unprofitable even against the highest participation consistent with the previous bounds, so action 1 is deleted. This proves the next-round bounds. The bounding cutoffs are comparison devices: they need not themselves be feasible surviving profiles at a finite round. The argument therefore does not assume that intermediate survivors are symmetric or monotone.

Limits of the cutoff iterations.

Continuity and the dominance regions imply that the agent at each finite-round boundary is exactly indifferent, giving

$$D_\sigma(\underline{c}_{n+1};\underline{c}_n)=0, \qquad D_\sigma(\bar{c}_{n+1};\bar{c}_n)=0. \tag{A9}$$

Passing to the limits yields

$$D_\sigma(\underline{c}_\sigma;\underline{c}_\sigma)=0, \qquad D_\sigma(\bar{c}_\sigma;\bar{c}_\sigma)=0. \tag{A10}$$

Hence $\mathbf{1}\{m \geq \underline{c}_\sigma\}$ and $\mathbf{1}\{m > \bar{c}_\sigma\}$ are best responses to themselves under opposite rules for what happens at exact indifference. It remains to verify that these endpoint profiles survive the full deletion process. Transfinite induction is the formal device for doing so even if deletion continues beyond every finite round. Each endpoint profile is admissible initially. If all of its prescribed actions survive through round γ, the profile itself is a measurable selector in $\mathcal{S}^\gamma$. Because every prescribed action is a best response to that selector, it cannot be strictly dominated against every admissible profile at round $\gamma+1$. At a limit ordinal, those same actions remain because they survived every earlier round. Thus both endpoint profiles survive every round. Every profile in the stable correspondence also survives each finite round, so taking $n \to \infty$ in the outer bounds places it between these endpoint profiles. Continuing the deletion process cannot move either bound.

Rank evaluation and the noise limit. Under the diffuse posterior kernel, an agent at the boundary $m_i = c$ represents the unknown state as $\theta = c - \sigma\varepsilon$, with $\varepsilon \sim F$. The probability-integral transform says that applying a continuous distribution function to a draw from that distribution produces a uniform random variable. It therefore makes $1 - F(\varepsilon)$ uniform on $[0,1]$, so the boundary payoff is

$$D_\sigma(c;c)=\int_0^1 u\big(c-\sigma F^{-1}(1-w),w\big)\,dw. \tag{A11}$$

This calculation is the source of the uniform rank used in the main theorem; uniform weighting is a consequence of the private-signal structure, not an arbitrary averaging choice. At the boundary signal, uncertainty about the state is equivalent, in the limit, to uncertainty about a uniformly distributed position along the recruitment path. For any sequence $c_\sigma \to c \in \mathcal{J}$, bounded payoffs again allow dominated convergence to give $D_\sigma(c_\sigma;c_\sigma) \to R(c)$. The contract payoff jumps when the threshold is crossed, but that crossing occurs at no more than one rank and therefore has no effect on the integral.

Under a proper prior, extend g by zero outside its compact support. The posterior density of $\varepsilon = (c_\sigma - \theta)/\sigma$ is proportional to $g(c_\sigma - \sigma\varepsilon)f(\varepsilon)$, and the

same payoff is

$$\frac{\int u\big(c_\sigma - \sigma\varepsilon, 1 - F(\varepsilon)\big) g(c_\sigma - \sigma\varepsilon) f(\varepsilon)\, d\varepsilon}{\int g(c_\sigma - \sigma\varepsilon) f(\varepsilon)\, d\varepsilon}. \tag{A12}$$

The denominator converges to $g(c) > 0$ and dominated convergence again gives $R(c)$. As in the diffuse calculation, the single moving rank at which the contract begins to publish has posterior probability zero and does not change the limit.

Both boundary sequences lie in the closed and bounded interval $\mathcal{J}$, so every subsequence has a convergent further subsequence. At any such limit c, boundary indifference and either rank calculation imply $R(c) = 0$. Strict single crossing says that R has only one zero, θ^*. Every convergent subsequence must therefore lead to that same point, which proves that the full lower and upper sequences converge to it:

$$\underline{c}_\sigma \longrightarrow \theta^*, \qquad \bar{c}_\sigma \longrightarrow \theta^*. \tag{A13}$$

Every surviving strategy is bracketed by these cutoffs. Hence, for every $\delta > 0$, all sufficiently small σ force participation above $\theta^* + \delta$ and nonparticipation below $\theta^* - \delta$, as claimed.

PROOF OF THEOREM 2:

For the open arm, expressive value arrives and exposure is incurred at every rank. The rank integral is therefore $e - \alpha \int_0^1 \ell(y(w;\theta), \mu)\, dw$. Setting it to zero gives the first rank-indifference value. For the contract, expressive value and exposure arrive only on ranks at or above w_τ, while friction is paid at every rank:

$$\int_0^1 u_{SAC}(\theta, w)\, dw = (1 - w_\tau)\, e - \alpha \int_{w_\tau}^1 \ell\big(y(w;\theta), \mu\big) dw - k, \tag{A14}$$

Setting this expression to zero gives the second rank-indifference value. To obtain G^{acct}, subtract the contract value from the open value, split the open integral into its below- and above-threshold parts, and write each integral as the length of its rank interval times the average over that interval. Within the monotone contract domain, the contract must pass both rank indifference and τ-protection. Its smallest admissible expressive value is therefore the larger of the two requirements, $e^{sel}_{SAC} = \max\{\bar{e}_{SAC}, \alpha\ell(\tau, \mu)\}$. Subtracting that maximum from $\bar{e}_{open}$ and using $x - \max\{y, z\} = \min\{x - y, x - z\}$ gives the formula for G^{sel}.

Under symmetric friction, subtracting k from the open payoff adds k to its expressive-benefit cutoff. The contract cutoff is unchanged, so adding k to the displayed baseline gap changes its friction term from $-k/(1 - w_\tau)$ to $-kw_\tau/(1 - w_\tau)$.

PROOF OF COROLLARY 1:

For part 1, take each (e, α) as indexing its own economy. If that economy

satisfies Assumption 1, Lemma 1 applies to both arms. The displayed inequalities then put its single marginal type above the contract cutoff e^{sel}_{SAC} and below the open-expression cutoff $\bar{e}_{open}$. The maximum in the lower bound enforces both rank indifference and τ-protection. The interval between the two cutoffs is nonempty exactly when $G^{sel} > 0$; equality at the recruitment-coverage boundary produces no strict interval. For part 2, removing safety in numbers makes the below- and above-threshold exposure averages equal, so their difference disappears. For part 3, as the seed approaches the threshold, $w_\tau \downarrow 0$ and the protected exposure tail disappears, leaving the baseline signing cost k. Part 4 follows by comparing the two rank averages and the protection requirement at the same fixed μ; neither a distribution of types nor a different audience index across institutions enters that calculation.

PROOF OF COROLLARY 2:

Under the common weighting H, open expression gives expected rank payoff $e - \alpha \, \mathrm{E}_H[L(W)]$, which yields $\bar{e}^H_{open}$. Under the contract, expressive value and exposure occur only when $W \geq w_\tau$, an event with probability p_H, while friction is always paid. Its expected rank payoff is therefore

$$p_H e - \alpha p_H \, \mathrm{E}_H[L(W) \mid W \geq w_\tau] - k, \tag{A15}$$

which yields $\bar{e}^H_{SAC}$. The law of total expectation gives

$$\mathrm{E}_H[L(W)] = (1 - p_H) \, \mathrm{E}_H[L(W) \mid W < w_\tau] + p_H \, \mathrm{E}_H[L(W) \mid W \geq w_\tau]. \tag{A16}$$

Substitution and subtraction give the stated identity. Because $y(w;\theta)$ rises in w and ℓ strictly decreases in coalition mass, exposure at every below-threshold rank exceeds exposure at every successful rank. Since $p_H \in (0,1)$, both regions receive positive probability and the difference between their conditional exposure averages is strictly positive.

PROOF OF PROPOSITION 1:

If $y^* = +\infty$, no coalition size protects the type. Safety then forces $R(y) = 0$ on all of $[0,1]$, which is the unique maximal safe rule. Otherwise, at every $y < y^*$, ex post safety forces every admissible binary rule to withhold the list. At every $y \geq y^*$, releasing the complete realized list is safe, so $R^*(y) = 1$. Hence R^* releases whenever any safe rule could release and weakly dominates every safe binary rule at each coalition size. Any other safe rule must withhold at some $y \geq y^*$ and is strictly less permissive there, proving uniqueness of the pointwise maximal rule. No partial-list release is available in the stated class.

PROOF OF COROLLARY 3:

The prior commitment forces $R(y) = 0$ for every $y < \tau$. Because $\tau \geq y^*$, every coalition at or above the chosen threshold is large enough to protect the type, so releasing the complete list there is safe. The rule $R(y) = \mathbf{1}\{y \geq \tau\}$ is therefore feasible and releases at every size permitted by the commitment. It

weakly dominates every other safe rule satisfying that commitment, with strict pointwise domination wherever another rule withholds at a mass $y \geq \tau$.

PROOF OF COROLLARY 4:

A belief-only intervention moves only the audience index from μ to some attainable μ'. It does not add names, so an isolated marginal expresser still appears beside only the seed mass $b(\theta)$. Exposure-dominance makes her payoff $e - \alpha\ell(b(\theta), \mu') < 0$ at every such index. Under the maintained payoff technology, making public association worthwhile at μ' therefore requires coalition mass at least $y^*(e, \alpha, \mu') > b(\theta)$.

PROOF OF LEMMA 2:

First suppose $R = \mathbf{1}\{y \geq \tau\}$. For a type whose success payoff is nonnegative wherever the list is released, participation pays $-k$ below the threshold, jumps weakly upward when the threshold is reached, and rises above it because ℓ strictly decreases in coalition mass. The participation payoff is therefore nondecreasing in coalition size.

For the converse, suppose R is not an upper set. Then there are coalition sizes $y_1 < y_2$ at which the rule releases at y_1 but withholds at y_2. Let $\underline{y} = \inf\{y : R(y) = 1\}$ be the lower edge of the release set. Richness supplies a recruited type with $e > \alpha\ell(\underline{y}, \mu)$. Because exposure falls with coalition size and every released mass is at least $\underline{y}$, this type has strictly positive success payoff at every released mass. It therefore belongs to the class considered in the lemma. Yet its participation payoff falls from $e - \alpha\ell(y_1, \mu) - k > -k$ at y_1 to $-k$ at the larger size y_2, where the list is withheld. This violates monotonicity for a type the rule actually recruits.

The richness condition is needed for this converse. Without it, a rule could release irregularly at low coalition sizes where publication benefits none of the recruited types. Those irrelevant releases need not create a payoff reversal for any type under consideration, even though the release region is not an upper set.

PROOF OF PROPOSITION 2:

Under complete failed-list disclosure, threshold success no longer changes whether a signer is named: names are public after both success and failure. By assumption, crossing the threshold also does not change the signer's expressive value or exposure technology. The only remaining payoff difference from open expression is therefore the contract-specific friction k, so $u^{FD}_{SAC} = u_{open} - k$ at every state and rank. Integrating over ranks gives

$$\int_0^1 u^{FD}_{SAC}(\theta, w)\, dw = e - \alpha \int_0^1 \ell\big(y(w;\theta), \mu\big)\, dw - k. \tag{A17}$$

Setting this expression to zero gives $\bar{e}^{FD}_{SAC} = \bar{e}_{open} + k$. If open expression also incurs k, subtracting the same friction from both arms makes their rank-indifference values equal.

REFERENCES

Arun, Venkat, Aniket Kate, Deepak Garg, Peter Druschel, and Bobby Bhattacharjee. 2020. "Finding Safety in Numbers with Secure Allegation Escrows." San Diego, CA, USA:Internet Society.

Ayres, Ian, and Cait Unkovic. 2012. "Information Escrows." *Michigan Law Review*, 111: 145.

Bagnoli, Mark, and Barton L. Lipman. 1989. "Provision of Public Goods: Fully Implementing the Core through Private Contributions." *The Review of Economic Studies*, 56(4): 583–601.

Bernheim, B. Douglas. 1994. "A Theory of Conformity." *Journal of Political Economy*, 102(5): 841–877.

Braghieri, Luca. 2024. "Political Correctness, Social Image, and Information Transmission." *American Economic Review*, 114(12): 3877–3904.

Braghieri, Luca, Leonardo Bursztyn, and Jan Fasnacht. 2026. "Threshold Disclosure in Collective Decisions." NBER Working Paper 34827.

Breuer, Florian. 2024. "SeCritMass: Threshold Secret Petitions." University of Newcastle.

Bursztyn, Leonardo, Alessandra L. González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review*, 110(10): 2997–3029.

Camenisch, Jan, and Anna Lysyanskaya. 2001. "An Efficient System for Non-transferable Anonymous Credentials with Optional Anonymity Revocation." In *Advances in Cryptology — EUROCRYPT 2001*. Vol. 2045, , ed. Gerhard Goos, Juris Hartmanis, Jan Van Leeuwen and Birgit Pfitzmann, 93–118. Berlin, Heidelberg:Springer Berlin Heidelberg.

Cantoni, Davide, David Y Yang, Noam Yuchtman, and Y Jane Zhang. 2019. "Protests as Strategic Games: Experimental Evidence from Hong Kong's Antiauthoritarian Movement." *The Quarterly Journal of Economics*, 134(2): 1021–1077.

Carlsson, Hans, and Eric van Damme. 1993. "Global Games and Equilibrium Selection." *Econometrica*, 61(5): 989–1018.

Cashman, Matthew. 2026. "Failure Privacy and Safe Collective Expression with Social Assurance Contracts." Working paper, arXiv:2607.05802.

Chassang, Sylvain, and Gerard Padró i Miquel. 2019. "Crime, Intimidation, and Whistleblowing: A Theory of Inference from Unverifiable Reports." *The Review of Economic Studies*, 86(6): 2530–2553.

Chwe, Michael Suk-Young. 1999. "Structure and Strategy in Collective Action." *American Journal of Sociology*, 105(1): 128–156.

Douceur, John R. 2002. "The Sybil Attack." 251–260. Berlin, Heidelberg:Springer.

Edmond, Chris. 2013. "Information Manipulation, Coordination, and Regime Change." *The Review of Economic Studies*, 80(4): 1422–1458.

Evans, David, Vladimir Kolesnikov, and Mike Rosulek. 2018. "A Pragmatic Introduction to Secure Multi-Party Computation." *Foundations and Trends in Privacy and Security*, 2(2-3): 70–246.

Frankel, David M., Stephen Morris, and Ady Pauzner. 2003. "Equilibrium Selection in Global Games with Strategic Complementarities." *Journal of Economic Theory*, 108(1): 1–44.

Goldstein, Itay, and Ady Pauzner. 2005. "Demand–Deposit Contracts and the Probability of Bank Runs." *The Journal of Finance*, 60(3): 1293–1327.

Halac, Marina, Ilan Kremer, and Eyal Winter. 2020. "Raising Capital from Heterogeneous Investors." *American Economic Review*, 110(3): 889–921.

Heese, Jonas, and Gerardo Pérez-Cavazos. 2021. "The Effect of Retaliation Costs on Employee Whistleblowing." *Journal of Accounting and Economics*, 71(2-3): 101385.

Independent Directors of the Board of Wells Fargo & Company. 2017. "Sales Practices Investigation Report." Independent Directors of the Board of Wells Fargo & Company.

Kosba, A., A. Miller, E. Shi, Z. Wen, and C. Papamanthou. 2016. "Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts." 839–858.

Kuran, Timur. 1997. *Private Truths, Public Lies: The Social Consequences of Preference Falsification.* Cambridge, MA:Harvard University Press.

Mangipudi, Easwar Vivek, Donghang Lu, Alexandros Psomas, and Aniket Kate. 2023. "Collusion-Deterrent Threshold Information Escrow." 584–599.

McCorry, Patrick, Siamak F. Shahandashti, and Feng Hao. 2017. "A Smart Contract for Boardroom Voting with Maximum Voter Privacy." 357–375. Cham:Springer International Publishing.

Morris, Stephen, and Hyun Song Shin. 2003. "Global Games: Theory and Applications." In *Advances in Economics and Econometrics*. , ed. Mathias Dewatripont, Lars Peter Hansen and Stephen J. Turnovsky, 56–114. Cambridge:Cambridge University Press.

Prentice, Deborah A, and Dale T Miller. 1993. "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *Journal of Personality and Social Psychology*, 64(2): 243–256.

Prentice, Deborah A., and Dale T. Miller. 1996. "Pluralistic Ignorance and the Perpetuation of Social Norms by Unwitting Actors." In *Advances in Experimental Social Psychology*. Vol. 28, , ed. Mark P. Zanna, 161–209. Academic Press.

Rezaei, Hadis, Mojtaba Eshghie, Karl Anderesson, and Francesco Palmieri. 2025. "SoK: Root Cause of $1 Billion Loss in Smart Contract Real-World Attacks via a Systematic Literature Review of Vulnerabilities."

Sákovics, József, and Jakub Steiner. 2012. "Who Matters in Coordination Problems?" *American Economic Review*, 102(7): 3439–3461.

Sonnino, Alberto, Mustafa Al-Bassam, Shehar Bano, Sarah Meiklejohn, and George Danezis. 2019. "Coconut: Threshold Issuance Selective Disclosure Credentials with Applications to Distributed Ledgers." *Proceedings 2019 Network and Distributed System Security Symposium*.

Strausz, Roland. 2017. "A Theory of Crowdfunding: A Mechanism Design Approach with Demand Uncertainty and Moral Hazard." *American Economic Review*, 107(6): 1430–1476.

Tabarrok, Alexander. 1998. "The Private Provision of Public Goods via Dominant Assurance Contracts." *Public Choice*, 96(3): 345–362.

U.S. Consumer Financial Protection Bureau. 2016. "Consumer Financial Protection Bureau Fines Wells Fargo $100 Million for Widespread Illegal Practice of Secretly Opening Unauthorized Accounts." *https://www.consumerfinance.gov/about-us/newsroom/consumer-financial-protection-bureau-fines-wells-fargo-100-million-widespread-illegal-practice-secretly-opening-unauthorized-accounts/*.

U.S. Department of Justice. 2020. "Wells Fargo Agrees to Pay $3 Billion to Resolve Criminal and Civil Investigations into Sales Practices Involving the Opening of Millions of Accounts without Customer Authorization." *https://www.justice.gov/archives/opa/pr/wells-fargo-agrees-pay-3-billion-resolve-criminal-and-civil-investigations-sales-practices*.

U.S. National Labor Relations Board. 2026*a*. "Conduct Elections | National Labor Relations Board." *https://www.nlrb.gov/about-nlrb/what-we-do/conduct-elections*.

U.S. National Labor Relations Board. 2026*b*. "Your Right to Form a Union | National Labor Relations Board." *https://www.nlrb.gov/about-nlrb/rights-we-protect/the-law/employees/your-right-to-form-a-union*.

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association*, 60(309): 63–69.

Werner, Sam, Daniel Perez, Lewis Gudgeon, Ariah Klages-Mundt, Dominik Harz, and William Knottenbelt. 2022. "SoK: Decentralized Finance (DeFi)." *Proceedings of the 4th ACM Conference on Advances in Financial Technologies*, 30–46.

Winter, Eyal. 2004. "Incentives and Discrimination." *American Economic Review*, 94(3): 764–773.

Supplemental Appendix

This Supplemental Appendix develops three parts of the argument that are useful but not needed for the article's central comparison. First, it replaces the continuum with a literal finite group and shows how a common signing cutoff works when one person can affect publication. Second, it distinguishes hiding a failed signer list from hiding the fact that a campaign was attempted, and explains when coalition size summarizes all the information in a named list. Third, it describes how existing institutional and cryptographic tools can implement conditional disclosure. The article contains the continuum selection proof and common-rank accounting result.

Appendix B. Finite Committed-Core Benchmark

This section replaces the continuum of the article with a literal group of N marginal agents. Unlike an infinitesimal agent in the continuum, each of these people contributes one whole signature and may be pivotal: adding her name can move the final count across the threshold. The section asks whether a simple cutoff rule remains meaningful when each person has this positive chance of affecting publication and does not know how many others will sign.

Let the state θ measure the underlying favorability of the environment. It lies in $[\underline{\theta}, \overline{\theta}]$. Before receiving private information, agents share a prior density g, which assigns positive probability density continuously across that range and describes how plausible they initially find each state. The integer publication threshold satisfies $K \geq 2$. A weakly increasing measurable function

$$B : [\underline{\theta}, \overline{\theta}] \longrightarrow \{0, 1, \ldots, K\} \tag{B1}$$

gives the number of committed or protected signers in each state. Thus $B(\theta)$ is a count, not a probability. These core signers participate independently of the marginal agents' decisions. The remaining N agents decide simultaneously whether to join the list. Because B is weakly increasing, more favorable states have at least as many committed signers as less favorable states.

The function B and the prior are common knowledge, but the realized state and hence the realized core count $B(\theta)$ are not observed before marginal agents act. During the simultaneous collection period, each marginal agent observes only her own signal: she does not observe other signals, other marginal decisions, or an interim count.

Agent i observes

$$x_i = \theta + \sigma\varepsilon_i, \tag{B2}$$

where $\sigma > 0$ controls signal noise. The shocks ε_i are independent across agents,

independent of θ, and identically distributed with continuous strictly increasing distribution F and positive density f; the structure satisfies monotone likelihood ratios. If agent i signs and Q_{-i} other marginal agents sign, publication occurs when

$$B(\theta) + 1 + Q_{-i} \geq K. \tag{B3}$$

Successful publication gives the signer the private value $v(x_i) \geq 0$ of being named on the released list, where v is continuous and strictly increasing. A non-signer receives no expressive or substantive payoff from another coalition's publication. This normalization keeps the decision focused on the private return from joining; a public benefit enjoyed by non-signers would add a separate motive to be the person whose signature causes publication. Signing costs $k > 0$ whether or not publication occurs. Ideal failure privacy gives no expressive value or exposure payoff after failure, so signing yields

$$u_i(1, Q_{-i}, \theta; x_i, k) = \mathbf{1}\{B(\theta) + 1 + Q_{-i} \geq K\} v(x_i) - k, \tag{B4}$$

whereas not signing yields zero. The agent pays k for certain and receives $v(x_i)$ only if the list is published.

A symmetric cutoff s has every marginal agent sign exactly when her signal is at least s; by convention, equality leads to signing. Allowing $s \in \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ includes the extreme rules "always sign" and "never sign." Under this conjecture, another marginal agent signs in state θ with probability

$$p_s(\theta) = 1 - F\left(\frac{s - \theta}{\sigma}\right). \tag{B5}$$

Use the conventions $p_{-\infty} = 1$ and $p_{+\infty} = 0$. Conditional on θ, each of the other $N - 1$ agents independently signs with this probability, so their total is $Q_{-i}(s) \sim \text{Binomial}(N - 1, p_s(\theta))$. After observing x, agent i holds the posterior density

$$g(\theta \mid x) = \frac{g(\theta) f((x - \theta)/\sigma)}{\int_{\underline{\theta}}^{\overline{\theta}} g(t) f((x - t)/\sigma)\, dt}. \tag{B6}$$

For a given state, define $r(\theta) = \max\{0, K - B(\theta) - 1\}$ as the minimum number of other marginal signers needed for publication after agent i joins. The state-

contingent success probability is the upper tail of the binomial distribution:

$$\pi(\theta; s) = \sum_{q=r(\theta)}^{N-1} \binom{N-1}{q} p_s(\theta)^q \{1 - p_s(\theta)\}^{N-1-q}, \tag{B7}$$
$$r(\theta) = \max\{0, K - B(\theta) - 1\},$$

with the sum interpreted as zero if $r(\theta) > N - 1$. The success probability conditional on agent i's signal is then derived from primitives as

$$\Pi(x; s) = \int_{\underline{\theta}}^{\overline{\theta}} \pi(\theta; s) g(\theta \mid x)\, d\theta. \tag{B8}$$

The function $\pi(\theta; s)$ answers: if the true state were known to be θ, how likely would the contract be to publish after agent i signs? The function $\Pi(x; s)$ averages that probability over the states the agent considers possible after seeing signal x. Because both the core $B(\theta)$ and each opponent's signing probability $p_s(\theta)$ rise with the state, $\pi(\theta; s)$ is weakly increasing. Monotone likelihood ratios make the posterior more favorable as x rises, so $\Pi(x; s)$ is weakly increasing in x. A signer's expected payoff is therefore

$$V(x; s, k) = \Pi(x; s) v(x) - k. \tag{B9}$$

This expression has a direct interpretation: probability of publication times value conditional on publication, minus the signing cost.

The following regularity assumption makes best responses and equilibrium cutoffs well behaved. It requires clear low- and high-signal dominance regions, a unique signal at which an agent is indifferent when everyone uses the same cutoff, and no reversals in the sign of that indifference condition.

ASSUMPTION B1 (Finite symmetric-cutoff regularity): *There are nonempty low- and high-state intervals on which $B(\theta) + N < K$ and $B(\theta) \geq K$, respectively. The first condition implies $N < K$: the model describes a marginal fringe that cannot publish without some core participation, while the core can publish by itself in favorable states. For each cutoff $s \in \overline{\mathbb{R}}$, $\Pi(x; s)$ is jointly continuous in (x, s) and $V(\cdot; s, k)$ is continuous and strictly increasing on the relevant signal interval. Uniformly over $s \in \overline{\mathbb{R}}$ and k in the compact interval $\mathcal{K} \subset (0, \infty)$, there are signals $x_L < x_H$ such that $V(x_L; s, k) < 0 < V(x_H; s, k)$. For every $k \in \mathcal{K}$, the diagonal payoff*

$$\Phi_k(s) \equiv V(s; s, k) \tag{B10}$$

is continuous, has a unique root $s^*(k)$, and satisfies $\Phi_k(s) < 0$ for $s < s^*(k)$ and $\Phi_k(s) > 0$ for $s > s^*(k)$.

The prior and signal structure generate Π and ensure that a better signal does not lower the chance of success. The remaining conditions in Assumption B1 control the feedback between one agent's decision and what she expects others to do. The diagonal payoff $\Phi_k(s) = V(s; s, k)$ evaluates an agent whose signal is exactly s when she expects every other marginal agent to use that same cutoff. A root of Φ_k is therefore a self-consistent boundary: the agent at the boundary is indifferent, agents above it prefer to sign, and agents below it prefer not to sign. The assumption requires this boundary to be unique and rules out a payoff that crosses zero, reverses direction, and crosses again. The prior and the existence of a core do not by themselves guarantee these properties. Uniform dominance, a strictly increasing best response, and the unique diagonal crossing are maintained regularity conditions. They play the same formal role as diagonal-crossing restrictions in monotone global-games selection arguments (Frankel, Morris and Pauzner, 2003).

PROPOSITION B1 (Finite cutoff characterization): *Under Assumption B1, the finite marginal-agent game has a unique symmetric monotone cutoff value. Under the maintained boundary convention, marginal agent i signs if and only if $x_i \geq s^*(k)$. Without that convention, cutoff strategies may differ only in the action assigned to the indifferent, zero-probability boundary signal. If $k_2 > k_1$ are in $\mathcal{K}$, then $s^*(k_2) > s^*(k_1)$. Consequently, for every state θ,*

$$\begin{aligned} &\Pr(B(\theta) + Q_N(s^*(k_2)) \geq K \mid \theta) \\ &\qquad \leq \Pr(B(\theta) + Q_N(s^*(k_1)) \geq K \mid \theta)\,. \end{aligned} \tag{B11}$$

The proposition has two economic conclusions. Within symmetric monotone strategies, agents sign when their signals exceed one common cutoff. A higher signing cost raises that cutoff, so a stronger signal is needed to make signing worthwhile. Fewer agents then sign, and the probability of publication cannot increase in any state. This finite result concerns participation and publication within the contract. It does not compare the contract with open expression or model how a signer's exposure changes with coalition size, so it does not prove the article's exposure-tail result for a finite population.

PROOF:

For any conjectured cutoff s, continuity, strict increase, and the uniform low- and high-signal inequalities make the best response another finite cutoff. The agent at its boundary is indifferent. In a symmetric monotone cutoff equilibrium, the conjectured cutoff and the best-response cutoff must be the same, so the marginal signal satisfies $V(s; s, k) = 0$. Assumption B1 gives the unique solution $s^*(k)$, and the weak-inequality convention says that the indifferent boundary agent signs. The proposition does not claim uniqueness among asymmetric or nonmonotone equilibria. It also assumes, rather than derives from the primitive prior and signal structure, that the diagonal payoff has a unique crossing.

Set $s_1 = s^*(k_1)$. Since $k_2 > k_1$,

$$\Phi_{k_2}(s_1) = \Phi_{k_1}(s_1) - (k_2 - k_1) < 0. \tag{B12}$$

The global sign pattern for Φ_{k_2} implies $s^*(k_2) > s_1$. Raising the cutoff weakly lowers the number of marginal signers for every realized collection of private signals. The total $B(\theta) + Q_N$ therefore falls *pathwise*: for each possible realization, not merely on average. An event that was below threshold cannot become successful when the cutoff rises, which gives the state-by-state publication-probability inequality.

Appendix C. Failed-List Privacy, Attempt Observability, and Exchangeability

Failure privacy contains two distinct informational restrictions that are easy to conflate. First, a failed contract does not disclose its signer list or count. Second, treating failure as a public non-event requires the campaign attempt itself to be unobserved by the outside audience whose reaction generates exposure costs. A mechanism can satisfy the first restriction without satisfying the second. In the ideal benchmark, that audience observes

$$o = \begin{cases} \mathcal{C}, & q \geq \tau, \\ \varnothing, & q < \tau, \end{cases} \tag{C1}$$

where $\mathcal{C}$ is the successful named coalition and $\varnothing$ means that the audience sees no list, count, failure announcement, or other indication that a campaign was attempted. Eligible agents may know that they received an offer and may infer after the deadline that publication did not occur. That private inference neither identifies other signers nor creates a public signal for the exposure-cost audience.

If the campaign's existence is publicly observable, a missing publication is an observed failed attempt rather than the public non-event represented by $\varnothing$. Failed-list privacy still protects names, but the audience can revise its beliefs after learning that a campaign was tried and fell short. Public attempt observability therefore defines a second information regime. A data breach or premature release is different: it exposes information that the promised rule was supposed to withhold.

The article also uses an exchangeability assumption to summarize a published coalition by its size: seeing Alice, Bob, and Carol rather than three other eligible people conveys no information beyond seeing a coalition of size three, so coalition size is a sufficient statistic for public inference.

LEMMA C1 (Exchangeable identities): *Suppose that, conditional on the state and the symmetric strategy, all coalitions of a given size have the same likelihood, and that identities carry no payoff-relevant public covariates beyond coalition*

membership. If publication reveals a named coalition $\mathcal{C}$, then coalition size is sufficient for public inference:

$$\Pr(\theta \mid \mathcal{C}, \text{publication}) = \Pr(\theta \mid |\mathcal{C}|, \text{publication}). \tag{C2}$$

If $\mu(y)$ denotes the resulting audience belief at coalition size y, signer exposure costs may consequently be represented as

$$c_i(\mathcal{C}) = \alpha_i \ell\big(|\mathcal{C}|, \mu(|\mathcal{C}|)\big). \tag{C3}$$

PROOF:

Under the stated exchangeability condition, the likelihood of a particular published coalition $\mathcal{C}$ depends on the state only through its size $|\mathcal{C}|$. Bayes' rule therefore gives the same state posterior for every coalition of that size. The cost representation follows from the additional restriction that audience response depends on the coalition only through its size and the induced belief.

Lemma C1 says when coalition size contains all the information that the audience learns from the identities on a published list. It does not show that the article's selection theorem continues to hold when the audience changes its beliefs as the coalition grows. That theorem holds μ fixed so that a larger coalition affects exposure directly through safety in numbers, not indirectly through a changing audience inference. If audience beliefs are instead written as $\mu(y)$, effective exposure becomes $\ell(y, \mu(y))$. Coalition size then works through both channels at once. Applying the theorem would require this combined function to remain bounded, continuous, and strictly decreasing, and would also require the dominance and single-crossing conditions to continue to hold. Exchangeability alone provides none of these additional properties. The appendix therefore makes no general selection claim for an audience whose response changes with coalition size.

Names still determine which agents become publicly associated with the statement and therefore who bears exposure. Exchangeability removes composition effects among coalitions of the same size. When signer attributes affect public inference, protection, or retaliation, coalition composition becomes an additional payoff-relevant state variable.

Appendix D. Implementation with Existing Tools

A fixed-population social assurance contract requires four operational functions. It must verify that each eligible person submits at most one endorsement, keep names and signatures confidential during collection, determine whether the threshold has been met, and release the complete authenticated list only after success. Existing cryptographic systems already perform close versions of each function. A proof-of-concept threshold secret petition accepts encrypted signatures from validated users, prevents duplicate signatures, divides decryption control among trustees, and makes the signatures readable after the threshold is

reached (Breuer, 2024). Secure allegation escrows divide confidential allegations and identities among independent parties and reveal them after designated matching thresholds are reached; these systems include formal security analyses and prototype implementations (Arun et al., 2020). Easwar Vivek Mangipudi, Donghang Lu, Alexandros Psomas and Aniket Kate (2023) develop a related threshold information escrow that deters trustee collusion. Each required function therefore has a close precedent. None of the cited systems implements and evaluates the exact combination described below under one set of assumptions about possible attacks and failures. A production system would still need to integrate the components, undergo a security audit and usability testing, and operate under clear governance rules.

D1. A Feasible Architecture

The least technical implementation delegates every function to an independent custodian such as a law firm, regulator, union, or ombudsperson. The custodian verifies participants against a roster, stores the list under access controls, and follows an audited release and destruction rule. Professional duties, contractual penalties, independent audits, and reputational capital can make this arrangement credible using ordinary institutional tools. Its tradeoff is concentrated trust: one organization can inspect, alter, suppress, or prematurely disclose the list.

When that concentration of trust is unacceptable, an alternative design separates five jobs. The *sponsor* fixes the contract but never receives submissions. A *registrar* verifies real-world eligibility. A *collection service* checks and counts encrypted submissions. An independent *auditor* resolves disputes about whether a valid submission was included. Finally, m independent *trustees* control whether submissions can be opened. The trustees jointly generate an encryption key used only for that campaign, and each retains only one share of it; no central operator ever possesses the complete key. At least t trustees must cooperate to reconstruct the power to decrypt, where $1 < t \leq m$. Existing institutions could fill these roles. Assigning the jobs to different institutions avoids giving one actor both the identities and the power to release the list.

Each cryptographic tool has a distinct job. A digital signature is the electronic counterpart of signing a particular document: it proves that the holder of a particular key endorsed an exact statement and that the statement has not since changed. Encryption turns the signed record into unreadable ciphertext. Threshold decryption divides the power to reverse that encryption among several trustees, so no trustee can reveal a record alone. A zero-knowledge proof lets the collector verify a fact about the encrypted record—for example, that it contains an eligible identity and a valid signature—without learning the hidden identity or signature. These tools answer different questions: authenticity, secrecy, shared release control, and verification without disclosure.

1) Before recruitment, the sponsor fixes the exact statement, an objectively

verifiable eligibility roster or rule, the endorsement threshold K_N, the closing time, and the success and failure disclosures. It also fixes any cancellation and emergency rules. A digital signature records assent to those terms; encryption makes the signed record unreadable during collection. Neither tool substitutes for the other.

2) The registrar verifies each member of $\mathcal{I}_N$ once and issues a cryptographically verifiable eligibility credential. The credential certifies either her identity or a public signing key that only she controls and that the registrar has linked to her identity. Registration can occur before any campaign so that approaching the registrar does not itself reveal an intent to sign. An anonymous credential lets a person prove eligibility to the collector without identifying herself at that stage; established systems also support one-use credentials and later opening under stated conditions (Camenisch and Lysyanskaya, 2001).

3) A participant signs the exact statement and a unique identifier for this contract with her identity-bound key. She submits an encrypted record containing that signature, her certified identity, and the identifier. She also submits a campaign-specific serial number and a zero-knowledge proof, which is evidence about hidden information that does not reveal the information itself. The proof establishes that the encrypted identity or key is the one certified by the registrar and that the record contains a valid signature on this contract. Anonymous-credential and private-contract systems supply the component proof techniques for these checks (Camenisch and Lysyanskaya, 2001; Kosba et al., 2016). The credential always produces the same serial number when used twice in this campaign but produces an unrelated number in a different campaign. The collector can therefore detect a second use without learning the participant's identity or linking her actions across campaigns. Coconut implements this eligibility and duplicate-prevention pattern in an electronic-petition application (Sonnino et al., 2019).

4) The collection service verifies the proof, rejects a repeated serial number, and records each accepted encrypted submission in a private authenticated log. Authentication makes later deletion or alteration detectable. The service returns a signed inclusion receipt. A participant who receives no receipt, or whose receipt is omitted from the count, can challenge the omission privately before an independent auditor. This makes suppression contestable without publishing a failed participant's name or action.

5) The committed core uses the same authenticated submission process, and its accepted submissions count toward K_N. Core participation is independent of the marginal participants' choices. Consistent with the finite benchmark, the service reveals no interim core or total count to marginal participants. At the deadline, the collector and trustees audit the accepted log. If no

operator should learn the count, secure multiparty computation can test whether private inputs reach K_N without giving all inputs to one party (Evans, Kolesnikov and Rosulek, 2018).

6) At the preannounced close, each trustee verifies the integrity and count of the same private accepted log, and the auditor resolves any timely challenge backed by an inclusion receipt. If the accepted count reaches K_N, at least t trustees sign a success certificate binding together the contract identifier, deadline, final count, and a cryptographic fingerprint of the exact accepted ciphertext set. A cryptographic fingerprint is a short value that changes if that set is altered, so the certificate identifies the precise encrypted records the trustees approved. The same quorum then combines its decryption-key shares. The resulting publication contains every accepted identity and signature, and any reader can verify both the registrar's identity binding and the participant's signature on the contract. The endorsement threshold K_N and trustee quorum t are different parameters: the first determines whether publication is authorized; the second determines how many custodians must cooperate to make the records readable. If the count falls short, the institution publishes nothing, the collector deletes the identity-bearing records, and the trustees erase the campaign-specific key shares. Failure privacy assumes that fewer than t trustees collude before that erasure and that a later coalition of t trustees has not retained copies of its shares.

This reference workflow creates no public transaction, log entry, or counter for an individual submission, and it issues no public failure notice. Only success produces a public certificate and named list. It matches the article's failure observation $\varnothing$ only if network traffic and recruitment also conceal the attempt from the audience that could punish signers. The requirements for hiding those traces are discussed below.

Several familiar technologies can replace parts of this workflow. Threshold cryptography divides decryption control so that no one trustee can open a record, while a specified quorum can combine its shares. Secure allegation escrows demonstrate distributed conditional revelation (Arun et al., 2020), and Easwar Vivek Mangipudi, Donghang Lu, Alexandros Psomas and Aniket Kate (2023) show how to deter premature trustee collusion. A smart contract is program code that a blockchain executes and records according to precommitted rules. Privacy-preserving voting systems show that such contracts can enforce eligibility, proof checks, stages, and tallying for a fixed group (McCorry, Shahandashti and Hao, 2017). More general systems use encrypted information and zero-knowledge proofs to enforce private contract logic without placing the underlying information in clear text on a public ledger (Kosba et al., 2016).

Public-counter variant. — A public ledger is a replicated record that many computers maintain and that is designed to make past entries hard to alter. It

can hold encrypted records or cryptographic fingerprints of them. Encryption can keep record contents secret, and fingerprints make alteration detectable. A fingerprint does not contain the underlying record, however, so copies stored elsewhere are still needed to ensure *availability*: the ability to retrieve the record when authorized. Public per-submission records also reveal that a campaign exists and may reveal its running count. Such a design implements failed-name privacy, but not the article's stronger public non-event unless both submission traffic and the count are hidden through fixed-size batches, decoy traffic, or another privacy layer. Putting identity-bearing ciphertexts on an immutable ledger also leaves permanent encrypted material that a future key compromise could expose. For the benchmark-compatible workflow, those ciphertexts therefore remain off the public ledger; a ledger may instead record the precommitted terms and, after success, the certificate and released list. Any public-ledger variant additionally relies on the security and availability of its consensus system, the rules by which participating computers agree on the ledger's valid state.

Separating these roles limits what any one actor can learn or do. The registrar can know who is eligible without learning who submitted. The collector can verify that submissions are unique without reading them, and no single trustee can open a failed list. The sponsor cannot change the statement or threshold after observing participation. Submission receipts let a private auditor detect omitted entries or cases in which the collector showed different participants different versions of the log. No component eliminates the need for reliable operation. The design still assumes that credentials are issued correctly or audited, participant devices and software work as intended, and all trustees use the same log. It also requires at least t trustees to remain available after success, fewer than t trustees to collude before success, and key shares to be erased after failure. Engineering controls, audits, and governance rules can support these conditions, but they must be evaluated for each deployment.

D2. Threats and Mitigations

Eligibility and malicious participation. — A Sybil attack occurs when one actor presents several identities and is therefore counted several times (Douceur, 2002). A fixed eligibility roster makes this problem much easier than proving that every account on the open internet belongs to a different person. The sponsor freezes the roster rule, the registrar gives one credential to each verified eligible person, and campaign-specific serial numbers let the collector reject repeated use. Auditable issuance, strong authentication, revocation for stolen credentials, and separation of registration from collection address the main risks. These controls prevent impersonation and duplicate counting; authentication verifies identity, not motive. An eligible saboteur may sign in order to push the list over the threshold and expose everyone else, so strategic sabotage remains a separate design risk.

Custody, collusion, and availability. — A centralized custodian can leak names, invent or discard entries, suppress a successful list, or yield to coercion. Distributed trustees, signed submission receipts, a log whose old entries cannot be silently rewritten, and a quorum release rule divide those powers. Distribution creates two opposing risks. A sufficiently large group of trustees may cooperate to open records early, while too few available trustees may prevent release after success. Choosing trustees from institutions with different interests, setting a quorum that tolerates some failures, protecting key shares, and precommitting legal duties and penalties reduce both risks. Collusion-deterrent escrow protocols show how the technical design can also make early cooperation against the rules less attractive to trustees (Easwar Vivek Mangipudi, Donghang Lu, Alexandros Psomas and Aniket Kate, 2023).

Software, keys, and governance. — A system built from sound cryptographic tools can still fail because of compromised devices, poor key storage, code defects, or an administrator's power to change the software. A participant's device might reveal her action or trick her into signing a substituted statement; a bug might reject valid submissions or apply the threshold incorrectly. Conservative software design, independent security audits, open-source review, hardware-backed or multiparty key custody, and clear recovery procedures can reduce these risks. The statement, roster rule, threshold, deadline, and release policy should be cryptographically tied to the contract identifier and should not be alterable by one administrator after collection begins. Decentralized-finance smart contracts held about $150 billion in total value as of April 2022, showing that related infrastructure operates at large scale (Werner et al., 2022). Attacks on these systems have also exposed recurring failures in design, governance, dependencies, and implementation (Hadis Rezaei, Mojtaba Eshghie, Karl Anderesson and Francesco Palmieri, 2025). Those failures provide a concrete list of risks for system design and audit.

Metadata and public observability. — Encryption hides the contents of a submission, not necessarily its *metadata*: where it came from, when it was sent, or the fact that it exists. If every encrypted submission is a public transaction, observers may learn the interim count and perhaps link a transaction to a participant even though they cannot read the encrypted contents. Grouping submissions into batches, routing them through intermediaries, delaying their recording, and collecting them privately rather than on a public chain can reduce those disclosures. A system concerned only with failed-name privacy may tolerate a public encrypted count. Implementing the article's stronger observation $\varnothing$ requires that the exposure-cost audience also see no count, failure announcement, or evidence that the campaign was attempted. The technical architecture can protect names without assuming that every deployment hides the attempt itself.

Legal and institutional controls remain useful throughout. Contracts can penalize misconduct, audits can expose deviations from the release rule, and administrators can build reputational capital by operating transparent open-source systems. These safeguards complement cryptography. The distributed design greatly reduces the authority entrusted to any one party: it replaces trust in one custodian with narrower, auditable assumptions about credential issuance, software, logs, and a trustee quorum. No single person or organization receives readable access to submissions, unilateral control of the count, and unilateral release authority.

D3. Networks and Deployment

Four networks matter for deployment, and they need not contain the same people or follow the same links. The *recruitment network* determines which eligible people learn of the contract. The *trust network* determines whether recipients believe the invitation and the custody arrangement. The *exposure network* identifies the audience able to impose material costs after names become public. The *observation network* determines whether outsiders learn that a campaign was attempted or failed.

Dense trusted groups can supply a plausible committed seed. An organizer can recruit a protected or already-public core through existing professional, organizational, or social ties before reaching more distant eligible participants. Bridges across departments or social circles determine who can be reached. In the model, these network features map into the seed $b(\theta)$, participating capacity $Y(\theta)$, and threshold reachability. Fragmented recruitment lowers capacity even when latent support is ample.

The exposure network answers the practical question of who can punish a signer. In an insular organization, supervisors or close peers may be the relevant audience; in a public campaign, employers, professional associations, or a wider political audience may matter. The homogeneous exposure benchmark applies when protection depends mainly on final coalition size. Network position, seniority, and subgroup membership instead map into the heterogeneous protection technology developed in the companion paper.

Finally, private recruitment and public attempt observability are separate design choices. Trusted invitations can spread through the recruitment network while the outside exposure audience sees no campaign. Public advertising or a visible progress counter may improve reach, but it also turns a failed campaign into a public event and can reveal how close the threshold came. The benchmark therefore describes a specific configuration: private recruitment of a credible seed, enough potential participation to reach the threshold in the state being evaluated, a seed that can clear it in favorable states, and an information policy that keeps failed lists and attempts from the exposure audience.

Implementation burden and distrust can raise the signing friction k by making participation harder or less attractive. Premature or partial disclosure is more

serious: it exposes signers before the coalition is large enough to protect them and violates the institution analyzed in the paper. A working system must provide the eligibility, custody, verification, and release functions described above under assumptions suited to the attacks and failures it may face. Existing institutional and cryptographic tools make such a system plausible. The security of a particular deployment depends on how those tools are combined and operated.