Failure Privacy and Safe Collective Expression with Social Assurance Contracts

By Matthew Cashman*

July 16, 2026 (latest version here)

Controversial views sometimes remain unspoken because they invite retaliation. However, a sufficiently large group could speak safely if only they spoke together. Speaking one-by-one may encourage others, but retaliation against early participants can stop the cascade before a protective group forms. A social assurance contract privately collects signed commitments and publishes them only when all signers will be safe. I show that such contracts can tunnel beneath this exposure barrier to safe coalitions the public speaking cascade cannot reach. Doing so requires private commitments and joint publication.
(JEL D71, D82, D83, J51) Keywords: Assurance contracts, disclosure design, failure privacy, coalition formation, collective action, retaliation

Many people hold views they will not share publicly if they must do so alone. Their views may fall outside the range of publicly acceptable opinion, be opposed by an employer, a regulator, or a regime, or concern allegations that invite retaliation. A person may be wrong about how many others agree with her. But even when she is not, being publicly identified as a supporter, signer, complainant, witness, dissident, or accuser can be dangerous unless enough others are identified at the same time.

This paper studies that problem as one of safe coalition formation. Classical accounts link beliefs about support to social sanctions: people may misperceive how many agree, while fear of isolation or retaliation makes public expression depend on those perceptions (Kuran, 1997; Noelle-Neumann, 1974). Third parties punish norm violations even when they were not directly harmed (Fehr and Fischbacher, 2004), and such sanctions can sustain inefficient norms (Bendor and Swistak, 2001). Information matters here too. But even if everyone knows that substantial private support exists, the supporters still have to make themselves public without leaving anyone exposed in a dangerously small group.

The exposure at issue is material retaliation: discharge, prosecution, blacklisting, harassment, or other punishments borne by identifiable individuals. A larger group can dilute or prevent these punishments, so a person's payoff from publishing a statement under her name depends on who is named alongside her. The *roster* is this set of named authors. Information may tell potential signers how many others would join, while a legal reform or a change in audience beliefs may alter which rosters are safe. The remaining problem is assembly. A large roster might protect everyone once it exists, yet be impossible to build by adding names publicly one at a time: the first few people would still face retaliation. Holding signatures until the whole coalition can speak together removes this path constraint. A lone raised hand can be punished, while a thousand raised together may be safe.

The distinction is whether signatures are visible as they accrue or held until release. It recurs in public statements, organizing, dissent, and misconduct reporting. An open letter on a controversial question can be assembled in either way. A publicly circulated version reveals signatures one at a time and stalls when the next signer would stand to lose too much. An embargoed version keeps each signature private from the retaliating audience and publishes the whole signed letter once the list is large enough to protect every signer; if it never meets that threshold, the signers' identities

* MIT Sloan, 100 Main Street, Cambridge, MA, USA. Email: cashman@mit.edu.

are never published to that audience.

On success, both procedures deliver the same signed letter to the same readers, so the counterfactual is procedural rather than informational. They differ while signatures accumulate and if the campaign falls short of its safety threshold. Changing when identities become known to the audience that can retaliate can therefore enlarge the set of coalitions that become public safely without changing the information conveyed on success.

To isolate the path constraint, the comparison holds fixed a finite eligible population, the people willing to participate, the audience able to retaliate, and each person's protection under every published roster. A legal reform or a change in audience beliefs creates a different protection environment. The paper compares what visible and hidden signature collection can achieve within the same environment.

For private release, the analysis also conditions on a realized set V of valid completed authorizations. Apart from the frictionless participation ceiling in Section III, the paper does not model formation of V under private information or positive signing friction.

Cascade versus tunnel.

I call a named coalition *self-protecting*, or simply safe, if every member is protected by exactly the membership that goes public. Under safety in numbers, the union of two self-protecting coalitions is also self-protecting, so there is a unique largest safe coalition. Public expression from committed early movers also has a unique stopping point, where no additional person can safely join alone. The question is whether an institution must expose people along the way or can collect signed permission to publish before making any signer public.

I give decentralized expression and the disclosure mechanism the same safety requirement: *regret-free safety* (robust ex post safety in mechanism-design terms). This means that no participant is ever named in a coalition too small to protect her, even in an unfavorable realization in which other intended participants fail to appear. Two further assumptions govern coalition formation. First, under safety in numbers (monotone protection), adding public participants cannot make an existing member less safe. Second, when people try to go public through separate acts, any one person's statement may appear without the others. This is unilateral-completion uncertainty. It concerns unverified public attempts, not signatures already deposited with a contract. The strict safety requirement reflects harms that are borne individually, cannot be contracted away, and may be compensated only incompletely and after a delay.

Between silence and safe collective expression stands the *exposure barrier*: a middle range of coalition sizes at which anyone visible is unsafe. In decentralized public expression, each participant must therefore be safe even if her statement is the only attempted statement that appears. She can move safely only when the people already public, together with her own name, protect her. The "cascade" advances one visible name at a time and stops at the first self-protecting coalition from which no further signer can join safely on her own. The exposure barrier is where it stops.

A *social assurance contract* instead gathers signed statements privately. Each signed statement is a completed authorization for the contract to publish it under the signer's name. The contract publishes the authorized statements together only when the resulting coalition protects every member. Safety is therefore required at the *destination*, not at each step along the way. An administrator, delegate, or protocol receives the signatures and keeps them hidden from the retaliating audience until release; the holder may know who signed. A companion paper studies entry and equilibrium selection in a homogeneous population, the effects of signing frictions and disclosing the entire signer list after failure, and implementation (Cashman, 2026).

Among the people who sign, the contract can implement the largest safe coalition. Where the public path is blocked by an exposure barrier, a social assurance contract tunnels under the mountain (Figure 1). Open expression is constrained by safe paths, while the contract is constrained

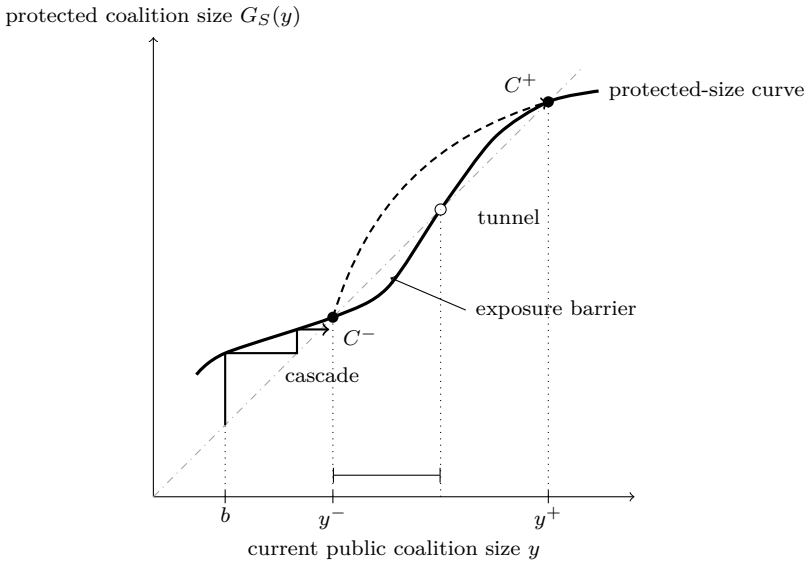

FIGURE 1. CASCADE VERSUS TUNNEL. THE HORIZONTAL AXIS y IS THE COALITION SIZE ALREADY PUBLIC. THE VERTICAL AXIS $G_S(y)$ IS THE SEED PLUS THE WILLING SUPPORTERS WHO WOULD BE PROTECTED AT THAT SIZE. CROSSINGS WITH THE DIAGONAL ARE FIXED POINTS: GROWTH RESUMES FROM THE HOLLOW CROSSING BUT NOT FROM THE FILLED CROSSINGS. STARTING FROM SEED b, OPEN EXPRESSION REACHES THE LOWER CROSSING y^- AND STALLS AT THE EXPOSURE BARRIER. THE SOCIAL ASSURANCE CONTRACT INSTEAD REACHES THE LARGEST SAFE COALITION AT y^+, ADDING COALITION SIZE $y^+ - y^-$. SECTION I DEFINES THE OBJECTS; APPENDIX S6 TREATS FINITE AGENTS.

by safe destinations. The contract moves the first-mover problem from public naming to private *signing*, where signatures can remain hidden. In the frictionless benchmark, a person never does worse by signing than by abstaining, and everyone signing is an equilibrium; this is the sense in which full signing is selected by dominance (Proposition 4). Signing costs, leakage, and distrust determine how much of that ceiling is reached (Section III; Cashman, 2026).

Failure privacy means that completed signatures remain private until a protecting group can be published together. Proposition 1 shows separately why privacy, completion, and conditionality each matter; joint publication is the release event that supplies protection. The name emphasizes that identities remain hidden from the retaliating audience when the campaign fails. The same technology also hides identities while the coalition is forming, which is a separate source of protection.

MAIN RESULTS.

Theorems 1–3 compare the two ways of assembling a coalition. Theorem 1 identifies the largest coalition that would protect all its members if published at once. Theorem 2 identifies the smaller coalition that can be built by safe public additions. Theorem 3 shows that conditional release reaches the larger coalition among people who have signed. The difference, $C^+ \setminus C^-$, is the set of *hidden supporters*: people who belong to the largest safe coalition but cannot join through the safe public path.

Theorem 4 gives the converse. Instead of asking what the contract can do, it asks what any zero-risk institution must do to move beyond the public cascade without exposing someone in an unsafe roster. When public acts complete individually, the relevant audience is fixed, and completion determines whose identities become known, the institution must first receive signed permissions while their authors remain private. It then publishes the protecting group in one act. Every named

participant other than the author of that release act must have authorized publication in advance. Lemmas 3–4 derive this structure from the act model in Section II.

The remaining results ask what changes when participants will accept some risk. Theorem 5 shows that, when delay is costless and requests can be repeated, privately collecting completed authorizations lowers exposure risk without reducing the final protected coalition. Propositions 2–3 show how a large public pact, or leakage from a private stage, changes that comparison. Theorem 6 explains why safely able public speakers act rather than wait, and Proposition 4 gives the corresponding benchmark for private signing. Section IV adds the fight-or-fold implication: making silent supporters visible can raise observed retaliation while reducing suppression.

Contribution and relation to the literature.

The paper combines fixed-point methods with a common safety requirement and explicit timing of public attribution. When the possible coalitions are ordered by inclusion, a monotone rule has smallest and largest fixed points (Topkis, 1998); threshold cascades grow from a seed to the smallest fixed point they can reach (Granovetter, 1978; Schelling, 1978); and in supermodular games adaptive dynamics eventually remain between the smallest and largest equilibria (Milgrom and Roberts, 1990). Requiring every completed public step to be safe turns the lower fixed point into a bound on public expression, not merely an equilibrium selected by learning.

Work on dissent emphasizes informational cascades, higher-order knowledge, and strategic control of information (Lohmann, 1994; Chwe, 1999; Edmond, 2013). Here latent support may already be known. Beliefs determine which rosters are safe; the institution determines whether a safe roster can be assembled without exposing its first members. Assurance contracts and provision-point mechanisms make monetary contributions conditional on aggregate participation and can return them after failure (Bagnoli and Lipman, 1989; Tabarrok, 1998; Zubrickas, 2014). Related commitment problems appear in contingent contracting and dynamic contribution games (Segal, 1999; Admati and Perry, 1991). The problem here differs from a threshold public good with refunds: a failed monetary contribution can be returned, while an identity-bound expression is irreversible and a failed public attempt may itself cause the harm. A refund cannot unname a complainant or employee, so the mechanism must control attribution rather than payment. Section VI develops the two closest analytical comparisons, with Gale (2001) and Braghieri, Bursztyn and Fasnacht (2026).

Appendix A contains the proofs of all results stated in the article body. The Supplemental Appendix contains extensions and auxiliary results.

I. The Model: Safe Coalitions, Cascade, and Tunnel

The model has three pieces. It records which people are willing to speak if protected, which public rosters would protect each person, and which of those rosters can be assembled without exposing anyone along the way. The first two pieces describe safe destinations. The third distinguishes the public cascade from the private tunnel.

Let N be a fixed finite set of potential supporters and let $S \subseteq N$ be the *consent set*: the people willing to publish a specified expression if the released coalition protects them. I call these people agents only to emphasize that each makes a choice. Consent is not yet a signature or commitment. Whether a consenter actually signs is studied in Section III.

A *seed* of mass $b \geq 0$ is already public and moves regardless of protection. It represents organizers or early speakers who accept exposure without the model's guarantee. Each person $i \in S$ has a positive mass a_i, meaning the numerical weight she adds to a coalition; if all people count equally, one can set every $a_i = 1$. The total mass of a coalition C is $m(C) = \sum_{i \in C} a_i$. The seed enters every safety calculation but is omitted from the roster notation C.

For each person i, the *protection family* $\mathcal{P}_i \subseteq \{C \subseteq S : i \in C\}$ is simply the collection of public rosters that leave her at least as well off speaking as remaining silent. Because retaliation is borne by identifiable individuals, safety is evaluated when both the statement and its signer become public. Thus C records whose specified statements are public, not merely how many supportive statements exist. The protection families and the retaliating audience remain fixed while a coalition forms. A legal reform, a change in audience beliefs, or a different opponent creates a new protection environment.

The seed gives public organizing a place to start. Without one, public expression cannot begin when no person is safe alone, even though a sufficiently large group could protect all of its members. The tunnel can still operate in that seedless case because it can release the group together. This is the sharpest example of the difference between a safe destination and a safe path.

ASSUMPTION 1 (Monotone protection): *For all i and all C, C' with $C \in \mathcal{P}_i$, $C \subseteq C'$, and $i \in C'$: $C' \in \mathcal{P}_i$.*

Monotone protection is safety in numbers: additional public co-participants never make an already-named member less safe. Section IV derives it from a fight-or-fold opponent facing a coalition whose resilience, the probability that it survives retaliation if attacked, weakly rises with size. Appendix S8.1 analyzes protection technologies in which an added name can make others less safe.

The assumption is about the realized roster, not merely its size. A senior employee, corroborating witness, or legally protected organizer may add more protection than another name of equal mass. The scalar threshold model below is therefore a benchmark, while the set-valued protection families carry heterogeneous coalition composition.

Monotonicity lets the mechanism admit new signers without rechecking the safety of everyone already included. Once a roster protects its members, accepting another authorized name cannot invalidate their protection. Without that property, the holder must re-evaluate every incumbent after each proposed addition, and prospective signers need roster-specific information to decide whether the resulting group still protects them. Those calculations become difficult in a large, open-ended campaign. Monotonicity fits reports matched on the same accused when another screened report corroborates and organizing against a capacity-constrained opponent when another participant dilutes targeting. Appendix S8.1 studies the nonmonotone cases created by stronger opponents, strategic allegations, and guilt by association.

Without monotone protection, there may be no largest self-protecting roster. Let $S = \{1, 2, 3\}$ and take

$$\mathcal{P}_1 = \{\{1, 2\}, \{1, 3\}\}, \qquad \mathcal{P}_2 = \{\{1, 2\}\}, \qquad \mathcal{P}_3 = \{\{1, 3\}\}.$$

The rosters $\{1, 2\}$ and $\{1, 3\}$ are both self-protecting, but neither contains the other, and their union is unsafe. In this sense the two safe rosters are incomparable: there is no single larger safe roster that combines them. After all three agents sign, the administrator must choose whether to exclude agent 2 or agent 3. The greatest-roster implementation in Theorems 1–3 therefore fails.

The distinction between a safe path and a safe destination still survives. Under the stronger batch-admissibility condition defined in Appendix S8.1, the *accessible kernel* is the collection of safe rosters that public organizing can actually reach while remaining safe after every completed batch. Proposition S8 shows that a safe roster outside this reachability set requires conditional release. Without a greatest roster, however, the administrator must choose among several safe outcomes that cannot be combined. Choosing the best one requires an application-specific objective and is generally computationally intractable (Proposition S9).

DEFINITION 1: *A coalition $C \subseteq S$ is* self-protecting *if $C \in \mathcal{P}_i$ for every $i \in C$. (The empty coalition is vacuously self-protecting.)*

The scalar special case reduces each person's protection family to one number, her protection requirement ρ_i. Person i is safe in coalition C exactly when the total public mass, including the seed, reaches that number: $C \in \mathcal{P}_i$ iff $b + m(C) \geq \rho_i$. The function $H_S(y) = m\{i \in S : \rho_i \leq y\}$ adds up the mass of all consenters whose requirements are no greater than public mass y.

For the graphical benchmark, agents form a continuum, so each individual has negligible mass, and H_S is continuous over the relevant interval. Continuity means that the protected mass does not jump when y changes by an arbitrarily small amount. It is a regularity condition on the distribution of requirements, separate from the assumption that no one individual has positive mass. A self-protecting coalition that has absorbed every person protected at its size satisfies

$$y = b + H_S(y), \tag{1}$$

the threshold-cascade equation of Granovetter (1978). The left side is the public mass that exists; the right side is the seed plus everyone who would be protected at that mass. Equality is therefore a point at which the coalition reproduces itself: no currently protected person is missing and no further person becomes protected. Such a self-consistent value is called a *fixed point*.

Figure 1 graphs the right side as $G_S(y) = b + H_S(y)$. Its crossings with the diagonal are the fixed points. Write $y^- = b + m(C^-)$ and $y^+ = b + m(C^+)$ for the smallest and largest crossings. A crossing is locally stable when a small increase above it does not trigger further growth, which occurs when G_S lies below the diagonal just above the crossing. With finitely many weighted people, entrant i contributes her own mass when she joins, so she is safe when $b + m(C) + a_i \geq \rho_i$. The public cascade is therefore governed by the lower activation threshold $\rho_i - a_i$, while the safety of a coalition released together is evaluated using the original requirement ρ_i (Appendix S6).

Two mathematical operators keep the path question separate from the destination question. An operator is simply a rule that takes one candidate roster as its input and returns another roster. The safe-expansion operator

$$T(C) = C \cup \{\, i \in S \setminus C : C \cup \{i\} \in \mathcal{P}_i \,\}$$

adds exactly the agents who become safe given the currently public set C: one round of public expression. The static-safety operator

$$\Gamma(C) = \{\, i \in S : C \cup \{i\} \in \mathcal{P}_i \,\}$$

collects everyone who would be safe if added to C: the static safety check. For $i \in C$, $C \cup \{i\} = C$, so C is self-protecting iff $C \subseteq \Gamma(C)$. It is a fixed point, $C = \Gamma(C)$, when it is self-protecting and no additional person would be safe to join.

The two operators answer different economic questions. Γ asks which names would be safe if a candidate roster could appear at once; its greatest fixed point is a destination test. T asks who can safely become public given only the realized public past; iterating it is a path test. Because a new participant may complete alone, T tests her against the existing public coalition plus her own name. A technology that guarantees joint completion changes this path test. That difference allows two institutions facing the same preferences and consent set to reach different coalitions.

LEMMA 1 (Monotonicity): *Under Assumption 1, $C \subseteq C'$ implies $\Gamma(C) \subseteq \Gamma(C')$ and $T(C) \subseteq T(C')$, and T is inflationary ($C \subseteq T(C)$).*

The word "inflationary" in Lemma 1 means that T never removes anyone. Starting from no nonseed participants, apply T once to add everyone who can safely move first, apply it again to add everyone newly protected by that first group, and continue. Because S is finite and each round only adds people, the sequence $\varnothing \subseteq T(\varnothing) \subseteq T^2(\varnothing) \subseteq \cdots$ stops changing in at most $|S|$ rounds.

Its endpoint is the *cascade closure* $C^-(S) = T^{|S|}(\varnothing)$. By contrast, define the largest self-protecting coalition by looking across all safe rosters at once:

$$C^+(S) = \bigcup \{\, C \subseteq S : C \text{ is self-protecting} \,\}.$$

A small example shows the difference. Because each finite entrant's own mass counts, the cascade uses the activation thresholds $\rho_i - a_i$ defined above. Take seed $b = 1$, four unit-mass consenters, and requirements $\rho = (2, 3, 5, 5)$. The first agent joins because public mass becomes $b + a_1 = 2 \geq 2$. The second then joins because public mass becomes $3 \geq 3$. The cascade stops there. A third entrant would require public mass 5, but only 4 would be public after she joined. Thus $C^- = \{1, 2\}$, with coalition mass $m(C^-) = 2$ and public mass $b + m(C^-) = 3$. The last two agents are safe only if they appear together, at public mass $5 = b + m(S)$. Hence $C^+ = S$. The two agents in $C^+ \setminus C^-$ cannot cross the exposure barrier one at a time, but the tunnel can publish their statements jointly.

THEOREM 1 (Largest Safe Coalition): *Under Assumption 1, the union of any family of self-protecting coalitions is self-protecting; hence $C^+(S)$ is the unique largest self-protecting coalition. Any consent-respecting disclosure rule—one that publishes no name without that agent's authorization—that satisfies regret-free safety (formalized in Section I.A) releases a subset of $C^+(S)$.*

Theorem 1 separates three questions. Which roster is safe? Can that roster be reached through public additions? Will the relevant people sign? The theorem answers only the first: under monotone protection, all safe rosters can be combined, so their union C^+ is itself safe and contains every other safe roster. The path results answer the second question, and the entry results answer the third.

A. *Regret-Free Safety*

Both institutions—open expression and conditional disclosure—face the same safety requirement; they differ in whether a completed authorization can be received before its author becomes public.

Write r_i for the exposure risk person i is willing to tolerate. An *admissible realization* is any pattern of completed and failed actions that the institution promises to handle. At $r_i = 0$, person i requires safety in every such pattern, however unfavorable. At $r_i > 0$, she accepts some chance of unsafe exposure in return for the benefit of speaking. Later, r_i equals the ratio of her expressive benefit to the harm from unsafe exposure. This section begins with zero tolerance; Section II then allows positive tolerance.

Zero risk can also be an institutional promise rather than a claim that every participant ranks safety above every possible benefit. An escrow, counsel, or organizer may undertake not to expose a depositor on a failed or under-protective path; participants facing irreversible harm may demand that guarantee rather than accept an uncertain chance of exposure. The theorem describes that service standard. Institutions and participants willing to trade a small exposure probability for reach belong to the $r_i > 0$ comparison.

Completion and verification.

Time is divided into dates $t = 1, 2, \ldots$. The set C_t contains everyone public by the end of date t, so irreversibility gives $\varnothing = C_0 \subseteq C_1 \subseteq \cdots$. At date t, the set $A_t \subseteq S \setminus C_{t-1}$ contains the people who attempt to go public. Some attempts may fail: $R_t \subseteq A_t$ is the subset that actually completes, and the new public set is $C_t = C_{t-1} \cup R_t$. A co-participant is *verified before exposure* when the institution has already received and certified her completed authorization rather than relying on her stated intention. Verification establishes that an eligible person authorized a particular statement and release rule; it does not establish that the statement is true. The assumptions below specify which patterns of completion and failure the institution must withstand.

ASSUMPTION 2 (Unilateral-completion uncertainty): *Unless the institution has verified co-participants before exposure, each attempted public action may complete without the others: for every $i \in A_t$, the realization in which i completes and no other unverified exposure completes (in particular $R_t = \{i\}$) is admissible.*

The assumption distinguishes trusted intentions from completed facts. In a rally, walkout, or scheduled post, one person's exposure can land while the others do not. A holding institution changes that technology by receiving completed authorizations before exposure. Correlated completion can reduce the risk of being left alone; exact joint completion is characterized in Lemma 4, while the positive-tolerance comparison allows imperfect coordination.

The assumption concerns possible completion patterns, not their relative probabilities. An institution promising perfect safety must withstand the outcome in which only one attempted action is completed. A public promise, poll response, or plan to move together leaves that outcome possible because none is a completed co-participation act. A platform that first receives completed authorizations removes it at release; Section II proves that doing so requires a holding stage.

One monitoring technology generates the assumption directly. An individual observes whether her own post, walkout, filing, or signature has completed only as it becomes attributable; before then she cannot certify that anyone else's separate act has completed, and after then her own exposure is irreversible. The intentions of the people planning to act with her may be sincere and highly correlated, but they are not collateral against the event in which her act lands and theirs do not. A device that receives completed authorizations before executing a release changes the technology rather than the agents' beliefs. It converts intended joint action into a fact on which exposure can be conditioned.

Organizers can test the room before acting. Deniable inquiries and polls change beliefs, repeated interaction creates trust, and costly pledges discipline later action. These channels change the protection families, reduce unilateral-completion risk, or create completed commitments. A zero-risk promise requires a channel that eliminates the realization in which one participant is stranded.

ASSUMPTION 3 (Regret-free safety): *For every date t, every admissible realization path, and every $i \in C_t$: $C_t \in \mathcal{P}_i$.*

No consenting agent is ever named in an under-protective coalition. Behaviorally, this is the $r = 0$ endpoint; institutionally, it is the disclosure guarantee just described. Assumption 2 requires only that unilateral completion be admissible; monotonicity makes that enough to test every realization.

LEMMA 2 (Reduction to singleton safety): *Consider a process in which every current attempt is unverified before exposure. Under Assumptions 1 and 2, the process is regret-free if and only if for every date t and every $i \in A_t$, $C_{t-1} \cup \{i\} \in \mathcal{P}_i$; equivalently, iff $C_t \subseteq T(C_{t-1})$ for every realization.*

Thus each newly public agent must be safe with the already-public coalition plus herself. Pure public accrual applies that test one name at a time. A planned simultaneous action applies the same test to every participant, because each must withstand completing alone. A social assurance contract instead holds completed authorizations tied to a specified statement. It publishes the authorized statements together only when the released coalition protects its members. A privately circulated letter published at a threshold is therefore a tunnel even without a formal platform; the private draft is its holding stage. What matters is whether completed authorization is received before exposure, not whether the institution uses software or a formal administrator.

There are three empirically recognizable cases. With public accrual, names arrive sequentially and the cascade stops at C^-. With simultaneous but unverified attempts, the calendar changes but the safety test does not: every participant must still be safe if she completes alone. With release gated by completed authorizations, the institution observes those authorizations while identities remain unattributed and can test the destination before making one public act. The theory classifies an embargoed letter or trusted delegate by this causal structure, not by whether it is called a contract.

B. Cascade and Tunnel

THEOREM 2 (Cascade Feasibility): *Under Assumptions 1 and 2:*

(i) (Bound.) *Every decentralized public process described in Section I.A, whose transitions use only prior public exposure and unverified current attempts, has $C_t \subseteq C^-(S)$ for every t and every realization when it satisfies regret-freeness.*

(ii) (Attainment.) *The greedy process $A_t = T(C_{t-1}) \setminus C_{t-1}$ satisfies regret-freeness and, in the failure-free realization, reaches exactly $C^-(S)$ in at most $|S|$ dates.*

(iii) (Lattice position.) *$C^-(S)$ is the least fixed point of Γ and $C^+(S)$ is the greatest; the fixed points of Γ form a complete lattice.*

Theorem 2 gives the exact feasibility limit for regret-free decentralized expression: C^- is both the upper bound and an attainable coalition. Section III shows why strategic public expression selects it.

On the mechanism side, a social assurance contract privately collects completed authorizations tied to a specified statement. An authorization is an action already taken and held by the contract, not an intention that may still fail at release. Given the realized completed-authorization set $V \subseteq S$, the rule jointly publishes the authorized statements of a self-protecting subset $D(V) \subseteq V$. It never publishes anything else, and unreleased authorizations remain private.

ASSUMPTION 4 (Atomic release): *The platform can jointly publish the authorized expressions of a released coalition under their signers' identities in a single completed event, not as a sequence of individual exposures subject to Assumption 2.*

Atomic release means one public event, not simultaneous individual attempts. The mechanism can execute that event because every statement in it arrives with a completed authorization. If the event fails, no partial publication appears. Lemma 4 shows that this capability must come from a private holding stage. A protocol, administrator, or trusted delegate can supply it; ordinary public actions subject to Assumption 2 cannot.

Atomicity is therefore where institutional power enters. Lemma 4 proves that any technology guaranteeing joint attribution has first collected and held the named agents' authorizations. The constructive and converse results therefore rely on the same distinction between a completed authorization and an unverified intention.

THEOREM 3 (Tunnel Implementation): *Under Assumptions 1 and 4, the rule $D(V) = C^+(V)$ satisfies regret-freeness for every realized participation V and releases $C^+(S)$ when all consenters commit. Every coalition released by a regret-free mechanism of any architecture is contained in $C^+(S)$. The expansion over the cascade bound is $m(C^+(S)) - m(C^-(S)) \geq 0$, strict iff $C^-(S) \subsetneq C^+(S)$ iff Γ has multiple fixed points.*

Open expression is constrained by safe paths; conditional disclosure by safe destinations. For the realized completed-authorization set V, the mechanism releases $C^+(V)$. Conditional on $V = S$, the expansion is $m(C^+(S)) - m(C^-(S))$. Proposition 4 supplies the frictionless participation ceiling under its stated assumptions; the companion paper studies the signing decision under private information and positive signing friction in a separate homogeneous threshold model (Cashman, 2026).

TABLE 1—RECURRING NOTATION.

Safe coalitions (Section I)	
N, S	potential supporters; consent set $S \subseteq N$
b	seed mass (committed early movers who act without the model's protection guarantee)
a_i	mass (size/weight) of agent i
$m(C) = \sum_{i \in C} a_i$	coalition mass
$\mathcal{P}_i$	protection family: coalitions that leave i safe
ρ_i	scalar protection threshold ($C \in \mathcal{P}_i$ iff $b + m(C) \geq \rho_i$)
C^-, C^+	cascade (least) and tunnel (greatest) self-protecting coalition
y	public mass

PROTECTION REQUIREMENTS AND AUTHORIZATION.

The families $\mathcal{P}_i$ describe the rosters that actually protect each person. To use the release rule, the holder must know those families or receive binding instructions from each signer that specify which rosters authorize publication. Conditional on valid inputs, $D(V) = C^+(V)$ identifies the greatest safe roster. Practical mechanisms instead use a preannounced, authenticated total-count trigger or a small vector of class-count quotas, such as a total count plus a minimum number from a protected class. These rules implement the scalar and low-dimensional cases directly. If safety depends on exactly who else signs, the authorizations must contain more detail and the mechanism needs a rule for choosing among possible rosters.

VALID INPUTS AND CREDIBLE HOLDING.

The set V is not a count of raw accounts. It contains at most one authorization from each eligible, authenticated person, and each authorization is tied to the statement, opponent, and release rule. Authentication establishes who authorized what. It does not establish sincerity, truth, or whether adding that authorization protects the roster. The model takes the separate problem of screening inputs as given: only authorizations that satisfy the announced eligibility and validity rules enter V. The model also requires the holder to keep unreleased valid authorizations hidden from the retaliating audience and follow the announced release rule. The credibility of an intermediary's announced rule is a separate mechanism-design problem (Akbarpour and Li, 2020). Automated, distributed social assurance contracts can greatly reduce the need to trust a single custodian. Mature cryptographic tools can prevent any single actor from inspecting or releasing a failed list, so premature disclosure requires collusion by enough operators to meet the system's threshold or a technical compromise. A companion paper provides a reference architecture and discusses its residual security and governance assumptions; neither paper solves strategic input screening (Cashman, 2026).

COUNT RELEASE AND ANONYMITY.

Certification may release an authenticated count rather than a roster. This is a different outcome: it supplies public evidence that a sufficiently large group exists, or it triggers an institutional action, without publicly naming the group. Proposition S1 (Supplemental Appendix) gives the mapping. Anonymous expression lies outside the named-coalition problem because it gives up the credibility or accountability that makes attribution valuable.

Table 1 collects the notation used in the remaining analysis.

II. The Holding Technology and Its Frontier

The tunnel has a simple sequence: receive signed permission, keep it private, check whether the assembled roster protects its members, and publish the authorized statements together only if it

does. This section asks whether some differently named or organized institution could guarantee the same outcome without following that sequence. The answer is no within the act technology defined below. The first lemma sorts all completed acts into public exposures or private commitments. The second shows that guaranteed joint publication requires someone or something to have received the private commitments before release.

A. *Representation and Zero-Risk Necessity*

Environment E1 begins with individual acts rather than a particular platform. Each person's statement is fixed and tied to her identity. When an act completes, either it makes that person identifiable to the retaliating audience or it does not. In the first case the act exposes her. In the second, a recipient may privately verify and hold the completed authorization for possible later publication. The environment fixes the relevant audience and each person's protection family, and it assumes that a completed act has a definite attribution consequence. Later results treat accidental disclosures and uncertain attribution as risks.

ASSUMPTION 5 (Causal act histories (environment E1)): *(i) Each attempted act has one author, completes or fails as an individual event, and is irreversible. Every subset of a date's attempted individual acts is admissible. Histories record completions in their realized order, including within a calendar date; public attribution updates at each completion before any device act at that date.*

(ii) Information moves through completed acts. Each act has a designated recipient that can verify its completion. Relative to a fixed retaliating audience, each act has a deterministic named set and attribution status, realized at completion. Receipt by a device is not itself attribution.

(iii) A device—a person, administrator, protocol, or platform—chooses its act and named set using only the public history and completed acts it received strictly before execution. If an unattributed act is absent from that received record, the admissibility correspondence contains a twin history with the same public and received record in which that act never completed; the device takes the same action at both histories.

(iv) Consent binds act by act: an act cannot name another agent unless an authorization by that agent has completed on the realized history.

The load-bearing clause is (iii): a device can condition its release only on what it has actually received, not on an authorization that completed elsewhere but never reached it. In E1, payoffs can depend on the sequence of public rosters, the roster ultimately released, and private action costs. A disclosure with no completed act is instead treated as a breach. Appendix S4 gives the formal treatment.

LEMMA 3 (The attribution dichotomy): *Fix any consent-respecting deterministic-attribution technology over this act space in E1. It has a realization-equivalent representation using the two act types in Definition 2. Realization by realization, the representation has the same acts, payoffs, public-coalition path, and released sets. Specifically: (i) a completion realizing attribution is an* exposure action*; (ii) a completed act remaining unattributed is a verifiable* private commitment*; and (iii) the induced admissibility correspondence contains Definition 2's unilateral completions and symmetric failures. Thus that definition represents, rather than restricts, consent-respecting technologies at fixed protection families.*

Here "realization-equivalent" means that the relabeling changes no event that matters: the same acts complete, the same people become public at the same times, the same rosters are released, and the same payoffs result. The representation matters because an institution need not call its

own actions "commitments." A pseudonymous filing, signed instruction to a delegate, or encrypted message can have arbitrary content. What classifies it is whether completion itself identifies its author to the retaliating audience. If it does, the act is exposure. If it does not, but a device receives and verifies it, the act is a held private commitment. Only a statement-bound commitment authorizes later publication; verification establishes completion, not truth. Acts that change legal protection or audience beliefs instead change the protection families and are evaluated in the resulting environment.

LEMMA 4 (Correlation requires holding): *Maintain the preceding primitives. Fix a history h, its attempted profile A_h, and the family $\mathcal{U}(A_h)$ of realizations the institution must withstand there. A realization $R \in \mathcal{U}(A_h)$ is a* joint-completion event *over B, $|B| \geq 2$, if it attributes all of B at one completion and no realization in $\mathcal{U}(A_h)$ attributes some but not all members of B.*

(a) Any consent-respecting technology realizing such an event operates a private conditional-release stage: one device act jointly attributes B, and before executing it the device has received a completed, still-unattributed authorization from every named agent other than its author. No device act names an agent at a history lacking that authorization.

(b) Conversely, a failure-private stage that holds completed, unattributed authorizations and releases them by one act realizes a joint-completion event at every release history whose newly attributed roster has at least two members.

Without holding, every first attribution is an individually completing act governed by Assumption 2. Holding is what makes atomic joint release possible.

Lemma 4 places several apparently different institutions within the same technology. A cryptographic bulletin board, embargoed letter, or delegate publishing a co-signed statement or matched reports performs one public release after receiving the named agents' completed authorizations. The institutional form differs, but the sequence does not: the participants first authorize publication privately; a holder receives those authorizations; the holder then controls the act that reveals the names. Separate posts or a common promise remain separate acts because any one can appear without the others. Guaranteed joint publication therefore requires holding.

The attribution dichotomy classifies acts by what the retaliating audience learns when they are completed. The holding result shows that guaranteed joint attribution requires completed private authorizations. Theorem 4 then applies the safety constraint. The private stage follows from eliminating partial attribution while respecting each named agent's consent.

REMARK 1 (Correlation and privacy risk): *The lemma concerns guaranteed jointness, because admissibility is a set of realizations rather than a probability law. Correlating separate attempts without eliminating unilateral completion changes risk only at positive tolerance. Proposition 2 quantifies risk under independent completion; arbitrary correlation remains an extension, favorable when it removes lone-completion risk and adverse when it concentrates it. Proposition 3 treats breaches of a true holding stage.*

The holder may be a platform, protocol, administrator, lawyer, journalist, or trusted organizer. The next definition uses the word "platform" for any such holder. The phrase "extensive game form" means only that the definition records who can act, what they can observe, and what can happen after each history. It describes every institution in E1 using two kinds of participant action: an action that makes its author public and an authorization that the platform can verify while keeping its author private.

DEFINITION 2 (Expression mechanisms and the public-only class): *Within E1, an* expression mechanism *is an extensive game form between a platform and S. At each history agents attempt*

exposure actions, *whose completion publishes their expressions under their identities, or* private commitments, *whose completion the platform verifies and holds without attribution. The platform observes completed acts; agents observe the nondecreasing public state* C_h. *Transitions condition only on completed acts. The admissibility correspondence includes unilateral completion and the failure of any subset of attempted actions. A mechanism is* public-only *if its publications condition only on completed exposure actions. Every mechanism is consent-respecting: an expression becomes public under an agent's identity only through her completed exposure action or a joint publication she authorized. Regret-free safety applies at every admissible realization.*

By Lemma 3, this definition includes cryptographic bulletin boards, pseudonymous stages, and delegated release. Consent and safety, rather than a platform's physical ability to print names, constrain the outcomes.

THEOREM 4 (Representation of Regret-Free Crossings): *Under Assumptions 1 and 2, with mechanisms as in Definition 2, suppose* $C^-(S) \subsetneq C^+(S)$.

(i) No public-only mechanism satisfying regret-freeness reaches, in any realization, a public coalition outside $2^{C^-(S)}$; *in particular none implements* $C^+(S)$.

(ii) Any regret-free mechanism that in some realization reaches a public coalition $D \not\subseteq C^-(S)$ *contains a* private conditional-release stage. *Consider a history* h *at which one completed release act first attributes a roster* R *containing an agent* i *with* $C_{h^-} \cup \{i\} \notin \mathcal{P}_i$ *but* $C_{h^-} \cup R \in \mathcal{P}_i$, *where* C_{h^-} *is the preceding public set. Before executing that act, its author or device has received a completed, still-non-public authorization from every member of* R *other than possibly the act's author. This held set is nonempty, and each such name in the release is conditioned on its authorization. Thus the mechanism jointly releases the protecting roster from completed private authorizations. Only the author of the release act may not have placed an authorization with the device in advance.*

The notation $2^{C^-(S)}$ in part (i) means all subsets of the cascade coalition, so part (i) rules out every public-only outcome containing a person beyond C^-. The theorem therefore says more than "coordination helps." To cross the public cascade safely, the administrator must hold completed authorizations, not reports that people intend or expect to participate. Each authorization remains hidden from the retaliating audience until the mechanism can release a protecting roster. The person or device performing the release may publish its own statement directly; every other statement in that release must rest on permission already received. Table 2 shows which ingredient familiar institutions possess or lack. Exposure lotteries and vanguards can cross by accepting the possibility that someone is stranded, whereas the theorem requires safety in every admissible realization.

The logic of part (ii): at the first completed act that crosses the cascade boundary, someone named is unsafe with the previously public coalition alone, so the same act must name enough others to protect her; consent requires their prior authorization, and the release device can condition only on authorizations it has received—completed, received, and still private. That sequence is the tunnel's holding stage.

B. *Institutional Features and Tightness*

This distinction separates institutions that can look similar before they are used. An anonymous poll keeps answers private but records preferences, not completed permission to publish a named statement. A pledge list may record completed permission, but publication on a fixed date can reveal a list that is too short. A certified count can trigger an institutional response without ever creating a named coalition. A matched allegation escrow supplies the full sequence for the named outcome: completed reports remain private, a match determines whether release occurs, and the

TABLE 2—FAMILIAR INSTITUTIONS AGAINST THE FOUR FEATURES OF THEOREM 4.

Institution	Private	Completed	Conditional	Joint	Consequence
Publicly accruing petition or letter	no	—	no	no	bounded by the cascade
Poll, "I'll go if you go" talk	yes	no	—	—	talk is not completed authorization; bounded by the cascade
Pledge list, fixed publication date	yes	yes	no	yes	failure exposes the shortfall roster
Commitment opened on a date	yes	yes	no	yes	same failure mode, stronger seal
Assurance contract (money)	yes	yes	yes	n/a	conditionality proven for funds, not names
Card drive; certified count	yes	yes	yes	n/a	unnamed count; outside T4's named-release premise
Targeted synchronized pact	no	no	no	no	positive-tolerance crossing; not regret-free
Embargoed letter (threshold release)	yes	yes	yes	yes	the tunnel (Section V)
Allegation escrow (matched release)	yes	yes	yes	yes	the tunnel at threshold two
Threshold voting w/ disclosure	yes	yes	yes	yes	conditions on the share, not the roster (Remark S2)
Informal trusted circle	yes	partial	yes	yes	reach bounded by pairwise trust

Notes: Private = commitments hidden from the retaliating audience before release; Completed = completed authorizations, not stated intentions; Conditional = release only if the assembled coalition is self-protecting; Joint = authorized expressions published with their signers' identities in one event. "n/a" means the institution certifies money or a count rather than publishing attributable expressions. Theorem 4(ii) requires the full bundle. Proposition 1 separately tests privacy, completed-authorization verification, and conditionality; joint publication is the protecting realization rather than a fourth independent perturbation.

matched reports are published under their authors' identities only when that roster protects them. The causal sequence, rather than the platform's label, determines which result applies.

PROPOSITION 1 (Tightness of the characterization): *On a two-agent instance with an empty seed, neither agent safe alone, and the pair safe together, each of the three properties—private, conditional, verified—is individually necessary: every mechanism retaining two but dropping the third fails to implement the pair regret-free. Thus the characterization's conjunction is irredundant on one fixed nontrivial instance.*

The phrase "each conjunct binds" means that none of the three requirements is redundant. The proof uses the same two-person example for all three omissions. If release is unconditional, a short list can be published. If verification is public, the first person to verify is exposed alone. If participation is merely intended rather than completed, one person's action can appear while the other's fails. The pair is implemented safely only when completed authorizations are both private and condition release.

Information can also make speaking safer by changing what the audience believes. Suppose an information intervention can produce one of the beliefs in a fixed set $\mathcal{M}$, and let $\mu \in \mathcal{M}$ denote the belief it produces. The family $\mathcal{P}_i(\mu)$ records which rosters protect person i under that belief. Changing μ changes the safe destinations themselves; holding changes which of those destinations can be reached without unsafe intermediate exposure.

COROLLARY 1 (Belief interventions and first-mover safety): *If $\{i\} \notin \mathcal{P}_i(\mu)$ for every $\mu \in \mathcal{M}$, no intervention whose only effect is to select a belief in $\mathcal{M}$ makes i safe as an isolated first mover. Information may still change the cascade by making someone else a safe first mover; if it makes i safe alone, the premise fails and information can suffice.*

Information and failure privacy are therefore complements. Information changes which coalitions are safe; holding changes which safe destination can be reached. In the scalar benchmark, lowering requirements from ρ_i to $\rho_i - \gamma\mu$, where the common coefficient $\gamma \geq 0$ measures how strongly favorable audience beliefs lower protection requirements, moves both fixed points. At each fixed belief, the path constraint still determines whether the outcome is the lower or upper crossing (Remark S1).

C. *Positive Tolerance and the Risk Frontier*

The zero-risk requirement treats even a very unlikely lone exposure as unacceptable. I now allow each person to accept some risk and ask whether the tunnel still improves on public attempts. Because exposure is irreversible, every history ends with a terminal public set C_∞. The mechanism's *reach* is the mass of people in that final set who are protected.

A mechanism is *interim-acceptable* at tolerance profile r if every person it might expose faces an unprotected-exposure probability no greater than her tolerance: $\Pr(\text{unprotected exposure} \mid i \text{ exposed}) \leq r_i$. The probability is conditional on being named, because the relevant question for a prospective participant is how often publication leaves her unprotected. Each attempted act completes independently with probability $1 - \varepsilon \in (0, 1)$, failed attempts may be tried again, the holding stage is leak-free, and $r_i < 1$ for every person. The comparison tracks two outcomes: each person's exposure risk and the total protected mass ultimately reached. It is a reliability-and-reach comparison, not a welfare ranking; delay, solicitation cost, entry, and administration are additional welfare considerations.

The platform may itself randomize. The zero-risk standard requires safety after every platform draw that occurs with positive probability and every finite pattern of completed acts following that draw. Because $\varepsilon \in (0, 1)$, each finite completion pattern has positive probability. Randomization therefore cannot make a genuine exposure risk count as zero merely by placing it behind a rare platform choice. Events to which the platform assigns probability exactly zero play no role in the stochastic comparison.

LEMMA 5 (Interim confinement): *Under Assumptions 1 and 2, with mechanisms as in Definition 2, completion as above, and $r_i \in [0, 1)$ for every i, let M be interim-acceptable and let $E(M)$ be the set of agents M exposes with positive probability. Then $E(M)$ is self-protecting, hence $E(M) \subseteq C^+(S)$: on almost every path, every exposed agent, including any agent exposed without protection, lies in $C^+(S)$.*

If person i accepts a risk below one, some protecting realization exposes her; everyone named there lies in $E(M)$, so monotone protection makes $E(M)$ itself protect her. The bound holds even though particular paths may leave particular people unsafe.

This turns a statement about one person's risk into a bound on the whole mechanism. An institution may use early exposures as stepping stones, expose different people on different paths, or accept occasional losses. As long as every exposed person's tolerance is below one, the union of everyone the mechanism ever exposes cannot extend beyond C^+.

The next theorem compares an unverified public trigger with a verified release in two stages. Part (i) replaces one risky trigger while holding later play fixed. Part (ii) uses confinement to compare complete mechanisms, where changing an early trigger may also change everything that happens afterward. In the statement, "almost surely" means with probability one.

THEOREM 5 (Verification dominance): *Maintain the mechanism class characterized by Lemmas 3–4, Assumptions 1 and 2, and the conventions above, with $r \in [0, 1)^S$. Parts (ii)–(iii) hold the consent set fixed and let solicited agents sign mechanically; strategic signing is the separate problem in Section III.*

(i) (Trigger comparison.) *Suppose an acceptable mechanism, at public state C, rests agent i's not-yet-safe exposure on same-date attempts by $B_i \neq \varnothing$, and let $T = C \cup \{i\} \cup B_i$. Suppose the full-completion target T protects every member of $\{i\} \cup B_i$ who is not already safe with C. Holding downstream play fixed, replace only that trigger by retried solicitation of the same agents' completed commitments and one joint release of T. Conditional on reaching this trigger, the replacement weakly lowers every participant's probability of unprotected exposure generated there; for i the reduction is at least $(1-\varepsilon)\varepsilon^{|B_i|}$ unconditionally over the trigger's completion draws, or $\varepsilon^{|B_i|}$ conditional on her attempt completing. On shared draws, the set released at the end of the replacement trigger contains every agent the original trigger exposes. Without retries only the risk comparison survives.*

(ii) (Global dominance.) *Every acceptable mechanism's terminal protected mass is at most $m(C^+(S))$ on almost every path. The tunnel with retried solicitation releases $C^+(S)$ almost surely in finite time with zero unprotected-exposure probability, and therefore weakly dominates every acceptable mechanism in risk, agent by agent, and in almost-sure reach.*

(iii) (Zero-risk endpoint.) *Among zero-risk mechanisms, the maximal almost-sure released mass is $m(C^+(S))$, attained by the tunnel. When $C^-(S) \subsetneq C^+(S)$, every zero-risk mechanism attaining that mass almost surely contains Theorem 4's private conditional-release stage. At the exact endpoint $r = 0$, acceptability is zero risk, so dominance becomes necessity. Along any sequence of acceptable mechanisms with tolerance profiles $r^n \to 0$, each agent's unprotected-exposure risk converges to zero; no strictly positive tolerance by itself forces zero risk.*

Theorem 4 is the zero-risk endpoint of this broader comparison. When solicitation can be retried for free, the tunnel can wait until every required authorization is actually in hand. Waiting removes the risk that only part of a proposed group appears, yet it does not reduce the final protected coalition. Interim confinement supplies the matching upper bound: no acceptable mechanism can expose people beyond C^+. Delay, repeated solicitation, administration, and entry matter for welfare once they are costly.

Part (i) isolates the mechanical value of verification: releasing only after every authorization is in hand removes the risk that one participant's act completes while her co-movers' acts fail. Part (ii) handles the harder comparison between whole mechanisms, where replacing one trigger may change the later public state and therefore later behavior.

Free retries isolate the value of information timing. They let the tunnel wait until the intended participants have completed authorization, so verification improves safety without discarding partially completed groups. Discounting and costly solicitation put a price on that waiting; Theorem 5 supplies the technological risk–reach limit, and Section III introduces costs and credibility.

REMARK 2 (Conditions for dominance): *The dominance comparison requires retries, tolerances below one, and an objective that ranks mechanisms only by exposure risk and final protected reach. With only one solicitation, a strict all-or-nothing release may discard a partially completed group that public attempts would expose. A person with $r_i \geq 1$ accepts certain unprotected exposure and can act as a new seed. Costs and delay can also make a faster but riskier mechanism preferable. Appendix S4 gives the local comparison and examples showing why it cannot simply be repeated trigger by trigger to compare whole mechanisms.*

Necessity under interim safety.

Regret-free safety gives exposure risk priority over every benefit of speaking. With positive tolerance, benefits and harms can instead be compared. Let e_i be person i's benefit from expression and h_i her loss from being exposed without protection. Her tolerance is $r_i = e_i/h_i$: greater expressive value or smaller retaliation harm makes her willing to accept more risk. Each attempted exposure

completes independently with probability $1-\varepsilon$, where $\varepsilon \in (0,1)$ is its failure probability. At a particular trigger, person i can receive either benefit or harm only if her own attempt completes, so the local calculation conditions on that event. For a complete mechanism, acceptability conditions on every route by which she might be exposed. (Subscripted e_i is expressive benefit; unsubscripted b remains seed mass.)

The fight-or-fold model in Section IV gives one economic setting in which these expressive benefits and exposure harms arise.

Proposition S3 (Supplemental Appendix) gives lower and upper bounds. Suppose person i relies on a batch B of other attempted participants. Conditional on her own attempt completing, all members of B fail with probability $\varepsilon^{|B|}$. If this trigger accounts for share α_i of all occasions on which she is exposed, it contributes at least $\alpha_i\varepsilon^{|B|}$ to her mechanism-wide risk, so acceptability requires $\alpha_i\varepsilon^{|B|} \le r_i$. Other routes that expose her safely can reduce the conditional importance of this trigger. Thus positive tolerance permits some unverified crossing; at exactly $r=0$, every unsafe trigger reached with positive probability is excluded and regret-free safety returns.

The all-fail event is deliberately conservative: a large group may still protect its members even when some attempts fail. The scalar benchmark can use the full distribution of the number who complete. A *targeted synchronized pact* asks everyone needed to cross the exposure barrier to attempt public expression at the same time. It has no private holding stage. Its protection comes from scale: if enough attempts are likely to complete, the realized group will usually clear the threshold even though no individual completion is guaranteed.

Formally, suppose G_S crosses the diagonal at $y^- < y^u < y^+$ and lies above it on (y^u, y^+). Every consenter with requirement in $(y^-, y_t]$ attempts exposure together for some target $y_t \in (y^u, y^+)$. The quantity $s(y_t) = G_S(y_t) - y_t$ is the *excess safe mass*: the amount by which the people protected at y_t exceed the target itself. The batch mass is $m_B = (y_t - y^-) + s(y_t)$, and $\Delta(y_t) = (1-\varepsilon)s(y_t) - \varepsilon(y_t - y^-)$ measures expected safety slack after allowing for failed attempts.

PROPOSITION 2 (The targeted synchronized pact): *In the scalar benchmark, with $y_t \in (y^u, y^+)$ and notation as above:*

(i) (Mean condition.) *The expected realized public mass is $y^- + (1-\varepsilon)m_B = y_t + \Delta(y_t)$: mean realized mass clears the marginal requirement exactly when $(1-\varepsilon)\, s(y_t) \ge \varepsilon\,(y_t - y^-)$. This is a condition on the mean; part (ii) bounds the probability of an unprotected outcome.*

(ii) (Concentration.) *With n equal-mass batch members and $\Delta = \Delta(y_t) > 0$, every member's unprotected probability conditional on her own completion is at most $\exp(-2n\Delta^2/m_B^2)$—the displayed bound absorbs both the conditioning on her own completion and the resulting slack $t^* \ge \Delta$—as is the probability that the realized mass falls short of y_t; the pact is interim-acceptable to every member once $n \ge (m_B^2/2\Delta^2)\ln(1/\underline{r})$, where $\underline{r} = \min_{i\in B} r_i > 0$ is the batch's minimum tolerance. As $n \to \infty$ at fixed $\Delta > 0$ the unprotected probability vanishes.*

(iii) (Composition.) *Read the post-pact continuation in the scalar nonatomic benchmark, with continuous H_S. On the complement of the shortfall event in (ii), the realized public mass lies in $[y_t,\, G_S(y_t)] \subseteq (y^u, y^+]$, the safe-expansion cascade resumes from the realized set, and its closure is y^+: pact-then-cascade reaches $C^+(S)$, so a public-only composite reaches the tunnel's destination except with probability at most $\exp(-2n\Delta^2/m_B^2)$.*

The pact is a third institution between cascade and tunnel. The exponential expression in part (ii) is a concentration bound: as the number of equal-sized participants grows, realized completion becomes tightly concentrated around its mean. Statistical concentration can therefore substitute for completed-authorization verification when excess safe mass, group size, and the completion rate are large enough. Failure privacy is most valuable where that statistical protection is weakest:

small groups, cases in which one missing participant matters, severe stakes, and adverse completion correlation.

This result also separates robust feasibility from high-probability coordination. A large pact can make the probability of unprotected exposure arbitrarily small while never removing the lone-completion realization from the set of admissible outcomes. Hence it approaches the tunnel in probability for every fixed positive tolerance but does not enter the regret-free class at $r = 0$. The discontinuity is economic rather than notational: a worker who tolerates a one-in-a-thousand chance of standing alone can use scale, while one who will not accept that realization at all needs completed-authorization holding.

The formula identifies where scale is and is not a substitute for the institution. Large batches help because realized completion concentrates around its mean. Slack helps because the target can lose some participants and remain protecting. Reliable completion raises that mean. If any of these is missing, concentration offers little: a pair of corroborating accusers has no law of large numbers; a coalition sitting just on its protection threshold has no slack; and severe retaliation makes even a small residual tail unacceptable. These are exactly the cases in which collecting completed authorizations before exposure has the greatest value. Completion correlation can help (eliminating lone completions) or hurt (concentrating them); the independent benchmark makes the tradeoff transparent.

In a public pact, the effective failure rate depends on whether participants strategically withdraw as well as on whether their attempts technically fail. A held commitment has already been completed, so the remaining strategic choice is whether to enter the holding stage. Proposition 4 gives the frictionless entry benchmark.

Real holding stages can leak. Let δ be an upper bound on the probability of a breach that comes from outside the modeled completion process and is independent of who completes and of any platform randomization. A stage is δ-failure-private if, whenever no breach occurs, it reveals commitments only through a protecting release. Not every breach is harmful: a breach that reveals an already protective roster does not strand anyone. Let δ_i^{h} denote the part of person i's mechanism-wide risk caused by breaches that name her in an under-protective set, conditional on her being exposed.

PROPOSITION 3 (The robust frontier): *Fix $\varepsilon \in (0,1)$ and tolerances r_i as above.*

(i) (Sufficiency.) *Run the tunnel on a self-protecting target C, soliciting exactly its members until every commitment is held and then releasing C. On a δ-failure-private stage, each $i \in C$ is exposed almost surely and is unprotected only on a harmful-leak event. Hence her conditional risk is $\delta_i^{\mathrm{h}} \le \delta$, and the mechanism is acceptable exactly when $r_i \ge \delta_i^{\mathrm{h}}$; the common bound $r_i \ge \delta$ is sufficient.*

(ii) (Necessity, with rates.) *Conversely, consider any mechanism acceptable to its participants. Suppose a no-leak trigger first exposes a not-yet-safe agent i through her own unverified attempt resting on a batch B_i of other attempted participants, as in Proposition S3. Let m_i be the probability, conditional on i being exposed, that this trigger is reached and her own attempt completes, measured before the other participants' same-date completion draws. Then*

$$\delta_i^{\mathrm{h}} + m_i\, \varepsilon^{|B_i|} \;\le\; r_i.$$

The harmful-leak rate and the unverified-crossing mass jointly draw down the tolerance budget. In particular $\delta_i^{\mathrm{h}} \le r_i$, and $m_i \le (r_i - \delta_i^{\mathrm{h}})/\varepsilon^{|B_i|}$. For disjoint unverified triggers, replace the second term by the sum of their trigger shares times their all-fail tails.

(iii) (Zero-risk endpoint.) *Suppose $C^-(S) \subsetneq C^+(S)$. At $r = 0$, every acceptable route past C^- conditions exposure on completed, verified, non-public authorization and joint release with*

probability one, providing failure privacy against harmful leaks and recovering Theorem 4. Along any sequence $r^n \to 0$, the harmful-leak contribution and every fixed trigger-share term in (ii) converge to zero, but may remain positive at every n.

The left side of the inequality in part (ii) adds two ways person i can be left unprotected. The term δ_i^{h} is harmful leakage from the private stage. The term $m_i\varepsilon^{|B_i|}$ is the risk from a public trigger at which her own act completes but every supporting act in B_i fails. Their sum must fit within her tolerance r_i. Short holding windows, rolling deletion, and minimal storage of identities reduce the first term. Larger and more reliable public groups reduce the second. Under monotone protection, a small targeted leak can be more harmful than a nearly complete disclosure because it may reveal too few people to protect one another.

The inequality gives the deployed institution a design target. Raw breach probability matters only through breaches that reveal an agent in an under-protective set. Deleting failed deposits, minimizing the time identities are jointly stored, and separating threshold certification from identity release all reduce that harmful component. Institutional security and behavioral coordination are therefore substitutes in the tolerance-budget inequality rather than unrelated safeguards.

A security audit should therefore track who can be exposed at each interim state, not only whether the database was accessed.

D. Scope of the Main Results

Theorem 4 anchors the zero-tolerance endpoint of a broader risk–reach tradeoff. Positive tolerance permits synchronized pacts, while institutional leaks consume some of that tolerance. Vanguards who knowingly move first, exposure lotteries, and groups sustained by trust can reach further by accepting risks that regret-free safety forbids. The analysis first derives the exact holding requirement at zero risk and then quantifies departures from it.

Regret-freeness functions either as an institutional constraint against trading a participant's safety for reach or as a behavioral description of people whose exposure harm dominates the benefit of taking that gamble. Vanguards, subsidized first movers, and actors protected by relational trust can seed a later cascade. For safety-seeking participants, the model identifies where public expression stalls and the institution that moves the coalition across the barrier.

III. Selection and the Participation Ceiling

The earlier results ask what a public process or a contract *can* reach if the relevant people participate. This section asks what they choose to do. It first shows that safely able public speakers act rather than wait, so the public process reaches the full cascade closure. It then studies the private signing decision and shows why full signing is a natural benchmark when signing is costless, hidden, and credible. Costs, leakage, and distrust create the richer entry problem studied in the companion paper (Cashman, 2026).

At each date, every person who is not yet public chooses whether to attempt public expression. Whether an attempt completes is represented by a binary random variable $z_{it} \in \{0,1\}$ for each agent–date pair. It is realized after that date's choices, with $\Pr(z_{it} = 1) = p_i \in (0,1)$. These variables are independent across agents and dates, and their distribution does not depend on the public state or history. Because S is finite, $p = \min_i p_i > 0$. Agent i's attempt completes exactly when $z_{it} = 1$; failure costs nothing and leaves the public roster unchanged. These public attempts are unverified: no co-participant's attempted action is received or certified before exposure. Once named, person i receives a continuing payoff $u_i(C)$ that weakly rises as the public coalition grows. This payoff is nonnegative exactly for rosters in her protection family; an unnamed person receives zero.

A strategy is *Markov* if it depends only on the current public roster, not on the full history that produced it. It is *robustly admissible* if no completion pattern allowed by the model exposes the person outside her protection family. A robust Markov-perfect equilibrium combines these restrictions: at every public roster, each person chooses her best Markov strategy among the robustly admissible strategies. The last condition, *strict safe surplus*, $u_i(C) > 0$ whenever $C \in \mathcal{P}_i$, means that speaking in a safe coalition is strictly better than remaining silent. This breaks the tie between acting safely now and waiting.

THEOREM 6 (Sequential safe-expression selection): *Payoffs are expected discounted sums of the per-period flows, with discount factor $\beta \in (0,1)$. Under the state-independent completion technology above, Assumptions 1 and 2, and strict safe surplus, every robust Markov-perfect equilibrium has the same action rule. At each public state C, exactly the agents in $T(C) \setminus C$ attempt. The public set therefore follows the greedy path, $C_t = T(C_{t-1})$, apart from completion failures and retries. No equilibrium path reaches outside C^-. Nor can an equilibrium* strategically *stall below it: at every public state, each agent who is safe to join attempts rather than waits. Because attempts complete with probability bounded below by the $p > 0$ above, the public coalition reaches C^- almost surely in finite time.*

Theorem 2 said that C^- is the largest coalition a regret-free public process can reach. The selection theorem adds that the process actually reaches it: whenever a person can safely join, strict surplus makes joining now better than waiting. The benchmark uses Markov strategies, no trust in promises of later action, no benefit to a silent person from other people's expression, costless failures that no one observes, and a binary distinction between being named and remaining private. Proposition S4 (Supplemental Appendix) shows how visible failed attempts can change the stopping point.

The strategic content is the exclusion of inaction. Every person in $T(C) \setminus C$ is safe even if her attempt is the only one that completes, so strict surplus makes acting immediately better than waiting. If safe expression yields exactly zero surplus, waiting can also be optimal. If people trust promises of later follow-through, that *ambient trust* can support coordinated movement beyond C^-. With no such trust, the selected public outcome is the cascade closure.

The selection result rules out a simple objection to the comparison. The difference $C^+ \setminus C^-$ does not arise because the paper chose an unusually inactive public equilibrium. Theorem 6 selects active behavior whenever public participation is safe, yet the public process still reaches only C^-. The remaining difference between C^- and C^+ is what private holding adds.

The tunnel outcome is conditional on authorization supply. Latent consenters form S, valid completed authorizations form V, and the mechanism releases $C^+(V)$. Conditional on $V = S$, the release is $C^+(S)$; for any other realized authorization set, the rule returns the greatest safe coalition within V. Apart from the frictionless participation ceiling below, the paper does not model formation of V under private information or positive signing friction.

At the frictionless benchmark, any difference between consenters S and actual signers V comes from strategic coordination rather than signing costs or fear of leakage. Consider a one-shot simultaneous *signing game*. Each person in S either submits a completed authorization tied to the statement or abstains. No one observes anyone else's choice before deciding. The resulting signer set is V, and the mechanism either releases $C^+(V)$ in one event or releases nothing.

The benchmark removes four possible frictions. Signing is costless. Signatures do not change anyone else's strategy, protection, or the audience's response before release. The stage does not leak and the administrator follows the announced rule. A published person earns $u_i(C)$, while an unpublished person earns zero. These assumptions isolate whether fear of being the only signer can itself block entry.

PROPOSITION 4 (Full signing at the frictionless benchmark): *Under Assumptions 1 and 4, in the frictionless signing game: (i) signing is weakly better than abstaining against every profile of others' choices, realization by realization. (ii) Under strict safe surplus, for every $i \in C^+(S)$ signing weakly dominates abstaining in the standard sense, and is the strict best response when all others sign; for $i \notin C^+(S)$ the two actions are payoff-equivalent against every profile, since such an agent is released under no profile. (iii) Everyone-signs is a Nash equilibrium in which no agent plays a weakly dominated strategy, and an equilibrium avoids weakly dominated strategies if and only if every member of $C^+(S)$ signs—in which case the release is exactly $C^+(S)$. When no agent is safe alone given the seed, all-abstain is also an equilibrium: the selection is by dominance, not uniqueness.*

Signing is a no-lose move here: it cannot place a person in an unsafe release, and abstaining only forgoes safe publication. Every equilibrium that avoids weakly dominated choices therefore releases $C^+(S)$, regardless of what people outside that coalition do. Small signing costs preserve full signing as an equilibrium but remove the dominance argument; observable signing, leakage, distrust, and benefits from others' expression also shape entry, with Proposition 3 pricing the leakage margin.

The companion paper studies the entry problem in a homogeneous threshold model. It asks how a small amount of private uncertainty can pin down signing behavior, how signing frictions change participation, how revealing the entire signer list after a failed campaign eliminates the benefit of failure privacy, and what implementation requires (Cashman, 2026). Those results characterize signing in a separate homogeneous environment. Here the release rule takes the realized valid authorization set V as an input and maps it into its largest safe coalition $C^+(V)$.

IV. Visibility and Suppression under a Fight-or-Fold Opponent

The protection families have so far been abstract lists of safe rosters. This section derives them from a simple opponent who chooses whether retaliation is worthwhile. The opponent may be an employer considering dismissal, an accused person considering intimidation, or a regime considering repression. A larger coalition is harder to suppress, but attacking it may still be worthwhile when the opponent's stake is high or retaliation is cheap.

Let $q(C) \in [0,1]$ be coalition C's probability of surviving retaliation, and assume that adding members weakly raises this resilience. The opponent loses a stake $d > 0$ if the coalition survives and pays retaliation cost $\kappa_T \in (0,d)$ if it fights. Fighting succeeds with probability $1-q(C)$, so its expected gain is $d[1-q(C)]$. The opponent fights when that gain exceeds κ_T and folds otherwise. Equivalently, each opponent has a break-even survival probability $q^* = 1-\kappa_T/d$: it fights coalitions less resilient than q^*. The distribution of opponent thresholds is F.

$$C \text{ draws no retaliation iff } q(C) \geq q^*; \qquad \Pr(C \text{ is fought}) = 1 - F\big(q(C)\big). \tag{2}$$

The probability of a fight is $1-F(q)$, and conditional on a fight the coalition is suppressed with probability $1-q$. Their product, $\pi(q) = (1-F(q))(1-q)$, is the overall suppression probability. As coalition size raises q, it both deters some opponents from fighting and makes attacks less likely to succeed. Person i, with expressive benefit e_i and suppression loss h_i, accepts publication when $e_i \geq \pi(q(C))h_i$. Larger coalitions therefore generate monotone protection families. Lemma 7 extends the earlier lattice results to these expected payoffs; strategic selection also requires expression to provide strictly positive expected surplus whenever the coalition protects the person. The benchmark holds the opponent's stake and retaliation cost schedule fixed across coalition sizes so that size operates through resilience.

The visibility–suppression reversal.

Failure privacy changes who enters the data. Hold the environment and the full-signing population fixed, and let $G = C^+ \setminus C^-$ be the hidden supporters reached by the tunnel but not by the public cascade. Under open organizing, these people remain silent. Their expression is therefore completely suppressed, but because they never become visible, the data record no retaliation against them. Under conditional release they become public as part of C^+. Opponents now fight them with probability $1 - F(q(C^+))$ and successfully suppress them with probability $\pi(q(C^+))$. For the proposition, assume the latter probability lies strictly between zero and one.

PROPOSITION 5 (Visibility–suppression reversal): *Fix the fight-or-fold primitives above, full signing, a nonempty set $G = C^+ \setminus C^-$ of hidden supporters, and $\pi(q(C^+)) \in (0,1)$. For the hidden supporters, the model comparison from open organizing to failure-private release* raises *observed retaliation—from 0 to $1 - F(q(C^+))$ in fight attempts, or to $\pi(q(C^+))$ in successful punishments—while* lowering *effective suppression of their expression from 1 to $\pi(q(C^+))$. The two move in opposite directions.*

COROLLARY 2 (Measurement reversal): *Among the hidden supporters identified by the model, more observed retaliation can accompany less suppression. The mechanism makes visible people who have no record on the open path, so recorded retaliation rises while suppression falls. This is a model implication for a latent subgroup, not a causal estimate from observed campaign data. Raw campaign counts can move either way because failure privacy changes both the visible population and the fight rate.*

The reversal is a denominator problem. Open-path records contain only people who became visible; they omit people whose fear of retaliation kept them silent. Identifying that missing group requires information beyond observed campaigns. Proposition S2 in the Supplemental Appendix gives the member-by-member version.

V. The Social Assurance Letter

The social assurance letter is the model's direct institutional counterpart. A statement on a contested question circulates privately. Signing is a completed authorization to publish that statement under the signer's name. The circulator holds the signatures without revealing them to the audience that can retaliate. Once the signatory list protects every person who will be named, the circulator publishes the letter and its signatures in one event. Publication itself is the expressive act. If the letter never assembles a protective list, it is not published and the retaliating audience learns no signer's identity.

The roster is the published signatory list. A signer's requirement ρ_i is the minimum public mass at which signing is no longer career-threatening. The class-count triggers in Section I give a practical rule: publish after receiving fifty signatures, including ten from tenured faculty. Organizers who are public from the start form the seed. The public counterpart is an accruing petition. Its visible signature list advances one name at a time and stops at the cascade closure; a letter assembled privately and released only after its threshold is met can reach signers that the petition cannot. In the model's terms, the letter uses private circulation, actual signatures rather than pledges, a threshold test, and one joint publication.

A dissent letter uses exactly this institution: the controversial position is the letter's content, not a different release technology. Misconduct reporting and union organizing fit less directly because the mechanism may produce a private notice or a certified count rather than publish names. An allegation escrow can hold reports until a match, but many systems notify a counselor rather than publish a named coalition (Ayres and Unkovic, 2012; Arun et al., 2020; Rajan et al., 2018). Screening

also matters because an additional report need not make the existing reports safer (Chassang and Padró i Miquel, 2019). Private union cards may certify support to a labor board, but certification often releases a count and triggers a legal response rather than completing a named expressive act. These count and intermediation outcomes are distinct from joint publication (Proposition S1). Appendix S7 develops these mappings and the dual-use collusion boundary.

Empirical interpretation.

A disclosure rule can change three distinct moments: whether identities become visible while commitments accumulate, whether a failed campaign reveals its participants, and whether a successful campaign publishes names. An empirical comparison should distinguish these margins. It should also measure separately how many previously silent people surface, how often opponents attempt retaliation, and how often those attempts succeed. The visibility–suppression reversal predicts that recorded retaliation can rise even while more hidden supporters speak and survive.

VI. Relation to the Literature

The closest structural comparison is Gale (2001). Both models study irreversible public participation in which one person's action makes participation more attractive for others. Proposition S5 constructs a game in Gale's domain where strategic expectations lead eventually to C^+, the largest safe coalition, even though the largest regret-free public path stops at C^-. The difference is follow-through. In Gale's game, an early participant accepts current exposure because equilibrium strategies prescribe that others will act later, and only the final action profile determines payoffs. This paper asks what can be reached when an early participant refuses to rely on that future action and which institution can replace the missing guarantee (Theorem 4).

Braghieri, Bursztyn and Fasnacht (2026) provide the closest mechanism comparison. They study and experimentally test a rule that reveals votes only when a voting threshold is met, protecting voters from social-image stigma after an unsuccessful vote. Here the identities released also determine whether the group can withstand material retaliation. Proposition S7 isolates that distinction: if safety depends only on a fully external statistic that the roster cannot change, the cascade and tunnel coalitions coincide. The proposition is a boundary case, not a version of their voting model. In their paper, voting behavior determines the vote share; here the boundary result assumes that the statistic governing safety is fixed outside the roster. The social-image literature (Bénabou and Tirole, 2006; Ali and Bénabou, 2020; Bursztyn, González and Yanagizawa-Drott, 2020) supplies belief-based exposure costs; this paper allows material protection to depend on roster composition.

The representation result also relates to fault-tolerant implementation. Eliaz (2002) allows some participants to take faulty actions. Failure to complete a promised public act is one possible fault in this application, but it is not the specific formal object of that paper.

VII. Conclusion

Publishing statements under their authors' names creates a problem of safe coalition formation. A group may be large enough to protect everyone once public but impossible to assemble through safe public additions. For a fixed audience and fixed protection environment, an institution must either expose participants while the coalition forms or hold completed authorizations until it can publish them together. Under safety in numbers, separate completion of public attempts, and the zero-risk service standard, open expression cannot pass the cascade coalition C^-. Under the robust Markov benchmark of Theorem 6, safely able speakers act, so the public path reaches C^- rather than stalling below it.

A social assurance contract can instead reach $C^+(V)$, the greatest safe roster among the people whose completed authorizations it holds. The representation theorem shows why. When public acts

complete individually and completion determines attribution, any zero-risk procedure that moves beyond the cascade must first receive permission while the relevant people remain private and then publish the protecting roster in one event. A sufficiently large and reliable group with positive risk tolerance can sometimes substitute a synchronized public pact for that private stage (Proposition 2). Holding is most valuable for small groups, coalitions barely above their protection threshold, and participants facing severe retaliation. Beyond the frictionless participation ceiling, how many valid authorizations are signed is a separate entry problem. The companion paper studies that problem in a homogeneous threshold environment with private information and signing frictions (Cashman, 2026).

The fight-or-fold model yields a measurement implication: for hidden supporters, recorded retaliation can rise while suppression falls. Performance measures must therefore distinguish whether supporters surface, whether opponents attack them, and whether those attacks succeed.

The central design principle is simple: when attribution is irreversible, hold completed consent until publication itself supplies protection.

Appendix

Proofs

This appendix proves every formal result stated in the article body. The proofs are ordered so that each supporting result appears before the argument that uses it. The Supplemental Appendix contains extensions and auxiliary results rather than proofs deferred from the article.

Proof of Lemma 1:

If $C \cup \{i\} \in \mathcal{P}_i$ and $C \subseteq C'$, then $C \cup \{i\} \subseteq C' \cup \{i\} \ni i$, so Assumption 1 gives $C' \cup \{i\} \in \mathcal{P}_i$; hence $i \in \Gamma(C')$ and (if $i \notin C'$) $i \in T(C')$. Inflation is immediate from the definition of T.

Proof of Theorem 1:

Let $\{C_j\}$ be self-protecting and $U = \bigcup_j C_j$. If $i \in U$ then $i \in C_j \in \mathcal{P}_i$ for some j, and $C_j \subseteq U \ni i$, so Assumption 1 gives $U \in \mathcal{P}_i$. Thus U is self-protecting and contains every self-protecting coalition, so $C^+(S)$ is well defined and largest. A safe rule must release a self-protecting coalition, hence a subset of $C^+(S)$.

Proof of Lemma 2:

($\Leftarrow$) Take any realization $R_t \ni i$. Then $C_{t-1} \cup \{i\} \subseteq C_{t-1} \cup R_t = C_t$, so Assumption 1 gives $C_t \in \mathcal{P}_i$; members of C_{t-1} remain safe because the public set only grows. Induct on t. ($\Rightarrow$) Because every current attempt is unverified before exposure, Assumption 2 makes the realization $R_t = \{i\}$ admissible for any $i \in A_t$, and regret-freeness at that realization requires $C_{t-1} \cup \{i\} \in \mathcal{P}_i$. The equivalence with $C_t \subseteq T(C_{t-1})$ is the definition of T.

Proof of Theorem 2:

(i) By Lemma 2, $C_t \subseteq T(C_{t-1})$; by Lemma 1 and induction, $C_t \subseteq T^t(\varnothing) \subseteq C^-$. (ii) Greedy intentions satisfy Lemma 2's condition by construction, and failure-free realizations give $C_t = T^t(\varnothing)$, stabilizing at C^- within $|S|$ steps. (iii) C^- is self-protecting (each member was singleton-safe at her addition date, and Assumption 1 preserves safety as the set grows) and closed (the iteration stabilized), hence a fixed point. *Leastness*: let $C^* = \Gamma(C^*)$; if $T^n(\varnothing) \subseteq C^*$ then for $i \in T^{n+1}(\varnothing)$ we have $T^n(\varnothing) \cup \{i\} \in \mathcal{P}_i$, so Assumption 1 gives $C^* \cup \{i\} \in \mathcal{P}_i$ and $i \in \Gamma(C^*) = C^*$; hence $C^- \subseteq C^*$. C^+ *greatest*: C^+ is self-protecting (Theorem 1), so $C^+ \subseteq \Gamma(C^+)$; conversely if $i \in \Gamma(C^+)$ then $C^+ \cup \{i\} \in \mathcal{P}_i$, so $C^+ \cup \{i\}$ is self-protecting (members stay safe by Assumption 1) and $C^+ \cup \{i\} \subseteq C^+$ by maximality, giving $i \in C^+$; thus $C^+ = \Gamma(C^+)$, and every fixed point, being self-protecting, lies in C^+. Tarski's theorem gives the complete lattice.

Proof of Theorem 3:

At release the public coalition jumps from the seed, the pre-release public state—no interim exposure, by failure privacy—to $C^+(V)$, which is self-protecting by Theorem 1 applied to consent set V; so every exposed agent is protected in every realization and regret-freeness holds. Monotonicity of $V \mapsto C^+(V)$ (Assumption 1) gives $C^+(V) \subseteq C^+(S)$, with equality at $V = S$; the upper bound over all mechanisms is Theorem 1. The strictness claim is Theorem 2(iii).

Proof of Lemma 3:

Claim 1 (exhaustive act partition). Fix an act a. Deterministic attribution in Assumption 5(ii) gives two cases. If a's completion attributes its author to the retaliating audience, that completion is an exposure action. If it does not, completion leaves its author non-public; a recipient can verify it by receipt without creating attribution, so it is a private commitment. Any later attribution is carried by a later act and is classified at that later completion. These cases are exhaustive and disjoint.

Claim 2 (pathwise translation). At every history, retain the same acts, authors, recipients, completion outcomes, and within-date order. Let the platform copy the device's action rule on the same public and received-act record. Claim 1 changes only the type label attached to each act.

Hence the translated mechanism has the same public-coalition path, released sets, and private act costs, realization by realization; no later result or payoff in E1 depends on the label itself.

Claim 3 (admissibility). Assumption 5(i) admits every subset of a date's attempted individual acts. In particular it contains each lone completion required by Assumption 2 and the complementary failure realizations required by Definition 2. Because Claim 2 preserves the completion correspondence, the translated mechanism inherits those realizations.

Claims 1–3 establish (i)–(iii) and the asserted realization-by-realization equivalence. The theorem and frontier results can therefore be read on the translated mechanism without changing their economic events.

Proof of Lemma 4:

Claim 1 (one carrier act). Let realization R contain the completion c that jointly attributes B. Individual completion means that c is the completion of one act a^* by one author w. Because attribution occurs at an act's own completion, every $j \in B \setminus \{w\}$ is named by a^* rather than by a separate self-attributing act. Separate acts cannot generate the stipulated event: Assumption 5(i) admits a realization in which one completes without the others, which would attribute only part of B.

Claim 2 (completed authorization). By act-by-act consent, for every $j \in B \setminus \{w\}$ some authorization by j completed before a^* named her. If $w \notin B$, this applies to all of B; if $w \in B$, it applies to $B \setminus \{w\}$. The relevant set is nonempty because $|B| \geq 2$.

Claim 3 (receipt, not mere completion). Fix such a j. Because c is j's first attribution, her earlier authorization is unattributed. Suppose the device had not received it before executing a^*. Assumption 5(iii) supplies a twin history with the same public and received-act record in which that authorization never completed. By the same assumption, the device takes the same action at the twin history, contradicting act-by-act consent. Thus the device received j's completed authorization strictly before executing a^*. The same argument shows that no device act names j at a history whose received record lacks her authorization.

Claim 4 (private holding). Completion c is j's first attribution. Her prior authorization therefore remained unattributed to the retaliating audience; receipt itself is not attribution. Lemma 3 classifies it as a completed private commitment. Claims 1–4 yield one device act, conditioned name by name on previously received, completed, still-unattributed authorizations. This proves (a).

Claim 5 (converse). At a release history of a failure-private stage, no member of the newly attributed roster D was public through her held commitment. The device attempts one release act. Its failure attributes none of D and its completion attributes all of D at that completion; Assumption 4 excludes a partial roster. Thus every release with $|D| \geq 2$ is a joint-completion event, proving (b).

Without holding, a device has no completed, unattributed authorization in its received record. Claims 2–4 then rule out a multi-name first-attribution act, leaving individually completing exposure acts and their admissible lone-completion realizations. This gives the lemma's final statement.

Proof of Theorem 4:

Claim 1 (public-only local bound). In a public-only mechanism, a newly public agent's exposure action can condition only on prior public exposure and its own completion. Assumption 2 admits the realization in which that action completes without the other current attempts. Regret-freeness therefore requires $C_{h^-} \cup \{i\} \in \mathcal{P}_i$ for each newly exposed i, so each public transition satisfies $C_h \subseteq T(C_{h^-})$.

Claim 2 (public-only induction). Starting from the empty nonseed public set, Claim 1 and monotonicity of T give $C_h \subseteq C^-$ at every history, exactly as in Theorem 2(i). This proves part (i).

Claim 3 (first crossing act). For part (ii), use Assumption 5(i)'s completed-act history, not merely its coarser calendar date. If several completions share a timestamp, their recorded realized order fixes the predecessor public set; this bookkeeping changes neither the acts nor payoffs. Let a^* be

the first completed act after which the public set is not a subset of C^-, let $C_{h^-} \subseteq C^-$ be the predecessor set, and let R be the roster first attributed by a^*. Choose $i \in R \setminus C^-$.

Claim 4 (the crossing requires another named participant). Agent i is unsafe with only the predecessor and herself. Otherwise $C_{h^-} \cup \{i\} \in \mathcal{P}_i$ and monotonicity would give $C^- \cup \{i\} \in \mathcal{P}_i$, so $i \in \Gamma(C^-) = C^-$, a contradiction. Regret-freeness at a^*'s completion instead gives $C_{h^-} \cup R \in \mathcal{P}_i$. Hence R contains at least one other named participant.

Claim 5 (represented release stage). Let w be the author of a^*. Act-by-act consent requires a completed authorization from every $j \in R \setminus \{w\}$. Each authorization remained unattributed, since a^* first attributes its member of R. If the device had not received one before execution, Assumption 5(iii)'s twin history would make the same act name j where her authorization never completed, violating consent. Each authorization was therefore received before execution. If $w \notin R$, all of R was held; if $w \in R$, all of $R \setminus \{w\}$ was held. Claim 4 makes that set nonempty. The roster is thus released in one act, name by name conditional on completed private authorization, which is the stage asserted in part (ii).

Proof of Proposition 1:

Take two agents a, b with an empty seed, neither safe alone and both safe together, so $C^- = \varnothing$ and $C^+ = \{a, b\}$. If the platform is private and verified but unconditional, the realization in which only a completes leaves her exposed alone. If it is conditional and verified but public, the first public act exposes its mover alone and Theorem 4(i) applies. If it is private and conditional but relies on an unverified intention, the other intended participant may fail, again leaving a singleton. Each construction retains two properties and drops only the named third.

Proof of Corollary 1:

Fix $\mu \in \mathcal{M}$. By the corollary's premise, $\{i\} \notin \mathcal{P}_i(\mu)$, so an intervention that changes only beliefs cannot make her isolated expression safe at that belief. This holds for every belief the intervention can select. Information can still start a cascade by protecting someone else; if it protects i alone, the premise no longer holds.

Proof of Lemma 5:

Fix $i \in E(M)$. Acceptability and $r_i < 1$ give positive conditional probability that i is protected at her first named date. The public set then takes finitely many values, so some $X_i \ni i$ with $X_i \in \mathcal{P}_i$ occurs with positive probability. Every member of X_i is exposed on that event, hence $X_i \subseteq E(M)$. Monotone protection gives $E(M) \in \mathcal{P}_i$. This holds for every $i \in E(M)$, so $E(M)$ is self-protecting and therefore contained in $C^+(S)$ by Theorem 1. Almost every realized exposed set is a subset of $E(M)$.

Proof of Proposition 2:

Number the batch members $1, \dots, n$ with equal masses $a = m_B/n$ and let $X_j \in \{0, a\}$ be j's completed mass, independent with $\Pr(X_j = a) = 1-\varepsilon$. The realized public mass is $Y = y^- + \sum_j X_j$. (i) $\mathbb{E}[Y] = y^- + (1-\varepsilon)m_B$, and since $m_B = (y_t - y^-) + s(y_t)$, $\mathbb{E}[Y] - y_t = (1-\varepsilon)s(y_t) - \varepsilon(y_t - y^-) = \Delta(y_t)$. (ii) Fix a member i with requirement $\rho_i \le y_t$. If $n = 1$, then conditional on $X_i = a$ the realized mass is $y^- + m_B = G_S(y_t) > y_t \ge \rho_i$, so her unprotected probability is zero and the stated bound holds. Now let $n \ge 2$. Conditional on $X_i = a$, her unprotection event is $\{y^- + a + \sum_{j \ne i} X_j < \rho_i\} \subseteq \{\sum_{j \ne i} X_j < (y_t - y^-) - a\}$. The sum has mean $(1-\varepsilon)(n-1)a$, and the gap between mean and threshold is $t^* = (1-\varepsilon)(n-1)a + a - (y_t - y^-) = \Delta + \varepsilon m_B/n \ge \Delta$. Hoeffding's inequality for $n-1$ independent variables with ranges $[0, a]$ gives

$$\Pr(\text{unprotected} \mid X_i = a) \;\le\; \exp\!\Big(-\tfrac{2t^{*2}}{(n-1)a^2}\Big) \;\le\; \exp\!\Big(-\tfrac{2n\Delta^2}{m_B^2}\Big),$$

using $t^* \ge \Delta$ and $(n-1)a^2 \le m_B^2/n$. The unconditional shortfall $\Pr(Y < y_t) = \Pr(\sum_j X_j < m_B - s)$ has mean-to-threshold gap exactly Δ over n such variables, giving the same bound $\exp(-2\Delta^2/(na^2)) = \exp(-2n\Delta^2/m_B^2)$. (With unequal masses $a_j \le \bar{a}$, the same argument gives

$\exp(-2t^{*2}/\sum_j a_j^2)$.) The conditional bound falls below $\underline{r}$ iff $n \geq (m_B^2/2\Delta^2)\ln(1/\underline{r})$, and vanishes as $n \to \infty$ at fixed $\Delta > 0$. (iii) On $\{Y \geq y_t\}$, $Y \in [y_t,\, y^- + m_B] = [y_t,\, G_S(y_t)] \subseteq (y^u,\, y^+]$, since G_S is nondecreasing with $G_S(y^+) = y^+$. From public mass Y, the iteration $y \mapsto G_S(y)$ is nondecreasing (Lemma 1 in scalar form) and, with G_S continuous in the nonatomic benchmark, $G_S > \mathrm{id}$ on (y^u, y^+), and $G_S(y^+) = y^+$, converges from Y to y^+, the least fixed point weakly above Y. Each round is realized by one-at-a-time public entries—including safe retries by pact members whose attempts failed—of consenters whose requirements the current mass already clears, each safe at the moment she moves, so the resumed path is regret-free and its closure is $y^+ = b + m(C^+(S))$: Theorem 2(i) in scalar form, with the realized set as the enlarged seed. Combining with the shortfall bound in (ii) gives the composite statement.

Proof of Proposition 3:

(i) On the no-leak event, the target C is eventually released and protects every member. Each $i \in C$ is exposed almost surely, either prematurely through a breach or in the protecting release. Her mechanism-wide conditional unprotected probability is therefore exactly the harmful-leak contribution δ_i^{h}. The bound on breach probability gives $\delta_i^{\mathrm{h}} \leq \delta$, proving both criteria. (ii) Condition throughout on $X_i = \{i \text{ is exposed}\}$. Let H_i be the no-leak trigger event whose conditional probability is m_i. Completion draws are independent of the trigger history, so on an $m_i\varepsilon^{|B_i|}$ share of X_i, agent i completes while every member of B_i fails. She is then named at a state $C \cup \{i\} \notin \mathcal{P}_i$. This event is disjoint from a harmful leak, whose conditional probability is δ_i^{h}. Thus

$$\Pr(i \text{ unprotected} \mid X_i) \;\geq\; \delta_i^{\mathrm{h}} + m_i\varepsilon^{|B_i|}.$$

Acceptability gives the displayed inequality. The same argument adds over disjoint trigger events. (iii) At $r = 0$, the nonnegative risk contributions in part (ii) vanish: no harmful leak and no unverified not-yet-safe trigger with positive conditional mass can occur. Since every completion pattern has positive probability under $\varepsilon \in (0, 1)$, probability-one safety coincides with safety at every admissible realization for almost every realization of the mechanism's randomization. With $C^- \subsetneq C^+$, Theorem 4(ii) then delivers the private conditional-release stage. Along $r^n \to 0$, the inequality implies that $\delta_{i,n}^{\mathrm{h}}$ and each fixed $m_{i,n}\varepsilon^{|B_i|}$ converge to zero; neither term must equal zero at positive tolerance.

Proof of Theorem 5:

All probability statements use the article's convention: platform randomization is assessed almost surely, while every finite completion pattern is covered conditional on each positive-probability realization of that randomization. (i) *Protection of T.* By hypothesis, $T \in \mathcal{P}_a$ for every not-yet-safe participant $a \in \{i\} \cup B_i$. Already-safe participants ($C \cup \{a\} \in \mathcal{P}_a$) are protected in T and in every realized superset of $C \cup \{a\}$ by Assumption 1, under either mechanism. *Risk.* Condition on reaching the trigger. Under the replacement, no participant is public before the release, and the release exposes each participant exactly once, into T, which protects her by the previous paragraph. The replacement therefore generates no unprotected exposure at this trigger. Under M, each participant's trigger-generated probability is nonnegative, so the comparison is weak. For i it is strict: on independent completion draws, her attempt completes while every member of B_i fails with probability $(1-\varepsilon)\,\varepsilon^{|B_i|}$. The public set when she is named is then $C \cup \{i\} \notin \mathcal{P}_i$. Conditional on her attempt completing, that harmful event has probability $\varepsilon^{|B_i|}$. Retries by M do not lower either local bound because the harm occurs when she is first named. The local comparison does not multiply by the probability of reaching C or combine this trigger with other routes through which i may be exposed. *Reach.* Give both processes the same agent-date completion draws. Whatever subset $R \subseteq \{i\} \cup B_i$ the trigger under M ever exposes, $C \cup R \subseteq T$; any member protected at M's terminal set $C \cup R$ is protected at T by Assumption 1. Under the replacement, re-solicited commitments complete in finite time almost surely because each new draw has completion probability $1 - \varepsilon > 0$. The terminal set is therefore T, and all its members are protected. The replacement's terminal

protected set contains M's in every realization and adds the members of T who did not complete under M. With a one-shot replacement, release occurs only when all authorizations complete, so this reach comparison fails after a partial completion. (ii) The tunnel rule solicits commitments from all of S, re-soliciting failures. The signed set reaches S almost surely in finite time ($|S|$ geometric variables). No one is public before release, and the single release act exposes exactly $C^+(S)$, which is self-protecting by Theorem 1. Every agent's unprotected-exposure probability is therefore zero, and the rule is acceptable at every profile. For any interim-acceptable M and almost any path, Lemma 5 gives terminal protected mass at most $m(C_\infty) \leq m(E(M)) \leq m(C^+(S))$. This proves both the risk and reach comparisons. (iii) *Ceiling and attainment.* Every completion realization has positive probability conditional on each positive-probability realization of the platform's randomization because $\varepsilon \in (0,1)$. By Proposition 3(iii), zero risk therefore implies regret-free safety for almost every realization of that randomization. Theorem 1 then bounds every released set by $C^+(S)$. By (ii), the tunnel attains $m(C^+(S))$ almost surely.

Only by. Suppose a zero-risk mechanism releases mass $m(C^+(S))$ almost surely. Because every agent has strictly positive mass, the released set must be $C^+(S)$ itself for almost every realization of the platform's randomization. Since $C^- \subsetneq C^+$, that set lies outside 2^{C^-}. Theorem 4(ii) therefore requires a private conditional-release stage.

Zero-risk endpoint. At $r = 0$, acceptability requires zero risk, so the preceding conclusion turns dominance into necessity. For a sequence M_n acceptable at $r^n \to 0$, acceptability gives $\Pr_{M_n}(i \text{ unprotected} \mid i \text{ exposed}) \leq r_i^n \to 0$ for every agent. This is convergence of risks, not an assertion that any M_n has zero risk at positive tolerance.

LEMMA 6 (Path coupling): *Under the state-independent completion technology of Section III, fix a public state C, an agent $i \in T(C) \setminus C$, and a Markov profile that is greedy at every public state strictly containing C and arbitrary but robustly admissible at C. Run two copies of the process from C on identical completion draws at every date: in copy A, agent i's date-0 attempt completes; in copy W, agent i waits whenever the state is C. Then the public sets satisfy $D_t^A \supseteq D_t^W \cup \{i\}$ for every $t \geq 0$, realization by realization.*

Proof of Lemma 6:

Index states at the end of each date. At date 0 both copies sit at C and every agent other than i takes the same profile action at C in both copies with shared draws; write R_0 for the common set of their date-0 completers, so $D_0^A = C \cup \{i\} \cup R_0 \supseteq C \cup R_0 \cup \{i\} = D_0^W \cup \{i\}$. Suppose $D_t^A \supseteq D_t^W \cup \{i\}$, and take any $j \neq i$ attempting at D_t^W in copy W with $j \notin D_t^A$. If $D_t^W = C$, robust admissibility of the profile at C forces $j \in T(C) \setminus C$; if $D_t^W \supsetneq C$, greediness gives $j \in T(D_t^W) \setminus D_t^W$. Either way $j \in T(D_t^W)$, so $j \in T(D_t^A) \setminus D_t^A$ by Lemma 1 and $D_t^W \subseteq D_t^A$; since $D_t^A \supseteq C \cup \{i\} \supsetneq C$, play in copy A is greedy there and j attempts as well. The completion draws are shared—indexed by agent and date, consulted only when that agent attempts—so every agent newly public in copy W at date $t+1$ is either newly public in copy A or already public there, and the inclusion is preserved.

Proof of Theorem 6:

Fix a robust Markov-perfect equilibrium and a public state C, and take $i \notin C$. If $i \notin T(C)$, then $C \cup \{i\} \notin \mathcal{P}_i$. The public attempts in this game are unverified, so Assumption 2 admits the realization in which only i completes, exposing her in $C \cup \{i\} \notin \mathcal{P}_i$. Attempting is therefore not robustly admissible, so no equilibrium has i attempt.

Now suppose $i \in T(C) \setminus C$, so $C \cup \{i\} \in \mathcal{P}_i$. By Lemma 2, every admissible realization $R \ni i$ satisfies $C \cup R \supseteq C \cup \{i\}$. Assumption 1 then gives $C \cup R \in \mathcal{P}_i$. Attempting is robustly admissible and safe in every realization, while waiting pays 0 this period.

Use backward induction on $|S \setminus C|$, taking greedy play at every strict superset of C as the induction hypothesis. Couple attempting and waiting with identical completion draws at every date. If i fails and no one else completes, the state remains C and the comparison repeats. If

another agent completes while i fails, the paths coincide thereafter because i's strategies differ only at C and public states never shrink. It remains to compare i's completion with waiting.

After i completes, Lemma 6 gives $D'_t \supseteq D_t \cup \{i\}$ at every later date; its greedy-play hypothesis at strict supersets of C is the induction hypothesis. When the waiting path first names i in $D_t \cup \{i\}$, the completion path pays $u_i(D'_t) \geq u_i(D_t \cup \{i\})$ by monotonicity of u_i. At every earlier date it pays $u_i(\cdot) \geq 0$ against waiting's 0. Assumption 1 preserves safety along the completion path from $C \cup \{i\} \in \mathcal{P}_i$. Completion is thus weakly better in every realization. With probability $p_i > 0$, attempting produces an immediate flow at least $u_i(C \cup \{i\}) > 0$ under strict safe surplus, so attempting is strictly optimal in expectation.

The unique equilibrium action at C therefore has exactly $T(C) \setminus C$ attempt. Hence $C_t \subseteq T(C_{t-1})$, and Lemma 1 implies by induction that the path remains within C^-. It cannot stop at $C \subsetneq C^-$. To see this directly, if $T(C) = C$, monotonicity of T and $\varnothing \subseteq C$ imply $T^n(\varnothing) \subseteq C$ for every n, and hence $C^- \subseteq C$. Thus no proper subset of C^- is T-closed: $T(C) \setminus C \neq \varnothing$, and those agents attempt. Since recurrent attempts have independent completion probability at least $p > 0$, each safe agent is named in finite time almost surely, and the public set reaches C^- a.s.

Proof of Proposition 4:

Fix i, the others' choices V_{-i}, and write $V = V_{-i} \cup \{i\}$. A preliminary observation (*pivotality*): $i \in C^+(W)$ iff some self-protecting roster containing i lies inside W; hence if $i \notin C^+(V)$, every self-protecting subset of V avoids i, so $C^+(V) = C^+(V_{-i})$. An unreleased signature never changes the release.

(i) Abstaining: by consent no act publishes an unsigned name, and the announced release is the only naming event, so i is unpublished and earns 0. Signing: if $i \notin C^+(V)$, her signature is never disclosed (failure privacy) and she earns 0; if $i \in C^+(V)$, she is published in $C^+(V)$, self-protecting by Theorem 1's union-closure clause applied to the self-protecting subsets of V, so $C^+(V) \in \mathcal{P}_i$ and she earns $u_i(C^+(V)) \geq 0$. Signing pays at least the abstain payoff against every profile, realization by realization. In the one-shot game, the no-conditioning premise rules out two effects: audience response to the signature itself and conditioning of strategies or protection families on signatures. Only in a dynamic extension would that premise also fix the realized V_{-i}.

(ii) For $i \in C^+(S)$: against $V_{-i} = S \setminus \{i\}$, $C^+(V) = C^+(S) \ni i$, so signing yields $u_i(C^+(S)) > 0$ under strict safe surplus against 0 from abstaining; with (i), signing weakly dominates in the standard sense. For $i \notin C^+(S)$: $C^+(V) \subseteq C^+(S)$ never contains i, so by pivotality both actions leave her unpublished and the release unchanged.

(iii) At $V = S$ the rule releases $C^+(S)$ (Theorem 3); members of $C^+(S)$ strictly lose by abstaining and outsiders are indifferent, so everyone-signs is Nash, and by (ii) no player uses a weakly dominated strategy. Abstaining is weakly dominated exactly for members of $C^+(S)$ (weakly worse everywhere by (i), strictly worse against $S \setminus \{i\}$), and signing is never weakly dominated, so an equilibrium avoids weakly dominated strategies iff every member of $C^+(S)$ signs. For any $V \supseteq C^+(S)$, monotonicity and Theorem 1 give $C^+(S) \subseteq C^+(V) \subseteq C^+(S)$: the release is exactly $C^+(S)$. Finally, all-abstain is Nash iff no lone deviation profits; a deviation by i releases $C^+(\{i\})$, which is $\{i\}$ if i is safe alone given the seed and empty otherwise, so under strict safe surplus all-abstain is Nash exactly when no agent is safe alone given the seed.

LEMMA 7 (Interim safety inherits the lattice): *Fix the interim criterion: i accepts release in C iff $e_i \geq \pi(q(C))\, h_i$, with π nonincreasing and q nondecreasing in the coalition. Define $\mathcal{P}_i^{\text{int}} = \{C \ni i : e_i \geq \pi(q(C))\, h_i\}$. Then each $\mathcal{P}_i^{\text{int}}$ satisfies Assumption 1, so Γ, T, C^-, and C^+ are defined as before. Theorems 1, 2, 3, and 4 apply with $\mathcal{P}_i^{\text{int}}$ in place of $\mathcal{P}_i$ under each theorem's original additional conditions: unilateral-completion uncertainty for the cascade; atomic release for the tunnel; and unilateral-completion uncertainty, E1, and Definition 2 for the representation theorem. The interim payoff $e_i - \pi(q(C))\, h_i$ is nondecreasing in the coalition. Theorem 6 therefore also carries over under its completion technology and separate strict-safe-surplus hypothesis, that*

is, if $e_i > \pi(q(C))\, h_i$ on every coalition treated as protecting i.

PROOF OF LEMMA 7:

If $C \subseteq C'$ and $i \in C$ with $e_i \geq \pi(q(C))h_i$, then $q(C') \geq q(C)$ and $\pi(q(C')) \leq \pi(q(C))$, so $e_i \geq \pi(q(C'))h_i$. Thus each $\mathcal{P}_i^{\text{int}}$ is closed under supersets containing i: Assumption 1 holds for the induced families. Substitution of these families changes none of the model's completion, atomic-release, E1, or mechanism primitives. The original proofs therefore apply whenever their other stated assumptions are maintained. The interim payoff $e_i - \pi(q(C))h_i$ is nondecreasing because $\pi \circ q$ is nonincreasing. This supplies the payoff-monotonicity premise of Theorem 6; its completion technology and strict-safe-surplus premise must still be imposed separately.

PROOF OF PROPOSITION 5:

Open organizing realizes C^-, and $G \cap C^- = \emptyset$, so each $i \in G$ never surfaces. Her expression does not both surface and survive (effective suppression 1), and she draws no retaliation because the opponent has no roster entry to act on. Failure-private release realizes C^+, which is self-protecting, so $e_i \geq \pi(q(C^+))\, h_i$ for every member and all of G accept. Each hidden supporter surfaces and is fought with probability $1 - F(q(C^+))$ by the resolve cutoff (eq. 2). Conditional on a fight, her expression is suppressed with probability $1 - q(C^+)$, for an unconditional suppression probability $(1-F(q(C^+)))(1-q(C^+)) = \pi(q(C^+))$. Comparing the two regimes, observed retaliation rises from 0 and effective suppression falls from 1 to $\pi(q(C^+))$; both changes are strict because $\pi(q(C^+)) \in (0,1)$.

PROOF OF COROLLARY 2:

For every hidden supporter, Proposition 5 raises recorded retaliation from zero to a positive probability while lowering effective suppression from one to $\pi(q(C^+)) \in (0,1)$. The corollary is this comparison for the latent subgroup G; it does not identify that subgroup from raw campaign records.

*

REFERENCES

Admati, Anat R., and Motty Perry. 1991. "Joint Projects without Commitment." *The Review of Economic Studies*, 58(2): 259.

Akbarpour, Mohammad, and Shengwu Li. 2020. "Credible Auctions: A Trilemma." *Econometrica*, 88(2): 425–467.

Ali, S. Nageeb, and Roland Bénabou. 2020. "Image versus Information: Changing Societal Norms and Optimal Privacy." *American Economic Journal: Microeconomics*, 12(3): 116–164.

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.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96(5): 1652–1678.

Bendor, Jonathan, and Piotr Swistak. 2001. "The Evolution of Norms." *American Journal of Sociology*, 106(6): 1493–1545.

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

Bronfenbrenner, Kate. 2009. "No Holds Barred: The Intensification of Employer Opposition to Organizing." Economic Policy Institute Briefing Paper 235, Washington, DC.

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.

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.

Cashman, Matthew. 2026. "Conditional Disclosure as a Coordination Device." Working paper, arXiv:2604.00874.

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.

Eaton, Adrienne E., and Jill Kriesky. 2001. "Union Organizing under Neutrality and Card Check Agreements." *Industrial and Labor Relations Review*, 55(1): 42–59.

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

Eliaz, Kfir. 2002. "Fault Tolerant Implementation." *The Review of Economic Studies*, 69(3): 589–610.

Fehr, Ernst, and Urs Fischbacher. 2004. "Third-Party Punishment and Social Norms." *Evolution and Human Behavior*, 25(2): 63–87.

Gale, Douglas. 2001. "Monotone Games with Positive Spillovers." *Games and Economic Behavior*, 37(2): 295–320.

Granovetter, Mark. 1978. "Threshold Models of Collective Behavior." *American Journal of Sociology*, 83(6): 1420–1443.

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

LaLonde, Robert J., and Bernard D. Meltzer. 1991. "Hard Times for Unions: Another Look at the Significance of Employer Illegalities." *The University of Chicago Law Review*, 58(3): 953–1014.

Lohmann, Susanne. 1994. "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989–91." *World Politics*, 47(1): 42–101.

McNicholas, Celine, Margaret Poydock, Julia Wolfe, Ben Zipperer, Gordon Lafer, and Lola Loustaunau. 2019. "Unlawful: U.S. Employers Are Charged with Violating Federal Law in 41.5% of All Union Election Campaigns." Economic Policy Institute, Washington, DC.

Milgrom, Paul, and John Roberts. 1990. "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities." *Econometrica*, 58(6): 1255.

Noelle-Neumann, Elisabeth. 1974. "The Spiral of Silence a Theory of Public Opinion." *Journal of Communication*, 24(2): 43–51.

Rajan, Anjana, Lucy Qin, David W. Archer, Dan Boneh, Tancrede Lepoint, and Mayank Varia. 2018. "Callisto: A Cryptographic Approach to Detecting Serial Perpetrators of Sexual Misconduct." *COMPASS '18*, 1–4. New York, NY, USA:Association for Computing Machinery.

Schelling, Thomas C. 1978. *Micromotives and Macrobehavior.* New York:W. W. Norton & Company.

Schmitt, John, and Ben Zipperer. 2009. "Dropping the Ax: Illegal Firings During Union Election Campaigns, 1951–2007." Center for Economic and Policy Research, Washington, DC.

Segal, I. 1999. "Contracting with Externalities." *The Quarterly Journal of Economics*, 114(2): 337–388.

Sherk, James. 2007. "The Truth About Improper Firings and Union Intimidation." Heritage Foundation WebMemo 1393, Washington, DC.

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

Topkis, Donald M. 1998. *Supermodularity and Complementarity.* Princeton, NJ:Princeton University Press.

Zubrickas, Robertas. 2014. "The Provision Point Mechanism with Refund Bonuses." *Journal of Public Economics*, 120: 231–234.

Except in the count-release variant below, the article's terminology applies: to "name" an agent is to publish her fixed, identity-bound statement, report, or other expression. A public roster records whose expressions have been published; it is not an announcement followed by a later choice to act.

Count Release: Supporting Result

This section records the count-release variant of the tunnel used in the applications. The practical trigger is an authenticated total count or a small vector of preannounced class-count quotas. The general protection-family rule identifies the greatest safe roster conditional on the supplied requirements. It does not claim that a platform can elicit or evaluate arbitrary requirements that depend on roster composition.

Count release.

Some deployed mechanisms release a *number* rather than jointly publishing statements under their authors' identities. The protective audience receives an authenticated count of sealed authorizations, while the retaliating audience sees no name even on success. Call a mechanism *count-release* if it holds completed authorizations as before but publishes only the certified mass of the released coalition. The roster goes only to the certifier, who is bound to keep it confidential. In the scalar benchmark, protection depends on mass: the threshold families $\mathcal{P}_i = \{C \ni i : b + m(C) \geq \rho_i\}$ are upper sets in mass and do not depend on identities. The certified count is therefore the statistic on which the rule conditions. Whether that certificate protects a later individual act lies outside the proposition.

PROPOSITION S1 (Count release): *In the nonatomic scalar benchmark with the threshold families above:*

(i) The threshold families are upper sets in mass, so Assumption 1 holds and Theorems 1 and 2 apply. Their least and greatest fixed points take the scalar form in equation (1): C^- and C^+ collect the consenters with requirements at or below y^- and y^+.

(ii) Conditional on $V = S$, the tunnel rule with count-only publication assembles and certifies the same mass y^+. Because count release never names an agent, the paper's naming-based safety requirement is vacuously satisfied. The rule certifies only masses for which $b + m \geq \rho_i$ for every member of $D(V)$. This self-protecting-roster calculation determines the mass certified but does not establish that the certificate protects a later self-identifying act, which lies outside the mechanism. The proposition assumes private accumulation and therefore does not cover observable signing. The companion paper discusses related exposure and observability risks, but neither paper solves a general observable-signing equilibrium (Cashman, 2026). Mere certifier knowledge is not a harmful leak. A breach that transmits an under-protective roster to the retaliating audience instead enters the harmful-breach term δ_i^{h} in Proposition 3.

(iii) Theorem 4 concerns histories that enlarge the public named coalition, which count release never does. Count release still uses completed private authorizations; it changes the released object from names to certified mass. It is therefore not a public-only route past C^-. Publishing only a count withholds the roster mechanically but does not guarantee anonymity. A small candidate set can concentrate the audience's posterior, an exact certified mass can identify a unique roster when masses are atomic, and a count equal to the whole candidate population identifies everyone. Inference depends on candidate sets, side information, and priors that the model does not track.

Proof of Proposition S1:

(i) Inclusion weakly increases mass. Thus $C \in \mathcal{P}_i$ and $C \subseteq C' \ni i$ imply $C' \in \mathcal{P}_i$, so Assumption 1 holds. Equation (1) gives the fixed-point specialization; finite-step attainment becomes monotone convergence in the nonatomic benchmark.

(ii) Run $D(V) = C^+(V)$ with the publication map replaced by the certified mass $b+m(D(V))$. No history names anyone to the retaliating audience, so Assumption 3 imposes no naming restriction. Conditional on $V = S$, the certified mass is $b + m(C^+(S)) = y^+$, with $C^+(S)$ self-protecting by Theorem 1. The rule certifies only self-protecting masses, so $b+m(D(V)) \geq \rho_i$ for every $i \in D(V)$. Whether that certificate protects an agent who later identifies herself, and the later act itself, lie outside the mechanism.

(iii) Both parts of Theorem 4 concern histories at which agents become newly *named*. Part (i) bounds the public named set, while part (ii) derives the private stage from the transition that first names an unsafe agent. Count release has no such history. It uses the same holding stage for a different object, a mass certificate.

The Fight-or-Fold Opponent: Per-Member Result

The fight-or-fold model in Section IV treats retaliation as all-or-none. This section shows that its reversal survives when the opponent makes the punishment decision member by member. Punishing one member of a released coalition of mass m costs $\kappa_T \psi(m)$, with ψ continuous, strictly increasing, and positive. Punishment prevents the fixed harm $d_i > 0$ that member i's participation would impose on the opponent. The opponent therefore punishes i exactly when $d_i \geq \kappa_T \psi(m)$. Its cost type κ_T is drawn once from a continuous, strictly increasing cdf Λ with full support on $(0, \infty)$. A punished member's expression is suppressed; an unpunished member's expression stands.

PROPOSITION S2 (Per-member reversal and selective retaliation): *Let $p_i(m) = \Lambda\big(d_i/\psi(m)\big)$ be the probability that member i of a released mass-m coalition is punished, let $h_i > 0$ and $e_i \geq 0$, and let i accept release iff $e_i \geq p_i(m)\, h_i$. Then: (i) acceptance is a mass threshold: i accepts iff $m \geq \mu_i$. When $e_i = 0$, $\mu_i = \infty$. When $0 < e_i < h_i$, $\mu_i = \psi^{-1}\big(d_i/\Lambda^{-1}(e_i/h_i)\big)$. If the argument exceeds ψ's range, set $\mu_i = \infty$, so the agent never accepts; if it falls below the range, set $\mu_i = 0$. When $e_i \geq h_i$, $\mu_i = 0$. At a common tolerance e/h, the cutoff μ_i is weakly increasing in d_i. The induced protection families are upward sets, and C^-, C^+ exist by Lemma 7; (ii) for every hidden supporter $i \in C^+ \setminus C^-$, replacing open organizing with release of C^+ raises observed retaliation on her from 0 to $p_i(m(C^+)) \in (0,1)$ and lowers suppression of her expression from 1 to the same interior number: the reversal holds member by member; (iii) every already-visible member $j \in C^-$ is punished with weakly lower probability after the release, $p_j(m(C^+)) \leq p_j(m(C^-))$; (iv) at any released mass, punishment concentrates on the most damaging members: $p_i(m)$ is increasing in d_i.*

Proof of Proposition S2:

(i) $e_i \geq p_i(m)h_i$ iff $\Lambda(d_i/\psi(m)) \leq e_i/h_i$. If $e_i = 0$, the left side is false at every finite mass because $d_i/\psi(m) > 0$ and full support makes $\Lambda(d_i/\psi(m)) > 0$, so $\mu_i = \infty$. If $e_i \geq h_i$ the inequality holds for every m since $\Lambda \leq 1$. For $0 < e_i < h_i$, Λ continuous and strictly increasing with full support gives: acceptance iff $d_i/\psi(m) \leq \Lambda^{-1}(e_i/h_i)$, i.e., iff $\psi(m) \geq d_i/\Lambda^{-1}(e_i/h_i)$, a mass threshold μ_i increasing in d_i and decreasing in e_i/h_i. Acceptance sets are upward in mass, hence upward in inclusion, and Lemma 7 delivers the lattice objects. (ii) Under open organizing a hidden supporter never surfaces: no punishment against her can be recorded, and her expression never stands, so observed retaliation is 0 and suppression is 1. Under release of C^+, which is self-protecting under the induced families, she accepts and is punished iff $\kappa_T \leq d_i/\psi(m(C^+))$, an event of probability $p_i(m(C^+)) = \Lambda(d_i/\psi(m(C^+))) \in (0,1)$ because $d_i > 0$ and Λ is interior on $(0,\infty)$. Her expression is suppressed exactly when she is punished. Observed retaliation thus rises from 0 to $p_i(m(C^+))$ while suppression falls from 1 to $p_i(m(C^+))$. (iii) $m(C^+) \geq m(C^-)$ and ψ increasing give $d_j/\psi(m(C^+)) \leq d_j/\psi(m(C^-))$; apply Λ. (iv) Immediate since Λ is increasing.

The coalition-level fight-or-fold choice in eq. 2 is the constrained case in which punishment must be all-or-none. The per-member model shows that the reversal does not depend on that constraint, and part (iv) gives the selective-discharge pattern documented in the union evidence. The proposition also shows that the minimum coalition mass at which a member accepts release, μ_i, rises with the harm d_i she imposes on the opponent.

The observable-attempt comparative static below uses one fixed-point lemma.

LEMMA S1 (Shifted crossings): *Let $b \le M$ and let $G, \tilde{G} : [b, M] \to [b, M]$ be nondecreasing with $\tilde{G} \le G$ pointwise and $G(b) \ge b$, $\tilde{G}(b) \ge b$.*

*(i) (*Monotonicity.*) Least fixed points exist (Tarski), and* $\mathrm{lfp}(\tilde{G}) \le \mathrm{lfp}(G)$.

*(ii) (*Continuity.*) Let $\{G_s\}_{s\in[0,\bar{s}]}$ be pointwise nonincreasing in s, with each G_s continuous on $[b, M]$ and $(y, s) \mapsto G_s(y)$ jointly continuous. Then $s \mapsto \mathrm{lfp}(G_s)$ is nonincreasing and right-continuous, and it is continuous at any s_0 at which the least fixed point is a* transversal *crossing—$G_{s_0} < \mathrm{id}$ on an interval immediately above $\mathrm{lfp}(G_{s_0})$. Without transversality, a new fixed point can appear by tangency below the previous least fixed point, causing a downward discontinuity from the left.*

Proof of Lemma S1:

(i) For nondecreasing G on the complete lattice $[b, M]$, Tarski gives $\mathrm{lfp}(G) = \inf\{y : G(y) \le y\}$; the set is nonempty because $G(M) \le M$. Since $\tilde{G} \le G$, every pre-fixed point of G is a pre-fixed point of $\tilde{G}$, ordering the infima. (ii) Monotonicity is (i). Write $\ell(s) = \mathrm{lfp}(G_s)$, which is nonincreasing. Fix s_0 and $\eta > 0$. For $y < \ell(s_0)$, $G_{s_0}(y) > y$, so $G_{s_0} - \mathrm{id}$ has a positive minimum on the compact $[b, \ell(s_0) - \eta]$ when it is nonempty. Joint continuity keeps $G_s > \mathrm{id}$ there for s near s_0, whence $\ell(s) > \ell(s_0) - \eta$; with monotonicity this gives right-continuity. Under transversality, choose $y^* \in (\ell(s_0), \ell(s_0) + \eta)$ with $G_{s_0}(y^*) < y^*$. Joint continuity keeps $G_s(y^*) < y^*$ nearby, so $\ell(s) \le y^* < \ell(s_0) + \eta$. The bounds give continuity. A tangency below the running least fixed point shows why transversality cannot be dropped.

Frontier Supporting Results

This section states and proves the interim bounds and the observable-attempt comparative static summarized in the article.

PROPOSITION S3 (Necessity at positive tolerance): *(i) If $C^-(S) \subsetneq C^+(S)$ and $r_i \ge 1-(1-\varepsilon)^{|B|}$ for every agent i of the nonempty batch $B = C^+(S) \setminus C^-(S)$, a single unverified synchronized public batch is acceptable to every member and realizes $C^+(S)$ with probability $(1-\varepsilon)^{|B|} > 0$: with tolerant participants, a private stage is not necessary to reach beyond the cascade. (ii) Fix a mechanism and agent i. Let X_i be the event that i is exposed. Suppose that at a trigger h, reached before the same-date completion draws, i is not safe with the current public set alone and her own exposure attempt rests on an unverified batch $B_h \ne \varnothing$. Let H_{ih} be the event that h is reached and i's attempt completes, and write $\alpha_{ih} = \Pr(H_{ih} \mid X_i)$. Then*

$$\Pr(i \text{ is exposed unprotected} \mid X_i) \;\ge\; \alpha_{ih}\varepsilon^{|B_h|}.$$

Consequently acceptability requires $\alpha_{ih}\varepsilon^{|B_h|} \le r_i$. For pairwise disjoint trigger events, the corresponding lower bounds add. Safe exposure on other branches can make $\alpha_{ih} < 1$, so $r_i < \varepsilon^{|B_h|}$ alone does not rule out the trigger. (iii) At the exact endpoint $r = 0$, every positive-mass not-yet-safe unverified trigger is excluded; interim acceptability is regret-freeness and Theorem 4 applies. Along any sequence of acceptable mechanisms with $r^n \to 0$, each agent's mechanism-wide unprotected-exposure risk converges to zero, although it may remain positive at every n.

PROOF OF PROPOSITION S3:

(i) Let $B = C^+(S) \setminus C^-(S)$ be announced and attempted at once from the public state $C^-(S)$. For $i \in B$, conditional on her own completion, the realizations leaving her unprotected are contained in the event that not all of $B \setminus \{i\}$ complete, of probability $1-(1-\varepsilon)^{|B|-1} \le 1-(1-\varepsilon)^{|B|} \le r_i$, so she accepts; if all complete, the realized public set is $C^+(S) \in \mathcal{P}_i$ for every member, an event of probability $(1-\varepsilon)^{|B|}$.

(ii) Condition on X_i. On H_{ih}, completion independence gives probability $\varepsilon^{|B_h|}$ that every member of B_h fails. In that realization i is named in the current public set plus herself, which is outside $\mathcal{P}_i$ by hypothesis. Hence the harmful event has conditional probability at least $\Pr(H_{ih} \mid X_i)\varepsilon^{|B_h|} = \alpha_{ih}\varepsilon^{|B_h|}$. Acceptability yields the inequality, and disjoint trigger events can be summed. Safe exposure on other branches lowers this trigger's share of agent i's total exposure, making $\alpha_{ih} < 1$, but does not change the probability that every member of B_h fails at the trigger.

(iii) At $r = 0$, part (ii) rules out every unverified not-yet-safe trigger with $\alpha_{ih} > 0$. Every completion pattern has positive probability, so zero mechanism-wide risk is regret-freeness on almost every branch of the mechanism's randomization; Theorem 4 then applies. Along $r^n \to 0$, acceptability directly bounds each mechanism-wide conditional risk by r_i^n, proving convergence without exact zero at any positive tolerance.

Now suppose failed attempts may be observed. Each *failed* attempt is observed by the retaliating audience with probability $\omega \in [0,1]$, independently across attempts. Observing one unit of failed-attempt mass raises each still-private consenter's future requirement by $\xi \ge 0$, but it does not change the protection of agents already named. The process follows the greedy path in Theorem 6 with the completion model of Proposition S3. Each attempt completes independently with probability $1-\varepsilon$ and is retried until it completes, so each unit of completed mass generates $\varepsilon/(1-\varepsilon)$ units of failed attempts in expectation. The fluid benchmark evaluates accumulated failures at this expectation. Building public mass y from the seed then raises future requirements by $\lambda(\omega)\,(y-b)$, where $\lambda(\omega) = \omega\xi\varepsilon/(1-\varepsilon)$. A consenter with requirement ρ_i is safe to join at mass y iff $y \ge \rho_i + \lambda(\omega)(y-b)$. The curve in equation (1) becomes

$$G_S^\omega(y) \;=\; b + H_S\big((1-\lambda(\omega))\,y + \lambda(\omega)\,b\big), \qquad \lambda(\omega) < 1,$$

still nondecreasing in y, and pointwise weakly falling in ω.

PROPOSITION S4 (Observability and the stall): *Consider the scalar nonatomic benchmark, in which H_S is continuous, and use the fluid interpretation above. Maintain the baseline crossing configuration used for the targeted pact: $y^- < y^u < y^+$ and $G_S < \mathrm{id}$ on (y^-, y^u). Let $\lambda(\omega) = \omega\xi\varepsilon/(1-\varepsilon) < 1$, with observed failures affecting only future requirements:*

(i) (The stall moves down; no jump at the onset.) *The open greedy path stalls at $y^-(\omega)$, the least fixed point of G_S^ω above the seed. On the range of ω with $\lambda(\omega) < 1$, $y^-(\omega)$ is nonincreasing in ω (and in ξ, and in ε through λ), right-continuous, and continuous at every ω whose stall is a transversal crossing—$G_S^\omega < \mathrm{id}$ immediately above $y^-(\omega)$—as at $\omega = 0$, where the standing configuration puts $G_S < \mathrm{id}$ on (y^-, y^u); at an interior ω where the rescaled curve instead touches the diagonal below the running stall, $y^-(\omega)$ can step down discretely (Lemma S1(ii)). $y^-(0) = y^-$ recovers Theorem 6's stall exactly, as does $\xi = 0$ (observed failures do not raise future requirements) or $\varepsilon = 0$ (nothing fails, so nothing is seen).*

(ii) (Marginal erosion.) *If H_S is continuously differentiable near y^- with $H_S'(y^-) < 1$, then*

$$\frac{dy^-}{d\omega}\Big|_{\omega=0} \;=\; -\,\frac{\xi\varepsilon}{1-\varepsilon}\cdot\frac{H_S'(y^-)\,\big(y^- - b\big)}{1-H_S'(y^-)} \;\le\; 0,$$

strict whenever $\xi\varepsilon > 0$, $H_S'(y^-) > 0$, and $y^- > b$: the stall prediction degrades smoothly in the observability of failure, not discretely.

When $\omega > 0$, a failed attempt raises the future requirements faced by the attempter and everyone else. The costless-failure condition in Theorem 6 therefore fails. The proposition characterizes the fluid greedy path, not equilibrium selection for $\omega > 0$. Equilibrium analysis and the finite-population stochastic path remain open.

Proof of Proposition S4:

(i) For $\lambda < 1$, the map $y \mapsto (1-\lambda)y + \lambda b$ is increasing. Hence G_S^ω is nondecreasing and maps $[b, b + m(S)]$ into itself, with $G_S^\omega(b) = b + H_S(b) \geq b$. For $y \geq b$, its argument $(1-\lambda)y + \lambda b = y - \lambda(y-b)$ is nonincreasing in λ. The parameter λ is increasing in ω, ξ, and ε on $[0,1)$, so G_S^ω is pointwise nonincreasing in each of them because H_S is nondecreasing.

The condition $y \geq \rho_i + \lambda(y-b)$ is equivalent to $\rho_i \leq (1-\lambda)y + \lambda b$. Because $\lambda < 1$, the right-hand threshold rises along the path. The mass joined by y is therefore $H_S((1-\lambda)y + \lambda b)$, and the fluid open path iterates G_S^ω from the seed. Continuity of H_S implies convergence to $y^-(\omega) = \mathrm{lfp}(G_S^\omega)$. In the nonatomic round-by-round process, a joiner's own failed attempts have zero mass, and expected cumulative requirement increases after each round equal λ times completed mass. Thus iteration of G_S^ω is exact for this fluid recursion.

Lemma S1(i)–(ii) now gives monotonicity and continuity; joint continuity in (y, ω) follows by composition. At $\omega = 0$, the stall is transversal because $G_S < \mathrm{id}$ on (y^-, y^u). At an interior ω, transversality must be checked for G_S^ω at its new stall. It does not follow from the benchmark because $y^-(\omega)$ may move into the region where $G_S > \mathrm{id}$. If the graph of H_S is tangent below the stall to the ray with slope $1/(1-\lambda)$, the least fixed point moves down discretely. The continuity claim therefore applies exactly at transversal values of ω. At $\lambda = 0$, the curve is G_S and its least fixed point is y^-. (ii) $F(y, \lambda) = y - b - H_S((1-\lambda)y + \lambda b)$ is C^1 near $(y^-, 0)$ with $F_y = 1 - H_S'(y^-) > 0$ there, so the implicit function theorem gives a C^1 branch of solutions through $(y^-, 0)$, which coincides with $\lambda \mapsto y^-(\lambda)$ near 0 by (i)'s continuity and the branch's local uniqueness. Differentiating along it: $\dot{y} = H_S'(y^-)\big[(b - y^-) + \dot{y}\big]$, so $\dot{y} = -H_S'(y^-)(y^- - b)/(1 - H_S'(y^-))$, and $d\lambda/d\omega = \xi\varepsilon/(1-\varepsilon)$.

Act-Space Representation and Risk–Reach Details

This section states the scope conditions used in Section II. Attribution is always relative to an audience. An act that reveals an attribute without identifying a participant does not enlarge the public coalition; it instead changes beliefs or future protection requirements, as in Proposition S4. An identity event carried by no completed act is a breach and enters the harmful-leak rate of Proposition 3. Environment E1 holds the protection families fixed and lets payoffs depend on the public-coalition path, released sets, and an agent's own private act costs. Legal protection, insurance, or belief campaigns that change the families are environment interventions; the results apply at the families they induce.

Lemma 3 represents only acts with deterministic attribution. Splitting one stochastically attributed act into simultaneous exposure and commitment acts would conflict with the primitive that every subset of attempted actions may complete. The split would add a both-complete realization absent from the original lottery. Stochastic attribution is therefore a risk extension, not an exact same-act equivalence. At zero tolerance every positive-probability attributing outcome must itself be safe; at positive tolerance its probability enters the participant's unprotected-exposure risk. Neither observation licenses the stronger two-type representation, which the paper does not claim.

Lemma 4 allows the device to be a person. A colleague who publishes a co-signed letter is a holding stage if she first receives the others' completed, unattributed authorizations and conditions her release act on them. The reach of that informal tunnel is limited by trust in the delegate rather than by a platform's scale. A merely correlated set of separate attempts is different: unless the

institution eliminates every proper-subset realization, unilateral completion remains in the set of admissible outcomes. Correlation can lower expected risk at positive tolerance but cannot create exact joint completion at zero tolerance.

The trigger comparison in Theorem 5(i) is local. At public state C, suppose an acceptable mechanism rests i's unsafe exposure on same-date attempts by B_i. Its completed-authorization replacement solicits the same agents until all authorizations are held and then releases $T = C \cup \{i\} \cup B_i$ as one act. The local comparison assumes this full-completion target protects every trigger participant not already safe in C; whole-mechanism acceptability does not imply that condition when an agent can be exposed on several branches. The replacement exposes no one before the protecting release, while the original leaves i unprotected when her act completes and every act in B_i fails, an event of probability $(1-\varepsilon)\varepsilon^{|B_i|}$ conditional on reaching the trigger, or $\varepsilon^{|B_i|}$ conditional additionally on i's completion. On shared draws the replacement's released set contains every agent the original trigger exposes. Without retries, however, strict conditionality discards partial batches and the reach comparison fails.

Three limits prevent this local replacement from proving the global result by iteration. First, in a one-shot environment a partially completed unverified batch may protect and expose more agents than a conditional release that fails. Second, at $r_i \geq 1$ an agent accepts certain unprotected exposure and acts as an endogenous seed, allowing acceptable reach outside $C^+(S)$. Third, replacing one trigger can increase entry into downstream states and thereby amplify third-party risk, or remove unsafe exposures that served as stepping stones. Theorem 5(ii) therefore uses the whole-mechanism confinement bound rather than iterating part (i). Act costs and delay are also outside its two-coordinate comparison and enter through the entry problem.

Relation to Dynamic Coordination

The cascade shares Gale (2001)'s primitives: actions are nondecreasing, irreversible, and publicly observed, with positive spillovers. The comparison nevertheless requires a payoff completion. The protection families determine only the sign of payoffs at binary rosters, whereas Gale's characterization is for continuous payoffs on interval action spaces. The next result supplies that bridge rather than treating the binary environment as an automatic specialization.

PROPOSITION S5 (A Gale completion versus regret-free safety): *Maintain Assumption 1, let $P = C^+(S)$ be nonempty, and set $n = |P|$. There exists an auxiliary continuous positive-spillover game on $[0,1]^P$ such that:*

(i) at every binary roster $C \subseteq P$, an exposed agent's terminal payoff is positive if C protects her and negative otherwise, while abstention pays zero;

(ii) full exposure $\mathbf{1}_P$, corresponding to P, is the unique SPE limit of Gale's monotone game;

(iii) under Assumption 2, the maximal roster reachable by regret-free public exposure remains $C^-(S)$. Hence, if $C^-(S) \subsetneq C^+(S)$, a continuous terminal-payoff equilibrium can reach the tunnel destination, while a public path safe at every exposure cannot.

Proof of Proposition S5:

Fix $\eta \in (0, 1/(n+1))$. For $i \in C \subseteq P$, define the binary terminal payoff

$$v_i(C) = \begin{cases} 1 + \eta|C|, & C \in \mathcal{P}_i, \\ -1 + \eta|C|, & C \notin \mathcal{P}_i. \end{cases}$$

For $x \in [0,1]^P$, let $R_{-i}(x)$ include each $j \neq i$ independently with probability x_j, and set

$$u_i(x) = x_i \, \mathbb{E}\big[v_i\big(\{i\} \cup R_{-i}(x)\big)\big] .$$

This multilinear extension is continuous. At a binary corner $\mathbf{1}_C$, it gives zero to $i \notin C$ and $v_i(C)$ to $i \in C$; because $\eta n < 1$, the latter is positive exactly when $C \in \mathcal{P}_i$. Adding another agent raises v_i by η if protection is unchanged and by $2+\eta$ if protection switches from unsafe to safe. Assumption 1 rules out the reverse switch, so u_i is nondecreasing in every other player's action: the completion has positive spillovers.

Since P is self-protecting, $u_i(\mathbf{1}_P) = 1 + \eta n$ for every i. If $x \neq \mathbf{1}_P$, then every player's payoff is strictly below $1+\eta n$: this follows immediately for $x_i < 1$; when $x_i = 1$, some $j \neq i$ has $x_j < 1$, and the strictly positive η increment from j makes the expectation strictly smaller. Thus $\mathbf{1}_P$ strictly Pareto-dominates every other action profile, and in particular every other *satiation point*, a profile from which no player can gain by increasing her own action. Gale (2001, Theorem 3) therefore makes $\mathbf{1}_P$ the unique SPE limit on these bounded interval action spaces. For part (iii), every safe-expansion iterate is self-protecting and hence is contained in P by Theorem 1; restricting the comparator's player set to P therefore preserves the original cascade endpoint. Theorem 2(i)–(ii) gives the claim.

The completion is an existence benchmark, not a behavioral identification claim. A fractional x_i is an irreversible participation-probability or intensity commitment, and the small $\eta|C|$ term deliberately creates strict Pareto dominance. Selection of C^+ therefore does not follow from the paper's ordinal protection primitives alone. The construction shows the narrower point the comparison needs: when only the terminal roster enters payoffs, equilibrium continuation can support the greatest safe destination even though some intermediate rosters are unsafe. The paper's regret-free requirement requires protection at every exposure and therefore stops the public path at C^-. Conditional release restores the terminal destination without asking agents to trust future action by others: it substitutes completed private authorization and joint release for equilibrium continuation (Theorem 4).

Heterogeneity and the Difference in Reach

Outside the homogeneous case, the protection-requirement distribution determines the difference between the cascade and tunnel. The results below give the heterogeneous fixed-point geometry and comparative statics used in the article. Let $H_S(y) = m\{i \in S : \rho_i \leq y\}$ and $G_S(y) = b + H_S(y)$. Let $y^- \leq y^+$ be the least and greatest fixed points of G_S, the public masses of C^- and C^+ in the scalar specialization of Theorems 1 and 2. The companion paper separately studies how signing frictions affect entry in a homogeneous population (Cashman, 2026).

PROPOSITION S6 (Exposure barriers and the difference in reach): *In the scalar nonatomic benchmark, with continuous H_S:*

(i) the contribution is the fixed-point distance, $m(C^+) - m(C^-) = y^+ - y^-$;

*(ii) (*Exposure barriers.*) The inequality $y^- < y^+$ is equivalent to G_S having multiple fixed points. An* exposure barrier *—a level $y > y^-$ at which the protected mass falls short of the mass needed to sustain it, $G_S(y) < y$, with a recrossing above it ($G_S(y') \geq y'$ for some $y' > y$)—is sufficient for $y^- < y^+$ and is also necessary when the lower endpoint is* locally blocked *($G_S(y) < y$ for all y in a right-neighbourhood of y^-). Without local blockage, multiple fixed points can instead arise through a critical-mass or tangency case, in which the cascade endpoint is not robust to an upward perturbation but no exposure barrier separates it from y^+;*

(iii) raising H_S only on (y^-, ∞) leaves y^- (hence C^-) unchanged and weakly raises y^+ (hence C^+); raising H_S on $[b, y^-]$, or raising the seed b, weakly raises y^-.

Proof of Proposition S6:

(i) $C^+ = \{i : \rho_i \le y^+\}$ and $C^- = \{i : \rho_i \le y^-\}$, so $m(C^+) = H_S(y^+) = y^+ - b$ and $m(C^-) = H_S(y^-) = y^- - b$ at the fixed points; subtract. (ii) The condition $y^- < y^+$ means that G_S has two distinct fixed points, directly from the definitions of $y^- \le y^+$. *Sufficiency*: let $y_0 > y^-$ satisfy $G_S(y_0) < y_0$ with $G_S(y') \ge y'$ for some $y' > y_0$. Since G_S is continuous in the nonatomic benchmark, the intermediate value theorem gives a fixed point in $(y_0, y']$, so $y^+ > y_0 > y^-$. *Necessity under local blockage*: if $y^- < y^+$ and $G_S(y) < y$ on $(y^-, y^- + \varepsilon)$, any $y_0 \in (y^-, y^- + \min\{\varepsilon, y^+ - y^-\})$ has $G_S(y_0) < y_0$ and recrosses at $y' = y^+$, an exposure barrier. Local blockage cannot be dropped: for $b = 1$, $G_S(y) = 1 + \sqrt{y-1}$ on $[1, 2]$, the fixed points are 1 and 2, yet $G_S(y) > y$ throughout $(1, 2)$. Thus $y^- < y^+$ without an exposure barrier. Here $y^- = b$, the cascade never leaves the seed, and any upward perturbation at the unstable lower fixed point can start it. (iii) The iteration to y^- evaluates G_S only at masses $\le y^-$, so y^- depends only on H_S restricted to $[b, y^-]$. A change confined to (y^-, ∞) leaves y^- fixed while weakly raising G_S elsewhere, so the greatest fixed point y^+ weakly rises. A change raising H_S on $[b, y^-]$, or a larger b, raises G_S on the iteration's range and weakly raises y^-.

This is the standard threshold-cascade geometry (Granovetter, 1978). The value of failure privacy is not total latent support but the mass beyond the exposure barrier that safe public expansion cannot reach. Part (iii) gives the comparative static: support above y^- is invisible to decentralized expression, so a movement can have broad private backing and a small cascade when support is concentrated at requirements the early mass cannot bridge. The contract's contribution grows as support concentrates there. In the union application, these are the most-exposed workers, reachable only by the tunnel. Theorem 6 shows that the decentralized game selects y^- rather than merely permitting a stop there. With ambient trust, agents may rely on promised follow-through and the public path may move beyond y^-. The model does not characterize the resulting endpoint.

Maximum protection shortfall.

Write

$$\varsigma \;=\; \sup_{y \in [y^-,\, y^+]} \big(y - G_S(y)\big)^+,$$

the largest amount by which protected mass falls short of public mass between y^- and y^+. A movement that crosses by incremental public exposure must reach an intermediate roster at which at least ς units of public mass are borne by participants whom that roster does not protect. Reaching y^+ from below makes public mass pass through every intermediate level; where the shortfall peaks, at most $G_S(y)$ of that mass is protected. At least ς is therefore exposed along the path, which regret-free safety forbids and which explains why safe incremental expression halts at C^-. Failure-private disclosure reaches y^+ with no such exposure. The shortfall is positive exactly when G_S dips below the diagonal somewhere in (y^-, y^+). A synchronized reveal avoids this incremental-path lower bound, but ς alone gives no upper bound on unsafe exposure in a particular realization: that depends on which agents complete. Proposition 2 instead bounds the probability of an unsafe realization from the completion law.

Real coalitions are atomic (each agent of positive, non-negligible size), and there the cascade and the destination are governed by different thresholds. With positive masses a_i, an agent can safely go public at existing public mass y iff $y + a_i \ge \rho_i$, that is iff y exceeds her *activation threshold* $\check{\rho}_i := \rho_i - a_i$; the cascade is therefore governed by the activation curve $K_S(y) = m\{i : \check{\rho}_i \le y\}$, while the tunnel destination C^+ is still governed by final requirements $H_S(y) = m\{i : \rho_i \le y\}$. The two coincide in the nonatomic limit $a_i \to 0$. The distinction has bite: a high-mass supporter can bootstrap the cascade despite a high requirement, because her own mass helps clear it.

As an illustration, take seed $b = 1$ and consenters $\{x, y, w_1, w_2\}$ with $a_x = a_y = 1$ and $a_{w_1} = a_{w_2} = 1.5$ (so $m(S) = 5$), requirements $\rho_x = 2$, $\rho_y = 3$, $\rho_{w_1} = \rho_{w_2} = 6$; activation and requirement order coincide here ($\check{\rho}_x = 1 < \check{\rho}_y = 2 < \check{\rho}_w = 4.5$). The cascade adds x (public mass $b + a_x = 2 \ge \rho_x$), then

y (public mass $3 \geq \rho_y$), then stalls: adding a w would give public mass $4.5 < 6$. So $C^- = \{x, y\}$, at public mass $y^- = 3$. But $C^+ = S$, at public mass $y^+ = b + m(S) = 6 \geq \rho_w$: the pair (w_1, w_2) is safe only jointly, which unverified simultaneity cannot deliver and the tunnel can. The contribution is $m(C^+) - m(C^-) = 5 - 2 = 3 = y^+ - y^-$.

1. *Information*

Audience beliefs μ enter through the protection families $\mathcal{P}_i(\mu)$ and hence through Γ_μ, $C^-(\mu)$, and $C^+(\mu)$. Information changes these protection families; conditional disclosure changes which fixed point can be reached at a fixed μ.

REMARK S1 (Information and the path constraint): *At fixed μ, a belief-only intervention that releases no named coalition leaves the safe-coalition fixed points unchanged: a public-only regret-free process remains bounded by $C^-(\mu)$, while the social assurance contract implements $C^+(\mu)$. The observation is immediate from the fixed-belief theorems and is recorded for the design discussion below.*

Corollary 1 distinguishes two cases. If an attainable belief makes a lone agent safe, information can start the public path. If no attainable belief does so, the same Corollary 1 shows that co-release remains necessary. Remark S1 holds audience beliefs fixed when comparing the paths. The paper does not solve the joint design of information and the release path, nor does it claim that conditional disclosure generally dominates information design.

2. *Exogenous Conditioning Statistics: a Boundary Case*

REMARK S2 (The exogenous-statistic limit): *Threshold majority voting with disclosure conditions release on a statistic rather than on a protecting roster (Braghieri, Bursztyn and Fasnacht, 2026). The realized vote share in that model is endogenous to participation and voting, so the proposition below does not formally nest it. The proposition instead isolates a stricter boundary: when the conditioning statistic is primitive and independent of who acts, the holding stage can coordinate release but cannot enlarge reach. Co-exposure-dependent safety restores roster composition, seed dependence, and a possible difference in feasible reach between cascade and tunnel.*

The following result is a boundary marker on this paper's primitives. It characterizes what disappears when safety is independent of the coalition and should not be read as a reduced form of an endogenous voting equilibrium.

PROPOSITION S7 (The exogenous-statistic limit): *Suppose each protection family conditions only on an exogenous statistic: there is a random variable σ, realized independently of the mechanism and of who is named, and thresholds σ_i^* with $\mathcal{P}_i = \{C \ni i : \sigma \geq \sigma_i^*\}$. Then for every consent set and seed: (i) $\Gamma(C) = C^*(\sigma) := \{i \in S : \sigma \geq \sigma_i^*\}$ for every C, the operator has the unique fixed point $C^*(\sigma)$, and $C^-(S) = C^+(S) = C^*(\sigma)$; (ii) outcomes are seed- and history-independent, pinned by σ alone; (iii) failure privacy adds no reachability—a disclosure mechanism can at most coordinate arrival at $C^*(\sigma)$. Conversely, $C^- \subsetneq C^+$ requires protection to depend on co-exposure.*

PROOF:

(i) For $C \ni i$, membership $C \in \mathcal{P}_i$ holds iff $\sigma \geq \sigma_i^*$, independently of C's composition; hence $\Gamma(C) = C^*(\sigma)$ for every C, the unique fixed point is $C^*(\sigma)$, and least and greatest coincide. Each $i \in C^*(\sigma)$ is safe alone, so $T(\varnothing) = C^*(\sigma)$: the cascade arrives in one round. (ii)–(iii) follow. For the converse, $C^- \subsetneq C^+$ requires some i with $C^+ \in \mathcal{P}_i$ and $C^- \cup \{i\} \notin \mathcal{P}_i$, impossible under composition-free membership.

The comparison with Braghieri, Bursztyn and Fasnacht (2026) is therefore institutional rather than a nesting claim. Both papers study threshold-contingent disclosure. Their vote share is chosen in equilibrium; this proposition takes σ as exogenous. Establishing a formal mapping would require carrying those endogenous voting choices into the coalition operator. The result here proves only that a fully exogenous, composition-free statistic cannot create a difference between the safe public path and the safe destination.

Additional Applications and Scope

Union organizing.

The retaliating audience is the employer; the protective audience is the labor board; the verifier is a confidentiality-bound certifier. The model isolates the *exposure window* during which support is visible and retaliable. Open organizing is a cascade. Authorization cards signed privately and held until certification are a tunnel, often releasing a count to the board rather than a roster to the employer. This supplies protection through certified mass without guaranteeing anonymity; in a small unit the count can still reveal who signed (Proposition S1). The comparison between elections and card-based recognition is therefore a comparison between exposure windows, not simply between voting rules. Employer opposition is common in supervised campaigns (Bronfenbrenner, 2009; McNicholas et al., 2019), while estimates of illegal organizer discharge remain contested (Schmitt and Zipperer, 2009; LaLonde and Meltzer, 1991; Sherk, 2007). Neutrality and card-check agreements reduce management campaigning (Eaton and Kriesky, 2001). The application is imperfect because certification can complete an institutional outcome with a count, whereas strikes and walkouts require continued action after identities are known.

Dissent under repression.

The retaliating audience is the authority and the holding stage may be a diaspora administrator, a privately circulated declaration, or an encrypted escrow. Open dissent is the cascade of Kuran (1997); the institutional margin here is completed authorization, not merely information that others agree. The application requires two scope conditions. First, larger coalitions must be no less resilient, as in Section IV's fight-or-fold benchmark. Second, signing must itself be appreciably safer than public expression. Under pervasive surveillance, the high-requirement agents the tunnel exists to reach may also have the highest signing exposure, shrinking the completed-authorization set V and its ceiling $C^+(V)$ toward or below C^-. Belief-based turnout effects can also be strategic substitutes (Cantoni et al., 2019); those effects operate through beliefs about participation, not through protection supplied by co-exposure.

Dual use: collusion.

A private holding stage can assemble harmful as well as beneficial coalitions. Firms contemplating a cartel may be unwilling to reveal intent before enough partners are committed. In that domain a large difference in feasible reach between cascade and tunnel is a social cost, and policy should attack the holding stage through subpoena, mandated disclosure, leniency, or infiltration. The mapping is not literal: cartel value comes from secrecy from regulators rather than public expression to a retaliating audience. Appendix S8 gives a conditional welfare accounting.

The Value of Reach: A Welfare Decomposition

Feasible reach depends on entry, valid inputs, and a credible administrator. Greater reach need not improve welfare; its effect depends on which participants the mechanism enables to act. Failure

privacy is dual-use: the same escrow that protects whistleblowers can protect price-fixers. Write the benchmark welfare measure as

$$\Omega(C) = P(C) + X(C),$$

where $P(C) = \sum_{i \in C} u_i(C)$ is participant surplus. It need not increase under arbitrary inclusion: a newly added member could contribute a negative payoff. It does increase along inclusions of self-protecting coalitions when each u_i is nondecreasing in the coalition, because old members weakly gain and each new member has nonnegative payoff. Thus $P(C^+) \geq P(C^-)$. This payoff monotonicity is stronger than Assumption 1; upward closure of the protection family does not supply it. The term $X(C)$ is the externality of the released coalition on outsiders. The object of interest is

$$\Delta\Omega = \Omega(C^+) - \Omega(C^-) = \underbrace{\Delta P}_{\geq 0} + \underbrace{X(C^+) - X(C^-)}_{\text{ambiguous}}.$$

The sign is deliberately unresolved. In an accountability application, truthful disclosures may create a positive externality; authentication and matching do not establish truth, so false or strategic reports can instead create harm. In a collusive application, added reach can impose losses on consumers or other outsiders. Nor does the difference between cascade and tunnel determine the sign: the externality depends on which people the tunnel reaches, not only on how many.

The displayed measure gives participant surplus unit weight and omits seed payoffs, signing and solicitation costs, delay, administration, and security costs. These omitted terms have no determinate net effect without a model of timing and payoffs. Giving participant surplus a domain-specific welfare weight $\eta \in [0, 1]$ yields

$$\Delta\Omega_\eta = \eta\, \Delta P + X(C^+) - X(C^-).$$

This matters in extractive domains: cartel profit may be a transfer rather than value created, so a consumer-welfare calculation may put little or no weight on it. A large difference in reach is therefore not itself a welfare metric. The geometry determines feasible reach; an application must supply the externalities and omitted costs needed to value that reach.

1. *Scope of the Monotone Benchmark*

The preceding results have a defined domain. The body's lattice needs monotone protection. Material-retaliation settings satisfy it only when additional public co-participants do not make incumbents less safe; material harm alone is not enough. The rest of this section treats environments in which monotone protection fails outright.

The nonmonotone frontier.

Two cases violate monotone protection. First, social-image models make exposure costs depend on audience beliefs and perceived norms (Bénabou and Tirole, 2006; Bursztyn, González and Yanagizawa-Drott, 2020). A co-signer whom the audience views negatively may then raise other members' exposure costs. Second, a member's record or vulnerability may lower the coalition's resilience $q(C)$ (Section IV), making the coalition easier to defeat. In either case, self-protecting coalitions need not be closed under union, and there may be no largest self-protecting roster. Accessibility then replaces the lattice.

Let C_0 be the irreversible seed and let $\Sigma \subseteq \{C : C_0 \subseteq C \subseteq C_0 \cup S\}$ be an arbitrary family of admissible public sets, with $C_0 \in \Sigma$. Every nonseed member of an admissible set is protected there; the committed seed remains exempt, as in Section I. A public step adds one agent and must leave

the resulting set in Σ. The *accessible kernel* $\mathcal{K}(C_0)$ is the set of coalitions that can be reached from C_0 through such safe one-agent steps.

Under monotone protection, Lemma 2 makes safety after each individual completion enough to ensure safety under every realization. That equivalence fails here, so the next assumption strengthens Assumption 2 for unverified public batches. It is an additional premise, not a consequence of nonmonotonicity.

ASSUMPTION S1 (Batch admissibility): *For an unverified public transition that attempts a set E, the admissible realization set contains, for every subset $E' \subseteq E$ and every ordering of E', the realization in which exactly the members of E' complete, one at a time, in that order. A device act releasing previously completed, held authorizations is instead governed by Assumption 4.*

Treat Σ as known. To implement a proposed coalition C, the administrator or protocol must know Σ, the seed C_0, and C. The proposition neither elicits protection sets nor chooses among incomparable admissible coalitions. Outside the monotone domain, there may be no greatest destination to select.

PROPOSITION S8 (Accessibility necessity): *Maintain the consent-respecting act space and causal timing preceding Lemma 3, mechanisms as in Definition 2, and Assumptions 4 and S1. For any admissible family Σ with irreversible seed C_0 as above: (i) a coalition is the outcome of some regret-free public-only process from C_0 if and only if it lies in the accessible kernel $\mathcal{K}(C_0)$; (ii) every $C \in \Sigma \setminus \mathcal{K}(C_0)$ requires a private conditional-release stage to implement, and holding authorizations from $C \setminus C_0$ before releasing those names in one atomic device act implements it. Failure privacy is therefore necessary and sufficient for exactly the admissible coalitions outside the kernel. When Σ is generated by monotone protection (Assumption 1), $\max \mathcal{K}(C_0) = C_0 \cup C^-$ and $\max \Sigma = C_0 \cup C^+$, recovering Theorem 4 with the seed made explicit.*

PROOF:

(i) A regret-free public-only process exposes names irreversibly, and its transitions condition only on completed exposures. Fix any transition that newly attempts a set E of agents. Because no held authorization gates a public-only transition, it is an unverified public batch governed by Assumption S1. Regret-freeness must hold along every admissible realization, so each prefix of each completion order, added to the current public set, lies in Σ. Expanding each realized batch transition into one such safe order and concatenating across transitions yields a unit-increment chain from C_0 to the outcome. Hence the outcome lies in $\mathcal{K}(C_0)$. Conversely, a chain witnessing $C \in \mathcal{K}(C_0)$ realizes C by singleton attempts, and every prefix remains in Σ, so the process is regret-free. (ii) If $C \in \Sigma \setminus \mathcal{K}(C_0)$, no unit-increment safe public chain reaches C. By (i), a transition leaving the kernel cannot consist only of unverified public attempts. In the maintained consent-respecting act space, any multi-name event that eliminates the partial-completion realizations must be one device act resting on completed, still-unattributed authorizations: Lemma 4. Thus any regret-free implementation of C contains a private conditional-release stage. Conversely, collect completed authorizations from every member of $C \setminus C_0$ while keeping them unattributed and release those names through one device act. Its failure leaves C_0 public and its completion makes C public; both lie in Σ, and Assumption 4 rules out a partial roster. This implements C. Under Assumption 1, restoring the suppressed seed identifies the maximal accessible and admissible public sets as $C_0 \cup C^-$ and $C_0 \cup C^+$.

PROPOSITION S9 (Intractability of the optimal coalition): *An instance consists of a finite set S, a Boolean circuit Q_I with $|S|$ input bits such that $C \in \Sigma_I$ iff $Q_I(\mathbf{1}_C) = 1$, nonnegative integer weights x_i, and a target k, with all integers binary encoded. Deciding whether some $C \in \Sigma_I$ has $\sum_{i \in C} x_i \geq k$ (in particular, whether some $C \in \Sigma_I$ has $|C| \geq k$) is NP-complete, and computing a maximum-weight admissible coalition is NP-hard.*

The circuit is a compact membership representation; an explicit enumeration of the family, which can be exponential in $|S|$, is never formed. Both the circuit and the objective weights are supplied to the planner. The result does not provide a way to learn either input from prospective members.

PROOF:

A coalition's incidence vector is a polynomial certificate: circuit evaluation verifies $C \in \Sigma_I$, and binary addition verifies the weight bound, so the problem is in NP. Hardness is the standard reduction from INDEPENDENT SET. Given $G = (V, E)$, set $S = V$, use unit weights, and construct

$$Q_G(z) = \bigwedge_{\{u,v\} \in E} (\neg z_u \vee \neg z_v).$$

The circuit has size $O(|V| + |E|)$ and accepts exactly the independent sets of G—equivalently, the protection construction $\mathcal{P}_i = \{C \ni i : \text{no neighbor of } i \text{ lies in } C\}$. Thus an accepted coalition of weight at least k exists iff G has an independent set of size at least k. The weighted optimization statement embeds MAXIMUM-WEIGHT INDEPENDENT SET.

PROPOSITION S10 (A tractable size-and-average-type subclass): *Suppose agent i is safe in C iff $|C| \geq n_i$ and the average type $\bar{\tau}(C) = \frac{1}{|C|} \sum_{j \in C} \tau_j \geq s$ for a common requirement s and individual size requirements n_i. Then a maximum-cardinality self-protecting coalition is computable in $O(n^2)$ time. If, in addition, social weight is nondecreasing in type—the* aligned *case—the maximum-weight self-protecting coalition is computable in the same way. Without alignment, with rational s, τ_i and integer social weights and target W encoded in binary, deciding whether some self-protecting coalition of a fixed size has weight at least W is an exact-cardinality knapsack and is weakly NP-hard; no polynomial bound is claimed. Restricting the nonmonotonicity to coalition size and average type thus restores, in the aligned case, the tractability lost in Proposition S9.*

PROOF:

A coalition of size m is self-protecting iff all its members satisfy $n_i \leq m$—so they are drawn from $A_m = \{i : n_i \leq m\}$ with $|A_m| \geq m$—and $\bar{\tau} \geq s$. Among size-m subsets of A_m, average type is maximized by the m highest-type members; if even that subset has $\bar{\tau} < s$, no size-m self-protecting coalition exists, and otherwise it is a self-protecting coalition of size m. Sorting types once costs $O(n \log n)$; for each m, one pass over the sorted list collects the m highest-type members of A_m and their running sum, an $O(n)$ scan per size and $O(n^2)$ in total, and the scan over $m = 1, \ldots, n$ returns the largest feasible m. When social weight is monotone in type, the same m highest-type members also maximize total weight among size-m subsets of A_m while being the easiest to satisfy the average constraint, so the identical scan returns the maximum-weight coalition. For the unaligned boundary, reduce exact-cardinality knapsack with costs c_i, values x_i, budget B, and target cardinality m: set every $n_i = 1$, choose $L \geq \max_i c_i$, put $\tau_i = L - c_i$ and $s = L - B/m$, and use x_i as social weight. Then a size-m coalition clears $\bar{\tau} \geq s$ iff its total cost is at most B, and its weight is the knapsack value. The binary-encoded problem is therefore weakly NP-hard (and pseudo-polynomial by the usual dynamic program), so the stated polynomial bound is claimed only for the aligned case.

Outside the monotone domain, implementation is governed by accessibility rather than a lattice, and the planner's problem is generally intractable. The reduction is standard; the result marks the boundary of the tractable benchmark. Tractability returns in the size-and-average-type subclass of Proposition S10 only for maximum cardinality or for supplied objective weights aligned with type. These results still do not select a unique admissible coalition. Choosing among incomparable coalitions requires the planner's objective and knowledge of the protection constraints.