LOOKING FOR BIDDING TEAMMATES: A GAME-THEORETIC MODEL OF STRANGER COLLUSION IN PEER REVIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Paper bidding is the entry point to reviewer assignment at large machine learning conferences: reviewers declare interest in submissions, and an optimizer combines those bids with automated affinity scores. Reviewers who have never met now recruit each other on public social platforms, exchange submission identifiers, bid on each other's papers, and reciprocate with inflated scores. Existing collusion models take the group as given and assume its members already trust one another. We give the first game-theoretic model of collusion *formation* in peer review: a four-stage game over recruitment, identifier exchange under the risk of being reported, bidding neither party can verify, and reciprocal reviewing. The model predicts the arrangement cannot form, because the benefit at stake is delivered by the partner's separate decision, so reciprocation is never individually rational for any payoffs and the game unravels to the recruitment post. What closes the gap is enforcement rather than incentives: authors always see the reviews of their own paper, deadlines recur, and the group remembers who reciprocated. We give the sustainability condition and the detection rate above which no partnership survives, and show review-writing effort enters both, so cheap machine-assisted reviewing enlarges the sustainable set with no change in detection or sanctions. In a calibrated end-to-end conference simulation no detector we test exceeds $F_1 = 0.322$ against an attacker who camouflages bids, distributes them around a ring, and manipulates affinity; a two-person arrangement is worth 3.2 percentage points of acceptance probability; and the harm is distributional rather than aggregate, with 70 honest papers displaced while mean accepted quality moves by 0.002. Randomized assignment is the one defense reaching the enforcement itself, making a partner who never bid indistinguishable from one who bid and lost.

1 INTRODUCTION

Machine learning conferences have outgrown the reviewer pool that serves them. NeurIPS received roughly 2,400 submissions a decade ago and more than 21,000 in the most recent cycle, with ICLR, ICML and AAAI on similar trajectories (Yang et al., 2025). No program chair can assign reviewers by hand at that scale, so assignment is delegated to an optimizer maximizing total reviewer-paper affinity subject to load and coverage constraints (Charlin & Zemel, 2013; Aksoy et al., 2023). It takes two inputs: affinity scores computed from text (Charlin & Zemel, 2013; Cohan et al., 2020), and bids in which reviewers declare which submissions they want. This is the pipeline ICLR itself runs. Bids are unpriced and unverified: a reviewer pays nothing to declare interest in a paper they have none in, and no part of the pipeline tests a declared interest against a genuine one.

That gap is now being exploited openly. On public social platforms, including WhatsApp groups, Telegram channels, and X, reviewers post messages whose subject lines say, in effect, *looking for bidding teammates* for a named conference, with the terms stated in the body: I bid on your submission, you bid on mine, and whoever is assigned returns a high score. Fig. 1 shows three such threads with a selection of the replies each drew. The posts are public, easy to find, and recur every submission cycle. Collusion rings among acquainted researchers were described publicly several years ago (Littman, 2021); what is new is that recruitment has moved into the open, and that the people answering are strangers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Three public recruitment threads soliciting reciprocal bidding partners, with selected replies (↪). Posts are typeset rather than screenshotted, translated, and paraphrased; no identifying detail appears. The right column is our annotation, naming what each line instantiates in §2. Full protocol in the ethics statement.

1.1 Why open recruitment is a different problem

The peer review collusion literature is small and almost entirely one line of work (Jecmen et al., 2020; 2022; 2023; 2024; Hsieh et al., 2025). Across it the colluding group is taken as given: reviewers who already know each other, whose problem is to be assigned to each other's papers without being caught. The defenses it proposes, randomized assignment (Jecmen et al., 2020), cycle-free constraints (Boehmer et al., 2021), detectors over bidding data (Jecmen et al., 2024), each constrain what the optimizer may output; none reaches the relationship that produced its inputs, and none asks how the group comes into being. Appendix B surveys this work and states our positioning axis by axis.

Recruitment among strangers breaks the premise that framing relies on. Two people who met an hour ago in a group chat have no shared history and no expectation of future contact; they are also, in the formal sense conflict-of-interest policies use, not in conflict, so no automated check separates them. Four problems follow that colleagues do not face: *first-mover exposure*, since somebody must send a submission identifier first, handing over authorship, intent, and an accusation ready to forward to a program chair; *sting risk*, since the reply may come from a genuine partner or from someone collecting evidence; *unverifiable bidding*, since bids are private and non-assignment is consistent both with an honest bid that lost and with no bid at all; and *absent repeated-game trust*.

Demand for such arrangements is easy to explain and does not explain their stability. For a colluder already assigned to a partner's paper, the benefit at stake is produced by the partner's separate decision about a separate paper, so the inflated review is pure cost; reporting honestly dominates, both parties anticipate as much, and the anticipation propagates backwards to the recruitment post itself. §3 shows this holds for every value of the payoffs, so the operative constraint is enforcement rather than incentives.

1.2 Contributions

1. **The Mutual Bidding Dilemma** (§2), the first game-theoretic model of collusion *formation* in peer review and the first to treat colluders as strangers: a four-stage extensive-form game over recruitment signaling, information exchange under sting risk, bidding as a hidden action, and reciprocal reviewing.
2. **An unraveling result and its resolution** (§3). The reviewing stage is a Prisoner's Dilemma for every parameter setting, so the one-shot game has a unique subgame-perfect equilib-

rium without collusion (Theorem 3) and observed collusion is evidence about enforcement technology rather than payoff magnitudes. We give the discount threshold $\delta^\star$ above which reciprocal inflation is sustainable (Theorem 4), the trust threshold $q^\star$ above which the first mover reveals (Proposition 6), and the detection rate $d^\star$ above which no discount factor sustains collusion (Theorem 7). Effort cost enters $\delta^\star$ directly, turning the claim that language models amplify collusion into a comparative static.

3. **Attack taxonomy and evasion analysis** (§4, §5.2): four bidding and four reviewing strategies against four detector families, with camouflaged and distributed bidding together defeating every single-signal detector we test.
4. **Impact and defense implications** (§5, §6): a calibrated end-to-end simulation measuring the acceptance boost colluders obtain, the honest papers they displace, and the residual harm under existing defenses.

2 The Mutual Bidding Dilemma

2.1 Setting and notation

A conference receives N submissions $\mathcal{P}$ and employs M reviewers $\mathcal{R}$. Reviewer i authors $\mathcal{P}_i \subset \mathcal{P}$ and is assigned at most L_i papers; each paper receives k reviews. Reviewers record bids over *eager*, *willing*, *neutral*, *not willing* and *conflict*, giving a bid matrix $\mathbf{B}$; a bid is cheap talk in the technical sense, costing nothing to submit or misreport and untestable against the reviewer's actual interest. The affinity matrix $\mathbf{A} \in [0,1]^{M\times N}$ comes from the venue's text-matching system (Charlin & Zemel, 2013; Cohan et al., 2020), and the assignment $\mathbf{X} = f(\mathbf{A}, \mathbf{B})$ solves

$$\begin{aligned} \max_{\mathbf{X}\in\{0,1\}^{M\times N}} \quad & \sum_{i,j} w(A_{ij}, B_{ij})\, X_{ij} \\ \text{s.t.} \quad & \textstyle\sum_j X_{ij} \le L_i \;\; \forall i, \quad \sum_i X_{ij} = k \;\; \forall j, \quad X_{ij} = 0 \;\; \forall (i,j) \in \mathcal{C}, \end{aligned} \tag{1}$$

following Taylor (2008) and Stelmakh et al. (2018), with $\mathcal{C}$ the conflict-of-interest set. The weight $w(\cdot)$ is monotone in both arguments and in most deployments gives bids substantial influence, so an eager bid can outweigh a sizable affinity deficit; that influence, with the sparsity of positive bids on most submissions, is what makes bidding worth manipulating. Papers are accepted if their aggregate score clears a threshold set to the venue's target rate. One further stage matters: venues release initial reviews to co-reviewers and to the paper's authors, normally analyzed as a quality mechanism (Stelmakh et al., 2021) but which §3 shows also functions as a monitoring device.

Definition 1 (Collusive dyad)**.** A pair of reviewers (i, j) forms a *collusive dyad* if they satisfy no declared conflict of interest, have no co-authorship or institutional relationship, have not interacted before the current recruitment episode, and agree to bid on each other's submissions with the intention of returning inflated scores conditional on assignment.

The third clause separates our object of study from the collusion modeled in prior work, and makes the analysis non-trivial by removing the trust those models assume.

2.2 The four-stage game

The dyad plays a four-stage extensive-form game ordered by the conference calendar, each stage under a distinct information structure, which is the substance of the model. Table 5 in Appendix C collects the parameters.

Stage 1: recruitment signaling. The initiator chooses whether to post a signal reaching n potential partners at cost c_s. The post is itself evidence: the venue attributes it with probability μ, carrying sanction π_s. Nature draws the responder's type, a genuine colluder with probability q, otherwise a sting, which the initiator cannot observe.

Stage 2: information exchange. Targeted bidding requires knowing which submission to bid on, and simultaneous exchange is unavailable to strangers without a commitment device, so exchange is sequential. The first mover chooses from $\{\texttt{reveal}, \texttt{withdraw}\}$; the responder, having received a real identifier, chooses from $\{\texttt{reciprocate}, \texttt{fake}, \texttt{report}\}$. This is a trust game with the roles reversed: what the first mover hands over is not money but evidence, transmitting identity,

Table 1: Stage 4 payoffs under mutual assignment. Row player is i; I is an inflated score, H an honest one.

	I	H
I	$(b-\kappa-c_e,\ b-\kappa-c_e)$	$(-\kappa-c_e,\ b)$
H	$(b,\ -\kappa-c_e)$	$(0,\ 0)$

authorship and proof of intent at once. A responder who reports obtains R; a reported first mover loses L.

Stage 3: bidding. Each party independently records a bid on the partner's paper at opportunity cost c_b. The action is hidden, since bids are visible only to the venue, and not decisive, since bidding yields assignment only with probability ρ: a moral hazard problem with a confounded signal, since a partner who fails to be assigned may never have bid, or may have bid and lost, an event of probability $1-\rho$ even under full cooperation.

Stage 4: reviewing. Conditional on assignment, each party chooses I (a score above their honest assessment by a margin sufficient to change the decision) or H (honest). An inflated review that survives co-reviewer discussion and meta-review costs c_e beyond the effort the reviewer owes anyway, and carries expected sanction $\kappa = d\pi$, giving Table 1 under mutual assignment.

Lemma 2 (Stage 4 is a Prisoner's Dilemma)**.** *For all $b > 0$ and $\kappa + c_e > 0$, the Stage 4 game of Table 1 is a Prisoner's Dilemma: H strictly dominates I for both players, mutual H is the unique Nash equilibrium, and mutual I Pareto-dominates it whenever $b > \kappa + c_e$.*

Fixing j's action, the difference i's payoff makes between H and I is $-(\kappa+c_e) < 0$ in both branches; all proofs are in Appendix A. The independence from b is the substantive content: b is delivered by the *partner's* action and cancels from the comparison governing one's own, so no acceptance benefit, however large, makes reciprocation individually rational.

Collecting the stages, a party who signals, exchanges identifiers with a genuine partner, bids, and expects reciprocation with probability σ obtains $U = \rho\sigma b - \rho(\kappa + c_e) - c_b - (c_s + \mu\pi_s)$. Under mutual cooperation $\sigma = 1$, and writing

$$g = \rho b, \qquad w = \rho(\kappa + c_e) + c_b, \tag{2}$$

the per-round net is $u_C = g - w$. Two parameters carry most of the paper's message: ρ, which the attacks of §4 raise and the defenses of §6 lower, and c_e. Writing a convincing favorable review by hand costs hours, because it must withstand co-reviewers who read the paper honestly; a language model costs minutes; a draft written by the paper's own authors costs nothing.

3 Equilibrium Analysis

3.1 The one-shot game unravels

Consider first the benchmark matching the phenomenon most directly: two strangers, one conference, and no observation of each other afterwards.

Theorem 3 (Unraveling)**.** *In the one-shot Mutual Bidding Dilemma with $c_s > 0$, $c_b > 0$ and $\kappa + c_e > 0$, the unique subgame-perfect equilibrium is: no recruitment signal is posted, no submission identifier is revealed, no collusive bid is placed, and both parties report honestly. Collusion does not occur for any values of b, ρ, q, or π.*

This is the pivot of the paper: collusion of exactly the kind in Fig. 1 is observed at increasing volume, and the one-shot model says it should never occur. The resolution cannot lie in payoff magnitudes, since Lemma 2 holds for every magnitude, so it must lie in the benchmark's two assumptions, that the interaction happens once, and that nothing is observed afterwards. Stranger collusion therefore requires an *enforcement technology*: a mechanism making failure to reciprocate costly outside the current stage game.

Three features of the review ecosystem supply one. *Reviews of one's own paper are always observed*, so Stage 4 is perfectly monitored ex post, unlike Stage 3, an asymmetry that drives

Proposition 5. *The conference calendar supplies repetition*, since deadlines are two to three months apart. And *the platform supplies reputation*: recruitment groups persist across cycles, and someone who takes a partner's bid and returns an honest score can be named inside the group. The platform performs for strangers what institutional proximity performs for colleagues, at near-zero cost.

3.2 Sustainability of reciprocal inflation

Let $\delta_r \in (0, 1)$ be the product of time discounting and the probability the relationship survives to the next conference round. Under grim trigger, cooperate iff no observed betrayal has occurred, a round in which the partner was assigned to one's own paper and reported honestly.

Theorem 4 (Sustainability)**.** *Mutual cooperation is a subgame-perfect equilibrium of the repeated Mutual Bidding Dilemma under grim trigger if and only if*

$$\delta_r \;\geq\; \delta^\star \;:=\; \frac{w}{g} \;=\; \frac{\kappa + c_e}{b} \;+\; \frac{c_b}{\rho\, b} \;=\; \frac{d\pi + c_e}{b} + \frac{c_b}{\rho b}. \tag{3}$$

Differentiating, $\delta^\star$ is increasing in d, π, c_e and c_b and decreasing in b and ρ. Monotonicity in c_e matters because writing a convincing inflated review is the only expensive step in the sequence and the one that has recently become cheap: $c_e \to 0$ gives $\delta^\star \to d\pi/b + c_b/(\rho b)$, so a dyad that could not sustain collusion at human effort levels may sustain it with machine assistance and nothing else changed. Monotonicity in ρ is what makes randomized assignment work, and Proposition 5 shows it works twice over.

3.3 Hidden bidding and the tolerance problem

The grim trigger punishes only observed betrayal, and Stage 3 produces no observation: a party who quietly stops bidding is never assigned, appears merely unlucky, and is never punished. Enforcement there requires punishing on non-assignment, a confounded signal occurring with probability $1 - \rho$ every round even under full cooperation, the structure Green and Porter identified in cartels observing only a noisy public signal of each other's conduct (Fudenberg & Tirole, 1991). The resolution is the same: punishment must be triggered by a statistic that sometimes fires on the equilibrium path. Let the colluders adopt a k-strike rule, terminating the relationship after k consecutive rounds in which the partner is not assigned to one's paper.

Proposition 5 (Tolerance)**.** *Under the k-strike rule, withholding the bid is deterred if and only if $\delta_r^k \geq \delta^\star$, that is*

$$k \;\leq\; k_{\max} \;=\; \frac{\ln \delta^\star}{\ln \delta_r}, \qquad \textit{while} \qquad k \;\geq\; k_{\min} \;=\; \frac{\ln \varepsilon}{\ln(1-\rho)} \tag{4}$$

is required to hold the per-round probability of terminating an honestly cooperating relationship below ε. A workable rule exists only if $k_{\min} \leq k_{\max}$. As $\rho \to 0$, $k_{\min} \to \infty$ and $k_{\max}$ falls, so the feasible set empties.

This identifies an effect of randomized assignment that has not, to our knowledge, been noted. Randomization is normally justified as reducing the expected return on manipulation and criticized for the assignment quality it costs (Jecmen et al., 2020; 2022). It also degrades the colluders' own monitoring technology, since a lower ρ forces a more forgiving rule, and a more forgiving rule cannot deter free-riding, at no cost beyond the first effect.

3.4 Who reveals first, and when it is worth posting

Let $W = (g - w)/(1 - \delta_r)$ be the continuation value of a genuine partnership, available whenever $\delta_r \geq \delta^\star$.

Proposition 6 (Trust threshold)**.** *At Stage 2, the responder reciprocates rather than reports if and only if $W \geq R$. Anticipating this, the first mover reveals a real submission identifier if and only if*

$$q \;\geq\; q^\star \;=\; \frac{L}{W + L}, \tag{5}$$

which is decreasing in W, with $q^\star \to 0$ as $\delta_r \to 1$.

This accounts for the behavior hardest to reconcile with intuition, people posting evidence of intent under identifiable handles, then sending identifiers to strangers. They are not misjudging the probability of a sting but responding correctly to a continuation value large enough that $q^\star$ is low. The lever a venue controls is L, which saturates; the effective lever is W.

Two further results are proved in Appendix A: a public group of even a few dozen members clears the posting threshold $n^\star$ by orders of magnitude, so monitoring intensity μ restrains posting where audience size does not (Proposition 9); and the number of partners a reviewer maintains has an interior optimum with $\partial m^\star/\partial c_e < 0$ (Proposition 10).

3.5 The equilibrium-breaking threshold

Theorem 7 (Detection threshold)**.** *For a fixed discount factor δ_r, mutual cooperation is sustainable if and only if the per-review detection probability satisfies*

$$d \;\le\; d^\star(\delta_r) \;=\; \frac{1}{\pi}\left(\delta_r b - c_e - \frac{c_b}{\rho}\right). \tag{6}$$

In particular, no discount factor sustains collusion once $d > (b - c_e - c_b/\rho)/\pi$.

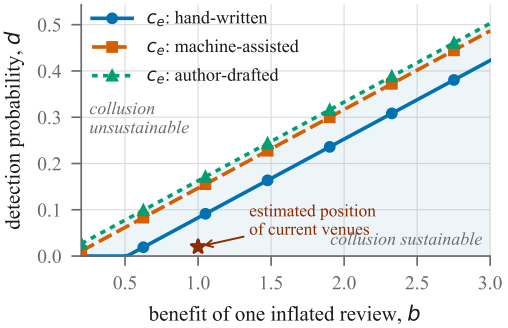


Figure 2: Stability regions in the (d, b) plane from (6). Below each boundary, reciprocal inflation is sustainable. Reducing c_e moves the boundary upward, enlarging the stable region with no change to detection.

The boundary is linear in b with slope δ_r/π, so the benefit of publication and the probability of detection trade off at a rate set entirely by the sanction and the durability of the relationship. The failure of existing detectors on camouflaged bidding (Jecmen et al., 2024), which §5 reproduces, places current venues far inside the stable region.

4 Attack Strategy Taxonomy

The equilibrium analysis treats ρ and d as parameters, but a colluder chooses them, within limits, by choosing how to bid and how to write. We catalog those choices because defenses cannot be evaluated against an unspecified adversary, and because the existing detector evaluation (Jecmen et al., 2024) considers a narrower strategy space than the recruitment posts describe. Table 2 gives the four bidding strategies we simulate. The four detector families we test them against, bid frequency, dense subgraph mining following Charikar (2000) and its camouflage-resistant refinements (Hooi et al., 2016), cycle detection, and text-similarity anomaly, are specified in Appendix C.

Affinity awareness is the only strategy raising effectiveness and lowering detectability at once: because the optimizer weighs affinity and bids together, raising A_{ij} raises ρ for a fixed bid and also makes the assignment defensible, since a reviewer who bid on a paper they *also* match strongly is not anomalous under any signal we know of. By Theorem 4 it lowers $\delta^\star$ as well.

The reviewing stage admits its own gradient, ordered by c_e rather than by the score produced: *uniform maximum*; *calibrated inflation*, high relative to the paper but ordinary relative to the reviewer's own history; *constructive positive*, surviving discussion at higher c_e; and *author-drafted*, in which the submitting authors write the review and the colluding reviewer submits it after superficial editing. Author-drafted attains the lowest c_e together with the highest resistance to content-based scrutiny,

Table 2: Bidding strategies; *defeats* names the detector family each is designed to evade, which §5.2 measures.

Strategy	Mechanism	Defeats
Naive	Bid *eager* on every partner's paper, behave normally otherwise.	—
Camouflaged	Place collusive bids inside a larger set of genuine bids on topically adjacent papers, sized so the collusive fraction matches the population.	Bid frequency
Distributed ring	Replace the dyad with a directed cycle of length $k \geq 3$ in which each member bids on one successor, so the bid graph contains no two-cycle.	Cycle detection
Affinity-aware	Curate the reviewer's publication record toward the partner's topic, or insert matched background text into the submission (Hsieh et al., 2025), raising A_{ij}.	Text similarity

and machine-text detectors do not apply because there is no machine text. It defeats the one policy response to §3's effort result: restricting machine assistance raises c_e only for reviewers who comply, while the author-drafted route reaches $c_e \approx 0$ without a machine at all.

5 Experimental Evaluation

5.1 Simulation framework and equilibrium validation

No conference releases bidding data, assignment logs, or affinity scores, so the quantities this paper needs cannot be measured on any existing dataset. We build an end-to-end simulator calibrated against what is public: submission and reviewer counts, review load, score distributions and acceptance rates from OpenReview (OpenReview, 2026), and bidding patterns from the simulated malicious bidding dataset of Jecmen et al. (2023). It generates submissions with a latent quality θ_p and a topic vector, computes affinity from topic overlap with noise, produces honest bids as a function of affinity, injects collusive bids per a chosen strategy from §4, solves (1), scores papers as noisy observations of θ_p with collusive scores inflated, and applies an acceptance threshold calibrated to the target rate. Because θ_p is generated rather than inferred, the simulator supplies the ground-truth quality no real dataset contains, which is what makes §5.3 possible at all. Venue parameters follow published ICLR and NeurIPS statistics at $N = 6000$, $M = 9000$, $k = 3$, $L_i = 5$; Appendix C gives the generative detail and Table 6 the full parameter set, marking which values are calibrated and which assumed.

Equilibrium validation. We first check the closed forms against a population game in which $20{,}000$ agents heterogeneous in b and δ_r best-respond to a fixed point, with more colluders competing for the same slots (lowering ρ) and larger subgraphs easier to detect (raising d). The measured transitions bracket their closed forms on the congestion-adjusted side: d at 0.079 against $d^\star = 0.113$, q at 0.627 against $q^\star = 0.571$. The effort sweep is the comparative static in c_e measured directly: 68.2% of the population can sustain collusion at hand-written effort ($c_e = 0.40\,b$), 91.9% at machine-assisted effort, and 95.2% at $c_e = 0$, with nothing else changed across those three points (Appendix E).

5.2 Detection evasion

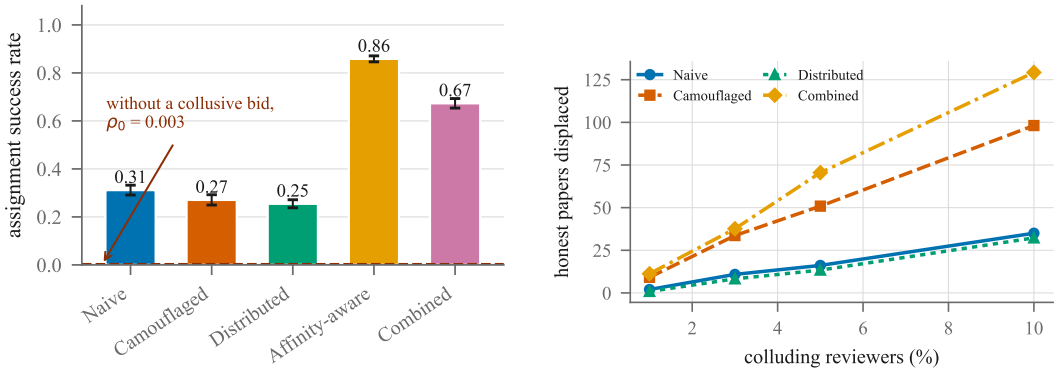
We inject colluders at 5% of the reviewer pool using each bidding strategy of §4 and run the four detector families over the resulting data, thresholding each at a common 5% false positive rate, since a chair's binding constraint is how many reviewers they can afford to investigate.

In Table 3 no detector family exceeds $F_1 = 0.322$ against any strategy. Cycle detection is the only signal identifying dyads at all ($F_1 = 0.156$ against naive bidding), and the distributed ring reduces it to 0.032 by adding one member. The bid-frequency test scores 0.000 against every dyadic strategy: once partners are screened for topical proximity, a collusive bid sits inside the reviewer's normal range and nothing anomalous remains to measure.

Fig. 3a gives the other half. A collusive bid raises the probability of landing on the target paper from $\rho_0 = 0.003$ to 0.312 for naive bidding and to 0.859 when affinity manipulation is added, a factor

Table 3: Detection F_1 at a common 5% false positive rate, mean over 8 instances. Lower is better for the attacker; best in each column bold. Precision, recall and other operating points in Appendix F.

Strategy	Bid freq.	Dense subgr.	Cycle det.	Text anom.
Naive	**0.000**	**0.069**	0.156	0.047
Camouflaged	**0.000**	0.164	0.322	**0.030**
Distributed	0.079	0.110	**0.032**	0.049
Affinity-aware	**0.000**	**0.069**	0.156	0.085
Combined	0.127	0.216	0.115	0.063



(a) Assignment success by strategy.

(b) Displacement against colluder fraction.

Figure 3: Attack effectiveness and downstream impact, each measured against the zero-collusion run on the same instance. The dashed line in (a) is ρ_0, assignment probability without a collusive bid.

of over two hundred; the combined strategy reaches 0.673 while holding every detector at or below 0.216. Evasion and effectiveness are usually in tension and here they are not, because affinity awareness raises ρ and lowers detectability for the same reason: it makes the assignment genuinely well matched as the system measures it. One result runs against the taxonomy's prediction, camouflage slightly *raises* the cycle-detection score, from 0.156 to 0.322, because a reviewer who bids more widely has more chances to form an incidental two-cycle with an honest author. That the countermeasures pull against each other is the strongest available argument for multi-signal detection.

5.3 Downstream Impact

We hold the conference instance fixed and vary the colluder fraction and ring size against a zero-collusion run on the same seed. Honest reviewers score honest papers identically across the two arms by construction, so every difference is attributable to the collusion.

A bare dyad, the arrangement the recruitment posts describe, raises the acceptance probability of a participant's paper from 25.0% to 28.2%, a gain of 3.2 percentage points, the quantity an individual reviewer weighs when deciding whether to answer a post, requiring no organization beyond one exchange of messages. The gain peaks at 4.8 points at ring size 4 and then falls, the interior optimum of Proposition 10 appearing in simulation: past a point, additional partners compete for the same three review slots and add detection exposure without adding inflated reviews. At a 5% colluder fraction the combined strategy pushes 70 honest papers out of the accepted set and camouflaged dyads 51; at 10% the combined figure is 129 (Fig. 3b).

The harm is distributional, not aggregate. What a venue would observe is essentially nothing (Fig. 6a): mean ground-truth quality of the accepted set falls by 0.002 on a standard-normal scale and the tenth percentile by 0.007, leaving the two distributions visually indistinguishable. We expected a thickened lower tail and did not find one, both sides of every swap sit near the acceptance threshold, so a displaced honest paper and its replacement have almost the same latent quality. This is not reassurance. No aggregate audit will detect the attack, and the entire cost falls on the identifiable authors of the 70 displaced papers. Collusion at these scales does not make the conference worse on average; it makes it arbitrary at the margin, where most submissions live.

6 DISCUSSION

6.1 WHAT EXISTING DEFENSES ACHIEVE

We apply cycle-free assignment at $k = 2$ (Boehmer et al., 2021), randomized assignment with a per-pair cap (Jecmen et al., 2020), and both together to the instances of §5; Appendix E reports every residual rate. Cycle-free assignment barely moves the combined strategy (0.673 to 0.665), and even against the dyad it was designed for it removes only about a fifth of collusive assignments, because forbidding a two-cycle requires *both* directions to be assigned and at a realized per-direction probability near 0.673 most dyads never reach that state. One-directional favors violate no cycle constraint, and Theorem 4 says they are exactly what a repeated relationship is built on. Randomized assignment instead reduces success across every strategy, the combined strategy falling to 0.311 at 96.8% retained affinity.

The more consequential quantity follows from Proposition 5. Under no defense $\rho = 0.673$ for the combined strategy, which by (4) lets colluders run a k-strike rule with $2.06 \leq k \leq 5.96$, so the partnership can police itself. Under randomization ρ falls to 0.311, pushing $k_{\min}$ to 6.19 against a $k_{\max}$ of 5.42: the feasible set is empty. No rule both deters a partner from quietly withholding their bid and survives unlucky runs of non-assignment, so the residual rates overstate what randomization leaves standing, they assume colluders cooperate, the assumption the mechanism undermines.

6.2 LEVERS AVAILABLE TO A PROGRAM CHAIR

Detection probability d is the lever the field has worked on, and Theorem 7 says it works once it clears $d^\star$ and does little below it: a detector twice as good as its predecessor, while both sit well below $d^\star = 0.113$, changes nothing about the equilibrium. Detection work should therefore be evaluated against the threshold rather than against relative improvement, and §5.2 places current venues far from it. *Assignment probability* ρ is the lever we would prioritize: it enters $\delta^\star$ directly and separately destroys the colluders' ability to police each other, the second effect free once the first is paid for. The literature prices randomized assignment against its quality cost while counting only the first (Jecmen et al., 2020; 2022), so the mechanism is better than its own analysis claims, and since it works through beliefs about ρ as much as through ρ, a venue adopting a cap should publish it. *Monitoring intensity* μ is the only lever acting before a bid exists (Proposition 9); the deterrent need not be effective, only credible. *Sanctions* saturate, since L appears in both numerator and denominator of $q^\star$, and *effort cost* c_e is least under control, since the author-drafted variant reaches $c_e \approx 0$ without a machine.

In the near term, then: adopt randomized assignment with an explicit per-pair cap, publish it, and state a monitoring policy for public recruitment channels. In the medium term, build detection fusing bidding, scoring, assignment and authorship signals, since each alone is evadable while the countermeasures are mutually constraining.

6.3 LIMITATIONS

The results rest on simulation, because no venue releases the bidding, assignment, or affinity data a direct measurement would need, the same evidentiary position as prior work here (Jecmen et al., 2023; 2024). We sweep the parameters no public source constrains rather than asserting them, and emphasize thresholds and the direction of each comparative static rather than absolute magnitudes. The model assumes agents best-respond, so the evasion results characterize the upper end of adversary capability rather than the typical case. Appendix G treats the fixed-pair repeated game, the single-venue assumption, and d as exogenous against an adversary who adapts.

7 CONCLUSION

Reviewers recruit strangers on public platforms to bid on each other's submissions and return inflated scores. Our model says the arrangement should not form at all, and that what sustains it is enforcement rather than incentives. The defense implication is therefore not a new detector: by Proposition 5, lowering the probability that a collusive bid produces an assignment attacks the colluders' ability to police each other, not merely their expected payoff.

ETHICS STATEMENT

This paper describes an attack on a system its authors depend on, so we set out the considerations in full.

Data collection. The recruitment threads in Fig. 1 are public. We collected no private messages, joined no groups under false pretenses, and conducted no interaction with the people posting. The replies shown are a selection: each thread carries more than we reproduce, and we kept those that bear on the model.

Reproduction. We typeset the threads rather than reproduce screenshots, because the platform is identifiable from its interface chrome alone. The text is translated and lightly paraphrased so that no thread can be recovered by searching a quoted string, which a verbatim transcription would still permit after the images were removed. Handles, avatars, timestamps, group identifiers, and venue names are redacted, we name no individuals or groups, and we report neither the platform nor the search terms that locate the threads. The threads serve as evidence that the phenomenon exists and as motivation for the model; no part of the analysis depends on their content.

Human subjects. No human subjects were recruited and no IRB review was sought or required. The study analyzes public posts without interacting with their authors, and every quantitative result comes from simulation.

Disclosure. The attack taxonomy of §4 describes strategies more effective than those the published detector evaluations assume. We include it because a defense cannot be evaluated against an unspecified adversary, and because the strategies are not novel to us: camouflage and ring topologies are discussed openly in the recruitment channels themselves, and affinity manipulation is already published (Hsieh et al., 2025). The disclosure balance is the standard one in security research, and we judge it to favor publication.

Sting operations. Lowering q, the probability that a responder is genuine, appears in the model as a lever and amounts to a venue running sting operations against its own reviewers. We note it in (5) and do not recommend it. It requires a venue to deceive its volunteers, it would poison the recruitment channels in ways that are hard to reason about, and Proposition 6 shows its effect is bounded by $q^\star$ already being low for durable partnerships.

REPRODUCIBILITY STATEMENT

Proofs of every formal claim in the main text are given in Appendix A. The simulator is specified in full in Appendix C: conference generation, honest bidding, collusion injection, the assignment program, the scoring and decision rule, and the four detector families, together with the parameter table (Table 6), which marks explicitly which values are calibrated to published venue statistics and which are assumed. Appendix D varies every assumed value and reports which conclusions move.

Two design choices make the experiments reproducible rather than merely repeatable. Review noise is a hash of the triple (seed, reviewer, paper) rather than a sequential draw from a shared generator, so an honest reviewer's score on an honest paper is identical across runs and a collusive run differs from its baseline only in the injected bids. Every number and every figure in the paper, including the bolding of per-column minima in Table 3, is generated from the committed experiment outputs by a single script, so no quantity in the text is transcribed by hand. The full suite runs in about twenty minutes on one core.

The simulator, the experiment configurations, and the scripts that generate every number, table and figure in this paper will be released publicly upon acceptance. Until then, Appendix C specifies the generative process, the assignment program and its sparsification, the four detector implementations and their operating points, and the determinism scheme, at the level of detail needed to reimplement the pipeline; Table 6 lists every parameter value used, separating those calibrated to published venue statistics from those assumed.

AI USE STATEMENT

We used generative AI tools for two purposes in this work: polishing the language of the manuscript, and assistance in implementing the simulation code. All AI-assisted output was reviewed by the

authors. The manuscript text was checked against the claims it reports, and the simulation code was verified against the closed forms of §3 and by the cross-experiment identities. We take responsibility for the final content of this work, including its text, its claims, and the artifacts it reports.

REFERENCES

Meltem Aksoy, Seda Yanik, and Mehmet Fatih Amasyali. Reviewer assignment problem: A systematic review of the literature. *Journal of Artificial Intelligence Research*, 76:761–827, 2023.

Federico Bianchi, Francisco Grimaldo, Giangiacomo Bravo, and Flaminio Squazzoni. The peer review game: an agent-based model of scientists facing resource constraints and institutional pressures. *Scientometrics*, 116(3):1401–1420, 2018.

Niclas Boehmer, Robert Bredereck, and André Nichterlein. Combating collusion rings is hard but possible. *arXiv preprint arXiv:2112.08444*, 2021.

Moses Charikar. Greedy approximation algorithms for finding dense components in a graph. In *International workshop on approximation algorithms for combinatorial optimization*, pp. 84–95. Springer, 2000.

Laurent Charlin and Richard Zemel. The toronto paper matching system: an automated paper-reviewer assignment system. In *ICML*, 2013.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 2270–2282, 2020.

Matteo Gioele Collu, Umberto Salviati, Roberto Confalonieri, Mauro Conti, and Giovanni Apruzzese. Publish to perish: Prompt injection attacks on LLM-assisted peer review. *arXiv preprint arXiv:2508.20863*, 2025.

Drew Fudenberg and Jean Tirole. *Game theory*. MIT press, 1991.

Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. FRAUDAR: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 895–904, 2016.

Jhih-Yi Janet Hsieh, Aditi Raghunathan, and Nihar B Shah. Vulnerability of text-matching in ML/AI conference reviewer assignments to collusions. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 5189–5208, 2025.

Ganesh Jawahar, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*, 2020.

Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12533–12545, 2020.

Steven Jecmen, Nihar B Shah, Fei Fang, and Vincent Conitzer. Tradeoffs in preventing manipulation in paper bidding for reviewer assignment. *arXiv preprint arXiv:2207.11315*, 2022.

Steven Jecmen, Minji Yoon, Vincent Conitzer, Nihar B Shah, and Fei Fang. A dataset on malicious paper bidding in peer review. In *Proceedings of the ACM Web Conference 2023*, pp. 3816–3826, 2023.

Steven Jecmen, Nihar B Shah, Fei Fang, and Leman Akoglu. On the detection of reviewer-author collusion rings from paper bidding. *arXiv preprint arXiv:2402.07860*, 2024.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *arXiv preprint arXiv:2403.07183*, 2024a.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024b.

Michael L Littman. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6):43–44, 2021.

Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373, 2005.

OpenReview. OpenReview API and public conference data. `https://openreview.net`, 2026.

Wenqi Shen, Yu Jeffrey Hu, and Jackie Rees Ulmer. Competing for attention: An empirical study of online reviewers' strategic behavior. *Mis Quarterly*, 39(3):683–696, 2015.

Ivan Stelmakh, Nihar B Shah, and Aarti Singh. Peerreview4all: Fair and accurate reviewer assignment in peer review. *arXiv preprint arXiv:1806.06237*, 2018.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–17, 2021.

Camillo J. Taylor. On the optimal assignment of conference papers to reviewers. Technical Report MS-CIS-08-30, Department of Computer and Information Science, University of Pennsylvania, 2008.

Yuanzhang Xiao, Florian Dörfler, and Mihaela Van Der Schaar. Incentive design in peer review: Rating and repeated endogenous matching. *IEEE Transactions on Network Science and Engineering*, 6(4):898–908, 2018.

Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper copilot: Tracking the evolution of peer review in ai conferences. *arXiv preprint arXiv:2510.13201*, 2025.

A PROOFS

Throughout, $g = \rho b$ and $w = \rho(\kappa + c_e) + c_b$ as in (2), with $\kappa = d\pi$.

A.1 PROOF OF LEMMA 2

Conditional on mutual assignment, the payoffs are those of Table 1. Fix j's action. If j plays I, then i's payoff from I is $b - \kappa - c_e$ and from H is b, a difference of $-(\kappa + c_e)$. If j plays H, then i's payoff from I is $-\kappa - c_e$ and from H is 0, again a difference of $-(\kappa + c_e)$. Since $\kappa + c_e > 0$, H strictly dominates I regardless of j's action and regardless of b. By symmetry the same holds for j, so (H, H) is the unique Nash equilibrium and it is in strictly dominant strategies. It yields $(0, 0)$, which is Pareto-dominated by $(b - \kappa - c_e,\ b - \kappa - c_e)$ whenever $b > \kappa + c_e$. □

Remark 8. The independence of the dominance argument from b is the substantive content. The benefit b enters a player's payoff only through the *partner's* action and therefore appears with the same coefficient in both branches of the comparison. Raising the value of a publication makes collusion more attractive to enter, by raising g, but does not make reciprocation individually rational once entered.

A.2 PROOF OF THEOREM 3

We solve the four stages by backward induction, establishing strictness at each step so that the equilibrium is unique.

Stage 4. By Lemma 2, H is strictly dominant at every node in the Stage 4 subgame, so both parties report honestly whenever the subgame is reached.

Stage 3. Consider party i choosing whether to place a collusive bid on P_j. By the previous step, j reports honestly if assigned, so i's expected benefit from the bid is $\rho \cdot 0 = 0$ while its cost is $c_b > 0$. Not bidding yields 0. Hence not bidding is strictly optimal. The same argument applies to j. Note this step does not depend on i's beliefs about whether j bid, so the conclusion holds at every node of i's Stage 3 information set.

Stage 2. By the previous step, no collusive bid is placed on any path, so the continuation value of a formed partnership is 0. The first mover's payoff from revealing is $q \cdot 0 - (1-q)L = -(1-q)L$, and from withdrawing is 0. For $q < 1$ and $L > 0$ withdrawing is strictly optimal; for $q = 1$ the two are equal and revealing is weakly optimal but yields nothing, so no collusive bid follows in either case. The responder's choice is off the equilibrium path; sequential rationality requires only that reporting be optimal there, which holds whenever $R > 0$.

Stage 1. The initiator's expected payoff from posting is at most $[1 - (1-\lambda)^n] \cdot 0 - c_s - \mu\pi_s < 0$, against 0 from not posting, so not posting is strictly optimal.

Each step is strict, so the profile is the unique subgame-perfect equilibrium. No step invoked the magnitudes of b, ρ, q, or π, so the conclusion holds for all of them. □

A.3 Proof of Theorem 4

Consider the infinitely repeated game with per-round discount factor $\delta_r \in (0, 1)$ and the grim-trigger profile: play the collusive action (bid, and report I if assigned) in round t if and only if no observed betrayal has occurred in rounds $1, \ldots, t-1$, where an observed betrayal is a round in which the partner was assigned to one's own paper and reported H; otherwise play the honest action forever.

Cooperation value. Under mutual cooperation the per-round expected payoff is

$$u_C = \underbrace{\rho b}_{g} - \underbrace{[\rho(\kappa + c_e) + c_b]}_{w},$$

since the partner is assigned to one's paper with probability ρ and inflates when assigned, while one's own bid costs c_b and produces an inflated review, with its sanction risk and effort, with probability ρ. Hence $V_C = (g - w)/(1 - \delta_r)$.

Best deviation. By the one-shot deviation principle it suffices to consider a single-round deviation. The available deviations are: withhold the bid; bid but report H if assigned; or both. Withholding the bid saves c_b and, since assignment probability drops to ρ_0, saves $\rho(\kappa + c_e)$ up to a term of order ρ_0, which we take to be zero. Reporting H when assigned saves $\kappa + c_e$ in that event. Doing both saves w in total, which is the maximum. In the deviation round the partner has already committed, so the deviator still collects g in expectation. Both forms of deviation are detected: reporting H when assigned is directly observed, and withholding the bid is caught by the assignment-based rule of Proposition 5 at $k = 1$. Continuation after the trigger is 0, since mutual honest play yields no collusive payoff and, by Theorem 3 applied to the continuation, no party re-enters. Hence $V_D = g$.

Condition. Cooperation is sustainable iff $V_C \geq V_D$:

$$\frac{g - w}{1 - \delta_r} \;\geq\; g \iff g - w \;\geq\; g(1 - \delta_r) \iff \delta_r g \;\geq\; w,$$

that is $\delta_r \geq w/g = \delta^\star$. Substituting (2),

$$\delta^\star = \frac{\rho(\kappa + c_e) + c_b}{\rho b} = \frac{\kappa + c_e}{b} + \frac{c_b}{\rho b} = \frac{d\pi + c_e}{b} + \frac{c_b}{\rho b}.$$

The profile is subgame perfect because the punishment phase is a Nash equilibrium of the stage game by Lemma 2. □

A.4 Comparative statics of $\delta^\star$

The signs quoted in §3 follow by differentiating $\delta^\star = (d\pi + c_e)/b + c_b/(\rho b)$: $\partial\delta^\star/\partial d = \pi/b > 0$, $\partial\delta^\star/\partial\pi = d/b > 0$, $\partial\delta^\star/\partial c_e = 1/b > 0$, $\partial\delta^\star/\partial c_b = 1/(\rho b) > 0$, $\partial\delta^\star/\partial\rho = -c_b/(\rho^2 b) < 0$, and $\partial\delta^\star/\partial b = -\left[(d\pi + c_e) + c_b/\rho\right]/b^2 < 0$. □

A.5 Proof of Proposition 5

Deterrence. Under the k-strike rule, a party who withholds the bid is assigned to the partner's paper with probability $\rho_0 \approx 0$, so the strike counter increments every round and the relationship terminates after exactly k rounds. During those k rounds the deviator collects g per round rather than $g - w$, since the partner continues to cooperate. Hence

$$V_{D3} = g \sum_{t=0}^{k-1} \delta_r^t = g \, \frac{1 - \delta_r^k}{1 - \delta_r}, \qquad V_C = \frac{g - w}{1 - \delta_r}.$$

Requiring $V_C \geq V_{D3}$ gives $g - w \geq g(1 - \delta_r^k)$, that is $\delta_r^k g \geq w$, or $\delta_r^k \geq \delta^\star$. Taking logarithms and noting $\ln \delta_r < 0$ reverses the inequality: $k \leq \ln \delta^\star / \ln \delta_r = k_{\max}$. Setting $k = 1$ recovers Theorem 4.

False termination. Under mutual cooperation the partner is not assigned in a given round with probability $1 - \rho$, independently across rounds. The rule fires on a run of k consecutive such rounds, which has probability $(1 - \rho)^k$. Requiring this to be at most ε gives $k \ln(1 - \rho) \leq \ln \varepsilon$, and since $\ln(1 - \rho) < 0$, $k \geq \ln \varepsilon / \ln(1 - \rho) = k_{\min}$.

Feasibility. A rule exists iff $k_{\min} \leq k_{\max}$, that is

$$\frac{\ln \varepsilon}{\ln(1 - \rho)} \;\leq\; \frac{\ln \delta^\star}{\ln \delta_r}.$$

As $\rho \to 0^+$ we have $\ln(1 - \rho) \to 0^-$, so $k_{\min} \to \infty$; simultaneously $\delta^\star \to \infty$ through the $c_b/(\rho b)$ term, so the sustainability condition of Theorem 4 fails outright. Both effects act in the same direction. □

A.6 Proof of Proposition 6

Responder. Having received a real identifier, the responder compares reciprocating, which yields the continuation value W of a genuine partnership, against reporting, which yields the one-time payoff R. Sharing a fake identifier yields 0, since without a real identifier from the responder the first mover cannot bid on the responder's paper and no partnership forms. Hence reciprocation is optimal iff $W \geq R$.

First mover. Anticipating reciprocation from a genuine responder, the first mover's expected payoff from revealing is $qW - (1 - q)L$ and from withdrawing is 0, where the signaling costs are sunk at this node. Revealing is optimal iff $qW \geq (1-q)L$, that is $q(W + L) \geq L$, or $q \geq L/(W + L) = q^\star$.

Monotonicity. $\partial q^\star / \partial W = -L/(W + L)^2 < 0$ and $\partial q^\star / \partial L = W/(W + L)^2 > 0$. Since $W = (g - w)/(1 - \delta_r) \to \infty$ as $\delta_r \to 1^-$ whenever $g > w$, we have $q^\star \to 0$. □

A.7 Platform size

This result and the next are stated here rather than in §3, which quotes their conclusions.

Proposition 9 (Platform size)**.** *An initiator whose post reaches n readers, each responding independently with probability λ, posts if and only if $n \geq n^\star = \ln(1 - (c_s + \mu\pi_s)/\Gamma)/\ln(1 - \lambda)$ with $\Gamma = qW - (1 - q)L > c_s + \mu\pi_s$. $n^\star$ is increasing in the venue's monitoring intensity μ and in the sanction π_s attached to an attributed post.*

A public recruitment group satisfies $n \geq n^\star$ comfortably at any plausible λ, which is why the posts are public rather than private. Appendix D quantifies $n^\star$ against λ.

The initiator meets at least one responder with probability $1 - (1 - \lambda)^n$, and conditional on meeting one obtains $qW - (1 - q)L$ by Proposition 6. Posting is optimal iff

$$[1 - (1 - \lambda)^n]\,(qW - (1 - q)L) \;\geq\; c_s + \mu\pi_s.$$

Write $\Gamma = qW - (1 - q)L$ and assume $\Gamma > c_s + \mu\pi_s$, which is necessary for any n to work. Rearranging, $(1 - \lambda)^n \leq 1 - (c_s + \mu\pi_s)/\Gamma$, and taking logarithms with $\ln(1 - \lambda) < 0$,

$$n \;\geq\; \frac{\ln(1 - (c_s + \mu\pi_s)/\Gamma)}{\ln(1 - \lambda)} \;=\; n^\star.$$

Both μ and π_s increase the numerator's argument toward 1 from below, hence increase $n^\star$. □

A.8 Optimal ring size

A reviewer may maintain m partners in one round. Detection risk grows with m, since a larger collusive subgraph is a stronger signal to any of the detectors of §4; we write $d(m) = d_0 + \beta m$ for the per-review detection probability.

Proposition 10 (Optimal ring size)**.** *The per-round payoff* $\Pi(m) = m\,[\rho(b - c_e) - c_b - \rho\pi d_0] - \rho\pi\beta m^2$ *is maximized at*

$$m^\star \;=\; \frac{\rho(b - c_e) - c_b - \rho\pi d_0}{2\rho\pi\beta}, \qquad \frac{\partial m^\star}{\partial c_e} = -\frac{1}{2\pi\beta} < 0. \tag{7}$$

The derivative in (7) is the formal version of the claim that language models enlarge collusion rings. Since β is in the denominator, a detector sharply more sensitive to larger rings limits ring size even when it cannot catch dyads.

With m partners and per-review detection probability $d(m) = d_0 + \beta m$, the per-round payoff is

$$\begin{aligned}\Pi(m) &= m\,[\rho b - \rho\,(\pi(d_0 + \beta m) + c_e) - c_b] \\ &= m\,\alpha - \rho\pi\beta m^2,\end{aligned}$$

where $\alpha = \rho(b - c_e) - c_b - \rho\pi d_0$. This is strictly concave in m with $\Pi'(m) = \alpha - 2\rho\pi\beta m$, so the unconstrained maximizer is $m^\star = \alpha/(2\rho\pi\beta)$, which is (7), and is positive iff $\alpha > 0$. Then $\partial m^\star/\partial c_e = -\rho/(2\rho\pi\beta) = -1/(2\pi\beta) < 0$, $\partial m^\star/\partial\beta < 0$, and $\partial m^\star/\partial b = 1/(2\pi\beta) > 0$. In practice m is an integer and bounded by the reviewer's load L_i, so the operative optimum is $\min\{L_i, \lfloor m^\star \rceil\}$. □

A.9 Proof of Theorem 7

By Theorem 4, cooperation is sustainable iff $\delta_r \geq (d\pi + c_e)/b + c_b/(\rho b)$. Multiplying by b and rearranging,

$$\begin{aligned} d\pi \;&\leq\; \delta_r b - c_e - \frac{c_b}{\rho} \\ \iff \quad d \;&\leq\; \frac{1}{\pi}\left(\delta_r b - c_e - \frac{c_b}{\rho}\right) = d^\star(\delta_r). \end{aligned}$$

Since $d^\star$ is increasing in δ_r and $\delta_r < 1$, the supremum over admissible discount factors is $d^\star(1) = (b - c_e - c_b/\rho)/\pi$, above which no discount factor sustains cooperation. □

B Extended Related Work

Collusion in peer review. Jecmen et al. (2020) introduced randomized reviewer assignment as a mitigation: capping the probability that any reviewer-paper pair is matched bounds the success probability of a collusive attempt, at a measured cost in assignment quality of 5–15%. A follow-up shows expertise matching and manipulation robustness cannot be maximized simultaneously (Jecmen et al., 2022). Cycle-free reviewing (Boehmer et al., 2021) instead forbids cycles of length up to k in the assignment graph; at $k = 2$ this rules out the direct reciprocity the recruitment posts describe, but the constraint is expensive for larger k and a ring of size $k + 1$ evades it by construction. On the detection side, Jecmen et al. (2023) released a simulated dataset of malicious bidding, and a subsequent study (Jecmen et al., 2024) found that most off-the-shelf fraud detection algorithms fail on it once colluders camouflage, placing collusive bids inside a larger set of genuine bids on topically similar papers. Our experiments reproduce that finding and extend it to strategy combinations the original evaluation did not consider. Separately, Hsieh et al. (2025) showed the text-matching component is manipulable independently of bidding: curating a reviewer's public publication list, or inserting background sentences matched to a target reviewer's corpus, moves the affinity score enough to make assignment likely. The two manipulations are complements, since raising A_{ij} raises the probability that a collusive bid succeeds, which enters our payoffs through ρ and by Theorem 4 makes collusion easier to sustain. All of this work takes the colluding set as given.

Table 4: Positioning relative to prior work. *Formation* means the paper models how a collusive group comes into being rather than taking it as given; *strangers* means it does not assume prior trust between colluders; *impact* means it quantifies downstream effects on acceptance decisions.

	Formation	Strangers	Game-theor.	Impact	Multi-stage
Randomized assign. (Jecmen et al., 2020)	✗	✗	✓	✗	✗
Manipulation tradeoffs (Jecmen et al., 2022)	✗	✗	✓	✗	✗
Malicious bid data (Jecmen et al., 2023)	✗	✗	✗	✗	✗
Ring detection (Jecmen et al., 2024)	✗	✗	✗	✗	✗
Text-match attack (Hsieh et al., 2025)	✗	✗	✗	✗	✗
Cycle-free review (Boehmer et al., 2021)	✗	✗	✗	✗	✗
Peer prediction (Miller et al., 2005)	✗	✗	✓	✗	✗
This paper	✓	✓	✓	✓	✓

Table 5: Game parameters used in §3, per-dyad and per-round. Signaling parameters (c_s, μ, π_s, λ) enter only Proposition 9; Table 6 gives the full set.

Symbol	Meaning
b	value to a colluder of one inflated review on their own paper
ρ	$\mathbb{P}$(assigned to target paper | collusive bid), against a baseline $\rho_0 \ll \rho$ without one
$d,\ \pi$	per-inflated-review detection probability, and sanction if detected; $\kappa = d\pi$
c_e	extra effort to write a convincing inflated review
c_b	opportunity cost of spending a bid slot collusively
q	$\mathbb{P}$(responder is a genuine colluder) rather than a sting
$L,\ R$	loss to a reported first mover; payoff to a responder who reports
δ_r	discount factor between consecutive submission deadlines

Mechanism design in peer review. Work on eliciting honest reports from reviewers with no interest in any specific outcome uses peer prediction and related scoring rules (Miller et al., 2005; Xiao et al., 2018), or studies strategic behavior where reviewers are themselves authors (Shen et al., 2015). Repeated-game models of reviewer behavior exist (Bianchi et al., 2018), but model the reviewer's relationship with the venue rather than with another reviewer. Our analysis rests on standard repeated-game machinery (Fudenberg & Tirole, 1991), including the observation that finite horizons need not unravel once types are uncertain.

Language models and review integrity. Language models can now produce review text that human readers find useful (Liang et al., 2024b), and submissions carrying hidden prompts aimed at machine reviewers have been documented (Collu et al., 2025). Estimates place the fraction of reviews at recent top venues showing signs of language model assistance between 15% and 21% (Liang et al., 2024a; Jawahar et al., 2020), and detection of machine-written review text is unreliable at the false positive rates a program chair can tolerate. Our interest is more specific: writing a favorable review that survives co-reviewers and a meta-reviewer is the one genuinely costly step in the collusion sequence, and the step language models make cheap. The author-drafted variant of §4 is harder still, and beyond the reach of any machine-text detector.

Table 4 states the axes on which this paper departs from all of the above.

C Extended Simulation Details

Table 6 lists the parameters that set the scale of the simulation and the ranges swept, and the subsections below give the generative detail behind them. Venue parameters are calibrated to published statistics, attack and game parameters are swept over the stated ranges, and the five marked *assumed* are the ones no public source constrains. Appendix D varies the assumed dispersions, which are the two that carry a conclusion.

C.1 Generating a conference

Papers and reviewers occupy a shared simplex over $T = 50$ subject areas. Two popularity vectors are drawn from a Dirichlet with concentration 1.2, one for submission volume and one for reviewer expertise, and the reviewer vector is mixed with the paper vector at weight 0.7. Reviewer supply

Table 6: Simulation parameters. Conference parameters are calibrated to published ICLR/NeurIPS statistics; game parameters are swept over the stated ranges except where a default is used.

Parameter	Value	Source
Venue, calibrated		
Submissions N	6000	ICLR 2024 (OpenReview, 2026)
Reviewers M	9000	ICLR 2024 (OpenReview, 2026)
Reviews per paper k	3	venue policy
Reviewer load L_i	5	venue policy
Subject areas	50	venue call for papers
Score scale	1–10	venue policy
Score noise σ	1.2	fitted (OpenReview, 2026)
Acceptance rate	25%	ICLR 2024
Bid weight (eager/willing)	1.0 / 0.5	(Charlin & Zemel, 2013)
Bid participation	60%	calibration target
Eager bids per reviewer	≤ 8	calibration target
Attack, swept		
Colluder fraction	1, 3, 5, 10%	swept
Ring size k	2, 4, 6, 8, 10	swept
Inflation Δs	$+2.0$	swept
Partner affinity screen	97th pct.	§5.1
Game, swept		
Detection prob. d	$[0, 0.30]$	swept
Effort cost c_e	$[0, 0.60]\, b$	swept
Trust q	$[0, 1]$	swept
Discount δ_r	0.85	4–6 deadlines/yr
Sanction π	$5\, b$	assumed
Bid cost c_b	$0.02\, b$	assumed
Report loss L	$3\, b$	assumed
Spread of b, σ_b	0.35	assumed, §D
Spread of δ_r, σ_δ	0.10	assumed, §D

therefore does not track submission volume area by area, which produces over-bid and under-bid regions of the conference. The under-bid region matters for the attack: it is where a collusive bid faces little competition and wins the assignment on a weaker margin than it would elsewhere.

Reviewer mixtures are Dirichlet draws around the reviewer popularity vector with concentration 0.35, low enough that reviewers are specialized. Authorship is decided before paper content: a fraction 0.75 of reviewers author one submission each, and a paper's mixture is drawn with concentration 2.0 around a convex combination of its author's mixture, at weight 0.75, and the global popularity vector. Generating papers independently of their authors would place an author's submission far from their own expertise, so a partner's paper would be one the colluder could not plausibly review, and every strategy in §4 would appear weaker than it is.

Affinity is the cosine similarity of the two mixtures plus Gaussian noise of standard deviation 0.05, clipped to $[0, 1]$. It stands in for the text-matching score a venue computes from publication records. Latent quality θ_p is standard normal and independent of topics, which is what allows §5.3 to measure displacement: no observational dataset contains ground-truth quality. The only conflict of interest the venue detects automatically is authorship of the paper under review. Dyad partners are strangers by Definition 1, so no declared conflict exists between them, which is what the attack relies on.

C.2 Honest bidding

Each reviewer inspects the 60 submissions with the highest affinity to their own profile. Interest in an inspected paper is its affinity plus Gaussian noise of standard deviation 0.12, so two reviewers with the same profile do not bid on the same papers. A reviewer bids *eager* on an inspected paper

with probability $\sigma(12(\text{interest} - 0.55))$ where σ is the logistic function, up to a cap of 8 eager bids, and the next papers by interest become *willing* bids. A fraction 0.40 of reviewers place no bids at all.

The participation rate and the cap are calibration targets rather than modeling conveniences. If every high-affinity reviewer bid on every paper in their area, a bid would confer no advantage in the assignment and there would be nothing for a colluder to manipulate. The targets are a mean near six eager bids per paper with roughly a quarter of papers receiving two or fewer, which reproduces the concentration reported for real venues.

C.3 Injecting collusion

Only reviewers who author a submission are eligible, since a colluder needs a paper to trade. Rings are formed among the eligible pool by a screen on mutual affinity: writing A_{ij} for reviewer i's affinity to j's submission, the pair (i, j) is admissible when the symmetrized affinity $\frac{1}{2}(A_{ij} + A_{ji})$ exceeds the 97th percentile of each side's own affinity distribution. Rings are then grown greedily from a random order, sampling each next member among the eight most similar admissible candidates rather than taking the most similar, so that rings are not deterministic in the seed.

The screen is a substantive constraint rather than a detail. A recruitment post names a subject area and the people who answer one work in it, so partners are topically adjacent while remaining strangers in the sense of Definition 1. Pairing uniformly at random would understate every strategy, because a bid on a topically distant paper loses the assignment on affinity alone and is conspicuous besides. The screen is also why not every willing reviewer finds a usable partner, and why the number of completed rings falls as the ring grows.

Which members a colluder bids on follows the strategy. Under the naive, camouflaged and affinity-aware strategies every member bids on every other member's submission; under the distributed and combined strategies the ring is a directed cycle in which each member bids on exactly one successor, so the bid graph contains no two-cycle.

Camouflaging strategies add r cover bids per collusive bid, $r = 4$ by default. Cover papers are ranked by $A_{i\cdot} + 1.5\, s.$, where s is the maximum topical similarity to any targeted paper. Both terms are necessary. Ranking on similarity to the target alone drags the whole bid set away from the reviewer's area and makes the profile more anomalous, not less; ranking on the reviewer's own affinity alone leaves the target as isolated as it was. The bridging term is what stops the collusive bid from being an outlier within its own bid set. Appendix F reports the effect of r.

Affinity-aware strategies add 0.15 to the affinity of a targeted pair, clipped at one, and the assignment is solved against the manipulated matrix. This models profile curation and matched background text as an additive boost rather than as an actual editing procedure, which is a limitation: it fixes the gain a colluder achieves instead of deriving it from what the text permits.

C.4 Assignment

Problem (1) is solved as a linear program over a sparsified candidate set: the 30 highest-scoring reviewers per paper under affinity-plus-bid-weight, together with every reviewer-paper pair carrying a bid, less all authorship pairs. Retaining every bid pair is required for validity. Dropping a low-affinity collusive bid during sparsification would defend against the attack being measured, and would do so silently.

Coverage is an equality constraint at $k = 3$ reviews per paper, load an inequality at $L_i = 5$, and edge variables are bounded by a per-pair cap. The constraint matrix of a bipartite b-matching is totally unimodular, so the relaxation has an integral optimum; HiGHS returns one and the assignment is read off at $x > 0.5$.

Randomized assignment keeps the fractional solution with the per-pair cap set to 0.35 and samples an integral assignment from it, processing papers in random order and drawing reviewers without replacement from the fractional marginals while skipping reviewers already at their load. The cap is therefore enforced approximately rather than exactly, since the repair needed for feasibility can perturb a marginal. Cycle-free assignment is a post-hoc repair that finds pairs of reviewers assigned to each other's submissions and drops the lower-scoring of the two assignments. It is a heuristic

at $k = 2$, not an exact integer program; the exact formulation is NP-hard, and §6.1 reports it as implemented.

C.5 Reviewing and decisions

An assigned reviewer reports $5.5 + 1.5\,\theta_p + 1.2\,\varepsilon$ clipped to the 1 to 10 scale. The noise term ε is not drawn sequentially. It is a SplitMix64 hash of the triple (seed, reviewer, paper) mapped to a standard normal through the Box-Muller transform, so an honest reviewer's score on an honest paper is identical across runs. Sequential draws from a shared generator would not have this property: injecting collusion consumes a different number of draws, the honest scores would diverge between a collusive run and its baseline, and every displacement measurement in §5.3 would be contaminated by that divergence rather than by the collusion.

A collusive review occurs wherever a colluder is assigned to a paper they placed a collusive bid on. Uniform maximum sets the score to the top of the scale, calibrated inflation adds $\Delta s = 2.0$, and constructive positive adds $0.8\,\Delta s$. Author-drafted reviewing is not a fourth scoring arm: it differs from constructive positive in effort cost and in resistance to content-based scrutiny rather than in the score it produces, so it enters the analysis through c_e in §3 and not through the score distribution. Decisions accept the top 25% of submissions by mean score, with unreviewed papers ranked last.

C.6 Detectors and operating points

Bid frequency scores a reviewer by the isolation of their most isolated eager bids. Isolation of a bid is the negative of its maximum topical similarity to any other paper in the same bid set, plus the shortfall of its affinity below the reviewer's own 90th percentile, and the reviewer's score is the sum of the two largest. This is a stronger test than counting low-affinity bids, and it is the reason a camouflaging colluder must draw cover bids from around the target rather than from their own usual area.

Dense subgraph builds the reviewer-author graph, joining r and a when r bids on a paper a authored and a also reviews, and scores a reviewer by the edge density of the subgraph induced by their neighborhood, weighted by $1+$ the number of neighbors who bid back. We score locally rather than running Charikar peeling globally because peeling returns one densest subgraph, and the object of interest here is a population of many small disjoint rings.

Cycle detection orients the same graph and counts short directed cycles through each reviewer. The default bound of 2 is the constraint that deployed cycle-free assignment enforces; Appendix F reports bounds 3 and 4.

Text anomaly scores an assignment by how far its realized affinity falls below what the reviewer normally receives, taking the reviewer's own top-L_i affinities as the reference distribution and summing the positive z-scores over their assigned papers. The reference has to be the top of the distribution rather than the reviewer's mean, since almost any assignment beats a reviewer's average paper and a mean-based test flags nothing.

Every family is thresholded at the $1 - \text{FPR}$ quantile of its scores over honest reviewers, which makes the families comparable at a fixed investigation budget. That is the right common currency because a program chair's binding constraint is how many reviewers they can afford to examine, not any intrinsic property of a score. Ties at zero are handled explicitly, by flagging only scores strictly above the threshold, so a family that assigns the same score to most of the pool receives no credit for flagging all of them. The implementations are deliberately reasonable rather than deliberately weak: where a family admits an obvious strengthening we take it, because the claim being made is that these signals fail against specific countermeasures, and that claim is only worth making against detectors that work on the naive adversary.

C.7 Determinism and cost

Conference generation, honest bidding, collusion injection and assignment draw from independent substreams of one seed sequence per run, and review noise is hashed as described above. Honest bids are therefore identical whether or not collusion is injected. A collusive run and its baseline share the generated conference, so the two arms differ in the injected bids and in nothing else. One

Table 7: Measured transitions and effort-regime collusion levels against the spread of b and of δ_r, over 20,000 agents. The three rightmost columns are the quantities §5.1 quotes. The closed forms are evaluated at the population mean and so do not move with either spread.

	d	q	Collusion at c_e (%)		
Spread	trans.	trans.	hand	machine	author
no heterogeneity in either					
0	0.060	0.702	100.0	100.0	100.0
spread of b, holding $\sigma_\delta = 0.10$					
0.000	0.081	0.641	87.4	99.8	100.0
0.175	0.080	0.640	79.5	98.1	99.3
0.350	0.079	0.627	68.2	91.9	95.2
0.700	0.076	0.573	51.9	75.1	80.8
spread of δ_r*, holding* $\sigma_b = 0.35$					
0.000	0.076	0.650	69.0	95.1	98.1
0.050	0.077	0.642	68.9	94.1	97.3
0.100	0.079	0.627	68.2	91.9	95.2
0.200	0.081	0.585	65.4	86.5	89.9
closed form, at the population mean					
any	0.113	0.571	—	—	—

conference at the reported size takes about 2.5 seconds and peaks near 1.8 GB; the full suite of experiments reported here runs in about 20 minutes on one core, and re-running it reproduces every macro in the paper exactly.

D Sensitivity Analyses

Four inputs to §5 are set by assumption rather than by evidence: the dispersion of b and δ_r across the population, the false positive rate at which the detector families are compared, the acceptance rate, and the per-reader response rate of Proposition 9. This appendix varies each in turn and reports which conclusions move.

D.1 Heterogeneity in b and δ_r

Table 7 varies the two spreads. Setting both to zero is the informative case, because it isolates the explanation §5.1 gives for the distance between the measured transitions and the closed forms. That explanation is congestion, and if it is right the distance should survive an identical population. It does, and it widens: the d transition falls to 0.060 against $d^\star = 0.113$. The mechanism is visible in the same row. With one agent type every agent makes the same choice, so wherever collusion is sustainable at all the fixed point is 100.0% of the population, and the congestion multiplier is evaluated at its maximum. Dispersion spreads agents across their individual thresholds, holds the fixed point below one, and weakens the congestion term.

Across every setting swept, the d transition stays within $[0.076, 0.081]$ and the q transition within $[0.573, 0.650]$, so neither panel of Fig. 5 depends on the assumed dispersion. The c_e transition does, ranging over $[0.280, 0.472]$, and almost all of that movement comes from the spread of b rather than of δ_r. The reason is in (3): effort enters as c_e/b, so dispersion in b rescales the effective effort cost across the population.

§5.1 quotes no c_e transition, but it does quote the collusion level at three effort regimes, and those move: hand-written effort sustains 87.4% of the population at $\sigma_b = 0$ and 51.9% at $\sigma_b = 0.70$. The ordering that Appendix A.4 asserts survives every configuration, including the homogeneous one, where all three regimes saturate and the comparative static is invisible at the population level while remaining exactly as $\partial\delta^\star/\partial c_e > 0$ states it pointwise. What the spread of b sets is the magnitude of the effort effect, not its direction, and the percentages in §5.1 should be read as conditional on $\sigma_b = 0.35$.

Table 8: Best F_1 over the four detector families, by false positive rate, mean over 8 instances. Rows are bidding strategies. The operating point changes only the threshold, so detector scores are computed once and re-thresholded.

Strategy	1%	2%	5%	10%	20%
Naive	0.139	0.156	0.156	0.567	0.567
Camouflaged	0.261	0.322	0.322	0.551	0.551
Distributed	0.079	0.079	0.110	0.145	0.158
Affinity-aware	0.139	0.156	0.156	0.567	0.567
Camo.+dist.+affinity	0.127	0.158	0.216	0.228	0.221

Table 9: Impact against the acceptance rate, camouflaged dyads at 5% of the pool, mean over 8 instances. Boost is in percentage points over the same colluders' baseline rate.

Acceptance rate	Colluder rate (%)	Boost (pp)	Displaced papers	Undeserved papers
15%	17.4	+2.4	39	43
20%	23.0	+3.1	46	50
25%	28.2	+3.2	51	55
30%	33.0	+3.5	59	64
40%	43.4	+4.7	68	72

D.2 Detector operating points

The main text compares the four families at a single 5% false positive rate. Table 8 gives the best F_1 any of the four attains, at budgets from 1% to 20% of the honest pool.

The bound the paper claims is not an artifact of the operating point. Against the combined strategy the best F_1 over all four families and all five budgets is 0.228, below the 0.322 of Table 3, and the distributed strategy stays lower still. A ring longer than the cycle bound produces no cycle to count at any threshold, and camouflage leaves nothing for the frequency test to find at any threshold, so loosening the budget does not recover either.

The detectability of the *unsophisticated* attacker does depend on the budget, sharply. Naive bidding moves from $F_1 = 0.156$ at 5% to 0.567 at 10%. The jump is a threshold effect with a specific cause: only 6.8% of honest reviewers carry any two-cycle at all (Table 14), so a budget below that fraction places the threshold inside the honest nonzero mass, while a budget above it places the threshold at zero and flags every reviewer holding a single cycle. The 5% convention therefore sits just below a discontinuity in what cycle detection can do. A venue willing to examine one reviewer in ten identifies naive and camouflaged dyads at more than triple the F_1 of Table 3, which is worth knowing, and gains nothing against the strategies the recruitment threads actually describe.

D.3 Acceptance rate

The accept/reject rule is a top-k on realized paper means, so the rate can be varied after the scores are fixed. Table 9 does that from one collusive run and one baseline run per instance, which means the arms differ in the threshold and in nothing else; the 25% row reproduces §5.3 exactly.

No rate in the range removes the effect. The boost grows with the acceptance rate, from 2.4 percentage points at 15% to 4.7 at 40%, and the count of displaced honest papers grows with it as well, from 39 to 68, since a larger accepted set offers more decisions to disturb. Selectivity therefore reduces the absolute harm a fixed population of colluders can do while leaving the colluders' advantage intact, and a venue at 15% still admits 17.4% of colluding submissions against 43.4% at 40%.

D.4 Audience size

Proposition 9 leaves the per-reader response rate λ free, and §3 asserts that a public group clears $n^\star$ at any plausible value. Table 10 quantifies that. Even at 1%, one reply per hundred readers, $n^\star = 4.9$ readers, and at 20% it is 0.2. A channel with tens of members clears the threshold by orders of magnitude, so audience size is not what restrains posting. The binding condition is $\Gamma > c_s + \mu\pi_s$,

Table 10: Minimum audience $n^\star$ of Proposition 9 against the per-reader response rate λ, at the default payoffs. Values below one mean a single expected reply justifies the post.

Response rate λ	$n^\star$ (readers)
1%	4.88
2%	2.43
5%	0.96
10%	0.47
20%	0.22

that a partnership be worth more than the cost of soliciting one, which is where a venue's monitoring intensity μ enters and why it is the only lever in the model that acts before any bid is placed.

D.5 Tolerance frontier

Fig. 4 plots the feasibility frontier of Proposition 5 over (ρ, δ_r). The region in which no k-strike rule both deters free-riding and survives unlucky runs of non-assignment is exactly the region a venue is trying to reach by lowering ρ, and it widens as δ_r falls.

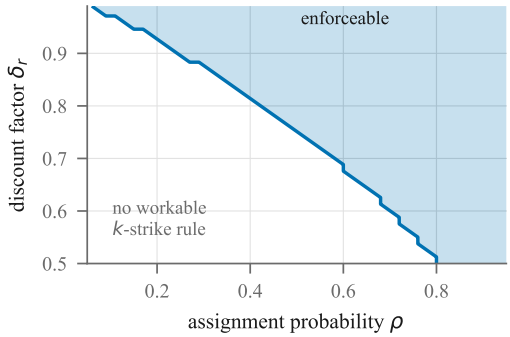
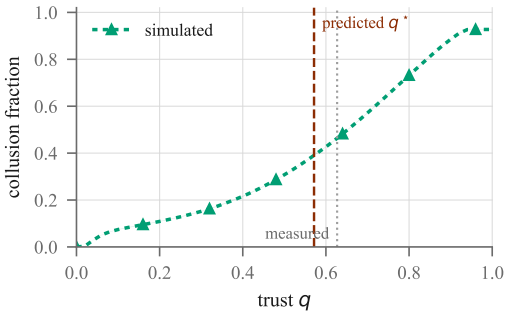


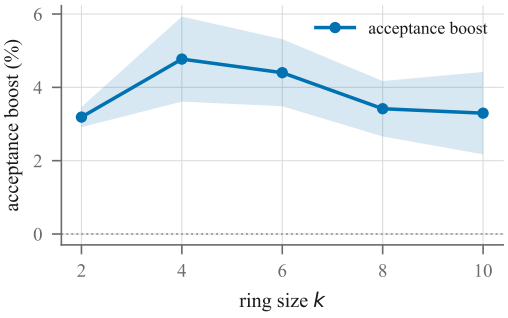
Figure 4: Feasibility frontier for the k-strike rule of Proposition 5. Inside the shaded region $k_{\min} \le k_{\max}$ and the partnership can police itself; outside it, no tolerance setting deters a partner who quietly withholds their bid without also destroying honestly cooperating partnerships.

E Additional Experimental Figures

The panels below are referenced from §5 and carry results the main text quotes numerically.

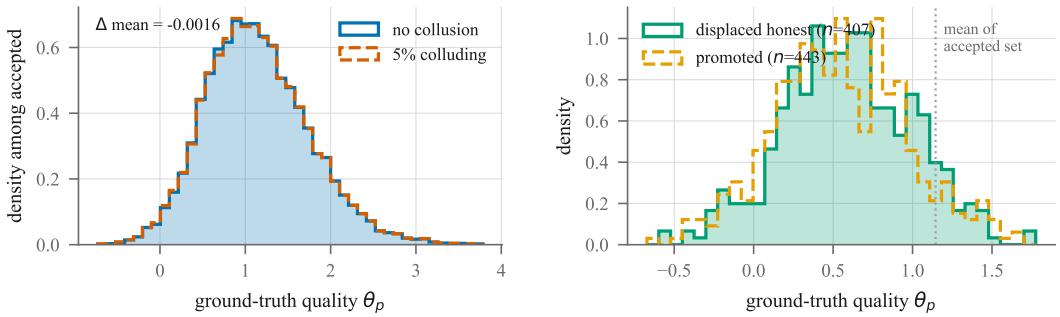


(a) Collusion fraction against the trust parameter q (§5.1). The dashed line is the closed-form $q^\star$ and the dotted line the measured transition.



(b) Acceptance gain against ring size (§5.3). The peak at $k = 4$ is the interior optimum of Proposition 10.

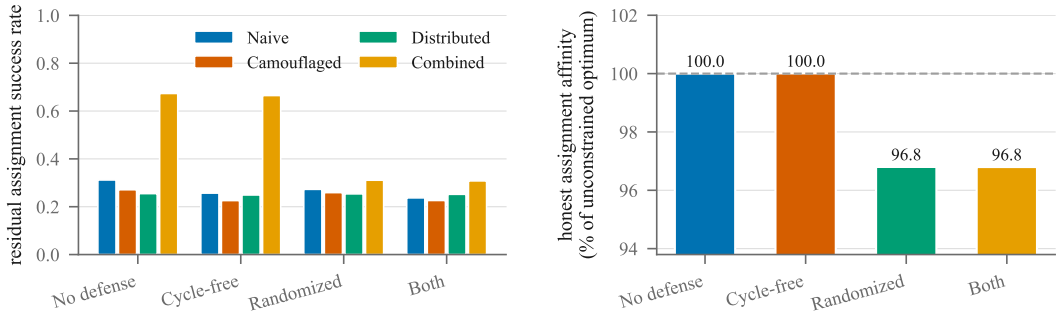
Figure 5: Equilibrium validation in q, and the ring-size optimum.



(a) The accepted set in aggregate.

(b) The papers that changed hands.

Figure 6: Ground-truth quality with and without collusion at a 5% colluder fraction (§5.3). In (a) the two distributions are almost identical and mean quality moves by 0.002, even though tens of papers changed hands. In (b) the displaced honest papers and the papers promoted in their place occupy the same narrow band at the acceptance threshold. The attack is invisible to any aggregate audit precisely because it operates where the two populations overlap.



(a) Residual assignment success by strategy.

(b) Cost to honest assignments.

Figure 7: Defense effectiveness over 8 instances (§6.1). Cycle-free assignment is nearly free (b) but only reaches dyads, and only those dyads whose two directions were both assigned. Randomization is the only defense in (a) that reduces the combined strategy.

Table 11: Strategy against detector family. ✓ denotes reliable evasion, ∼ partial evasion, ✗ detection. Entries are predictions from the model; measured F_1 appears in Table 3.

	Bid freq.	Dense subgr.	Cycle det.	Text anom.
Naive	✗	✗	✗	∼
Camouflaged	✓	∼	✗	∼
Distributed	✓	∼	✓	∼
Affinity-aware	∼	✗	✗	✓
Camo. + dist. + affinity	✓	✓	✓	✓

F Additional Detection Results

Table 11 states the outcome the model predicts for each pairing of strategy and detector family. Table 3 in §5.2 reports the measured counterpart.

F.1 Precision and Recall

Table 12 separates the two errors that Table 3 compresses into F_1. The distinction is not cosmetic, because two entries in the table would be read differently by a program chair than their F_1 suggests. The bid-frequency family reaches precision 1.000 against both the distributed and the combined strategy, at recall 0.041 and 0.068. When it fires it is never wrong, and it almost never fires. Its F_1 of 0.079 and 0.127 invites discarding a signal that at this operating point produces no false accusations

Table 12: Precision, recall and F_1 for every strategy against every detector family, at a common 5% false positive rate, mean over 8 instances. Table 3 reports the F_1 columns.

	Bid frequency			Dense subgraph			Cycle detection			Text anomaly		
Strategy	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Naive	0.000	0.000	0.000	0.063	0.075	0.069	0.362	0.100	0.156	0.043	0.052	0.047
Camouflaged	0.000	0.000	0.000	0.148	0.185	0.164	0.530	0.232	0.322	0.028	0.033	0.030
Distributed	1.000	0.041	0.079	0.098	0.127	0.110	0.092	0.019	0.032	0.044	0.056	0.049
Affinity-aware	0.000	0.000	0.000	0.063	0.075	0.069	0.362	0.100	0.156	0.077	0.094	0.085
Camo.+dist.+affinity	1.000	0.068	0.127	0.186	0.261	0.216	0.254	0.075	0.115	0.056	0.071	0.063

Table 13: Camouflage dose-response: cover bids per collusive bid against detection F_1 at 5% false positive rate and assignment success, dyads at 5% of the pool, mean over 8 instances. Honest reviewers place 4.8 eager bids on average. Ratio 0 is the naive strategy.

Cover	Collusive	Eager	F_1 by family				Assign.
bids	share	bids	freq.	dense	cycle	text	success
$1\times$	0.258	6.7	0.000	0.096	0.205	0.038	0.296
$2\times$	0.186	7.6	0.000	0.117	0.251	0.032	0.285
$4\times$	0.124	9.5	0.000	0.164	0.322	0.030	0.271
$8\times$	0.078	13.4	0.000	0.245	0.416	0.034	0.232

at all, which is the error a venue is least able to afford. Cycle detection against camouflage is the opposite shape, precision 0.530 at recall 0.232, and it holds the highest F_1 in the table.

F.2 How much camouflage

§5.2 notes that camouflage raises the cycle-detection score. Table 13 gives the dose-response and shows the effect is monotone over the whole range. As cover bids per collusive bid rise from 1 to 8, the collusive share of the bid vector falls from 0.258 to 0.078, which is what camouflage is for, while F_1 rises for both graph-based families, from 0.096 to 0.245 for dense subgraph and from 0.205 to 0.416 for cycle detection. The frequency test stays at 0.000 throughout and the text test is unmoved.

Two mechanisms drive the graph columns. Each cover bid adds an edge to the reviewer-author graph, so a heavily camouflaged colluder has more chances to close an incidental cycle with an honest author; and the bid vector grows to 13.4 entries against an honest mean of 4.8, so the neighborhood whose density the dense-subgraph test measures is larger than any honest reviewer's. Camouflage also costs assignment success, which falls from 0.296 to 0.232, because cover bids are eager bids competing for the same reviewer load.

Camouflage therefore has an interior optimum and it lies at the low end. The attacker's best choice is the smallest ratio that zeroes the frequency test, which is the smallest ratio we test, and heavier camouflage is worse on every axis measured: more visible to two families, no less visible to the other two, and less likely to win the assignment. The description of camouflage as cheap in §4 holds at the low ratio; the assumption that more of it is better does not. We also note a signal none of the four families uses. Bid volume is what heavy camouflage inflates, and a test on volume alone would separate the $8\times$ row from the honest population immediately. That is a concrete instance of the multi-signal argument of §6.

F.3 Cycle detection beyond two

A distributed ring evades a two-cycle constraint by construction, and the obvious reply is to raise the bound. Table 14 sweeps bounds 2 through 4 against rings of 2 through 6 members, reporting F_1 alongside the fraction of honest reviewers who carry at least one cycle at that bound.

Raising the bound works when it matches the ring. A three-member ring goes from 0.032 at bound 2, which is evasion, to 0.248 at bound 3. It does not keep working. The same ring falls to 0.144 at bound 4, and a four-member ring reaches only 0.141 even though the bound matches it. The reason is in the honest column: the share of honest reviewers carrying a cycle rises from 0.068 at bound 2 to 0.180 at bound 4. Longer cycles are ordinary in honest bidding, so a longer bound admits far

Table 14: Cycle detection at bound L against ring size, distributed bidding at 5% of the pool, 5% false positive rate, mean over 8 instances. *honest* is the fraction of honest reviewers holding at least one cycle at that bound, which is what limits the threshold.

	$L = 2$		$L = 3$		$L = 4$	
Ring size	F_1	honest	F_1	honest	F_1	honest
2	0.156	0.068	0.228	0.119	0.138	0.180
3	0.032	0.067	0.248	0.118	0.144	0.180
4	0.041	0.068	0.106	0.118	0.141	0.180
6	0.031	0.068	0.098	0.119	0.095	0.180

more honest mass than collusive signal, and a fixed false positive budget then buys a threshold that separates less than it did.

The exchange favors the attacker on both sides. Lengthening the ring costs one additional partner from the same recruitment thread, which Proposition 10 prices, and it is available to anyone who can read the thread. Matching it costs the venue an enumeration whose time per instance grows from 0.01 to 0.61 seconds here and grows faster than that with the bound, together with a false positive base that more than doubles. At bound 4 a ring of 6 still sits at 0.095. Cycle detection is a bound-guessing game, and raising the bound alone does not win it.

F.4 On temporal dispersion

A strategy that spreads collusive bids across the bidding period is not reported here, and the reason is a property of the detector families rather than a gap in the sweep. None of the four uses bid timing: each reads the bid set, the bid graph, or the affinity of the resulting assignment, and all three are invariant to the order in which bids arrive. A dispersed variant of any strategy in Table 12 would reproduce its row exactly. Measuring dispersion needs a fifth family built on timing and a calibrated arrival process for honest bids, and no venue publishes bid timestamps. We record it in Appendix G as future work, and note that arrival time is one of the few coordination problems a pair cannot solve without communicating further, so a timing-based family would attack the part of the arrangement that §2 identifies as most exposed.

G Extended Limitations and Future Work

The repeated game is between a fixed pair across rounds. Real partnerships form, dissolve, and reform through the platform, and a network formation model would capture the reputation mechanism we describe informally in §3 rather than folding it into δ_r. We also model a single venue: cross-venue collusion, in which a ring operates simultaneously at several conferences that share no data, both raises the effective number of rounds and lowers the effective detection probability, and we discuss it without modeling it. Finally, we treat detection probability d as exogenous, whereas a venue that deploys a detector faces an adversary who adapts, and the fixed point of that interaction is not the fixed point we compute.

Several parts of this paper are deliberately shallow and are being taken up separately: a multi-signal detection framework built on the observation that the countermeasures in §4 are mutually constraining; an adaptive adversary analysis in which detector and strategy co-evolve; the expertise-familiarity trade-off, which governs how much room a venue has to exclude suspicious matches before review quality suffers; cross-venue collusion, which our model says should be strictly easier to sustain than single-venue collusion; and the empirical question of whether author-drafted reviews leave a stylometric signature.

Two specific detector questions are left open by Appendix F. Bid timing is a signal none of the four families uses, and it is the one dimension along which a pair cannot coordinate without communicating further, so a timing-based family would bear on the stage of the arrangement the model identifies as most exposed. Evaluating one requires a calibrated arrival process for honest bids, which no venue publishes. Bid volume is a signal our frequency family also does not use, and it is what heavy camouflage inflates; the dose-response in Table 13 suggests that volume and topical

isolation together constrain the attacker more tightly than either does alone, which is the multi-signal argument stated for a case where we can measure both sides of it.