SHAPLEY DATA VALUATION UNDER COMMITMENT: ADAPTIVE MANIPULATION AND INTERACTION LAUNDERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Shapley-based data valuation can be manipulated when a provider adapts its submission to the coalition being evaluated. However, it remains unclear how restricting such adaptation affects manipulation and whether sufficiently strong commitment can eliminate overvaluation. We study this question using commitment partitions, which specify which coalition evaluations must reuse the same provider action. We characterize the loss in maximal manipulation gain caused by commitment and show that stronger commitment reduces, but does not eliminate, manipulation. In particular, even under one fixed submission, a multi-artifact provider can increase its Shapley payment while no coalition utility improves. We identify the underlying mechanism as *interaction laundering*, where interaction value shifts toward interactions in which the provider owns a larger share, and realize this behavior in both entity-matching and learned-representation settings. Finally, we show that adaptive and committed manipulation require different defenses. Reversible adaptive manipulation calls for removing exposed payment sensitivity, while one-sided committed degradation can be addressed by *Best-Bundle*, which applies free disposal before provider-level allocation.

1 INTRODUCTION

Shapley-based data valuation (Shapley, 1953; Ghorbani & Zou, 2019; Jia et al., 2019) assigns payments from the utilities of many counterfactual coalitions. This creates a strategic problem when the evaluated data are supplied by providers who can choose how their artifacts are submitted or represented. Recent work has shown that a provider can increase its Shapley payment by adapting its report to the coalition being evaluated, even while lowering the utility of selected coalition evaluations (Zheng et al., 2025a). Such attacks rely on a highly adaptive interface in which different coalition evaluations may receive different provider actions. This raises two related questions: how much manipulation is removed when the same action must be reused across multiple coalition evaluations, and is one fixed submission sufficient to prevent overvaluation?

To model this restriction, we introduce a commitment partition that specifies which coalition evaluations must reuse the same provider action. At one extreme, a provider may choose its action separately for each coalition evaluation; at the other, it must reuse one fixed action throughout valuation. This framework separates how the valuation rule responds to utility changes from which changes a provider can induce under a given level of commitment.

We then characterize how commitment limits manipulation. Allowing more coalition-specific adaptation can only increase the maximal manipulation gain. We further derive an exact commitment gap that measures the loss caused by forcing several coalition evaluations to share one action. The gap is zero exactly when all evaluations in each block share a common optimal action, and is positive otherwise. Thus, commitment reduces manipulation by coupling decisions across coalition evaluations, but need not eliminate it.

We next show that full commitment does not, in general, eliminate overvaluation under Atomic Shapley valuation. The key case arises when one provider owns multiple artifacts and its payment is the sum of their Shapley values. A coalition containing only part of the provider's portfolio can then enter the provider's total payment with a negative coefficient. Consequently, one fixed action

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.

can make such a coalition less useful while another artifact of the same provider restores the lost utility when both artifacts are present. In our minimal construction, no coalition utility improves and the grand coalition remains unchanged, yet the provider's payment increases. We call this phenomenon *interaction laundering*. In the Harsanyi representation, the same mechanism appears as a transfer of interaction value toward interactions in which the provider owns a larger fraction of the participating artifacts. We realize this behavior with fixed submissions in both entity-matching and learned-representation settings.

The distinction between adaptive and committed manipulation also leads to different defenses. Under reversible coalition-adaptive manipulation, robustness requires the payment rule to be insensitive to feasible manipulation directions. The committed attack above is instead one-sided, so such complete insensitivity is not necessary. We therefore define *Best-Bundle*, which selects the best available sub-bundle before allocating value across providers. For the committed attacks considered here, where bundles excluding the provider are unchanged and bundles using its artifacts can only degrade, Best-Bundle prevents such degradation from increasing the provider's payment.

Our contributions are as follows.

- We separate the payment sensitivity of a linear valuation rule from the joint coalition-utility changes induced by provider actions, and introduce commitment partitions to model how actions must be reused across coalition evaluations. We characterize how commitment limits manipulation and derive an exact commitment gap determined by whether evaluations forced to share an action admit a common optimal action.
- We show that full commitment does not eliminate overvaluation under Atomic Shapley valuation with multi-artifact ownership. We identify *interaction laundering*, characterize it through Harsanyi owner density, and realize it with fixed submissions in both entity-matching and learned-representation settings.
- We show that adaptive and committed manipulation require different defenses. Reversible adaptive manipulation requires zero payment sensitivity along feasible reversible directions, while *Best-Bundle* combines free disposal with provider-level allocation and prevents payment increases when bundles excluding the provider are unchanged and bundles using its artifacts can only degrade.

2 Problem Setup

We consider a valuation setting in which data artifacts are evaluated through coalition utilities and are supplied by strategic providers. A provider's action may change the utilities of several coalitions, and the valuation interface determines whether the provider must reuse the same action across coalition evaluations or may adapt its action to the coalition being evaluated. We first describe how coalition utilities determine provider payments and then formalize the provider's actions, commitment, and manipulation gain.

Atomic Shapley valuation. Let $\mathcal{E}$ be a fixed set of n artifacts and $v : 2^{\mathcal{E}} \to \mathbb{R}$ a deterministic, normalized utility game with $v(\varnothing) = 0$. Let $\mathcal{P}$ be the set of verified providers. Each provider $i \in \mathcal{P}$ owns an artifact set $E_i \subseteq \mathcal{E}$, and $\{E_i : i \in \mathcal{P}\}$ forms a partition of $\mathcal{E}$. Artifact e has Shapley value (Shapley, 1953)

$$\phi_e(v) = \sum_{S \subseteq \mathcal{E} \setminus \{e\}} \frac{|S|!(n - |S| - 1)!}{n!} \big[v(S \cup \{e\}) - v(S)\big].$$

The Atomic payment to provider i is the sum of the Shapley values of its artifacts, $\Phi_i(v) = \sum_{e \in E_i} \phi_e(v)$. Expanding the Shapley sums and collecting terms by coalition gives

$$\Phi_i(v) = \sum_{S \subseteq \mathcal{E}} \beta_i(S) v(S),$$

where $\beta_i(S)$ is the coefficient of coalition utility $v(S)$ in provider i's payment. Thus, a change $\Delta(S)$ in $v(S)$ contributes $\beta_i(S)\Delta(S)$ to the payment change. In particular, if $\beta_i(S) < 0$, lowering $v(S)$ increases provider i's payment when the other coalition utilities are held fixed.

Some of our results require only that the payment be linear in coalition utilities. For these results, we consider the general linear payment rule

$$F_i(v) = c_i + \sum_{S \subseteq \mathcal{E}} \beta_i(S) v(S) = c_i + \langle \beta_i, v \rangle. \tag{1}$$

Atomic Shapley is recovered by setting $c_i = 0$ and using its corresponding coefficient vector β_i.

Provider actions and induced utility changes. Fix the platform's evaluation rule and the other providers' actions. Let Θ_i be the physical action set of provider i, and let $\theta_i^0 \in \Theta_i$ denote its honest action. The honest action induces the utility game v^0. An alternative action $\theta \in \Theta_i$, such as an upload, schema, or fixed portfolio, induces a game v^θ and the utility change

$$\Delta_\theta(S) = v^\theta(S) - v^0(S), \qquad S \subseteq \mathcal{E}.$$

Each action $\theta \in \Theta_i$ induces a joint pattern of utility changes across coalition evaluations, with the honest action inducing zero change. Application constraints can be encoded in Θ_i; for example, external invariance requires $\Delta_\theta(S) = 0$ whenever $S \cap E_i = \varnothing$, while a protected grand coalition requires $\Delta_\theta(\mathcal{E}) = 0$.

Commitment across coalition evaluations. The attainable utility changes depend not only on the physical action set Θ_i, but also on whether the provider may choose different actions for different coalition evaluations. We use a commitment partition to represent this restriction. A commitment partition $\Pi = \{Q_1, \ldots, Q_m\}$ partitions the coalition evaluations into blocks that must reuse the same provider action. For each block $Q \in \Pi$, provider i chooses one action $\theta_Q \in \Theta_i$, which determines $\Delta(S) = \Delta_{\theta_Q}(S)$ for every $S \in Q$.

We write $\mathcal{D}_i(\Pi)$ for the set of utility-change vectors attainable under partition Π. A vector Δ belongs to $\mathcal{D}_i(\Pi)$ if each block $Q \in \Pi$ can be assigned one action $\theta_Q \in \Theta_i$ such that $\Delta(S) = \Delta_{\theta_Q}(S)$ for every $S \in Q$. This basic partition model permits the block actions $\{\theta_Q : Q \in \Pi\}$ to be chosen independently. The two extreme partitions correspond to the two main settings considered in this paper. The one-block partition $\Pi_{\text{fix}} = \{2^{\mathcal{E}}\}$ requires one fixed action to be reused across all coalition evaluations. We use $\mathcal{D}_i := \mathcal{D}_i(\Pi_{\text{fix}})$ for this fixed-action case. At the other extreme, $\Pi_{\text{ad}} = \big\{\{S\} : S \subseteq \mathcal{E}\big\}$ allows the provider to choose its action separately for every coalition evaluation and corresponds to coalition-conditioned reporting. Intermediate partitions represent partial adaptation.

Manipulation gain. For any attainable utility change $\Delta \in \mathcal{D}_i(\Pi)$, linearity of equation 1 gives

$$F_i(v^0 + \Delta) - F_i(v^0) = \sum_{S \subseteq \mathcal{E}} \beta_i(S) \Delta(S) = \langle \beta_i, \Delta \rangle. \tag{2}$$

The maximal payment gain under commitment partition Π is therefore $\Gamma_i(\Pi; \beta_i) = \sup_{\Delta \in \mathcal{D}_i(\Pi)} \langle \beta_i, \Delta \rangle$. Equation 2 shows that manipulation depends on both the payment rule and the utility changes available to the provider. The coefficient vector β_i determines how coalition-utility changes affect payment, while $\mathcal{D}_i(\Pi)$ describes which joint changes are attainable under commitment partition Π. Section 3 studies how the attainable manipulation gain changes as the commitment partition varies.

3 Commitment and Manipulation

Using the commitment partitions defined in Section 2, this section quantifies how action reuse changes manipulation gain and then studies the fully adaptive and fully committed endpoints.

3.1 Effect of Commitment

We first study how the commitment partition affects the maximal manipulation gain defined in Section 2. A finer partition allows the provider to use different actions in more coalition evaluations. If Π' refines Π, every utility change attainable under Π remains attainable under Π'. Therefore,

$\mathcal{D}_i(\Pi) \subseteq \mathcal{D}_i(\Pi')$ and $\Gamma_i(\Pi;\beta_i) \le \Gamma_i(\Pi';\beta_i)$, so relaxing commitment can only increase the maximal manipulation gain.

This ordering is immediate, but it does not quantify how much gain is lost when several coalition evaluations must reuse one action. To characterize this loss, suppose Θ_i is finite. For each coalition S, define $q_S(\theta) = \beta_i(S)\Delta_\theta(S)$, $q_S^\star = \max_{\theta\in\Theta_i} q_S(\theta)$, and $\Theta_{i,S}^\star = \arg\max_{\theta\in\Theta_i} q_S(\theta)$. Here, $q_S(\theta)$ is the payment gain contributed by coalition S when action θ is used, while $q_S^\star$ is the largest gain that S could contribute if its action were chosen independently.

Because all coalitions in the same block $Q \in \Pi$ must reuse one action, their maximal joint contribution is $\max_{\theta\in\Theta_i} \sum_{S\in Q} q_S(\theta)$. Therefore,

$$\Gamma_i(\Pi;\beta_i) = \sum_{Q\in\Pi} \max_{\theta\in\Theta_i} \sum_{S\in Q} q_S(\theta).$$

Under the fully adaptive partition, each coalition chooses its action independently, giving $\Gamma_i(\Pi_{\mathrm{ad}};\beta_i) = \sum_{S\subseteq\mathcal{E}} q_S^\star$. We define the commitment gap as

$$\mathrm{Gap}_i(\Pi) = \Gamma_i(\Pi_{\mathrm{ad}};\beta_i) - \Gamma_i(\Pi;\beta_i).$$

Theorem 1 (Commitment gap)**.** *Under the finite-action blockwise model above,*

$$\mathrm{Gap}_i(\Pi) = \sum_{Q\in\Pi} \left[\sum_{S\in Q} q_S^\star - \max_{\theta\in\Theta_i} \sum_{S\in Q} q_S(\theta) \right] = \sum_{Q\in\Pi} \min_{\theta\in\Theta_i} \sum_{S\in Q} \left[q_S^\star - q_S(\theta) \right] \ge 0.$$

Moreover, $\mathrm{Gap}_i(\Pi) = 0$ *if and only if* $\bigcap_{S\in Q} \Theta_{i,S}^\star \neq \varnothing$ *for every* $Q \in \Pi$.

The theorem shows that the commitment gap arises exactly when coalitions in the same block do not share an optimal action. If Π' refines Π, then $0 \le \mathrm{Gap}_i(\Pi') \le \mathrm{Gap}_i(\Pi)$. We next examine the two endpoints: coalition-adaptive reporting under Π_{ad} and one fixed action under Π_{fix}.

3.2 Adaptive Setting

At the fully adaptive endpoint, each coalition evaluation may use a different provider action. In the FL attack of Zheng et al. (2025a), the server requests a separate retraining for each block coalition S, and a client may report a different subset $\widehat{D}_i^S$ in each request. This corresponds to Π_{ad}.

Let $\Phi_{-i} = \sum_{j\neq i} \Phi_j$ denote the total payment to the other providers, with coefficient vector $\beta_{-i} = \sum_{j\neq i} \beta_j$. Under a manipulation Δ, the attacker's payment changes by $\beta_i^\top \Delta$, while the other providers' total payment changes by $\beta_{-i}^\top \Delta$. The deterministic analogue of their criterion requires, for every $\Delta \in \mathcal{D}_i(\Pi_{\mathrm{ad}})$,

$$\beta_i^\top \Delta \le 0, \qquad \beta_{-i}^\top \Delta \ge 0. \tag{3}$$

When the adaptive interface allows isolated changes to individual coalition utilities, these conditions reduce to the sign restrictions of Zheng et al. (2025a). In particular, if the adaptive interface allows $v(S)$ to be increased or decreased independently of all other coalition utilities, then $\beta_i(S) = \beta_{-i}(S) = 0$. If it allows only an independent decrease in $v(S)$, then $\beta_i(S) \ge 0$ and $\beta_{-i}(S) \le 0$. These sign conditions may not hold when changing $v(S)$ also changes other coalition utilities. We next turn to the opposite endpoint, where one fixed action is reused throughout valuation.

3.3 Committed Setting and Interaction Laundering

We now turn to the fully committed endpoint Π_{fix}, where one fixed action is reused across all coalition evaluations. Full commitment does not eliminate profitable manipulation. A multi-artifact provider can still increase its payment through the coupled utility changes induced by one fixed action, a phenomenon we call *interaction laundering*.

A minimal committed example. Let provider A own artifacts $\{r, c\}$ and provider B own b. Provider A makes one fixed modification to r, while c reverses this modification whenever it accompanies r. Suppose the induced utility change satisfies

$$\Delta(r) = -\epsilon_1, \qquad \Delta(rb) = -\epsilon_2, \qquad \Delta(S) = 0 \text{ otherwise},$$

where $\epsilon_2 > \epsilon_1 > 0$. Thus, r becomes slightly worse alone and more strongly incompatible with b, while coalitions containing the repair artifact c, including the grand coalition, remain unchanged. No coalition utility improves, and the same action is reused throughout valuation. Nevertheless, A's Atomic payment can increase.

The coefficient view. Recall that the Atomic payment to provider i is $\Phi_i(v) = \sum_{e \in E_i} \phi_e(v)$. We now derive the coefficient $\beta_i(S)$ of coalition utility $v(S)$ in provider i's Atomic payment. Write $m_i = |E_i|$, $s = |S|$, and $k_i = |S \cap E_i|$.

Proposition 2 (Coalition coefficient)**.** *For $0 < s < n$, $\beta_i(S) = \frac{(s-1)!(n-s-1)!}{n!}\left(nk_i - m_i s\right)$, with $\beta_i(\varnothing) = -m_i/n$ and $\beta_i(\mathcal{E}) = m_i/n$. Hence,*

$$\beta_i(S) < 0 \quad \Longleftrightarrow \quad \frac{k_i}{m_i} < \frac{s}{n}.$$

Proposition 2 shows why multi-artifact ownership matters. If provider i owns a single artifact and coalitions excluding that artifact are unchanged, every coalition affected by i has a positive coefficient. Degrading the provider's contribution therefore cannot increase its Atomic payment. With multiple owned artifacts, however, a coalition containing only part of the provider's portfolio can have a negative coefficient.

For the minimal example above, $\beta_A(r) = 1/6$ and $\beta_A(rb) = -1/6$. Therefore, $\Phi_A(v^0 + \Delta) - \Phi_A(v^0) = \frac{\epsilon_2 - \epsilon_1}{6} > 0$. The larger utility loss on rb outweighs the smaller loss on r in A's payment, even though no coalition utility increases. For example, $\epsilon_1 = .05$ and $\epsilon_2 = .30$ give a payment increase of .0417.

The interaction view. The coefficient view identifies which coalition utilities drive the payment increase. To understand what the fixed action changes at the interaction level, we turn to the Harsanyi representation (Harsanyi, 1959). Let $h_v(T)$ denote the Harsanyi dividend of interaction T, and define provider i's owner density in T as $\rho_i(T) = \frac{|T \cap E_i|}{|T|}$. The Atomic payment can be written as

$$\Phi_i(v) = \sum_{\varnothing \neq T \subseteq \mathcal{E}} \rho_i(T) h_v(T).$$

For the minimal example, the corresponding dividend changes are $\delta h(r) = -\epsilon_1$, $\delta h(rc) = +\epsilon_1$, $\delta h(rb) = -(\epsilon_2 - \epsilon_1)$, and $\delta h(rcb) = +\epsilon_2 - \epsilon_1$. The first transfer moves value from r to rc. Both interactions have A-density one, so this transfer does not change A's payment. The second transfer moves value from rb, where A owns one half of the interaction, to rcb, where A owns two thirds. Its payment effect is $(\epsilon_2 - \epsilon_1)\left(\frac{2}{3} - \frac{1}{2}\right) = \frac{\epsilon_2 - \epsilon_1}{6}$, which matches the coefficient calculation.

This example suggests a more general pattern. For a grand-utility-preserving committed change Δ, let $\delta h(T) = h_{v^0 + \Delta}(T) - h_{v^0}(T)$ denote the change in the Harsanyi dividend of interaction T. Define $L = \{T \neq \varnothing : \delta h(T) < 0\}$ and $H = \{T \neq \varnothing : \delta h(T) > 0\}$ as the interactions that lose and gain value, respectively. We call Δ *interaction laundering* if the lost interaction value can be redistributed to the gaining interactions through nonnegative flows $f_{TT'}$. Specifically, each losing interaction must send exactly its lost value, and each gaining interaction must receive exactly its gain

$$\sum_{T' \in H} f_{TT'} = -\delta h(T) \quad (T \in L), \qquad \sum_{T \in L} f_{TT'} = \delta h(T') \quad (T' \in H).$$

Positive flow is allowed only when $\rho_i(T') \geq \rho_i(T)$, with a strict increase for at least one positive-flow pair.

Proposition 3 (Owner-density transport)**.** *Every interaction-laundering change satisfies*

$$\Phi_i(v^0 + \Delta) - \Phi_i(v^0) = \sum_{T \in L} \sum_{T' \in H} f_{TT'} \left[\rho_i(T') - \rho_i(T)\right] > 0.$$

Proposition 3 shows that moving value between interactions with the same owner density does not change the provider's payment. A gain arises when interaction value moves toward interactions in which the provider owns a larger share. The incompatibility-and-repair example above is one

realization of this pattern. A fixed laundering action can be selected using validation feedback and then frozen before paid evaluation; Appendix C.4 gives the search procedure.

Adaptive manipulation exploits coalition-specific action choices, whereas interaction laundering exploits the coupled utility changes induced by one fixed action. Section 4 develops defenses for these two settings.

4 Defenses for Adaptive and Committed Manipulation

This section develops defenses for the two attack regimes studied above. We first consider sensitivity restriction for coalition-adaptive manipulation, and then study robustness under one-sided committed degradation.

4.1 Sensitivity Restriction for Adaptive Manipulation

At the fully adaptive endpoint, coalition-specific actions can expose individual coalition utilities to independent manipulation. We ask what robustness requires when these changes are reversible. Suppose every attainable utility change can also be reversed, so that $\mathcal{D}_i(\Pi_{\mathrm{ad}}) = -\mathcal{D}_i(\Pi_{\mathrm{ad}})$. Robustness must then hold for both d and $-d$. For a linear payment, this forces $\beta_i^\top d = 0$ for every attainable d.

Theorem 4 (Robustness under reversible manipulation)**.** *Suppose* $F_i(v) = c_i + \langle \beta_i, v \rangle$ *and* $\mathcal{D}_i(\Pi_{\mathrm{ad}}) = -\mathcal{D}_i(\Pi_{\mathrm{ad}})$. *Then*

$$F_i(v^0 + d) \le F_i(v^0) \quad \textit{for all } d \in \mathcal{D}_i(\Pi_{\mathrm{ad}}) \quad \Longleftrightarrow \quad \beta_i \in \operatorname{span}\big(\mathcal{D}_i(\Pi_{\mathrm{ad}})\big)^{\perp}.$$

If the adaptive interface can independently increase or decrease the utility of a partial-owner coalition S while leaving the other coalition utilities unchanged, Theorem 4 requires $\beta_i(S) = 0$. Truth-Shapley/Owen follows this principle by valuing combinations of complete provider portfolios rather than manipulable partial-owner coalitions (Zheng et al., 2025a; Owen, 1977).

This conclusion relies on reversibility. The committed manipulation in Section 3.3 is instead one-sided, so robustness need not eliminate the corresponding payment sensitivity. We next consider a defense for this setting.

4.2 Best-Bundle for Committed Manipulation

The committed manipulation in Section 3.3 is one-sided: a provider can make coalitions using its artifacts less useful while leaving coalitions that do not use its artifacts unchanged. This suggests a different defense from the sensitivity restriction used in the adaptive setting. Before allocating value across providers, the platform can first discard artifacts that reduce coalition utility.

Best-Bundle. For a provider coalition $C \subseteq \mathcal{P}$, let $E(C) = \bigcup_{i \in C} E_i$ denote the artifacts owned by its members. The platform may choose any sub-bundle $T \subseteq E(C)$. We define

$$V_v^\star(C) = \max_{T \subseteq E(C)} v(T), \qquad \Psi_i^{\mathrm{BB}}(v) = \mathrm{Shapley}_i(V_v^\star).$$

We call Ψ_i^{BB} *Best-Bundle*.

This construction is equivalent to first applying the classical free-disposal closure $v^+(S) = \max_{T \subseteq S} v(T)$ (Peleg & Sudhölter, 2007; Solymosi, 2008), since $V_v^\star(C) = v^+(E(C))$. Thus, Best-Bundle first removes harmful artifacts when beneficial and then allocates the resulting value across verified providers.

Proposition 5 (Monotonicity under one-sided degradation)**.** *Suppose provider* i *changes the artifact game from* v *to* $\widetilde{v}$ *such that* $\widetilde{v}(T) = v(T)$ *whenever* $T \cap E_i = \varnothing$ *and* $\widetilde{v}(T) \le v(T)$ *whenever* $T \cap E_i \neq \varnothing$. *Then*

$$\Psi_i^{\mathrm{BB}}(\widetilde{v}) \le \Psi_i^{\mathrm{BB}}(v).$$

The proposition follows because provider coalitions excluding i have exactly the same available bundles, while the best utility available to a coalition containing i cannot increase. Therefore, every marginal contribution of provider i in the Best-Bundle provider game weakly decreases. A provider cannot increase its Best-Bundle payment by making bundles that use its own artifacts worse.

Why provider-level allocation matters. Free disposal alone is not sufficient. If Shapley value is still allocated at the artifact level after applying v^+, partial-owner coalitions remain in the payment and interaction laundering can persist. Best-Bundle instead performs the final allocation across verified providers.

Let $\mathcal{B} = \{E_i : i \in \mathcal{P}\}$ denote the verified ownership partition. If $\omega_e(v^+, \mathcal{B})$ denotes the Owen value of artifact e in the free-disposal game, coalitional efficiency gives

$$\Psi_i^{\mathrm{BB}}(v) = \sum_{e \in E_i} \omega_e(v^+, \mathcal{B}).$$

This relation separates three mechanisms. Merged/Truth-Shapley uses owner-aware allocation on the original game. Atomic Free Disposal applies free disposal but retains artifact-level allocation. Best-Bundle combines free disposal with provider-level allocation. The latter removes the partial-owner structure exploited by interaction laundering while still allowing useful sub-bundles to be selected.

Representation invariance. Best-Bundle depends on the best utility that a provider can supply in each external context, rather than on how its artifacts are internally represented.

Proposition 6 (Representation invariance)**.** *Let $E_{-i} = \bigcup_{j \neq i} E_j$. Suppose replacing E_i by E_i' leaves all utilities involving only E_{-i} unchanged and, for every $X \subseteq E_{-i}$,*

$$\max_{T \subseteq E_i} v(X \cup T) = \max_{T' \subseteq E_i'} v'(X \cup T').$$

Then the induced Best-Bundle provider games are identical, and hence all provider payments are unchanged.

Thus, reorganizing a provider's artifacts does not change its total Best-Bundle payment as long as the best attainable utility is preserved in every external context, although the allocation among individual artifacts may change. The monotonicity guarantee also extends to fixed restricted candidate ranges; Appendix B.7 gives the result, while Appendix D.6 reports the corresponding computational trade-offs.

5 Experiments

We evaluate three questions: how commitment affects manipulation gain, whether one fixed submission can realize interaction laundering, and whether the corresponding defenses remove the resulting gain. The main text reports the key held-out results, while full experimental protocols, additional controls, and detailed statistics are provided in Appendix C and Appendix D.

5.1 Commitment narrows but does not eliminate gain

In a six-artifact entity game, A owns r, c and four external artifacts expose field views. The 15 active contexts contain r, no c, and a nonempty external subset. Every task has six equal-cost actions (honest plus five field permutations on one fixed shard), exactly optimized under nested partitions with $1, 2, 4$, and 15 blocks. Validation selects and hashes the complete policy before held-out scoring.

The left panels of Figure 1 show that the fully committed action already obtains mean validation/test gains 0.00566/0.00521 on DBLP–ACM and 0.00115/0.00102 on Walmart–Amazon. Coalition-adaptive selection raises the validation optima to 0.00759 and 0.00133—a 34.1%/15.2% uplift—with frozen test gains 0.00730/0.00114. Every validation seed obeys nesting, yet commitment retains positive gain. The exact committed gaps are 0.00193/0.00017; all 40/40 response matrices lack a common optimizer, matching Theorem 1 (Appendix Table 3). Thus the experiment establishes more than the predicted ordering: positive gain remains even when one action is shared across all contexts. Separately, the absence of a common optimizer explains why the adaptive validation optimum is strictly larger than the committed optimum.

We additionally verify the coalition-adaptive endpoint with a centralized horizontal-block replay on Breast Cancer and Digits; all held-out runs yield positive gain and the observed payment changes match $\beta_A^\top \Delta$ to numerical precision (Appendix C.2).

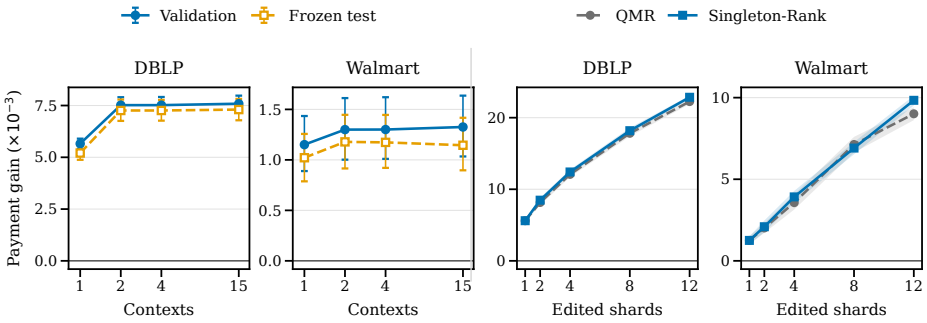

Figure 1: Commitment and fixed-submission attacks. Panels (a,b) show Atomic owner-A gain across commitment levels; (c,d) show held-out gain versus edit budget for QMR and Singleton-Rank. Curves average 20 seeds with 95% seed-bootstrap intervals; repaired grand utility is unchanged in (c,d).

5.2 Fixed submissions realize interaction laundering

We first establish that interaction laundering can be realized by one fixed submission on held-out data. We then test two alternative explanations: whether the observed gain depends on our search procedure, and whether it is specific to named schema manipulations.

On structured DBLP–ACM and Walmart–Amazon (Mudgal et al., 2018), A supplies table r, B supplies b, and A's artifact c provides the inverse field alignment. Validation selects one fixed multi-edit submission, which is frozen before held-out evaluation. Applying c restores the complete rcb representation exactly to the honest one. Across 20 seeds and five edit budgets, Singleton-Rank obtains mean held-out Atomic owner-A gains of 0.01351 and 0.00480, corresponding to 6.04% and 9.96% of the honest payment. All 200 held-out cells have positive gain while $\Delta v(rcb) = 0$. Because only rb deteriorates in this construction and $\beta_A(rb) = -1/6$, the payment increase is exactly its utility loss divided by six, matching the interaction-laundering mechanism in Section 3.3. Full protocols and diagnostics are reported in Appendices C.3, C.4, and D.

We next compare Singleton-Rank with query-matched random search under identical validation-query budgets and selected edit cardinalities. Query-matched random also produces positive held-out gain in all 200 cells. On DBLP–ACM, the mean gain is 0.01318 for query-matched random and 0.01351 for Singleton-Rank; on Walmart–Amazon, the corresponding values are 0.00459 and 0.00480. The paired held-out intervals exclude zero, showing that singleton information provides a modest improvement at the same feedback cost. More importantly, successful manipulation does not depend on Singleton-Rank. The right panels of Figure 1 show the gain across edit budgets, and Appendix C.4 gives the matched-search protocol.

Finally, we replace named fields with a learned representation interface. Using an ImageNet-pretrained ResNet-18 representation on CIFAR-10 (He et al., 2016; Krizhevsky, 2009), A publishes an invertibly reparameterized representation, while c provides the inverse alignment required by B's frozen consumer. Across 20 seeds, the held-out Atomic owner-A payment increases by 0.1113, while the unaligned rb accuracy falls from 82.5% to 58.2%. A's standalone utility and the repaired composite remain unchanged. Thus, the same interaction-laundering pattern persists without a named schema. Implementation details and complete audits are reported in Appendix C.5 and Table 7.

5.3 Defense Evaluation

Across all real-data schema-sensitive Singleton-Rank cells, Atomic Free Disposal retains the same gain as Atomic. Merged/quotient owner totals, Truth-Shapley/Owen owner totals, a validation-selected Best-Bundle policy evaluated after freezing, and the committed-range version have zero gain up to numerical tolerance (Appendix Table 6). Both committed games satisfy the one-sided degradation condition of Proposition 5: external bundles are unchanged, and every bundle using A weakly decreases or remains unchanged. The proposition therefore guarantees that the exact Best-Bundle payment cannot increase.

The controlled targeted-mismatch grid makes the attack-to-defense response visible in Figure 2. Across three datasets, increasing the mismatch fraction monotonically increases mean Atomic rent, while Merged and Best-Bundle suppress the entire curve.

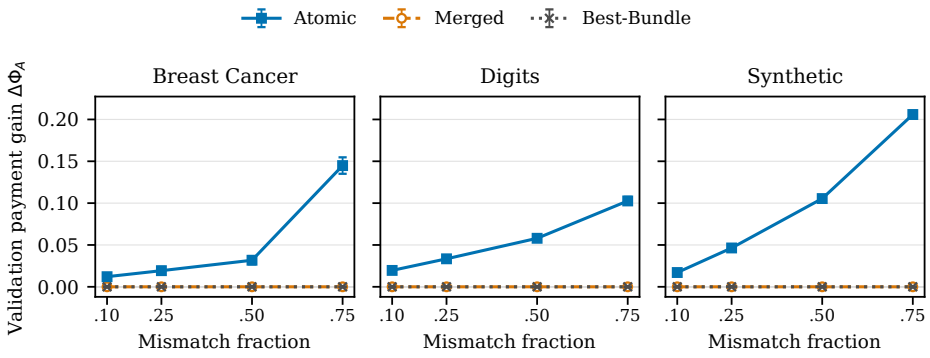

Figure 2: Targeted mismatch dose response. Atomic gain increases with mismatch, while Merged and Best-Bundle remain zero in all 240 cells. Means are over 20 seeds with 95% seed-bootstrap intervals. Full repair preserves grand utility; Proposition 5 covers 239 cells, with the remaining zero empirical.

In controls, Best-Bundle discards mislabeled overlays, rewards an independent repair in 168/180 controlled cells, and stays invariant under eight within-owner fragments. A Sybil repair restores gain 0.01167, confirming the identity boundary. Appendix D reports partial repair, selection cost, and model controls. Atomic Free Disposal is insufficient because it still allocates value at the artifact level. Best-Bundle instead combines free disposal with provider-level allocation. The independent-repair and Sybil controls show that this preserves value for genuine third-party repair while preventing a provider from profiting from repairing degradation created by its own artifacts.

6 Related work

Shapley-based data valuation has been studied for training examples, data sources, and structured or fragmented ownership (Ghorbani & Zou, 2019; Jia et al., 2019; Liu et al., 2023; Zheng et al., 2025b). Recent attacks manipulate attribution through example modification, synthetic distributed data, or coalition-dependent reporting (Wang et al., 2024; Gao et al., 2026; Zheng et al., 2025a). Our work focuses on how restricting such adaptation changes manipulation and shows that overvaluation can persist even under one fixed submission.

Related work on group valuation, replication, and false-name robustness studies how ownership and player grouping affect Shapley allocation (Owen, 1977; Ohrimenko et al., 2019; Lee et al., 2025; Burnat & Davidson, 2026). Best-Bundle uses the classical free-disposal closure (Peleg & Sudhölter, 2007; Solymosi, 2008) before provider-level allocation. Our contribution is not the closure itself, but its robustness connection to one-sided self-degradation and interaction laundering. We provide a more detailed comparison with related work in Appendix A.

7 Conclusion and limitations

In this work, we studied how commitment constrains manipulation in Shapley-based data valuation. Using commitment partitions, we showed that restricting coalition-specific adaptation can only reduce the maximal manipulation gain and characterized the exact loss through the commitment gap. However, full commitment does not eliminate overvaluation: a multi-artifact provider can still increase its payment through *interaction laundering*, which shifts value toward interactions in which it owns a larger share. We realized this behavior with fixed submissions in both entity-matching and learned-representation settings. Finally, we showed that adaptive and committed manipulation require different defenses: reversible adaptive manipulation calls for restricting payment sensitivity, while one-sided committed degradation can be addressed by Best-Bundle, which combines free disposal with provider-level allocation.

AI USE STATEMENT

Generative AI tools were used only for language polishing and editorial assistance. All research ideas, theoretical results, experiments, analyses, and claims were developed and verified by the authors.

ETHICS STATEMENT

This work does not involve new human-subject data collection. Experiments use public benchmark datasets and controlled or synthetic settings.

REPRODUCIBILITY STATEMENT

Appendix C specifies datasets, utility construction, all hyperparameters, task counts, confidence intervals, and stopping rules. Appendix B gives complete proofs. The code is available at `https://anonymous.4open.science/r/shapley-commitment-73D8`.

REFERENCES

Florian A. D. Burnat and Brittany I. Davidson. Quotient semivalues for false-name-resistant data attribution, 2026. URL https://arxiv.org/abs/2605.07663.

Ramón Flores, Elisenda Molina, and Juan Tejada. Evaluating groups with the generalized shapley value. *4OR*, 17(2):141–172, 2019. doi: 10.1007/s10288-018-0380-8.

Xian Gao, Bo Hui, Min-Te Sun, and Wei-Shinn Ku. On the fragility of data attribution when learning is distributed, 2026. URL https://arxiv.org/abs/2605.15520.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251. PMLR, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

John C. Harsanyi. A bargaining model for the cooperative n-person game. In Albert W. Tucker and R. Duncan Luce (eds.), *Contributions to the Theory of Games IV*, volume 40 of *Annals of Mathematics Studies*, pp. 325–355. Princeton University Press, 1959.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the shapley value. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1167–1176. PMLR, 2019. URL https://proceedings.mlr.press/v89/jia19a.html.

Piotr Kępczyński and Oskar Skibski. Sequential elimination and union shapley value for group assessment in coalitional games, 2025.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kiljae Lee, Ziqi Liu, Weijing Tang, and Yuan Zhang. Faithful group shapley value. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://neurips.cc/virtual/2025/poster/115078.

Zhihong Liu, Hoang Anh Just, Xiangyu Chang, Xi Chen, and Ruoxi Jia. 2D-shapley: A framework for fragmented data valuation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21730–21755. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23s.html.

Sidharth Mudgal, Han Li, Theodoros Rekatsinas, AnHai Doan, Youngchoon Park, Ganesh Krishnan, Rohit Deep, Esteban Arcaute, and Vijay Raghavendra. Deep learning for entity matching: A design space exploration. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pp. 19–34. ACM, 2018. doi: 10.1145/3183713.3196926.

Aniket Murhekar, Jiaxin Song, Parnian Shahkar, Bhaskar Ray Chaudhury, and Ruta Mehta. You get what you give: Reciprocally fair federated learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45289–45310. PMLR, 2025. URL https://proceedings.mlr.press/v267/murhekar25a.html.

Olga Ohrimenko, Shruti Tople, and Sebastian Tschiatschek. Collaborative machine learning markets with data-replication-robust payments, 2019. URL https://arxiv.org/abs/1911.09052.

Guillermo Owen. Values of games with a priori unions. In Rudolf Henn and Otto Moeschlin (eds.), *Mathematical Economics and Game Theory*, volume 141 of *Lecture Notes in Economics and Mathematical Systems*, pp. 76–88. Springer, 1977. doi: 10.1007/978-3-642-45494-3_7.

Bezalel Peleg and Peter Sudhölter. *Introduction to the Theory of Cooperative Games*, volume 34 of *Theory and Decision Library C*. Springer, Berlin, Heidelberg, 2 edition, 2007. doi: 10.1007/978-3-540-72945-7.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games II*, volume 28 of *Annals of Mathematics Studies*, pp. 307–317. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Tamás Solymosi. Bargaining sets and the core in partitioning games. *Central European Journal of Operations Research*, 16(4):425–440, 2008. doi: 10.1007/s10100-008-0070-2.

Xinhe Wang, Pingbang Hu, Junwei Deng, and Jiaqi W. Ma. Adversarial attacks on data attribution, 2024. URL https://arxiv.org/abs/2409.05657.

Dengji Zhao, Yiqing Huang, Liat Cohen, and Tal Grinshpoun. Coalitional games with stochastic characteristic functions defined by private types. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*, pp. 2086–2088. International Foundation for Autonomous Agents and Multiagent Systems, 2020. doi: 10.5555/3398761.3399083. URL https://arxiv.org/abs/1910.11737.

Shuyuan Zheng, Sudong Cai, Chuan Xiao, Yang Cao, Jianbin Qin, Masatoshi Yoshikawa, and Makoto Onizuka. Data overvaluation attack and truthful data valuation in federated learning, 2025a. URL https://arxiv.org/abs/2502.00494.

Xi Zheng, Yinghui Huang, Xiangyu Chang, Ruoxi Jia, and Yong Tan. Rethinking data value: Asymmetric data shapley for structure-aware valuation in data markets and machine learning pipelines, 2025b. URL https://arxiv.org/abs/2511.12863.

A Additional Related Work

Data valuation and manipulation. Shapley-based data valuation has been widely used to quantify the contribution of training examples and data sources (Ghorbani & Zou, 2019; Jia et al., 2019). Later work extends this view to settings with richer data structure or ownership. For example, 2D-Shapley studies fragmented data valuation across multiple dimensions (Liu et al., 2023), while asymmetric variants incorporate directional or structural differences among data contributions (Zheng et al., 2025b). A separate line of work studies strategic manipulation of data attribution. Existing attacks can modify examples to inflate attribution (Wang et al., 2024), introduce synthetic distributed batches (Gao et al., 2026), or adapt a federated-learning report to the coalition being evaluated (Zheng et al., 2025a). The last setting is the closest to our adaptive endpoint. Our focus is different: rather than assuming fully coalition-dependent reporting, we study how manipulation changes as actions must be reused across evaluations, and show that overvaluation can persist even when the provider is restricted to one fixed submission. Our federated-learning experiment is therefore used only to reproduce the coalition-adaptive utility geometry studied in prior work; the new committed attack is evaluated separately through fixed-submission interfaces.

Ownership, replication, and group valuation. Several lines of work study how ownership and grouping affect cooperative data valuation. Owen's value evaluates games with an exogenous coalition structure, while generalized group values and Union Shapley provide related ways to aggregate or assess groups of players (Owen, 1977; Flores et al., 2019; Kępczyński & Skibski, 2025). These constructions are relevant when several artifacts belong to one provider. In particular, summing Owen values within an ownership block agrees with the corresponding owner-level quotient-game payment. Other work studies strategic issues created by replication, splitting, and pseudonymous identities, including replication-robust payments, faithful group valuation, and quotient semivalues for false-name resistance (Ohrimenko et al., 2019; Lee et al., 2025; Burnat & Davidson, 2026). Strategic federated-learning reward mechanisms have also modeled contribution costs and reciprocal incentives (Murhekar et al., 2025). These lines of work motivate our assumption that ownership is verified and fixed. Our threat model is different from replication or false-name manipulation: the provider keeps the same registered identity and ownership set, but changes the joint utilities induced by its submitted artifacts. Interaction laundering exploits the resulting partial-owner coalition structure rather than creating duplicate or pseudonymous players.

Cooperative-game constructions and free disposal. Our Best-Bundle construction uses two classical cooperative-game ideas rather than introducing a new allocation primitive. First, the free-disposal or monotonic-cover operation replaces an artifact game v by $v^+(S) = \max_{T \subseteq S} v(T)$, allowing harmful artifacts to be discarded (Peleg & Sudhölter, 2007; Solymosi, 2008). Second, allocation is performed at the provider level, which is closely related to applying the Owen value under the verified ownership partition (Owen, 1977). This distinguishes Best-Bundle from two nearby constructions. Merged or Truth-Shapley/Owen allocation respects provider ownership but treats each provider portfolio as inseparable, whereas Atomic Free Disposal first applies v^+ but continues to allocate value at the artifact level. Best-Bundle combines free disposal with provider-level allocation, retaining sub-bundle selection while removing the partial-owner allocation structure exploited by interaction laundering. Our contribution is therefore not the classical free-disposal closure or Owen allocation themselves, but the connection between this combination and robustness to one-sided self-degradation, together with the corresponding monotonicity guarantee and its role in defending against committed manipulation.

Endogenous and private characteristic functions. Our formulation is also related more broadly to cooperative games in which the characteristic function is not treated as exogenously fixed. Games with stochastic characteristic functions defined by private types provide one such connection (Zhao et al., 2020). In our setting, the source of endogeneity is different: the platform fixes the evaluation procedure, while a provider's physical action induces a joint change across multiple coalition utilities. Commitment partitions then determine which coalition evaluations must share the same action. This action-reuse constraint is what links the adaptive and fixed submission regimes studied in this paper.

B PROOFS AND FURTHER MECHANISM PROPERTIES

B.1 COMMITMENT GAP AND ADAPTIVE ROBUSTNESS

If Π' refines Π, any action assignment that is constant on a coarse block remains feasible after that block is subdivided: copy the same action to each child block. Hence

$$\mathcal{D}_i(\Pi) \subseteq \mathcal{D}_i(\Pi'), \qquad \Gamma_i(\Pi;\beta_i) \le \Gamma_i(\Pi';\beta_i),$$

which proves the refinement ordering stated in Section 3.1.

Under the blockwise model of Section 2 and finite Θ_i, the objective separates across blocks:

$$\Gamma_i(\Pi;\beta_i) = \sum_{Q\in\Pi} \max_{\theta\in\Theta_i} \sum_{S\in Q} q_S(\theta).$$

Under the singleton partition, each coalition forms its own block, so

$$\Gamma_i(\Pi_{\mathrm{ad}};\beta_i) = \sum_{S\subseteq\mathcal{E}} \max_{\theta\in\Theta_i} q_S(\theta) = \sum_{S\subseteq\mathcal{E}} q^\star_S.$$

Subtracting the first display from the second gives the first identity in Theorem 1. For each block,

$$\sum_{S\in Q} q^\star_S - \max_{\theta\in\Theta_i} \sum_{S\in Q} q_S(\theta) = \min_{\theta\in\Theta_i} \sum_{S\in Q} \big[q^\star_S - q_S(\theta)\big].$$

Every regret in brackets is nonnegative. Because the finite action set attains the minimum, a block gap is zero if and only if one action makes every regret zero, equivalently if and only if it belongs to $\bigcap_{S\in Q} \Theta^\star_{i,S}$. Summing proves the zero-gap characterization. Refinement gives $\Gamma_i(\Pi;\beta_i) \le \Gamma_i(\Pi';\beta_i) \le \Gamma_i(\Pi_{\mathrm{ad}};\beta_i)$, and subtracting from the adaptive gain gives $0 \le \mathrm{Gap}_i(\Pi') \le \mathrm{Gap}_i(\Pi)$. □

The honest action is available to every block, so every maximal manipulation gain is nonnegative. The finite-action assumption can be replaced by compactness and continuity when maxima are attained. Cross-block resource constraints are not part of the partition model in Section 2; adding them can destroy the blockwise decomposition and requires a separate feasible assignment set.

For the adaptive sign statement, apply equation 3 to an isolated feasible change $+\epsilon e_S$ and its reverse $-\epsilon e_S$. The two directions force $\beta_i(S) = \beta_{-i}(S) = 0$. If only $-\epsilon e_S$ is feasible, the same inequalities require $\beta_i(S) \ge 0$ and $\beta_{-i}(S) \le 0$. These coordinate signs require isolated changes; coupled changes instead impose joint inner-product constraints.

For Theorem 4, robustness gives $\langle \beta_i, d\rangle \le 0$ for every $d \in \mathcal{D}_i(\Pi_{\mathrm{ad}})$. Central symmetry also makes $-d$ feasible, so $-\langle \beta_i, d\rangle \le 0$ and hence $\langle \beta_i, d\rangle = 0$. Linearity extends this equality to $\mathrm{span}(\mathcal{D}_i(\Pi_{\mathrm{ad}}))$. Conversely, orthogonality to this span makes every feasible payment change zero and therefore implies robustness. □

B.2 DERIVATION OF THE COALITION COEFFICIENT

Fix provider i and coalition S, and write $s = |S|$ and $k_i = |S \cap E_i|$. For an artifact $e \in S \cap E_i$, $v(S)$ occurs as the positive term in the marginal contribution with predecessor $S \setminus \{e\}$, with weight $(s-1)!(n-s)!/n!$. There are k_i such artifacts. For $e \in E_i \setminus S$, $v(S)$ occurs as the negative predecessor term, with weight $-s!(n-s-1)!/n!$. There are $m_i - k_i$ such artifacts. For $0 < s < n$, summing gives

$$\begin{aligned}\beta_i(S) &= k_i \frac{(s-1)!(n-s)!}{n!} - (m_i - k_i)\frac{s!(n-s-1)!}{n!} \\ &= \frac{(s-1)!(n-s-1)!}{n!}(nk_i - m_i s).\end{aligned}$$

The factorial multiplier is positive, so the sign is that of $nk_i - m_i s$, yielding the ratio criterion. At $S = \varnothing$, every $e \in E_i$ contributes the negative empty-predecessor coefficient $-1/n$; at $S = \mathcal{E}$, every $e \in E_i$ contributes the positive grand-coalition coefficient $1/n$. This proves the two boundary values. □

B.3 Harsanyi Representation and Proof of Proposition 3

Every finite game has the unique unanimity decomposition

$$v = \sum_{T \subseteq \mathcal{E}} h_v(T) u_T, \qquad u_T(S) = \mathbf{1}\{T \subseteq S\}.$$

For nonempty T, symmetry and efficiency assign the dividend of u_T equally to its $|T|$ members and zero to all other artifacts. Summing over E_i therefore assigns provider i the fraction $\rho_i(T) = |T \cap E_i|/|T|$. By linearity,

$$\Phi_i(v) = \sum_{\varnothing \neq T \subseteq \mathcal{E}} \rho_i(T) h_v(T).$$

If the only dividend changes are $-\delta$ at T and $+\delta$ at T', their sum, and hence the grand-coalition utility, is unchanged. The corresponding payment change is

$$\delta\big[\rho_i(T') - \rho_i(T)\big].$$

□

For Proposition 3, use the two flow-balance constraints in the definition of interaction laundering to rewrite the owner-payment change as

$$\begin{aligned}\sum_T \rho_i(T)\delta h(T) &= -\sum_{T \in L} \rho_i(T) \sum_{T' \in H} f_{TT'} + \sum_{T' \in H} \rho_i(T') \sum_{T \in L} f_{TT'} \\ &= \sum_{T \in L} \sum_{T' \in H} f_{TT'} \big[\rho_i(T') - \rho_i(T)\big].\end{aligned}$$

Each term is nonnegative because positive flow only goes to interactions with weakly larger owner density, and at least one term is positive by definition. Hence the payment change is strictly positive. □

For the minimal committed response in Section 3.3, the coalition changes imply $\delta h(r) = -\epsilon_1$, $\delta h(rc) = +\epsilon_1$, $\delta h(rb) = -(\epsilon_2 - \epsilon_1)$, and $\delta h(rcb) = +\epsilon_2 - \epsilon_1$. The $r \to rc$ transfer leaves the owner-A density unchanged, whereas the $rb \to rcb$ transfer increases it from $1/2$ to $2/3$. Hence the latter determines A's payment increase, $(\epsilon_2 - \epsilon_1)/6$.

B.4 Proof of the Singleton-Rank Approximation Bound

Let $\mathcal{A}$ be the candidate edits, $\mathcal{F}_k$ the feasible edit sets under budget k, and $J(M)$ the validation-payment increase of edit set M. From empty-state and singleton queries, Singleton-Rank constructs

$$\widehat{J}(M) = J(\varnothing) + \sum_{e \in M} [J(\{e\}) - J(\varnothing)], \qquad \epsilon_k = \max_{M \in \mathcal{F}_k} |J(M) - \widehat{J}(M)|.$$

Let $M^\star \in \arg\max_{M \in \mathcal{F}_k} J(M)$ and $\widehat{M} \in \arg\max_{M \in \mathcal{F}_k} \widehat{J}(M)$. By the definition of ϵ_k,

$$\begin{aligned}J(\widehat{M}) &\geq \widehat{J}(\widehat{M}) - \epsilon_k \\ &\geq \widehat{J}(M^\star) - \epsilon_k \\ &\geq J(M^\star) - 2\epsilon_k.\end{aligned}$$

Thus,

$$J(\widehat{M}) \geq \max_{M \in \mathcal{F}_k} J(M) - 2\epsilon_k.$$

Under our shard-conflict constraint, the additive surrogate can be maximized exactly by selecting the highest-weight feasible edits. Exhaustive tiny games in Appendix D.4 enumerate all feasible states and compute ϵ_k exactly, providing an ex post certificate for this bound. □

B.5 Best-Bundle Monotonicity and Owen Representation

For Proposition 5, first consider a provider coalition C excluding i. Every $T \subseteq E(C)$ is disjoint from E_i, so $\widetilde{v}(T) = v(T)$ and therefore

$$V_{\widetilde{v}}^{\star}(C) = V_v^{\star}(C).$$

If $i \in C$, every available bundle is either disjoint from E_i and unchanged, or intersects E_i and weakly degrades. Thus $\widetilde{v}(T) \leq v(T)$ for every $T \subseteq E(C)$, which implies

$$V_{\widetilde{v}}^{\star}(C) \leq V_v^{\star}(C).$$

Writing the provider-level Shapley value in marginal form gives

$$\Psi_i^{\mathrm{BB}}(\widetilde{v}) - \Psi_i^{\mathrm{BB}}(v) = \sum_{C \subseteq \mathcal{P} \setminus \{i\}} w(C) \left[V_{\widetilde{v}}^{\star}(C \cup \{i\}) - V_v^{\star}(C \cup \{i\}) \right] \leq 0,$$

because the predecessor terms for C are unchanged and every Shapley weight $w(C)$ is positive. This proves the proposition. □

The external-invariance premise is necessary. In a two-provider game with $V(A) = 0$, $V(B) = 1$, and $V(AB) = 1$, provider A receives zero. Lowering the external baseline $V(B)$ to zero while fixing $V(AB)$ raises A's Shapley payment to $1/2$.

For completeness, let $\mathcal{B} = \{E_i : i \in \mathcal{P}\}$ be the verified owner partition. Coalitional efficiency of the Owen value states that the sum of the Owen payoffs inside a block equals that block's Shapley value in the quotient game. Applied to v^+, the quotient game is $C \mapsto v^+(E(C)) = V_v^{\star}(C)$. Hence

$$\sum_{e \in E_i} \omega_e(v^+, \mathcal{B}) = \mathrm{Shapley}_i(V_v^{\star}) = \Psi_i^{\mathrm{BB}}(v),$$

which proves the Owen representation used in Section 4.2.

B.6 Proof of Proposition 6

For a provider coalition C excluding i, every available bundle contains only external artifacts, so the two provider games agree by assumption. For $C \ni i$, write any available bundle as $X \cup T$, where $X \subseteq E(C \setminus \{i\})$ and $T \subseteq E_i$. Then

$$V_v^{\star}(C) = \max_{X \subseteq E(C \setminus \{i\})} \max_{T \subseteq E_i} v(X \cup T).$$

The assumed inner equality holds for every external context X, so the outer maxima also agree after replacing E_i by E_i'. Thus the two Best-Bundle provider games are identical for every $C \subseteq \mathcal{P}$, and their Shapley payments are identical. □

Matching only singleton and grand utilities would not suffice. With three providers, fix all external-only coalitions, $W(A) = 0$, and $W(ABC) = 1$; lowering $W(AB)$ from one to zero, with other unspecified values zero, changes A's Shapley payment from $1/2$ to $1/3$.

B.7 Committed candidate ranges

For each provider coalition C, fix a nonempty candidate family $\mathcal{R}(C) \subseteq 2^{E(C)}$ independently of the evaluated game, and define

$$V_v^{\mathcal{R}}(C) = \max_{T \in \mathcal{R}(C)} v(T).$$

Under the conditions of Proposition 5, $V_{\widetilde{v}}^{\mathcal{R}}(C) = V_v^{\mathcal{R}}(C)$ when $i \notin C$: every candidate bundle is external to i and therefore unchanged. When $i \in C$, every candidate bundle is unchanged or weakly degraded, so $V_{\widetilde{v}}^{\mathcal{R}}(C) \leq V_v^{\mathcal{R}}(C)$. The same provider-level Shapley argument therefore gives

$$\mathrm{Shapley}_i(V_{\widetilde{v}}^{\mathcal{R}}) \leq \mathrm{Shapley}_i(V_v^{\mathcal{R}}).$$

A random range fixed independently of the utilities and shared between the two games preserves the guarantee conditionally and in expectation. By contrast, utility-dependent search may generate different candidate ranges; such paired ranges do not inherit this result automatically and are audited empirically.

B.8 Optional artifact-level allocation

The mechanism first fixes the safe owner total $\Psi_i^{\mathrm{BB}}(v)$. If artifact-level credits are required, first define

$$V^\star(C;T) = \max_{\substack{U_i \subseteq T \\ U_j \subseteq E_j,\ j \in C}} v\left(U_i \cup \bigcup_{j \in C} U_j\right)$$

for $C \subseteq \mathcal{P} \setminus \{i\}$. Then, for $T \subseteq E_i$, let

$$G_i(T) = \sum_{C \subseteq \mathcal{P} \setminus \{i\}} w(C)\left[V^\star(C;T) - V^\star(C)\right],$$

where $w(C) = |C|!(|\mathcal{P}| - |C| - 1)!/|\mathcal{P}|!$ is the provider-level Shapley coefficient, and $V^\star(C;T)$ permits provider i to use only artifacts in T and lets other providers freely select their sub-bundles. Then $G_i(\varnothing) = 0$ and $G_i(E_i) = \Psi_i^{\mathrm{BB}}(v)$. Applying Shapley within E_i allocates this fixed owner payment. Proposition 5 protects the owner total under its stated conditions, while individual credits may vary.

C Experimental protocol and reproducibility

C.1 Commitment-continuum protocol

We derive a six-artifact game from each entity-matching benchmark. Provider A owns raw source/schema artifact r and exact inverse metadata c. Four external field artifacts $b_1, \ldots, b_4$ expose the three semantic text-field comparison blocks and the scalar year/price block. A coalition has nonzero utility only if it contains r and at least one field artifact; including c mechanically restores honest feature bytes. The 15 manipulable contexts are therefore $\{r\} \cup B$ for every nonempty $B \subseteq \{b_1, \ldots, b_4\}$, without c. For A's Atomic payment, their coefficients for $|B| = 1, 2, 3, 4$ are respectively $1/60, 0, -1/60, -2/15$. The remaining 49 coalition contexts have zero response under every action and can be placed in any partition block without changing the objective. Thus, the reported block counts refer to the 15-dimensional active subspace of the full 2^6-coalition partition model.

Every task uses the same equal-cost finite action library: the honest schema and the five nonidentity permutations of the three text fields, each applied to predeclared shard 0. These action definitions remain fixed across contexts. The four nested partitions are: one block for all contexts (committed); two size bins $|B| \leq 2$ and $|B| \geq 3$; four exact-size blocks; and one block per coalition. For each partition, we enumerate all six actions and optimize each block independently according to the blockwise objective in Section 3.1. The complete context-to-action policy is then serialized and SHA-256 hashed before any test feature is read. Held-out evaluation uses only the frozen policy. Feasible-set nesting under partition refinement guarantees monotonicity of the validation optima, while the held-out curve measures transfer and need not preserve the same ordering.

We use the same 20 predeclared shard partitions per dataset as in the main attack. Each seed task materializes $6 \times 15 = 90$ validation action responses, checks the direct Shapley payment change against $\beta^\top \Delta$, verifies grand/repaired feature identity, and records the first test-access event after policy freeze. Confidence intervals resample the 20 shard-partition seeds and are conditional on the fixed official benchmark split. As a post-processing audit, we compute every block term in Theorem 1 from the stored 6×15 validation response matrix and intersect context-wise argmax sets at tolerance 10^{-12}. This uses no additional model queries or test observations.

C.2 Adaptive horizontal-block replay

We additionally isolate the coalition-conditioned endpoint on Breast Cancer Wisconsin and Digits. Each of ten predeclared seeds defines a stratified 60/20/20 train/validation/test split and a stratified six-way partition of the training rows. Provider A owns blocks a_1, a_2; the remaining four blocks are treated as the other-owner portfolio. Classification accuracy is coalition utility, with empty utility fixed at zero. Every non-boundary coalition can select from the same three actions: use its declared blocks honestly; add both A blocks whenever A participates; or reflect every feature of its

participating A rows around the full-training range midpoint. Coalitions without A and the grand coalition are forced to be honest.

For each of the 2^6 coalitions, validation exactly maximizes $\beta_A(S)\Delta(S)$ over the three actions. The full 64-action policy and complete validation response table are hashed and durably written before test evaluation is unlocked. Test then evaluates only the honest game and frozen policy. The 20 tasks materialize 1,280 policy rows and 3,840 complete validation action responses. Audits check pointwise optimality, direct Shapley–$\beta^\top\Delta$ equality, empty/external/grand identity, and the freeze-before-test event ordering. The experiment uses a centralized logistic solver and an in-process freeze/test ledger to measure reporting geometry and access order. End-to-end FL evaluation additionally includes communication, secure aggregation, convergence, and cryptographic isolation. Models use standardized logistic regression with $C = 1$, the `lbfgs` solver, and a 2,000-iteration limit.

C.3 Entity-matching data

We use the structured versions of two public DeepMatcher benchmarks. The local copies are verified against a download manifest before use.

Table 1: Entity-matching benchmark sizes. Positive counts are in parentheses. Official candidate-pair splits are pair-disjoint and may share entities across splits.

Dataset	A records	B records	Train pairs	Validation	Test
DBLP–ACM	2,616	2,294	7,417 (1,332)	2,473 (444)	2,473 (444)
Walmart–Amazon	2,554	22,074	6,144 (576)	2,049 (193)	2,049 (193)

For each candidate pair and each aligned text field, we compute token Jaccard, character-trigram Jaccard, sequence similarity, normalized length similarity, and two missingness indicators. DBLP–ACM additionally uses year equality and scaled difference; Walmart–Amazon uses price equality and relative difference. A standardized logistic regression with $C = 1$, the `lbfgs` solver, and a 2,000-iteration limit is trained on the official training pairs. The utility is $v = \ell_{\text{prior}} - \ell_{\text{model}}$, where ℓ is log loss and the null predictor uses the training positive rate. Log-loss improvement determines payment; average precision, F1 at threshold 0.5, and recall provide model-quality diagnostics. Non-logistic models appear in the separate controlled robustness matrix.

The three artifact players are A's table and advertised shard schema r, A's inverse schema metadata c, and B's table b. Pair labels and split membership belong to the platform. The repair payload has a closed type containing only a shard integer, a permutation enum, and that enum's exact inverse; mechanical checks reject extra fields, record values, record IDs, and pair labels. Coalition utilities are zero unless both source tables are present. The repaired grand coalition is checked for byte-identical train, validation, and test features relative to the honest game.

C.4 Attack search and held-out discipline

Fixed-submission search. Stable SHA-256 hashing assigns A's records to 20 shards. The partition seed is included in the salt, giving 20 predeclared shard assignments for each fixed benchmark split. Five nonidentity permutations of the three text fields are available for each shard, yielding 100 candidate singleton edits. For budget $k \in \{1, 2, 4, 8, 12\}$, at most one edit may be selected per shard. Let $\mathcal{F}_k$ denote the feasible edit sets and let $J(M)$ be A's validation payment increase for state M.

Singleton-Rank evaluates the empty state and all feasible singletons and constructs the additive surrogate

$$\widehat{J}(M) = J(\varnothing) + \sum_{e \in M} \big[J(\{e\}) - J(\varnothing)\big].$$

It then maximizes $\widehat{J}(M)$ over $M \in \mathcal{F}_k$ and, for $k > 1$, evaluates the selected combined state once. If $\epsilon_k = \max_{M \in \mathcal{F}_k} |J(M) - \widehat{J}(M)|$, the returned state $\widehat{M}$ satisfies $J(\widehat{M}) \geq \max_{M \in \mathcal{F}_k} J(M) - 2\epsilon_k$. The proof is given in Appendix B.4.

Query-matched baseline. Query-matched random uses exactly the same validation-query budget and selected edit cardinality as Singleton-Rank. At $k = 1$, both methods use 101 unique queries:

the empty state and all 100 singleton states. For $k > 1$, both use 102 queries. Singleton-Rank uses the empty state, 100 singletons, and one selected combined state, whereas query-matched random uses the empty state and 101 distinct feasible full-budget states sampled without replacement and retains the validation maximum. Every selected state contains exactly k edits, so query count and action cardinality match in every dataset, seed, and budget cell. Each budget is treated as a separate attack instance, and queries are not shared across budgets. A one-draw unguided random baseline is reported only in the supplementary results.

At this feedback budget, a coordinate method that first evaluates all 100 single edits has only one combined-state query remaining and therefore reduces to Singleton-Rank. Additional coordinate rounds would require a larger query budget.

Held-out discipline. All attack selection uses validation feedback only. After the selected schema state is frozen, the model is refit on the training pairs and evaluated on the held-out test split. The dataset loader verifies the test-file checksum at startup, and a fail-closed ledger records that policy freeze occurs before the first test access. Budgets, candidate edits, and stopping rules are fixed in advance.

C.5 Pretrained representation and alignment interface

CIFAR-10 images (Krizhevsky, 2009) pass through a frozen ImageNet-1K V1 ResNet-18 (He et al., 2016) with its official resize-256, center-crop-224, normalization, and pretrained weight hash. We sample 300 training-pool and 100 test images per class. Within each of 20 split seeds, 50 training-pool images per class form validation and the remaining 250 split equally into disjoint A/B sets. A alone fits a 64-dimensional PCA bottleneck and its logistic probe; B fits a second fusion probe in A's honest coordinates. Rows are ℓ_2 normalized, and probes use $C = 10$, `lbfgs`, tolerance 10^{-10}, and 1,000 iterations. Utility is cross-entropy improvement over uniform logits.

Artifact r is the frozen encoder/bottleneck and A branch, b is B's fixed consumer branch, and c is a selectable interface alignment. A's committed action publishes zQ for one of 16 dense QR-orthogonal bases (plus identity), counter-rotates its own probe, and defines $c = Q^\top$. Validation Atomic payment selects Q; its matrix and hash are frozen before test labels enter any scoring computation. Without c, B consumes the wrong coordinates; with c, the original composite is recovered algebraically. A fresh logistic probe is also refit on zQ from the same training labels. Its held-out probabilities/accuracy are compared to a fresh probe on z, while Gram matrices, basis orthogonality, every coalition utility, the $\Delta\Phi_A = [v^0(rb) - v^Q(rb)]/6$ identity, and owner-level defenses are audited. CIFAR batch files, pretrained weights, code, candidate tables, frozen bases, and result tables are covered by SHA-256 manifests; raw CIFAR images are not redistributed.

C.6 Valuation mechanisms

All eight artifact coalitions are materialized for each three-player real game, and attack actions are frozen before held-out evaluation. Atomic values are exact. Exact Best-Bundle, as defined in Section 4.2, maximizes over the utility game being valued and is covered by Proposition 5. To avoid held-out selection in the schema study, Atomic Free Disposal and the reported Best-Bundle policy instead select bundles on validation and evaluate those frozen choices on the held-out game; their held-out zero-gain result is therefore empirical. The representation diagnostic instead evaluates each exact mechanism on the held-out utility game after Q is frozen; no held-out outcome enters the attack selection. Merged/quotient and Truth-Shapley/Owen have identical owner totals (the latter may allocate that total differently within a portfolio). We report Union Shapley as a group score; owner-payment comparisons use budget-balanced mechanisms. The committed-range mechanism generates a utility-independent set of two nested bundle paths per provider coalition, reuses the exact same fingerprint in honest and attacked games, and audits its payment-change identity and selection regret. Because this fixed range is shared across the two games, the restricted-range result in Appendix B.7 applies formally.

C.7 CONTROLLED SUITE

The final-code controlled matrix contains 2,260 seeded study tasks plus three boundary-control jobs (2,263 unique runner hashes):

Study	Runner jobs	Fixed design
Mismatch type/strength	720	3 datasets × 3 modes × 4 fractions × 20 seeds
Partial repair	300	3 datasets × 5 repair fractions × 20 seeds
Model robustness	450	3 datasets × 5 models × 3 fractions × 10 seeds
Harmful artifact	120	3 datasets × 4 strengths × 10 seeds
Third-party repair	180	3 datasets × 3 fractions × 20 seeds
Coupled low-rank interface	400	4 ranks × 5 strengths × 20 seeds
Scalability	90	9 sizes × 2 game types × 5 seeds
Boundary controls	3 groups	refinement, Sybil, additive null providers

Controlled tabular games. We use a balanced 1,500-example binary synthetic problem with 16 features (eight informative, two redundant, one cluster per class, class separation 1.5, and 1% label noise), Breast Cancer Wisconsin (569 examples, 30 features, two classes), and Digits (1,797 examples, 64 features, ten classes). Each seed defines a stratified 60/20/20 train/validation/test split. ANOVA F-statistics computed on training data rank features; the top half form A's r and the rest form B's b. Every classifier is a StandardScaler pipeline, and every coalition is retrained. Utility is log-loss improvement over the training-class-frequency predictor. Coalitions containing neither r nor b have zero utility.

Mismatch and model grids. A mismatch changes r's advertised cross-source alignment while leaving its standalone data fixed. Targeted mismatch pairs different-class examples; random mismatch pairs records without labels; the configuration named *natural* is a structured, label-blind proxy that orders records by A's strongest feature and pairs neighbors. Artifact c restores a seeded fraction $q \in \{0, .25, .5, .75, 1\}$ of corrupted pairs; at $q = 1$, the grand training matrix is exactly honest. The main mismatch grid uses logistic regression, seeds 0–19, the three modes, and fractions $\{.1, .25, .5, .75\}$. The partial-repair grid fixes targeted mismatch at .75. The model grid uses seeds 0–9, targeted fractions $\{.25, .5, .75\}$, and logistic regression ($C = 1$, `lbfgs`, 1,000 iterations), calibrated LinearSVC ($C = 1$, three folds), RBF SVM ($C = 1$, `gamma=scale`), 50-tree gradient boosting, and a one-hidden-layer 32-unit MLP with early stopping, 250 iterations, and patience 15.

Harmful and independent repair artifacts. The harmful-overlay grid applies no join mismatch. A seeded half of training rows receives shifted labels $(y + 1) \bmod K$ and is appended with strength $\{.5, 1, 2, 3\}$; deterministic repetition supplies strengths above one. Artifact h is owned by A but supplies no standalone feature space. In the third-party grid, provider C owns c; seeds 0–19 and targeted fractions $\{.25, .5, .75\}$ give 180 cells. Repair costs $\{0, .001, .005, .01, .025, .05, .1, .2, .4\}$ are applied after valuation and never change utility or retraining.

Coupled interface and scalability. The low-rank control uses balanced three-class Gaussian mixtures in 12 dimensions with 192/96/96 independently generated train/validation/test examples, noise 0.9, ranks $\{1, 2, 4, 6\}$, strengths $\{0, .5, 1, 1.5, 2\}$, and seeds 0–19. Full-batch float64 up-projections are fit for 50 Adam steps at learning rate .08. At zero strength $Q = I$; otherwise $Q = -\alpha I$, r publishes $zQ^\top$, its up-map is changed by $Q^{-\top}$, and c applies Q^{-1} before b consumes the code. These transformations define the coupled interface evaluated here.

The scalability grid pairs artifact and provider counts, respectively, as

$$n \in \{4, 6, 8, 10, 12, 14, 16, 18, 20\},$$
$$p \in \{2, 3, 4, 5, 6, 7, 8, 8, 8\}.$$

We cross seeds 0–4 with random and structured Harsanyi games. Random games draw every nonempty coalition utility from $\mathrm{Unif}[-.25, 1]$. Structured games draw singleton dividends from $\mathrm{Unif}[.04, .28]$, retain pair interactions with probability .25 (Gaussian standard deviation .08, mean .025 within owner and $-.005$ across owners), and retain triples with probability .05 (Gaussian mean zero, standard deviation .035), before the zeta transform. We then reduce every P0-containing

coalition by $.1(.5 + U)$ for $U \sim \mathrm{Unif}[0, 1]$. Exact subset-max dynamic programming, strict forward greedy, beam width four, and a utility-independent committed range with two nested paths are evaluated on pre-materialized honest and degraded games. Paired call counts sum both games; selector timings exclude game generation and are per-task measurements under concurrent workers. Boundary studies use refinement counts $\{1, 2, 4, 8\}$, one Sybil control, and additive-provider counts $\{0, 1, 2, 4, 6\}$.

C.8 Statistics, software, and integrity gates

Task matrices, seeds, and attack budgets are fixed by JSON configurations. Within a dataset/attack, headline means are over seed–budget cells. Confidence intervals use a deterministic nonparametric bootstrap that samples partition seeds as clusters and retains all five budgets within the relevant policy; we use 10,000 replicates and percentile 95% intervals. Paired comparisons first subtract policies within seed and budget and then resample seeds while retaining all five paired differences. These intervals are conditional on the fixed benchmark split. The predeclared seed set remains fixed across analyses.

Experiments use Python 3.13.2, NumPy 2.2.4, and scikit-learn 1.6.1 with one BLAS thread per worker; the coupled control uses CPU PyTorch 2.8.0. Entity, commitment, and controlled suites use up to eight process workers; adaptive replay uses two, on Linux 5.15/glibc 2.35. Checkpoints are appended atomically per task. The final controlled run contains 2,263/2,263 unique task hashes, 13,650 long-form rows, no nonfinite cells, and passes all automated result audits. Entity runs also check attack-budget feasibility, the analytic $\beta = -1/6$ payment identity, raw source/label fingerprint equality, clean repair metadata, grand-feature identity, test-after-freeze access, and zero schema-control gain.

D Additional results

D.1 Commitment continuum

This section reports the complete numerical results underlying the commitment-continuum experiment in Section 5. Table 2 gives the validation and frozen-test payment gains for all four commitment partitions, while Table 3 reports the corresponding commitment-gap diagnostic on validation.

Table 2: Exact means and shard-partition bootstrap intervals for the left panels of Figure 1. Validation selects each policy; test evaluates the frozen validation-selected context–action map.

Dataset	Partition	Validation [95% CI]	Test [95% CI]
DBLP–ACM	Committed	0.00566 [0.00541,0.00590]	0.00521 [0.00488,0.00556]
DBLP–ACM	Two size bins	0.00752 [0.00714,0.00790]	0.00726 [0.00676,0.00779]
DBLP–ACM	Exact size	0.00752 [0.00713,0.00791]	0.00726 [0.00677,0.00777]
DBLP–ACM	Coalition-adaptive	0.00759 [0.00719,0.00798]	0.00730 [0.00679,0.00782]
Walmart–Amazon	Committed	0.00115 [0.00089,0.00143]	0.00102 [0.00079,0.00126]
Walmart–Amazon	Two size bins	0.00130 [0.00100,0.00161]	0.00118 [0.00092,0.00145]
Walmart–Amazon	Exact size	0.00130 [0.00101,0.00162]	0.00117 [0.00092,0.00145]
Walmart–Amazon	Coalition-adaptive	0.00133 [0.00103,0.00164]	0.00114 [0.00090,0.00142]

The full results confirm the two effects highlighted in the main text. First, positive gain remains under full commitment on both datasets. Second, refinement weakly improves the validation optimum, and the committed gap is positive whenever the grouped contexts fail to share a common maximizing action. The held-out values evaluate the frozen validation-selected policies and therefore need not preserve the same monotone ordering.

D.2 Adaptive horizontal-block replay

All 20 held-out gains are positive. Mean test gain across both tasks is 0.04109, while the other-owner total changes by -0.04109. All 334 selected non-honest coalition actions have positive realized $\beta_A(S)\Delta(S)$ on validation; the maximum direct Shapley–coefficient residual is 1.87×10^{-16} and every protected boundary gap is exactly zero.

Table 3: Exact commitment-gap diagnostic on validation. The gap is the coalition-adaptive optimum minus the indicated partition optimum. A task is binding when at least one block has no common context-wise maximizing action (argmax tolerance 10^{-12}).

Dataset	Contexts	Mean gap to adaptive	Binding tasks
DBLP–ACM	1	0.00193	20/20
DBLP–ACM	2	0.00007	20/20
DBLP–ACM	4	0.00007	20/20
DBLP–ACM	15	0	0/20
Walmart–Amazon	1	0.00017	20/20
Walmart–Amazon	2	0.00003	18/20
Walmart–Amazon	4	0.00003	18/20
Walmart–Amazon	15	0	0/20

Table 4: Atomic owner-A payment gain in the coalition-adaptive replay. Intervals bootstrap the ten predeclared split/block seeds of each fixed dataset. Validation selects the coalition policy and test evaluates the frozen map.

Dataset	Validation [95% CI]	Test [95% CI]	Positive test
Breast Cancer	0.03333 [0.02691,0.03987]	0.03197 [0.02728,0.03741]	10/10
Digits	0.04774 [0.04274,0.05353]	0.05020 [0.04643,0.05411]	10/10

D.3 Real entity matching

The payment-directed edits reduce held-out quality of the unrepaired rb matcher. Mean AP/F1 changes from $0.996 \to 0.931/0.978 \to 0.851$ on DBLP–ACM and from $0.568 \to 0.449/0.427 \to 0.279$ on Walmart–Amazon (F1 threshold 0.5). Payment and attack search optimize log-loss utility; AP and F1 diagnose the resulting matcher quality. The repaired grand features remain byte-identical to honest features.

Table 5: Cluster-bootstrap intervals for atomic payment gain. Each cluster is one shard-partition seed and contains all five attack budgets. Difference is paired Singleton-Rank minus query-matched random.

Dataset	Attack	Split	Mean [95% CI]
DBLP–ACM	Query-matched random	Validation	0.01580 [0.01565,0.01597]
DBLP–ACM	Query-matched random	Test	0.01318 [0.01293,0.01339]
DBLP–ACM	Singleton-Rank	Validation	0.01687 [0.01665,0.01710]
DBLP–ACM	Singleton-Rank	Test	0.01351 [0.01326,0.01374]
DBLP–ACM	Singleton minus QMR	Validation	0.00106 [0.00087,0.00125]
DBLP–ACM	Singleton minus QMR	Test	0.00033 [0.00012,0.00053]
Walmart–Amazon	Query-matched random	Validation	0.00566 [0.00550,0.00583]
Walmart–Amazon	Query-matched random	Test	0.00459 [0.00440,0.00479]
Walmart–Amazon	Singleton-Rank	Validation	0.00680 [0.00661,0.00699]
Walmart–Amazon	Singleton-Rank	Test	0.00480 [0.00455,0.00507]
Walmart–Amazon	Singleton minus QMR	Validation	0.00113 [0.00099,0.00128]
Walmart–Amazon	Singleton minus QMR	Test	0.00021 [0.00001,0.00041]

D.4 Tiny exhaustive attack certificate

On DBLP–ACM, for seeds 0–9, four shards, and two candidate permutations per shard, the complete feasible state counts through budgets 1–4 are 9, 33, 65, and 81. We persist the true J and singleton-additive $\widehat{J}$ for every state. Table 8 reports Singleton-Rank's observed gap and the computed $2\epsilon_k$ certificate. Here ϵ_k is obtained ex post by exhaustive enumeration. The exhaustive enumeration certifies these tiny runs; full-scale tables report observed Singleton-Rank performance. *Exact/cert.* gives exact optima and certified cells out of ten.

Table 6: Mechanism comparison on Singleton-Rank cells. Entries are mean held-out owner-A payment changes; maximum absolute grand-utility gaps are zero.

Mechanism	DBLP–ACM	Walmart–Amazon
Atomic	0.01351	0.00480
Atomic Free Disposal	0.01351	0.00480
Merged (quotient)	0	0
Truth-Shapley / Owen total	0	0
Best-Bundle	0	0
Committed range	0	0

Table 7: Pretrained representation/basis-alignment experiment over 20 split seeds. Validation selects one fixed basis before held-out evaluation.

Quantity	Held-out result
Setting	CIFAR-10; ImageNet-pretrained ResNet-18; learned 64-D PCA
Atomic owner-A gain	0.1113 (95% CI [0.1071, 0.1152])
$r+b$ accuracy	82.5% $\rightarrow$ 58.2%
Fresh-probe accuracy gap	0.000
Grand-coalition utility gap	0
Best-Bundle owner-A gain	0

Table 8: Tiny exhaustive post-hoc certificate. *Relative* normalizes the observed gap by the exact optimum; *Mean bound* is the additive certificate.

k	States	Mean/max gap	Relative	Mean bound	Exact/cert.
1	9	0/0	0%	0	10/10; 10/10
2	33	0.000732/0.001805	2.95%	0.017153	2/10; 10/10
3	65	0.000873/0.001659	2.86%	0.035807	0/10; 10/10
4	81	0.000720/0.001191	2.13%	0.056495	0/10; 10/10

D.5 Controlled mechanism studies

Table 9: Controlled studies under the final implementation. Positive gain means the manipulated provider is paid more.

Study	Atomic	Merged	Best-Bundle
Full-repair mismatch: mean val. gain	0.04032	0	0.00010
Validation pointwise-valid: positive BB gain	–	–	0/706
Harmful overlay: excluded	–	0/120	120/120
Third-party repairer's mean gross payment	0.1661	0.1661	0.0114
Portfolio refinement, 1 to 8 fragments	$0.3750 \rightarrow 0.8611$	–	$0.3750 \rightarrow 0.3750$

Pointwise nonimprovement holds in 706 of 720 full-repair cells, and Atomic gains in 706/720 cells. Best-Bundle has 0/706 positive changes within the covered set; its 12/720 positive changes all occur among the remaining cells. Under targeted mismatch at fractions $0.10, 0.25, 0.50, 0.75$, mean Atomic validation gains are respectively $0.0121, 0.0193, 0.0317, 0.1449$ on Breast Cancer; $0.0196, 0.0335, 0.0580, 0.1026$ on Digits; and $0.0171, 0.0464, 0.1055, 0.2059$ on the synthetic dataset. Figure 2 plots these targeted cells by mechanism; Merged and Best-Bundle are exactly zero throughout. The 720-cell summary pools all three mismatch modes, whereas the figure isolates the 240 targeted-mismatch cells. Thus, a nonzero Best-Bundle mean over the full matrix is compatible with the zero targeted curve.

With partial repair, Atomic gain is often negative; full repair restores grand identity. Exact Best-Bundle gain is nonpositive in all 300 validation pointwise-degradation cells. For a repair artifact owned by an independent third party, its gross payment is positive in 180/180 cells under Atomic/Merged and

168/180 under Best-Bundle; applying an external repair cost exposes the expected trade-off between suppressing self-created rents and retaining repair-supply incentives.

Across the five classifiers in the controlled robustness grid, Atomic gain is positive in 444/450 validation cells and 444/450 test cells. In the synthetic coupled-interface control, the 320 positive-strength cells have mean Atomic gain 0.7030/0.6928 on validation/test, and the unrepaired rb utility falls jointly on both splits in 317/320 cells.

D.6 Selection cost and mechanism guarantees

At 20 artifacts, exact dynamic programming reads 2,097,152 entries across the cached honest and degraded tables (2^{20} each) and takes 13.11 seconds after game materialization, with zero regret. These are not additional learner fits: exact Atomic Shapley already requires the two complete artifact-game tables in our generic exhaustive evaluator, and the subset-max transform reuses them. The count excludes the transform's $O(n2^n)$ comparison operations. Greedy uses 4,180 accesses and 0.16 seconds with mean regret 0.02845; beam-4 uses 27,074 calls and 0.59 seconds with mean regret 0.002631. The committed range uses 311,040 calls and takes 14.43 seconds in this unoptimized implementation. It evaluates fewer bundles than exact dynamic programming and takes slightly longer. Timings measure selector execution after game materialization. Relative to merged owner-level Shapley, exact Best-Bundle can nevertheless be much more expensive: merged valuation needs only 2^p full-portfolio utilities, whereas the exact subbundle envelope may need all 2^n. Across 90 games, greedy and beam show zero observed degradation violations; the committed range carries the formal monotonicity guarantee proved in Appendix B.7 under the conditions of Proposition 5.

Finally, correct owner grouping holds A's Best-Bundle payment fixed at 0.5 in the three-artifact attack. If c is registered as a new owner, the controlled A-plus-Sybil total increases by 0.01167. Best-Bundle therefore requires a verified owner partition together with false-name-resistant provenance.