Faithful Group Shapley Value

Kiljae Lee*, Ziqi Liu*, Weijing Tang, and Yuan Zhang

The Ohio State University, Carnegie Mellon University

*Equal Contribution, They are co-first authors and were listed alphabetically.

NEURAL INFORMATION PROCESSING SYSTEMS

Data Valuation via Shapley Value

THE OHIO STATE UNIVERSITY

Carnegie Mellon University

- In **data valuation**, Ghorbani and Zou (2019) used the **Shapley value** to quantify each data point's contribution to model performance.

Definition [Data Shapley]

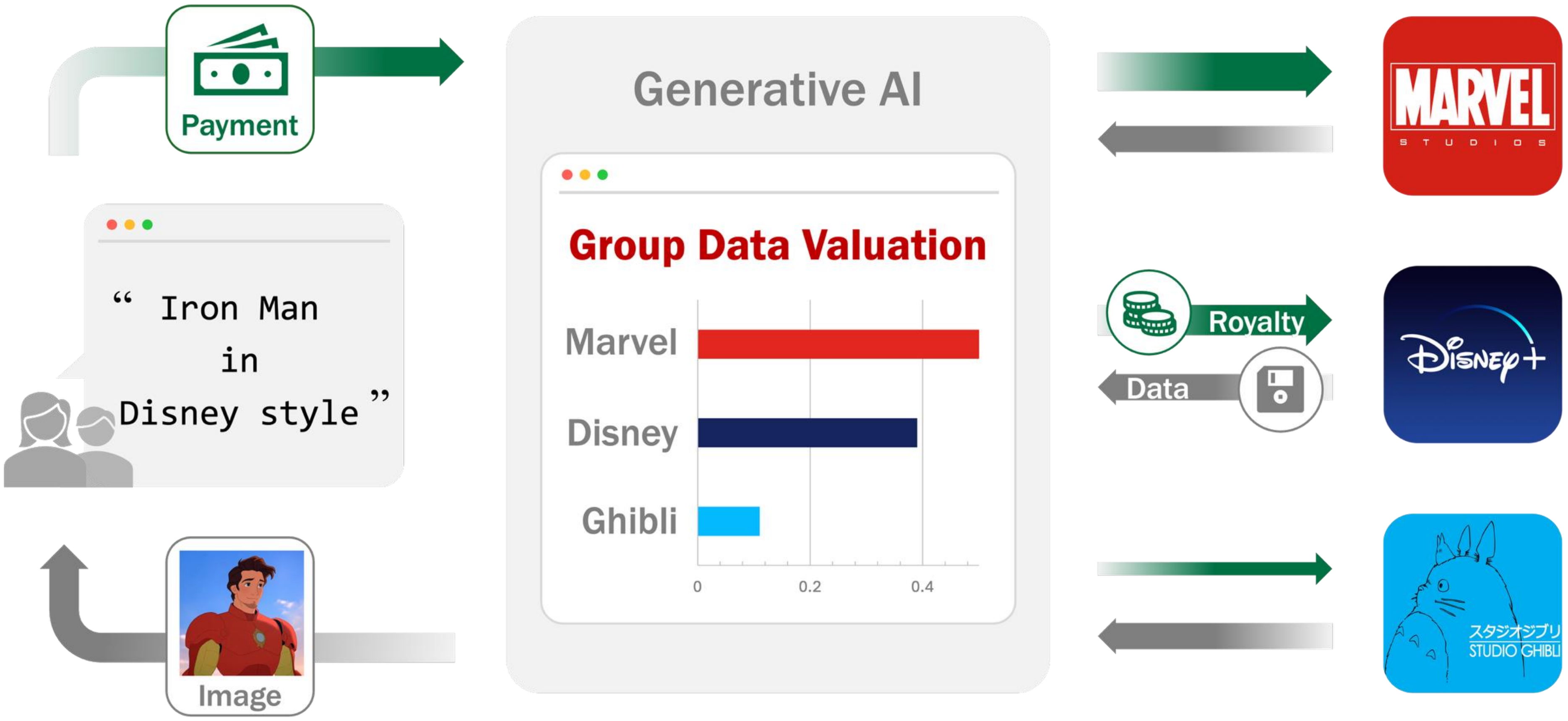
For a dataset $\mathcal{D} = \{z_1, \dots, z_n\}$ and a utility function $U: 2^{\mathcal{D}} \to \mathbb{R}$, the (individual) ***Shapley Value*** *of i-th player is defined as*

$$SV(z_i) = \sum_{S \subseteq \mathcal{D} \setminus \{z_i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} [U(S \cup \{z_i\}) - U(S)].$$

- $U(S)$ measures model performance when trained on the subset S.

Group Data Valuation

- In many applications, data naturally **form groups**.



Group Shapley Value

- The standard baseline on **group data valuation** treats each group as a single player.

Definition [Group Shapley Value (GSV)]
For a dataset $\mathcal{D} = \{z_1, \dots, z_n\}$, *its partition* $\Pi = \{S_1, \dots, S_K\}$, *and a utility function* $U: 2^{[n]} \to \mathbb{R}$, *the* ***Group Shapley Value*** *of* j*-th group is defined as*

$$GSV(S_j) = \sum_{S \subseteq \Pi \setminus \{S_j\}} \frac{|S|!\,(K - |S| - 1)!}{K!} \left[U\big(agg(S) \cup S_j\big) - U\big(agg(S)\big)\right],$$

where $agg(S) = \bigcup_{T \in S} T$.

- Note that when each $S_k \in \Pi$ is a singleton, GSV reduces to the individual Shapley value.

Key Observation: Shell Company Attack

NEURAL INFORMATION PROCESSING SYSTEMS

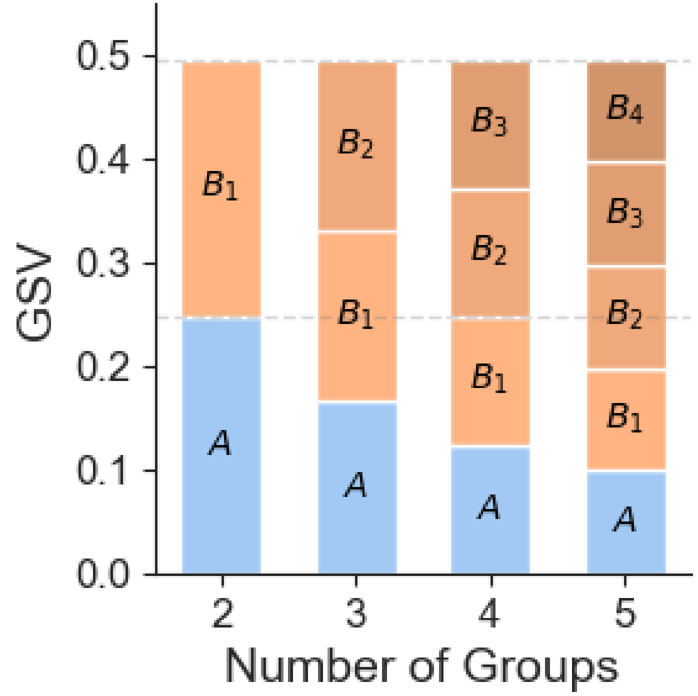
THE OHIO STATE UNIVERSITY

Carnegie Mellon University

- However, we identify that GSV can be **manipulated** by splitting a real group into many small "shell" subgroups.

Example [Shell Company Attack (informal)]
Suppose data consist of two groups: A and B. Under GSV, splitting B into smaller subgroups $B_1, B_2, \ldots$ ***inflates the total share*** *B receives:*

$$GSV(B_1) + GSV(B_2) + \cdots > GSV(B)$$

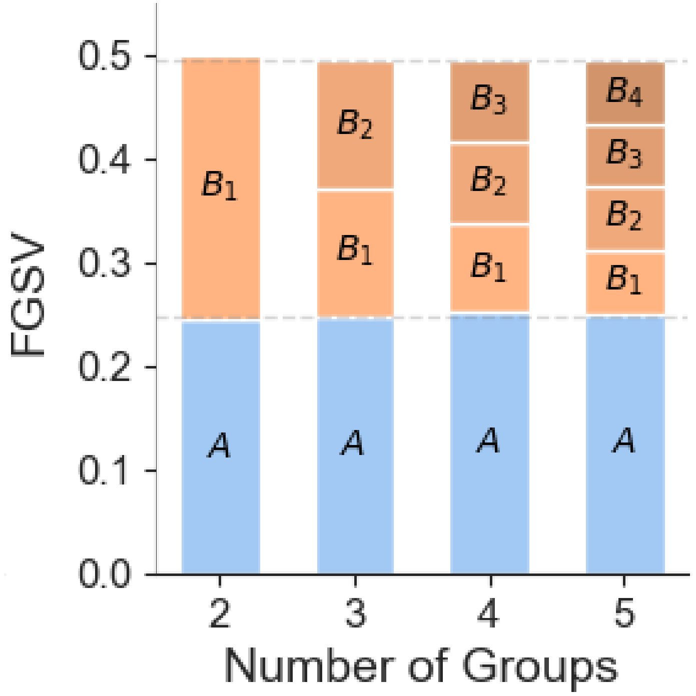


Our Method: Faithful Group Shapley Value

- We propose **Faithful Group Shapley Value (FGSV)** to defend the shell company attack.

Definition [Faithful Group Shapley Value (FGSV)]
Under the group data valuation setting, the ***Faithful Group Shapley Value*** *of j-th group is defined as the sum of individual Shapley values within the group*

$$FGSV(S_j) = \sum_{i \in S_j} SV(i)$$



Theory: Uniqueness under Axiomatization

NEURAL INFORMATION PROCESSING SYSTEMS

The Ohio State University

Carnegie Mellon University

- FGSV is **uniquely** characterized by following five **axioms:**

1. Efficiency
2. Linearity
3. Symmetry
4. Null Player

→ Classical Shapley Axioms (but "group" version)

5. **Faithfulness** → New Axiom for fending off the shell company attack

Axiom [Faithfulness]
If S is a common member of two different partitions Π_1, Π_2, of $\mathcal{D}$, then the group valuation function ν satisfies

$$\nu(S; \Pi_1) = \nu(S; \Pi_2)$$

Theorem [Uniqueness of FGSV]
The group valuation function ν ***satisfies the above five axioms*** *if and only if*

$$\nu(S_j) = FGSV(S_j)$$

Computation: Direct Estimation of FGSV

- It is possible to **directly estimates the sum** of individual Shapley values for each group.
- The key idea is FGSV's probabilistic representation:

Lemma [Alternative Representation]

Let $s_0 = |S_j|$. *Then,*

$$FGSV(S_j) = \frac{s_0}{n}[U(\mathcal{D}) - U(\emptyset)] + \sum_{s=1}^{n-1} \mathbb{E}_{s_1 \sim HG(n,s_0,s)}\left[\frac{n}{n-s}\left(\frac{s_1}{s} - \frac{s_0}{n}\right)\mu(s_1, s)\right],$$

where $\mu(s_1, s)$ *is the average utility of* S *satisfying* $s = |S|$ *and* $s_1 = |S \cap S_j|$.

- This allows us to **approximate FGSV** before estimation by using the exponential tail of hypergeometric r.v.'s.

Computation: Direct Estimation of FGSV

- Our method achieves (ϵ, δ)-approximation with much fewer utility evaluations.

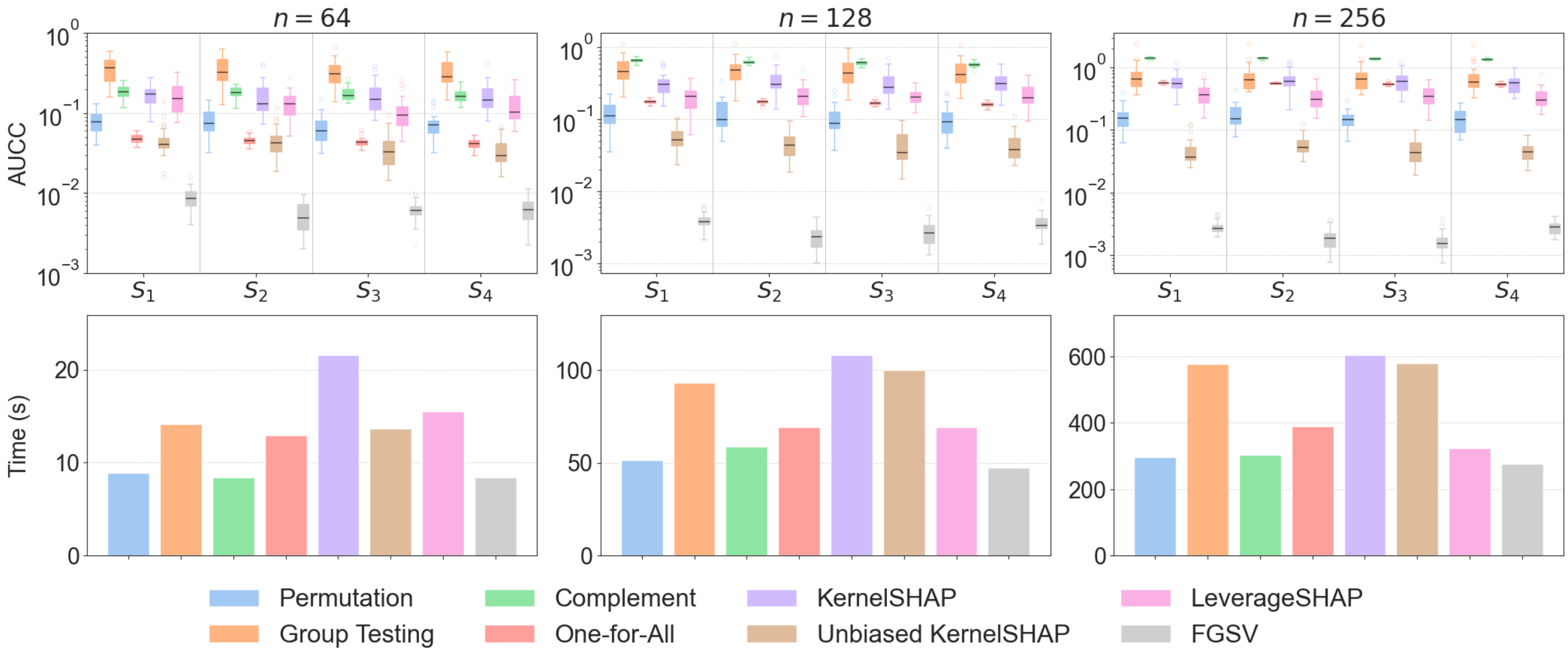
Definition [(ϵ, δ)-Approximation]

An estimator $\hat{\theta}$ for θ is called an (ϵ, δ)-approximation if

$$\mathbb{P}\left(\|\hat{\theta} - \theta\|_2 \geq \epsilon\right) \leq \delta$$

FGSV: $\mathcal{O}(n(\log n)^3)$ vs. SOTA by summing SV's: $\mathcal{O}(n^2 Poly(\log n))$

Empirical Results



Application: Copyright Sharing in GenAI