SUPPLEMENTARY MATERIAL TO JOINT SHAPLEY VALUES: A MEASURE OF JOINT FEATURE IMPORTANCE

Anonymous authors
Paper under double-blind review

In this supplementary material we discuss the arrival order interpretation, present the proofs of all results, and discuss a notion of joint symmetry obtained by removing conditions 2 and 3 from JSY.

A ARRIVAL ORDER INTERPRETATION

As discussed in the paper, the joint Shapley value can be viewed in terms of the worth brought by 'arriving' agents but, rather than arriving one at time, they can now also arrive in coalitions. To be precise, consider this procedure: at time 0, no agents have arrived; at each $t \in \{1, 2, \ldots\}$, the next set of agents to arrive is chosen uniformly from the set of non-empty subsets of size at most k of the remaining (yet to arrive) agents. Then ϕ_T^J is the expected worth brought by coalition T when it arrives (a coalition is assigned zero worth if it does not arrive at any time). To see this, denote by A_i the coalition to arrive at time i, by B_i the union of all coalitions that have arrived up to time i: $B_i = \bigcup_{j \leq i} A_j$, and by p_T the probability that at some time coalition T arrives, $p_T = \mathbb{P}(\exists\, i : B_i = T)$. We have the recursive relationship:

$$p_T = \sum_{i=1}^{n} \sum_{\substack{S \subsetneq T: \\ |S| \geq |T|-k}} \mathbb{P}(B_{i-1} = S)\mathbb{P}(A_i = T \setminus S \mid B_{i-1} = S)$$

$$= \sum_{i=1}^{n} \sum_{\substack{S \subsetneq T: \\ |S| \geq |T|-k}} \mathbb{P}(B_{i-1} = S) \left(\sum_{r=1}^{(n-|S|)\wedge k} \binom{n-|S|}{r} \right)^{-1} = \sum_{s=(|T|-k)\vee 0}^{|T|-1} \binom{t}{s} p_S \left(\sum_{r=1}^{(n-s)\wedge k} \binom{n-s}{r} \right)^{-1},$$

where in the last line S is any set of size s. Thus we see that p_T only depends on T through its cardinality, and defining

$$\hat{p}_t := p_T \left(\sum_{r=1}^{(n-t)\wedge k} \binom{n-t}{r} \right)^{-1}$$

for any T with $|T| = t$, the expected worth brought by coalition T under this procedure is

$$\sum_{i=1}^{n} \sum_{S \subseteq N \setminus T} \mathbb{P}(B_{i-1} = S,\, A_i = T)[v(S \cup T) - v(S)] = \sum_{S \subseteq N \setminus T} \hat{p}_{|S|}[v(S \cup T) - v(S)].$$

Further, we have the relationship

$$\hat{p}_t = \frac{\sum_{s=(t-k)\vee 0}^{t-1} \binom{t}{s} p_s}{\sum_{r=1}^{(n-t)\wedge k} \binom{n-t}{r}}$$

and since we know that $p_0 = 1$ (we start with no agents having arrived) and $p_n = 1$ (we finish with all agents having arrived) we also have the identities:

$$1 = \sum_{s=n-k}^{n-1} \binom{n}{s} \hat{p}_s, \quad 1 = \hat{p}_0 \sum_{r=1}^{k} \binom{n}{s}.$$

By comparing with the identities defining $(q_0, \ldots, q_{n-1})$, we deduce that $\hat{p}_t = q_t$ for all $t \in \{0, \ldots, n-1\}$, which verifies this arrival interpretation.

B PROOFS

Proof of Lemma 2. For the first statement, for each $\varnothing \neq S \subseteq N \setminus T$ consider the game

$$v_S(R) = \begin{cases} 1 & \text{if } R = S \text{ or } R = S \cup T, \\ 0 & \text{otherwise.} \end{cases}$$

Then for such S, by JNU, $\phi_T(v_S) = 0$. By this equality, Lemma 1, and the definition of v_S,

$$0 = \phi_T(v_S) = \sum_{R \subseteq N} a_R^T v_S(R) = a_S^T + a_{S \cup T}^T.$$

For the second statement, for each $\varnothing \neq H \subsetneq T$ let α_H be a constant and for $S \subseteq N \setminus T$, consider the game

$$x_S^{\boldsymbol{\alpha}}(R) = \begin{cases} \alpha_H & \text{if } R = S \cup H \text{ for some } \varnothing \neq H \subsetneq T, \\ 0 & \text{otherwise.} \end{cases}$$

By JNU, $\phi_T(x_S^{\boldsymbol{\alpha}}) = 0$ for every $S \subseteq N \setminus T$. Thus by Lemma 1

$$0 = \phi_T(x_S^{\boldsymbol{\alpha}}) = \sum_{R \subseteq N} a_R^T x_S^{\boldsymbol{\alpha}} = \sum_{\varnothing \neq H \subsetneq T} a_{S \cup H}^T x_S^{\boldsymbol{\alpha}}(S \cup H) = \sum_{\varnothing \neq H \subsetneq T} a_{S \cup H}^T \alpha_H.$$

Since this holds for every choice of constants α_H, it follows that $a_{S \cup H}^T = 0$ for all $S \subseteq N \setminus T$ and $\varnothing \neq H \subsetneq T$, as required. □

Proof of Proposition 1. By Lemma 1 there exist constants $\{a_S^T\}_{S \subseteq N,\, \varnothing \neq T \subseteq N}$ such that for every v and $\varnothing \neq T \subseteq N$,

$$\begin{aligned} \phi_T(v) &= \sum_{S \subseteq N} a_S^T v(S) = \sum_{S \subseteq N \setminus T} \left(a_S^T v(S) + \sum_{\varnothing \neq H \subsetneq T} a_{S \cup H}^T v(S \cup H) + a_{S \cup T}^T v(S \cup T) \right) \\ &= \sum_{S \subseteq N \setminus T} (a_S^T v(S) + a_{S \cup T}^T v(S \cup T)) = \sum_{S \subseteq N \setminus T} a_{S \cup T}^T [v(S \cup T) - v(S)]; \end{aligned}$$

where the last two equalities owe to Lemma 2. The proof is complete by setting $p^T(S) = a_{S \cup T}^T$. □

Proof of Proposition 2. Suppose ϕ satisfies axioms JLI, JNU and JEF. Then by Proposition 1 the constants $\{p^T(S)\}$ exist, such that for every v and $\varnothing \neq T \subseteq N$,

$$\phi_T(v) = \sum_{S \subseteq N \setminus T} p^T(S)[v(S \cup T) - v(S)].$$

Now for each $\varnothing \neq R \subseteq N$ consider the identity game

$$w_R(S) = \begin{cases} 1 & \text{if } S = R, \\ 0 & \text{otherwise.} \end{cases}$$

Then for every $\varnothing \neq T \subseteq N$ with $|T| \leq k$,

$$\phi_T(w_R) = \sum_{S \subseteq N \setminus T} p^T(S)[w_R(S \cup T) - w_R(S)].$$

Note that the term $w_R(S \cup T) - w_R(S)$ in the above sum is equal to 1 only when $S \subsetneq R$ and $T = R \setminus S$, i.e. only when $S = R \setminus T$ and $\varnothing \neq T \subseteq R$. Further note that this term is equal to -1 only when $S = R$ and $T \neq \varnothing$, i.e. when $S = R$ and $\varnothing \neq T \subseteq N \setminus R$ (as must have $S \cap T = \varnothing$). In all other cases, this term is 0. Hence we deduce from JEF that

$$\delta_N(R) = w_R(N) = \sum_{\substack{\varnothing \neq T \subseteq R: \\ |T| \leq k}} p^T(R \setminus T) - \sum_{\substack{\varnothing \neq T \subseteq N \setminus R: \\ |T| \leq k}} p^T(R).$$

Now we show the implication in the other direction. If $\phi_T(v) = \sum_{S \subseteq N \setminus T} p^T(S)[v(S \cup T) - v(S)]$ then it is immediate that JLI and JNU are satisfied. For JEF, we wish to show that for every v,

$$\sum_{\substack{\varnothing \neq T \subseteq N: \\ |T| \leq k}} \sum_{S \subseteq N \setminus T} p^T(S)[v(S \cup T) - v(S)] = v(N).$$

Note that for each $\varnothing \neq R \subseteq N$, the coefficient of $v(R)$ on the left-hand side in the above equation is

$$\sum_{\substack{\varnothing \neq T \subseteq R: \\ |T| \leq k}} p^T(R \setminus T) - \sum_{\substack{\varnothing \neq T \subseteq N \setminus R: \\ |T| \leq k}} p^T(R).$$

But by equation (4), this is equal to $\delta_N(R)$. □

Proof of Proposition 3. In light of Proposition 2 we just have to consider JAN.

Only if: Suppose ϕ satisfies JLI, JNU, JEF and JAN. First, we shall establish that

$$p^T(S) = p^T(S') \; \forall \varnothing \neq T \subseteq N, \; S, S' \subseteq N \setminus T \text{ s.t. } s = s'. \tag{8}$$

Fix such a T, S and S'. Consider again the identity game, w_S and let σ be a self-inverse permutation such that $S \mapsto S'$, $S' \mapsto S$, and $\sigma(\{i\}) = \{i\}$ for all $i \notin S \cup S'$. As $T \subseteq N \setminus (S \cup S')$ and $\sigma(T) = T$, we have by JAN

$$\phi_T(w_S) = \phi_{\sigma^{-1}(T)}(w_S) = \phi_T(\sigma w_S)$$

where

$$\sigma w_S(R) = w_S\left(\sigma^{-1}(R)\right) = \left\{ \begin{array}{ll} 1 & \text{if } R = \sigma(S) = S' \\ 0 & \text{otherwise} \end{array} \right\} = w_{S'}(R).$$

Hence we obtain $\phi_T(w_S) = \phi_T(\sigma w_S) = \phi_T(w_{S'})$. Next, from Proposition 1 we have

$$\phi_T(w_S) = \sum_{Q \subseteq N \setminus T} p^T(Q)\left[w_S(Q \cup T) - w_S(Q)\right] = -p^T(S),$$

and similarly $\phi_T(w_{S'}) = -p^T(S')$. Hence we obtain $p^T(S) = p^T(S')$, showing (8).

Using induction on s, we now establish that (5) holds. Fix T and T' of the same size. For the base case, suppose $s = s' = n - t$. For $S \subseteq N \setminus T$ and $S' \subseteq N \setminus T'$, this forces $S = N \setminus T$ and $S' = N \setminus T'$. Now consider the game

$$x_n(R) = \begin{cases} 1 & \text{if } r = n \\ 0 & \text{otherwise,} \end{cases}$$

so that $x_n(R) = 1$ if and only if $R = N$. Define a self-inverse permutation σ so that $\sigma(T) = T'$, $\sigma(T') = T$ and $\sigma(\{i\}) = \{i\}$ for all $i \notin (T \cup T')$. Then by JAN and as $\sigma x_n = x_n$,

$$\phi_T(x_n) = \phi_{\sigma^{-1}(T')}(x_n) = \phi_{T'}(\sigma x_n) = \phi_{T'}(x_n).$$

Next, from Proposition 1,

$$\phi_T(x_n) = \sum_{Q \subseteq N \setminus T} p^T(Q)\left[x_n(T \cup Q) - x_n(Q)\right] = p^T(N \setminus T),$$

and similarly $\phi_{T'}(w_n) = p^{T'}(N \setminus T')$. Hence we obtain $p^T(N \setminus T) = p^{T'}(N \setminus T')$ which establishes the base case.

We now suppose that $p^T(S) = p^{T'}(S')$ for all $s = s' \geq n - c$ where $S \subseteq N \setminus T$, $S' \subseteq N \setminus T'$ and c is a positive integer. We shall show that $p^T(S) = p^{T'}(S')$ for all $s = s' \geq n - c - 1$ where where $S \subseteq N \setminus T$ and $S' \subseteq N \setminus T'$. To this end, consider the game

$$x(R) = \begin{cases} 1 & \text{if } r \geq n - c - 1 + t, \\ 0 & \text{otherwise} \end{cases}.$$

Thus, as before, we may write

$$\phi_T(x) = \sum_{Q \subseteq N \setminus T} p^T(Q)\left[x(Q \cup T) - x(Q)\right] = \sum_{\substack{Q \subseteq N \setminus T \\ n-c-1 \leq q < n-c-1+t}} p^T(Q).$$

Similarly,

$$\phi_{T'}(x) = \sum_{\substack{Q' \subseteq N \setminus T' \\ n-c-1 \leq q' < n-c-1+t}} p^T(Q).$$

Again, by JAN, we prove that $\phi_T(x) = \phi_{T'}(x)$. Define a self-inverse permutation σ so that $\sigma(T) = T'$, $\sigma(T') = T$ and $\sigma(\{i\}) = \{i\}$ for all $i \notin (T \cup T')$. As worth in game x depends only on a coalition's cardinality, we have $\sigma x = x$. Thus, by JAN, $\phi_T(x) = \phi_{\sigma(T)}(\sigma x) = \phi_{\sigma(T)}(x) = \phi_{T'}(x)$.

However,

$$\phi_T(x) = \sum_{\substack{Q \subseteq N \setminus T \\ q = n-c-1}} p^T(Q) + \sum_{\substack{Q \subseteq N \setminus T \\ n-c-1 < q < n-c-1+t}} p^T(Q) = \sum_{\substack{Q \subseteq N \setminus T \\ q = n-c-1}} p^T(Q) + \sum_{\substack{Q' \subseteq N \setminus T' \\ n-c-1 < q' < n-c-1+t}} p^{T'}(Q');$$

and

$$\phi_{T'}(x) = \sum_{\substack{Q' \subseteq N \setminus T' \\ q' = n-c-1}} p^{T'}(Q') + \sum_{\substack{Q' \subseteq N \setminus T' \\ n-c-1 < q' < n-c-1+t}} p^{T'}(Q')$$

which gives, by the inductive hypothesis and $\phi_T(x) = \phi_{T'}(x)$,

$$\sum_{\substack{Q \subseteq N \setminus T \\ q = n-c-1}} p^T(Q) = \sum_{\substack{Q' \subseteq N \setminus T' \\ q' = n-c-1}} p^{T'}(Q').$$

But by (8), $p^T(Q) = p^T(Q')$ if $q = q'$. Thus the above equation becomes

$$\binom{n-t}{n-c-1} p^T(Q) = \binom{n-t}{n-c-1} p^{T'}(Q')$$

for any $Q \subseteq N \setminus T$ and $Q' \subseteq N \setminus T'$ with $q = q' = n-c-1$. Thus, $p^T(Q) = p^{T'}(Q')$, completing the inductive step, and the 'only if' statement.

If: Suppose (5) is satisfied. Fix a permutation σ on N and game $v \in \mathcal{G}^N$. Then for any $\varnothing \neq T \subseteq N$,

$$\begin{aligned}\phi_T(\sigma v) &= \sum_{S \subseteq N \setminus T} p^T(S)\left[\sigma v(S \cup T) - \sigma v(S)\right] = \sum_{S \subseteq N \setminus T} p^T(S)\left[v\left(\sigma^{-1}(S \cup T)\right) - v\left(\sigma^{-1}(S)\right)\right] \\ &= \sum_{S \subseteq N \setminus T} p^T(S)\left[v\left(\sigma^{-1}(S) \cup \sigma^{-1}(T)\right) - v\left(\sigma^{-1}(S)\right)\right].\end{aligned}$$

Defining the set $S' = \sigma^{-1}(S)$ allows us to rewrite the above as

$$\begin{aligned}\phi_T(\sigma v) = \cdots &= \sum_{S' \subseteq N \setminus \sigma^{-1}(T)} p^T(\sigma(S'))\left[v\left(S' \cup \sigma^{-1}(T)\right) - v(S')\right] \\ &= \sum_{S' \subseteq N \setminus \sigma^{-1}(T)} p^{\sigma^{-1}(T)}(S')\left[v\left(S' \cup \sigma^{-1}(T)\right) - v(S')\right] = \phi_{\sigma^{-1}(T)}(v),\end{aligned}$$

with the penultimate step due to condition (5). □

Proof of Proposition 4. In light of Proposition 2 we just have to consider JSY.

Only if: Suppose ϕ satisfies JLI, JNU, JEF and JSY, fix $\varnothing \neq T, T' \subseteq N$, and consider again the identity game w_R. Then for any $\varnothing \neq R \subseteq N \setminus (T \cup T')$,

- $w_R(S \cup T) = 0 = w_R(S \cup T')$ for all $S \subseteq N \setminus (T \cup T')$,

- $w_R(S \cup T) = 0 = w_R(S)$ for all $S \subseteq N \setminus T$ such that $S \cap T' \neq \varnothing$,
- $w_R(S \cup T') = 0 = w_R(S)$ for all $S \subseteq N \setminus T'$ such that $S \cap T \neq \varnothing$.

Hence by JSY $\phi_T(w_R) = \phi_{T'}(w_R)$. But $\phi_T(w_R) = p^T(R)$ and $\phi_{T'}(w_R) = p^T(R)$. This shows that $p^T(R) = p^{T'}(R)$ for all $\varnothing \neq R \subseteq N \setminus (T \cup T')$. To show that $p^T(\varnothing) = p^{T'}(\varnothing)$ we consider the game

$$w^*(S) = \begin{cases} 1 & \text{if } S \neq \varnothing, \\ 0 & \text{otherwise.} \end{cases}$$

Then

- $w^*(S \cup T) = 1 = w^*(S \cup T')$ for all $S \subseteq N \setminus (T \cup T')$,
- $w^*(S \cup T) = 1 = w^*(S)$ for all $S \subseteq N \setminus T$ such that $S \cap T' \neq \varnothing$ (since then $S \neq \varnothing$),
- $w^*(S \cup T') = 1 = w^*(S)$ for all $S \subseteq N \setminus T'$ such that $S \cap T \neq \varnothing$ (since then $S \neq \varnothing$).

It thus follows by JSY that $\phi_T(w^*) = \phi_{T'}(w^*)$. However, $\phi_T(w^*) = p^T(\varnothing)$ and $\phi_{T'}(w^*) = p^{T'}(\varnothing)$, which gives the required identity and shows that (6) holds.

If: Now we show the implication in the other direction. Suppose (6) holds and $v \in \mathcal{G}^N$ satisfies the three conditions in JSY. Then

$$\begin{aligned} \phi_T(v) &= \sum_{S \subseteq N \setminus T} p^T(S)[v(S \cup T) - v(S)] = \sum_{S \subseteq N \setminus (T \cup T')} p^T(S)[v(S \cup T) - v(S)] \\ &= \sum_{S \subseteq N \setminus (T \cup T')} p^{T'}(S)[v(S \cup T') - v(S)] = \sum_{S \subseteq N \setminus T'} p^{T'}(S)[v(S \cup T') - v(S)] = \phi_{T'}(v). \end{aligned}$$

Hence JSY is satisfied. □

Proof of Theorem 1. We have to show that there exists exactly one choice of constants $\{p^T(S)\}$ which satisfy equations (4)–(6). Notice that satisfying (5) and (6) is equivalent to satisfying

$$p^T(S) = p^{T'}(S) \, \forall \, S \subseteq N \setminus T, \, S' \subseteq N \setminus T' \text{ s.t. } s = s'.$$

Thus $p^T(S)$ does not depend on T at all, and only depends on the cardinality of S. Let q_s denote $p^T(S)$ for any $S \subseteq N \setminus T$. Then we can re-write equation (4) in terms of q_s as

$$1 = \sum_{i=n-k}^{n-1} \binom{n}{i} q_i, \tag{9}$$

$$q_s = \frac{\sum_{i=(s-k)\vee 0}^{s-1} \binom{s}{i} q_i}{\sum_{i=1}^{k \wedge (n-s)} \binom{n-s}{i}} \quad \forall \, s \in \{1, \ldots, n-1\}. \tag{10}$$

Note that for any q_0, equation (10) fully determines all other q_i, for $i \in \{1, \ldots, n-1\}$ and q_0 is then determined by (9). Thus there is at most one solution. However, we have already identified (see the arrival-order discussion in Appendix A) that a solution to this recurrence is given by $(q_0, \ldots, q_{n-1}) = (\hat{p}_0, \ldots, \hat{p}_{n-1})$, for which $\hat{p}_0 = \left(\sum_{i=1}^{k} \binom{n}{i} \right)^{-1}$. □

C STRONG JOINT SYMMETRY

We examine the effect of removing conditions 2 and 3 from JSY. As it turns out, this leads to the non-existence of an interaction index. To be precise, we consider replacing axioms JAN and JSY with:

SJS *strong joint symmetry* : fix $\varnothing \neq T, T' \subseteq N$. Then

$$\begin{aligned} v(S \cup T) = v(S \cup T') \, \forall S \subseteq N \setminus (T \cup T') \\ \Rightarrow \phi_T(v) = \phi_{T'}(v). \end{aligned}$$

Proposition C.1. *There is no interaction index ϕ satisfying axioms JLI, JNU, JEF, and SJS that is guaranteed to exist for all games.*

Proof. Since ϕ satisfies JLI, JNU, and JEF, by Proposition 2,

$$\phi_T(v) = \sum_{S \subseteq N \setminus T} p^T(S)[v(S \cup T) - v(S)]$$

with $\{p^T(S)\}$ satisfying (4), for any game $v \in \mathcal{G}^N$. We consider two games $v_1, v_2 \in \mathcal{G}^{\{1,2\}}$. As $N = \{1, 2\}$, (4) gives $p^{\{1\}}(\varnothing) = p^{\{2\}}(\{1\})$, $p^{\{2\}}(\varnothing) = p^{\{1\}}(\{2\})$, and $p^{\{1,2\}}(\varnothing) + p^{\{1\}}(\varnothing) + p^{\{2\}}(\varnothing) = 1$.

Suppose $v_1(\{1\}) = v_1(\{1, 2\}) = 1$, $v_1(\{2\}) = 0$. SJS thus gives that $\phi_{\{1\}}(v_1) = \phi_{\{1,2\}}(v_1)$, i.e. $p^{\{1,2\}}(\varnothing) = p^{\{1\}}(\varnothing) + p^{\{2\}}(\varnothing)$ which implies $p^{\{1,2\}}(\varnothing) = 1/2$.

Suppose also that $v_2(\{1\}) = v_2(\{1, 2\}) = v_2(\{2\}) = 1$. SJS gives that $\phi_{\{1\}}(v_2) = \phi_{\{2\}}(v_2) = \phi_{\{1,2\}}(v_2)$, i.e. $p^{\{1,2\}}(\varnothing) = p^{\{1\}}(\varnothing) = p^{\{2\}}(\varnothing)$ which implies $p^{\{1,2\}}(\varnothing) = 1/3$, giving a contradiction. □

Thus, SJS is too strong a notion of symmetry, imposing linear restrictions on sets of unequal sizes.