Flow-based Attribution in Graphical Models: A Recursive Shapley Approach (Online Supplement)

Raghav Singal [1] **George Michailidis** [1 2] **Hoiyi Ng** [1]

A. Details on Existing Node-based Approaches

We elaborate on our presentation in §3 and show how selected existing node-based attribution approaches are captured by our graphical model framework. In particular, we discuss three existing approaches: (1) independent Shapley value (§A.1), (2) conditional Shapley value (§A.2), and (3) asymmetric Shapley value (§A.3).

A.1. Independent Shapley Value (ISV)

ISV is arguably the simplest application of Shapley value (Shapley, 1953) to the posited attribution problem and has appeared in various works (Strumbelj & Kononenko, 2010; Sun & Sundararajan, 2011; Sundararajan et al., 2020; Janzing et al., 2020; Sundararajan & Najmi, 2020). In the underlying game, the set of players is N. The empty coalition corresponds to the background input $\mathbf{X}^{(1)}$. If a player (node) $i \in \mathsf{N}$ is added to a coalition, then X_i changes from $X_i^{(1)}$ to $X_i^{(2)}$. Hence, a coalition $N \subseteq \mathsf{N}$ maps to a corresponding value of $\mathbf{X}$, i.e.,

$$\mathbf{X}^{\text{ISV}}(N) := (\mathbf{X}_N^{(2)}, \mathbf{X}_{\mathsf{N}\setminus N}^{(1)}).$$

It is possible for $\mathbf{X}^{\text{ISV}}(N)$ to violate the local relations in F and we interpret these violations via a do-operator (Pearl, 2009). For instance, in the model corresponding to Example 1 (Figure A below), if $N = \{1\}$, then $\mathbf{X}^{\text{ISV}}(N) = (X_1^{(2)}, X_2^{(1)}) = (1, 0)$, which violates $X_2 = X_1$. The characteristic function is defined as

$$v^{\text{ISV}}(N) := f(\mathbf{X}^{\text{ISV}}(N)),$$

which equals $f_{n+1}(\mathbf{X}^{\text{ISV}}_{\mathsf{P}_{n+1}}(N))$ given our do-operator interpretation. This results in node $i \in \mathsf{N}$ receiving an attribution (Shapley value) of

$$\pi_i^{\text{ISV}} = \sum_{N \subseteq \mathsf{N}\setminus\{i\}} w_{\mathsf{N}}(N) \times \left\{ v^{\text{ISV}}(N \cup \{i\}) - v^{\text{ISV}}(N) \right\},$$

where $w_{\mathsf{N}}(N) := \frac{|N|!(|\mathsf{N}|-|N|-1)!}{|\mathsf{N}|!}$ is the SV weight function. In general, given the do-operator interpretation, ISV attributes all value to the parents of the output node, i.e., $\sum_{i \in \mathsf{P}_{n+1}} \pi_i^{\text{ISV}} = Y^{(2)} - Y^{(1)}$ and $\pi_i^{\text{ISV}} = 0 \; \forall i \in \mathsf{N} \setminus \mathsf{P}_{n+1}$. This formalizes the following statement of Heskes et al. (2020) regarding ISV: "root causes with strong indirect effects (e.g. genetic markers) are ignored".

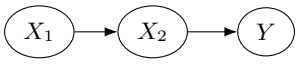


Figure A. The graphical model for Example 1. The source variable X_1 is set exogenously, $X_2 = X_1$, and $Y = X_2$. We consider the background value of $X_1^{(1)} = 0$ and the foreground value of $X_1^{(2)} = 1$. Hence, $X_2^{(1)} = Y^{(1)} = 0$ and $X_2^{(2)} = Y^{(2)} = 1$.

[1]Amazon [2]University of Florida. Correspondence to: RS <rs3566@columbia.edu>, GM <gmichail@ufl.edu>, HN <nghoiyi@amazon.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

As an illustration, in Example 1 introduced in §3 (Figure A), ISV attributes all the value to node 2, i.e., $(\pi_1^{\text{ISV}}, \pi_2^{\text{ISV}}) = (0, 1)$:

$$\begin{aligned}
\pi_1^{\text{ISV}} &= \sum_{N \subseteq \mathsf{N} \setminus \{1\}} w_{\mathsf{N}}(N) \times \left\{ v^{\text{ISV}}(N \cup \{1\}) - v^{\text{ISV}}(N) \right\} \\
&= \frac{1}{2} \left\{ v^{\text{ISV}}(\{1\}) - v^{\text{ISV}}(\emptyset) \right\} + \frac{1}{2} \left\{ v^{\text{ISV}}(\{1,2\}) - v^{\text{ISV}}(\{2\}) \right\} \\
&= \frac{1}{2} \left\{ f(\mathbf{X}^{\text{ISV}}(\{1\})) - f(\mathbf{X}^{\text{ISV}}(\emptyset)) \right\} + \frac{1}{2} \left\{ f(\mathbf{X}^{\text{ISV}}(\{1,2\})) - f(\mathbf{X}^{\text{ISV}}(\{2\})) \right\} \\
&= \frac{1}{2} \left\{ f(X_1^{(2)}, X_2^{(1)}) - f(X_1^{(1)}, X_2^{(1)}) \right\} + \frac{1}{2} \left\{ f(X_1^{(2)}, X_2^{(2)}) - f(X_1^{(1)}, X_2^{(2)}) \right\} \\
&= \frac{1}{2} \left\{ f_3(X_2^{(1)}) - f_3(X_2^{(1)}) \right\} + \frac{1}{2} \left\{ f_3(X_2^{(2)}) - f_3(X_2^{(2)}) \right\} \\
&= \frac{1}{2} \{0 - 0\} + \frac{1}{2} \{1 - 1\} = 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\pi_2^{\text{ISV}} &= \sum_{N \subseteq \mathsf{N} \setminus \{2\}} w_{\mathsf{N}}(N) \times \left\{ v^{\text{ISV}}(N \cup \{2\}) - v^{\text{ISV}}(N) \right\} \\
&= \frac{1}{2} \left\{ f(\mathbf{X}^{\text{ISV}}(\{2\})) - f(\mathbf{X}^{\text{ISV}}(\emptyset)) \right\} + \frac{1}{2} \left\{ f(\mathbf{X}^{\text{ISV}}(\{1,2\})) - f(\mathbf{X}^{\text{ISV}}(\{1\})) \right\} \\
&= \frac{1}{2} \left\{ f(X_1^{(1)}, X_2^{(2)}) - f(X_1^{(1)}, X_2^{(1)}) \right\} + \frac{1}{2} \left\{ f(X_1^{(2)}, X_2^{(2)}) - f(X_1^{(2)}, X_2^{(1)}) \right\} \\
&= \frac{1}{2} \left\{ f_3(X_2^{(2)}) - f_3(X_2^{(1)}) \right\} + \frac{1}{2} \left\{ f_3(X_2^{(2)}) - f_3(X_2^{(1)}) \right\} \\
&= \frac{1}{2} \{1 - 0\} + \frac{1}{2} \{1 - 0\} = 1.
\end{aligned}$$

A.2. Conditional Shapley Value (CSV)

CSV has been studied by Štrumbelj & Kononenko (2014); Datta et al. (2016); Lundberg & Lee (2017); Aas et al. (2019); Frye et al. (2020a). Similar to ISV, the set of players in the underlying game is N. However, the mapping from a coalition $N \subseteq \mathsf{N}$ to the input $\mathbf{X}$ is different. In particular,

$$\mathbf{X}^{\text{CSV}}(N) := (\mathbf{X}_N^{(2)}, \mathbf{X}_{\mathsf{N} \setminus N} \mid \mathbf{X}_N^{(2)}).$$

That is, instead of setting the "missing" nodes $\mathsf{N} \setminus N$ to the background value $\mathbf{X}_{\mathsf{N} \setminus N}^{(1)}$, their value is conditioned on the values of the nodes that are present in N. This ensures the local relations in F are satisfied. In our deterministic setup, if a *source* node $i \in \mathsf{N}_0 \setminus N$ is an ancestor (parent, grandparent, etc.) of a node in N, then it takes on its foreground value, i.e., $X_i^{\text{CSV}}(N) = X_i^{(2)}$. The remaining *source* nodes in $\mathsf{N}_0 \setminus N$ take on their background values and the non-source nodes in $\mathsf{N} \setminus N$ are determined via the equations in F. The characteristic function is

$$v^{\text{CSV}}(N) := f(\mathbf{X}^{\text{CSV}}(N))$$

and node $i \in \mathsf{N}$ receives an attribution of

$$\pi_i^{\text{CSV}} = \sum_{N \subseteq \mathsf{N} \setminus \{i\}} w_{\mathsf{N}}(N) \times \left\{ v^{\text{CSV}}(N \cup \{i\}) - v^{\text{CSV}}(N) \right\}.$$

CSV splits the value equally in Example 1: $(\pi_1^{\text{CSV}}, \pi_2^{\text{CSV}}) = (1/2, 1/2)$. Thus, it violates source efficiency. Computations are as follows:

$$\begin{aligned}
\pi_1^{\text{CSV}} &= \sum_{N \subseteq \mathsf{N} \setminus \{1\}} w_{\mathsf{N}}(N) \times \left\{ v^{\text{CSV}}(N \cup \{1\}) - v^{\text{CSV}}(N) \right\} \\
&= \frac{1}{2} \left\{ f(\mathbf{X}^{\text{CSV}}(\{1\})) - f(\mathbf{X}^{\text{CSV}}(\emptyset)) \right\} + \frac{1}{2} \left\{ f(\mathbf{X}^{\text{CSV}}(\{1,2\})) - f(\mathbf{X}^{\text{CSV}}(\{2\})) \right\} \\
&= \frac{1}{2} \left\{ f(X_1^{(2)}, X_2^{(2)}) - f(X_1^{(1)}, X_2^{(1)}) \right\} + \frac{1}{2} \left\{ f(X_1^{(2)}, X_2^{(2)}) - f(X_1^{(2)}, X_2^{(2)}) \right\} \\
&= \frac{1}{2} \{1 - 0\} + \frac{1}{2} \{1 - 1\} = \frac{1}{2}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\pi_2^{\text{CSV}} &= \sum_{N \subseteq \mathsf{N} \setminus \{2\}} w_{\mathsf{N}}(N) \times \left\{ v^{\text{ISV}}(N \cup \{2\}) - v^{\text{ISV}}(N) \right\} \\
&= \frac{1}{2} \left\{ f(\mathbf{X}^{\text{CSV}}(\{2\})) - f(\mathbf{X}^{\text{CSV}}(\emptyset)) \right\} + \frac{1}{2} \left\{ f(\mathbf{X}^{\text{CSV}}(\{1,2\})) - f(\mathbf{X}^{\text{CSV}}(\{1\})) \right\} \\
&= \frac{1}{2} \left\{ f(X_1^{(2)}, X_2^{(2)}) - f(X_1^{(1)}, X_2^{(1)}) \right\} + \frac{1}{2} \left\{ f(X_1^{(2)}, X_2^{(2)}) - f(X_1^{(2)}, X_2^{(2)}) \right\} \\
&= \frac{1}{2} \{1 - 0\} + \frac{1}{2} \{1 - 1\} = \frac{1}{2}.
\end{aligned}$$

A.3. Asymmetric Shapley Value (ASV)

ASV has been proposed by Frye et al. (2020b). In their words, "if X_i is known to be the deterministic causal ancestor of X_j, one might want to attribute all the importance to X_i and none to X_j". To formalize this intuition, Frye et al. (2020b) consider a game with the set of players being N, but re-define the weight function $w_{\mathsf{N}}(\cdot)$ to account for the structure of the graph. To facilitate comparison with ISV and CSV, we define ASV in an alternate but equivalent way, while keeping $w_{\mathsf{N}}(\cdot)$ as in §A.1. We define the set of players to be just the source nodes N_0. Empty coalition corresponds to $\mathbf{X}_{\mathsf{N}_0}^{(1)}$. If a node $i \in \mathsf{N}_0$ is added to a coalition, then X_i changes from $X_i^{(1)}$ to $X_i^{(2)}$. The mapping from a coalition $N_0 \subseteq \mathsf{N}_0$ to the input $\mathbf{X}$ is

$$\mathbf{X}^{\text{ASV}}(N_0) := (\mathbf{X}_{N_0}^{(2)}, \mathbf{X}_{\mathsf{N}_0 \setminus N_0}^{(1)}, \mathbf{X}_{\mathsf{N} \setminus \mathsf{N}_0} \mid (\mathbf{X}_{N_0}^{(2)}, \mathbf{X}_{\mathsf{N}_0 \setminus N_0}^{(1)})).$$

Hence, the rest of the graph is determined via the values at source nodes and the equations in F. This ensures the local relations in F are satisfied. The characteristic function is

$$v^{\text{ASV}}(N_0) := f(\mathbf{X}^{\text{ASV}}(N_0))$$

and node $i \in \mathsf{N}_0$ receives an attribution of

$$\pi_i^{\text{ASV}} = \sum_{N_0 \subseteq \mathsf{N}_0 \setminus \{i\}} w_{\mathsf{N}_0}(N_0) \times \left\{ v^{\text{ASV}}(N_0 \cup \{i\}) - v^{\text{ASV}}(N_0) \right\}.$$

In Example 1, ASV attributes all the value to node 1: $(\pi_1^{\text{ASV}}, \pi_2^{\text{ASV}}) = (1, 0)$. Though ASV obeys source efficiency, it does not tell how the effect flows through the graph. Computations are as follows:

$$\begin{aligned}
\pi_1^{\text{ASV}} &= \sum_{N_0 \subseteq \mathsf{N}_0 \setminus \{1\}} w_{\mathsf{N}_0}(N_0) \times \left\{ v^{\text{ASV}}(N_0 \cup \{1\}) - v^{\text{ASV}}(N_0) \right\} \\
&= f(\mathbf{X}^{\text{ASV}}(\{1\})) - f(\mathbf{X}^{\text{ASV}}(\emptyset)) \\
&= f(X_1^{(2)}, X_2^{(2)}) - f(X_1^{(1)}, X_2^{(1)}) \\
&= 1 - 0 = 1.
\end{aligned}$$

Node 2 is not a player in the game underlying ASV since the set of players only contains the source nodes N_0. Hence, node 2 receives zero attribution and $(\pi_1^{\text{ASV}}, \pi_2^{\text{ASV}}) = (1, 0)$.

Note that our definition of ASV corresponds to the "distal" variation in Frye et al. (2020b) (attribution to the "root causes" N_0). The "proximate" variation in Frye et al. (2020b) is the same as ISV (given our problem definition from §2) and it attributes to the "immediate causes" P_{n+1}. Irrespective of the variation, ASV does not provide a flow-based view. Also, it is not the case that Shapley values "ignore all causal structure", as stated in Frye et al. (2020b). As we show in §4 and §5, if the underlying game is defined appropriately (via a sequence of recursive games), then Shapley values capture the causal structure and in fact, generalize ASV (cf. Proposition 3 in §5 and Proposition B in Appendix G).

Remark A (Causal SV)**.** *Though we do not define the causal SV approach (Heskes et al., 2020), we note that "symmetric causal SV" attributes 1/2 to both nodes in Example 1 (violating source efficiency), whereas "asymmetric causal SV" attributes 1 to node 1 and 0 to node 2 (not a flow-based view). This follows from Figure 1 in Heskes et al. (2020).*

B. Details on Shapley Flow (SF)

In this appendix, we provide details on Shapley Flow (SF) (Wang et al., 2021) and highlight how it differs from the proposed flow-based approach in this work, recursive Shapley value (RSV), which is formally defined in §4. There are three significant differences between the two. For illustration, consider the graphical model in Figure B.

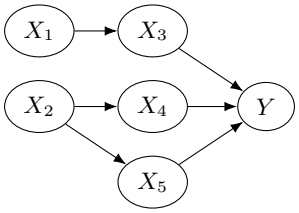


Figure B. The graphical model used to illustrate SF. There are two source nodes (X_1 and X_2), three intermediate nodes (X_3, X_4, and X_5), and one output node (Y). Suppose the structural equations are as follows: $X_3 = X_1$, $X_4 = \sqrt{X_2}$, $X_5 = \sqrt{X_2}$ and $Y = X_3X_4X_5$.

There are three unique paths that connect a source node to the output node in this graph:

1. $P_1 : 1 \to 3 \to 6$
2. $P_2 : 2 \to 4 \to 6$
3. $P_3 : 2 \to 5 \to 6$,

where "6" is used to denote the output node. We use $\mathbf{P} := \{P_1, P_2, P_3\}$ to denote the set of all such paths. In contrast to the *edge-based* RSV, SF is *path-based* (difference #1) as it considers each path in $\mathbf{P}$ as a player. With background $(X_1^{(1)}, X_2^{(1)}) = (0, 0)$ and foreground $(X_1^{(2)}, X_2^{(2)}) = (1, 1)$, the output function $Y_{\text{SF}}(\cdot)$ under path-based SF equals 1 if all three paths are present and 0 otherwise. There are 6 possible orderings of the players:

$$\Pi := \{P_1P_2P_3, P_1P_3P_2, P_2P_1P_3, P_2P_3P_1, P_3P_1P_2, P_3P_2P_1\}.$$

Thus, given the set of players $\mathbf{P}$ and the output function $Y_{\text{SF}}(\cdot)$, the Shapley value (SV) for path $P \in \mathbf{P}$ is

$$\frac{1}{|\Pi|} \sum_{\pi \in \Pi} \left\{ Y_{\text{SF}}(Q \in \mathbf{P} : \pi_Q \leq \pi_P) - Y_{\text{SF}}(Q \in \mathbf{P} : \pi_Q < \pi_P) \right\},$$

where $\pi_Q < \pi_P$ denotes whether path Q comes before path P in ordering π. However, SF modifies this definition by only considering the four orderings that are consistent with a depth first search (DFS):

$$\Pi_{\text{DFS}} := \{P_1P_2P_3, P_1P_3P_2, P_2P_3P_1, P_3P_2P_1\}.$$

That is, orderings $P_2P_1P_3$ and $P_3P_1P_2$ are not present in Π_{DFS}. The SF attribution to path $P \in \mathbf{P}$ equals:

$$\frac{1}{|\Pi_{\text{DFS}}|} \sum_{\pi \in \Pi_{\text{DFS}}} \left\{ Y_{\text{SF}}(Q \in \mathbf{P} : \pi_Q \leq \pi_P) - Y_{\text{SF}}(Q \in \mathbf{P} : \pi_Q < \pi_P) \right\}.$$

Given this modification (Π vs. Π_{DFS}), it is unclear what connection SF exhibits with SV (if any). To the best of our knowledge, there is no underlying "game" for which SF is the SV of. On the other hand, the edge-based RSV comes out naturally from a well-defined sequence of games, without any ad hoc modifications (difference #2).

The third difference concerns the definition of output $Y(\cdot)$ as a function of input edges. RSV uses the notion of "active / inactive edges" to define $Y(E)$ for a subset $E \subseteq \mathsf{E}$ of edges. This notion is foundational in causality theory (see, for example, Figure 3 in Pearl (2001)). On the other hand, SF uses the notion of "history", which is a *list* (as opposed to a *subset*) of edges, where an edge can appear twice. This notion is rather unnatural, since it can result in unrealistic counterfactuals. For instance, in Figure 3 of the manuscript (chain graph), if "history" equals $[(2, 3), (1, 2)]$, then the output function in SF will equal its background value even though all the edges are present.

C. RSV Computations for Example 1

In terms of the computations for Example 1, recall that RSV inserts a super-source node 0, as shown in Figure C.

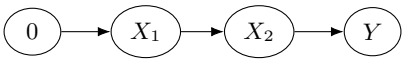


Figure C. The graphical model for Example 1 with a super-source node. X_1 is set exogenously, $X_2 = X_1$, and $Y = X_2$. We consider the background value of $X_1^{(1)} = 0$ and the foreground value of $X_1^{(2)} = 1$. Hence, $X_2^{(1)} = Y^{(1)} = 0$ and $X_2^{(2)} = Y^{(2)} = 1$.

First, consider the game at node 0. The set of players is $\mathsf{E}_0 = \{(0,1)\}$ and the characteristic function given coalition $E_0 \subseteq \mathsf{E}_0$ equals $v_0(E_0) = Y(E_0, \mathsf{E}_1, \ldots, \mathsf{E}_n)$, where $Y(\cdot)$ follows (1) and (2). Hence, the attribution received by edge $(0,1)$ equals

$$\begin{aligned}
\pi_{01}^{\text{RSV}} &= \pi_{01}(\mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3) \\
&= \sum_{E_0 \subseteq \mathsf{E}_0 \setminus \{(0,1)\}} w_{\mathsf{E}_0}(E_0) \times \{v_0(E_0 \cup \{(0,1)\}) - v_0(E_0)\} \\
&= v_0(\{(0,1)\}) - v_0(\emptyset) \\
&= Y(\{(0,1)\}, \mathsf{E}_1, \mathsf{E}_2) - Y(\emptyset, \mathsf{E}_1, \mathsf{E}_2) \\
&= Y^{(2)} - Y^{(1)} = 1 - 0 = 1.
\end{aligned}$$

Second, consider the game at node 1. The set of players is $\mathsf{E}_1 = \{(1,2)\}$ and the characteristic function given coalition $E_1 \subseteq \mathsf{E}_1$ equals equals $v_1(E_1) = \pi_1(\mathsf{E}_0, E_1, \mathsf{E}_2) = \pi_{01}(\mathsf{E}_0, E_1, \mathsf{E}_2)$. Hence, the attribution received by edge $(1,2)$ equals

$$\begin{aligned}
\pi_{12}^{\text{RSV}} &= \pi_{12}(\mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3) \\
&= \sum_{E_1 \subseteq \mathsf{E}_1 \setminus \{(1,2)\}} w_{\mathsf{E}_1}(E_1) \times \{v_1(E_1 \cup \{(1,2)\}) - v_1(E_1)\} \\
&= v_1(\{(1,2)\}) - v_1(\emptyset) \\
&= \pi_{01}(\mathsf{E}_0, \{(1,2)\}, \mathsf{E}_2) - \pi_{01}(\mathsf{E}_0, \emptyset, \mathsf{E}_2) \\
&= \pi_{01}^{\text{RSV}} - \pi_{01}(\mathsf{E}_0, \emptyset, \mathsf{E}_2) \\
&= (Y^{(2)} - Y^{(1)}) - 0 = 1.
\end{aligned}$$

Above, we used the fact that $\pi_{01}(\mathsf{E}_0, \emptyset, \mathsf{E}_2) = 0$, which is true because

$$\begin{aligned}
\pi_{01}(\mathsf{E}_0, \emptyset, \mathsf{E}_2) &= \sum_{E_0 \subseteq \mathsf{E}_0 \setminus \{(0,1)\}} w_{\mathsf{E}_0}(E_0) \times \{v_0(E_0 \cup \{(0,1)\} \mid E_1 = \emptyset) - v_0(E_0 \mid E_1 = \emptyset)\} \\
&= v_0(\{(0,1)\} \mid E_1 = \emptyset) - v_0(\emptyset \mid E_1 = \emptyset) \\
&= Y(\{(0,1)\}, \emptyset, \mathsf{E}_2) - Y(\emptyset, \emptyset, \mathsf{E}_2) \\
&= Y^{(1)} - Y^{(1)} = 0.
\end{aligned}$$

Third, consider the game at node 2. The set of players is $\mathsf{E}_2 = \{(2,3)\}$ and the characteristic function given coalition $E_2 \subseteq \mathsf{E}_2$ equals equals $v_2(E_2) = \pi_2(\mathsf{E}_0, \mathsf{E}_1, E_2) = \pi_{12}(\mathsf{E}_0, \mathsf{E}_1, E_2)$. Hence, the attribution received by edge $(2,3)$ equals

$$\begin{aligned}
\pi_{23}^{\text{RSV}} &= \pi_{23}(\mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3) \\
&= \sum_{E_2 \subseteq \mathsf{E}_2 \setminus \{(2,3)\}} w_{\mathsf{E}_2}(E_2) \times \{v_2(E_2 \cup \{(2,3)\}) - v_2(E_2)\} \\
&= v_2(\{(2,3)\}) - v_2(\emptyset) \\
&= \pi_{12}(\mathsf{E}_0, \mathsf{E}_1, \{(2,3)\}) - \pi_{12}(\mathsf{E}_0, \mathsf{E}_1, \emptyset) \\
&= \pi_{12}^{\text{RSV}} - \pi_{12}(\mathsf{E}_0, \mathsf{E}_1, \emptyset) \\
&= (Y^{(2)} - Y^{(1)}) - 0 = 1.
\end{aligned}$$

Above, we used the fact that $\pi_{12}(\mathsf{E}_0, \mathsf{E}_1, \emptyset) = 0$, which is true because

$$\begin{aligned}
\pi_{12}(\mathsf{E}_0, \mathsf{E}_1, \emptyset) &= \sum_{E_1 \subseteq \mathsf{E}_1 \setminus \{(1,2)\}} w_{\mathsf{E}_1}(E_1) \times \{v_1(E_1 \cup \{(1,2)\} \mid E_2 = \emptyset) - v_1(E_1 \mid E_2 = \emptyset)\} \\
&= v_1(\{(1,2)\} \mid E_2 = \emptyset) - v_1(\emptyset \mid E_2 = \emptyset) \\
&= \pi_{01}(\mathsf{E}_0, \mathsf{E}_1, \emptyset) - \pi_{01}(\mathsf{E}_0, \emptyset, \emptyset) \\
&= \sum_{E_0 \subseteq \mathsf{E}_0 \setminus \{(0,1)\}} w_{\mathsf{E}_0}(E_0) \times \{v_0(E_0 \cup \{(0,1)\} \mid E_2 = \emptyset) - v_0(E_0 \mid E_2 = \emptyset)\} \\
&\quad - \sum_{E_0 \subseteq \mathsf{E}_0 \setminus \{(0,1)\}} w_{\mathsf{E}_0}(E_0) \times \{v_0(E_0 \cup \{(0,1)\} \mid E_1 = \emptyset, E_2 = \emptyset) - v_0(E_0 \mid E_1 = \emptyset, E_2 = \emptyset)\} \\
&= (Y(\mathsf{E}_0, \mathsf{E}_1, \emptyset) - Y(\emptyset, \mathsf{E}_1, \emptyset)) - (Y(\mathsf{E}_0, \emptyset, \emptyset) - Y(\emptyset, \emptyset, \emptyset)) \\
&= (Y^{(1)} - Y^{(1)}) - (Y^{(1)} - Y^{(1)}) = 0.
\end{aligned}$$

Hence, $(\pi_{01}^{\text{RSV}}, \pi_{12}^{\text{RSV}}, \pi_{23}^{\text{RSV}}) = (1, 1, 1)$.

D. Proof of Theorem 1

Theorem 1. $[\pi^{\text{RSV}}_{jk}]_{(j,k)\in\mathsf{E}}$ is the unique solution to the flow-based axioms.

Proof of Theorem 1. First, consider the game at node 0. The set of players is E_0 and the characteristic function given coalition $E_0 \subseteq \mathsf{E}_0$ is $v_0(E_0) = Y(E_0, \mathsf{E}_1, \ldots, \mathsf{E}_n)$. RSV attributes the Shapley values of this game to the source edges, i.e.,

$$\pi^{\text{RSV}}_{0k} = \sum_{E_0 \subseteq \mathsf{E}_0 \setminus \{(0,k)\}} w_{\mathsf{E}_0}(E_0) \times \{v_0(E_0 \cup \{(0,k)\}) - v_0(E_0)\} \;\; \forall (0,k) \in \mathsf{E}_0.$$

Invoking the uniqueness result of the classical Shapley value (Shapley, 1953) implies that $[\pi^{\text{RSV}}_{0k}]_{(0,k)\in\mathsf{E}_0}$ is the unique solution to the flow-based axioms at node 0. In particular, flow symmetry, flow nullity, and flow linearity are equivalent to the corresponding axioms (symmetry, nullity, and linearity) of the classical Shapley value. Furthermore, the efficiency axiom of the classical Shapley value states $\sum_{k\in\mathsf{C}_0} \pi^{\text{RSV}}_{0k} = v_0(\mathsf{E}_0) - v_0(\emptyset)$, which is equivalent to flow conservation at node 0 since

$$\begin{aligned} v_0(\mathsf{E}_0) - v_0(\emptyset) &= Y(\mathsf{E}_0, \mathsf{E}_1, \ldots, \mathsf{E}_n) - Y(\emptyset, \mathsf{E}_1, \ldots, \mathsf{E}_n) \\ &= Y^{(2)} - Y^{(1)}. \end{aligned}$$

The last equality follows the definition of $Y(\cdot)$ (see (1) and (2)). In particular, $Y(\mathsf{E}_0, \mathsf{E}_1, \ldots, \mathsf{E}_n) = Y^{(2)}$ and $Y(\emptyset, \mathsf{E}_1, \ldots, \mathsf{E}_n) = Y^{(1)}$.

Second, consider the game at node $j \in \mathsf{N} \setminus \{0\}$. The set of players is E_j and the characteristic function given coalition $E_j \subseteq \mathsf{E}_j$ is $v_j(E_j) = \sum_{i\in\mathsf{P}_j} \pi_{ij}(\mathsf{E}_0, \ldots, E_j, \ldots, \mathsf{E}_n)$. RSV attributes the Shapley values of this game to edges E_j, i.e.,

$$\pi^{\text{RSV}}_{jk} = \sum_{E_j \subseteq \mathsf{E}_j \setminus \{(j,k)\}} w_{\mathsf{E}_j}(E_j) \times \{v_j(E_j \cup \{(j,k)\}) - v_j(E_j)\} \;\; \forall (j,k) \in \mathsf{E}_j.$$

Invoking the uniqueness result of the classical Shapley value (Shapley, 1953) implies that $[\pi^{\text{RSV}}_{jk}]_{(j,k)\in\mathsf{E}_j}$ is the unique solution to the flow-based axioms at node j. In particular, flow symmetry, flow nullity, and flow linearity are equivalent to the corresponding axioms (symmetry, nullity, and linearity) of the classical Shapley value. Furthermore, the efficiency axiom of the classical Shapley value states $\sum_{k\in\mathsf{C}_j} \pi^{\text{RSV}}_{jk} = v_j(\mathsf{E}_j) - v_j(\emptyset)$, which is equivalent to flow conservation at node j since

$$\begin{aligned} v_j(\mathsf{E}_j) - v_j(\emptyset) &= \sum_{i\in\mathsf{P}_j} \pi_{ij}(\mathsf{E}_0, \ldots, \mathsf{E}_j, \ldots, \mathsf{E}_n) - \sum_{i\in\mathsf{P}_j} \pi_{ij}(\mathsf{E}_0, \ldots, \emptyset, \ldots, \mathsf{E}_n) \\ &= \sum_{i\in\mathsf{P}_j} \pi^{\text{RSV}}_{ij}. \end{aligned}$$

The last equality is true because $\pi_{ij}(\mathsf{E}_0, \ldots, \mathsf{E}_j, \ldots, \mathsf{E}_n) = \pi^{\text{RSV}}_{ij} \; \forall i \in \mathsf{P}_j$ (by definition) and $\pi_{ij}(\mathsf{E}_0, \ldots, \emptyset, \ldots, \mathsf{E}_n) = 0 \; \forall i \in \mathsf{P}_j$ since given $E_j = \emptyset$, node j passes no new information to its children (follows the definition of $Y(\cdot)$ as in (1) and (2)) and hence, edge (i,j) will be a null player in the upstream game at each node $i \in \mathsf{P}_j$. Putting the uniqueness of $[\pi^{\text{RSV}}_{jk}]_{k\in\mathsf{C}_j}$ w.r.t. the axioms at each node $j \in \mathsf{N}$ implies that $[\pi^{\text{RSV}}_{jk}]_{(j,k)\in\mathsf{E}}$ is the unique solution to the flow-based axioms. This completes the proof. □

E. RSV Characterization under a Linear Model

In this appendix, we generalize our discussion around Example 2 and characterize RSV under a linear model. §E.1 presents the characterization and §E.2 discusses a technical lemma we use to prove the characterization.

E.1. Characterization

Consider an arbitrary DAG G (containing node 0) with linear equations F:

$$X_j = \sum_{i \in \mathsf{P}_j} a_{ij} X_i \; \forall j \in \mathsf{N}^+ \setminus \{\mathsf{N}_0 \cup 0\}. \tag{E.1}$$

Denote by $\mathsf{E}^{\text{in}}_j := \{(i,j) : i \in \mathsf{P}_j\}$ the incoming edges of node $j \in \mathsf{N}^+ \setminus \{0\}$. Define the *forward-looking* weights $\boldsymbol{c}$ as follows:

$$c_{j,n+1} := a_{j,n+1} \qquad \forall (j, n+1) \in \mathsf{E}^{\text{in}}_{n+1} \tag{E.2a}$$

$$c_{ij} := a_{ij} \sum_{k \in \mathsf{C}_j} c_{jk} \qquad \forall (i,j) \in \mathsf{E} \setminus \mathsf{E}^{\text{in}}_{n+1}, \tag{E.2b}$$

where $a_{0j} := 1 \; \forall (0,j) \in \mathsf{E}_0$. Similarly, define the *backward-looking* weights $\boldsymbol{b}$ as follows:

$$b_{0j} := 1 \qquad \forall (0,j) \in \mathsf{E}_0 \tag{E.3a}$$

$$b_{jk} := \sum_{i \in \mathsf{P}_j} b_{ij} a_{jk} \qquad \forall (j,k) \in \mathsf{E} \setminus \mathsf{E}_0. \tag{E.3b}$$

Since G is a DAG, weights (E.2) and (E.3) are well-defined. Then, the RSV attribution is as stated in Proposition A. (We assume $X_i^{(1)} = 0$ and $X_i^{(2)} = 1 \; \forall i \in \mathsf{N}_0$ to keep the presentation clean. The generalization is straightforward.)

Proposition A. *Consider DAG* G *and linear* F *as in* (E.1)*, with* $\boldsymbol{c}$ *and* $\boldsymbol{b}$ *as in* (E.2) *and* (E.3)*. Suppose* $X_i^{(1)} = 0$ *and* $X_i^{(2)} = 1 \; \forall i \in \mathsf{N}_0$*. Then,*

$$\pi^{RSV}_{0j} = c_{0j} \qquad \forall (0,j) \in \mathsf{E}_0$$

$$\pi^{RSV}_{jk} = \sum_{i \in \mathsf{P}_j} b_{ij} c_{jk} \qquad \forall (j,k) \in \mathsf{E} \setminus \mathsf{E}_0.$$

Proof of Proposition A. Plugging in $\mathcal{E}$ equal to E in Lemma A (§E.2) gives the following attributions to all outgoing edges $(0,j) \in \mathsf{E}_0$ of node 0:

$$\begin{aligned}
\pi_{0j}(\mathsf{E}) &= \sum_{k:(j,k) \in \mathcal{E}_j} b_{0j}(\mathsf{E}_0) c_{jk}(\mathsf{E}_j, \ldots, \mathsf{E}_n) \\
&= \sum_{k:(j,k) \in \mathcal{E}_j} b_{0j} c_{jk} \\
&= \sum_{k:(j,k) \in \mathcal{E}_j} c_{jk}.
\end{aligned}$$

The second equality is true because $\boldsymbol{b} = \boldsymbol{b}(\mathsf{E})$ (see (E.3) and (E.5)) and $\boldsymbol{c} = \boldsymbol{c}(\mathsf{E})$ (see (E.2) and (E.4)) and the third equality is true because $b_{0j} = 1 \; \forall (0,j) \in \mathsf{E}_0$. Similarly, plugging in $\mathcal{E}$ equal to E in Lemma A gives the following attributions to

$(j,k) \in \mathsf{E} \setminus \mathsf{E}_0$:

$$\begin{aligned}
\pi_{jk}(\mathsf{E}) &= \sum_{\ell:(k,\ell)\in\mathsf{E}_k} b_{jk}(\mathsf{E}_0,\dots,\mathsf{E}_j)c_{k\ell}(\mathsf{E}_k,\dots,\mathsf{E}_n) \\
&= \sum_{\ell\in\mathsf{C}_k} b_{jk}c_{k\ell} \\
&= \sum_{\ell\in\mathsf{C}_k}\sum_{i\in\mathsf{P}_j} b_{ij}a_{jk}c_{k\ell} \\
&= \sum_{i\in\mathsf{P}_j} b_{ij}a_{jk}\sum_{\ell\in\mathsf{C}_k} c_{k\ell} \\
&= \sum_{i\in\mathsf{P}_j} b_{ij}c_{jk}.
\end{aligned}$$

The third and fifth equalities follow the definitions of $\boldsymbol{b}$ and $\boldsymbol{c}$, respectively (see (E.3) and (E.2)). Recall that π_{jk}^{RSV} is defined as $\pi_{jk}(\mathsf{E})$ for all $(j,k) \in \mathsf{E}$. The proof is now complete. □

E.2. Details on Lemma A

The proof of Proposition A leverages a more general result (Lemma A), which we present now. We generalize the definitions of the forward-looking and backward-looking weights $\boldsymbol{c}$ and $\boldsymbol{b}$. In §E.1 (see Equations (E.2) and (E.3)), we implicitly assumed all edges E to be active. We now define these weights as a function of subset $\mathcal{E} = (\mathcal{E}_0,\dots,\mathcal{E}_n) \subseteq \mathsf{E}$. We assume wlog that the DAG G is topologically sorted, i.e., there is no edge $(i,j) \in \mathsf{E}$ with $i > j$ (Cormen et al., 2009). The generalized forward-looking weights $\boldsymbol{c}(\mathcal{E})$ are defined as follows:

$$c_{j,n+1}(\mathcal{E}_j) := a_{j,n+1}\mathbb{I}\{(j,n+1) \in \mathcal{E}_j\} \qquad \forall (j,n+1) \in \mathsf{E}^{\text{in}}_{n+1} \tag{E.4a}$$

$$c_{ij}(\mathcal{E}_i,\dots,\mathcal{E}_n) := a_{ij}\mathbb{I}\{(i,j) \in \mathcal{E}_i\} \sum_{k:(j,k)\in\mathcal{E}_j} c_{jk}(\mathcal{E}_j,\dots,\mathcal{E}_n) \qquad \forall (i,j) \in \mathsf{E} \setminus \mathsf{E}^{\text{in}}_{n+1}, \tag{E.4b}$$

where $a_{0j} := 1\ \forall (0,j) \in \mathsf{E}_0$ as before. Similarly, the generalized backward-looking weights $\boldsymbol{b}(\mathcal{E})$ are defined as follows:

$$b_{0j}(\mathcal{E}_0) := \mathbb{I}\{(0,j) \in \mathcal{E}_0\} \qquad \forall (0,j) \in \mathsf{E}_0 \tag{E.5a}$$

$$b_{jk}(\mathcal{E}_0,\dots,\mathcal{E}_j) := \sum_{i:(i,j)\in\mathcal{E}_i} b_{ij}(\mathcal{E}_0,\dots,\mathcal{E}_i)a_{jk}\mathbb{I}\{(j,k) \in \mathcal{E}_j\} \qquad \forall (j,k) \in \mathsf{E} \setminus \mathsf{E}_0. \tag{E.5b}$$

Note that plugging in $\mathcal{E}$ as E recovers the original weights of (E.2) and (E.3), i.e., $\boldsymbol{b} = \boldsymbol{b}(\mathsf{E})$ and $\boldsymbol{c} = \boldsymbol{c}(\mathsf{E})$. We are now in a position to present Lemma A.

Lemma A. *Consider DAG G and linear F as in (E.1), with $\boldsymbol{c}(\cdot)$ and $\boldsymbol{b}(\cdot)$ as in (E.4) and (E.5). Suppose $X_i^{(1)} = 0$ and $X_i^{(2)} = 1\ \forall i \in \mathsf{N}_0$. Then, given $\mathcal{E} \subseteq \mathsf{E}$, for all $i \in \mathsf{N}$,*

$$\pi_{ij}(\mathcal{E}) = \sum_{k:(j,k)\in\mathcal{E}_j} b_{ij}(\mathcal{E}_0,\dots,\mathcal{E}_i)c_{jk}(\mathcal{E}_j,\dots,\mathcal{E}_n) \qquad \forall j : (i,j) \in \mathcal{E},$$

where $\mathcal{E} = (\mathcal{E}_0,\dots,\mathcal{E}_n)$ and the graph G is assumed to be topologically sorted (wlog).

Proof of Lemma A. First, consider the super-source node $i = 0$. By definition, $\pi_{0j}(\mathcal{E})$ corresponds to the Shapley value of the following game. The set of players is $\mathcal{E}_0$ and for a given coalition $E_0 \subseteq \mathcal{E}_0$, characteristic function equals

$$\begin{aligned}
v_0(E_0 \mid \mathcal{E}_{-0}) &= Y(E_0,\mathcal{E}_1,\dots,\mathcal{E}_n) \\
&= \sum_{j:(0,j)\in E_0} c_{0j}(E_0,\mathcal{E}_1,\dots,\mathcal{E}_n) \\
&= \sum_{j:(0,j)\in E_0} a_{0j}\mathbb{I}\{(0,j) \in E_0\} \sum_{k:(j,k)\in\mathcal{E}_j} c_{jk}(\mathcal{E}_j,\dots,\mathcal{E}_n) \\
&= \sum_{j:(0,j)\in E_0}\ \sum_{k:(j,k)\in\mathcal{E}_j} c_{jk}(\mathcal{E}_j,\dots,\mathcal{E}_n).
\end{aligned}$$

The second equality follows the definition of $\boldsymbol{c}(\cdot)$ (see (E.4)) and $Y(\cdot)$ (see (1) and (2)). The third equality follows the definition of $\boldsymbol{c}(\cdot)$ (see (E.4)). The final equality holds since $a_{0j} = 1$ and $\mathbb{I}\{(0,j) \in E_0\} = 1$ for $(0,j) \in E_0$. Given the separability of $v_0(E_0 \mid \mathcal{E}_{-0})$ over the players $(0,j)$ in $\mathcal{E}_0$, it directly follows that the Shapley value for player $(0,j) \in \mathcal{E}_0$ in such a game equals

$$\begin{aligned}\pi_{0j}(\mathcal{E}) &= \sum_{k:(j,k)\in\mathcal{E}_j} c_{jk}(\mathcal{E}_j, \ldots, \mathcal{E}_n) \\ &= \sum_{k:(j,k)\in\mathcal{E}_j} b_{0j}(\mathcal{E}_0) c_{jk}(\mathcal{E}_j, \ldots, \mathcal{E}_n),\end{aligned}$$

where the second equality is true because $b_{0j}(\mathcal{E}_0) = 1$ for $(0,j) \in \mathcal{E}_0$ by definition (see (E.5)). This completes the "base case".

Given the DAG structure, it suffices to show the statement holds at node $j \in \mathsf{N}$ by assuming the statement to hold at each of its parent nodes i s.t. $(i,j) \in \mathcal{E}_i$. That is, it suffices to show that

$$\pi_{ij}(\mathcal{E}) = \sum_{k:(j,k)\in\mathcal{E}_j} b_{ij}(\mathcal{E}_0, \ldots, \mathcal{E}_i) c_{jk}(\mathcal{E}_j, \ldots, \mathcal{E}_n) \; \forall j : (i,j) \in \mathcal{E} \tag{E.6}$$

implies

$$\pi_{jk}(\mathcal{E}) = \sum_{\ell:(k,\ell)\in\mathcal{E}_k} b_{jk}(\mathcal{E}_0, \ldots, \mathcal{E}_j) c_{k\ell}(\mathcal{E}_k, \ldots, \mathcal{E}_n) \; \forall k : (j,k) \in \mathcal{E}.$$

By definition, $\pi_{jk}(\mathcal{E})$ corresponds to the Shapley value of the following game. The set of players is $\mathcal{E}_j$ and for a given coalition $E_j \subseteq \mathcal{E}_j$, characteristic function equals

$$\begin{aligned}v_j(E_j \mid \mathcal{E}_{-j}) &= \sum_{i:(i,j)\in\mathcal{E}_i} \pi_{ij}(\mathcal{E}_0, \ldots, E_j, \ldots, \mathcal{E}_n) \\ &= \sum_{i:(i,j)\in\mathcal{E}_i} \sum_{k:(j,k)\in E_j} b_{ij}(\mathcal{E}_0, \ldots, \mathcal{E}_i) c_{jk}(E_j, \ldots, \mathcal{E}_n) \\ &= \sum_{i:(i,j)\in\mathcal{E}_i} \sum_{k:(j,k)\in E_j} b_{ij}(\mathcal{E}_0, \ldots, \mathcal{E}_i) c_{jk}(\mathcal{E}_j, \ldots, \mathcal{E}_n) \\ &= \sum_{k:(j,k)\in E_j} \sum_{i:(i,j)\in\mathcal{E}_i} b_{ij}(\mathcal{E}_0, \ldots, \mathcal{E}_i) c_{jk}(\mathcal{E}_j, \ldots, \mathcal{E}_n).\end{aligned}$$

The second equality follows (E.6). The third equality holds since $c_{jk}(E_j, \ldots, \mathcal{E}_n) = c_{jk}(\mathcal{E}_j, \ldots, \mathcal{E}_n)$ for $(j,k) \in E_j$ (see (E.4)). Given the separability of the characteristic function over the players, it directly follows that the Shapley value for player $(j,k) \in \mathcal{E}_j$ in such a game equals

$$\begin{aligned}\pi_{jk}(\mathcal{E}) &= \sum_{i:(i,j)\in\mathcal{E}_i} b_{ij}(\mathcal{E}_0, \ldots, \mathcal{E}_i) c_{jk}(\mathcal{E}_j, \ldots, \mathcal{E}_n) \\ &= \sum_{i:(i,j)\in\mathcal{E}_i} b_{ij}(\mathcal{E}_0, \ldots, \mathcal{E}_i) a_{jk} \mathbb{I}\{(j,k) \in \mathcal{E}_j\} \sum_{\ell:(k,\ell)\in\mathcal{E}_k} c_{k\ell}(\mathcal{E}_k, \ldots, \mathcal{E}_n) \\ &= b_{jk}(\mathcal{E}_0, \ldots, \mathcal{E}_j) \sum_{\ell:(k,\ell)\in\mathcal{E}_k} c_{k\ell}(\mathcal{E}_k, \ldots, \mathcal{E}_n) \\ &= \sum_{\ell:(k,\ell)\in\mathcal{E}_k} b_{jk}(\mathcal{E}_0, \ldots, \mathcal{E}_j) c_{k\ell}(\mathcal{E}_k, \ldots, \mathcal{E}_n),\end{aligned}$$

where the second and third equalities follow the definitions of $\boldsymbol{c}(\cdot)$ (see (E.4)) and $\boldsymbol{b}(\cdot)$ (see (E.5)), respectively. This completes the proof. □

F. Further Details on the Additional Properties of RSV

In this appendix, we provide further details on the additional properties of RSV from §5: proof of Proposition 1 (§F.1), implementation invariance example (§F.2), and proof of Proposition 2 (§F.3).

F.1. Proof of Proposition 1

Proposition 1. RSV obeys implementation invariance, sensitivity(a), and sensitivity(b).

Proof of Proposition 1. For implementation invariance, recall from §4 that RSV attributes to edges $[\pi_{ij}^{\text{RSV}}]_{(i,j)\in\mathsf{E}}$ and the node attributions are defined as $\pi_j^{\text{RSV}} := \sum_{i\in\mathsf{P}_j} \pi_{ij}^{\text{RSV}}$ for all $j \in \mathsf{N}\setminus\{0\}$. Furthermore, RSV adds a super-source node 0 with an edge directed to each source node. Hence, $\pi_j^{\text{RSV}} = \pi_{0j}^{\text{RSV}}$ for all $j \in \mathsf{N}_0$. Now, for $j \in \mathsf{N}_0$, observe that

$$
\begin{aligned}
\pi_j^{\text{RSV}}(\mathsf{M}_1) &= \pi_{0j}^{\text{RSV}}(\mathsf{M}_1) \\
&= \sum_{E_0\subseteq\mathsf{E}_0\setminus\{(0,j)\}} w_{\mathsf{E}_0}(E_0)\times\{v_0(E_0\cup\{(0,j)\}\mid\mathsf{M}_1) - v_0(E_0\mid\mathsf{M}_1)\} \\
&= \sum_{E_0\subseteq\mathsf{E}_0\setminus\{(0,j)\}} w_{\mathsf{E}_0}(E_0)\times\{Y(E_0\cup\{(0,j)\},\mathsf{E}_{-0}\mid\mathsf{M}_1) - Y(E_0,\mathsf{E}_{-0}\mid\mathsf{M}_1)\} \\
&= \sum_{E_0\subseteq\mathsf{E}_0\setminus\{(0,j)\}} w_{\mathsf{E}_0}(E_0)\times\left\{g(\mathbf{X}^{(2)}_{N_0\cup\{j\}},\mathbf{X}^{(1)}_{\mathsf{N}_0\setminus\{N_0\cup\{j\}\}}\mid\mathsf{F}_1) - g(\mathbf{X}^{(2)}_{N_0},\mathbf{X}^{(1)}_{\mathsf{N}_0\setminus N_0}\mid\mathsf{F}_1)\right\} \\
&\overset{(\star)}{=} \sum_{E_0\subseteq\mathsf{E}_0\setminus\{(0,j)\}} w_{\mathsf{E}_0}(E_0)\times\left\{g(\mathbf{X}^{(2)}_{N_0\cup\{j\}},\mathbf{X}^{(1)}_{\mathsf{N}_0\setminus\{N_0\cup\{j\}\}}\mid\mathsf{F}_2) - g(\mathbf{X}^{(2)}_{N_0},\mathbf{X}^{(1)}_{\mathsf{N}_0\setminus N_0}\mid\mathsf{F}_2)\right\} \\
&= \sum_{E_0\subseteq\mathsf{E}_0\setminus\{(0,j)\}} w_{\mathsf{E}_0}(E_0)\times\{Y(E_0\cup\{(0,j)\},\mathsf{E}_{-0}\mid\mathsf{M}_2) - Y(E_0,\mathsf{E}_{-0}\mid\mathsf{M}_2)\} \\
&= \sum_{E_0\subseteq\mathsf{E}_0\setminus\{(0,j)\}} w_{\mathsf{E}_0}(E_0)\times\{v_0(E_0\cup\{(0,j)\}\mid\mathsf{M}_2) - v_0(E_0\mid\mathsf{M}_2)\} \\
&= \pi_{0j}^{\text{RSV}}(\mathsf{M}_2) \\
&= \pi_j^{\text{RSV}}(\mathsf{M}_2).
\end{aligned}
$$

The notation $v_0(\cdot\mid\mathsf{M})$ and $Y(\cdot\mid\mathsf{M})$ captures the dependence of the characteristic function $v_0(\cdot)$ and the output $Y(\cdot)$ on the model M. Given E_0, $N_0 := \{k : (0,k)\in E_0\}$ denotes the children nodes of 0 that are present in E_0. To be thorough, we should use $N_0(E_0)$ but we omit the dependence on E_0 for conciseness. The key step is $(\star)$, which is true because $g(\cdot\mid\mathsf{F}_1) = g(\cdot\mid\mathsf{F}_2)$. All other equalities directly follow the corresponding definitions. This proves RSV obeys implementation invariance.

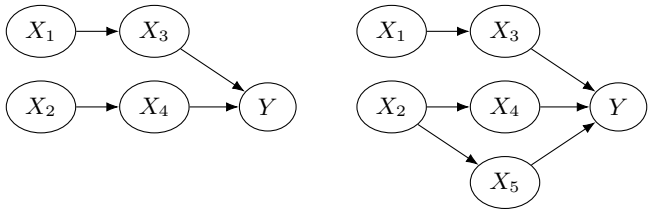
To see why RSV obeys sensitivity(a), suppose the qualifiers in Definition 4 hold. Then, it follows from the definition of the game at the super-source node 0 that $\pi_{0j}^{\text{RSV}} = 0$ for all $j\in\mathsf{N}_0\setminus\{i\}$ (flow nullity) and $\pi_{0i}^{\text{RSV}} = Y^{(2)} - Y^{(1)}$ (flow conservation). Since $\pi_j^{\text{RSV}} = \pi_{0j}^{\text{RSV}}$ for all $j\in\mathsf{N}_0$ (by definition), $\pi_i^{\text{RSV}} = \pi_{0i}^{\text{RSV}} = Y^{(2)} - Y^{(1)} \neq 0$ and hence, sensitivity(a) holds. Similarly, for sensitivity(b), under the condition in Definition 5, flow nullity implies $\pi_i^{\text{RSV}} = 0$. This completes the proof. □

F.2. Implementation Invariance Example

We use the example from §8.2 of Dhamdhere et al. (2018), which shows that backpropagation-based approaches such as DeepLIFT (Shrikumar et al., 2017), LRP (Binder et al., 2016), and DeepSHAP (Lundberg & Lee, 2017) do not satisfy implementation invariance.

Example A. *Consider the two models in Figure D. Both the models have the set of source nodes. Furthermore, the input-output function is the same in both:* $g(X_1,X_2) = X_1X_2$. *The key difference between them is in the "internal" structure of the graph. In particular, node 4 is split into two nodes in model 2, while preserving the input-output mapping. Consider background* $(X_1^{(1)},X_2^{(1)}) = (0,0)$ *and foreground* $(X_1^{(2)},X_2^{(2)}) = (1,1)$. *Under backpropagation-based approaches such as DeepLIFT, LRP, and DeepSHAP, the source nodes 1 and 2 receive an attribution of* $(1/2,1/2)$ *in model 1 and* $(1/3,2/3)$ *in model 2. This illustrates their lack of robustness. On the other hand, our top-down approach attributes*

$(\pi_1^{RSV}, \pi_2^{RSV}) = (1/2, 1/2)$ *in both the models. (Recall from §4 that (a) RSV attributes to edges* $[\pi_{ij}^{RSV}]_{(i,j)\in\mathsf{E}}$ *and the node attributions are defined as* $\pi_j^{RSV} := \sum_{i\in\mathsf{P}_j} \pi_{ij}^{RSV}$ *for all* $j \in \mathsf{N} \setminus \{0\}$*, and (b) RSV adds a super-source node 0. Hence,* $\pi_1^{RSV} = \pi_{01}^{RSV}$ *and* $\pi_2^{RSV} = \pi_{02}^{RSV}$*. The computation of* π_{01}^{RSV} *and* π_{02}^{RSV} *is straightforward in both the models.)*



Model 1 Model 2

Figure D. The two models corresponding to Example A. There are two source variables in both the models, X_1 and X_2. In model 1, the structural equations are as follows: $X_3 = X_1$, $X_4 = X_2$, and $Y = X_3X_4$. On the other hand, in model 2, $X_3 = X_1$, $X_4 = \sqrt{X_2}$, $X_5 = \sqrt{X_2}$ and $Y = X_3X_4X_5$.

F.3. Proof of Proposition 2

Proposition 2. RSV obeys DM and ASI.

Proof of Proposition 2. To see why RSV obeys DM, recall the notation from Definition 6 and observe that

$$\begin{aligned}
\pi_i^{\text{RSV}} &= \pi_{0i}^{\text{RSV}} \\
&= \sum_{E_0 \subseteq \mathsf{E}_0 \setminus \{(0,i)\}} w_{\mathsf{E}_0}(E_0) \times \{v_0(E_0 \cup \{(0,i)\}) - v_0(E_0)\} \\
&= \sum_{E_0 \subseteq \mathsf{E}_0 \setminus \{(0,i)\}} \underbrace{w_{\mathsf{E}_0}(E_0)}_{\geq 0} \times \left\{ \underbrace{g(\mathbf{X}_{N_0}^{(2)}, x_i, \mathbf{X}_{\mathsf{N}_0 \setminus \{N_0 \cup \{i\}\}}^{(1)})}_{\text{monotone in } x_i} - \underbrace{g(\mathbf{X}_{N_0}^{(2)}, \mathbf{X}_{\mathsf{N}_0 \setminus N_0}^{(1)})}_{\text{constant w.r.t. } x_i} \right\}.
\end{aligned}$$

$N_0 := \{k : (0,k) \in E_0\}$ denotes the children nodes of 0 that are present in E_0. To be thorough, we should use $N_0(E_0)$ but we omit the dependence on E_0 for conciseness. The proof for ASI follows from first principles as well. □

G. Further Details on RSV's Connection to Existing Node-based Approaches

In this appendix, we provide further details on RSV's connection to existing node-based approaches. In particular, we provide a proof of Proposition 3 in §G.1 (hence, verifying RSV's connection to distal ASV), followed by highlighting RSV's connection to ISV and proximate ASV in §G.2.

G.1. Proof of Proposition 3

Proposition 3. Source nodes receive the same attribution under RSV and distal ASV, i.e., $\pi_j^{\text{RSV}} = \pi_j^{\text{ASV}}\ \forall j \in \mathsf{N}_0$.

Proof of Proposition 3. It suffices to show that the games underlying $[\pi_j^{\text{RSV}}]_{j\in\mathsf{N}_0}$ and $[\pi_j^{\text{ASV}}]_{j\in\mathsf{N}_0}$ are equivalent. Recall from §5 that $g(\cdot)$ denotes the mapping from source nodes to the output, i.e., $Y = g(\mathbf{X}_{\mathsf{N}_0})$ (assuming all downstream edges $(\mathsf{E}_1, \ldots, \mathsf{E}_n)$ to be active). For distal ASV, it follows from the definition in Appendix A.3 that the underlying game is as follows. Set of players is N_0. Given coalition $N_0 \subseteq \mathsf{N}_0$, characteristic function is $g(\mathbf{X}^{(2)}_{N_0}, \mathbf{X}^{(1)}_{\mathsf{N}_0\setminus N_0})$. For $j \in \mathsf{N}_0$, distal ASV π_j^{ASV} is the corresponding Shapley value of this game. For RSV, it follows from the definition in §4.2 that the underlying game is as follows. Set of players is E_0 (outgoing edges of the super-source node 0). Given coalition $E_0 \subseteq \mathsf{E}_0$, characteristic function is $g(\mathbf{X}^{(2)}_{N_0}, \mathbf{X}^{(1)}_{\mathsf{N}_0\setminus N_0})$, where $N_0 := \{k : (0,k) \in E_0\}$ denotes the children nodes of 0 that are present in E_0. To be thorough, we should use $N_0(E_0)$. For $j \in \mathsf{N}_0$, RSV π_{0j}^{RSV} is the corresponding Shapley value of this game and by definition, $\pi_j^{\text{RSV}} = \pi_{0j}^{\text{RSV}}$. Clearly, the two games are indentical and hence, $\pi_j^{\text{RSV}} = \pi_j^{\text{ASV}}\ \forall j \in \mathsf{N}_0$. This completes the proof. □

G.2. Connection Between RSV and ISV / Proximate ASV

In this subsection, we establish a connection between our flow-based RSV and the node-based ISV (Appendix A.1) / proximate ASV (Frye et al., 2020b). Recall from Appendix A that ISV and proximate ASV attribute all the value to the parent nodes P_{n+1} of the output. However, such an attribution is only apt if the graph is "flat" since otherwise, it violates source efficiency. Interestingly, RSV recovers both ISV and proximate ASV under "flat" graphs, which we define next.

Definition A (Flat graph)**.** *We say a graph* G *(without super-source node 0) is flat if there are no edges between the nodes in* N*. In other words, each edge is directed to the output node.*

An example is provided in Figure E and we state our claim in Proposition B.

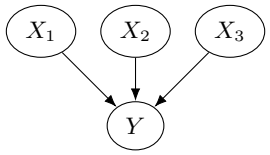


Figure E. An example of a flat graph.

Proposition B. *Suppose the graph* G *is flat. Then, RSV recovers ISV, i.e.,* $\pi_j^{RSV} = \pi_j^{ISV}\ \forall j \in \mathsf{N}$.

The notation $[\pi_j^{\text{ISV}}]_{j\in\mathsf{N}}$ is defined in Appendix A.1. It is straightforward to prove Proposition B. In particular, it is easy to verify that if the graph is flat, then the game corresponding to ISV (defined in Appendix A.1) is equivalent to the node 0 game in RSV (defined in §4.2). Furthermore, since proximate ASV is the same as ISV given our problem definition from §2, equivalence between RSV and proximate ASV under flat graphs follows directly.

H. Details on the Numerics Corresponding to Example 3

In this appendix, we provide details on the numerics corresponding to the causal unfairness example (Example 3) from §6. We first discuss the underlying data generating process (§H.1), followed by our estimation procedure (§H.2). We then discuss our computation of RSV in the estimated probabilistic graphical model (§H.3). Finally, we present some sensitivity analysis (§H.4).

H.1. Data Generating Process

Recall that the true graph is as in Figure F. The data generating process is as follows. Sensitive attribute $X_1 \in [0,1]$ is drawn from a Uniform$[0,1]$ distribution. Test score $X_2 \in \mathbb{R}$ follows a standard normal distribution, i.e., Normal$(0,1)$. If $X_1 \leq 1/2$, then the department choice X_3 equals 0 w.p. 4/5 and 1 w.p. 1/5 whereas if $X_1 > 1/2$, then X_3 equals 0 w.p. 1/5 and 1 w.p. 4/5. Similarly, if $X_1 \leq 1/2$, then the unreported referral X_4 equals 0 w.p. 4/5 and 1 w.p. 1/5 whereas if $X_1 > 1/2$, then X_4 equals 0 w.p. 1/5 and 1 w.p. 4/5. Hence, an applicant with a higher value of the sensitive attribute (X_1) is more likely to apply to department 1 and is more likely to have a referral. The admit outcome Y is a Bernoulli random variable with mean equal to $\Phi(a_2X_2 + a_3X_3 + a_4X_4)$, where $\Phi(\cdot)$ denotes the standard normal CDF and $(a_2, a_3, a_4) \in \mathbb{R}^3_+$ are the probit weights. By construction, a_4 captures the level of unfair influence.

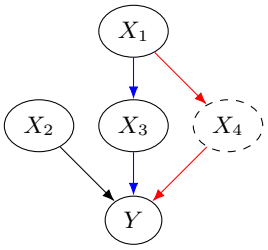


Figure F. True graph for Example 3.

The three parameters in our data generating process are the probit weights (a_2, a_3, a_4). We fix $a_2 = a_3 = 1$ and generate multiple datasets by varying $a_4 \in \{0, 0.5, \ldots, 6\}$. For each of the 13 values of a_4, we generate 100 datasets by using 100 different seeds ($s = 1, \ldots, 100$). Hence, we generate a total of 1300 datasets. In each dataset, we sample 1000 data points $\{(x_{1i}, x_{2i}, x_{3i}, x_{4i}, y_i)\}_{i=1}^{1000}$ using the process defined above.

H.2. Estimation Procedure

For a given $a_4 \in \{0, 0.5, \ldots, 6\}$, we estimate a separate model for each of the 100 datasets to account for the randomness due to finite data. Recall that the observed graph is as in Figure G, i.e., the referral data is missing. Accordingly, consider a dataset without the unobserved variable X_4: $\{(x_{1i}, x_{2i}, x_{3i}, y_i)\}_{i=1}^{1000}$.

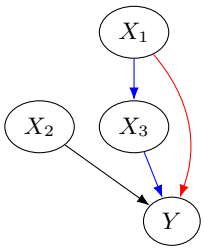


Figure G. Observed graph for Example 3.

We use this dataset to estimate two quantities: (1) the distribution of X_3 as a function of X_1 and (2) the distribution of Y as a function of (X_1, X_2, X_3). To do so, we fit a probit regression model at each of the nodes. Denote the estimated coefficient of X_1 in the model at X_3 by $\hat{b}_{13}$ and the estimated coefficients of (X_1, X_2, X_3) in the model at Y by $(\hat{b}_{15}, \hat{b}_{25}, \hat{b}_{35})$.

Accordingly, our estimated distribution for X_3 (as a function of X_1) is summarized by the following probit model:

$$\hat{X}_3 = \begin{cases} 1 \text{ if } \hat{b}_{13} X_1 + U_3 > 0 \\ 0 \text{ if } \hat{b}_{13} X_1 + U_3 \leq 0 \end{cases} \tag{H.7a}$$

$$U_3 \sim \text{Normal}(0, 1). \tag{H.7b}$$

Similarly, our estimated distribution for Y (as a function of (X_1, X_2, X_3)) is summarized by the following probit model:

$$\hat{Y} = \begin{cases} 1 \text{ if } \hat{b}_{15} X_1 + \hat{b}_{25} X_2 + \hat{b}_{35} X_3 + U_5 > 0 \\ 0 \text{ if } \hat{b}_{15} X_1 + \hat{b}_{25} X_2 + \hat{b}_{35} X_3 + U_5 \leq 0 \end{cases}$$

$$U_5 \sim \text{Normal}(0, 1).$$

U_3 and U_5 denote the error terms. Note that given (X_1, X_2, X_3), the expected value of $\hat{Y}$ equals

$$\hat{\mu} := \Phi(\hat{b}_{15} X_1 + \hat{b}_{25} X_2 + \hat{b}_{35} X_3), \tag{H.8}$$

which denotes the admit probability (under our estimated model).

H.3. RSV Computation

Consider the estimated model corresponding to an arbitrary $a_4 \in \{0, 0.5, \ldots, 6\}$ and an arbitrary seed $s \in \{1, \ldots, 100\}$. To understand if the sensitive attribute (X_1) has an unfair influence on the outcome, we consider the following two applicants: $(X_1^{(1)}, X_2^{(1)}) = (0, 0)$ (background) and $(X_1^{(2)}, X_2^{(2)}) = (1, 0)$ (foreground), i.e., different value of X_1 but same score. Observe that the framework presented in the paper is for a deterministic model and our estimated model in this example is probabilistic. To attribute in this probabilistic model, we use structural equations model with errors as follows. We insert the error term U_3 from (H.7) as a parent of node 3 (see Figure H). Furthermore, we replace the output node $Y \in \{0, 1\}$ by the estimated model's admit probability $\hat{\mu} \in [0, 1]$ (see (H.8)) since we are interested in understanding the difference in probabilities. We can choose to understand the difference in $\hat{Y}$ too by simply adding the error term U_5 as a parent of the output node.

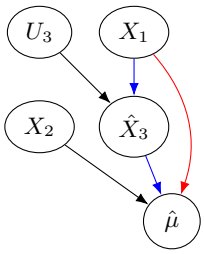


Figure H. Estimated graph with error term. U_3 and $\hat{X}_3$ as in (H.7). Output $\hat{\mu}$ as in (H.8). Given U_3, the model is deterministic.

Figure H corresponds to a model in which the structural equations F are deterministic. In particular, given (X_1, X_2) and a sample of $U_3 \sim \text{Normal}(0, 1)$, (X_1, U_3) map deterministically to $\hat{X}_3$ via (H.7) and $(X_1, X_2, \hat{X}_3)$ map deterministically to $\hat{\mu}$ via (H.8). Defining $Y^{(1)} := \mathbb{E}_{U_3}[\hat{\mu} \mid (X_1^{(1)}, X_2^{(1)})]$ and $Y^{(2)} := \mathbb{E}_{U_3}[\hat{\mu} \mid (X_1^{(2)}, X_2^{(2)})]$, we are interested in understanding $Y^{(2)} - Y^{(1)}$, which denotes the difference in the expected admit probabilities between applicant 2 and applicant 1 (under the estimated model). To compute RSV, we generate 1000 samples of U_3 and recycle our deterministic machinery since given U_3, the system is deterministic. Hence, we compute RSV for each sample of U_3 and output the average RSV (averaged over U_3 samples), which is justified since RSV obeys the linearity axiom (Theorem 1). Note that we use the same sample of U_3 for both the background and the foreground, i.e., $U_3^{(1)} = U_3^{(2)}$. Hence, the edge from U_3 to $\hat{X}_3$ satisfies flow nullity.

In Figure 11 of the paper, we show the RSV corresponding to each value of a_4, where the error bars ($\pm$ 1 standard deviation) correspond to the uncertainty over 100 seeds. Note that given a seed (and hence, an estimated model) and a sample of U_3, we compute RSV exactly (as opposed to a Monte-Carlo estimation), which is possible in this relatively small graph.

H.4. Sensitivity Analysis

Though model estimation is not our focus, we perform the following robustness check. We change the data generating process to a sigmoid, but fit a probit model as before (model mismatch). In particular, we use the following data generating process. Sensitive attribute (X_1), test score (X_2), department choice (X_3), and unreported referral (X_4) follow the same distribution as in §H.1. The admit outcome Y follows a Bernoulli distribution but with mean equal to $\sigma(a_2X_2 + a_3X_3 + a_4X_4)$, where $\sigma(x) := 1/(1 + \exp(-x))$ denotes the sigmoid function and $(a_2, a_3, a_4) \in \mathbb{R}^3_+$ are the sigmoid weights.

As before, we fix $a_2 = a_3 = 1$ and generate multiple datasets by varying $a_4 \in \{0, 0.5, \ldots, 6\}$ and use 100 seeds for each value of a_4. Our estimation procedure and the RSV computation remains the same as in §H.2 and §H.3, respectively. In Figure I, we show attributions to the fair $(X_1 \to X_3 \to Y)$ and the unfair $(X_1 \to Y)$ channels as a function of a_4. Similar to Figure 11, attribution to the unfair channel increases with a_4, which is logical. Furthermore, the unfair channel receives zero attribution when there is no unresolved discrimination ($a_4 = 0$) and the same attribution as the fair channel when the two exert the same influence ($a_3 = a_4 = 1$).

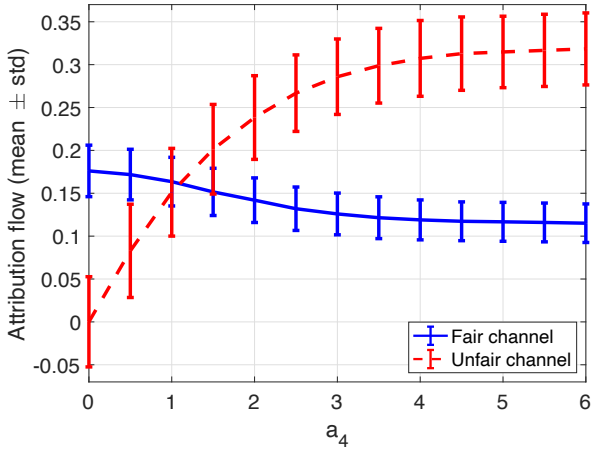


Figure I. Sensitivty analysis for Example 3. We change the data generating process to a sigmoid but fit a probit model. As before, attribution to the unfair channel increases with the level of unfair influence a_4.

References

Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers. In *International Conference on Artificial Neural Networks*, pp. 63–71. Springer, 2016.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to Algorithms*. MIT press, 2009.

Datta, A., Sen, S., and Zick, Y. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 598–617. IEEE, 2016.

Dhamdhere, K., Sundararajan, M., and Yan, Q. How Important Is a Neuron? *arXiv preprint arXiv:1805.12233*, 2018.

Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., and Feige, I. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020a.

Frye, C., Rowat, C., and Feige, I. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33, 2020b.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. *Advances in Neural Information Processing Systems*, 33, 2020.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020.

Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

Pearl, J. Direct and Indirect Effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 32: 411–420, 2001.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Shapley, L. S. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3145–3153. PMLR, 2017.

Strumbelj, E. and Kononenko, I. An Efficient Explanation of Individual Classifications using Game Theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

Štrumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

Sun, Y. and Sundararajan, M. Axiomatic Attribution for Multilinear Functions. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, pp. 177–178, 2011.

Sundararajan, M. and Najmi, A. The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9269–9278. PMLR, 2020.

Sundararajan, M., Dhamdhere, K., and Agarwal, A. The Shapley Taylor Interaction Index. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9259–9268. PMLR, 2020.

Wang, J., Wiens, J., and Lundberg, S. Shapley Flow: A Graph-based Approach to Interpreting Model Predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 721–729. PMLR, 2021.