Title

Axiomatic effect propagation in structural causal models

Authors

Singal, Raghav
Michailidis, George

Publication Date

2024

Journal of Machine Learning Research 25 (2024) 1-71 Submitted 3/22; Revised 2/24; Published 3/24

Axiomatic effect propagation in structural causal models

Raghav Singal SINGAL@DARTMOUTH.EDU
Operations and Management Science
Tuck School of Business at Dartmouth College
Hanover, NH 03755, USA

George Michailidis GMICHAIL@UCLA.EDU
Department of Statistics and Data Science
University of California, Los Angeles
Los Angeles, CA 90095, USA

Editor: Mladen Kolar

Abstract

We study effect propagation in a causal directed acyclic graph (DAG), with the goal of providing a *flow-based decomposition* of the effect (i.e., change in the outcome variable) as a result of changes in the source variables. We first compare various ideas on causality to quantify effect propagation, such as direct and indirect effects, path-specific effects, and degree of responsibility. We discuss the shortcomings of such approaches and propose a flow-based methodology, which we call *recursive Shapley value* (RSV). By considering a broader set of counterfactuals than existing methods, RSV obeys a *unique* adherence to four desirable flow-based axioms. Further, we provide a general path-based characterization of RSV for an arbitrary non-parametric structural equations model (SEM) defined on the underlying DAG. Interestingly, for the special class of linear SEMs, RSV exhibits a simple and tractable characterization (and hence, computation), which recovers the classical method of path coefficients and is equivalent to path-specific effects. For non-parametric SEMs, we use our general characterization to develop an unbiased Monte-Carlo estimation procedure with an exponentially decaying sample complexity. We showcase the application of RSV on two challenging problems on causality (causal overdetermination and causal unfairness).

Keywords: Structural causal model, attribution, effect decomposition, Shapley value, Monte-Carlo estimation

1. Introduction

A fundamental problem in causal analysis is to quantify effect propagation; namely, given a change in the outcome variable (*effect*) that is driven by changes at the source nodes in a causal graph, how does the effect *flow* through the graph? This question relates to various foundational ideas on causality, including direct and indirect effects (Pearl, 2001), path analysis (Wright, 1934; Duncan, 1966; Goldberger, 1972) and path-specific effects (Pearl, 2001), degree of responsibility (Chockler and Halpern, 2004), explanations (Halpern and Pearl, 2005), mediation analysis (Baron and Kenny, 1986; MacKinnon et al., 2007; Imai et al., 2010), information flow (Ay and Polani, 2008), causal influence (Janzing et al., 2013), and effect decomposition (Kitagawa, 1955; Blinder, 1973; Oaxaca, 1973; Alwin and

Hauser, 1975; Fortin et al., 2011; Chernozhukov et al., 2013). In this work, we study the effect propagation problem on a causal directed acyclic graph (DAG) with arbitrary non-parametric structural equations. Such a fairly general framework captures a broad spectrum of causal systems and has found applications in political science (Imai et al., 2011), epidemiology (VanderWeele et al., 2012), ecology (Weible et al., 2004) and wage inequality (Firpo and Pinto, 2016), to name a few.

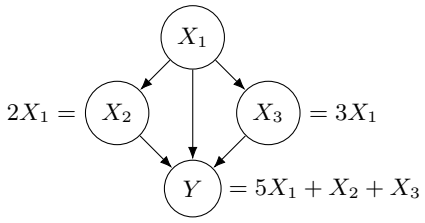


Figure 1: An example of a causal graph with linear structural equations. The source variable X_1 (without parents) is set exogenously, $X_2 = 2X_1$, $X_3 = 3X_1$, and $Y = 5X_1 + X_2 + X_3$.

To build intuition, consider the causal graph in Figure 1, where the underlying structural equations are linear and deterministic (no error / noise terms present). Suppose the source variable X_1 changes from a value of 0 (*background*) to 1 (*foreground*), causing the outcome variable Y to change from 0 to 10. How does this effect (10) of the change in X_1 propagate through the DAG? Given the linear equations, an intuitive answer is to express the effect in terms of the corresponding edge weights / path coefficients (Wright, 1918, 1934; Alwin and Hauser, 1975):

- Path $X_1 \to X_2 \to Y$ carries 2 units of the total effect.
- Path $X_1 \to Y$ carries 5 units of the total effect.
- Path $X_1 \to X_3 \to Y$ carries 3 units of the total effect.

This decomposition of $(2, 5, 3)$ along the three channels sums up to the total effect of 10, which is desirable. However, quantifying effect propagation is not as straightforward when the underlying structural equations are non-linear and earlier work, including Wright (1934) and Alwin and Hauser (1975), assumes linear equations. Motivated by this shortcoming, Pearl (2001) introduced natural direct and indirect effects, which are well-defined for non-linear equations and recover the above-mentioned seemingly natural decomposition for linear equations. However, the natural direct and indirect effects do not necessarily sum to the total effect, which is rather "strange" (quoted from Pearl (2001)).

To see this, consider the causal graph in Figure 2 with the following non-linear equations:

$$\begin{aligned} X_1 &\text{ exogenous} \\ X_2 &= X_1 \\ Y &= X_1 X_2. \end{aligned}$$

In particular, the outcome involves an interaction of X_1 and X_2. Similar to the previous example, suppose the source variable X_1 changes from 0 to 1, causing Y to change from 0 to 1. How does this effect (change in Y) propagate through the two channels $X_1 \to X_2 \to Y$

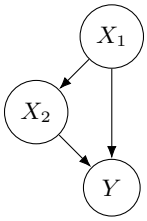


Figure 2: An example of a two-channeled graph. $X_1 \to X_2 \to Y$ is the *indirect* channel with X_2 being the *mediator* and $X_1 \to Y$ is the *direct* channel.

(indirect) and $X_1 \to Y$ (direct)? As we elaborate in §3, both the natural direct and indirect effects equal 0, implying such a decomposition explains none of the total effect. The key reason driving this behavior is that direct (indirect) effect only considers one counterfactual, i.e., when computing effect propagated through the direct (indirect) channel, edges on the indirect (direct) channel are assumed to be inactive. As we will see, by allowing for a broader set of counterfactuals, the proposed approach will overcome this limitation and always explain 100% of the effect, irrespective of the underlying equations being linear or non-linear.

A further challenge to quantify effect propagation is to allow for the possibility of *multiple source variables changing simultaneously* and letting the effect flow through *multiple mediators*, as opposed to a single block of mediators (Zhang and Bareinboim (2018b) call it "en bloc"), as shown in Figure 3. For instance, in domains such as school admissions, credit approval, and wage discrimination, it is often of interest to quantify causal unfairness (Kilbertus et al., 2017; Kusner et al., 2017; Chiappa, 2019), which can be driven by simultaneous changes in *multiple protected attributes*. A desirable property in such a setting is flow conservation, i.e., the total effect (due to changes at all the source nodes) flows down the DAG while obeying flow-in equals flow-out at each node (formally defined in §2). To the best of our knowledge, most state-of-the-art approaches are either not well-defined in such a setting or fail to obey flow conservation, primarily due to their inability to account for multiple counterfactuals, which is important in applications with multiple moving pieces (e.g., multiple source variables changing or multiple variables mediating the effect). From both a foundational and a practical point-of-view, this is unsatisfying.

One possibility to characterize the direct effect of multiple source variables changing simultaneously is via k-way interaction effects; i.e., let the changed values of race and sex propagate only through the red direct links in Figure 3 (2-way effect) and compute the corresponding difference in the outcome variable (income). However, such a way of thinking leads to *exponentially* many effects (one for each possible combination of the source nodes). In addition, when one sums up all of these interaction effects, they do not necessarily equal the total effect, which is again undesirable. To see this, consider the Figure 4 setup.

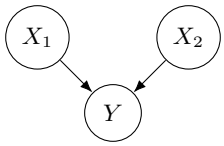


Figure 4: A simple graph to illustrate 2-way interaction effects.

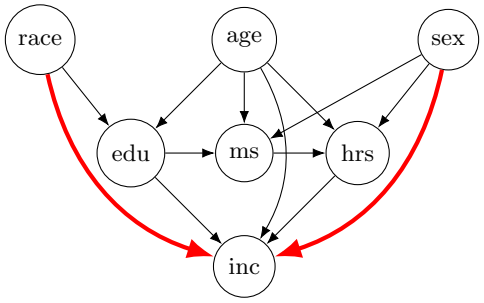


Figure 3: A DAG with multiple source variables (*race*, *age*, and *sex*) and multiple mediators (*edu*, *ms*, and *hrs*). We re-visit this DAG in §7 to illustrate how RSV quantifies causal unfairness (effect propagated via the "unfair" thick red edges) when multiple sensitive attributes (*race* and *sex*) change simultaneously and the underlying dynamics are non-linear.

Let $y = f(x_1, x_2)$ with the source variables (x_1, x_2) changing from $(0, 0)$ to $(1, 1)$. The 1-way effects correspond to the two direct effects (one for X_1 and one for X_2):

$$\theta_1 := f(1, 0) - f(0, 0)$$
$$\theta_2 := f(0, 1) - f(0, 0).$$

There is one 2-way effect, which corresponds to changing both the source variables:

$$\theta_{12} := f(1, 1) - f(0, 0).$$

There are no k-way effects for $k \geq 3$ since there are only 2 source variables. On adding all of these effects, we get

$$\theta_1 + \theta_2 + \theta_{12} = f(1, 0) + f(0, 1) + f(1, 1) - 3f(0, 0).$$

However, the latter does not necessarily equal the total effect, which is $f(1, 1) - f(0, 0)$. Naturally, θ_{12} equals the total effect, but such a view provides no insight into how to decouple this interaction effect in terms of the underlying contributing factors. Even in a setting as simple as Figure 4 (i.e., no mediators), being able to decouple the total effect is important and analogous to Shapley value, which is a well-accepted concept in economics. In particular, as discussed in Friedenberg and Halpern (2019), one potential application of such a decomposition is in "legal ascription of responsibility". In fact, this has been discussed in the legal literature as well, where Ferey and Dehez (2016) motivate it as "fairness of the apportionment" from an *"ex post perspective"*. To dive deeper, consider the "Tragedy of the Commons" example from Friedenberg and Halpern (2019):

> 100 fishermen live by a lake. If at least 10 of them overfish this year, then the entire fish population of the lake will die out. Each fisherman overfishes. By the end of the year, the entire fish population has died out.

Here, we can define k-way interaction effects for $k \in \{1, \cdots, 100\}$. All of them will equal 0 for $k < 10$ and equal 1 for $k \geq 10$. However, such a view gives no insight into individual-

level responsibility.[1] On the other hand, the Shapley value assigns a responsibility equal to 1/100 to each fisherman, which given their symmetric behavior, is a fair apportionment from an *ex post perspective*.

This work aims to develop a principled framework to quantify effect propagation in causal DAGs with linear / non-linear structural equations. While obeying intuitive properties such as the effect decomposition adding up to the total effect (*source efficiency*) and obeying flow conservation, our framework is flexible enough to allow for the possibility of simultaneous changes at an arbitrary number of source nodes. Note that such properties represent one possible way to operationalize the notion of fair apportionment from an ex post perspective. We reckon there might be other sets of properties[2] (and context-specific knowledge could be useful in determining which properties are relevant), but we show uniqueness to four desirable axioms under the posited properties. While doing so, we do not impose any additional assumption on the graph topology, besides being a DAG, and hence, allow for multiple mediators. When the graph has no mediators, our framework recovers Shapley value. In addition, it goes a step beyond and is able to decouple the total effect in more involved settings (i.e., with mediators), while accounting for a multitude of non-trivial interaction effects. As such, our proposal can be seen as a generalization of SV to graphs with mediators. The main contributions of this work are as follows.

First, we use the language of causal graphs to formally state the underlying attribution / effect propagation problem. The framework is general enough to capture an arbitrary causal graph as long as the underlying graph topology contains no cycles (DAG). We use our problem definition to highlight some intuitive properties (source efficiency and flow conservation) that any methodology to quantify effect propagation should satisfy.

Second, we develop a flow-based methodology to quantify effect propagation, which we call *recursive Shapley value* (RSV). As illustrated in Figure 5, RSV employs a top-down principle by first attributing to the source nodes and then quantifying how the effect flows down the DAG. Such a top-down principle generalizes a number of existing node-based attribution techniques. On the foundational front, we establish that by considering a broad spectrum of counterfactuals, RSV *uniquely* obeys a set of natural flow-based axioms (Theorem 4). We further show that adherence to such axioms result in RSV satisfying a number of intuitive properties (implementation invariance, sensitivity, monotonicity, and affine scale invariance) discussed in recent literature (Sundararajan et al., 2017; Sundararajan and Najmi, 2020). On the implementation front, we provide a closed-form characterization for RSV under both linear (Theorem 6) and non-linear [3] (Theorem 8) SEMs. Our characterization for the linear model recovers the classical method of path coefficients (Wright, 1918, 1934) and is equivalent to path-specific effects advocated by Pearl (2001), whereas for non-parametric (possibly non-linear) models, the decomposition provided by RSV sums up

1. Of course, this is not meant to diminish the importance of such interaction effects from an *ex ante* perspective, where one is interested in the impact of (possibly multiple) interventions on the outcome variable. However, that is not our objective in this work.
2. In fact, Friedenberg and Halpern (2019) discuss this caveat as well: "words like "blame" have a wide variety of nuanced meanings in natural language. While we think that the notion that we are trying to capture is useful, it corresponds at best to only one way that the word "blame" is used by people." So do Chockler and Halpern (2004): "We cannot say that a definition is "right" or "wrong"."
3. All results for the non-linear case hold more generally for an arbitrary non-parametric SEM, as it will become clear in §5.

to the total effect (an implication of the more fundamental flow-based axioms). We use our characterization for the non-parametric models to develop a Monte-Carlo estimation scheme and analyze the quality of our estimator (Proposition 10), which is illustrated through a numerical study. Subsequently, we show how RSV facilitates non-linear mediation analysis to quantify causal overdetermination and causal unfairness and contrast it with existing ideas such as degree of responsibility and path-specific effects.

Third, we provide a comprehensive view of existing approaches to effect propagation. In particular, we show how various existing techniques in the causality literature operate and shed light on their shortcomings via simple examples. Interestingly, most existing approaches do not result in the decomposition adding up to the total effect. To the best of our knowledge, this is the first work to understand the effect propagation literature using a common framework.

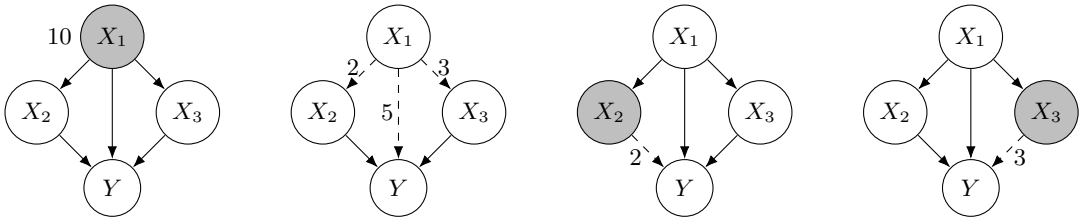


Figure 5: Illustrating RSV on the causal graph from Figure 1. First (see the leftmost plot), RSV attributes the total effect of 10 to the source node 1. Second (see plot #2), node 1's attribution (10) is decomposed among its outgoing edges by evaluating their contributions via the following counterfactual questions: how much attribution would node 1 have received had edge $(1, 2)$ not propagated the change at node 1? This results in edge $(1, 2)$ receving a flow of 2, edge $(1, Y)$ a flow of 5, and edge $(1, 3)$ a flow of 3. Third (see plot #3), to propagate down the flow node 2 receives, RSV evaluates the attribution node 2 would have received had edge $(2, Y)$ not communicated the change at node 2. Finally (see the rightmost plot), RSV repeats the exercise at node 3.

Outline The remainder of the paper is organized as follows. In §2, we formally state the effect propagation problem and use such formalism to discuss existing approaches and their shortcomings in §3. We then propose the recursive Shapley value (RSV) approach and establish its desirability in §4, in addition to discussing its connections with the existing works. We characterize RSV for both linear and non-parametric (possibly non-linear) structural equations models and discuss its exact computation in §5, followed by developing a Monte-Carlo estimation scheme in §6. We then showcase applications in §7 and discuss concluding remarks in §8. Technical details are deferred to the appendices.

2. Problem formulation

We start by formulating the problem under consideration. First, we define the causal graph of interest (§2.1) and subsequently, formally present the effect propagation problem (§2.2).

2.1 Causal graph setting

Let $\mathsf{G} = (\mathsf{N}^+, \mathsf{E})$ denote the underlying directed acyclic graph (DAG), with node set $\mathsf{N}^+ := \{1, \ldots, n, n+1\}$ and edge set E. A node without parents is called a *source node* and we allow for multiple such nodes, with N_0 denoting their collection. The outcome variable corresponds to the *sink node* (node without children) and we let it correspond to node $n+1$ without loss of generality. We focus on a single sink node and note that it is straightforward to generalize our framework to multiple sink nodes (multivariate outcome). We let $\mathsf{N} := \mathsf{N}^+ \setminus \{n+1\} = \{1, \ldots, n\}$ be the set of non-sink nodes. Given an edge (i, j) in E, node i is referred to as a *parent* of node j whereas node j is a *child* of node i. For an arbitrary node $j \in \mathsf{N}^+$, we define P_j to be the set of its parents:

$$\mathsf{P}_j := \{i \in \mathsf{N} : (i, j) \in \mathsf{E}\} \ \forall j \in \mathsf{N}^+.$$

Naturally, nodes in N_0, i.e., source nodes, have no parents: $\mathsf{P}_j = \emptyset$ for all $j \in \mathsf{N}_0$. Similarly, for node $i \in \mathsf{N}^+$, we define C_i to be the set of its children:

$$\mathsf{C}_i := \{j \in \mathsf{N}^+ : (i, j) \in \mathsf{E}\} \ \forall i \in \mathsf{N}^+,$$

and note that node $n+1$, i.e., the sink node, has no children, i.e., $\mathsf{C}_{n+1} = \emptyset$.

Corresponding to each non-outcome node $i \in \mathsf{N}$ is a variable [4] $X_i \in \mathbb{R}$ whereas the variable corresponding to the outcome node $n+1$ is denoted by $Y \in \mathbb{R}$. For ease of notation, we will sometimes denote Y by X_{n+1}. We define the vector of all non-outcome variables as $\mathbf{X} := (X_1, \ldots, X_n)$ and a subset of this collection as $\mathbf{X}_N := (X_i)_{i \in N}$ for $N \subseteq \mathsf{N}$. Each non-source variable X_i is a function of its parents $\mathbf{X}_{\mathsf{P}_i}$ and a corresponding noise variable U_i, leading to the following *structural equations* $\mathsf{F} := [f_i(\cdot)]_{i \in \mathsf{N}^+ \setminus \mathsf{N}_0}$ in the DAG:

$$x_i = f_i(\boldsymbol{x}_{\mathsf{P}_i}, u_i) \ \forall i \in \mathsf{N}^+ \setminus \mathsf{N}_0.$$

Non-capitalized symbols are used to denote the realizations of the corresponding random variables. The noise variables $\mathbf{U} := (U_i)_{i \in \mathsf{N}^+ \setminus \mathsf{N}_0}$ are drawn from some exogenous distribution D, which is *independent* of $(\mathbf{X}, Y)$. The values of the source variables $\mathbf{X}_{\mathsf{N}_0}$ are set *exogenously and independently* of each other. *Mediating variables* or *mediators* (variables that are neither source nor outcome, i.e., $\mathbf{X}_{\mathsf{N} \setminus \mathsf{N}_0}$) propagate the source nodes values down to the outcome node $n+1$. Given source variables $\mathbf{X}_{\mathsf{N}_0} = \boldsymbol{x}_{\mathsf{N}_0}$ and noise variables $\mathbf{U} = \boldsymbol{u}$, we denote the observed outcome at node $n+1$ by

$$f(\boldsymbol{x}_{\mathsf{N}_0}, \boldsymbol{u}).$$

Clearly, conditioned on the noise $\mathbf{U} = \boldsymbol{u}$, the model is deterministic and the source variables realizations $\mathbf{X}_{\mathsf{N}_0} = \boldsymbol{x}_{\mathsf{N}_0}$ are sufficient to determine the outcome. Hence, we denote the expected outcome by

$$g(\boldsymbol{x}_{\mathsf{N}_0}) := \mathbb{E}[f(\boldsymbol{x}_{\mathsf{N}_0}, \boldsymbol{u})],$$

4. Our framework is well-defined and all our results go through for multi-dimensional variables as well, i.e., each non-outcome node $i \in \mathsf{N}$ corresponding to a multi-dimensional variable. However, to be consistent with the literature on causal graphs, we treat each variable as being one-dimensional.

where the expectation is over the distribution of $\boldsymbol{u}$. Similar to Definition 7.1.1 in Pearl (2009), the *structural causal model* (or *causal graph*) is denoted by $\mathsf{M} := (\mathsf{G}, \mathsf{F}, \mathsf{D})$ [5]. Given our focus on effect propagation, we assume M to be fixed and known throughout the paper, as do related notions such as direct, indirect, and path-specific effects (Pearl, 2001) and degree of responsibility (Chockler and Halpern, 2004).

2.2 The effect propagation problem

Next, we use the setup previously defined to formally state the effect propagation problem. Similar to Pearl (2001), we are interested in two realizations of the source variables $\mathbf{X}_{\mathsf{N}_0}$:

$$\boldsymbol{x}^{(1)}_{\mathsf{N}_0} := (x^{(1)}_i)_{i \in \mathsf{N}_0} \qquad (\textit{background})$$
$$\boldsymbol{x}^{(2)}_{\mathsf{N}_0} := (x^{(2)}_i)_{i \in \mathsf{N}_0}. \qquad (\textit{foreground})$$

There is an underlying temporal / interventional aspect such that the background value changes to the foreground value. To tie our terminology with existing literature, note that our "background" is similar to the "reference" value x^* in Pearl (2001). Given noise realization $\mathbf{U} = \boldsymbol{u}$, the structural causal model is deterministic and hence, the changes at source nodes dictate changes in the entire graph. The background value defines the baseline outcome

$$y^{(1)}_{\boldsymbol{u}} := f(\boldsymbol{x}^{(1)}_{\mathsf{N}_0}, \boldsymbol{u}),$$

and analogously

$$y^{(2)}_{\boldsymbol{u}} := f(\boldsymbol{x}^{(2)}_{\mathsf{N}_0}, \boldsymbol{u}).$$

The super-script in $y^{(t)}_{\boldsymbol{u}}$ captures the dependence on the source values $\boldsymbol{x}^{(t)}_{\mathsf{N}_0}$ for $t \in \{1, 2\}$ whereas the sub-script captures the dependence on the noise $\boldsymbol{u}$. Hence, the change in the outcome equals

$$\delta_{\boldsymbol{u}} := y^{(2)}_{\boldsymbol{u}} - y^{(1)}_{\boldsymbol{u}},$$

and the *expected* change equals

$$\delta := \mathbb{E}[\delta_{\boldsymbol{u}}] = \mathbb{E}[y^{(2)}_{\boldsymbol{u}} - y^{(1)}_{\boldsymbol{u}}] = \mathbb{E}[y^{(2)}_{\boldsymbol{u}}] - \mathbb{E}[y^{(1)}_{\boldsymbol{u}}] = y^{(2)} - y^{(1)},$$

where $y^{(t)} := \mathbb{E}[y^{(t)}_{\boldsymbol{u}}] = g(\boldsymbol{x}^{(t)}_{\mathsf{N}_0})$ for $t \in \{1, 2\}$ with the expectation over $\boldsymbol{u}$. Note that this delta effect, i.e., the (expected) change in the outcome, is *entirely* driven by the changes at the source nodes and their propagation through the DAG. This is because all the exogenous sources of variation are captured by the source nodes.

The question here is how to attribute the expected change in Y (i.e., δ) to the changes at the source nodes and their propagation through the edges of the graph. If the graph has

5. Note that Pearl (2009) defines a structural causal model (Definition 7.1.1) as the triplet $(\mathbf{V}, \mathsf{F}, \mathbf{U})$ where $\mathbf{V} := (\mathbf{X}, Y)$, whereas we define it as $(\mathsf{G}, \mathsf{F}, \mathsf{D})$. The two definitions are equivalent as we use G instead of $\mathbf{V}$ and D instead of $\mathbf{U}$.

only one node, i.e., $n = 1$, such a task is straightforward since the only node is completely responsible for the expected change and thus, receives 100% attribution, which propagates through the only edge in the graph. However, if the graph has multiple nodes, i.e., $n > 1$, then non-trivial interactions are possible. Given an arbitrary causal graph, is there a systematic way to disentangle such interactions and quantify effect propagation via *flow-based* attribution, i.e., the amount of effect each edge carries?

In order to quantify effect propagation, we introduce the following notation. We use $[\pi_i]_{i\in\mathsf{N}}$ to capture the attribution to nodes and $[\pi_{ij}]_{(i,j)\in\mathsf{E}}$ to denote the attribution to edges (*attribution flow*). The scale of both is in terms of δ. As discussed above, all the exogenous variation is captured by the source nodes. Hence, a desirable property is *source efficiency*:

$$\sum_{i\in\mathsf{N}_0} \pi_i = \delta.$$

In addition, it seems natural to ensure conservation of flow at each node, that is, *flow in equals flow out.*

Definition 1 (Flow conservation) *At each internal node $j \in \mathsf{N} \setminus \mathsf{N}_0$,*

$$\sum_{i\in\mathsf{P}_j} \pi_{ij} = \sum_{k\in\mathsf{C}_j} \pi_{jk}.$$

For source nodes,

$$\sum_{i\in\mathsf{N}_0} \sum_{j\in\mathsf{C}_i} \pi_{ij} = \delta.$$

Finally, at sink node,

$$\sum_{i\in\mathsf{P}_{n+1}} \pi_{i,n+1} = \delta.$$

Despite being a desirable feature, flow conservation does not map to a unique attribution flow $[\pi_{ij}]_{(i,j)\in\mathsf{E}}$ in general, making the task of attribution non-trivial. As mentioned earlier, consistent with the related literature (Pearl, 2001; Chockler and Halpern, 2004), the underlying causal graph $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$ is assumed to be fixed and known in this work. It is worth emphasizing that attribution, i.e., quantifying effect propagation, is a challenging problem in itself, which complements the ongoing developments in causal discovery (Peters et al., 2017; Glymour et al., 2019).

Remark 2 (Individual-level effect) *In the problem formulation above, we focus on the* population-*level effect (or the* average *treatment effect) as we take an expectation over $\boldsymbol{u}$ and aim to decompose δ (as opposed to $\delta_{\boldsymbol{u}}$). We note that our framework works for* individual-*level effect as well since given $\boldsymbol{u}$, all of our machinery can be recycled to propagate $\delta_{\boldsymbol{u}}$ through the graph. However, a practical challenge might be that $\boldsymbol{u}$ is unobserved. To overcome this, we can use the "abduction, action, and prediction" framework of Pearl et al. (2016). We elaborate further in the context of the causal unfairness application (see Remark 11).*

3. Existing approaches and their limitations

As stated in §1, quantifying effect propagation relates to various concepts on causality. To be concise, we review direct and indirect effects (§3.1), path-based techniques (§3.2), and degree of responsibility (§3.3) as these notions seem to be the closest to our work. As we illustrate below, each of these notions either fails to satisfy the seemingly natural property of source efficiency or is not even well-defined beyond the class of linear SEMs. On the other hand, in addition to satisfying source efficiency, our flow-based approach (§4) uniquely obeys a set of desirable axioms (e.g., flow conservation and attributing zero flow to a redundant edge) and is well-defined for a non-parametric SEM with arbitrary DAG structure.

It is also worth highlighting ideas such as Blinder-Oaxaca decomposition (Kitagawa, 1955; Blinder, 1973; Oaxaca, 1973), which attempt to quantify effect decomposition and understand direct and indirect discrimination in a regression setting. Our flow-based approach (§4) naturally aids such quantification in a much more general setting than regression. We discuss one such application in §7 (causal unfairness).

We note that in addition to the causality literature, our work intersects with the burgeoning explainable AI / interpretable ML literature. We provide an in-depth discussion of this literature (and its connections to the proposed approach) in our preliminary work (Singal et al., 2021). Therefore, for brevity, we primarily center our discussion here on the causality literature.

3.1 Direct and indirect effects

Pearl (2001) considers a graph with two channels (see Figure 2 in §1) and quantifies *direct* (effect propagated through the direct channel $X_1 \to Y$) and *indirect* (effect mediated through the indirect channel $X_1 \to X_2 \to Y$) effects. Given background $x_1^{(1)}$ and foreground $x_1^{(2)}$ with noise (u_2, u_3), Pearl (2001) defines *natural direct effect* as the following difference:

$$f_3(x_1^{(2)}, f_2(x_1^{(1)}, u_2), u_3) - f_3(x_1^{(1)}, f_2(x_1^{(1)}, u_2), u_3). \tag{1}$$

That is, in the output equation $f_3(x_1, x_2, u_3)$, the source variable changes from $x_1^{(1)}$ to $x_1^{(2)}$ while the intermediate variable X_2 is held at its background value $f_2(x_1^{(1)}, u_2)$. On the other hand, for the *natural indirect effect*, the intermediate variable changes from $f_2(x_1^{(1)}, u_2)$ to $f_2(x_1^{(2)}, u_2)$ while the source variable is held at its background value $x_1^{(1)}$:

$$f_3(x_1^{(1)}, f_2(x_1^{(2)}, u_2), u_3) - f_3(x_1^{(1)}, f_2(x_1^{(1)}, u_2), u_3). \tag{2}$$

Figure 6 provides a visual description of (1) and (2) in terms of active (solid black) and inactive (dashed red) edges (formally defined in §4). As alluded to in §1, it should be clear that such a notion only allows for one possible counterfactual. The *average* direct and indirect effects are defined by taking an expectation over the noise (u_2, u_3).

A key limitation of such a definition is that it fails to capture the interactions between the two channels, resulting in it violating source efficiency. We illustrate this in Example 1.

Example 1 (Interaction) *Consider the graph in Figure 2 with the following structural equations:* $X_2 = X_1$ *and* $Y = X_1 X_2$ *(interaction). Suppose background* $x_1^{(1)} = 0$ *and*

(a) Natural direct effect **(b)** Natural indirect effect

Figure 6: Illustration of (a) direct and (b) indirect effects (Pearl, 2001). Under the direct effect, as shown in the first graph, the foreground value $x_1^{(2)}$ propagates through the direct edge $(1, 3)$ (active) but not through the indirect edge $(1, 2)$ (inactive). (Node 3 denotes the outcome Y.) Direct effect corresponds to the difference between the first two graphs. Under the indirect effect, as shown in the third graph, the foreground value $x_1^{(2)}$ propagates through the indirect edge $(1, 2)$ (active) but not through the direct edge $(1, 3)$ (inactive). Indirect effect corresponds to the difference between the last two graphs.

foreground $x_1^{(2)} = 1$. *Hence,* $x_2^{(1)} = y^{(1)} = 0$ *and* $x_2^{(2)} = y^{(2)} = 1$. *Both natural direct and indirect effects as in* (1) *and* (2) *equal 0, even though the change in the outcome equals 1. Hence, the sum of the two effects is less than the total effect. (We assumed the model in this example to be deterministic for ease of illustration and note that it is straightforward to introduce noise variables.)*

In fact, realizing that natural direct and indirect effects do not necessarily sum up to the total effect, Pearl (2001) himself stated such relationships to be "strange" (see discussion below Equation (23) in Pearl (2001)). The notion of reverse indirect effect (Pearl, 2010) can fix this limitation. However, as we discuss in Appendix B, it is rather an ad hoc fix and can result in different attributions depending on the order in which one adds the channels. In fact, the recent work of Plecko and Bareinboim (2024) points this out as well (see §5.1 in their paper) and highlights the averaging we propose in this work as a solution.

We mention in passing the work of Zhang and Bareinboim (2018a), which generalizes such notions to account for confounders. However, under the absence of counfounders (which is the focus of this work), the proposal of Zhang and Bareinboim (2018a) suffers from the above-mentioned limitations as it reduces to Pearl (2001).

Beyond the simple graphical structure as in Figure 6 (one source variable and one mediator), it is possible to generalize the notion of direct and indirect effects to graphs with multiple source variables and multiple mediators. In fact, Pearl (2001) does so via path-specific effects, which we discuss next.

3.2 Path-based techniques

Path-based techniques decompose the total effect by attributing it to the underlying paths in the DAG and date back to the method of *path coefficients*, introduced over a century ago (Wright, 1918) and discussed in various fields, including mathematical statistics (Wright, 1934), sociology (Duncan, 1966; Alwin and Hauser, 1975; Fox, 1980), and econometrics (Goldberger, 1972). As alluded to in §1, though the path coefficients method handles the

case of linear SEMs (by appropriately defining path coefficients using the underlying edge weights, as we did for the example in Figure 1), its key limitation is the inability to generalize beyond the class of linear SEMs.

Recognizing this limitation, Pearl (2001) defined *path-specific effects*, where given a path from a source node to the outcome node, one only allows the foreground value of the source node to be propagated along the edges in the path (and all other edges are inactive). The corresponding path-specific effect is the difference between the resulting (expected) outcome value and the baseline outcome value $y^{(1)}$. Despite being well-defined for non-parametric SEMs, path-specific effects can suffer from the same key limitation as direct and indirect effects, i.e., the sum of path-specific effects might not add up to the total effect. For instance, in Example 1, the path-specific effect of both the direct path ($X_1 \to Y$) and the indirect path ($X_1 \to X_2 \to Y$) equals 0, since the absence of either of the paths results in a failure to capture the interaction effect X_1X_2. As with direct and indirect effects, the key reason for this shortcoming is that path-specific effects only allow for one counterfactual, i.e., all other edges are assumed to be inactive.

Recent path-based works include Zhang and Bareinboim (2018b) and Henshaw et al. (2020). Though non-parametric in nature, Zhang and Bareinboim (2018b) tackle a different problem, as their goal is to decompose the *covariance* between a source variable and the outcome variable as a sum over the unblocked paths. On the other hand, though we do not discuss the *path-specific causal derivative / selection gradient* (Henshaw et al., 2020) in detail, it is straightforward to verify that it suffers from the same limitation as path-specific effects (decomposition not adding up to the total effect).

3.3 Degree of responsibility

Motivated by "causality typically [being] treated an all-or-nothing concept", Chockler and Halpern (2004) study a similar problem and propose *degree of responsibility* to quantify attribution to nodes. To illustrate the intuition behind degree of responsibility, consider the following example, which is adapted from Chockler and Halpern (2004).

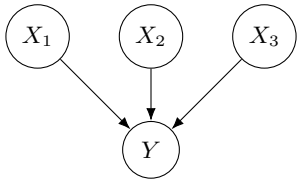


Figure 7: The graph corresponding to the voting example (Example 2).

Example 2 (Voting) *There are two candidates (1 and 2) in an election and three voters. For each voter* $i \in \{1, 2, 3\}$*, source variable* $X_i \in \{0, 1, 2\}$ *denotes the candidate voter* i *votes for, where* 0 *means the voter is undecided. We are interested in the outcome* $Y \in \{0, 1\}$*, which denotes whether the first candidate wins:*

$$Y = \begin{cases} 1 & \text{if } \sum_{i=1}^{3} \mathbb{I}\{X_i = 1\} > \sum_{i=1}^{3} \mathbb{I}\{X_i = 2\} \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{I}\{\cdot\}$ *denotes the indicator function and the underlying graph is shown in Figure 7. Consider the following two scenarios:*

1. *Background* $\boldsymbol{x}^{(1)} = (0, 0, 0)$ *and foreground* $\boldsymbol{x}^{(2)} = (1, 1, 1)$*, i.e., all three voters vote for candidate 1 and hence, candidate 1 wins the election 3-0.*
2. *Background* $\boldsymbol{x}^{(1)} = (0, 0, 0)$ *and foreground* $\boldsymbol{x}^{(2)} = (1, 1, 2)$*, i.e., the first two voters pick candidate 1 and the third voter picks candidate 2, meaning candidate 1 wins 2-1.*

In the words of Chockler and Halpern (2004),

> *"If someone wins an election [3-0], then each person who votes for [them] is less responsible for the victory than if [they] had won [2-1]."*

To operationalize this intuition, Chockler and Halpern (2004) define degree of responsibility of A for B as $1/(\kappa+1)$*, where* κ *"is the minimal number of changes that have to be made to obtain a contingency where B counterfactually depends on A". In the first scenario (3-0), this notion attributes a value of 1/2 to each of the 3 voters, as 1 change is needed for a vote to be critical. On the other hand, in the second scenario (2-1), the degree of responsibility of each of the 2 voters (for candidate 1) is 1, as each vote is critical (and attribution to the remaining 1 voter is 0).*

Similar to Pearl (2001), a key limitation of such a notion is that it violates source efficiency, as should be clear from Example 2. In fact, Chockler and Halpern (2004) themselves stated that a better name for their notion may have been "degree of criticality" (see their discussion in §5) by pointing out the following example:

> "For example, consider a voter who voted for [candidate 1] in the case of a 1-0 vote and a voter who voted for [candidate 1] in the case of a 100-99 vote. In both case, that voter has degree of responsibility 1. While it is true that, in both cases, that voter's vote was ciritical, in the second case, the voter may believe that his responsibility is more diffuse."

Recognizing this, Chockler and Halpern (2004) discussed the possibility of using Shapley value to define responsibility, which motivates RSV.

4. The recursive Shapley value (RSV) approach

We first provide the intuition behind the proposed approach (§4.1) and then formally introduce it (§4.2). Subsequently, we establish the flow-based axioms (§4.3) obeyed by RSV. Finally, we discuss how RSV connects to existing approaches in the literature (§4.4). We note that the presentation here generalizes our preliminary work (Singal et al., 2021) since the model here has noise variables $\mathbf{U}$, which were absent before.

Before providing intuition, we introduce some notation. Without loss of generality, we insert node 0 (*super-source*) to the DAG G. This node has no parent but has $|\mathsf{N}_0|$ outgoing edges (one to each of the source node): $(0, j)$ for $j \in \mathsf{N}_0$. The set of edges is re-defined to include these edges: $\mathsf{E} \leftarrow \mathsf{E} \cup \{(0, j) : j \in \mathsf{N}_0\}$. Similarly, the set of nodes is re-defined: $\mathsf{N} \leftarrow \mathsf{N} \cup \{0\}$ and $\mathsf{N}^+ \leftarrow \mathsf{N}^+ \cup \{0\}$. The original set of source nodes is still denoted by N_0

and this collection does not contain the super-source node 0. For each non-outcome node $i \in \mathsf{N}$, we define $\mathsf{E}_i := \{(i,j) : j \in \mathsf{C}_i\}$ as the set of its outgoing edges, implying $(\mathsf{E}_0, \ldots, \mathsf{E}_n)$ forms a partition of E. The proposed flow-based approach will quantify edge attributions $[\pi_{ij}]_{(i,j)\in\mathsf{E}}$. These edge attributions are used to define the node attributions $[\pi_j]_{j\in\mathsf{N}\setminus\{0\}}$ as the corresponding incoming flow:

$$\pi_j := \sum_{i\in\mathsf{P}_j} \pi_{ij} \ \forall j \in \mathsf{N} \setminus \{0\}.$$

It will help to formalize the notion of an *active* or *inactive* edge. In our causal graph setting, the role of an edge is to propagate information. As a result, motivated by Pearl (2001) (recall Figure 6), if an edge $(i,j) \in \mathsf{E}$ is active, then it propagates the updated value x_i from node i to node j, whereas if it is inactive, then it does not propagate the updated value, meaning node j receives the background value. An illustration is provided in Figure 8.

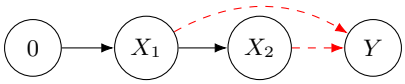


Figure 8: Illustrating active / inactive edges. Here, $\mathsf{E} = \{(0,1),(1,2),(1,3),(2,3)\}$ is the set containing all edges in the DAG (active and inactive). Node 3 denotes the outcome Y. There are two inactive edges: $(1,3)$ and $(2,3)$ (dashed red lines). Consider arbitrary noise value $\mathbf{U} = (u_2, u_3)$. Since edge $(0,1)$ is active, we set node 1 to its foreground value $x_1^{(2)}$. As edge $(1,2)$ is active *and* X_1 is set to $x_1^{(2)}$, node 2 receives $x_1^{(2)}$. However, since edges $(1,3)$ and $(2,3)$ are inactive, node 3 (Y) receives $x_1^{(1)}$ (from node 1) and $x_2^{(1)} := f_2(x_1^{(1)}, u_2)$ (from node 2). Therefore, $X_1 = x_1^{(2)}$, $X_2 = f_2(x_1^{(2)}, u_2)$, and $Y = f_3(x_1^{(1)}, x_2^{(1)}, u_3)$. Note that if edge $(0,1)$ had been inactive, then the variables would have been set as follows: $X_1 = x_1^{(1)}$, $X_2 = f_2(x_1^{(1)}, u_2)$, and $Y = f_3(x_1^{(1)}, x_2^{(1)}, u_3)$.

In general, we consider a subset $E \subseteq \mathsf{E}$ of active edges and arbitrary noise $\mathbf{U} = \boldsymbol{u}$. Let $\boldsymbol{x}^{(1)}(\boldsymbol{u})$ denote the background corresponding to $\boldsymbol{u}$, i.e., $x_i^{(1)}(\boldsymbol{u})$ denotes the value at node $i \in \mathsf{N}^+ \setminus \{0\}$ given $(\boldsymbol{x}_{\mathsf{N}_0}^{(1)}, \boldsymbol{u})$. Then, each source node $i \in \mathsf{N}_0$ is set as a function of E as follows:

$$x_i(E) := \begin{cases} x_i^{(1)} & \text{if } (0,i) \notin E \\ x_i^{(2)} & \text{if } (0,i) \in E. \end{cases} \tag{3a}$$

Leveraging the underlying structural equations, we set each non-source node $j \in \mathsf{N}^+ \setminus \{\mathsf{N}_0 \cup \{0\}\}$ as a functon of active edges E as follows:

$$x_j(E, \boldsymbol{u}) = f_j((x_{ij}(E, \boldsymbol{u}))_{i\in\mathsf{P}_j}, u_j), \tag{3b}$$

where for all $i \in \mathsf{P}_j$,

$$x_{ij}(E, \boldsymbol{u}) := \begin{cases} x_i^{(1)}(\boldsymbol{u}) & \text{if } (i,j) \notin E \\ x_i(E, \boldsymbol{u}) & \text{if } (i,j) \in E. \end{cases} \tag{3c}$$

We note that such a construction is motivated by Pearl (2001) (see Figure 3 in Pearl (2001) for example). Notation x_{ij} is new and $x_{n+1}(E, \boldsymbol{u})$ corresponds to the outcome, which we denote by $y_{\boldsymbol{u}}(E)$. Analogous to the notation in §2.2, we define $y(E) := \mathbb{E}[y_{\boldsymbol{u}}(E)]$ by taking the expectation over $\boldsymbol{u}$. Thus, given arbitrary $E \subseteq \mathsf{E}$, the notation $y(E)$ is well-defined, using which we provide intuition next. This notation would be critical to define RSV, since unlike existing notions, RSV allows for a broad spectrum of counterfactuals in terms of which edges are active. (Note that the discussion below assumes basic knowledge of Shapley value (Shapley, 1953), a brief primer for which is presented in Appendix A.)

4.1 Intuition guiding the RSV approach

First, the proposed RSV approach attributes to the source nodes. Then, it quantifies effect propagation by flowing down the source node attributions on the outgoing edges (top-down). To convey intuition, we illustrate the RSV mechanics on a relatively simple DAG (see Figure 9) with arbitrary F (structural equations) and D (noise distribution).

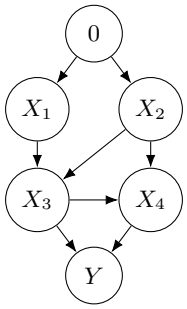


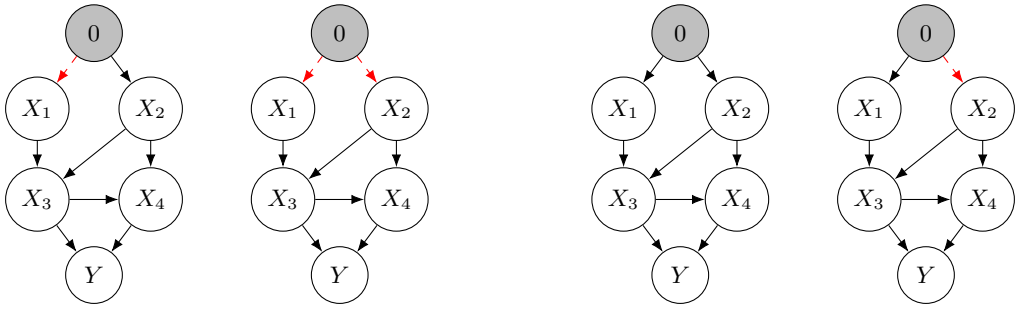
Figure 9: DAG used to illustrate the intuition behind the proposed RSV approach.

Recall that $\mathbf{X}_{\mathsf{N}_0}$ changes from $\boldsymbol{x}^{(1)}_{\mathsf{N}_0}$ to $\boldsymbol{x}^{(2)}_{\mathsf{N}_0}$, resulting in the expected outcome changing from $y^{(1)} = \mathbb{E}[y^{(1)}_{\boldsymbol{u}}]$ to $y^{(2)} = \mathbb{E}[y^{(2)}_{\boldsymbol{u}}]$. Since the source nodes $\mathsf{N}_0 = \{1, 2\}$ capture all the exogenous source of variation, the expected change in Y is entirely driven by the propagation of changes at (X_1, X_2). As a result, we begin by attributing to the source nodes, which is done via the following game at the super-source node 0. The collection of outgoing edges from node 0, i.e., $\mathsf{E}_0 = \{(0, 1), (0, 2)\}$, is the set of players. Given a subset of players (coalition) $E_0 \subseteq \mathsf{E}_0$, we define the corresponding characteristic function as

$$v_0(E_0) := y(E_0, \mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3, \mathsf{E}_4).$$

Recall the notation $y(\cdot)$ is defined in (3). The downstream edges $(\mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3, \mathsf{E}_4)$ are set to be active in this game. Then, the attributions received by edges $(0, 1)$ and $(0, 2)$, i.e., π_{01} and π_{02}, equal the SVs of this game (illustrated in Figure 10). To explicitly capture the dependence on $\mathsf{E} = (\mathsf{E}_0, \mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3, \mathsf{E}_4)$ in our notation, we denote these attributions as $\pi_{01}(\mathsf{E})$, $\pi_{02}(\mathsf{E})$, $\pi_1(\mathsf{E})$, and $\pi_2(\mathsf{E})$.

We then move down to the children of node 0, i.e., nodes 1 and 2, to understand how the effect flows through the graph. Since node 1 has only one outgoing edge, the effect propagation through node 1 is trivial (cf. flow conservation): $\pi_{13} = \pi_1(\mathsf{E})$. On the other hand, quantifying the propagation of $\pi_2(\mathsf{E})$ through node 2 is challenging as there are two



(a) Edge $(0,1)$ inactive

(b) Edge $(0,1)$ active

Figure 10: Illustration of the flow π_{02}. The effect propagated by edge $(0,2)$ is quantified via its value-add, as determined by the following counterfactual question: how much would have been the effect (expected change in Y) had edge $(0,2)$ been inactive? Note that such a counterfactual question involves two possibilities: (a) edge $(0,1)$ being inactive and (b) edge $(0,1)$ being active. The difference between the first two graphs captures possibility (a) (value-add of edge $(0,2)$ given edge $(0,1)$ being inactive). On the other hand, the difference between the last two graphs captures possibility (b) (value-add of edge $(0,2)$ given edge $(0,1)$ being active). Edge $(0,2)$ receives an attribution equal to a (weighted) average of these two counterfactuals, where the weights are determined via SV ($1/2$ in this example). We note that all downstream edges $(\mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3, \mathsf{E}_4)$ are set to be active in both the possibilities.

outgoing edges: $\mathsf{E}_2 = \{(2,3),(2,4)\}$. Our goal is to decouple the flow received by node 2 into its two outgoing edges. To do so, we aim to understand how much does the presence of each of these two edges contributes to $\pi_2(\mathsf{E})$. Recall that the upstream game at node 0 assumed all the downstream edges to be active. Thus, node 2 receives an attribution of $\pi_2(\mathsf{E}_0, \mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3, \mathsf{E}_4)$ *if* both the edges in E_2 are active. To decompose π_2 into π_{23} and π_{24}, following counterfactual questions seem natural:

- Had both the edges in E_2 been inactive, how much flow would have propagated through node 2?
- What if edge $(2,3)$ was inactive and $(2,4)$ active (and vice versa)?

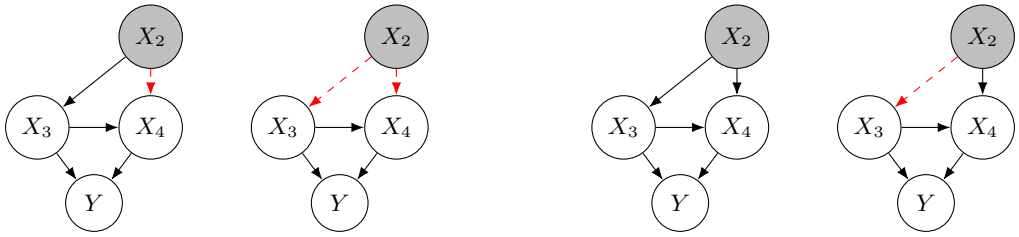
In the upstream game at node 0, had both the edges $(2,3)$ and $(2,4)$ been inactive, then the resulting characteristic function would have been $y(E_0, \mathsf{E}_1, \emptyset, \mathsf{E}_3, \mathsf{E}_4)\ \forall E_0 \subseteq \mathsf{E}_0$, meaning edge $(0,2)$ (and thus, node 2) would have received no flow at all, i.e., $\pi_2(\mathsf{E}_0, \mathsf{E}_1, \emptyset, \mathsf{E}_3, \mathsf{E}_4) = 0$. Therefore, edges $(2,3)$ and $(2,4)$ are entirely responsible for the flow $\pi_2(\mathsf{E})$ through node 2. As a result, flow conservation seems logical: $\pi_2(\mathsf{E}) = \pi_{23} + \pi_{24}$. However, this does not uniquely determine the flow on the two edges: π_{23} and π_{24}.

An intuitive scenario is if an edge is redundant, i.e., the attribution received by node 2 is independent of the edge being active or inactive. Then, assigning zero flow on such an edge seems apt (nullity). A second intuitive scenario corresponds to both the edges in E_2 being equivalent. For instance, if the value attributed to node 2 equals $\pi_2(\mathsf{E})/2$ if *either* of the two edges is active, then decoupling $\pi_2(\mathsf{E})$ equally between them aligns with intuition (symmetry).

We operationalize such intuition by considering the following game at node 2. The set of outgoing edges (E_2) maps to the set of players. For arbitrary subset of players (coalition) $E_2 \subseteq \mathsf{E}_2$, we define the characteristic function as the attribution received by node 2 from the upstream game at node 0 (hence, recursive):

$$v_2(E_2) := \pi_2(\mathsf{E}_0, \mathsf{E}_1, E_2, \mathsf{E}_3, \mathsf{E}_4).$$

The flows received by edges $(2,3)$ and $(2,4)$, i.e., π_{23} and π_{24}, are defined to be the SVs of this game (illustrated in Figures 11 and 12).



(a) Edge $(2,4)$ inactive

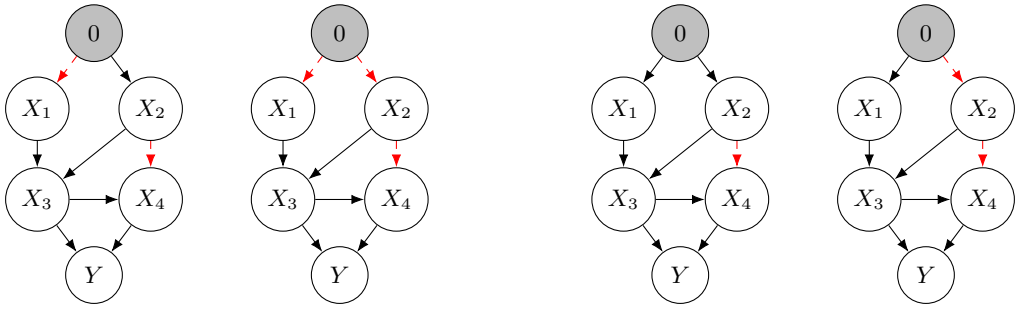
(b) Edge $(2,4)$ active

Figure 11: Illustration of the flow π_{23}. The effect propagated by edge $(2,3)$ is quantified via its value-add, as determined by the following counterfactual question: how much would have been the attribution received by node 2 had edge $(2,3)$ been inactive? Note that such a counterfactual question involves two possibilities: (a) edge $(2,4)$ being inactive and (b) edge $(2,4)$ being active. The difference between the first two graphs captures to possibility (a) (value-add of edge $(2,3)$ given edge $(2,4)$ being inactive). On the other hand, the difference between the last two graphs captures possibility (b) (value-add of edge $(2,3)$ given edge $(2,4)$ being active). Edge $(0,2)$ receives an attribution equal to a (weighted) average of these two counterfactuals, where the weights are determined via SV ($1/2$ in this example). Note that the attribution node 2 receives in each of these four graphs still needs to be evaluated. Figure 10 illustrated this evaluatation for the third graph here (both $(2,3)$ and $(2,4)$ active). In Figure 12 below, we show it for the first graph ($(2,3)$ active but $(2,4)$ inactive). Computations for the second and fourth graphs are similar (not shown to be concise).

Having computed the flow through nodes 1 and 2, we move down to the next layer of children: nodes 3 and 4. To compute the flow through these two nodes, we recycle our logic (in a recursive manner). For example, at the node 3 game, the players are the corresponding outgoing edges E_3. Given arbitrary coalition $E_3 \subseteq \mathsf{E}_3$, we define the characteristic function to be the incoming flow at node 3 from the upstream games: $v_3(E_3) := \pi_3(\mathsf{E}_0, \mathsf{E}_1, \mathsf{E}_2, E_3, \mathsf{E}_4)$.

4.2 Recursive Shapley value based attribution

The formal definition of RSV involves n recursions, with each recursion being initialized at the corresponding node $j \in \mathsf{N} \setminus \{0\}$ via the following game. The set of players includes all the outgoing edges E_j. Given a subset of players (coalition) $E_j \subseteq \mathsf{E}_j$, we define the characteristic function to be the node j attribution from the upstream games with all other



(a) Edge $(0,1)$ inactive

(b) Edge $(0,1)$ active

Figure 12: Illustration of the flow π_{02} (attribution node 2 receives) when one of its outgoing edges (edge $(2,4)$) is inactive. The four graphs here replicate the ones in Figure 10 but with edge $(2,4)$ being inactive. The steps to compute π_{02} are identical to the ones in the caption of Figure 10.

edges being active:

$$v_j(E_j) := \pi_j(\mathsf{E}_0, \ldots, E_j, \ldots, \mathsf{E}_n) = \sum_{i \in \mathsf{P}_j} \pi_{ij}(\mathsf{E}_0, \ldots, E_j, \ldots, \mathsf{E}_n). \tag{4}$$

Given this game, the flow attributed to edge $(j,k) \in \mathsf{E}_j$ is the corresponding SV:

$$\pi_{jk}(\mathsf{E}) = \sum_{E_j \subseteq \mathsf{E}_j \setminus \{(j,k)\}} w_{\mathsf{E}_j}(E_j) \times \left\{ v_j(E_j \cup \{(j,k)\}) - v_j(E_j) \right\}, \tag{5}$$

where

$$w_{\mathsf{E}_j}(E_j) := \frac{|E_j|!\,(|\mathsf{E}_j| - |E_j| - 1)!}{|\mathsf{E}_j|!}$$

is the SV weight function as in Appendix A. It should be clear from (4) and (5) that $\pi_{jk}(\mathsf{E}_0, \ldots, \mathsf{E}_n)$ is a function of $\pi_{ij}(\mathsf{E}_0, \ldots, E_j, \ldots, \mathsf{E}_n)$ for $E_j \subseteq \mathsf{E}_j$, where node i is a parent of node j. The latter defines a recursion since to compute $\pi_{ij}(\mathsf{E}_0, \ldots, E_j, \ldots, \mathsf{E}_n)$ at an arbitrary upstream node $i \in \mathsf{P}_j$, we use

$$\pi_i(\mathsf{E}_0, \ldots, E_i, \ldots, E_j, \ldots, \mathsf{E}_n) \; \forall E_i \subseteq \mathsf{E}_i$$

as the characteristic function of the corresponding game at node i, as opposed to

$$\pi_i(\mathsf{E}_0, \ldots, E_i, \ldots, \mathsf{E}_j, \ldots, \mathsf{E}_n).$$

In other words, all other edges are not assumed to be active *within* a recursion, but only at the *initialization*. The pseudocode in Algorithm 1 formalizes (and clarifies) this. When necessary, we will use the notation $v_j(E_j \mid E_{-j})$ to highlight this difference, where $E_{-j} := (E_0, \ldots, E_{j-1}, E_{j+1}, \ldots, E_n)$ and $E_\ell \subseteq \mathsf{E}_\ell \; \forall \ell \in \mathsf{N}$. Recall that the default notation $v_j(\cdot)$ assumes all other edges E_{-j} to be active, i.e., $v_j(\cdot) = v_j(\cdot \mid \mathsf{E}_{-j})$ for all $j \in \mathsf{N}$.

Each recursion breaks at node 0 via the following game. The set of players is E_0 and we define this game conditioned on an arbitrary collection of downstream edges being active: $(E_1, \ldots, E_n)$ where $E_\ell \subseteq \mathsf{E}_\ell,\ \forall \ell = 1, \ldots, n$. For a given coalition $E_0 \subseteq \mathsf{E}_0$, we define the characteristic function to be the expected outcome with (E_0, E_{-0}) being the active edges (recall (3)):

$$v_0(E_0 \mid E_{-0}) := y(E_0, E_{-0}).$$

Given this game, the flow on edge $(0,k) \in \mathsf{E}_0$ is the corresponding SV:

$$\pi_{0k}(\mathsf{E}_0, E_{-0}) = \sum_{E_0 \subseteq \mathsf{E}_0 \setminus \{(0,k)\}} w_{\mathsf{E}_0}(E_0) \times \left\{ v_0(E_0 \cup \{(0,k)\} \mid E_{-0}) - v_0(E_0 \mid E_{-0}) \right\}.$$

Observe that this definition is non-recursive and hence, it breaks every recursion. The flow attributed to super-source edges E_0 equals the SVs of the node 0 game defined above but with all downstream edges being active, i.e., E_{-0} is set to E_{-0}.

Given the DAG structure, all of the n recursions are well-defined and end up evaluating the expected outcome $y(E)$ for various $E \subseteq \mathsf{E}$. We formally state our recursive definition of RSV in Algorithm 1 (and sub-routine Algorithm 2). Algorithm 1 outputs RSV: $\pi^{\text{RSV}}_{jk} = \pi_{jk}(\mathsf{E})\ \forall (j,k) \in \mathsf{E}$. Note that the two inputs N and E provided to Algorithm 1 are assumed to include the super-source node 0 and the corresponding edges E_0, i.e., $\mathsf{N} = \{0, \ldots, n\}$ and $\mathsf{E} = (\mathsf{E}_0, \ldots, \mathsf{E}_n)$. Note that the (j,k) notation in the sub-script of $\texttt{RSV}_{jk}$ represents an input to Algorithm 2. Also, the notation E, $\mathcal{E}$, and E is used to distinguish three different sets of edges since the recursion primarily revolves around them.

Algorithm 1 $\texttt{RSV}(\mathsf{N}, \mathsf{E})$

for $(j,k) \in \mathsf{E}$
  $\pi^{\text{RSV}}_{jk} = \texttt{RSV}_{jk}(\mathsf{E}_0, \ldots, \mathsf{E}_n)$
end for
return $[\pi^{\text{RSV}}_{jk}]_{(j,k)\in\mathsf{E}}$

Algorithm 2 $\texttt{RSV}_{jk}(\mathcal{E}_0, \ldots, \mathcal{E}_n)$

if $j > 0$ % recursion
  return $\sum_{E_j \subseteq \mathcal{E}_j \setminus \{(j,k)\}} w_{\mathcal{E}_j}(E_j) \times$
    $\sum_{i \in \mathsf{P}_j} \{\texttt{RSV}_{ij}(\mathcal{E}_0, \ldots, E_j \cup \{(j,k)\}, \ldots, \mathcal{E}_n) - \texttt{RSV}_{ij}(\mathcal{E}_0, \ldots, E_j, \ldots, \mathcal{E}_n)\}$
else % base case
  return $\sum_{E_0 \subseteq \mathcal{E}_0 \setminus \{(0,k)\}} w_{\mathcal{E}_0}(E_0) \times \{y(E_0 \cup \{(0,k)\}, \mathcal{E}_1, \ldots, \mathcal{E}_n) - y(E_0, \mathcal{E}_1, \ldots, \mathcal{E}_n)\}$
end if

4.3 Flow-based axioms

Having defined RSV, we now establish its desirability by proving uniqueness to a set of seemingly natural flow-based axioms. The first axiom is conservation of flow (recall Definition 1 in §2). The next two axioms are flow symmetry and flow nullity, as touched upon

in §4.1. Intuitively, they require equivalent outgoing edges to receive equal flow and a redundant edge to receive no flow, respectively. The final axiom advocates for flow linearity, which we shed light upon after formally defining these axioms.

Definition 3 (Flow-based axioms) *The flow-based axioms are as follows:*

1. *Flow conservation:* $\sum_{k\in\mathsf{C}_0}\pi_{0k} = y^{(2)} - y^{(1)}$ *and* $\sum_{i\in\mathsf{P}_j}\pi_{ij} = \sum_{k\in\mathsf{C}_j}\pi_{jk}\ \forall j \in \mathsf{N}\setminus\{0\}$.
2. *Flow symmetry: For node* $j \in \mathsf{N}$*, if* $(j,k)\in\mathsf{E}_j$ *and* $(j,\ell)\in\mathsf{E}_j$ *are such that* $v_j(E_j\cup\{(j,k)\}) = v_j(E_j\cup\{(j,\ell)\})\ \forall E_j\subseteq\mathsf{E}_j\setminus\{(j,k),(j,\ell)\}$*, then* $\pi_{jk}=\pi_{j\ell}$.
3. *Flow nullity: For node* $j\in\mathsf{N}$*, if* $v_j(E_j\cup\{(j,k)\}) = v_j(E_j)\ \forall E_j\subseteq\mathsf{E}_j\setminus\{(j,k)\}$*, then* $\pi_{jk}=0$.
4. *Flow linearity: For node* $j\in\mathsf{N}$*, consider characteristic functions* $v_j(\cdot)$ *and* $v'_j(\cdot)$*. Linearity requires* $\pi_{jk}(v_j+v'_j) = \pi_{jk}(v_j)+\pi_{jk}(v'_j)\ \forall(j,k)\in\mathsf{E}_j$*. (Notation* $\pi_{jk}(v_j)$ *captures the dependence of* π_{jk} *on* $v_j(\cdot)$*.)*

Though conservation of flow is stated slightly differently than before (Definition 1), it is easy to establish equivalence between the two. Flow linearity can be interpreted as follows. Note that the characteristic functions $[v_j(\cdot)]_{j\in\mathsf{N}}$ (defined in §4.2) in the end correspond to the structural equations F as given arbitrary $j\in\mathsf{N}$, $v_j(\cdot)$ ends up evaluating $v_0(\cdot)$, which ultimately is a function of F. Now, let $[v'_j(\cdot)]_{j\in\mathsf{N}}$ correspond to a different population of variables $\mathbf{X}$, governed by a different set of structural equations (say F') and noise (say D'). Linearity is the requirement that the attribution should be robust to mixing such heterogeneous populations. In other words, one can either (1) mix the two populations first $(v_j+v'_j)$ and compute attribution using the mixture $(\pi_{jk}(v_j+v'_j))$ or (2) compute attributions on the two populations first $(\pi_{jk}(v_j)$ and $\pi_{jk}(v'_j))$ and mix them later $(\pi_{jk}(v_j)+\pi_{jk}(v'_j))$. Under linearity, both (1) and (2) output the same attributions.

It is possible to interpret flow symmetry and nullity using the model primitive $y(\cdot)$. To do so, observe that that given node $j\in\mathsf{N}$, if $(j,k)\in\mathsf{E}_j$ and $(j,\ell)\in\mathsf{E}_j$ obey

$$y(E\cup\{(j,k)\}) = y(E\cup\{(j,\ell)\})\ \forall E\subseteq\mathsf{E}\setminus\{(j,k),(j,\ell)\},$$

then symmetry requires the two edges two receive same flow, i.e., $\pi_{jk}=\pi_{j\ell}$. Similarly, if

$$y(E\cup\{(j,k)\}) = y(E)\ \forall E\subseteq\mathsf{E}\setminus\{(j,k)\},$$

then nullity requires edge (j,k) to receive zero flow, i.e., $\pi_{jk}=0$. In Theorem 4, we establish RSV is the unique solution to these four flow-based axioms, with a proof in Appendix C.

Theorem 4 *Given structural causal model* $\mathsf{M}=(\mathsf{G},\mathsf{F},\mathsf{D})$*, the RSV* $[\pi^{RSV}_{jk}]_{(j,k)\in\mathsf{E}}$ *defined via Algorithm 1 is the unique solution to the flow-based axioms.*

A direct corollary of Theorem 4 is that RSV overcomes the source efficiency violation of most existing approaches discussed in §3. In fact, Theorem 4 implies that if one believes in the seemingly natural flow-based axioms, then RSV is the *only* metric they should use to quantify effect propagation (since it is the *unique* solution). To showcase the desirability of the flow-based axioms, we illustrate them on a DAG equipped with linear structural equations F (Example 3).

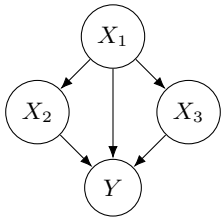


Figure 13: The graph for Example 3. (Node 0 is not included for brevity.)

Example 3 (Linear SEM) *Consider the graph in Figure 13. There is one source variable X_1. The structural equations are linear in $\mathbf{X}$ and have an additive noise $\mathbf{U}$:*

$$\begin{aligned} X_2 &= a_{12}X_1 + U_2 \\ X_3 &= a_{13}X_1 + U_3 \\ Y &= a_{14}X_1 + a_{24}X_2 + a_{34}X_3 + U_4. \end{aligned}$$

Suppose the noise is mean-zero, i.e., $\mathbb{E}[\mathbf{U}] = 0$ and consider the background $x_1^{(1)} = 0$ and the foreground $x_1^{(2)} = 1$. Then, the background and foreground expected outcomes equal:

$$\begin{aligned} y^{(1)} &= 0 \\ y^{(2)} &= a_{14} + a_{12}a_{24} + a_{13}a_{34}. \end{aligned}$$

RSV outputs the following flow:

$$\begin{aligned} \pi_{12}^{RSV} = \pi_{24}^{RSV} &= a_{12}a_{24} \\ \pi_{14}^{RSV} &= a_{14} \\ \pi_{13}^{RSV} = \pi_{34}^{RSV} &= a_{13}a_{34}. \end{aligned}$$

Interpreting conservation of flow is straightforward (flow in equals flow out). To see the intuition behind symmetry, let $a_{12} = a_{13}$ and $a_{24} = a_{34}$. Then, the change at node 1 is propagated identically through edges $(1, 2)$ and $(1, 3)$. As a result, these two edges receive the same flow. To understand nullity, consider an arbitrary $(j, k) \in \mathsf{E}$ and let $a_{jk} = 0$. Clearly, such as edge is redundant in terms of effect propagation and RSV attributes zero flow to it. Finally, to see the mechanics behind linearity, consider a different model with the same underlying DAG G but suppose the coefficients are different: $[a'_{jk}]_{(j,k)\in\mathsf{E}}$. RSV under the mixture model ($\tilde{a} = (a + a')/2$) is the same as averaging the RSVs under the individual models. For instance, edge $(1, 4)$ receives a flow of $(a_{14} + a'_{14})/2$ under the mixture model and a_{14} and a'_{14} under the individual models.

The above example also highlights RSV's ability to decouple *direct* and *indirect* effects (Pearl, 2001). In particular, the direct effect of the source X_1 on the outcome Y corresponds to the flow a_{14} on edge $(1, 4)$ whereas the indirect effect is the sum of flow on edges $(1, 2)$ and $(1, 3)$: $a_{12}a_{24} + a_{13}a_{34}$. In addition, RSV goes a step beyond by splitting the indirect effects into the underlying edge-specific flows and thus, facilitates *mediation analysis*: "a mediating variable transmits the effect of an independent variable on a dependent variable"

(MacKinnon et al., 2007). Connecting this notion to Figure 13, the independent variable X_1 changes from a value of 0 (background) to 1 (foreground), causing Y (the dependent variable) to change from an expected value of 0 to $a_{14} + a_{12}a_{24} + a_{13}a_{34}$ (total effect). RSV provides a crisp decomposition of this total effect. In particular, the direct channel ($X_1 \to Y$) carries a_{14} whereas the two mediating channels ($X_1 \to X_2 \to Y$) and ($X_1 \to X_3 \to Y$) transmit $a_{12}a_{24}$ and $a_{13}a_{34}$, respectively. (Note that it is possible to generalize the intuition here to an arbitrary DAG G with linear structural equations F and mean-zero noise. We do so in §5 (Theorem 6).)

Before concluding this subsection, we note that RSV's adherence to the flow-based axioms results in it respecting a mix of desirable properties (implementation invariance, sensitivity, monotonicity, and affine scale invariance) discussed in the interpretable ML literature (Sundararajan et al., 2017; Sundararajan and Najmi, 2020). For instance, implementation invariance requires the attributions to be robust to internal changes in the graph whereas sensitivity / monotonicity require the attributions to respect basic independence and monotonicity relations. Our preliminary work (Singal et al., 2021) provides a detailed discussion of such additional properties along with formal propositions, which we omit here for brevity. We emphasize here that such properties come out naturally as implications of our more fundamental flow-based axioms.

4.4 Connections with existing approaches in the causality literature

Having established RSV's desirable properties, we now highlight connections between RSV and existing approaches on causality. As before, we refer the reader to our preliminary work (Singal et al., 2021) for the connections to the interpretable ML literature, where we rigorously show how the *edge*-based RSV generalizes a number of existing *node*-based approaches.

Direct and indirect effects To understand the connection between natural direct / indirect effects (Pearl, 2001) and RSV, we revisit the two-channeled graph introduced in Figure 2 and illustrate the computation of direct effect under RSV in Figure 14. When computing the direct effect, i.e., attribution to edge $(1, 3)$, RSV allows for two possibilities with respect to the indirect channel: (a) edge $(1, 2)$ inactive and (b) edge $(1, 2)$ active. In particular, RSV averages over these two possibilities, allowing it to decouple the interaction effect (which only appears in possibility (b)). On the other hand, natural direct effect (Pearl, 2001) only considers possibility (a) and hence, is unable to capture the interaction. The connection between indirect effect under the two definitions is analogous. We exemplify this discussion by revisiting Example 1 next.

Revisiting Example 1 *Recall from Example 1 (interaction) that a key limitation of the proposal in Pearl (2001) is it does not explain interaction effects, which leads to a source efficiency violation (both channels receiving zero attribution in Example 1). RSV, on the other hand, obeys source efficiency (cf. flow conservation axiom). In Example 1, RSV attributes 1/2 to each of the two channels (cf. flow symmetry) and hence, captures the interaction effect. Thus, by allowing for a richer set of counterfactuals, RSV is able to decouple interaction effects in a principled manner, ensuring the direct and indirect effects add up to the total effect. Further, recall from Appendix B that although the notion of*

(a) Edge (1,2) inactive **(b)** Edge (1,2) active

Figure 14: Direct effect under RSV for the graph introduced in Figure 2. The direct channel $(1, 3)$ receives an attribution equal to its value-add, which is evaluated via the following counterfactual question: how much would have been the effect (change in Y) had edge $(1, 3)$ been inactive? Note that there are two possibilities in such a counterfactual question: (a) edge $(1, 2)$ is inactive and (b) edge $(1, 2)$ is active. The difference between the first two plots corresponds to possibility (a), i.e., value-add of edge $(1, 3)$ when edge $(1, 2)$ is inactive. Similarly, the difference between the last two plots corresponds to possibility (b), i.e., value-add of edge $(1, 3)$ when edge $(1, 2)$ is active. The attribution received by edge $(1, 3)$ is the weighted average of these two value-adds, where the weights come from the classical SV (1/2 for this instance). Recall that the natural direct effect of Pearl (2001) places all the weight on possibility (a).

reverse effects obeys efficiency, it lacks uniqueness since it results in either the direct or the indirect channel getting all the attribution depending on the order in which one adds the channels. RSV avoids this issue by averaging over these two orderings.

Path-based techniques The connection between path-specific effects (Pearl, 2001) / path-specific selection gradient (Henshaw et al., 2020) and RSV is similar. In particular, to quantify the effect propagated through the direct path $X_1 \to Y$, both path-specific effects and selection gradient put all the weight on possibility (a) and hence, do not account for the interaction effect, which only appears in possibility (b).

Remark 5 *It is worth noting that under the class of linear SEMs, a multitude of existing approaches (direct and indirect effects, path coefficients, path-specific effects, and path-specific selection gradient) result in the same intuitive quantification as discussed around Figure 1. In particular, all these approaches boil down to using the edge weights to compute the corresponding path coefficients and attribute accordingly* [6]*. As we formalize in §5 (Theorem 6), for linear SEMs, RSV recovers the same flow. Accordingly, RSV is consistent with most existing techniques in the relatively simple linear setting but provides a new perspective under non-linearity, where it is systematically able to account for interaction effects.*

Degree of responsibility RSV operationalizes the intuition of Chockler and Halpern (2004) (see the quotes "if someone wins ..." and "... responsibility is more diffuse" stated in §3.3) while adhering to a set of desirable axioms, which imply source efficiency. To illustrate this, we revisit Example 2.

6. Though some of these approaches are not clearly defined for the setting with multiple source nodes and / or multiple mediators, the implicit decoupling in the linear SEM case (due to no interaction effects) enables their computations.

Revisiting Example 2 *Recall that a key limitation of degree of responsibility is it violates source efficiency as well, as shown in Example 2 (voting). RSV, on the other hand, attributes 1/3 to each of the 3 voters in scenario 1 whereas in scenario 2, RSV attributes 2/3 to each of the 2 voters of candidate 1 and* $-1/3$ *to the voter of candidate 2. If, in scenario 2, we consider the background of* $(0, 0, 2)$ *(as opposed to* $(0, 0, 0)$*) while keeping the foreground as* $(1, 1, 2)$*, then voters 1 and 2 receive an attribution of 1/2 each, whereas voter 3 receives zero attribution (nullity).*

Blameworthiness It is worth discussing the work of Friedenberg and Halpern (2019), which is complementary to our work in the sense that they propose a definition of blame in a structural causal model. Their key contribution is to extend the framework of Halpern and Kleiman-Weiner (2018) to a multi-agent setting and they do so using Shapley value. Saying that, their definition is fundamentally different than ours since it depends on the notion of an epistemic state (distribution over causal models) to "capture an agent's beliefs about the effects that actions may have". RSV is independent of the individual-level beliefs but only operates on a single given causal model. To dive deeper, consider the illustrative example presented by Friedenberg and Halpern (2019) (see §3.4 of their paper):

> "Consider a scenario where a committee of 7 people, ag_1 through ag_7, vote for whether or not to pass a bill. If at least 4 agents vote `yes`, then the bill will pass. Everyone agrees that it would be better for the bill to pass, but there are external reasons (such as opinions of constituents) that might result in agents benefiting from voting `no` as long as the bill is passed. The committee votes and agents ag_1 through ag_5 all vote `no`, so the bill does not pass. How blameworthy is each agent for this outcome?"

As discussed in Friedenberg and Halpern (2019), the degree of blameworthiness varies as a function of the agent's beliefs. For instance,

> "ag_1 believed that each of the 6 other agents started with a 60% chance of voting `yes`. For any coalition of n agents, ag_1 also believed that for a cost of $n \times 100$ each agent's probability of voting `yes` (including that of agents not in the coalition) could be increased by $n \times 5\%$ by applying social pressure. In addition, if ag_1 herself was in the coalition, then for an additional cost of 2000 she would have switched her vote to `yes`. Given these beliefs, the degree of blameworthiness for the entire group is ≈ 0.390, while ag_1's degree of blameworthiness is ≈ 0.073."

On the other hand, RSV is independent of individual-level beliefs and simply attributes the same value to the 5 agents who voted `no` (cf. symmetry axiom). Degree of blameworthiness requires much more information (individual-level beliefs) than RSV, which can be challenging to gather. Clearly, there is a fundamental difference in how the two approaches attribute ex post blame.

We conclude this section by summarizing our foundational developments. So far, we have focussed on the definitional / philosophical aspects of flow-based attribution. Our goal has been to convince the reader that leaving computational tractability aside, RSV is an attractive metric for quantifying effect propagation. It recovers the existing approaches when appropriate and overcomes their limitations when they "fail", with a unique adherence

to a set of seemingly natural flow-based axioms. Having established such desirability of RSV, we shed light on its computational tractability next.

5. Characterization and computation of RSV

Next, we focus on the computational aspects of RSV and consider the following two cases: (i) all the structural equations $\mathsf{F} = [f_i(\cdot)]_{i \in \mathsf{N}^+ \setminus \mathsf{N}_0}$ in model $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$ are linear (§5.1) and (ii) F is non-parametric and hence, possibly non-linear (§5.2). For both settings, we provide a closed-form characterization of RSV and discuss the implications on its computation.

5.1 Linear SEM

Consider an arbitrary DAG G (containing super-source node 0) with linear structural equations F and mean-zero noise:

$$X_j = \sum_{i \in \mathsf{P}_j} a_{ij} X_i + U_j \;\; \forall j \in \mathsf{N}^+ \setminus \{\mathsf{N}_0 \cup 0\} \tag{6a}$$

$$\mathbb{E}[\mathbf{U}] = 0. \tag{6b}$$

Note that this is a classical setting for structural equations in DAGs (see Peters et al. (2017) for example). Denote by $\mathsf{E}_j^{\text{in}} := \{(i,j) : i \in \mathsf{P}_j\}$ the incoming edges of node $j \in \mathsf{N}^+ \setminus \{0\}$. Define the *forward-looking* weights $\boldsymbol{c}$ as follows:

$$c_{j,n+1} := a_{j,n+1} \qquad \forall (j, n+1) \in \mathsf{E}_{n+1}^{\text{in}} \tag{7a}$$

$$c_{ij} := a_{ij} \sum_{k \in \mathsf{C}_j} c_{jk} \qquad \forall (i,j) \in \mathsf{E} \setminus \mathsf{E}_{n+1}^{\text{in}}, \tag{7b}$$

where $a_{0j} := x_j^{(2)} - x_j^{(1)} \;\forall (0,j) \in \mathsf{E}_0$. Similarly, define the *backward-looking* weights $\boldsymbol{b}$ as follows:

$$b_{0j} := x_j^{(2)} - x_j^{(1)} \qquad \forall (0,j) \in \mathsf{E}_0 \tag{8a}$$

$$b_{jk} := \sum_{i \in \mathsf{P}_j} b_{ij} a_{jk} \qquad \forall (j,k) \in \mathsf{E} \setminus \mathsf{E}_0. \tag{8b}$$

Since G is a DAG, weights (7) and (8) are well-defined. In such a linear setting, RSV can be characterized as in Theorem 6.

Theorem 6 *Consider structural causal model $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$, wherein the structural equations F are linear and the noise distribution D has zero mean (see (6)), with $\boldsymbol{c}$ and $\boldsymbol{b}$ as in (7) and (8). Then,*

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

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

A proof is given in Appendix D and illustrations are provided in Example 3 (§4.3) and the motivating example (§1). Furthermore, this characterization establishes a tight connection between RSV and most existing approaches, as alluded to in Remark 5 earlier.

Remark 7 (Computational implication of Theorem 6) *Given the characterization in Theorem 6, it follows that RSV can be computed in linear time and exhibits linear space complexity. Note that topological sorting of a DAG requires $\mathcal{O}(|\mathsf{N}|+|\mathsf{E}|)$ time. Given a topologically sorted graph, $\boldsymbol{c}$ can be computed in $\mathcal{O}(|\mathsf{N}|+|\mathsf{E}|)$ time by proceeding in a reverse topological order and storing local computations at each node ($\mathcal{O}(|\mathsf{N}|)$ additional space):*

$$c_j := \sum_{k \in \mathsf{C}_j} c_{jk} \ \forall j \in \mathsf{N}.$$

Similarly, $\boldsymbol{b}$ can be computed in $\mathcal{O}(|\mathsf{N}|+|\mathsf{E}|)$ time. Finally, given $\boldsymbol{c}$ and $\boldsymbol{b}$, RSV can be computed in $\mathcal{O}(|\mathsf{N}|+|\mathsf{E}|)$ time by using the characterization in Theorem 6. This highlights the tractability of RSV for linear models.

5.2 Non-parametric SEM

We now characterize RSV for an arbitrary set of structural equations (possibly non-linear). We start by illustrating the intuition of the derivation, through the DAG depicted in Figure 15. Our discussion highlights the role of paths in the DAG and the notion of "sister edges" that prove critical in the calculation of RSV.

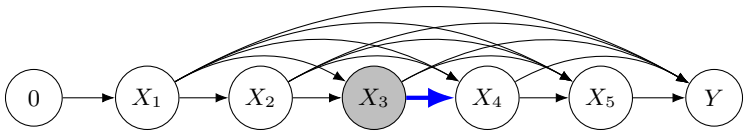


Figure 15: DAG for illustrating the path-based characterization of RSV.

Pick an arbitrary edge (say the thick blue edge $(3,4)$ in Figure 15) and suppose we are interested in computing the corresponding RSV $\pi_{34}^{\text{RSV}}(\mathsf{E})$, with $\mathsf{E} = (\mathsf{E}_0, \ldots, \mathsf{E}_5)$ denoting the set of all edges. By definition (recall from §4.2), $\pi_{34}^{\text{RSV}}(\mathsf{E})$ is the SV of the following game. The set of players is $\mathsf{E}_3 = \{(3,4),(3,5),(3,6)\}$ and given coalition $E_3 \subseteq \mathsf{E}_3$, the characteristic function is given by

$$\begin{aligned} v_3(E_3) &= \pi_3(\mathsf{E}_0, \mathsf{E}_1, \mathsf{E}_2, E_3, \mathsf{E}_4, \mathsf{E}_5) \\ &= \pi_3(E_3, \mathsf{E}_{-3}) \\ &= \pi_{13}(E_3, \mathsf{E}_{-3}) + \pi_{23}(E_3, \mathsf{E}_{-3}). \end{aligned}$$

Hence, $\pi_{34}^{\text{RSV}}(\mathsf{E})$ equals

$$\begin{aligned} &= \sum_{E_3 \subseteq \mathsf{E}_3 \setminus \{(3,4)\}} w_{\mathsf{E}_3}(E_3) \times \{v_3(E_3 \cup \{(3,4)\}) - v_3(E_3)\} \\ &= \sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_3}(E_3) \times \{v_3(E_3 \cup \{(3,4)\}) - v_3(E_3)\} \\ &= \underbrace{\sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_3}(E_3) \times \{\pi_{13}(E_3 \cup \{(3,4)\}) - \pi_{13}(E_3)\}}_{(\star)} \end{aligned}$$

$$+ \underbrace{\sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_3}(E_3) \times \{\pi_{23}(E_3 \cup \{(3,4)\}) - \pi_{23}(E_3)\}}_{(\square)}, \tag{9}$$

where we use the short-hand notation $\mathsf{E}_i^j := \mathsf{E}_i \setminus \{(i,j)\}$ for arbitrary node $i \in \mathsf{N}^+$ and its child node $j \in \mathsf{C}_i$.

Next, we analyze the first term in (9), i.e., $(\star)$. Given $E_3 \subseteq \mathsf{E}_3$, observe that $\pi_{13}(E_3)$ denotes the flow (RSV) on edge $(1,3)$ when only E_3 edges are active at node 3, i.e.,

$$\begin{aligned}\pi_{13}(\mathsf{E}_0, \mathsf{E}_1, \mathsf{E}_2, E_3, \mathsf{E}_4, \mathsf{E}_5) &= \sum_{E_1 \subseteq \mathsf{E}_1^3} w_{\mathsf{E}_1}(E_1) \times \{v_1(E_1 \cup \{(1,3)\} \mid E_3) - v_1(E_1 \mid E_3)\} \\ &= \sum_{E_1 \subseteq \mathsf{E}_1^3} w_{\mathsf{E}_1}(E_1) \times \{\pi_{01}(E_1 \cup \{(1,3)\}, E_3) - \pi_{01}(E_1, E_3)\}. \end{aligned} \tag{10}$$

Plugging (10) into $(\star)$, we get

$$\begin{aligned}(\star) &= \sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_3}(E_3) \times \{\pi_{13}(E_3 \cup \{(3,4)\}) - \pi_{13}(E_3)\} \\ &= \sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_3}(E_3) \Bigg\{ \sum_{E_1 \subseteq \mathsf{E}_1^3} w_{\mathsf{E}_1}(E_1) \{\pi_{01}(E_1 \cup \{(1,3)\}, E_3 \cup \{(3,4)\}) \\ &\qquad - \pi_{01}(E_1, E_3 \cup \{(3,4)\})\} - \sum_{E_1 \subseteq \mathsf{E}_1^3} w_{\mathsf{E}_1}(E_1) \{\pi_{01}(E_1 \cup \{(1,3)\}, E_3) - \pi_{01}(E_1, E_3)\} \Bigg\} \\ &= \sum_{E_1 \subseteq \mathsf{E}_1^3} \sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_1}(E_1) w_{\mathsf{E}_3}(E_3) \{\pi_{01}(E_1 \cup \{(1,3)\}, E_3 \cup \{(3,4)\}) - \pi_{01}(E_1, E_3 \cup \{(3,4)\}) \\ &\qquad - \pi_{01}(E_1 \cup \{(1,3)\}, E_3) + \pi_{01}(E_1, E_3)\}. \end{aligned} \tag{11}$$

Now, given $E_1 \subseteq \mathsf{E}_1$ and $E_3 \subseteq \mathsf{E}_3$, observe that $\pi_{01}(E_1, E_3)$ equals $\pi_{01}(\mathsf{E}_0, E_1, \mathsf{E}_2, E_3, \mathsf{E}_4, \mathsf{E}_5)$, which equals

$$\begin{aligned} &= \sum_{E_0 \subseteq \mathsf{E}_0^1} w_{\mathsf{E}_0}(E_0) \times \{v_0(E_0 \cup \{(0,1)\} \mid E_1, E_3) - v_0(E_0 \mid E_1, E_3)\} \\ &= \sum_{E_0 \subseteq \mathsf{E}_0^1} w_{\mathsf{E}_0}(E_0) \times \{y(E_0 \cup \{(0,1)\}, E_1, \mathsf{E}_2, E_3, \mathsf{E}_4, \mathsf{E}_5) - y(E_0, E_1, \mathsf{E}_2, E_3, \mathsf{E}_4, \mathsf{E}_5)\}, \end{aligned} \tag{12}$$

where $y(E)$ is the expected outcome as a function of the set of active edges $E \subseteq \mathsf{E}$ (see (3)). Plugging (12) into (11) and re-arranging gives us

$$\begin{aligned}(\star) = \sum_{E_0 \subseteq \mathsf{E}_0^1} \sum_{E_1 \subseteq \mathsf{E}_1^3} \sum_{E_3 \subseteq \mathsf{E}_3^4} & w_{\mathsf{E}_0}(E_0) w_{\mathsf{E}_1}(E_1) w_{\mathsf{E}_3}(E_3) \\ & \times \{y(E_0 \cup \{(0,1)\}, E_1 \cup \{(1,3)\}, E_3 \cup \{(3,4)\}) - y(E_0 \cup \{(0,1)\}, E_1 \cup \{(1,3)\}, E_3) \\ & \quad - y(E_0 \cup \{(0,1)\}, E_1, E_3 \cup \{(3,4)\}) - y(E_0, E_1 \cup \{(1,3)\}, E_3 \cup \{(3,4)\}) \\ & \quad + y(E_0 \cup \{(0,1)\}, E_1, E_3) + y(E_0, E_1 \cup \{(1,3)\}, E_3) + y(E_0, E_1, E_3 \cup \{(3,4)\}) \\ & \quad - y(E_0, E_1, E_3)\}. \end{aligned} \tag{13}$$

There is a total of eight $y(\cdot)$ terms in (13). Observe the following pattern among them. The first term

$$y(E_0 \cup \{(0,1)\}, E_1 \cup \{(1,3)\}, E_3 \cup \{(3,4)\})$$

includes each of the three edges $\{(0,1),(1,3),(3,4)\}$ (3 choose 3) and has a positive sign in front of it. Each of the next three terms

$$\begin{aligned} &- y(E_0 \cup \{(0,1)\}, E_1 \cup \{(1,3)\}, E_3), \\ &- y(E_0 \cup \{(0,1)\}, E_1, E_3 \cup \{(3,4)\}), \text{ and} \\ &- y(E_0, E_1 \cup \{(1,3)\}, E_3 \cup \{(3,4)\}) \end{aligned}$$

include two of the three edges (3 choose 2) and has a negative sign. The next three terms

$$\begin{aligned} &+ y(E_0 \cup \{(0,1)\}, E_1, E_3), \\ &+ y(E_0, E_1 \cup \{(1,3)\}, E_3), \text{ and} \\ &+ y(E_0, E_1, E_3 \cup \{(3,4)\}) \end{aligned}$$

include one of the three edges (3 choose 1) and are positive. The last term

$$-y(E_0, E_1, E_3)$$

includes none of the three edges (3 choose 0) and is negative.

The previous illustration reveals certain patterns that hold in the general case. First, we introduce concise notation to summarize the previous calculation. Clearly, $(\star)$ corresponds to the path $0 \to 1 \to 3$ from source node 0 to node 3. Accordingly, define $W := (0,1,3)$ to be the corresponding (ordered) set of nodes in this path. We use the short-hand

$$\sum_{E_{013} \subseteq \mathsf{E}_{013}^{134}} \kappa_{013}(E_{013}) := \sum_{E_0 \subseteq \mathsf{E}_0^1} \sum_{E_1 \subseteq \mathsf{E}_1^3} \sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_0}(E_0) w_{\mathsf{E}_1}(E_1) w_{\mathsf{E}_3}(E_3). \tag{14}$$

for the summation over the weights. In the illustrative example we have here, $\mathsf{E}_{013}^{134} := \mathsf{E}_0^1 \cup \mathsf{E}_1^3 \cup \mathsf{E}_3^4$, $E_{013} := (E_0, E_1, E_3)$, and $\kappa_{013}(E_{013}) := w_{\mathsf{E}_0}(E_0) w_{\mathsf{E}_1}(E_1) w_{\mathsf{E}_3}(E_3)$.

Next, in the general case, given path W, we use the notation

$$\sum_{E_W \subseteq \mathsf{E}_W^{W \cup \{j\} \setminus \{0\}}} \kappa_W(E_W),$$

where j denotes the end node (4 in the above example) of the edge we are interested in ($(3,4)$ in the example). As a side note, observe that

$$\sum_{E_W \subseteq \mathsf{E}_W^{W \cup \{j\} \setminus \{0\}}} \kappa_W(E_W) = 1 \tag{15}$$

since each of the $w_{\mathsf{E}_k}(E_k)$ term decouples and sums to 1 (by definition of the SV weight function). The latter fact is useful for developing a Monte-Carlo estimation scheme (§6). Next, we express the 2^3 "$y(\cdot)$" terms in $(\star)$ (recall (13)). Recycling $W = (0,1,3)$ and

defining $\texttt{Edge}(W) := \{(0,1),(1,3)\}$ to be the set of edges in path W, these $y(\cdot)$ terms can be expressed as follows (cf. the pattern discussed above):

$$\sum_{V \subseteq \texttt{Edge}(W \cup \{4\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathsf{E}_{-W}).$$

In summary, given path $W = (0,1,3)$ corresponding to $(\star)$, we have

$$(\star) = \sum_{E_W \subseteq \mathsf{E}_W^{W \cup \{4\} \setminus \{0\}}} \kappa_W(E_W) \sum_{V \subseteq \texttt{Edge}(W \cup \{4\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathsf{E}_{-W}). \tag{16}$$

Similarly, $(\square)$ from (9) equals

$$(\square) = \sum_{E_W \subseteq \mathsf{E}_W^{W \cup \{4\} \setminus \{0\}}} \kappa_W(E_W) \sum_{V \subseteq \texttt{Edge}(W \cup \{4\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathsf{E}_{-W}), \tag{17}$$

where the only difference is that path W equals $(0,1,2,3)$, instead of $(0,1,3)$. Putting (16) and (17) together, we get

$$\pi_{34}^{\text{RSV}}(\mathsf{E}) = \sum_{W \in \mathsf{W}_3} \sum_{E_W \subseteq \mathsf{E}_W^{W \cup \{4\} \setminus \{0\}}} \kappa_W(E_W) \sum_{V \subseteq \texttt{Edge}(W \cup \{4\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathsf{E}_{-W}), \tag{18}$$

where $\mathsf{W}_3 := \{(0,1,3),(0,1,2,3)\}$ is the set of all unique paths from node 0 to node 3. As we prove in Appendix E, this characterization holds in general and we state it formally in Theorem 8.

Theorem 8 *Given structural causal model $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$, the RSV of edge $(i,j) \in \mathsf{E}$ exhibits the following characterization:*

$$\pi_{ij}^{RSV}(\mathsf{E}) = \sum_{W \in \mathsf{W}_i} \sum_{E_W \subseteq \mathsf{E}_W^{W \cup \{j\} \setminus \{0\}}} \kappa_W(E_W) \sum_{V \subseteq \textit{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathsf{E}_{-W}).$$

Hence, the recursion in RSV (recall Algorithms 1 and 2) can be boiled down to a path-based expression, wherein we sum over each unique path $W \in \mathsf{W}_i$ from node 0 to node i. For each such path W, we consider the subset of edges E_W that excludes the edges in this path, i.e., E_W contains the "sister edges" of the edges in W. We weigh E_W by a corresponding factor $\kappa_W(E_W)$ and evaluate the value-add of all possible subsets V of the edges in W (to E_W). That is, how much value the edges in the path W add to their "sister edges" E_W.

Remark 9 (Computational implication of Theorem 8) *It should be clear from the characterization in Theorem 8 that in the worst case, computing RSV takes an exponential number of operations in the number of edges. For example, consider the characterization for edge $(5,6)$ in Figure 15 and focus on the path $W = (0,1,2,3,4,5)$. The $\sum_{E_W \subseteq \mathsf{E}_W^{W \cup \{j\} \setminus \{0\}}}$ sum expands into roughly $2^{|\mathsf{E}|}$ terms (all possible subsets of E). Since a dense DAG has $\mathcal{O}(n^2)$ edges (where n is the number of nodes), we get the worst-case run-time to be $\mathcal{O}(2^{n^2})$. Further, techniques such as dynamic programming can not reduce this worst-case runtime since each of the $\mathcal{O}(2^{n^2})$ term can be unique in general.*

Naturally, such a high runtime is undesirable and the following two questions are of interest:

1. Can we develop a computationally tractable characterization for a special class of models?

2. Can we *estimate* RSV and reduce the exponential dependence on the number of edges?

For the first question, Theorem 6 provides a positive answer for the class of linear SEMs. However, as we discuss in Appendix F, developing a tractable characterization even for a linear model with two-way interactions [7] is challenging. Therefore, we focus on developing an estimation scheme for RSV next, which leverages the path-based characterization in Theorem 8.

6. Estimation of RSV

Our estimation scheme along with its properties (unbiasedness, rate of decay of its variance, and sample complexity) is discussed in §6.1 followed by a numerical illustration in §6.2. Note that the model M is assumed to be known (as it is throughout the paper) and our focus here is on estimating the RSV (for a given model M).

6.1 Estimation scheme and properties

Suppose we are interested in estimating $\pi_{ij}^{\text{RSV}}(\mathsf{E})$ for an arbitrary edge $(i,j) \in \mathsf{E}$. Recalling the path-based characterization (Theorem 8), consider an arbitrary path $W \in \mathsf{W}_i$.

The key computational bottleneck is the sum $\sum_{E_W \subseteq \mathsf{E}_W^{W\cup\{j\}\setminus\{0\}}}$ since it enumerates all possible subsets of the "sister edges". To estimate this sum, we leverage the fact that the kappa factor sums to 1 (recall (15)) and is non-negative (and hence, can be interpreted as a probability), which leads to a natural Monte-Carlo estimation scheme. Define

$$\mu_W^j := \sum_{E_W \subseteq \mathsf{E}_W^{W\cup\{j\}\setminus\{0\}}} \kappa_W(E_W) \sum_{V \subseteq \texttt{Edge}(W\cup\{j\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathsf{E}_{-W}) \tag{19}$$

so that π_{ij}^{RSV} equals $\sum_{W\in\mathsf{W}_i} \mu_W^j$ (cf. Theorem 8). We focus on estimating μ_W^j, for which we propose the following strategy:

1. Instead of summing over all $E_W \subseteq \mathsf{E}_W^{W\cup\{j\}\setminus\{0\}}$ (which can be $\mathcal{O}(2^{n^2})$), sample E_W with corresponding "probability" $\kappa_W(E_W)$. To do so, we can sample using the independent structures in $\mathsf{E}_W^{W\cup\{j\}\setminus\{0\}}$ (each independent structure corresponds to a node in path W) and use the SV weight function (recall the relation in (14)) to leverage existing SV estimation techniques (Castro et al., 2009; Maleki et al., 2013; Castro et al., 2017; Mitchell et al., 2022). For example, given $W = (0,1,3)$ in Figure 15, observe from

7. That is, $f(X_j) = \sum_{i\in\mathsf{P}_j} a_i^j X_i + \sum_{i,i'\in\mathsf{P}_j} a_{ii'}^j X_i X_{i'}$; arguably the "closest" non-linear model to a linear model, since the model remains linear in the coefficients $\boldsymbol{a}$ and quadratic in the variables $\mathbf{X}$. The superscripts in the coefficients (e.g., "j" in a_i^j) do not represent exponents but are used to indicate the child node of the corresponding edge.

(14) that the sum $\sum_{E_W \subseteq \mathsf{E}_W^{W\cup\{j\}\setminus\{0\}}} \kappa_W(E_W)$ decomposes as follows ("independent structures"):

$$\begin{aligned}\sum_{E_W \subseteq \mathsf{E}_W^{W\cup\{j\}\setminus\{0\}}} \kappa_W(E_W) &= \sum_{E_{013} \subseteq \mathsf{E}_{013}^{134}} \kappa_{013}(E_{013}) \\ &= \sum_{E_0 \subseteq \mathsf{E}_0^1} \sum_{E_1 \subseteq \mathsf{E}_1^3} \sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_0}(E_0) w_{\mathsf{E}_1}(E_1) w_{\mathsf{E}_3}(E_3) \\ &= \sum_{E_0 \subseteq \mathsf{E}_0^1} w_{\mathsf{E}_0}(E_0) \sum_{E_1 \subseteq \mathsf{E}_1^3} w_{\mathsf{E}_1}(E_1) \sum_{E_3 \subseteq \mathsf{E}_3^4} w_{\mathsf{E}_3}(E_3).\end{aligned}$$

Accordingly, we can first sample $E_0 \subseteq \mathsf{E}_0^1$ using the SV weight $w_{\mathsf{E}_0}(E_0)$, then $E_1 \subseteq \mathsf{E}_1^3$ using the SV weight $w_{\mathsf{E}_1}(E_1)$, followed by $E_3 \subseteq \mathsf{E}_3^4$ using the SV weight $w_{\mathsf{E}_3}(E_3)$ and finally, use (E_0, E_1, E_3) as a sample E_W. Since each of the sampling probabilities ($w_{\mathsf{E}_0}(E_0)$, $w_{\mathsf{E}_1}(E_1)$, and $w_{\mathsf{E}_3}(E_3)$) corresponds to the SV weight function, we can use existing Monte-Carlo techniques to generate such samples of E_0, E_1, and E_3 (Castro et al., 2009; Maleki et al., 2013; Castro et al., 2017; Mitchell et al., 2022). As one possibility, observing that for an arbitrary function $h(\cdot)$,

$$\begin{aligned}\sum_{E_1 \subseteq \mathsf{E}_1^3} w_{\mathsf{E}_1}(E_1) \times h(E_1) &= \sum_{E_1 \subseteq \mathsf{E}_1^3} \frac{|E_1|!\,(|\mathsf{E}_1| - |E_1| - 1)!}{|\mathsf{E}_1|!} \times h(E_1) \qquad [|\mathsf{E}_1^3| = |\mathsf{E}_1| - 1] \\ &= \frac{1}{|\mathsf{E}_1|} \sum_{k=0}^{|\mathsf{E}_1^3|} \frac{1}{\binom{|\mathsf{E}_1^3|}{k}} \sum_{\{E_1 \subseteq \mathsf{E}_1^3 : |E_1| = k\}} h(E_1),\end{aligned}$$

we can sample a coalition size first from a uniform distribution ("$1/|\mathsf{E}_1|$" factor) and then a coalition (among the coalitions of the sampled size) from a uniform distribution ("$1/(|\mathsf{E}_1^3|$ choose k)" factor). That is, to sample $E_1 \subseteq \mathsf{E}_1^3$, we will first sample the coalition size

$$k \sim \texttt{Uniform}(\{0, \ldots, |\mathsf{E}_1^3|\})$$

and then sample a coalition

$$E_1 \sim \texttt{Uniform}(\mathsf{E}_1^3(k))$$

from the set of all coalitions of this size, i.e., $\mathsf{E}_1^3(k) := \{\mathcal{E}_1 \subseteq \mathsf{E}_1^3 : |\mathcal{E}_1| = k\}$. Note that the set $\mathsf{E}_1^3(k)$ contains $|\mathsf{E}_1^3|$ choose k elements. Fortunately, we do not need to enumerate all possible elements to sample E_1. Instead, to generate a uniform sample of size k from E_1^3, we can uniformly sample k elements from E_1^3 without replacement. Repeating the process to sample E_0 and E_3 results in (E_0, E_1, E_3), which is an unbiased sample of E_W.

2. Use the sampled E_W to compute the inner term in (19):

$$\hat{\mu}_W^j(s) := \sum_{V \subseteq \texttt{Edge}(W \cup \{j\})} (-1)^{|W| - |V|} \times y(E_W \cup V, \mathsf{E}_{-W}),$$

where s denotes the Monte-Carlo sample number (corresponding to the sample E_W).

3. Repeat steps 1 and 2 for $s = 1, \ldots, S$ (for relatively large S, the number of Monte-Carlo samples) and return the average, which we denote by

$$\hat{\mu}_W^j := \frac{1}{S} \sum_{s=1}^{S} \hat{\mu}_W^j(s).$$

Since π_{ij}^{RSV} equals $\sum_{W \in \mathsf{W}_i} \mu_W^j$, a natural estimate for π_{ij}^{RSV} is

$$\hat{\pi}_{ij}^{\text{RSV}} := \sum_{W \in \mathsf{W}_i} \hat{\mu}_W^j.$$

We summarize our estimation scheme in Algorithm 3 along with the sub-routine Algorithm 4 to sample "sister edges".

Algorithm 3 Estimating π_{ij}^{RSV} for a given edge $(i,j) \in \mathsf{E}$ using S Monte-Carlo samples

for $W \in \mathsf{W}_i$
  for $s = 1, \ldots, S$
    $E_W = \texttt{sample\_sister\_edges}(W \cup \{j\})$
    $\hat{\mu}_W^j(s) = \sum_{V \subseteq \texttt{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathsf{E}_{-W})$
  end for
end for
$\hat{\mu}_W^j = \frac{1}{S} \sum_{s=1}^{S} \hat{\mu}_W^j(s)$
return $\hat{\pi}_{ij}^{\text{RSV}} = \sum_{W \in \mathsf{W}_i} \hat{\mu}_W^j$

Algorithm 4 $\texttt{sample\_sister\_edges}(W \cup \{j\})$

for node $h \in W$
  $k \sim \texttt{Uniform}(\{0, \ldots, |\mathsf{E}_h| - 1\})$
  $E_h \sim \texttt{Uniform}(\mathsf{E}_h^{h^+}(k))$ % h^+ denotes the node that follows h in path $W \cup \{j\}$
end for
return $E_W = (E_h)_{h \in W}$

By construction of our Monte-Carlo scheme, $\hat{\pi}_{ij}^{\text{RSV}}$ is an unbiased estimator of π_{ij}^{RSV} and its variance decays at a rate of $1/S$. It is also possible to provide a high probability concentration bound on the quality of our estimate. We summarize such properties in Proposition 10, with a proof in Appendix G.

Proposition 10 *For edge $(i,j) \in \mathsf{E}$, $\hat{\pi}_{ij}^{RSV}$ is an unbiased estimator of π_{ij}^{RSV} with variance decaying at a rate of $1/S$. Furthermore, for all $t \geq 0$,*

$$\mathbb{P}\left\{\left|\pi_{ij}^{RSV} - \hat{\pi}_{ij}^{RSV}\right| \geq t\right\} \leq 2\exp\left(-\beta_{ij} S t^2\right)$$

for some $\beta_{ij} > 0$ finite. (The probability measure $\mathbb{P}\{\cdot\}$ here denotes the Monte-Carlo sampling distribution of the estimator $\hat{\pi}_{ij}^{RSV}$.)

An exponentially decaying sample complexity (w.r.t. number of samples S) is highly desirable. The proposed Monte-Carlo scheme tackles the key computational bottleneck in the path-based characterization, i.e., the sum $\sum_{E_W \subseteq \mathsf{E}_W^{W\cup\{j\}\setminus\{0\}}}$. Though this eliminates the $\mathcal{O}(2^{n^2})$ dependence, we still require $\mathcal{O}(2^n)$ number of operations in the worst-case, since each of the remaining two sums $\sum_{W\in\mathsf{W}_i}$ and $\sum_{V\subseteq\texttt{Edge}(W\cup\{j\})}$ can be such. We note that it is possible to develop polynomial-time Monte-Carlo estimation schemes to estimate both the sums via techniques such as importance sampling. For example, to estimate the sum $\sum_{W\in\mathsf{W}_i}$, we can multiply and divide by $|\mathsf{W}_i|$ and use the "probability" weight $1/|\mathsf{W}_i|$ to sample W. This requires us to sample uniformly from W_i, which can be done in linear time and space (see Appendix H for one procedure that does so). Similar desirable properties such as in Proposition 10 apply to such sequential Monte-Carlo estimation as well. Furthermore, it is worth mentioning that estimating SV is an active area of research (see Mitchell et al. (2022) for a recent work). Given our estimation scheme for RSV involves sampling SV coalitions ("sister edges") as a sub-routine (recall Algorithm 4), such advances can be leveraged to further enhance our strategy.

6.2 Numerical illustration of Algorithms 3 and 4

We consider a fully connected layered graph with L layers and M nodes per layer (see Figure 16 for example). We assume each non-source node to be a product of all its parents, e.g., $X_4 = X_1X_2X_3$ (non-linear SEM) in Figure 16. In general,

$$X_j = \prod_{i\in\mathsf{P}_j} X_i \;\forall j \in \mathsf{N}^+ \setminus \{\mathsf{N}_0 \cup 0\}.$$

For each source node in N_0, we use a background and foreground of 0 and 1, respectively. This stylized setup enables us to know the exact RSV without resorting to a brute force calculation. In particular, the axioms of flow conservation and flow symmetry prove useful. For example, in Figure 16, the effect (change in Y) equals 1, with each outgoing edge of node 0 receiving a flow of 1/3, each outgoing edge in layer 1 (edges coming out of nodes 1, 2, and 3) receiving a flow of 1/9, and each edge in the last layer receiving 1/3. In general, given the setup, the exact RSV equals $1/M$ for the edges coming out of node 0 and going in to node Y, whereas it equals $1/M^2$ for all other edges. Knowing the exact RSV is useful for our numerical exercise since it enables us to evaluate the quality of our estimate. (Though not the focus of this subsection, note that essentially all existing approaches attribute zero to every edge or path, again highlighting their limitation to explain the total effect in the presence of interactions.)

The goal is to evaluate the quality of the proposed RSV estimation scheme given a layered DAG of size (L, M). For ease of presentation, we focus on the top-right edge in the DAG (e.g., the blue thick edge $(4, Y)$ in Figure 16). Note that this edge is most computationally demanding, since compute time grows with the depth of the DAG.

Figure 17a depicts the Monte-Carlo distribution of the estimate (white bars) and the exact RSV (blue vertical line) for a layered graph with $(L, M) = (3, 3)$ for $S \in \{100, 200, \ldots, 900\}$ Monte-Carlo samples. The distribution is based on 100 runs for each value of S. It can be seen that mean estimate is very close to the theoretical RSV of 1/3, even for low values of S, which is in accordance with the unbiased nature of the estimate. Further, the vari-

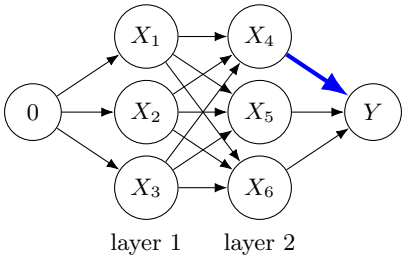
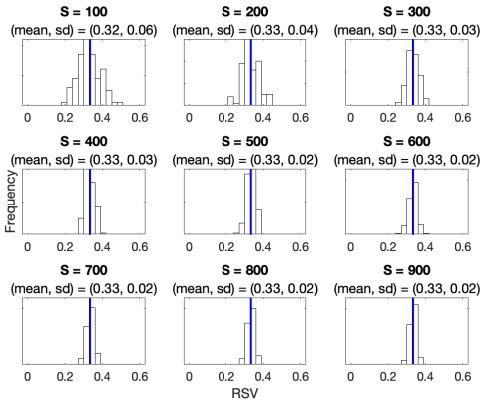
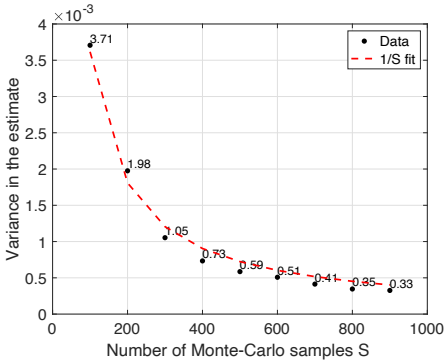


Figure 16: A layered graph with $L = 2$ layers and $M = 3$ nodes per layer. For conciseness, we always evaluate the quality of our estimate on the top-right edge in the graph (e.g., the blue thick edge $(4, Y)$ in this figure).

ance decays as S increases, as seen in Figure 17b and the rate of decay is $1/S$ as stated in Proposition 10.





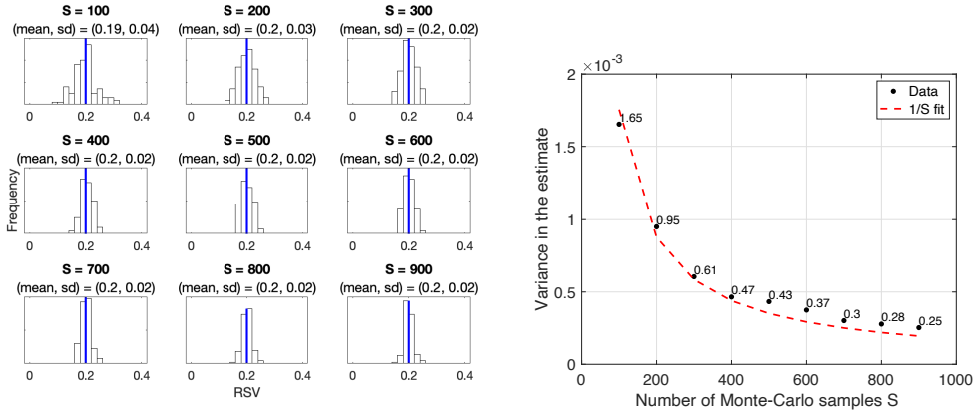
(a) Estimator distribution **(b)** Variance as a function of S

Figure 17: Numerical illustration of our estimation scheme for $(L, M) = (3, 3)$. In plot (a), we show the Monte-Carlo distribution of our estimate (white bars) and the exact RSV (blue vertical line) for $S \in \{100, 200, \ldots, 900\}$ Monte-Carlo samples. The exact RSV (of the top-right edge) equals $1/M$, which equals $1/3$ in this case. At the top of each subplot, we report the mean and the standard deviation of the corresponding distribution. In plot (b), we show the variance of the Monte-Carlo estimate as a function of samples S. The black dots correspond to the square of the standard deviation numbers reported in plot (a) (at the top of each subplot). The dotted red line is a regression fitted on the black dots ($y = \frac{\beta}{x}$), highlighting that the variance decays at a rate of $1/S$.

In Figures 18 and 19, we report the results for $M = 5$ and $M = 7$, respectively (while holding $L = 3$). All previous findings regarding the unbiased nature of the Monte-Carlo estimate and the rate of decay of its variance as a function of S continue to hold. Results from experiments involving an increasing number of layers from $L = 3$ to $L \in \{4, 5\}$ (for $M \in \{3, 5, 7\}$ as above) are given in Appendix I. The additional experimental results

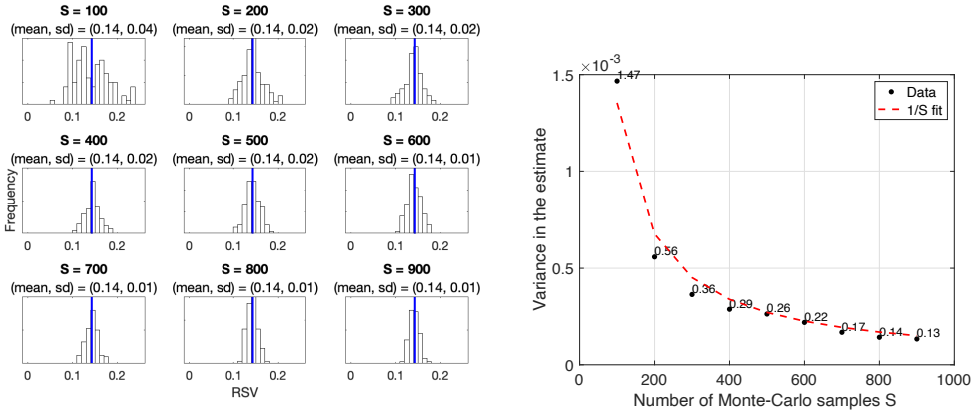
highlight the robustness of the proposed estimation scheme to the number of nodes per layer and number of layers, respectively.

Note that even for $(L, M) = (3, 7)$, a brute force calculation of the RSV requires a long computing time, since the latter is exponential in the number of edges which is around 100. On the other hand, for the same setting, the Monte-Carlo estimate takes ~ 10 seconds for $S = 100$ and ~ 90 seconds for $S = 900$, which highlights the computational benefit to estimate the $\sum_{E_W \subseteq \mathsf{E}_W^{W\cup\{j\}\setminus\{0\}}}$ sum. Note that the compute time of our estimation scheme increases (exponentially) in the depth L (e.g., around 7 hours for $(L, M) = (5, 7)$ with $S = 900$), primarily due to the other two sums in the path-based characterization ($\sum_{W\in\mathsf{W}_i}$ and $\sum_{V\subseteq\mathtt{Edge}(W\cup\{j\})}$). As previously mentioned, it can be controlled via importance sampling enhancements. Also worth noting, the Monte-Carlo estimate is trivially parallelizable and the compute times can be significantly reduced in multi-core environments.



(a) Estimator distribution **(b)** Variance as a function of S

Figure 18: Numerical illustration of our estimation scheme for $(L, M) = (3, 5)$. Note that the exact RSV (of the top-right edge) equals $1/M$, which equals $1/5$ in this case (blue vertical line in plot (a)). The estimator's mean is close to 0.2 for all values of S and the variance decays at a rate of $1/S$.



(a) Estimator distribution **(b)** Variance as a function of S

Figure 19: Numerical illustration of our estimation scheme for $(\boldsymbol{L}, \boldsymbol{M}) = (\mathbf{3}, \mathbf{7})$. Note that the exact RSV (of the top-right edge) equals $1/M$, which equals $1/7$ in this case (blue vertical line in plot (a)). The estimator's mean is close to 0.14 for all values of S and the variance decays at a rate of $1/S$.

7. Applications

We showcase the application of the RSV framework on two challenging problems in causality, namely, causal overdetermination in §7.1 and causal unfairness in §7.2 and benchmark it with appropriate existing methodologies.

7.1 Causal overdetermination

The rock-throwing example is a textbook application introduced by Hall (2004) that helps illustrate the robustness of RSV to model changes (implementation invariance) and compare it with existing methods on causality. In the words of Chockler and Halpern (2004), the setup is as follows:

> "Suppose that Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle."

Each throw is assumed to be perfectly accurate and hence, is sufficient but not necessary to shatter the bottle (*causal overdetermination*). A corresponding DAG is shown in Figure 20a with the following structural equation for the outcome BS (bottle shatters):

$$BS = \mathbb{I}\{ST = 1 \text{ or } BT = 1\}.$$

$ST \in \{0, 1\}$ denotes whether Suzy throws (1) or not (0) whereas $BT \in \{0, 1\}$ denotes whether Billy throws (1) or not (0), and $\mathbb{I}\{\cdot\}$ denotes the indicator function. The background for the source variables (ST, BT) is $(0, 0)$ and the foreground is $(1, 1)$.

As discussed in §1, in this work, we care about the cause of the bottle shattering from a blame / ex post perspective. In this simple model, the degree of responsibility (Chockler

(a) Model 1 **(b)** Model 2

Figure 20: Two models for the rock-throwing application.

and Halpern, 2004) of Suzy is 1/2 and of Billy is 1/2, which is consistent with RSV[8]. However, techniques such as direct and indirect effects (Pearl, 2001), path-specific effects (Pearl, 2001), and path-specific selection gradient (Henshaw et al., 2020) attribute 1 to both Suzy and Billy, resulting in an over-allocation, i.e., they attribute more than the total effect. Note that in this example, such approaches give the same decomposition as RSV up to a normalization constant, but this is not true in general (cf. Example 1 in §3).

Chockler and Halpern (2004) also discuss another way to model this setting (Figure 20b). They "implicitly assume a context where Suzy throws first, so there is an edge from SH [Suzy hits] to BH [Billy hits], but not one in the other direction". The underlying structural equations are as follows:

$$\begin{aligned} SH &= \mathbb{I}\{ST = 1\} \\ BH &= \mathbb{I}\{BT = 1 \text{ and } SH = 0\} \\ BS &= \mathbb{I}\{SH = 1 \text{ or } BH = 1\}. \end{aligned}$$

Interestingly, the degree of responsibility for Suzy remains at 1/2, but for Billy drops to 0 under the "restriction to allowable settings" (Chockler and Halpern, 2004)[9]. Clearly, the decomposition does not add up to the total effect of 1. In a more extreme setting with $k-1$ people throwing rocks in addition to Suzy (as opposed to just Billy), the degree of responsibility equals $1/k$ for Suzy and 0 for all others. As $k \to \infty$, the degree of responsibility over all people sums to 0. To put it another way, if an infinite number of people harm a person, then according to the degree of responsibility framework, no one is responsible! This is rather strange.

There are two possible ways to interpret the dynamics here. In the first one, to capture the fact that "Suzy throws first", we can model the change in source variables as a sequence of two changes, i.e., the source variables (ST, BT) change from $(0,0)$ to $(1,0)$ to $(1,1)$, as opposed to a direct change from $(0,0)$ to $(1,1)$. Under such a view, the outcome BS changes from 0 to 1 to 1 and it is easy to verify that all approaches attribute 1 to Suzy (as a result of the first change) and 0 to Billy (as a result of the second change), irrespective

8. In this simple example, a standard Shapley value framing would suffice and hence, RSV can be seen as a generalization of SV to graphs with mediators. We discussed this in §1 as well. The fact that RSV recovers SV when appropriate should be seen in positive light. In fact, we formalized such connections in the preliminary version of this work; see Proposition 3 and Proposition B in Singal et al. (2021).
9. From §5 of Chockler and Halpern (2004): "If Suzy's rock hits first and it requires only one rock to shatter the bottle then, as we have seen, Suzy has degree of responsibility 1 or 1/2 (depending on whether we consider only allowable settings) and Billy has degree of responsibility 0."

of the underlying model one uses (Figure 20a or Figure 20b). However, this case does not highlight the difference between RSV and existing techniques.

To understand the difference, we discuss the second interpretation next. Both Suzy and Billy throw the rocks at the same time, but Suzy's throw is faster than Billy's. Hence, the source variables (ST, BT) change from $(0, 0)$ to $(1, 1)$ without an intermediate value of $(1, 0)$. Under this view, both model 1 (Figure 20a) and model 2 (Figure 20b) are correct, but model 2 is more fine-grained. As discussed above, Suzy's degree of responsibility is $1/2$ under both models, but Billy's degree of responsibility drops from $1/2$ (model 1) to 0 (model 2). Despite the two models being equivalent (in terms of the relationship between the source variables and the outcome variable), the degree of responsibility changes. We posit that such lack of robustness to how the underlying mechanics are implemented is undesirable. As we established in our preliminary work (Proposition 1 in Singal et al. (2021)), RSV obeys *implementation invariance*, i.e., source variables receive the same attribution under equivalent models. For instance, RSV for both Suzy and Billy equals $1/2$ under both models [10]. Further, path-based approaches including path-specific effects and path-specific selection gradient attribute 1 to both Suzy and Billy under both models, resulting in an over-allocation irrespective of the model one uses. In particular, paths $ST \to SH \to BS$ and $BT \to BH \to BS$ have a path-specific effect of 1, whereas the only remaining path $ST \to SH \to BH \to BS$ has a path-specific effect of 0.

7.2 Causal unfairness

As alluded to in §1 (recall Figure 3), we next illustrate an application of RSV on non-linear mediation analysis to quantify *causal unfairness*, which has been a topic of great interest recently (Kilbertus et al., 2017; Kusner et al., 2017; Chiappa, 2019). The goal is to understand the influence exerted by sensitive attributes on the outcome, through both fair channels (effect mediated by resolving variables) and unfair channels. As we show next, RSV provides a crisp flow-based decomposition of the total effect even when multiple sensitive attributes (e.g., race and sex) change and the underlying dynamics are non-linear. Such a flow-based accounting naturally quantifies causal unfairness, with one possible context being an employee (or a class of employees) suing their employer for unjust treatment and the court trying to apportion legal responsibility to the employer.

Setup To illustrate causal unfairness, we use a graph structure motivated by real data, but assume intuitive relationships between the variables for ease of exposition. In particular, we focus on the DAG shown in Figure 21, which is motivated by the graph topologies discovered in recent works (Zhang et al., 2017; Wu et al., 2019), by employing the PC algorithm (Spirtes et al., 2000) on the `adult` dataset in the UCI repository (Dua and Graff, 2017). Given source variables *race* (sensitive attribute #1), *age*, and *sex* (sensitive attribute #2), we assume the following non-linear structural equations (with noise):

$$edu = \begin{cases} 4\log(age) + N(0, 3^2) & \text{if } race = 0 \\ 5\log(age) + N(0, 3^2) & \text{if } race = 1 \end{cases} \tag{20a}$$

10. In the setting with $k-1$ people throwing rocks in addition to Suzy, every person's RSV equals $1/k$, and hence, the sum equals the total effect.

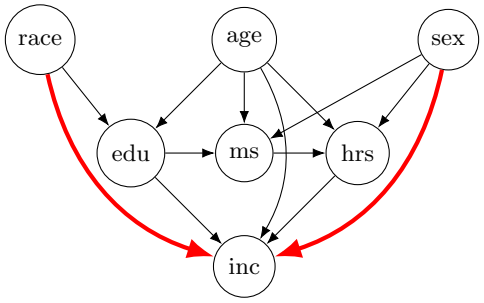


Figure 21: DAG used to illustrate causal unfairness. There are three source variables (*race*, *age*, and *sex*), two of which are sensitive (*race* and *sex*). There are four non-source variables: *edu* stands for years of education, *ms* for marital status (1 for married and 0 for not), *hrs* for average hours spent working per day, and *inc* for whether the person's annual income is above a threshold (e.g., \$50,000 in the `adult` dataset (Dua and Graff, 2017)). The two red thick edges show instances of unfair channels / unresolved discrimination (overt sexism and racism) and the corresponding unfairness parameters a_1 and a_3 are highlighted in red in the corresponding structural equation (20d). Note that although the mediated pathways could be said to reflect systemic racism (e.g., unequal access to education), they can also map to no unfairness if the difference is due to voluntary behavior of the underlying population. A well-known example is the Berkeley admissions case study (Bickel et al., 1975) where one category of *sex* (sensitive attribute) voluntarily chooses to apply to a more competitive *department* (mediating variable) and hence, is less likely to be *admitted* (outcome variable). We discussed such an example in our preliminary version (see §6 in Singal et al. (2021)). Irrespective of whether the difference is systemic or voluntary, we note that RSV is well-defined.

$$ms = \begin{cases} 1 & \text{w.p. } p(age, sex, edu) \\ 0 & \text{w.p. } 1 - p(age, sex, edu) \end{cases} \tag{20b}$$

$$hrs = 8 - ms - 0.05age + sex + N(0, 2^2) \tag{20c}$$

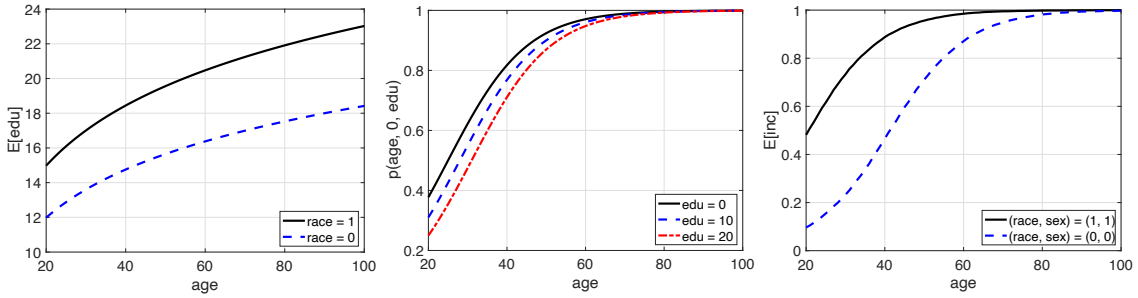
$$inc = \begin{cases} 1 & \text{w.p. } \frac{1}{1+\exp\{-0.1(age-50)-0.05edu-0.02hrs-a_1 race-a_3 sex\}} \\ 0 & \text{w.p. } 1 - \frac{1}{1+\exp\{-0.1(age-50)-0.05edu-0.02hrs-a_1 race-a_3 sex\}}. \end{cases} \tag{20d}$$

The *ms* probability $p(age, sex, edu)$ in (20b) is governed by the following logistic equations such that $sex = 1$ population is 90% less likely to be married relative to the $sex = 0$ population with all else being equal (i.e., same *age* and *edu*):

$$p(age, sex, edu) = \begin{cases} \frac{1}{1+\exp\{-0.1(age-25)+0.03edu\}} & \text{if } sex = 0 \\ \frac{0.9}{1+\exp\{-0.1(age-25)+0.03edu\}} & \text{if } sex = 1. \end{cases} \tag{20e}$$

To provide intuition, we visualize these relationships in Figure 22. In Figure 22a, we plot the expected value of *edu* (from (20a)) as we vary *age* from 20 to 100 for $race \in \{0, 1\}$. The monotonicity and concavity is encoded to reflect reality and the $race = 1$ population is expected to have a higher level of education. Figure 22b depicts the *ms* probability (from (20b) and (20e)) as a function of *age* for three levels of *edu* with $sex = 0$. Again, the monotonicity with respect to *age* captures reflects reality (more likely to be married as *age* increases) and we further encode an inverse relation with *edu*. We assume a linear function

for *hrs* (see (20c)) with a baseline value of 8, which decreases by 1 if $ms = 1$ (married), decreases in *age*, and increases by 1 if *sex* equals 1. Finally, in Figure 22c, we show expected *inc* (from (20d)) as a function of *age* for two settings of the sensitive attributes: background $(race^{(1)}, sex^{(1)}) = (0, 0)$ and foreground $(race^{(2)}, sex^{(2)}) = (1, 1)$. In this plot, we set the unfairness parameters a_1 and a_3 equal to 1 and the intermediate variables (*edu*, *ms*, and *hrs*) are set according to the corresponding relationships in (20). To compute the expected outcome for a given instance of the source variables (*race*, *age*, and *sex*), we generate 10,000 samples of the noise, which we found to be large enough to provide a stable estimate. As before, the monotonicity with *age* is desirable. Further, the functional form stated above in (20d) makes it clear that the expected outcome is increasing in *edu* and *hrs* as well, and the level of unfairness increases as we increase a_1 and a_3.



(a) Education **(b)** Marital status **(c)** Income

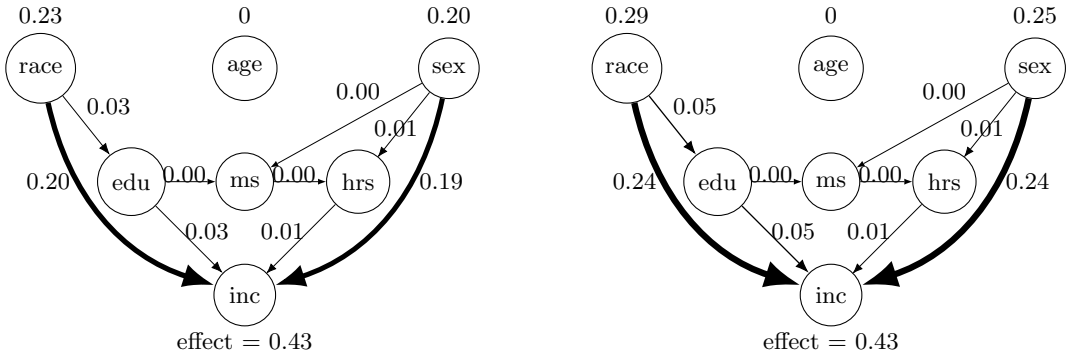
Figure 22: Visualizing the underlying structural equations presented in (20).

Our goal is to understand causal unfairness, which relates to the change in the expected outcome as a result of changes in sensitive attributes. For instance, in Figure 22c, for $age = 40$, the expected outcome roughly equals 0.46 and 0.89 for the two scenarios of sensitive attributes. Accordingly, the delta corresponding to an $(race, age, sex)$ background and foreground of $(0, 40, 0)$ and $(1, 40, 1)$ equals the difference $0.89 - 0.46 = 0.43$. However, not all of this 0.43 can be credited as unfair since some of this effect is propagated through mediators (*edu*, *ms*, and *hrs*), as seen in Figure 21. The question of interest is how to decompose the total effect of 0.43 along the various edges / pathways in order to quantify causal unfairness.

Results In Figure 23a, we show the RSV [11] corresponding to the SEM of (20) with a background $(race^{(1)}, age^{(1)}, sex^{(1)}) = (0, 40, 0)$ and a foreground $(race^{(2)}, age^{(2)}, sex^{(2)}) = (1, 40, 1)$ and unfairness parameters $(a_1, a_3) = (1, 1)$. Clearly, RSV obeys flow conservation and thus, provides an exact decomposition of the total effect of 0.43, which is in contrast to the over-allocation under path-specific effects (Figure 23b) [12]. The zero flow attributed

11. Given the relatively small size of the DAG, we computed the RSV using Algorithm 1 (brute-force). With 10,000 noise samples (used to evaluate expected outcome), it took around 10 minutes on a 3.8 GHz 8-core Intel i7 machine with 16 GB memory to compute RSV for all the edges.
12. We benchmark RSV with the path-specific effects approach, since it can be perceived as a generalization of direct and indirect effects and is very similar to path-specific selection gradient in this setup. Further, it

to the edges coming out of *age* is an instance of the flow nullity axiom, since those edges carry no new information (recall *age* has the same background and foreground).



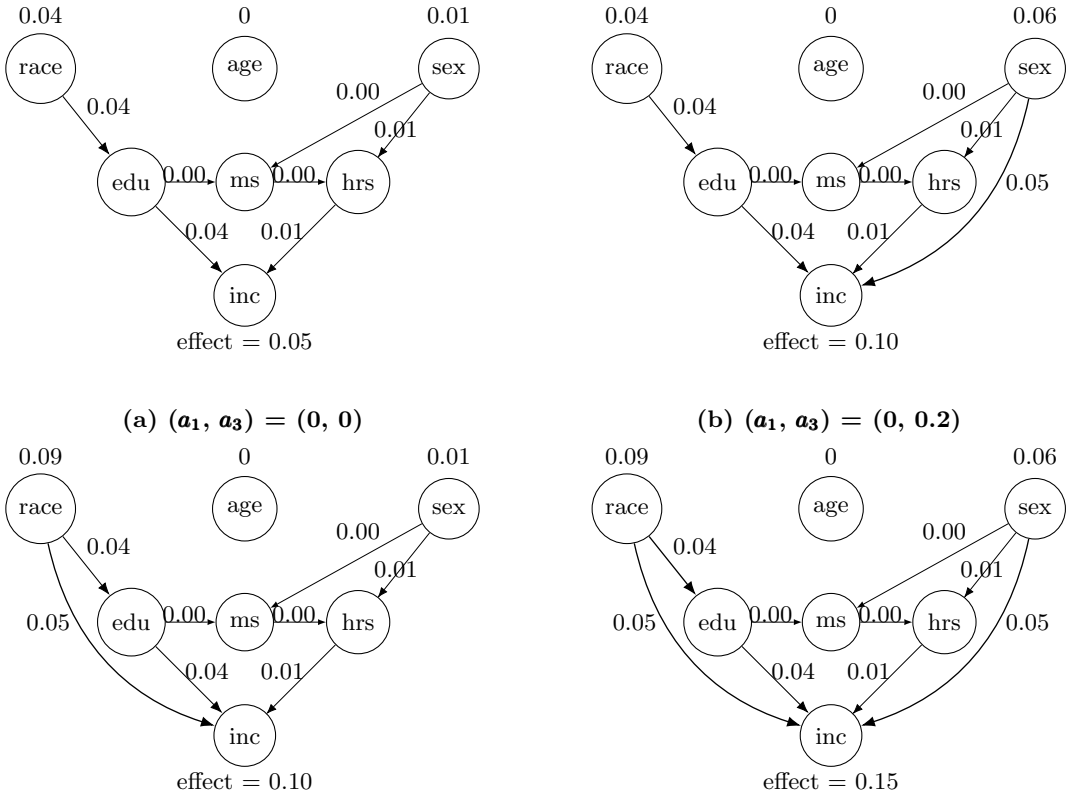
(a) RSV **(b)** Path-specific effects

Figure 23: RSV and path-specific effects for the SEM of (20) with a background $(race^{(1)}, age^{(1)}, sex^{(1)}) = (0, 40, 0)$ and a foreground $(race^{(2)}, age^{(2)}, sex^{(2)}) = (1, 40, 1)$ and unfairness parameters $\boldsymbol{(a_1, a_3) = (1, 1)}$. Expected *inc* under the background equals 0.89 and under the foreground equals 0.46 and hence, the total effect $inc^{(2)} - inc^{(1)}$ equals 0.43. Path-specific effects add up to 0.54, which is more than the total effect. The thickness of each edge is proportional to the flow.

Hence, with a relatively large value of $(a_1, a_3) = (1, 1)$, around 90% of the effect is propagated via the unfair edges and the remainder of around 10% is mediated through resolving variables. Of course, these numbers would vary as we change the unfairness parameters a_1 and a_3. In Figure 24, we show the RSV for $(a_1, a_3) \in \{(0, 0), (0, 0.2), (0.2, 0), (0.2, 0.2)\}$. When $(a_1, a_3) = (0, 0)$, it follows from (20d) that there is no causal unfairness since the sensitive attributes do not exert any direct influence on the outcome. In fact, the corresponding edges $(race, inc)$ and (sex, inc) obey the flow nullity axiom and as a result, they receive an attribution equal to 0 (see Figure 24a). The total effect equals 0.05 and all of it is propagated through intermediate variables (0.04 by *edu* and 0.01 by *hrs*). When we increase a_3 to 0.2 (see Figure 24b), edge (sex, inc) is no longer redundant and all of the 0.05 increment in the total effect flows through this unfair channel. A similar observation holds for the $(a_1, a_3) = (0.2, 0)$ setting (see Figure 24c), with edge $(race, inc)$ no longer being redundant and carrying the 0.05 increment in the total effect. Finally, at $(a_1, a_3) = (0.2, 0.2)$ (see Figure 24d), the total effect jumps to 0.15, with each of the unfair channels carrying 0.05 units and the remainder 0.05 units being propagated by the mediating variables, as in the $(a_1, a_3) = (0, 0)$ case.

As it should be clear, there is nothing special about keeping the *age* variable at a constant value of 40 in both background and foreground. Our framework is flexible to allow changes in *age* as well. In Figure 25, we show the RSV when the source variables change from a

is unclear how one would compute the degree of responsibility for this setting. Note that the path-specific effects approach outputs a value corresponding to each source-to-sink path, and these path-specific values uniquely map to edge-specific values via simple aggregation, i.e., for each edge, consider all the paths it appears in and add those path values. We did such an aggregation in our results to facilitate comparison with RSV, e.g., in Figure 23.



(a) $(a_1, a_3) = (0, 0)$ **(b) $(a_1, a_3) = (0, 0.2)$**

(c) $(a_1, a_3) = (0.2, 0)$ **(d) $(a_1, a_3) = (0.2, 0.2)$**

Figure 24: RSV for the SEM of (20) with a background $(race^{(1)}, age^{(1)}, sex^{(1)}) = (0, 40, 0)$ and a foreground $(race^{(2)}, age^{(2)}, sex^{(2)}) = (1, 40, 1)$ and unfairness parameters $\boldsymbol{(a_1, a_3)} \in \{\mathbf{(0, 0)}, \mathbf{(0, 0.2)}, \mathbf{(0.2, 0)}, \mathbf{(0.2, 0.2)}\}$.

background of $(0, 40, 0)$ to a foreground of $(1, 45, 1)$, i.e., *age* changes as well (from 40 to 45). The unfair parameters are set at $(a_1, a_3) = (0.2, 0.2)$, as in Figure 24d (to facilitate comparison). The expected outcome in the background equals 0.46 (as before) and in the foreground equals 0.73 and hence, the total effect equals the difference 0.27, an increase of 0.12 when compared to the total effect of 0.15 in Figure 24d. This increase of 0.12 in total effect is essentially attributed to the outgoing edges of *age*, as seen in Figure 25. (Note that the outgoing edges of *age* are attributed 0.13, which is not exactly 0.12. This is primarily because we rounded to two decimal places and there is small noise in our numbers since we used 10,000 noisy samples to evaluate the expected outcome.)

Connecting this application back to our motivation discussed in §1, RSV is naturally able to quantify effect propagation even when *multiple source variables change simultaneously* and the effect is propagated through *multiple mediators*, that too in a *non-linear* manner. Further, it does so while obeying a set of four desirable flow-based axioms, primarily because of its ability to consider a broader set of counterfactuals than existing approaches. On the contrary, approaches such as direct / indirect effects and path-specific effects only allow

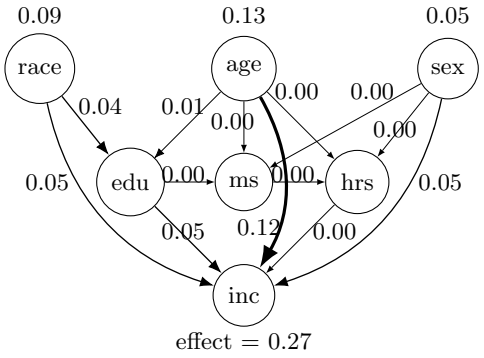


Figure 25: RSV for the SEM of (20) with a background $(race^{(1)}, age^{(1)}, sex^{(1)}) = (0, \mathbf{40}, 0)$ and a foreground $(race^{(2)}, age^{(2)}, sex^{(2)}) = (1, \mathbf{45}, 1)$ and unfairness parameters $(a_1, a_3) = (0.2, 0.2)$. As opposed to before, the edges coming out of *age* are no longer redundant.

for the possibility of one counterfactual (all other edges being inactive), resulting in them violating arguably the most primitive axiom (flow conservation).

RSV can in fact handle more involved scenarios. As an illustration, we can have an unfair path from *race* to *inc* that goes through one of the intermediate variables (*edu*). As shown in Figure 26, we insert a new node denoting geographical location of the individual ($loc \in \{0, 1\}$), which depends on *race* and influences *edu* and *inc*. We modify (20) as follows:

$$loc = \begin{cases} 1 \text{ w.p. } p(race) \\ 0 \text{ w.p. } 1 - p(race) \end{cases}$$

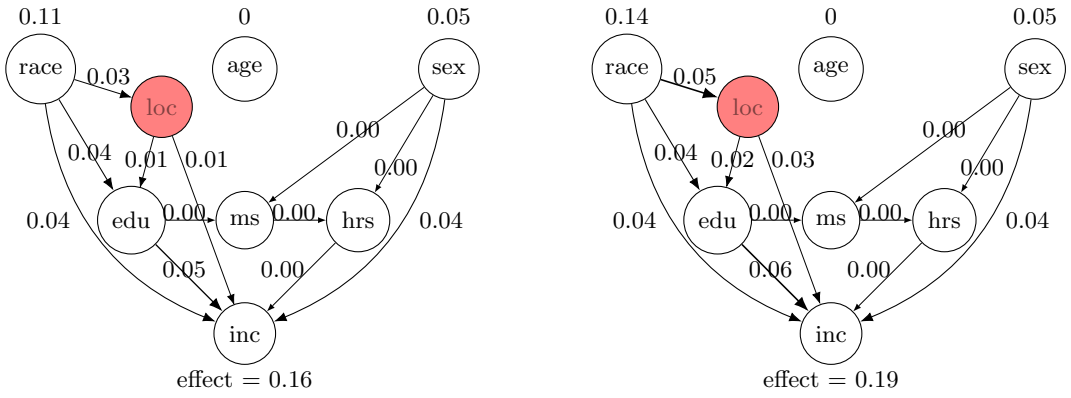
$$edu = \left(4 + \mathbb{I}\{race = 1\} + \underbrace{\mathbb{I}\{loc = 1\}}_{\text{new term}} \right) \log(age) + N(0, 3^2)$$

$$inc = \begin{cases} 1 \text{ w.p. } \frac{\overbrace{0.9 + 0.1\mathbb{I}\{loc = 1\}}^{\text{modification}}}{1+\exp\{-0.1(age-50)-0.05edu-0.02hrs-a_1X_1-a_3X_3\}} \\ 0 \text{ w.p. } 1 - \frac{0.9+0.1\mathbb{I}\{loc=1\}}{1+\exp\{-0.1(age-50)-0.05edu-0.02hrs-a_1X_1-a_3X_3\}}, \end{cases}$$

where $p(race)$ is parameterized by $p \in [0, 1]$ such that $p(1) = p$ and $p(0) = p/10$, i.e., *race* 1 is 10 times more likely to live in *loc* 1 (than *race* 0). The equation for *edu* is a generalization of that in (20) and simply boosts the *edu* level, if *loc* equals 1. As such, *race* 1 is more likely to live in the location with better access to education and the $loc \to edu$ edge can be seen as the unfair channel that goes through an intermediate variable (*edu*). The unfairness here is due to a lower access to education in location 0. Further, we modify the equation for *inc* by changing the numerator from 1 to $0.9 + 0.1\mathbb{I}\{loc = 1\}$, so that location 1 results in a direct boost to the income as well. This might not be unfair, e.g., location 1 might have a higher cost of living and hence, the incomes are higher as well. All other components of the SCM remain the same as before[13].

13. Note that in Figure 26, we do not show the edges coming out of *age* since the background and foreground values of *age* equal each other and hence, the corresponding edges will have an RSV of 0. As such, to

In addition to the unfairness discussed already (i.e., on edges $race \to inc$ and $sex \to inc$), it is also of interest to understand the unfairness mediated through loc (i.e., on the edge $loc \to edu$). We provide such an understanding in Figure 26. RSV attributes 1/16th of the total effect to such unfairness (0.01 out of 0.16) for $p = 0.3$ and 2/19th for $p = 0.6$. As we increase p from 0 to 1, the total effect increases from 0.13 to 0.22. RSV values for all edges except the ones corresponding to loc do not change much. This aligns with intuition (as increasing p results in more effect being propagated through loc). Accordingly, in Figure 27, we show how RSV evolves as a function of p for the two edges coming out of loc. For $p = 0$, both the links are attributed a value of 0, which makes sense (since they do not propagate any effect when $p = 0$). The attributions increase with p, with a clear delineation of how much of the effect flows through the unfair channel of interest ($loc \to edu$).



(a) $p = 0.3$ **(b)** $p = 0.6$

Figure 26: RSV for the modified SCM of with a background $(race^{(1)}, age^{(1)}, sex^{(1)}) = (0, 40, 0)$ and a foreground $(race^{(2)}, age^{(2)}, sex^{(2)}) = (1, 40, 1)$ with unfairness parameters $(a_1, a_3) = (0.2, 0.2)$ and location parameter $p \in \{0.3, 0.6\}$. (Note that the reported numbers do not necessarily obey flow conservation as they are rounded to 2 decimal places.)

Remark 11 (Individual-level counterfactual) *Note that in our illustration of causal unfairness, we only fix the data corresponding to source variables* $(race, age, sex)$ *and do not condition on any realization of the non-source variables but let them by determined by the structural equations* (20)*. We did this for ease of illustration and we note that it is straightforward to account for* individual-level *data observed at non-source nodes. Motivated by Kusner et al. (2017), this can be done by following the standard three-step framework of Pearl et al. (2016) (Chapter 4): (a) abduction, (b) action, and (c) prediction. In particular, given individual-level realization* $(\mathbf{X}, Y) = (\boldsymbol{x}, y)$*, we can first compute the posterior of the noise (abduction)* $\mathbf{U} \mid (\boldsymbol{x}, y)$*. Second, as we did above, we can consider the two cases of interest (action): setting the source variables at background and foreground. Finally, we can compute the RSV-based effect propagation (prediction) as we did above but by using the posterior* $\mathbf{U} \mid (\boldsymbol{x}, y)$ *(instead of the prior). Furthermore, this can be done*

prevent clutter, we omit showing those edges but note that they exist and the corresponding structural equations are the same as in (20).

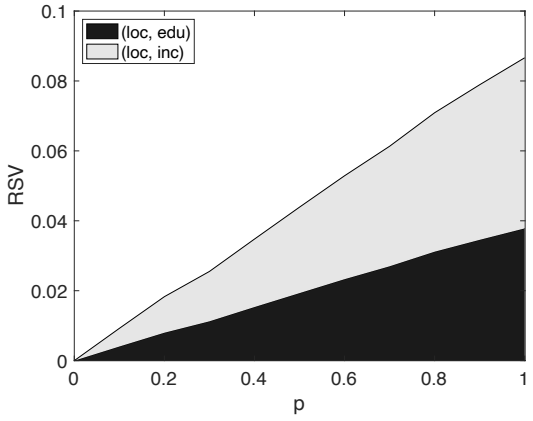


Figure 27: RSV on the two outgoing edges of *loc* (i.e., $loc \to edu$ and $loc \to inc$) as we vary $p \in \{0, 0.1, \ldots, 1\}$. The rest of the setup remains the same as in Figure 26.

for realizations corresponding to any subset of $(\mathbf{X}, Y)$*. For example, one might only observe* $(race, age, sex, edu, ms, inc)$ *but not the* hrs *of an individual, which will result in the corresponding noise posterior.*

8. Concluding remarks

Before concluding, we elaborate on a number of interesting extensions for the posited framework. Instead of having the quantity of attribution (i.e., the δ) as the difference in the *expected* outcomes (i.e., $\mathbb{E}_{\boldsymbol{u}}[y^{(2)}_{\boldsymbol{u}}] - \mathbb{E}_{\boldsymbol{u}}[y^{(1)}_{\boldsymbol{u}}]$), we can define it using other functionals of the noise distribution, e.g., the quantile or the variance. Our key results (axiomatic support in Theorem 4, path-based characterization for non-parametric SEMs in Theorem 8, and the Monte-Carlo estimation scheme in §6 along with the properties in Proposition 10) hold for such a generalization. To see this, observe that given a subset $E \subseteq \mathsf{E}$ of active edges, the definition of $y_{\boldsymbol{u}}(E)$ remains the same as discussed around (3) in §4. The key definition that changes is that of $y(E)$. Instead of defining $y(E)$ as $\mathbb{E}_{\boldsymbol{u}}[y_{\boldsymbol{u}}(E)]$ (as we did below (3)), we now define it by using an arbitrary operator of interest, denoted by $\mathbb{F}_{\boldsymbol{u}}[\cdot]$ (e.g., quantile or variance), i.e.,

$$y(E) := \mathbb{F}_{\boldsymbol{u}}[y_{\boldsymbol{u}}(E)].$$

Accordingly, given background and foreground source values $\boldsymbol{x}^{(1)}_{\mathsf{N}_0}$ and $\boldsymbol{x}^{(2)}_{\mathsf{N}_0}$, the background and foreground outcome equals $y^{(1)} := \mathbb{F}_{\boldsymbol{u}}[y^{(1)}_{\boldsymbol{u}}]$ and $y^{(2)} := \mathbb{F}_{\boldsymbol{u}}[y^{(2)}_{\boldsymbol{u}}]$, respectively, where the super-script captures the dependence on source values as before. Hence, the quantity of attribution equals the difference $y^{(2)} - y^{(1)}$ as before. For this modification, the intuition provided in §4.1 holds as it is and the definition of RSV remains as in §4.2. Further, the aforementioned results (Theorem 4, Theorem 8, and Proposition 10) go through without any additional modification (proofs are identical). This highlights the flexibility of the proposed framework to accommodate various metrics of interest for attribution purposes. For example, given the widespread usage of quantile regression (Koenker and Bassett Jr, 1978; Koenker and Hallock, 2001; Meinshausen, 2006), one might be interested in understand-

ing the change in a certain p-quantile of the outcome, for some $p \in (0, 1)$. In fact, if the underlying SEM is linear (as in (6)), then our closed-form characterization in Theorem 6 holds for quantiles as well. The reason for this is that due to the additive nature of the noise in a linear SEM, the p-quantile of the outcome can be represented as the outcome's expected value plus some constant (with respect to which edges are active). This constant gets cancelled out under the RSV computations and hence, the solution ends up being the same as the one for the expectation operator.

In summary, we formulate a generic problem to study effect propagation in causal DAGs, and use it to provide a comprehensive view on existing approaches such as direct and indirect effects, path-specific effects, and degree of responsibility. In addition to highlighting limitations of such techniques, we propose an *axiomatic flow-based methodology* (RSV) to quantify attribution. RSV operates on a top-down principle and flows down the effect from the source nodes via a sequence of recursive games, which allow RSV to consider a broader spectrum of counterfactuals than existing approaches. For linear SEMs, RSV admits a tractable closed-form characterization, which recovers the classical method of path coefficients and is equivalent to ideas such as path-specific effects. For non-parametric SEMs, we provide a path-based characterization of RSV, which leads to an unbiased Monte-Carlo estimation scheme with an exponentially decaying sample complexity. Our principled approach to effect propagation provides a new perspective on challenging problems on causality, such as causal overdetermination and causal unfairness.

Next, we outline several directions worth further exploring. First, although our path-based characterization leads to an *estimation* scheme for non-parametric SEMs, one can potentially develop tractable *exact* characterizations of RSV, especially beyond the class of linear SEMs. We found it challenging to do so even for the case of a linear model with two-way interactions, as mentioned at the end of §5 (recall Footnote 7). Even though if such characterizations are intractable, they might lead to efficient estimation schemes. Second, our framework assumes knowledge of the underlying structural causal model and it would be useful to understand RSV's *robustness* to model estimation and presence of hidden confounders. For example, in the case of a linear SEM, one can possibly use the characterization in Theorem 6 to understand the sensitivity of RSV to the model parameters $[a_{ij}]_{(i,j)\in\mathsf{E}}$ (i.e., the edge weights). Understanding such robustness is critical before employing RSV in real-world applications, especially for sensitive applications such as causal unfairness. Third, given the applicability of causal attribution in practical domains such as advertising (Dalessandro et al., 2012; Singal et al., 2022) and legal studies (Ferey and Dehez, 2016), applying RSV to such real-world applications is of interest. Finally, this work proposed one set of desirable axioms for effect propagation and it is of interest to explore alternative axioms.

Acknowledgments

We thank the Action Editor and two anonymous referees for the careful reading of the paper and many constructive comments and suggestions. The work of GM was supported in part by NSF grant DMS 2348640.

Appendix A. Primer on Shapley value (SV)

In this appendix, we provide a brief primer on Shapley value (SV) and refer the reader to Chapter 8 of Peleg and Sudhölter (2007) for details. In cooperative game theory, SV (Shapley, 1953) is a well-accepted solution concept to fairly attribute the total value in a *game*. In particular, given a finite set $\mathcal{P}$ of *players*, let the value generated by any subset of players (*coalition*) $P \subseteq \mathcal{P}$ be denoted by the *characteristic function* $v(P)$. Accordingly, the total value in the game equals $v(\mathcal{P})$ and the goal is to distribute this value back to the individual players in $\mathcal{P}$. To do so, SV attributes the following to player $r \in \mathcal{P}$:

$$\pi_r := \sum_{P \subseteq \mathcal{P} \setminus \{r\}} w_{\mathcal{P}}(P) \times \{v(P \cup \{r\}) - v(P)\},$$

where the weight function is defined as

$$w_{\mathcal{P}}(P) := \frac{|P|!\,(|\mathcal{P}|-|P|-1)!}{|\mathcal{P}|!}.$$

Intuitively speaking, SV of player r computes its value-add to each coalition P that does not contain player r and weights it by a corresponding "probability" [14] $w_{\mathcal{P}}(P)$. Hence, it can be seen as the "expected" value-add of player r. SV is desirable since it is the *unique* solution to obey the following four axioms:

1. **Efficiency**: Total value is distributed, i.e.,

$$\sum_{r \in \mathcal{P}} \pi_r = v(\mathcal{P}) - v(\emptyset).$$

2. **Symmetry**: Equivalent players receive same attribution, i.e., $\pi_r = \pi_{r'}$ if players $r, r' \in \mathcal{P}$ obey the following:

$$v(P \cup \{r\}) = v(P \cup \{r'\})\ \forall P \subseteq \mathcal{P} \setminus \{r, r'\}.$$

3. **Linearity**: Given two characteristic functions $v_1(\cdot)$ and $v_2(\cdot)$, linearity requires robustness to mixing and scaling, i.e., for all $r \in \mathcal{P}$,

$$\pi_r(v_1 + v_2) = \pi_r(v_1) + \pi_r(v_2)$$
$$\pi_r(\alpha v_1) = \alpha \pi_r(v_1)\ \forall \alpha \in \mathbb{R}.$$

4. **Null player**: Useless player gets zero attribution, i.e., $\pi_r = 0$ if player $r \in \mathcal{P}$ obeys

$$v(P \cup \{r\}) = v(P)\ \forall P \subseteq \mathcal{P} \setminus \{r\}.$$

14. Note that $\sum_{P \subseteq \mathcal{P} \setminus \{r\}} w_{\mathcal{P}}(P)$ equals 1 and each of the weight term is non-negative.

Appendix B. Reverse effects

As in our direct and indirect effects discussion (§3.1), consider the setup of Figure 2. To understand a limitation of *reverse indirect effect* (Pearl, 2010), we first visualize direct effect (DE) and reverse indirect effect ($-IE_r$) in Figure 28. Simply put, DE starts off with the baseline of both direct and indirect channels being inactive (subplot 2) and computes the value-add of making the direct channel active (subplot 1). $-IE_r$ starts off with the baseline of the direct channel being active (subplot 4) and computes the value-add of making the indirect channel active (subplot 3). When we add up DE and $-IE_r$, subplots 1 and 4 in Figure 28 cancel out and we are left with the total effect TE (difference between subplots 3 and 2). Clearly, $DE = 0$ and $-IE_r = 1$, suggesting all of the effect is mediated by the indirect channel. However, we can similarly define *reverse direct effect* ($-DE_r$) by flipping the order in which we proceed. In particular, as we show in Figure 29, we can first compute indirect effect (IE) by starting off with the baseline of both direct and indirect channels being inactive (subplot 2) and computing the value-add of making the indirect channel active (subplot 1). Then, we can start off with the baseline of the indirect channel being active (subplot 4) and compute the value-add of making the direct channel active (subplot 3), which we call $-DE_r$ (analogous to $-IE_r$). Similar to $TE = DE - IE_r$, we have $TE = IE - DE_r$. Doing so results in $IE = 0$ and $-DE_r = 1$, meaning all of the effect is propagated by the direct channel, which is the opposite of what we concluded above. As such, the order in which we proceed leads to a different conclusion. RSV avoids this issue by averaging over these two orderings:

$$TE = \underbrace{(DE - DE_r)/2}_{\text{RSV of direct channel}} + \underbrace{(IE - IE_r)/2}_{\text{RSV of indirect channel}} . \tag{21}$$

Note that RSV simplifies to (21) only for the *very specific* graph structure under consideration. To see this, observe that (21) is *path*-based in that it decouples the total effect along two paths (direct and indirect), which is different from the *edge*-based RSV. Futhermore, (21) provides a very coarse decomposition of the total effect as it only decouples it along *two* paths, which is very different from RSV's ability to provide a granular flow-based decomposition along *every* edge. Given the clean decomposition in (21), it is of interest to understand if RSV can be expressed in a similar manner for more general graphs. In fact, this was partly our motivation behind the path-based characterization of RSV in Theorem 8, but RSV's intricate nature leads to a much more involved expression in general. Irrespective, RSV can be seen as generalizing (21) to more involved settings as it recovers (21) in simple settings but has the ability to provide a more granular decomposition of the total effect.

This issue is exacerbated in bigger graphs as there are more possible choices of which order to proceed in. For example, in the graph shown in Figure 30, there are 6 possible orderings. Depending on the order, we end up attributing different values, and there is no unique right order. As such, though the notion of reverse effects can restore efficiency (decomposition adding up to the total effect), it feels rather ad hoc and lacks uniqueness. This is in sharp contrast to RSV, which comes out uniquely from four desirable axioms.

(a) Direct effect (DE) **(b)** Reverse indirect effect ($-IE_r$)

Figure 28: Visualizing DE and $-IE_r$. DE corresponds to the difference between the two graphs in subplot (a). $-IE_r$ corresponds to the difference between the two graphs in subplot (b).

(a) Indirect effect (IE) **(b)** Reverse direct effect ($-DE_r$)

Figure 29: Visualizing IE and $-DE_r$. IE corresponds to the difference between the two graphs in subplot (a). $-DE_r$ corresponds to the difference between the two graphs in subplot (b).

Appendix C. Proof of Theorem 4

Theorem 4 *Given structural causal model* $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$*, the RSV* $[\pi^{RSV}_{jk}]_{(j,k)\in\mathsf{E}}$ *defined via Algorithm 1 is the unique solution to the flow-based axioms.*

Proof First, consider the node 0 game. E_0 is the set of players with the characteristic function $v_0(E_0) = y(E_0, \mathsf{E}_1, \ldots, \mathsf{E}_n)$ for coalition $E_0 \subseteq \mathsf{E}_0$. Under RSV, flow received by each edge in E_0 equals the corresponding SV of this game:

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

The uniqueness result of SV w.r.t. the original SV axioms (Shapley, 1953) implies that $[\pi^{\text{RSV}}_{0k}]_{(0,k)\in\mathsf{E}_0}$ uniquely satisfies the four flow-based axioms (at node 0). Observe that the symmetry, nullity, and linearity axioms of the classical SV are equivalent to the flow symmetry, flow nullity, and flow linearity axioms. In addition, flow conservation at node 0 is implied by the efficiency axiom of the classical SV. To see this, observe that the classical efficiency axiom states $\sum_{k\in\mathsf{C}_0} \pi^{\text{RSV}}_{0k} = v_0(\mathsf{E}_0) - v_0(\emptyset)$. This is equivalent to the conservation of flow at node 0:

$$v_0(\mathsf{E}_0) - v_0(\emptyset) = y(\mathsf{E}_0, \mathsf{E}_1, \ldots, \mathsf{E}_n) - y(\emptyset, \mathsf{E}_1, \ldots, \mathsf{E}_n)$$

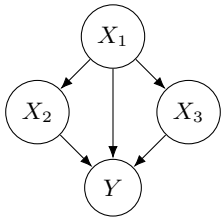


Figure 30: A graph with one direct and two indirect channels. There are 6 possible orderings, which can be understood as follows. We start with the baseline of all 3 channels being inactive and hence, we have the option of adding any of the 3 channels first. Then, we activate any of the remaining 2 channels. Finally, we activate the remaining 1 channel. Note that $3 \times 2 \times 1 = 6$.

$$= y^{(2)} - y^{(1)}.$$

The final equality is due to the definition of $y(\cdot)$ (recall (3)):

$$y(\mathsf{E}_0, \mathsf{E}_1, \ldots, \mathsf{E}_n) = y^{(2)}$$
$$y(\emptyset, \mathsf{E}_1, \ldots, \mathsf{E}_n) = y^{(1)}.$$

Next, consider the node $j \in \mathsf{N} \setminus \{0\}$ game. The players are E_j with the characteristic function $v_j(E_j) = \sum_{i \in \mathsf{P}_j} \pi_{ij}(\mathsf{E}_0, \ldots, E_j, \ldots, \mathsf{E}_n)$ for coalition $E_j \subseteq \mathsf{E}_j$. Under RSV, flow received by each edge in E_j equals the corresponding SV of this game:

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

The uniqueness result of SV w.r.t. the original SV axioms (Shapley, 1953) implies that $[\pi_{jk}^{\text{RSV}}]_{(j,k) \in \mathsf{E}_j}$ uniquely satisfies the four flow-based axioms (at node j). Observe that the symmetry, nullity, and linearity axioms of the classical SV are equivalent to the flow symmetry, flow nullity, and flow linearity axioms. In addition, flow conservation at node j is implied by the efficiency axiom of the classical SV. To see this, observe that the classical efficiency axiom states $\sum_{k \in \mathsf{C}_j} \pi_{jk}^{\text{RSV}} = v_j(\mathsf{E}_j) - v_j(\emptyset)$. This is equivalent to the conservation of flow at node j:

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

The last equality holds since

$$\pi_{ij}(\mathsf{E}_0, \ldots, \mathsf{E}_j, \ldots, \mathsf{E}_n) = \pi_{ij}^{\text{RSV}} \ \forall i \in \mathsf{P}_j$$
$$\pi_{ij}(\mathsf{E}_0, \ldots, \emptyset, \ldots, \mathsf{E}_n) = 0 \ \forall i \in \mathsf{P}_j.$$

The first statement is by definition and the second statement is true because for $E_j = \emptyset$, no new information is passed via node j to its children (see the definition of $y(\cdot)$ as in (3)) and therefore, edge (i,j) becomes a null player in the node $i \in \mathsf{P}_j$ upstream game. This completes the proof. □

Appendix D. Proof of Theorem 6

Theorem 6 *Consider structural causal model* $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$ *, wherein the structural equations* F *are linear and the noise distribution* D *has zero mean (see* (6)*), with* $\boldsymbol{c}$ *and* $\boldsymbol{b}$ *as in* (7) *and* (8)*. Then,*

$$\begin{aligned} \pi_{0j}^{RSV} &= c_{0j} && \forall (0,j) \in \mathsf{E}_0 \\ \pi_{jk}^{RSV} &= \sum_{i \in \mathsf{P}_j} b_{ij} c_{jk} && \forall (j,k) \in \mathsf{E} \setminus \mathsf{E}_0. \end{aligned}$$

Proof Seeting $\mathcal{E} = \mathsf{E}$ in Lemma 12 (stated below) gives the following flow on each edge $(0,j) \in \mathsf{E}_0$:

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

The second equality holds due to $\boldsymbol{b} = \boldsymbol{b}(\mathsf{E})$ (see (8) and (23)) and $\boldsymbol{c} = \boldsymbol{c}(\mathsf{E})$ (see (7) and (22)), the third equality holds as $b_{0j} = x_j^{(2)} - x_j^{(1)} = a_{0j}\ \forall (0,j) \in \mathsf{E}_0$, and the fourth equality directly follows the definition of $c_{0j}\ \forall (0,j) \in \mathsf{E}_0$ (see (7)). Similarly, setting $\mathcal{E} = \mathsf{E}$ in Lemma 12 gives the following attributions to each edge $(j,k) \in \mathsf{E} \setminus \mathsf{E}_0$:

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

The third and fifth equalities directly follow the definitions of $\boldsymbol{b}$ and $\boldsymbol{c}$ (see (8) and (7)). Recalling that π_{jk}^{RSV} is defined as $\pi_{jk}(\mathsf{E})$ for all $(j,k) \in \mathsf{E}$, the proof is complete. □

Details on Lemma 12

The proof of Theorem 6 uses a more general result (Lemma 12), which we develop next. It will help to generalize the definitions of both the forward-looking weights $\boldsymbol{c}$ and the

backward-looking weights $\boldsymbol{b}$. In §5.1 (recall Equations (7) and (8)), when defining these weights, we implicitly assumed all edges E as being active. We now let these weights vary as a function of an arbitrary subset $\mathcal{E} = (\mathcal{E}_0, \ldots, \mathcal{E}_n) \subseteq \mathsf{E}$ of edges. We assume topologically sorting wlog, i.e., there does not exist an edge $(i,j) \in \mathsf{E}$ such that $i > j$ (Cormen et al., 2009). We define the generalized forward-looking weights $\boldsymbol{c}(\mathcal{E})$ as follows:

$$c_{j,n+1}(\mathcal{E}_j) := a_{j,n+1}\mathbb{I}\{(j,n+1) \in \mathcal{E}_j\} \qquad \forall (j,n+1) \in \mathsf{E}^{\text{in}}_{n+1} \tag{22a}$$

$$c_{ij}(\mathcal{E}_i, \ldots, \mathcal{E}_n) := a_{ij}\mathbb{I}\{(i,j) \in \mathcal{E}_i\} \sum_{k:(j,k)\in\mathcal{E}_j} c_{jk}(\mathcal{E}_j, \ldots, \mathcal{E}_n) \qquad \forall (i,j) \in \mathsf{E} \setminus \mathsf{E}^{\text{in}}_{n+1}, \tag{22b}$$

where $a_{0j} := x_j^{(2)} - x_j^{(1)}$ $\forall (0,j) \in \mathsf{E}_0$ as before. Similarly, we define the generalized backward-looking weights $\boldsymbol{b}(\mathcal{E})$ as follows:

$$b_{0j}(\mathcal{E}_0) := (x_j^{(2)} - x_j^{(1)})\mathbb{I}\{(0,j) \in \mathcal{E}_0\} \qquad \forall (0,j) \in \mathsf{E}_0 \tag{23a}$$

$$b_{jk}(\mathcal{E}_0, \ldots, \mathcal{E}_j) := \sum_{i:(i,j)\in\mathcal{E}_i} b_{ij}(\mathcal{E}_0, \ldots, \mathcal{E}_i) a_{jk}\mathbb{I}\{(j,k) \in \mathcal{E}_j\} \qquad \forall (j,k) \in \mathsf{E} \setminus \mathsf{E}_0. \tag{23b}$$

Observe that setting in $\mathcal{E} = \mathsf{E}$ recovers the original weights as in (7) and (8), i.e., $\boldsymbol{b} = \boldsymbol{b}(\mathsf{E})$ and $\boldsymbol{c} = \boldsymbol{c}(\mathsf{E})$. We next present Lemma 12.

Lemma 12 *Consider structural causal model $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$ where the structural equations F are linear and the noise D is mean-zero (see (6)), with $\boldsymbol{c}(\cdot)$ and $\boldsymbol{b}(\cdot)$ as in (22) and (23). Then, given $\mathcal{E} \subseteq \mathsf{E}$, for all $i \in \mathsf{N}$,*

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

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

Proof First, consider the node $i = 0$ (super-source). Recall that $\pi_{0j}(\mathcal{E})$ is the SV of the following game. The players are the elements of $\mathcal{E}_0$ and given an arbitrary coalition $E_0 \subseteq \mathcal{E}_0$, characteristic function is

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

The second equality directly follows the definition of $\boldsymbol{c}(\cdot)$ (see (22)) and $y(\cdot)$ (see (3)). The third equality follows the definition of $\boldsymbol{c}(\cdot)$ (see (22)). The last equality holds since $a_{0j} = x_j^{(2)} - x_j^{(1)}$ and it also leverages the definition of $b_{0j}(\cdot)$ (see (23)). Given $v_0(E_0 \mid \mathcal{E}_{-0})$

is separabile over the players $(0,j)$ in $\mathcal{E}_0$, we have that the SV of each player $(0,j) \in \mathcal{E}_0$ equals

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

The base case is complete.

Due to the DAG structure, it is sufficient to show the claim holds at node $j \in \mathsf{N}$ by assuming it to hold at each of its parent nodes i such that $(i,j) \in \mathcal{E}_i$. In other words, it suffices to show that

$$\pi_{ij}(\mathcal{E}) = \sum_{k:(j,k)\in\mathcal{E}_j} b_{ij}(\mathcal{E}_0,\ldots,\mathcal{E}_i) c_{jk}(\mathcal{E}_j,\ldots,\mathcal{E}_n) \;\forall j : (i,j) \in \mathcal{E} \tag{24}$$

implies

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

Recall that $\pi_{jk}(\mathcal{E})$ corresponds to the SV of the game with $\mathcal{E}_j$ as the set of players and the following characteristic function given coalition $E_j \subseteq \mathcal{E}_j$:

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

The second equality follows (24), the third is true as $c_{jk}(E_j,\ldots,\mathcal{E}_n) = c_{jk}(\mathcal{E}_j,\ldots,\mathcal{E}_n)$ for $(j,k) \in E_j$ (see (22)). Given the characteristic function is separabile over the underlying players, we have that the SV of player $(j,k) \in \mathcal{E}_j$ 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 from the definitions of $\boldsymbol{c}(\cdot)$ (see (22)) and $\boldsymbol{b}(\cdot)$ (see (23)), respectively. This completes the proof. □

Appendix E. Proof of Theorem 8

Theorem 8 *Given structural causal model* $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$*, the RSV of edge* $(i,j) \in \mathsf{E}$ *exhibits the following characterization:*

$$\pi_{ij}^{RSV}(\mathsf{E}) = \sum_{W \in \mathsf{W}_i} \sum_{E_W \subseteq \mathsf{E}_W^{W \cup \{j\} \setminus \{0\}}} \kappa_W(E_W) \sum_{V \subseteq \texttt{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathsf{E}_{-W}).$$

Proof Directly follows when we plug in $\mathcal{E}$ equal to E in Lemma 13 (see below). □

Details on Lemma 13

The proof of Theorem 8 leverages a more general result (Lemma 13), which we present now. We generalize the definitions of various primitives as a function of a subset $\mathcal{E} = (\mathcal{E}_0, \ldots, \mathcal{E}_n) \subseteq \mathsf{E}$. For instance, the set of parents of node $i \in \mathsf{N}^+$ is denoted as $\mathsf{P}_i(\mathcal{E})$, the set of all unique paths to node $i \in \mathsf{N}^+$ is denoted as $\mathsf{W}_i(\mathcal{E})$, flow (RSV) on edge $(i,j) \in \mathcal{E}$ is denoted as $\pi_{ij}^{\text{RSV}}(\mathcal{E})$ (or simply $\pi_{ij}(\mathcal{E})$ for conciseness), the SV weight as $w_{\mathcal{E}_i}(\cdot)$ for the game at node $i \in \mathsf{N}$, the kappa factor (given path W) as $\kappa_W^{\mathcal{E}}(\cdot)$, etc. We assume wlog that the DAG is topologically sorted, i.e., there is no edge $(i,j) \in \mathsf{E}$ with $i > j$ (Cormen et al., 2009). Note that plugging in $\mathcal{E}$ as E recovers the original primitives. We are now in a position to present Lemma 13.

Lemma 13 *Given structural causal model* $\mathsf{M} = (\mathsf{G}, \mathsf{F}, \mathsf{D})$ *and a subset* $\mathcal{E} = (\mathcal{E}_0, \ldots, \mathcal{E}_n) \subseteq \mathsf{E}$ *of edges, the RSV (corresponding to input* $\mathcal{E}$*) of edge* $(i,j) \in \mathcal{E}$ *exhibits the following characterization:*

$$\pi_{ij}^{RSV}(\mathcal{E}) = \sum_{W \in \mathsf{W}_i(\mathcal{E})} \sum_{E_W \subseteq \mathcal{E}_W^{W \cup \{j\} \setminus \{0\}}} \kappa_W^{\mathcal{E}}(E_W) \sum_{V \subseteq \texttt{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathcal{E}_{-W}).$$

(Note that $\pi_{ij}^{RSV}(\mathcal{E})$ *corresponds to the following call of Algorithm 1:* $\texttt{RSV}(\mathsf{N}, \mathcal{E})$*. In particular, the second input is* $\mathcal{E}$ *instead of* E*.)*

Proof 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

$$\begin{aligned} v_0(E_0 \mid \mathcal{E}_{-0}) &= y(E_0, \mathcal{E}_1, \ldots, \mathcal{E}_n) \\ &= y(E_0, \mathcal{E}_{-0}). \end{aligned}$$

Accordingly, the Shapley value for player $(0,j) \in \mathcal{E}_0$ equals

$$\pi_{0j}(\mathcal{E}) = \sum_{E_0 \subseteq \mathcal{E}_0 \setminus \{(0,j)\}} w_{\mathcal{E}_0}(E_0) \times \left\{ y(E_0 \cup \{(0,j)\}, \mathcal{E}_{-0}) - y(E_0, \mathcal{E}_{-0}) \right\}. \tag{25}$$

Observing that $\mathsf{W}_0(\mathcal{E}) = (0)$, the RHS of the claim for $(i,j) = (0,j)$ equals:

$$= \sum_{W \in \mathsf{W}_0(\mathcal{E})} \sum_{E_W \subseteq \mathcal{E}_W^{W \cup \{j\} \setminus \{0\}}} \kappa_W^{\mathcal{E}}(E_W) \sum_{V \subseteq \texttt{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathcal{E}_{-W})$$

$$
\begin{aligned}
&= \sum_{E_0 \subseteq \mathcal{E}_0^j} \kappa_0^{\mathcal{E}}(E_0) \sum_{V \subseteq \{(0,j)\}} (-1)^{1-|V|} \times y(E_0 \cup V, \mathcal{E}_{-0}) \\
&= \sum_{E_0 \subseteq \mathcal{E}_0 \setminus \{(0,j)\}} w_{\mathcal{E}_0}(E_0) \times \{y(E_0 \cup \{(0,j)\}, \mathcal{E}_{-0}) - y(E_0, \mathcal{E}_{-0})\} \\
&= \pi_{0j}(\mathcal{E}).
\end{aligned}
$$

The first equation is the RHS of the claim, the second is obtained by plugging in $\mathsf{W}_0(\mathcal{E}) = (0)$ and simplifying, the third follows the definition of the kappa factor and expanding the sum over V, and the final equality follows (25). This completes the base case.

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

$$
\pi_{ij}(\mathcal{E}) = \sum_{W \in \mathsf{W}_i(\mathcal{E})} \sum_{E_W \subseteq \mathcal{E}_W^{W \cup \{j\} \setminus \{0\}}} \kappa_W^{\mathcal{E}}(E_W) \sum_{V \subseteq \mathtt{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathcal{E}_{-W}) \tag{26}
$$

for all $j : (i,j) \in \mathcal{E}$ implies

$$
\pi_{jk}(\mathcal{E}) = \sum_{W \in \mathsf{W}_j(\mathcal{E})} \sum_{E_W \subseteq \mathcal{E}_W^{W \cup \{k\} \setminus \{0\}}} \kappa_W^{\mathcal{E}}(E_W) \sum_{V \subseteq \mathtt{Edge}(W \cup \{k\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathcal{E}_{-W})
$$

for all $k : (j,k) \in \mathcal{E}$. Consider arbitrary edge $(j,k) \in \mathcal{E}$. Observe that

$$
\begin{aligned}
\pi_{jk}(\mathcal{E}) &= \sum_{E_j \subseteq \mathcal{E}_j \setminus \{(j,k)\}} w_{\mathcal{E}_j}(E_j) \times \{v_j(E_j \cup \{(j,k)\} \mid \mathcal{E}_{-j}) - v_j(E_j \mid \mathcal{E}_{-j})\} \\
&= \sum_{E_j \subseteq \mathcal{E}_j \setminus \{(j,k)\}} w_{\mathcal{E}_j}(E_j) \sum_{i \in \mathsf{P}_j(\mathcal{E})} \{\pi_{ij}(E_j \cup \{(j,k)\}, \mathcal{E}_{-j}) - \pi_{ij}(E_j, \mathcal{E}_{-j})\} \\
&= \sum_{E_j \subseteq \mathcal{E}_j \setminus \{(j,k)\}} w_{\mathcal{E}_j}(E_j) \sum_{i \in \mathsf{P}_j(\mathcal{E})} \sum_{W \in \mathsf{W}_j(\mathcal{E})} \sum_{E_W \subseteq \mathcal{E}_W^{W \cup \{j\} \setminus \{0\}}} \kappa_W^{\mathcal{E}}(E_W) \\
&\quad \times \sum_{V \subseteq \mathtt{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \\
&\quad \times \big\{y(E_W \cup V, E_j \cup \{(j,k)\}, \mathcal{E}_{-(W \cup j)}) - y(E_W \cup V, E_j, \mathcal{E}_{-(W \cup j)})\big\} \\
&= \sum_{i \in \mathsf{P}_j(\mathcal{E})} \sum_{W \in \mathsf{W}_j(\mathcal{E})} \sum_{E_j \subseteq \mathcal{E}_j \setminus \{(j,k)\}} \sum_{E_W \subseteq \mathcal{E}_W^{W \cup \{j\} \setminus \{0\}}} w_{\mathcal{E}_j}(E_j) \kappa_W^{\mathcal{E}}(E_W) \\
&\quad \times \sum_{V \subseteq \mathtt{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \\
&\quad \times \big\{y(E_W \cup V, E_j \cup \{(j,k)\}, \mathcal{E}_{-(W \cup j)}) - y(E_W \cup V, E_j, \mathcal{E}_{-(W \cup j)})\big\} \\
&= \sum_{i \in \mathsf{P}_j(\mathcal{E})} \sum_{W \in \mathsf{W}_j(\mathcal{E})} \sum_{E_j \subseteq \mathcal{E}_j \setminus \{(j,k)\}} \sum_{E_W \subseteq \mathcal{E}_W^{W \cup \{j\} \setminus \{0\}}} w_{\mathcal{E}_j}(E_j) \kappa_W^{\mathcal{E}}(E_W) \\
&\quad \times \sum_{V \subseteq \mathtt{Edge}(W \cup \{j\})} (-1)^{|W|-|V|} \sum_{V' \subseteq \{(j,k)\}} (-1)^{1-|V'|} y(E_W \cup V, E_j \cup V', \mathcal{E}_{-(W \cup j)})
\end{aligned}
$$

$$= \sum_{W \in \mathsf{W}_j(\mathcal{E})} \sum_{E_W \subseteq \mathcal{E}_W^{W \cup \{k\} \setminus \{0\}}} \kappa_W^{\mathcal{E}}(E_W) \sum_{V \subseteq \texttt{Edge}(W \cup \{k\})} (-1)^{|W|-|V|} \times y(E_W \cup V, \mathcal{E}_{-W}).$$

The first equality follows the definition of SV, second follows the recursive definition of the characteristic function $v_j(\cdot)$, third invokes (26), and the last three equalities use basic algebra. This completes the proof. □

Appendix F. RSV characterization for a linear model with interactions

Consider the DAG in Figure 31 for illustration.

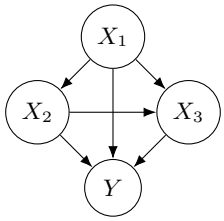


Figure 31: DAG used to illustrate the RSV characterization for a linear model with two-way interactions. We have excluded node 0 for simplicity.

Suppose X_1 is set exogenously and the true (deterministic) relations are linear with two-way interactions:

$$\begin{aligned} X_2 &= a_1^2 X_1 \\ X_3 &= a_1^3 X_1 + a_2^3 X_2 + a_{12}^3 X_1 X_2 \\ Y &= a_1^4 X_1 + a_2^4 X_2 + a_3^4 X_3 + a_{12}^4 X_1 X_2 + a_{13}^4 X_1 X_3 + a_{23}^4 X_2 X_3. \end{aligned}$$

The super-scripts in the coefficients (e.g., "2" in a_1^2) do not represent exponents but are used to indicate the child node of the corresponding edge (e.g., "2" in a_1^2 denotes node 2). In §F.1, we compute the RSV for this relatively simple example and observe a pattern, which we generalize in §F.2. Finally, we discuss its computational implications in §F.3. The key message of this section is that developing a computationally tractable RSV characterization even for a linear model (in the coefficients) with two-way interactions in the variables is challenging.

F.1 RSV computation

For simplicity [15], assume background $X_1^{(1)} = 0$ and foreground $X_1^{(2)} = 1$. Then, $Y^{(1)} = 0$ and

$$\begin{aligned} Y^{(2)} &= a_1^4 X_1^{(2)} + a_2^4 X_2^{(2)} + a_3^4 X_3^{(2)} + a_{12}^4 X_1^{(2)} X_2^{(2)} + a_{13}^4 X_1^{(2)} X_3^{(2)} + a_{23}^4 X_2^{(2)} X_3^{(2)} \\ &= a_1^4 + a_2^4 X_2^{(2)} + a_3^4 X_3^{(2)} + a_{12}^4 X_2^{(2)} + a_{13}^4 X_3^{(2)} + a_{23}^4 X_2^{(2)} X_3^{(2)} \\ &= a_1^4 + a_2^4 a_1^2 + a_3^4 X_3^{(2)} + a_{12}^4 a_1^2 + a_{13}^4 X_3^{(2)} + a_{23}^4 a_1^2 X_3^{(2)} \end{aligned}$$

15. As we show, even under this simplification, a computationally tractable characterization of RSV is challenging.

$$= a_1^4 + a_2^4 a_1^2 + a_3^4(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2) + a_{12}^4 a_1^2 + a_{13}^4(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2) + a_{23}^4 a_1^2(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2).$$

First equality is by definition. Second plugs in $X_1^{(2)} = 1$. Third plugs in $X_2^{(2)} = a_1^2$. Fourth plugs in $X_3^{(2)} = a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2$.

First, RSV considers a game at node 1, where the players are the outgoing edges $\mathsf{E}_1 = \{(1,2),(1,3),(1,4)\}$. All downstream edges are assumed to be active. The RSV of edge $(1,2)$ equals:

$$\begin{aligned}\pi_{12}^{\text{RSV}} = &\frac{1}{3}\left(v_1(\{(1,2)\}) - v_1(\emptyset)\right) + \frac{1}{6}\left(v_1(\{(1,2),(1,3)\}) - v_1(\{(1,3)\})\right)\\ &+ \frac{1}{6}\left(v_1(\{(1,2),(1,4)\}) - v_1(\{(1,4)\})\right)\\ &+ \frac{1}{3}\left(v_1(\{(1,2),(1,3),(1,4)\}) - v_1(\{(1,3),(1,4)\})\right).\end{aligned}$$

Observe that the coalition values are as follows (via bruteforce computation):

$$\begin{aligned}v_1(\emptyset) &= Y^{(1)} = 0\\ v_1(\{(1,2)\}) &= a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2\\ v_1(\{(1,3)\}) &= a_3^4 a_1^3\\ v_1(\{(1,4)\}) &= a_1^4\\ v_1(\{(1,2),(1,3)\}) &= a_2^4 a_1^2 + a_3^4(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2) + a_{23}^4 a_1^2(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2)\\ v_1(\{(1,2),(1,4)\}) &= a_1^4 + a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{12}^4 a_1^2 + a_{13}^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2\\ v_1(\{(1,3),(1,4)\}) &= a_1^4 + a_3^4 a_1^3 + a_{13}^4 a_1^3\\ v_1(\{(1,2),(1,3),(1,4)\}) &= Y^{(2)} = a_1^4 + a_2^4 a_1^2 + a_3^4(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2) + a_{12}^4 a_1^2\\ &\qquad + a_{13}^4(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2) + a_{23}^4 a_1^2(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2).\end{aligned}$$

Plugging these in π_{12}^{RSV}, we get π_{12}^{RSV} equals

$$\begin{aligned}&= \frac{1}{3}\left(a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2\right)\\ &+ \frac{1}{6}\left(a_2^4 a_1^2 + a_3^4(a_2^3 a_1^2 + a_{12}^3 a_1^2) + a_{23}^4 a_1^2(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2)\right)\\ &+ \frac{1}{6}\left(a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{12}^4 a_1^2 + a_{13}^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2\right)\\ &+ \frac{1}{3}\left(a_2^4 a_1^2 + a_3^4(a_2^3 a_1^2 + a_{12}^3 a_1^2) + a_{12}^4 a_1^2 + a_{13}^4(a_2^3 a_1^2 + a_{12}^3 a_1^2) + a_{23}^4 a_1^2(a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2)\right)\\ &= a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2 + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{2} + \frac{a_{12}^4 a_1^2}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2}\\ &+ \frac{a_{13}^4 a_{12}^3 a_1^2}{3}.\end{aligned}$$

We have arranged the terms using the denominators 1, 2, and 3. Note that each of the numerator term appears in $Y^{(2)}$ and hence, π_{12}^{RSV} splits $Y^{(2)}$ via these term-specific denominators. Interestingly, there is an underlying pattern behind these denominators. Observe

that the denominator corresponds to the number of edges (out of the 3 edges of interest in this game, i.e., $\mathsf{E}_1 = \{(1,2),(1,3),(1,4)\}$) that play a role in the term. For example, in $a_{13}^4 a_{12}^3 a_1^2$, all three edges of interest appear:

- a_{13}^4: $(1,4)$ and $(3,4)$
- a_{12}^3: $(1,3)$ and $(2,3)$
- a_1^2: $(1,2)$.

Hence, the corresponding denominator equals 3. Extrapolating this pattern, we can decompose $Y^{(2)}$ as follows:

$$\begin{aligned}
\pi_{12}^{\text{RSV}} &= a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2 + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{2} + \frac{a_{12}^4 a_1^2}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} \\
&\quad + \frac{a_{13}^4 a_{12}^3 a_1^2}{3} \\
\pi_{13}^{\text{RSV}} &= a_3^4 a_1^3 + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{13}^4 a_1^3}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{13}^4 a_{12}^3 a_1^2}{3} \\
\pi_{14}^{\text{RSV}} &= a_1^4 + \frac{a_{12}^4 a_1^2}{2} + \frac{a_{13}^4 a_1^3}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} + \frac{a_{13}^4 a_{12}^3 a_1^2}{3}.
\end{aligned}$$

By construction of this decomposition, the sum $(\pi_{12}^{\text{RSV}} + \pi_{13}^{\text{RSV}} + \pi_{14}^{\text{RSV}})$ equals $Y^{(2)}$. As a sanity check, let's verify the value of π_{14}^{RSV} via first principles:

$$\begin{aligned}
\pi_{14}^{\text{RSV}} &= \frac{1}{3}\left(v_1(\{(1,4)\}) - v_1(\emptyset)\right) + \frac{1}{6}\left(v_1(\{(1,4),(1,3)\}) - v_1(\{(1,3)\})\right) \\
&\quad + \frac{1}{6}\left(v_1(\{(1,4),(1,2)\}) - v_1(\{(1,2)\})\right) \\
&\quad + \frac{1}{3}\left(v_1(\{(1,4),(1,2),(1,3)\}) - v_1(\{(1,2),(1,3)\})\right) \\
&= \frac{1}{3}\left(a_1^4\right) + \frac{1}{6}\left(a_1^4 + a_{13}^4 a_1^3\right) + \frac{1}{6}\left(a_1^4 + a_{12}^4 a_1^2 + a_{13}^4 a_2^3 a_1^2\right) \\
&\quad + \frac{1}{3}\left(a_1^4 + a_{12}^4 a_1^2 + a_{13}^4 (a_1^3 + a_2^3 a_1^2 + a_{12}^3 a_1^2)\right) \\
&= a_1^4 + \frac{a_{13}^4 a_1^3}{2} + \frac{a_{12}^4 a_1^2}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} + \frac{a_{13}^4 a_{12}^3 a_1^2}{3}.
\end{aligned}$$

This matches the pattern above. We don't need to verify π_{13}^{RSV} due to efficiency (flow conservation).

Having computed RSV at the node 1 game, let's flow it down. Consider node 2. It receives an inflow of π_{12}^{RSV}, which is split into the outflow $\pi_{23}^{\text{RSV}} + \pi_{24}^{\text{RSV}}$. Recall from above that

$$\begin{aligned}
\pi_{12}^{\text{RSV}} &= a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2 + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{2} + \frac{a_{12}^4 a_1^2}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} \\
&\quad + \frac{a_{13}^4 a_{12}^3 a_1^2}{3}.
\end{aligned}$$

Extrapolating the pattern from above, we get the following flow decomposition:

$$\pi_{23}^{\text{RSV}} = a_3^4 a_2^3 a_1^2 + \frac{a_{23}^4 a_1^2 a_2^3 a_1^2}{2} + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{4} + \frac{a_{13}^4 a_2^3 a_1^2}{2} + \frac{a_{13}^4 a_{12}^3 a_1^2}{3}$$
$$\pi_{24}^{\text{RSV}} = a_2^4 a_1^2 + \frac{a_{23}^4 a_1^2 a_2^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{4} + \frac{a_{12}^4 a_1^2}{2}.$$

Only two inflow terms, $a_{23}^4 a_1^2 a_2^3 a_1^2$ and $\frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{2}$, involve both the edges of interest ($(2,3)$ and $(2,4)$) and hence, are split between the two. All other terms involve just one of the two edges and hence, are allocated accordingly. As a sanity check, let's verify the value of π_{24}^{RSV} via first principles:

$$\pi_{24}^{\text{RSV}} = \frac{1}{2}\left(v_2(\{(2,4)\}) - v_2(\emptyset)\right) + \frac{1}{2}\left(v_2(\{(2,3),(2,4)\}) - v_2(\{(2,3)\})\right).$$

Observe that the coalition values are as follows:

$$v_2(\emptyset) = 0$$
$$v_2(\{(2,3)\}) = a_3^4 a_2^3 a_1^2 + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} + \frac{a_{13}^4 a_{12}^3 a_1^2}{3}$$
$$v_2(\{(2,4)\}) = a_2^4 a_1^2 + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{12}^4 a_1^2}{2}$$
$$v_2(\{(2,3),(2,4)\}) = \pi_{12}^{\text{RSV}} = a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2 + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{2} + \frac{a_{12}^4 a_1^2}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} + \frac{a_{13}^4 a_{12}^3 a_1^2}{3}.$$

Plugging these in π_{24}^{RSV}, we get π_{24}^{RSV} equals

$$= \frac{1}{2}\left(a_2^4 a_1^2 + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{12}^4 a_1^2}{2}\right) + \frac{1}{2}\left(a_2^4 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2 + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{2} + \frac{a_{12}^4 a_1^2}{2}\right)$$
$$= a_2^4 a_1^2 + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{12}^4 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_2^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{4}.$$

It matches the decomposition above. We do not need to verify π_{23}^{RSV} due to flow conservation. Hence, the pattern seems to hold when we flow down the attribution. Note that the due to flow conservation, attribution through node 3 is trivial ($\pi_{13}^{\text{RSV}} = \pi_{34}^{\text{RSV}}$) and hence,

$$\pi_{34}^{\text{RSV}} = a_3^4 a_1^3 + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{13}^4 a_1^3}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{13}^4 a_{12}^3 a_1^2}{3}.$$

This completes the computation of RSV for the example at hand.

F.2 General expression

Continuing with the example from above, consider the decomposition of Y that one obtains by substituting all of the foreground inputs:

$$Y = a_1^4 + a_2^4 a_1^2 + a_3^4 a_1^3 + a_3^4 a_2^3 a_1^2 + a_3^4 a_{12}^3 a_1^2 + a_{12}^4 a_1^2 + a_{13}^4 a_1^3 + a_{13}^4 a_2^3 a_1^2 + a_{13}^4 a_{12}^3 a_1^2$$

$$+ a_{23}^4 a_1^2 a_1^3 + a_{23}^4 a_1^2 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_{12}^3 a_1^2. \tag{27}$$

Since the background equals 0, (27) denotes the total effect. The decomposition has 12 terms and we will use the notation $\mathtt{a}$ to denote an individual term and the notation "$\sum_{\mathtt{a}\in Y}$" to sum over all such terms in Y (notation abuse since Y is not a set). Given this decomposition, RSV for the topmost game (edges $(1,2)$, $(1,3)$, and $(1,4)$) is as follows:

$$\pi_{1j}^{\text{RSV}} = \sum_{\mathtt{a}\in Y} \frac{\mathtt{a} \times \mathbb{I}\{(1,j) \in \mathtt{a}\}}{\mathtt{unique}_1(\mathtt{a})} \ \forall j \in \{2,3,4\},$$

where $\mathbb{I}\{(1,j) \in \mathtt{a}\}$ denotes whether edge $(1,j)$ appears in term $\mathtt{a}$ or not (another notation abuse since $\mathtt{a}$ is not a set) and $\mathtt{unique}_1(\mathtt{a})$ denotes the number of *unique* edges in $\mathtt{a}$ out of the outgoing edges of node 1 (sub-script "1" in $\mathtt{unique}_1(\cdot)$ captures this dependence on node 1). For example, a_1^4 has only edge $(1,4)$ and $\mathtt{unique}_1(a_1^4) = 1$ whereas $a_3^4 a_{12}^3 a_1^2$ has four edges ($(3,4)$, $(1,3)$, $(2,3)$, and $(1,2)$) but only two of these four emit from node 1 and hence, $\mathtt{unique}_1(a_3^4 a_{12}^3 a_1^2) = 2$.

Flowing down this attribution can be done using the same formula. In particular, everything remains the same except the value that is split. For example, at node 2, instead of splitting Y, we split the value it receives from the top, i.e., π_2^{RSV}, which equals π_{12}^{RSV} in this example. The formula above gives us a decomposition of π_{12}^{RSV} (and hence, π_2^{RSV}):

$$\pi_2^{\text{RSV}} = a_2^4 a_1^2 + a_3^4 a_2^3 a_1^2 + a_{23}^4 a_1^2 a_2^3 a_1^2 + \frac{a_3^4 a_{12}^3 a_1^2}{2} + \frac{a_{23}^4 a_1^2 a_1^3}{2} + \frac{a_{23}^4 a_1^2 a_{12}^3 a_1^2}{2} + \frac{a_{12}^4 a_1^2}{2} + \frac{a_{13}^4 a_2^3 a_1^2}{2} + \frac{a_{13}^4 a_{12}^3 a_1^2}{3}.$$

The decomposition has 9 terms and as before, we will use the notation $\mathtt{a}$ to denote an individual term (including the denominator that comes with it). Given this decomposition, RSV for the game at node 2 (edges $(2,3)$ and $(2,4)$) is as follows:

$$\pi_{2j}^{\text{RSV}} = \sum_{\mathtt{a}\in\pi_2^{\text{RSV}}} \frac{\mathtt{a} \times \mathbb{I}\{(2,j) \in \mathtt{a}\}}{\mathtt{unique}_2(\mathtt{a})} \ \forall j \in \{3,4\}.$$

Note that we use $\mathtt{unique}_2(\cdot)$ (sub-script changed from 1 to 2), which captures the number of unique edges out of the outgoing edges at node 2. The general pattern follows (formal proof omitted for brevity).

Given this characterization, we have a handle on computing RSV without using the bruteforce recursions in Algorithms 1 and 2. However, as we discuss next, the such a procedure runs into computational tractability issues (primarily due to the interaction terms).

F.3 Runtime

It should be clear that the runtime of this technique is primarily driven by the number of terms in the decomposition of Y as in (27), which we denote by $|Y|$. In particular, the runtime is upper bounded by $\mathcal{O}(|\mathsf{E}|\times|Y|)$, where $|\mathsf{E}|$ denotes the number of edges in the graph. This is because computing π_{ij}^{RSV} for an arbitrary $(i,j) \in \mathsf{E}$ involves a sum over at most $|Y|$ terms. Accordingly, the question of interest is as follows: how many terms does Y have?

In this direction, consider a fully dense DAG with n input nodes $(X_1, \ldots, X_n)$ and outcome node Y. Then, the structural equations for a linear model with two-way interactions are:

$$\begin{aligned}
&X_1 \text{ exogenous} \\
&X_2 = a_1^2 X_1 \\
&X_3 = a_1^3 X_1 + a_2^3 X_2 + a_{12}^3 X_1 X_2 \\
&\quad \vdots \\
&X_n = \sum_{i=1}^{n-1} a_i^n X_i + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} a_{ij}^n X_i X_j \\
&Y = \sum_{i=1}^{n} a_i^{n+1} X_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij}^{n+1} X_i X_j.
\end{aligned}$$

Obtaining the decomposition of Y (especially when X_1 equals 1) is conceptually straightforward (though quite mechanical): starting from X_1, keep performing forward-substituition.

Denote by κ_j the number of terms that appear in X_j for $j = 1, \ldots, n+1$. That is, $|Y|$ equals κ_{n+1}. Given the structural equations as above, the sequence $\{\kappa_1, \ldots, \kappa_{n+1}\}$ obeys the following recursion:

$$\kappa_j = \sum_{i=1}^{j-1} \kappa_i + \sum_{h=1}^{j-2} \sum_{i=h+1}^{j-1} \kappa_h \kappa_i \;\; \forall j = 2, \ldots, n+1.$$

$\kappa_1 = 1$ initializes the recursion. Unfortunately, this sequence explodes quite fast, as shown in Figure 32. Hence, the technique of decomposing Y and using it to attribute via the $\texttt{unique}(\cdot)$ characterization seems computationally intractable. Of course, this is not an impossibility result in general for SEMs with linear main effects and 2-way interaction terms, as there might some other characterization that might be computationally tractable. However, it seems rather challenging to come up with one, primarily due to the presence of the interaction terms.

Appendix G. Proof of Proposition 10

Proposition 14 *For edge $(i,j) \in \mathsf{E}$, $\hat{\pi}_{ij}^{RSV}$ is an unbiased estimator of π_{ij}^{RSV} with variance decaying at a rate of $1/S$. Furthermore, for all $t \geq 0$,*

$$\mathbb{P}\left\{\left|\pi_{ij}^{RSV} - \hat{\pi}_{ij}^{RSV}\right| \geq t\right\} \leq 2\exp\left(-\beta_{ij} S t^2\right)$$

for some $\beta_{ij} > 0$ finite. (The probability measure $\mathbb{P}\{\cdot\}$ here denotes the Monte-Carlo sampling distribution of the estimator $\hat{\pi}_{ij}^{RSV}$.)

Proof The unbiasedness and $1/S$ rate of variance decay directly follow our construction of the Monte-Carlo scheme. For the sample complexity, first, observe that given S samples, Hoeffding's inequality (Hoeffding, 1994) implies the following for $W \in \mathsf{W}_i$: for all $t \geq 0$,

$$\mathbb{P}\left\{\left|\mu_W^j - \hat{\mu}_W^j\right| \geq t\right\} \leq 2\exp\left(\frac{-2S^2t^2}{\sum_{s=1}^{S}(\max(\hat{\mu}_W^j(s)) - \min(\hat{\mu}_W^j(s)))^2}\right),$$

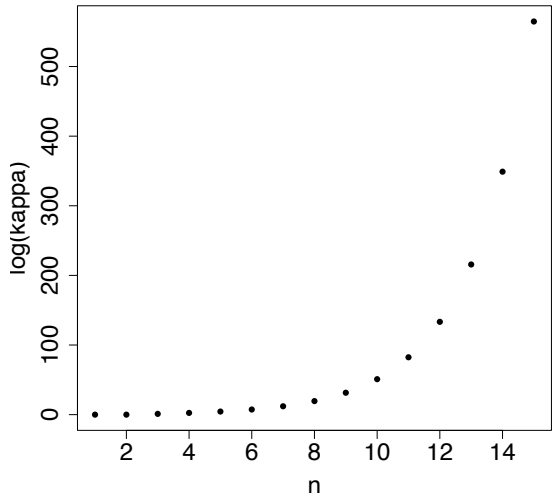


Figure 32: Growth of κ_n (on the log scale) as a function of n.

where $\min(\hat{\mu}_W^j(s))$ and $\max(\hat{\mu}_W^j(s))$ capture the range of $\hat{\mu}_W^j(s)$ over s, i.e., $\hat{\mu}_W^j(s) \in [\min(\hat{\mu}_W^j(s)), \max(\hat{\mu}_W^j(s))]$ for all s. Since the range is independent of s, we get

$$\mathbb{P}\left\{\left|\mu_W^j - \hat{\mu}_W^j\right| \geq t\right\} \leq 2\exp\left(\frac{-2St^2}{(\max(\hat{\mu}_W^j) - \min(\hat{\mu}_W^j))^2}\right). \tag{28}$$

Note that (28) bounds the gap between μ_W^j and $\hat{\mu}_W^j$ (i.e., for a given path $W \in \mathsf{W}_i$) whereas we are ultimately interested in estimating π_{ij}^{RSV}, which equals $\sum_{W \in \mathsf{W}_i} \mu_W^j$. It is possible to leverage (28) along with a corresponding concentration inequality for sub-Gaussian random variables to obtain a similar concentration bound for the difference between π_{ij}^{RSV} and $\hat{\pi}_{ij}^{\text{RSV}}$. With the error equal to

$$\pi_{ij}^{\text{RSV}} - \hat{\pi}_{ij}^{\text{RSV}} = \sum_{W \in \mathsf{W}_i} \underbrace{\left(\mu_W^j - \hat{\mu}_W^j\right)}_{=: Z_W^j},$$

Theorem 2.6.2 of Vershynin (2018) implies the following sample complexity: for all $t \geq 0$,

$$\mathbb{P}\left\{\left|\pi_{ij}^{\text{RSV}} - \hat{\pi}_{ij}^{\text{RSV}}\right| \geq t\right\} = \mathbb{P}\left\{\left|\sum_{W \in \mathsf{W}_i} Z_W^j\right| \geq t\right\} \leq 2\exp\left(\frac{-2c_{ij}St^2}{\sum_{W \in \mathsf{W}_i} ||Z_W^j||_{\psi_2}^2}\right), \tag{29}$$

where

$$c_{ij} := \min_{W \in \mathsf{W}_i}\left\{\frac{1}{(\max(\hat{\mu}_W^j) - \min(\hat{\mu}_W^j))^2}\right\}$$

and for a random variable Z, the norm

$$||Z||_{\psi_2} := \inf \left\{ a \geq 0 : \mathbb{E}\left[\exp(Z^2/a^2)\right] \leq 2 \right\},$$

which is finite if and only if Z is sub-Gaussian. For us, note that Z_W^j is sub-Gaussian due to (28) for all $W \in \mathsf{W}_i$. Setting

$$\beta_{ij} = \frac{2c_{ij}}{\sum_{W \in \mathsf{W}_i} ||Z_W^j||_{\psi_2}^2}$$

in (29) completes the proof. □

Appendix H. Sampling paths in linear time and space

Here, we supplement our discussion at the end of §6.1 by illustrating how we can sample paths uniformly at random from W_j (set of all paths from node 0 to j) in linear time and space for node $j \in \mathsf{N}^+ \setminus \{0\}$. For simplicity, we showcase the procedure on the topologically sorted DAG in Figure 33a. (Note that topological sorting a DAG takes linear time.)

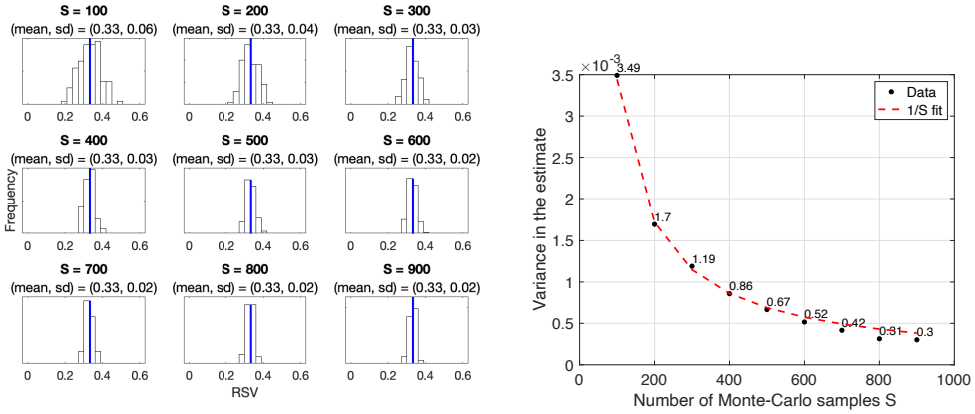
(a) DAG **(b)** Cumulative values $[d_j]_{j \in \mathsf{N}^+}$

Figure 33: The DAG and the corresponding cumulative values used to illustrate our sampling procedure.

We focus on node 6 (Y) and show how we can generate uniform samples from W_6. To do so, we compute the cumulative values (Figure 33b):

$$\begin{aligned} d_0 &:= 1 \\ d_j &:= \sum_{i \in \mathsf{P}_j} d_i \ \forall j \in \mathsf{N}^+ \setminus \{0\}. \end{aligned}$$

Observe that, by construction, d_j equals $|\mathsf{W}_j|$ (number of paths from node 0 to node j) for all $j \in \mathsf{N}^+ \setminus \{0\}$. Hence, to generate a uniform sample from W_6, we can employ the following procedure:



(a) Estimator distribution **(b)** Variance as a function of S

Figure 34: Numerical illustration of our estimation scheme for $(\boldsymbol{L}, \boldsymbol{M}) = (\mathbf{4}, \mathbf{3})$. Note that the exact RSV (of the top-right edge) equals $1/M$, which equals $1/3$ in this case (blue vertical line in plot (a)). The estimator's mean is close to 0.33 for all values of S and the variance decays at a rate of $1/S$.

1. Start at node 6 and sample a node from its set of parents $\mathsf{P}_6 = \{1, 4, 5\}$ with sampling weights $(1, 3, 6)$, i.e., the cumulative values (d_1, d_4, d_5) of the parents.

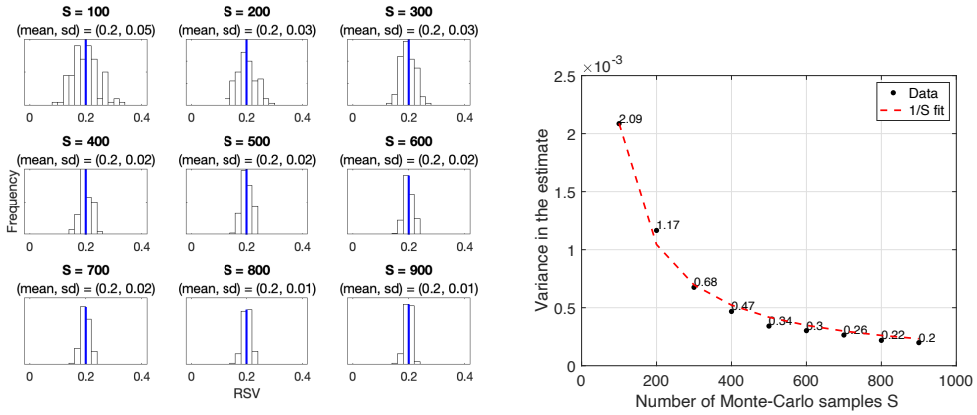
2. Suppose the sampled node is $j \in \mathsf{P}_6$. Repeat the procedure at node j by sampling from P_j with weights $(d_i)_{i \in \mathsf{P}_j}$.

3. Keep doing so until node 0 is reached and stitch together the sampled edges as a sample path from W_6.

Since the DAG is topologically sorted, we will reach node 0 in at most n steps and the entire procedure runs in linear time and space, without enumerating W_6. It should be clear that the procedure samples paths uniformly at random from W_6 and the same idea can be used to sample from W_j for an arbitrary node $j \in \mathsf{N}^+ \setminus \{0\}$. Furthermore, generalizing this procedure to an arbitrary DAG is straightforward.

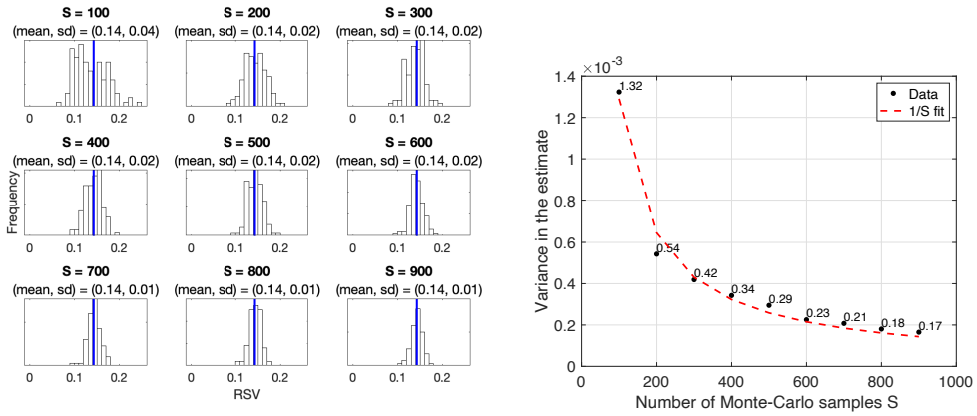
Appendix I. Additional plots for the numerical illustration of our estimation scheme

See Figures 34 to 39.



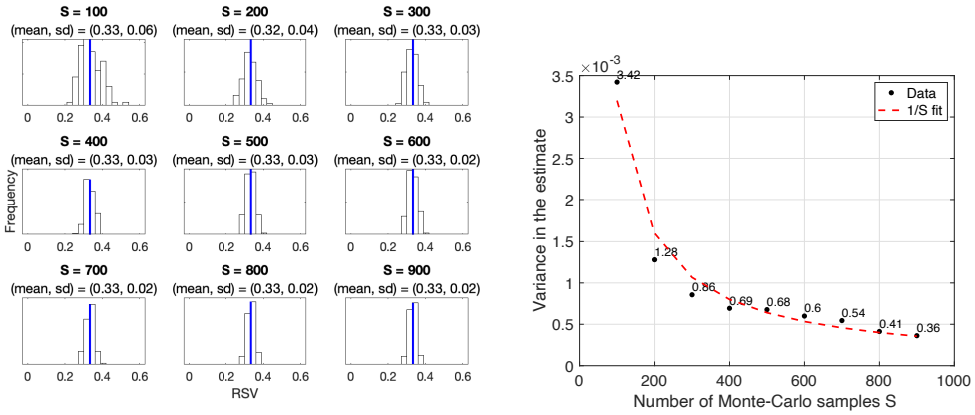
(a) Estimator distribution **(b)** Variance as a function of S

Figure 35: Numerical illustration of our estimation scheme for $(\boldsymbol{L}, \boldsymbol{M}) = (\mathbf{4}, \mathbf{5})$. Note that the exact RSV (of the top-right edge) equals $1/M$, which equals $1/5$ in this case (blue vertical line in plot (a)). The estimator's mean is close to 0.20 for all values of S and the variance decays at a rate of $1/S$.



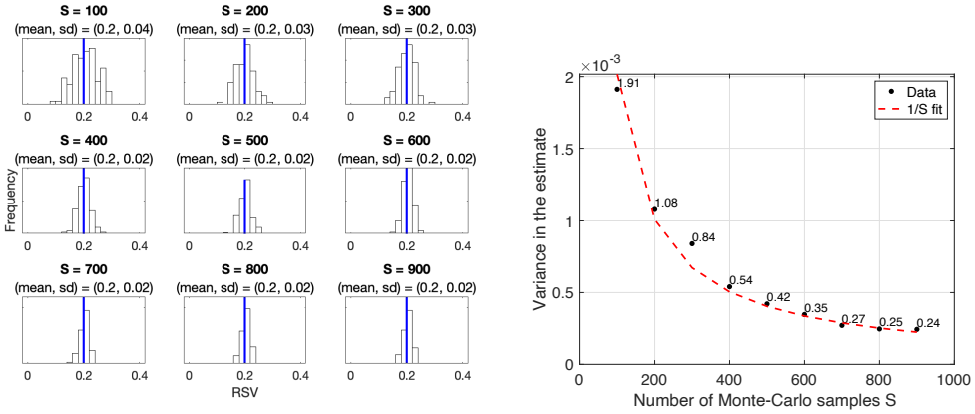
(a) Estimator distribution **(b)** Variance as a function of S

Figure 36: Numerical illustration of our estimation scheme for $(\boldsymbol{L}, \boldsymbol{M}) = (\mathbf{4}, \mathbf{7})$. Note that the exact RSV (of the top-right edge) equals $1/M$, which equals $1/7$ in this case (blue vertical line in plot (a)). The estimator's mean is close to 0.14 for all values of S and the variance decays at a rate of $1/S$.



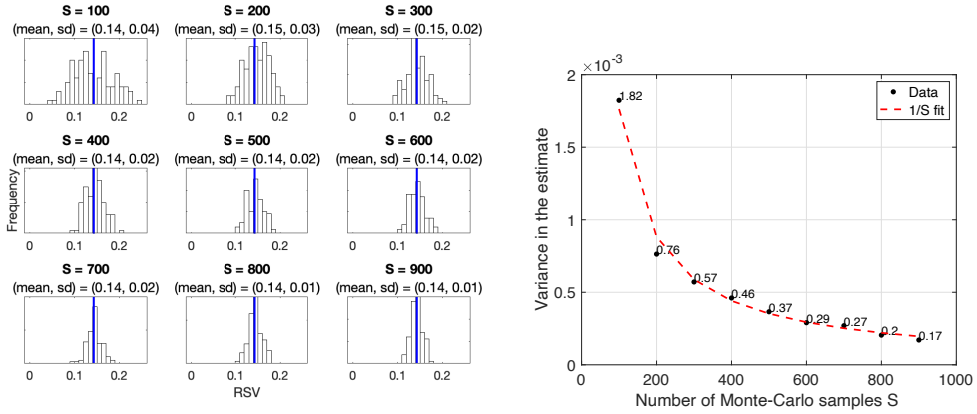
(a) Estimator distribution **(b)** Variance as a function of S

Figure 37: Numerical illustration of our estimation scheme for $(\boldsymbol{L}, \boldsymbol{M}) = (\mathbf{5}, \mathbf{3})$. Note that the exact RSV (of the top-right edge) equals $1/M$, which equals $1/3$ in this case (blue vertical line in plot (a)). The estimator's mean is close to 0.33 for all values of S and the variance decays at a rate of $1/S$.



(a) Estimator distribution **(b)** Variance as a function of S

Figure 38: Numerical illustration of our estimation scheme for $(\boldsymbol{L}, \boldsymbol{M}) = (\mathbf{5}, \mathbf{5})$. Note that the exact RSV (of the top-right edge) equals $1/M$, which equals $1/5$ in this case (blue vertical line in plot (a)). The estimator's mean is close to 0.20 for all values of S and the variance decays at a rate of $1/S$.



(a) Estimator distribution **(b)** Variance as a function of S

Figure 39: Numerical illustration of our estimation scheme for $(\boldsymbol{L}, \boldsymbol{M}) = (\mathbf{5}, \mathbf{7})$. Note that the exact RSV (of the top-right edge) equals $1/M$, which equals $1/7$ in this case (blue vertical line in plot (a)). The estimator's mean is close to 0.14 for all values of S and the variance decays at a rate of $1/S$.

References

Duane F Alwin and Robert M Hauser. The decomposition of effects in path analysis. *American Sociological Review*, pages 37–47, 1975.

Nihat Ay and Daniel Polani. Information flows in causal networks. *Advances in Complex Systems*, 11(01):17–41, 2008.

Reuben M Baron and David A Kenny. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173, 1986.

Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404, 1975.

Alan S Blinder. Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resources*, pages 436–455, 1973.

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Javier Castro, Daniel Gómez, Elisenda Molina, and Juan Tejada. Improving polynomial estimation of the Shapley value by stratified random sampling with optimum allocation. *Computers & Operations Research*, 82:180–188, 2017.

Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

Hana Chockler and Joseph Y Halpern. Responsibility and Blame: A Structural-Model Approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT press, 2009.

Brian Dalessandro, Claudia Perlich, Ori Stitelman, and Foster Provost. Causally motivated attribution for online advertising. In *Proceedings of the 6th International Workshop on Data Mining for Online Advertising and Internet Economy*, pages 1–9, 2012.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Otis Dudley Duncan. Path analysis: Sociological examples. *American Journal of Sociology*, 72(1):1–16, 1966.

Samuel Ferey and Pierre Dehez. Multiple causation, apportionment, and the Shapley value. *The Journal of Legal Studies*, 45(1):143–171, 2016.

Sergio Firpo and Cristine Pinto. Identification and estimation of distributional impacts of interventions using changes in inequality measures. *Journal of Applied Econometrics*, 31 (3):457–486, 2016.

Nicole Fortin, Thomas Lemieux, and Sergio Firpo. Decomposition methods in economics. In *Handbook of Labor Economics*, volume 4, pages 1–102. Elsevier, 2011.

John Fox. Effect analysis in structural equation models: Extensions and simplified methods of computation. *Sociological Methods & Research*, 9(1):3–28, 1980.

Meir Friedenberg and Joseph Y Halpern. Blameworthiness in multi-agent settings. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, volume 33, pages 525–532, 2019.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10:524, 2019.

Arthur S Goldberger. Structural equation methods in the social sciences. *Econometrica*, pages 979–1001, 1972.

Ned Hall. Two concepts of causation. *Causation and Counterfactuals*, pages 225–276, 2004.

Joseph Halpern and Max Kleiman-Weiner. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Joseph Y Halpern and Judea Pearl. Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, 56(4): 889–911, 2005.

Jonathan M Henshaw, Michael B Morrissey, and Adam G Jones. Quantifying the causal pathways contributing to natural selection. *Evolution*, 74(12):2560–2574, 2020.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309, 2010.

Kosuke Imai, Luke Keele, Dustin Tingley, and Teppei Yamamoto. Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, 105(4):765–789, 2011.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf, et al. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

Evelyn M Kitagawa. Components of a Difference Between Two Rates. *Journal of the American Statistical Association*, 50(272):1168–1194, 1955.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, pages 33–50, 1978.

Roger Koenker and Kevin F Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, 2001.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30, 2017.

David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation Analysis. *Annual Review of Psychology*, 58:593–614, 2007.

Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. Bounding the estimation error of sampling-based Shapley value approximation. *arXiv preprint arXiv:1306.4265*, 2013.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(6), 2006.

Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling Permutations for Shapley Value Estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022.

Ronald Oaxaca. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, pages 693–709, 1973.

Judea Pearl. Direct and Indirect Effects. *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, 32:411–420, 2001.

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

Judea Pearl. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–62, 2010.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Bezalel Peleg and Peter Sudhölter. *Introduction to the Theory of Cooperative Games*, volume 34. Springer Science & Business Media, 2007.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. The MIT Press, 2017.

Drago Plecko and Elias Bareinboim. Causal Fairness Analysis. *Foundations and Trends in Machine Learning*, 17(3):1–238, 2024.

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

Raghav Singal, George Michailidis, and Hoiyi Ng. Flow-based attribution in graphical models: a recursive Shapley approach. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021.

Raghav Singal, Omar Besbes, Antoine Desir, Vineet Goyal, and Garud Iyengar. Shapley meets uniform:: An axiomatic framework for attribution in online advertising. *Management Science*, 68(10):7457–7479, 2022.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

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

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3319–3328. PMLR, 2017.

Tyler J VanderWeele, Sunni L Mumford, and Enrique F Schisterman. Conditioning on intermediates in perinatal epidemiology. *Epidemiology*, 23(1):1, 2012.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Christopher Weible, Paul A Sabatier, and Mark Lubell. A comparison of a collaborative and top-down approach to the use of science in policy: Establishing marine protected areas in California. *Policy Studies Journal*, 32(2):187–207, 2004.

Sewall Wright. On the nature of size factors. *Genetics*, 3(4):367, 1918.

Sewall Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5 (3):161–215, 1934.

Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making – the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018a.

Junzhe Zhang and Elias Bareinboim. Non-parametric path analysis in structural causal models. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018b.

Lu Zhang, Yongkai Wu, and Xintao Wu. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.