Causal Graph Pruning with Do-Shapley: Enhancing Efficiency and Visualization

Bodun Du

24.1.2025

Abstract

The first part of this paper focuses on summarizing the development of Shapley Value in the context of explainable artificial intelligence (XAI), both in terms of its practical applications and historical evolution. It provides a comprehensive overview of the "all about Shapley Value" story in XAI.

The second part of this paper shifts its focus to integrating Shapley Value with causal graph theory, leading to the derivation of the do-Shapley method, which is applied innovatively to a simple pruning example. By preserving the core principles of fairness and consistency inherent in Shapley Value, this method leverages interventional causal inference to achieve precise quantification of causal contributions in high-dimensional feature spaces.

Through a pruning strategy based on reasonable thresholds, the approach not only enhances the readability of causal graphs but also effectively reduces computational complexity while maintaining the fairness of feature importance distribution. These characteristics make the method well-suited for a variety of complex systems and scenarios in XAI, extending the traditional Shapley Value's applicability to contexts with significant dependency and causal interactions.

1 Background: The Implementation of Shapley Analysis in XAI

1.1 Why XAI Needs Shapley Value

In the field of explainable artificial intelligence (XAI), the importance of Shapley Value lies primarily in its fairness and consistency when measuring feature contributions (these two properties will be detailed at the end of the paragraph).

Shapley Value originates from cooperative game theory, where it was initially used to measure the payoffs allocated to each beneficiary. Over time, this same methodology has been directly applied to measure the contribution of each participant to the overall outcome—referred to as feature contribution allocation.

This makes Shapley Value a reliable "measurement of contribution," which is also the focal point of this paper.

Example 1: Consider a decision-making process where a resolution is passed if at least two voters agree. Let $N = \{A, B, C\}$ represent three voters, and S denote a coalition supporting the resolution (e.g., if both A and B support, then $S = \{A, B\}$). The decision rule can be expressed as:

$$v(S) = \begin{cases} 1 & \text{if } |S| \geq 2 \\ 0 & \text{if } |S| < 2 \end{cases}$$

To quantify the contribution of voter A to the outcome $v(S)$, we list A's marginal contributions and their weights:

S	$v(S)$	$v(S \cup \{A\})$	$v(S \cup \{A\}) - v(S)$
$\emptyset$	0	0	0
$\{B\}$	0	1	1
$\{C\}$	0	1	1
$\{B, C\}$	1	1	0

S	$\lvert S\rvert$	Weight: $\frac{\lvert S\rvert!\cdot(\lvert N\rvert-\lvert S\rvert-1)!}{\lvert N\rvert!}$
$\emptyset$	0	$\frac{0!\cdot 2!}{3!} = \frac{2}{6}$
$\{B\}$	1	$\frac{1!\cdot 1!}{3!} = \frac{1}{6}$
$\{C\}$	1	$\frac{1!\cdot 1!}{3!} = \frac{1}{6}$
$\{B, C\}$	2	$\frac{2!\cdot 0!}{3!} = \frac{2}{6}$

The "measurement of contribution" for voter A, denoted as $\phi_A(v)$, is calculated by multiplying marginal contributions with their respective weights and summing them:

$$\phi_A(v) = \sum_{S \subseteq N \setminus \{A\}} \text{Weight} \times \text{Marginal Contribution}$$

$$\phi_A(v) = \frac{2}{6} \cdot 0 + \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 1 + \frac{2}{6} \cdot 0 = \frac{2}{6} = \frac{1}{3}.$$

For B and C, due to the symmetry of the voting system, the results are identical. Each voter contributes $\frac{1}{3}$ to the overall decision.

This calculation demonstrates that Shapley Value, as defined, provides a robust representation of contributions. It aligns well with our expectations for an ideal "measurement of contribution" in similar or even more complex scenarios.

Axiomatic Properties: From this example, we outline the conditions under which Shapley Value operates:

- **Additivity/Linearity:** $\phi_i(v_1 + v_2) = \phi_i(v_1) + \phi_i(v_2)$

- **Symmetry:** $\phi_i(v) = \phi_j(v), \quad \text{if } v(S \cup \{i\}) - v(S) = v(S \cup \{j\}) - v(S), \forall S.$
- **Null Player Property:** $\phi_i(v) = 0, \quad \text{if } v(S \cup \{i\}) = v(S), \forall S \subseteq N.$
- **Efficiency:** $\phi_A(v) + \phi_B(v) + \phi_C(v) = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1.$

These properties not only encompass the example above but also apply to scenarios with larger numbers of participants or weighted voting systems, such as network bandwidth allocation or profit sharing. Regardless of scale, Shapley Value can quantify each party's contribution through the same computation method.

More broadly, these properties serve as the axiomatic foundation of Shapley Value, as defined in:

- Rozemberczki, B., Watson, L., Bayer, P., Yang, H.T., Kiss, O., Nilsson, S., and Sarkar, R. (2022). *The Shapley Value in Machine Learning.* arXiv preprint arXiv:2202.05594.
- Aumann, R.J., and Shapley, L.S. (1974). *Values of Non-Atomic Games.* Princeton University Press.

Applications in Machine Learning: Through the constraints and axioms outlined above, Shapley Value has been successfully introduced into machine learning to provide a unified method for measuring feature contributions across various tasks (e.g., regression, classification, and deep learning models). By defining the scope of applicability, the guarantees of "fairness" and "consistency" are evident through logical derivation:

- **Fairness:** Symmetry and the null player property ensure that "equal contributors" receive equal values and "non-contributors" are assigned a value of zero.
- **Consistency:** The additivity principle ensures that if a feature contributes more in all subsets, it is assigned a higher Shapley Value.

These properties establish Shapley Value as a method for feature importance measurement that is both fair and consistent, providing an objective and reproducible interpretability metric. This, in turn, strengthens model transparency and result trustworthiness.

1.2 Shapley Value for Understanding Complex Systems and Its Improvements in Ensemble Networks and Federated Networks

Although the theoretical foundation of Shapley Value assumes linear additivity of feature contributions, it remains capable of providing stable feature importance estimates in nonlinear systems through approximation methods such as local linearization and weighted sampling. Particularly in ensemble networks

and federated networks, these local approximations ensure its adaptability in high-dimensional scenarios. The use of Shapley Value in these contexts brings the following benefits:

- **Theoretical Robustness Leading to Practical Generality:** Shapley Value provides a standardized, axiom-based mathematical method. Even in complex systems where nonlinearities or high-dimensional interactions cannot be fully captured, it still offers a globally interpretable perspective on feature contributions. It does not rely on explicit model assumptions or structural requirements, making it applicable to most machine learning models or game-theoretic environments.
- **Approximation Advantages:** While Shapley Value only provides approximations, these approximations highlight key feature importance, offering a practical layer of interpretability for systems. In many real-world applications, such approximate explanations are sufficient to identify critical drivers in the system. For example, in ensemble networks, the SHAP method employs sampling and local linear approximations to identify major features accurately while reducing computational complexity.

Thus, when assessing the importance of individual features to overall system decisions, Shapley Value serves as a measurement tool that balances collaboration and fair allocation, solidifying its position in both industry and academia.

Given the high-dimensional, dynamic, and cross-model characteristics of ensemble networks and federated networks in industrial and academic applications, researchers have optimized various algorithms to enhance the usability and performance of Shapley Value in these domains. Among the most representative approaches is SHAP (SHapley Additive exPlanations).

SHAP reduces the computational cost of full cooperative game calculations on large-scale, high-dimensional data by approximating feature importance through local approximations. It employs linearization and approximation strategies, such as assuming local additivity or sampling marginal contributions, to accelerate the explanation process.

Additionally, in federated networks, where data is distributed and subject to privacy constraints, researchers have proposed Shapley Value approximation methods based on secure multi-party computation or cryptographic techniques. These methods enable secure and accurate feature importance evaluation across nodes.

As a result, with improvements and optimizations tailored to specific domain requirements, Shapley Value is gradually becoming a core tool that combines rigor and scalability in the interpretability and usability of complex systems.

1.3 Shapley Values Are Widely Used in XAI

1.3.1 How the SHAP Family Performs Well

In recent years, the SHAP methodology, designed to reduce computational complexity, has demonstrated outstanding interpretability and stability across vari-

ous applications, including deep learning models, tree models, and hybrid models. Due to the SHAP family's ability to provide intuitive explanations for both local and global model decisions based on the properties of Shapley Value, it has proven to be portable and highly accurate across different models and data scales. This has led to its widespread adoption in high-risk or high-value domains, such as medical diagnosis, financial risk management, and public service evaluation, with substantial testing and deployment by numerous research groups.

For example, KernelSHAP effectively reduces computational complexity in high-dimensional data scenarios through random sampling of feature subsets and linear regression fitting. TreeSHAP, on the other hand, leverages path information from decision tree structures to compute SHAP values quickly and accurately within tree models. These optimization techniques significantly enhance the scalability of Shapley Value in complex systems.

- **KernelSHAP:** Randomly samples feature subsets and approximates their contributions using linear regression fitting. Applicable to general black-box models and retains Shapley Value's theoretical guarantees. However, it can be computationally expensive for high-dimensional data and depends on the quality of the sampling strategy.
- **TreeSHAP:** Accurately calculates marginal contributions within decision trees by optimizing along tree split paths. This method is fast and scalable for large datasets, but it is only applicable to structured models such as decision trees.
- **DeepSHAP:** Utilizes gradient and backpropagation principles to provide hierarchical explanations for neural networks. It is efficient for deep neural network structures but may be influenced by factors such as network architecture, activation functions, and regularization techniques.
- **GradientSHAP:** Performs random interpolation of input samples and uses gradients to approximate integrals. It is well-suited for differentiable models without explicit tree structures but requires significant sampling to reduce variance and faces challenges in high-dimensional inputs.
- **PartitionSHAP:** Clusters features based on dependency relationships and computes local contributions for each cluster. It reduces computational cost by grouping dependent features, while preserving theoretical consistency. However, the accuracy of the results depends on the correctness of the feature dependency structure.

References: WANG, Zhaohua, LIU, Jie, WANG, Bo, DENG, Nana, NIE, Fuhua. *Research on mining and applications of individual heterogeneity factors in resident demand response by integrating machine learning and SHAP value algorithm.* Systems Engineering - Theory & Practice, 2024, 44(7): 2247-2259. (https://doi.org/10.12011/SETP2023-0677).

Table 1: Comparison of SHAP Methods

Method	Key Idea	Applicable Mo
KernelSHAP	Approximates contributions via linear regression on sampled subsets	General black-bo
TreeSHAP	Uses tree split paths to compute contributions	Tree models (e.g
DeepSHAP	Leverages gradients and backpropagation for layered explanation	Neural networks
GradientSHAP	Uses random interpolation and gradients for integral approximation	General different
PartitionSHAP	Clusters features based on dependencies for local contribution	General black-bo

2 Shapley Value

2.1 Mathematical Basis and Transferable Utility (TU) Setting

Under the framework defined by the four core axioms mentioned earlier, the total contribution can be entirely redistributed as individual contributions. Participants' contributions can be freely transferred among them without affecting the total contribution. This assumed game setting is referred to as a Transferable Utility (TU) cooperative game.

Reference: The Shapley Value in Machine Learning. The original text uses the general term "collective value" to describe the total value of a coalition, referring to profit or cost. In the context of Shapley Value, this total value is reinterpreted as the fundamental metric for marginal contribution allocation.

Thus, the Shapley Value defined above can be more accurately categorized in the XAI framework as a "measurement of marginal contribution" method. While previous sections introduced several practical applications using local approximations and algorithmic optimizations, the mathematical rules underlying Shapley Value reveal inherent limitations in adapting to complex XAI systems, particularly in scenarios with high feature dependency or causal interactions.

Illustrative Example: Consider a simple XAI scenario: suppose three features $\{X_1, X_2, X_3\}$ are used for binary classification. If $\{X_1, X_2\}$ exhibit strong multicollinearity or potential causal links in the training data, the SHAP method based on the TU setting (which assumes that each feature independently contributes marginal increments to the model output) may result in misattribution or redundant measurement of $\{X_1, X_2\}$'s individual contributions. Conversely, incorporating more refined conditional distributions or causal structures could address such issues but imposes higher demands on data collection and modeling, while also making the explanation process significantly more complex.

This raises a core question: in practical applications, how can one balance "operability" and "explanatory precision"? For scenarios with a small number of features and relatively reasonable independence assumptions, marginalization-based explanation methods have clear advantages. However, when faced with complex dependency structures or genuine causal mechanisms, relying solely on simple marginal settings to achieve Shapley allocation becomes inadequate. In such cases, more flexible and sophisticated models are required to capture nonlinear interactions among features and deliver more credible explanations.

Beyond Marginal Contributions: In these situations, the contributions of certain players (or features) are no longer purely "marginal increments" but include inseparable synergistic effects or nonlinear interactions. For example, the inclusion of player X_1 might trigger cooperative changes or nonlinear benefits with X_2, exceeding the scope of "marginal contribution" as defined by Shapley Value. This limitation undermines the explanatory capability of existing methods.

To mathematically adapt to such scenarios, adjustments to the axiomatic framework may be necessary (e.g., considering conditional distributions or embedding causal assumptions). Alternatively, more advanced correction mechanisms, such as conditional Shapley Value, can be introduced to account for the effects of non-marginal contributions.

2.2 Both TU Setting and Other Settings Work in XAI

The SHAP methodologies mentioned earlier succeed in applying Shapley Value to complex systems through local approximations and substitutions that adhere to the axioms outlined previously. Another approach, however, focuses on improving Shapley Value mathematically, paving the way for its integration into the next generation of XAI systems equipped with precise descriptive capabilities.

2.3 Significant Advantages in Simple Cases, but Challenges Persist with Dependency and Real Causality

While the SHAP methodologies demonstrate significant advantages in simple cases, they face inherent uncertainties and biases when applied to scenarios with strongly dependent features or genuine causal relationships. Treating unselected features as random or independent in these contexts can lead to inaccuracies.

Driven by the demands of XAI scenarios, there is a growing need to develop a more precise version of Shapley Value capable of capturing causal relationships. This advancement would enable Shapley Value to better adapt to the challenges posed by dependency structures and causal mechanisms in complex systems.

3 Do-Shapley

3.1 Extending Shapley Value to Account for Dependency and Causality

The evolution of Shapley Value in cooperative game theory and explainable artificial intelligence (XAI) reflects a progression from emphasizing external coalitions and productive components to integrating causal inference frameworks.

Conditional Shapley Value: One of the earliest extensions, Conditional Shapley Value, introduces conditional distributions when calculating feature contributions. Unlike the classical Shapley Value, which assumes feature independence, Conditional Shapley Value addresses the problem of misleading interpretations caused by statistical dependencies among features. Its advantage lies in more accurately reflecting the roles of features in real-world data. However, Conditional Shapley Value presents challenges in practical application:

- **High Data Requirements:** Accurately estimating conditional distributions requires a large number of data samples. As the number of features increases, the computation and data demands grow exponentially.
- **Estimation Errors:** Conditional distributions are often approximated using limited samples in practice. Inaccurate estimates can lead to biased Conditional Shapley Value calculations.
- **High-Dimensional Problems:** As the number of features increases, conditional distributions become increasingly complex. The curse of dimensionality renders Conditional Shapley Value computation infeasible in such cases.

Early research on Conditional Shapley Value and its axiomatization includes works by von Neumann and Morgenstern (1944) and Maschler (1992).

Limitations of Conditional Shapley Value: While Conditional Shapley Value handles statistical dependencies, it is fundamentally correlation-based and lacks the ability to distinguish causal relationships. In certain applications, such as model interpretability or economic analysis, this limitation can lead to misjudgments about feature importance, as statistical correlation may obscure true causal relationships.

Causal Shapley Value: With the maturation of causal inference methods (e.g., Causal Shapley Value), researchers increasingly favor directly modeling causal relationships instead of relying solely on correlations. This shift marks a gradual decline in the prominence of Conditional Shapley Value. Causal inference, deeply rooted in philosophical inquiry, has become a mainstream branch of contemporary knowledge science.

As modeling tools like causal graphs have matured, researchers have integrated them with Shapley Value to more accurately identify the true contributions of features to model outputs and to differentiate between direct and indirect effects. Specifically, Shapley Value is introduced as the final step in causal inference to quantify causal effects, giving rise to **Causal Shapley Value**.

Causal Inference Workflow: Causal Shapley Value should not be seen as merely a mathematical modification but should be understood as part of the complete causal inference workflow, which typically includes:

1. Constructing a causal graph,
2. Identifying causal effects,
3. Estimating causal effects.

During the causal effect estimation phase, Shapley Value's principle of fair allocation is applied within the causal framework. By distinguishing between direct and indirect effects, it provides an accurate allocation scheme for each feature's contribution to "post-intervention output changes." In other words, Causal Shapley Value does not independently rely on mathematical formulas for causal contribution allocation. Instead, it integrates Shapley Value into the final effect quantification phase after the completion of causal graph construction and causal effect identification, ensuring that the allocation results are consistent, theoretically grounded, and traceable.

Significance of Causal Shapley Value: The emergence of Causal Shapley Value represents a key evolution from Conditional Shapley Value. By retaining Shapley Value's core interpretability and game-theoretic stability, while incorporating the strengths of causal inference, Causal Shapley Value addresses challenges posed by complex dependency structures and multiple effects. This advancement broadens the applicability of causal inference in machine learning and decision science, opening up new frontiers for its integration into high-impact applications.

References:

- Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- Jung, Yonghan, Shiva Prasad Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum and Elias Bareinboim. "On Measuring Causal Contributions via do-interventions." International Conference on Machine Learning (2022).
- Physics Evolution and Fusion: Bridging Traditional and Wolfram's Computational Theories into a New Era (Author's own paper)

3.2 Why Do-Shapley Could Be Considered the Final Version of Shapley Value

The initial versions of Causal Shapley Value retained the excellent interpretability of Shapley Value while providing a new perspective for distinguishing correlation from causality in high-dimensional and nonlinear machine learning models. However, the theoretical framework of classical Causal Shapley Value mainly targeted partially accessible models or idealized Markovian causal graphs, leaving challenges regarding theoretical completeness. To further refine methods for quantifying causal contributions, researchers proposed new approaches such as ICC (Intrinsic Causal Contribution) and Do-Shapley.

Intrinsic Causal Contribution (ICC): The ICC method captures the direct contribution of features in a causal graph through structure-based interventions. It performs well in cases where the causal graph is Markovian and the structural functions are invertible, providing a more granular view of feature contributions. However, ICC lacks axiomatic support and cannot fully distinguish between the effects of direct and indirect paths on the target variable.

Do-Shapley: In contrast, Do-Shapley introduces a rigorous causal axiomatic framework based on do-interventions. It explicitly satisfies the classical game-theoretic properties of Shapley Value, such as completeness, causal symmetry, causal irrelevance, and causal approximation. Do-Shapley can compute causal contributions in semi-Markovian causal graphs, distinguish between direct and indirect contributions of features, and does not require complete access to the model outputs. This establishes a closer theoretical connection between causal graph modeling and Shapley Value, further improving the accuracy and operability of feature causal contribution quantification.

Key Properties of Causal Shapley Values: In the context of Causal Shapley Values, the following four properties are particularly relevant (using English for compatibility with international literature):

Version	Assignment	Symmetry	Irrelevance	Approximation
Basic Causal Shapley	✓	*maybe*	*maybe*	*notsupporting*
ICC (Intrinsic Causal Contribution)	✓	*maybe*	*maybe*	*maybe*
Do-Shapley	✓	✓	✓	✓

- **Assignment:** Whether each variable can be assigned a clear causal contribution value.
- **Symmetry:** Whether variables with equal contributions are treated fairly.
- **Irrelevance:** Whether irrelevant variables are excluded.
- **Approximation:** Whether reasonable approximations can be provided under imperfect information.

Advantages of Do-Shapley: Do-Shapley is currently the most comprehensive and theoretically grounded method, satisfying all four properties. These properties can be seen as an extension of the classical axiomatic framework of Shapley Value, tailored to causal inference and practical applications:

- **Preserving Core Interpretability and Fairness:** The properties of Symmetry and Irrelevance represent causal adaptations of the classical axioms of "symmetry" and "null player." They ensure intuitive and fair allocation in a causal context.
- **Enhancing Feasibility and Practicality for Causal Inference:** The properties of Assignment and Approximation allow causal contribution methods to be applied in real-world scenarios with imperfect information, complementing the idealized nature of classical Shapley axioms.
- **Providing a New Criterion for Method Selection:** When a causal Shapley method satisfies these four properties, it often demonstrates strong theoretical validity and practical value (e.g., Do-Shapley). This provides a clear benchmark for evaluating candidate methods and encourages further refinement to meet these critical properties.

Applications in Complex Causal Scenarios: In fields such as healthcare, economics, social science, and environmental science, causal inference is a core research focus. Fair, interpretable, and feasible causal contribution allocation often presents a critical challenge. These four properties provide a unified framework for the development of Causal Shapley methods, enabling them to remain rigorous even in complex application scenarios.

Conclusion: In summary, Do-Shapley stands out among various versions of Causal Shapley Value because it fulfills the four key properties: Assignment, Symmetry, Irrelevance, and Approximation. These properties represent a natural extension of the classical axiomatic framework of Shapley Value in the context of causal inference. Do-Shapley combines mathematical rigor with practical feasibility, offering promising prospects for both theoretical and applied research.

Uniqueness of Do-Shapley: Notably, Do-Shapley is the only method that satisfies all four properties. Any causal allocation method meeting these properties can be considered a form of Do-Shapley, as only the combination of Shapley's incremental calculation with causal interventions can fulfill these requirements. Although the most common implementation is value-function-based Do-Shapley, other implementations (e.g., those based on different effect decomposition approaches) may also exist, demonstrating the theoretical uniqueness and practical flexibility of Do-Shapley.

References:

- Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.
- *On Measuring Causal Contributions via Do-Shapley* (Author's own paper).

3.3 How Causality Works and Why It Is Precise: From Classical Experimental Data to Observational Data in XAI

The Workflow of Do-Shapley: Do-Shapley is implemented through the following workflow:

1. Creating the Causal Graph

In this phase, researchers rely on domain knowledge and theoretical assumptions to identify key variables and their interactions, constructing a causal structure model. Common methods include Structural Equation Modeling (SEM) and Directed Acyclic Graphs (DAGs). This process is critical, as the accuracy of the causal graph directly affects the identification and estimation of causal effects.

2. Causal Effect Identification

Once the causal graph is established, the next task is to determine whether the target intervention effect (e.g., $P(Y|do(X))$) can be identified from observational data. This involves the following steps:

1. **Determining the Intervention Method:** Decide whether to perform a **do-intervention** or infer the intervention effect from **identifiable observation**:
   - **Do-intervention:** External manipulation directly fixes the value of a variable, severing its connection with its causal parents.
   - **Identifiable Observation:** Use causal inference rules (e.g., the backdoor criterion) to transform observational distributions into intervention distributions.
2. **Handling C-Components:** Identify which variables belong to C-components (i.e., groups of variables connected by bidirectional edges). These variables may be affected by latent confounders, making direct identification of intervention effects infeasible. C-components need to be decomposed or removed to ensure model identifiability.
3. **Running Causal Effect Identification Algorithms:** After addressing C-components, causal effect identification algorithms (e.g., Pearl's ID algorithm or its extensions) can be used to determine whether the intervention effect is uniquely identifiable from observational data. This process

systematically decomposes causal paths and computes intervention effects based on the causal graph's structure.

Through this process, the target causal effect's identifiability is ensured while mitigating confounding paths and handling unobservable influences.

3. Causal Effect Estimation

Once the causal effect has been successfully identified, the next step is to estimate and quantify it. The primary task in this phase is to calculate causal contributions and allocate them to individual features (or variables), providing a causal explanation of feature importance. However, directly computing Causal Shapley Values in practice is infeasible because the definition requires enumerating the intervention effects $E[Y|do(S)]$ for all subsets $S \subseteq V$. The computational complexity grows exponentially with the number of features ($O(2^n)$).

To address this challenge, the following three efficient and robust estimation methods are proposed:

- **IPW (Inverse Probability Weighting):** Adjusts sample weights to approximate intervention effects by simulating the distribution of variables after intervention through reweighted observational data.
- **REG (Regression-Based Estimation):** Uses regression models to estimate intervention effects by fitting functional relationships between variables and directly predicting changes in the target variable post-intervention.
- **DML (Double/Debiased Machine Learning):** Combines the strengths of IPW and REG by employing machine learning techniques to debias and correct model errors, offering stronger robustness and convergence, especially in scenarios with model misspecification or noise.

These methods not only meet the need for precise characterization of feature importance but also maintain fidelity to causal relationships. By combining IPW, REG, and DML, the computational efficiency is optimized, reducing complexity in high-dimensional scenarios.

The Role of the Backdoor Criterion

Although the definition of Do-Shapley relies on the intervention distribution $P(Y|do(S))$, the backdoor criterion allows researchers to infer intervention effects from observational data. By controlling for confounding variables Z, the observational distribution $P(Y, X, Z)$ can be transformed into the intervention distribution $P(Y|do(X))$:

$$P(Y|do(X)) = \sum_{Z} P(Y|X, Z)P(Z).$$

This method enables causal inference without performing actual intervention experiments, instead using observational data to simulate intervention effects.

This is particularly important in XAI, which often depends on generated data rather than experimental designs. By leveraging the backdoor criterion, Do-Shapley can quantify the causal contribution of features to model predictions, offering reliable causal explanations for complex, high-dimensional models.

Reference: Tian, Jin and Judea Pearl. *On the Identification of Causal Effects*. (2015).

4 Demo: From Causal Graph to a Cleaner, More Readable Causal Graph via Do-Shapley Pruning

4.1 Scenario Description

Suppose we have a causal graph annotated with Do-Shapley values, describing how an individual's study behavior (**Study**) affects their exam score (**Exam Score**) through multiple mediating factors.

$$\text{Study } (0.5) \rightarrow \text{Sleep } (0.3) \rightarrow \text{Exam Score } (1.0)$$

$$\text{Study } (0.5) \rightarrow \text{Stress } (0.1) \rightarrow \text{Exam Score } (1.0)$$

$$\text{Study } (0.5) \rightarrow \text{Coffee } (0.05)$$

$$\text{Coffee } (0.05) \rightarrow \text{Sleep } (0.3)$$

$$\text{Coffee } (0.05) \rightarrow \text{Stress } (0.1)$$

Causal Contributions:

- **Study:** 0.5 (Directly impacts the exam score and also indirectly influences it through other variables).
- **Sleep:** 0.3 (Indirectly improves the exam score by enhancing the individual's condition).
- **Stress:** 0.1 (High stress reduces the exam score).
- **Coffee:** 0.05 (Has a minimal impact on the exam score, only indirectly influencing it).

4.2 Pruning the Causal Graph Using Do-Shapley Values

Based on the Do-Shapley value assessment of causal contributions, we can prune edges or nodes that have a negligible causal effect on the target variable (**Exam Score**).

Pruning Rules:

- Set a contribution threshold (e.g., 0.1).
- Remove edges or nodes with causal contributions below the threshold.

Pruned Causal Graph:

After pruning, the simplified causal graph retains only the significant paths:

$$\text{Study } (0.5) \rightarrow \text{Sleep } (0.3) \rightarrow \text{Exam Score } (1.0)$$

$$\text{Study } (0.5) \rightarrow \text{Stress } (0.1) \rightarrow \text{Exam Score } (1.0)$$

Result: The pruned graph is cleaner and more readable, focusing only on the significant contributions to the exam score. Paths or nodes with negligible influence (e.g., Coffee with 0.05 contribution) have been removed, ensuring the graph emphasizes the most critical relationships.

4.3 Comparison with Traditional Causal Graph Pruning

The pruning operation demonstrated here is a practical application of the desirable properties of Do-Shapley values. By quantifying the causal contribution of each node to the target variable, we can mathematically and rigorously remove paths and nodes with negligible causal impact, thereby optimizing the structure of the causal graph. Shapley values clearly illustrate the relative importance of each path, facilitating a quantitative analysis of both direct and indirect causal effects.

Compared to traditional pruning methods, causal graphs with Do-Shapley values retain traces of minor factors even after pruning. Instead of outright removing such nodes, Do-Shapley can annotate hidden or minor factors, showing how much they still influence intermediate variables like Stress or Sleep, even if their overall contribution to the target variable (e.g., Exam Score) is small.

Preserving Scientific Rigor: This approach explicitly marks the causal graph as a pruned version, ensuring that while the graph is simplified, it retains scientific rigor. For instance, traditional pruning methods might directly delete the Coffee node, whereas Do-Shapley provides annotations to indicate that although Coffee has been pruned due to its minimal contribution, it still has a measurable influence on Stress and Sleep.

Scalability to Large-Scale Causal Graphs: While the example provided here is a simplified, small-scale causal graph, real-world applications often involve much larger causal graphs, such as those used in explainable AI (XAI) to interpret the behavior of deep learning models. For example, Harvard University's GraphXAI project offers a dataset generator called ShapeGGen, capable of creating graph datasets with thousands of nodes and edges to evaluate

the explainability of graph neural networks (GNNs). These large-scale causal graphs help researchers identify which features critically influence decisions and prune redundant features, thereby improving model interpretability and computational efficiency.

Benefits of Pruning at Scale:

- **Reduced Computational Complexity:** Pruning redundant causal paths significantly reduces the computational burden of analyzing large-scale graphs, enabling efficient engineering-level analysis.
- **Improved Visualization and Interpretability:** Pruned causal graphs serve as intuitive tools for visualizing causal structures. Experts and non-experts alike can better understand the causal logic of models, avoiding distractions from insignificant paths. This has the potential to transform how we interpret XAI and fine-tune expert systems.

Implications for Trust and Practical Applications: This method not only enhances the transparency and interpretability of complex systems but also lays a solid foundation for building trust in AI technologies and expanding their practical applications. By offering cleaner, more comprehensible causal graphs, Do-Shapley enables researchers and practitioners to bridge the gap between model complexity and human understanding.

Reference: Harvard University GraphXAI Project: (https://zitniklab.hms.harvard.edu/projects/GraphXAI/

4.4 Conclusion

This pruning process, guided by Do-Shapley values, improves the interpretability of causal graphs by eliminating low-contribution paths and nodes. It ensures that the resulting graph provides a clear and concise representation of the most influential factors affecting the target variable, enhancing both usability and readability in XAI applications.

5 Summary and Extension

- **Shapley Value** has a well-established position in XAI (Explainable Artificial Intelligence). The SHAP family effectively mitigates computational challenges in large-scale scenarios.
- **Causal Shapley** provides more precise measurements for problems with complex dependencies and genuine causal interactions.
- **Do-Shapley** satisfies key properties of causal axiomatization, enabling fine-grained estimation of feature causal contributions even in partially observable or identifiable data scenarios.

- Utilizing Do-Shapley for causal graph pruning can yield more interpretable and readable structures, enhancing the understanding of complex models.
- Future research could focus on optimizing estimation algorithms and developing a complete, systematic process for visualization, further promoting real-world applications in the industry.

6 Comparison of SHAP-Approximations and the Estimation Methods (REG, IPW, DML) of Do-Shapley

6.1 Overview of Methods

This section compares SHAP-approximations with the estimation methods of Do-Shapley, including REG (Regression-Based Estimation), IPW (Inverse Probability Weighting), and DML (Double/Debiased Machine Learning). Due to time constraints, detailed results are not yet presented. However, the structure below provides placeholders for future data.

6.2 Comparison Table

Table 2: Comparison of SHAP-Approximations and Do-Shapley Estimation Methods

Method	**Computation Efficiency**	**Scalability**	**Robustness**	**Causal Precision**
SHAP (Kernel)	Moderate	High	Moderate	Low
SHAP (Tree)	High	Very High	Moderate	Low
REG (Do-Shapley)	High	Moderate	High	High
IPW (Do-Shapley)	Moderate	Moderate	Moderate	High
DML (Do-Shapley)	High	High	Very High	High

6.3 Test Results (Placeholder)

Future experiments will evaluate the methods in the following dimensions:

- **Computation Efficiency:** Measured by the average runtime under different feature dimensions and sample sizes.
- **Scalability:** Assessed by the performance when scaling to high-dimensional data or large datasets.
- **Robustness:** Evaluated under varying levels of noise, missing data, or model misspecification.
- **Causal Precision:** Assessed by the accuracy in estimating causal contributions in synthetic and real-world datasets.