EXPLANATIONS OF BLACK-BOX MODELS BASED ON DIRECTIONAL FEATURE INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

As machine learning algorithms are deployed ubiquitously to a variety of domains, it is imperative to make these often black-box models transparent. Several recent works explain black-box models by capturing the most influential features for prediction per instance; such explanation methods are univariate, as they characterize importance per feature. We extend univariate explanation to a higher-order; this enhances explainability, as bivariate methods can capture feature interactions in black-box models, represented as a directed graph. Analyzing this graph enables us to discover groups of features that are equally important (i.e., interchangeable), while the notion of directionality allows us to identify the most influential features. We apply our bivariate method on Shapley value explanations, and experimentally demonstrate the ability of directional explanations to discover feature interactions. Using post-hoc accuracy, we show the superiority of our method against state-of-the-art on CIFAR10, IMDB, Census, Divorce, Drug, and gene data.

1 INTRODUCTION

The ability to interpret and understand the reasoning behind black box decision-making increases user trust; it provides insights into how a model is working and, as a consequence, how a model can be improved. This has led to a large body of work on the development of explanation methods (Ribeiro et al., 2016; Chen et al., 2018; Yoon et al., 2018; Lundberg & Lee, 2017) applied to black-box models. Intuitively, such methods aim to explain black-box behavior by understanding how individual features influence prediction outcomes. Recently, Covert et al. Covert et al. (2020a) proposed a unifying mathematical framework capturing a broad array of explainability techniques, termed *Removal-based Explanation* methods. Nevertheless, the overwhelming majority of existing literature on explainability has a significant drawback: present methods only provide univariate explanations and, as a result, they do not take into account *feature interactions*. This is problematic precisely because many black box models, such as deep neural networks, perform well by creating complex structures and combining features in their latent layers. To address this, recent methods have been proposed to learn the interaction between features (Sundararajan et al., 2020a; Maas et al., 2011). Their definition of interaction assumes features affect each other symmetrically; however, in many real-world applications, feature interactions may be asymmetrical. We also observe this experimentally (Fig. 1), and argue of the importance of developing black-box interactions that not only capture interactions, but also incorporate asymmetry. Overall, we make the following contributions:

- We propose a method to extend any given univariate removal-based explanation to a bivariate explanation model that can capture asymmetrical feature interactions, represented as a directed graph. Our method is general, and can be applied to a broad array of univariate removal-based explanations, as defined by Covert et al. (2020a).
- We show that analyzing this graph gives a semantically-rich interpretation of black boxes. In particular, beyond the ability to identify most influential features, the graph can identify *directionally redundant* features, i.e., features whose presence negates the influence of other features, as well as *mutually redundant* features, i.e., features that are interchangeable. These two concepts cannot be captured by neither univariate nor symmetric bivariate explanations in existing literature.
- We systematize the analysis of the directed explanation graph, providing both formal definitions of the aforementioned notions as well as algorithms for scrutinizing and explaining black-box model behavior. We also provide theoretical justification for these definitions in the context of SHAP, the Shapley value explanation map introduced by Lundberg & Lee (2017).

- Finally, extensive experiments on MNIST, CIFAR 10, IMDB, Census, Divorce, Drug, and gene expression data show that our explanation graph outperforms prior symmetrical interaction explainers as well as univariate explainers with respect to post-hoc accuracy, AUC, and time.

2 Related work

Many methods have been proposed for explaining black box models (Guidotti et al., 2018). For instance, LIME (Ribeiro et al., 2016) explains the prediction of a model by learning a linear model locally around a sample, through which it quantifies feature influences on the prediction. SHAP (Lundberg & Lee, 2017) learns an influence score for each feature based on the Shapley value (Shapley, 2016). L2X (Chen et al., 2018) learns a set of most influential features per sample based on the mutual information between features and labels. Recently, Covert et al. (2020a) unified many of such explanation methods under a single framework; we present this in detail in Sec. 3.

However, all these methods only capture the univariate influences of each feature; i.e., they do not explain feature interactions. Here, we go from univariate to multivariate explanations, enabling discovery of feature interactions. Discovering feature interactions has drawn recent interest in machine learning (Bondell & Reich, 2008; Chormunge & Jena, 2018; Zeng & Figueiredo, 2014; Janizek et al., 2021; Zhang et al., 2021; Tsang et al., 2018; 2020). Tsang et al. (2017) proposed a framework to detect statistical interactions in a feed-forward neural network by directly interpreting the weights of the model. Cui et al. (2019) proposed a non-parametric probabilistic method to detect global interactions. However, most such methods study feature interactions globally (for all instances). In contrast, our work detects interactions per individual instance. The work more related to explainability of a black box via feature interaction is Shapley interaction. Grabisch & Roubens (1999) proposed a Shapley interaction value to explore the interaction between features rather than feature influence. Lundberg et al. (2018) and Sundararajan et al. (2020a) applied Shapley interaction value in order to explain black box predictions. Instance-wise Feature Grouping (Masoomi et al., 2020) explored the effect of feature interaction by allocating features in a different group based on the similarity of their contribution to the prediction task. These methods assume a symmetrical interaction between features; in contrast, our method provides instance-wise explanations that can capture asymmetrical (directional) interactions.

Another type of explainers are Graph Neural Network (GNN) explainers (Yuan et al., 2021; Duval & Malliaros, 2021; Luo et al., 2020; Ying et al., 2019). Such explainers only work for GNN as the black box needs access to a prior graph as its input. In contrast, our method allows the black box to be any type of function (e.g., CNN, RNN, Random Forest). Our method works with a generic neural network or a graph neural network and does not assume access to a graph structure, rather we learn the feature interactions directly from the data. Causal methods provide explanations through feature influence by utilizing knowledge about causal dependencies. Frye et al. (2020) generalized the Shapley-value framework to enable the incorporation of causality. In particular, they provided a new formulation for computing Shapley value when a partial causal understanding of data is accessible, which they called Asymmetric Shapley values. Wang et al. (2021) extend this idea to incorporate the entire casual graph to reason about the feature influence on the output prediction. Causal methods rely on prior access to casual relationships; in contrast, our method learns the asymmetrical interaction between features rather than the casual dependencies.

3 Background

In general, explainability methods aim to discover the reason why a black box model makes certain predictions. In the local interpretability setting (Chen et al., 2018; Lundberg & Lee, 2017; Ribeiro et al., 2016; Sundararajan et al., 2017), which is our main focus, explanations aim to interpret predictions made by the model on an individual sample basis. Typically, this is done by attributing the prediction to sample features that were most influential on the model's output. The latter is discovered through some form of input perturbation (Lundberg & Lee, 2017; Zeiler & Fergus, 2014). In a recent paper, Covert et al. (2020a) proposed a unifying framework for a variety of different explainability methods. As our work generalizes this framework, we formally describe the removal-based explanation framework by Covert et al. below.

Univariate Removal-Based Explanations. The unifying framework of Covert et al. identifies three stages in a removal-based explanation method. The first, *feature-removal*, defines how the method perturbs input samples by removing features; the second, termed *model-behavior*, captures the effect that this feature removal has on the black-box model predictions; finally, the *summary* stage abstracts the impact of feature subset selection to a numerical score for each feature, capturing the overall influence of the feature in the output.

Formally, a black-box model is a function $f : \mathcal{X} \to \mathcal{Y}$, mapping input features $x \in \mathcal{X} \subseteq \mathbb{R}^d$ to labels $y \in \mathcal{Y}$. Let $D \equiv \{1, \ldots, d\}$ be the coordinates of the feature space. Given an input $x \in \mathcal{X}$ and a subset of features $S \subseteq D$, we denote by $x_S = [x_i]_{i \in S} \in \mathbb{R}^d$ the projection of x to the features in S. In the local interpretability setting, we are given a black-box model f, an input $x \in X$ and in some methods the additional ground truth label $y \in \mathcal{Y}$, and wish to interpret the output $f(x)$ produced by the model. The three stages of the removal based model are defined by a triplet of functions (F, u, E), which we define below.

First, the feature removal stage is defined by a *subset function*

$$F : \mathcal{X} \times P(D) \to \mathcal{Y}, \tag{1}$$

where $P(D) = 2^D$ is the power set of D. Given an input $x \in \mathcal{X}$, and a set of features $S \subseteq D$, the map $F(x, S)$ indicates the label generated by the model when feature subset S is given. For example, several interpretability methods (Yoon et al., 2018; Chen et al., 2018) set $F(x, S) = f([\mathbf{0}; x_S])$, i.e., replace the "removed" coordinates with zero. Other methods (Lundberg et al., 2020; Covert et al., 2020c) remove features by marginalizing them out using their conditional distribution $p(X_{\bar{S}}|X_S = x_S)$, where $\bar{S} = D \setminus S$, i.e., $F(x, S) = \mathbb{E}[f(X)|X_S = x_s]$.

Having access to the subset function F, the model behavior stage defines a *utility* function

$$u : P(D) \to \mathbb{R} \tag{2}$$

quantifying the utility of a subset of features $S \subseteq D$. For instance, some methods (Covert et al., 2020c; Schwab & Karlen, 2019) calculate the prediction loss between the true label y for an input x, using a loss function ℓ, i.e., $u(S) = -\ell(F(x, S), y)$. Other methods (Chen et al., 2018; Yoon et al., 2018) compute the expected loss for a given input x using the label's conditional distribution, i.e., $u(S) = -\mathbb{E}_{p(Y|X=x)}[\ell(F(x, S), Y)]$.

The utility function can be difficult to interpret due to the exponential number of possible feature subsets. This is addressed by the summary stage as follows. Let $\mathcal{U} = \{u : P(D) \to \mathbb{R}\}$ be the set of all possible utility functions. An *explanation map* is a function

$$E : \mathcal{U} \to \mathbb{R}^d \tag{3}$$

mapping a utility function to a vector of scores, one per feature in D. These scores succinctly summarize each feature's value and are the final explanations produced by the removal-based explainability algorithm (F, u, E). For instance, some methods (Zeiler & Fergus, 2014; Petsiuk et al.; Schwab & Karlen, 2019) define $E(u)_i = u(D) - u(D \setminus \{i\})$, or $E(u)_i = u(\{i\}) - u(\emptyset)$ (Guyon & Elisseeff, 2003). Some methods learn $E(u)$ by solving an optimization (Chen et al., 2018; Ribeiro et al., 2016; Yoon et al., 2018). For instance L2X defines $E(u) = \operatorname*{argmax}_{S:|S|=k} u(S)$, for a given k.

The Shapley Value Explanation Map. In our experiments, we focus on Shapley value explanation maps. Shapley (Shapley, 2016) introduced the Shapley value in coalition/cooperative games as a means to compute "payouts", i.e., ways to distribute the value of a coalition to its constituent members. Treating features as players in such a coalition, the Shapley value has been used by many research works on explainability to compute feature influence (Datta et al., 2016; Lundberg et al., 2020; Covert et al., 2020b). For instance, Lundberg & Lee (2017) proposed SHAP as a unified measure of feature influence which uses Shapley value for summarization. They also showed that explainers such as LIME , DeepLIFT (Shrikumar et al., 2017), LRP (Bach et al., 2015), QII (Datta et al., 2016) can all be described by SHAP under different utility functions. Formally, given a utility function u, the SHAP explanation map $E(u)$ has coordinates:

$$E(u)_i = \textstyle\sum_{S \subseteq D \setminus \{i\}} \frac{|S|!\,(d-|S|-1)!}{d!} (u(S \cup \{i\}) - u(S)) \tag{4}$$

where $S \in P(D)$. Direct computation of Eq. (4) is challenging, as the summands grow exponentially as the number of features increases. Štrumbelj & Kononenko (2014) proposed an approximation

with Monte Carlo sampling, known as Shapley sampling values. Later on, Lundberg & Lee (2017) introduced KernelSHAP and DeepSHAP to compute the shapley values using kernel and neural network approaches respectively, and showed that such methods require fewer evaluations of the original model compared to other methods to obtain similar approximation accuracy.

4 From Univariate to Multi-Variate Explanations

Univariate removal-based explanation methods presented in the previous section share a similar limitation: they do not explain feature interactions in the black box model. Given a *univariate* explanation map $E : \mathcal{U} \to \mathbb{R}^d$, we propose a method to extend E to its *Bivariate* explanation map which discovers feature interactions. Let $u \in \mathcal{U}$ be the utility function $u : P(D) \to \mathbb{R}$. Given this $u \in \mathcal{U}$, we define the bivariate explanation $E^2 : \mathcal{U} \to \mathbb{R}^{d\times d}$ as a $d \times d$ matrix: the ith column of this matrix is $E(u_i)$, i.e., the univariate explanation applied to utility function u_i, where u_i is defined as:

$$u_i : P(D) \to \mathbb{R} \quad \text{s.t.} \quad \forall S \in P(D),\, u_i(S) = \begin{cases} u, & \text{if } i \in S, \\ 0, & \text{if } i \notin S. \end{cases} \tag{5}$$

This approach can also be directly generalized beyond bivariate to *multivariate* explanation maps (see App E). Intuitively, the utility function u_i is a restriction of u to sets in which when i is an included feature. As a result, $E(u_i) \in \mathbb{R}^d$ determines a feature's influence *conditioned on the presence of feature* i. We denote the jth element by $E(u_i)_j = E^2(u)_{ji}$ which represents the importance of feature j conditioned on feature i being present.

Bivariate Shapley Explanation Map. As a motivating example, let explanation map $E : \mathcal{U} \to \mathbb{R}^d$ be the Shapley map , defined in Eq. (4). Applying the *Bivariate explanation* extension (5) to the Shapley value we obtain:

$$E^2(u)_{ij} = \sum_{j \in S \subseteq D \setminus \{i\}} \frac{|S|!\ (d-|S|-1)!}{d!} (u(S \cup \{i\}) - u(S)). \tag{6}$$

We provide the derivation of this formula in App D in the supplement. An important feature of the above bivariate explanation is that it is *not symmetric*: in general, $E^2(u)_{ij} \neq E^2(u)_{ji}$. In other words, feature i may influence j differently than how feature j influences i. This is in sharp contrast with other bivariate explanation methods, such as, e.g., interaction Shapley (Grabisch & Roubens, 1999; Sundararajan et al., 2020a), that are symmetric. Hence, $E^2(u)$ is an asymmetric matrix, that can be represented by a weighted directed graph denoted by $\mathcal{G} = (V_\mathcal{G}, E_\mathcal{G}, W_\mathcal{G})$, where weights $W_\mathcal{G}(i, j) = E^2(u)_{ji}$ (see App A for a brief review of graph terminology). We call $\mathcal{G}$ the *directed explanation graph*. The directionality of G/asymmetry of E^2 has important implications for explainability, which we illustrate next with an example.

Illustrative Example (Univariate Shapley vs. Bivariate Shapley). In order to motivate the bivariate explanation map, highlight its difference from the univariate Shapley explanation and illustrate the importance of directionality, we study the directed explanation graph of one sample of "The Large Movie Review Dataset (IMDB) (Maas et al., 2011)". This is a dataset of movie reviews with labels indicating positive or negative sentiment. We used a Recurrent Neural Network (RNN) as the black box and SHAP as the univariate explainer. In particular we used default choices for (F, u, E) defined in SHAP. In particular, given a black box model f and a point x, SHAP chooses $F(x, S) = \mathbb{E}[f(X)|X_S = x_s]$, $u = F(x, S)$, and E to be the explanation map using Shapley value We compute both SHAP, the univariate Shapley explanation E, as well as the directed explanation graph G/bivariate explanation E^2 for the sentence, "The movie is the worst; surprisingly awful". This sentence is predicted to be negative by the black box. Both explanations are shown in Figure 1. We observe the following differences:

The influence of word '*surprisingly*': In Fig. 1(a), we observe that SHAP explanation E identifies 'awful', 'worst', 'surprisingly' as the most influential features. The negative Shapley value for 'Surprisingly' indicates that this feature affects prediction strongly in favor of a positive label (opposite of the black box outcome). However, looking at E^2, we realize this explanation is in fact misleading. The absence of an edge from 'worst' $\to$ 'surprisingly' suggests that *in the presence of 'worst' the word 'surprisingly' has no influence.* Interestingly, the reverse edge does exist; hence, presence of 'surprisingly' does not remove the influence of 'worst', which still influences the classification outcome. This lack of symmetry is informative, and would not be detected from either a univariate or a bivariate but symmetric explanation.

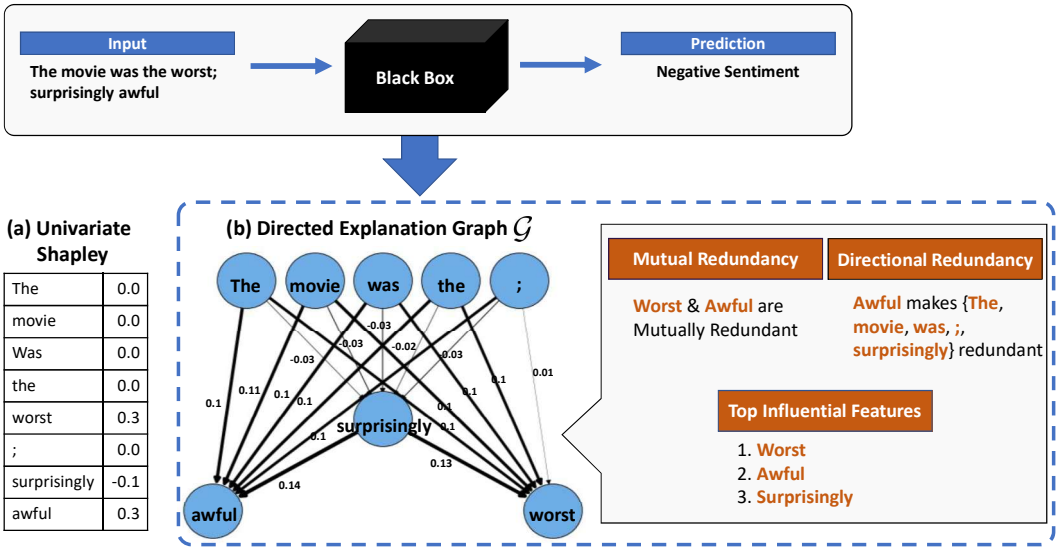

Figure 1: **a)** Represents univariate Shapley value, the values suggest that 'surprisingly', 'worst', 'awful' are the most influential features, however it does not explain the feature interaction **b)** Using Shapley explanation map E, we plot Directed Explanation graph $\mathcal{G}$ using our method. An edge $i \rightarrow j$ represents the conditional influence of word j when word i is present. We can then use the properties of Graph $\mathcal{G}$ to derive notions of Mutual and Directional Redundancy, as well as Influential Features.

'*awful*' vs '*worst*': The univariate explanation E suggests that 'awful' and 'worst' are both influential. However, from the Shapley graph, we observe the absence of an edge from 'awful' $\rightarrow$ 'worst' and vice versa. This indicates that *the presence of either word negates the influence of the other on the classification outcome, making it redundant*. Another way to interpret this is that 'awful' and 'worst' are interchangeable. This is aligned with our understanding that awful and worst have a similar meaning, and is an observation that is not evident from the univariate SHAP E alone.

Least Influential Features: Words like 'The', 'movie' have Shapley value zero. Such words are *sources* in the directed explanation graph G, i.e., have only outgoing edges towards 'awful', 'worst', and 'surprisingly' but no incoming edges at all. This suggests that they are, overall, not influential; more generally, there is a consistency between features that E identifies as non-influential and *sources* in G.

Most Influential Features: The support of E (i.e., words 'awful', 'worst', 'surprisingly') represents the words that have the greatest influence. In graph G, we observe that 'awful' and 'worst' are the *sinks* of the graph (have only incoming edges, and no outgoing edges); even though 'surprisingly' is not a sink, it still has many nodes point into it. This is intuitive: sinks, that have no outgoing edges, indicate that all other words lose their influence when these words are present. 'Surprisingly' is important, but there are still other words that negate its influence (namely, 'awful' and 'worst').

In summary, the above example illustrates how E^2 and G reveal more information than was present in E alone. Although in a more nuanced way, the most and least influential features again stand out; however, in addition to this, features deemed as influential by SHAP can be discovered to be less influential when observed in conjunction with other features; similarly, groups of influential but mutually redundant features may also arise. Both enhance our understanding of how the black box makes predictions, and neither are observable from univariate explanations.

4.1 Analyzing the Directional Explanation Graph

Motivated by the above observations, we turn our attention to means of systematizing the analysis and interpretation of the Directional Explanation graph. In particular, we identify ways to discover the following notions from the graph $\mathcal{G}$: (a) the most influential features for the black box model, taking their interaction into account, and the extent to which they act as *sinks* of $\mathcal{G}$; (b) inferring

redundancies between the features, by formally defining and detecting *Directional Redundancy* (e.g. 'surprisingly' and 'awful') and *Mutual Redundancy* (e.g. 'awful' and 'worst'); and (c) the transitivity properties of mutually redundant features: if mutual redundancy is transient, it implies the existence of entire equivalence classes of features that are all interchangeable. Our goal in this section is to both state such concepts formally, but also describe algorithms through which these concepts can be used to scrutinize and interpret E^2 and $\mathcal{G}$.

Most Influential Features in Graph $\mathcal{G}$. Most influential features in E^2 can be identified as sinks in G. These are easy to identify in polynomial time (as the nodes with no outgoing edges). Nevertheless, a more graduated approach seems warranted, to identify nodes that are 'almost' sinks (like 'surprisingly' in Fig. 3). This can be accomplished by the use of a random-walk inspired harmonic function on $\mathcal{G}$, like the classic PageRank (Page et al., 1999) or HITS (Kleinberg, 1999).[1] These typically correspond to the steady state distribution over a random walk on the graph with random restarts; as such it indeed offers a more graduated version of the notion of "sinkness", as desired. Variants of this, such as personalized PageRank (Page et al., 1999), can also be used, whereby the random restart is to a node sampled from a predetermined distribution over the vertices of $\mathcal{G}$. Setting this to be, e.g., proportional to the magnitude of the univariate Shapley value E can create an interpolation between the univariate map (that captures univariate influence but not directionality) and E^2 (that captured directionality and "sinkness").

Directional Redundancy and Mutual Redundancy. In the example stated above, using a bivariate explanation map enabled us to discover which features are redundant with respect to other features. One of these examples was symmetric (e.g. 'awful' and 'worst') and one was one-sided (e.g. 'awful' makes 'surprisingly' redundant, but not vice-versa). Motivated by this, we define:

Definition 4.1. Given $i, j \in D$, i is *directionally redundant* with respect to feature j if $E^2(u)_{ij} = 0$.

Directionality arises in Def. 4.1 because $E^2(u)$ in general is not symmetric, i.e., $E^2(u)_{ij} \neq E^2(u)_{ji}$. Nevertheless, we can have features i, j that have the same influence on the model (e.g. 'awful' and 'worst' in the example). We formalize this idea to the E^2 explanation through the following definition:

Definition 4.2. Given $i, j \in D$, features i, j are *mutually redundant* if $E^2(u)_{ij} = E^2(u)_{ji} = 0$.

Transitivity of Mutually Redundant Features. Given that mutual redundancy is symmetric, it is natural to ask if it is also transitive: if 'bad' and 'awful' are mutually redundant, and so are 'awful' and 'terrible', would that imply that 'bad' and 'terrible' are also mutually redundant? This behavior is natural, and suggest that *groups of features* may act interchangeably (here, indeed corresponding to variants of 'bad'). Clearly, identifying such groups is important for the interpretability of the model, as it exhibits an invariant behavior under the exchange of such features. We therefore turn our attention to studying the transitivity properties of mutual redundancy. To do this, we define unweighted directed graph $\mathcal{H} = (V_{\mathcal{H}}, E_{\mathcal{H}})$, where $V_{\mathcal{H}} = V_{\mathcal{G}}$ and $E_H = \{(i, j) \in E_{\mathcal{G}} | \, W_{\mathcal{G}}(i, j) = 0\}$. Graph $\mathcal{H}$ captures all the redundancies between any two pair of features and existence of an edge from i to j (i.e., $i \to j$) represents that feature j is directionally redundant with respect to feature i. We denote $\mathcal{H}$ as *Redundancy Graph*. In practice, we may use the relaxed version of redundancy. Given a redundancy threshold γ, we define $\mathcal{H}_\gamma = (V_{\mathcal{H}}, E^\gamma_{\mathcal{H}})$ to be a graph where $V_{\mathcal{H}} = V_{\mathcal{G}}$ and $E^\gamma_{\mathcal{H}} = \{(i, j) \in E_{\mathcal{G}} : |W_{\mathcal{G}}(i, j)| \leq \gamma\}$. Intuitively, if the presence of a feature makes the influence of the other less than the threshold γ, we still declare the latter to be redundant. If mutual redundancy is transitive, Graph $\mathcal{H}$ would also be *transitive*; formally:

Definition 4.3. An unweighted directed graph $\mathcal{H}$ is *transitive* if $(i, j), (j, k) \in E_{\mathcal{H}}$, then $(i, k) \in E_{\mathcal{H}}$.

In other words, a transitive graph comprises a collection of cliques/fully connected graphs, captured by mutual redundancy, along with possible 'appendages' (pointing out) that are due to the non-symmetry of directed redundancy. Not every explanation graph $\mathcal{G}$ leads to a transitive $\mathcal{H}$. In the following theorem however, we prove that the Shapley explanation map E indeed leads to the transitivity of mutual redundancy and, thereby, the graph $\mathcal{H}$:

Theorem 1. *For $i, j, k \in D$, assume that $\max_{j \in S \subseteq N} |u(S \cup \{i\}) - u(S)| \leq \varepsilon_j$ and $\max_{i \in S \subseteq N} |u(S \cup \{k\}) - u(S)| \leq \varepsilon_i$. Then, the following inequalities hold: $|E^2(u)_{ij}| \leq \frac{d!}{2}\varepsilon_j$, $|E^2(u)_{ki}| \leq \frac{1}{2}\varepsilon_i$, and $|E^2(u)_{kj}| \leq \frac{d!}{2}(2\varepsilon_j + \varepsilon_i)$.*

[1] In the context of HITS (Kleinberg, 1999), "almost" sources and sinks correspond to "hubs" and "authorities", respectively.

In short, Theorem 1 states that for a given path $i \to j \to k$ the summation of upper bounds for edges $i \to j$ and $j \to k$ can be used to upper bound the weight of the edge $i \to k$. An immediate implication of this "triangle-inequality"-like theorem is that if all $\epsilon = 0$, as is the case in directed redundancy, the weight of edge $i \to k$ must also be zero. In other words:

Corollary 1.1. *Graph $\mathcal{H}$ for Shapley explanation map with a monotone utility function is transitive.*

Most importantly, Theorem 1 proves something stronger than that. In fact, it allows for 'almost transitivity' of graph $\mathcal{H}^\gamma$ in the case of approximate Shapley value computation: even if we set the threshold γ to a non-zero value, short paths (of length γ/ϵ) will also be guaranteed to be transitive. The proofs for Thm. 1 and Corollary 1.1 are provided in the supplement.

Sources and Sinks in the Redundancy Graph. Beyond identifying classes of mutually redundant features, the redundancy graph $\mathcal{H}$ can also be used to identify (classes of) features under which all remaining features become redundant. This can be accomplished in the following fashion. First, the strongly connected components/classes of mutually redundant features need to be discovered. As a consequence of Thm. 1, such strongly connected components will be cliques if the exact Shapley explanation map is used, or almost cliques if the threshold γ is non-zero. Collapsing these connected components we obtain a DAG, also known as the quotient graph (see App A). The sources of this quotient graph (which may correspond to entire classes) correspond to feature classes that make all other features redundant (possibly through transitivity). Note the distinction between sinks in $\mathcal{G}$ and sources in $\mathcal{H}$, that may be a different set of nodes, and capture a different kind of importance. Again, rather than determining importance simply from the fact that a node in $\mathcal{H}$ is a source, a more graduated approach, using a harmonic function of nodes like Pagerank (over $\mathcal{H}$) could be used.

5 Experiments

In this section, we investigate the ability of Bivariate Shapley for discovering (1) mutual redundancy, (2) directional redundancy, and (3) influential features, from black-box models. Fully evaluating shapley values is computationally expensive; we implement Bivariate Shapley using two different Shapley approximation methods for comparison, Shapley Sampling (BivShap-S) and KernelSHAP (BivShap-K). The algorithms are outlined in App F.1.1. The KernelSHAP approximation significantly reduces computational time (tbl. 1) at the cost of slightly reduced Post-hoc accuracy results (tbl. 3). In our method comparisons, We take 500 test samples from each dataset (less if the test set is smaller than 500) and generate their respective $\mathcal{G}$ and $\mathcal{H}$ graphs. We select $\gamma = 10^{-5}$ for the threshold in converting $\mathcal{G} \to \mathcal{H}$, which generally corresponds to 50% average graph density across the datasets (see App. F.2.5 for details). All source code will be made publicly available.

Data. We evaluate our methods on a 1077 RNAseq COPDGene data (Regan et al., 2010), CIFAR10 (Krizhevsky, 2009) and MNIST (LeCun & Cortes, 2010) image data, IMDB (Maas et al., 2011) movie review text data, and on three tabular UCI datasets (Drug, Divorse, and Census) (Dua & Graff, 2017). We train separate black-box models for each dataset. Image and text datasets, as well as COPDGene and Divorce use neural network architectures; Census and Divorce use tree-based models (XGBoost (Chen & Guestrin, 2016) and Random Forest). Full dataset details and model settings can be found in App. F.1.3.

Competing Methods. We compare our method against both univariate and bivariate, instance-wise, black-box explanation methods. Univariate methods Shapley sampling values (Sh-Sam), KernelSHAP (kSHAP), and L2X are used to identify the top important features, either through a feature ranking or by choosing a subset of features. Second-order methods Shapley Interaction Index (Sh-Int) (Owen, 1972), Shapley-Taylor Index (Sh-Tay) (Sundararajan et al., 2020b), and Shapley Excess (Sh-Exc) (Shoham & Leyton-Brown, 2008) capture symmetric interactions, for which we apply the same PageRank algorithm as Bivariate Shapley to derive a feature ranking. Further details regarding control parameter settings of these competing methods and licensing information are provided in App F.1.

(1) Mutual Redundancy Evaluation. We evaluate the validity of mutually redundant features by evaluating the change in model accuracy after masking redundant features, with post-hoc accuracy results shown in Figure 2. We identify such features as groups of strongly connected nodes in graph $\mathcal{H}$, which we find using a depth-first search on $\mathcal{H}$ using Tarjan's algorithm (Tarjan, 1972). After finding the mutually redundant groups, we test their exchangeability by randomly selecting subsets of features within the group to mask. We evaluate post-hoc accuracy at different levels of

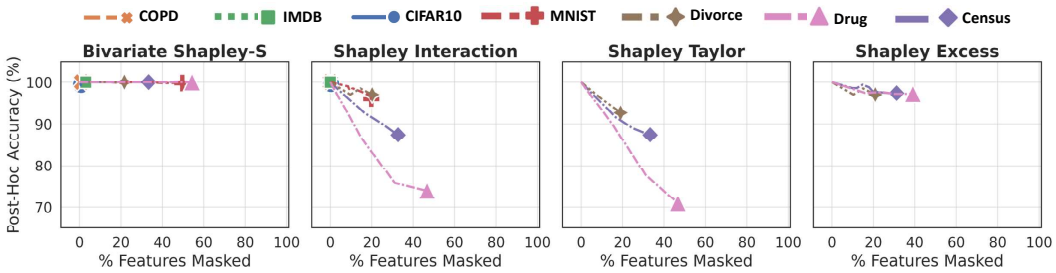

Figure 2: Posthoc Accuracy evaluated on Mutual Redundancy masking derived from graph $\mathcal{H}$. Strongly connected nodes in $\mathcal{H}$ are randomly masked with increasing mask sizes until a single node remains, represented by the final marker for each dataset. Note that we can only run Sh-Tay and Sh-Exc methods on the smaller datasets Divorce, Drug, and Census due to their computational issues.

Time Complexity: Seconds per Sample						
Dataset	#Features	BivShap-S	BivShap-K	Sh-Int	Sh-Tay	Sh-Exc
COPD	1077	5942	36	2877	112900	838200
CIFAR10	255	218	2.5	101	2819	6267
MNIST	196	116	1.5	48	1194	2350
IMDB	$\leq$400	207	1.9	160	1279	1796
Census	12	2.7	0.20	2.6	11.6	5.3
Divorce	54	18.2	0.34	6.5	63.2	93.3
Drug	6	2.3	0.07	1.21	181	0.96

Table 1: Time comparisons of the methods. Results are calculated on the time to produce the interaction matrix (including all features) for a single sample, as measured by seconds per sample.

	PH-Accy		% Feat Masked	
Dataset	Sink Masked	Source Masked	Sink Masked	Source Masked
COPD	99.5	62.7	1.5	98.5
CIFAR10	94.6	15.0	6.2	93.8
MNIST	100.0	13.4	77.7	22.3
IMDB	100.0	54.0	3.5	96.5
Census	100.0	82.0	23.8	76.2
Divorce	100.0	51.5	22.2	77.8
Drug	100.0	48.5	43.5	56.5

Table 2: Posthoc-accy of BivShap-S after masking $\mathcal{H}$-source nodes, representing features with minimal redundancies, and $\mathcal{H}$-sink nodes, representing directionally redundant features.

masking. Masking all but one feature results in the quotient graph $\mathcal{S}$, which is represented by each dataset's marker in Figure 2. Note that these groups can be discovered by other second-order methods by similarly interpreting the resulting feature interactions as an adjacency matrix; we include the results for applying the same algorithm to Sh-Int, Sh-Tay and Sh-Exc. We observe that masking an arbitrary number of mutually redundant features has minimal impact to accuracy under the BivShap method. In contrast, the groups identified by the other methods do not share the same level of feature interchangeability. In terms of finding mutual redundancy, this suggests that incorporating directionality is critical in identifying mutually redundant features; undirected methods do not capture the full context of feature interactions.

(2) Directional Redundancy Evaluation. We validate directional redundancy using post-hoc accuracy as shown in tbl. 2. We identify directionally redundant features as $\mathcal{H}$-sinks and $\mathcal{H}$-sources, which we identify using PageRank. We collapse the strongly connected components in graph $\mathcal{H}$ to its quotient graph $\mathcal{S}$, then apply PageRank to each connected graph in $\mathcal{S}$. The maximum and minimum ranked nodes in each connected graph correspond to the sinks and sources in $\mathcal{S}$, respectively. The condensed sinks and sources are expanded to form $\mathcal{H}$-sinks and $\mathcal{H}$-sources. We then evaluate the relevance of the identified nodes using post-hoc accuracy of masked samples. To validate our claims of directionality, we compare the results of masking $\mathcal{H}$-source nodes and $\mathcal{H}$-sink nodes, where $\mathcal{H}$-sink nodes represent the directionally redundant features. We observe that masking sinks have little effect on accuracy, suggesting that these features contain redundant information given the unmasked source nodes. In contrast, masking source nodes and keeping sinks results in large decreases in accuracy, indicating that prediction-relevant information is lost during masking.

(3) Influential Feature Evaluation. While mutual and directional redundancy can be investigated individually using graph $\mathcal{H}$, we can combine both concepts in a continuous ranking of features based on the steady-state distribution of a random walk on graph $\mathcal{G}$. We add a small $\epsilon \approx 0$ value to the adjacency matrix of $\mathcal{G}$ to ensure that all nodes are connected, then directly apply PageRank. The resulting feature scores are ranked, with high-scoring features representing high importance. We observe the experimental results in Figure 3. BivShap-S consistently performs the best across data sets, including against BivShap-K, which suggests some small accuracy tradeoff for the faster

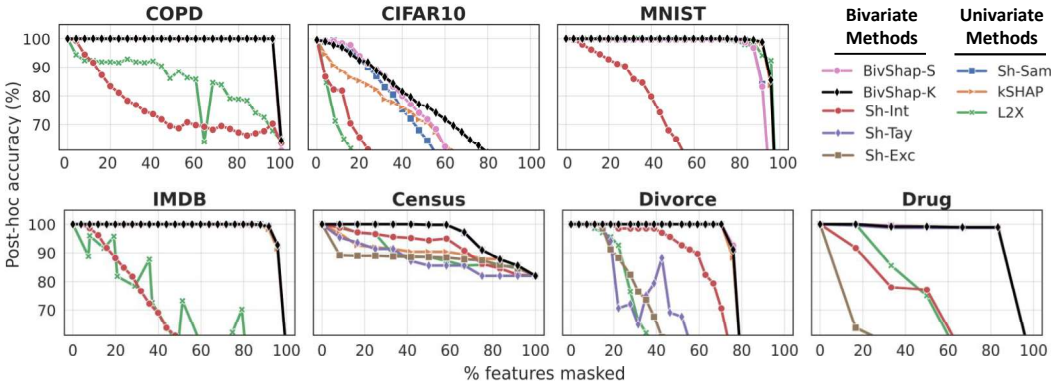

Figure 3: Comparison of explanation methods on a feature removal task. We apply PageRank to graph $\mathcal{G}$ to derive a univariate ranking based on feature redundancy. We compare to other explanation methods by iteratively masking the lowest ranked features. Note that we can only run Sh-Tay and Sh-Exc methods on the smaller datasets Divorce, Drug, and Census due to their computational issues.

approximation. Note that Bivariate Shapley does not explicitly utilize feature importance, but rather each feature's steady-state distribution. However we can also incorporate feature importance information, defined by univariate Shapley, through Personalized PageRank. Personalization increases the steady state probabilities for important features.

Analysis of COPDGene Data. Chronic Obstructive Pulmonary Disease (COPD) is a chronic inflammatory lung disease. The relation between COPD and smoking is well-established; it has been shown that smoking increases the risk of developing lung disease through a variety of ways, such as increasing lung inflammation (Arnson et al., 2010). Here, we investigate the relation between gene expression data and smoking status in COPDGene data. We show the interpretation power of our methods by relating our most influential genes to biological pathways which correspond to smoking. We performed Gene Set Enrichment Analysis (GSEA) using the GenePattern web interface (Reich et al., 2006) on the ranking of influential features, which we generate as follows. We first calculate graph $\mathcal{G}$ locally, as in Alg. 1. We then create the global $\mathcal{G}$ graph for each subgroup, smokers and non-smokers, by averaging the $\mathcal{G}$ adjacency matrix over all samples within each subgroup (Fig. 6). We directly apply the ranking algorithm outlined in Section F.1.1 to obtain subgroup-specific importance scores. We use the list of 1,079 unique gene names with their associated importance score as input into the GenePattern interface. Gene set enrichment for these two groups was calculated using the GSEAPreranked module with 1000 permutations, using the Hallmark (h.all.v7.4.symbols.gmt) and Immunologic gene sets (c7.all.v7.4.symbols.gmt). We observed genetic pathways corresponding to Macrophages as statistically significant at a q-value ≤ 0.05 (the pathway table is in the App., Table. 7). Macrophages are a type of immune cells that can initiate inflammation, and they also involve the detection and destruction of bacteria in the body. The relation between such cells and smoking has been observed in biological domain; many studies have observed that smoking induces changes in immune cell function in COPD patients (Yang & Chen, 2018; Strzelak et al., 2018).

Time Comparisons and Additional Results. Table 1 reports the time in seconds per sample to calculate feature interactions. Note that BivShap-K performs the best among feature interaction methods. In the appendix we also provide a number of illustrative interpretability examples from MNIST, CIFAR10, and IMDB, as well as other additional evaluation results based on AUC.

6 Conclusion

We extend the removal-based explanation to bivariate explanation that captures asymmetrical feature interactions per sample, which we represented as a directed graph. Using our formulation, we define two concepts of redundancy between features denoted as mutual redundancy and directional redundancy. Our theoretical contribution leads to a systematic approach for detecting such redundancies through finding the cliques and sources in graph $\mathcal{H}$. Experiments show the benefit of capturing directional redundancies and the superiority of our method against competing methods.

REFERENCES

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and Sabine Süsstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, November 2012. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2012.120.

Yoav Arnson, Yehuda Shoenfeld, and Howard Amital. Effects of tobacco smoke on immunity, inflammation and autoimmunity. *Journal of autoimmunity*, 34(3):J258–J265, 2010.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Jørgen Bang-Jensen and Gregory Z Gutin. *Digraphs: theory, algorithms and applications*. Springer Science & Business Media, 2008.

Roderick Bloem, Harold N Gabow, and Fabio Somenzi. An algorithm for strongly connected component analysis in n log n symbolic steps. *Formal Methods in System Design*, 28(1):37–56, 2006.

Thomas Bonald, Nathan de Lara, Quentin Lutz, and Bertrand Charpentier. Scikit-network: Graph analysis in python. *Journal of Machine Learning Research*, 21(185):1–6, 2020. URL http://jmlr.org/papers/v21/20-412.html.

Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 883–892. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/chen18j.html.

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, August 2016. doi: 10.1145/2939672.2939785.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014. URL http://arxiv.org/abs/1406.1078.

Smita Chormunge and Sudarson Jena. Correlation based feature selection with clustering for high dimensional data. *Journal of Electrical Systems and Information Technology*, 5(3):542–549, 2018.

Ian Covert, Scott Lundberg, and Su-In Lee. Feature removal is a unifying principle for model explanation methods. *CoRR*, abs/2011.03623, 2020a. URL https://arxiv.org/abs/2011.03623.

Ian Covert, Scott Lundberg, and Su-In Lee. Understanding global feature contributions through additive importance measures. *arXiv preprint arXiv:2004.00668*, 2020b.

Ian Covert, Scott M Lundberg, and Su-In Lee. Understanding global feature contributions with additive importance measures. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17212–17223. Curran Associates, Inc., 2020c. URL https://proceedings.neurips.cc/paper/2020/file/c7bf0b7c1a86d5eb3be2c722cf2cf746-Paper.pdf.

Tianyu Cui, Pekka Marttinen, and Samuel Kaski. Learning global pairwise interactions with bayesian neural networks. *arXiv preprint arXiv:1901.08361*, 2019.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pp. 598–617. IEEE, 2016.

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

Alexandre Duval and Fragkiskos D Malliaros. Graphsvx: Shapley value explanations for graph neural networks. *arXiv preprint arXiv:2104.10482*, 2021.

E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. The Five Factor Model of personality and evaluation of drug consumption risk. June 2015.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1229–1239. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0d770c496aa3da6d2c3f2bd19e7b9d6b-Paper.pdf.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman (eds.), *Proceedings of the 7th Python in Science Conference*, pp. 11 – 15, Pasadena, CA USA, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

Jon M Kleinberg. Hubs, authorities, and communities. *ACM computing surveys (CSUR)*, 31(4es): 5–es, 1999.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles. pp. 9.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *arXiv preprint arXiv:2011.04573*, 2020.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, and Jennifer Dy. Instance-wise feature grouping. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13374–13386. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/9b10a919ddeb07e103dc05ff523afe38-Paper.pdf.

W Meeusen and L Cuyvers. Clique detection in directed graphs: A new algorithm. *Journal of Computational and Applied Mathematics*, 1(3):185–203, 1975.

Mark Newman. *Networks*. Oxford university press, 2018.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5-Part-2):64–79, 1972. URL https://EconPapers.repec.org/RePEc:inm:ormnsc:v:18:y:1972:i:5-part-2:p:64-79.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models.

Elizabeth A. Regan, John E. Hokanson, James R. Murphy, Barry Make, David A. Lynch, Terri H. Beaty, Douglas Curran-Everett, Edwin K. Silverman, and James D. Crapo. Genetic epidemiology of COPD (COPDGene) study design. *COPD*, 7(1):32–43, February 2010. ISSN 1541-2563. doi: 10.3109/15412550903499522.

Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P. Mesirov. GenePattern 2.0. *Nature Genetics*, 38(5):500–501, May 2006. ISSN 1061-4036. doi: 10.1038/ng0506-500.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3ab6be46e1d6b21d59a3c3a0b9d0f6ef-Paper.pdf.

Lloyd S Shapley. *17. A value for n-person games*. Princeton University Press, 2016.

Yoav Shoham and Kevin Leyton-Brown. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press, 2008. doi: 10.1017/CBO9780511811654.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

Agnieszka Strzelak, Aleksandra Ratajczak, Aleksander Adamiec, and Wojciech Feleszko. Tobacco smoke induces and alters immune responses in the lung triggering inflammation, allergy, asthma and other lung diseases: a mechanistic review. *International journal of environmental research and public health*, 15(5):1033, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International Conference on Machine Learning*, pp. 9259–9268. PMLR, 2020a.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The Shapley Taylor Interaction Index. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9259–9268. PMLR, November 2020b.

R. Tarjan. Depth-first search and linear graph algorithms. *SIAM J. Comput.*, 1:146–160, 1972.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977*, 2017.

Michael Tsang, Youbang Sun, Dongxu Ren, and Yan Liu. Can i trust you more? model-agnostic hierarchical explanations. *arXiv preprint arXiv:1812.04801*, 2018.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. *arXiv preprint arXiv:2006.10965*, 2020.

Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 721–729. PMLR, 2021.

David C Yang and Ching-Hsien Chen. Cigarette smoking-mediated macrophage reprogramming: mechanistic insights and therapeutic implications. *Journal of nature and science*, 4(11), 2018.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32:9240, 2019.

Mustafa Kemal Yöntem, Kemal Adem, Tahsin İlhan, and Serhat Kiliçarslan. DIVORCE PREDICTION USING CORRELATION BASED FEATURE SELECTION AND ARTIFICIAL NEURAL NETWORKS. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi*, 9(1):259–273, June 2019.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. *arXiv preprint arXiv:2102.05152*, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Xiangrong Zeng and Mário AT Figueiredo. Solving oscar regularization problems by fast approximate proximal splitting algorithms. *Digital Signal Processing*, 31:124–135, 2014.

Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10877–10886, 2021.