2016 IEEE Symposium on Security and Privacy

Algorithmic Transparency via Quantitative Input Influence:
Theory and Experiments with Learning Systems

Anupam Datta Shayak Sen Yair Zick
Carnegie Mellon University, Pittsburgh, USA
{danupam, shayaks, yairzick}@cmu.edu

***Abstract*—Algorithmic systems that employ machine learning play an increasing role in making substantive decisions in modern society, ranging from online personalization to insurance and credit decisions to predictive policing. But their decision-making processes are often opaque—it is difficult to explain why a certain decision was made. We develop a formal foundation to improve the transparency of such decision-making systems. Specifically, we introduce a family of *Quantitative Input Influence (QII)* measures that capture the degree of influence of inputs on outputs of systems. These measures provide a foundation for the design of transparency reports that accompany system decisions (e.g., explaining a specific credit decision) and for testing tools useful for internal and external oversight (e.g., to detect algorithmic discrimination).**

Distinctively, our *causal* QII measures carefully account for correlated inputs while measuring influence. They support a *general* class of transparency queries and can, in particular, explain decisions about individuals (e.g., a loan decision) and groups (e.g., disparate impact based on gender). Finally, since single inputs may not always have high influence, the QII measures also quantify the *joint influence* of a set of inputs (e.g., age and income) on outcomes (e.g. loan decisions) and the *marginal influence* of individual inputs within such a set (e.g., income). Since a single input may be part of multiple influential sets, the average marginal influence of the input is computed using principled aggregation measures, such as the Shapley value, previously applied to measure influence in voting. Further, since transparency reports could compromise privacy, we explore the transparency-privacy tradeoff and prove that a number of useful transparency reports can be made differentially private with very little addition of noise.

Our empirical validation with standard machine learning algorithms demonstrates that QII measures are a useful transparency mechanism when black box access to the learning system is available. In particular, they provide better explanations than standard associative measures for a host of scenarios that we consider. Further, we show that in the situations we consider, QII is efficiently approximable and can be made differentially private while preserving accuracy.

I. Introduction

Algorithmic decision-making systems that employ machine learning and related statistical methods are ubiquitous. They drive decisions in sectors as diverse as Web services, healthcare, education, insurance, law enforcement and defense [1], [2], [3], [4], [5]. Yet their decision-making processes are often opaque. *Algorithmic transparency* is an emerging research area aimed at explaining decisions made by algorithmic systems.

The call for algorithmic transparency has grown in intensity as public and private sector organizations increasingly use large volumes of personal information and complex data analytics systems for decision-making [6]. Algorithmic transparency provides several benefits. First, it is essential to enable identification of harms, such as discrimination, introduced by algorithmic decision-making (e.g., high interest credit cards targeted to protected groups) and to hold entities in the decision-making chain accountable for such practices. This form of accountability can incentivize entities to adopt appropriate corrective measures. Second, transparency can help detect errors in input data which resulted in an adverse decision (e.g., incorrect information in a user's profile because of which insurance or credit was denied). Such errors can then be corrected. Third, by explaining why an adverse decision was made, it can provide guidance on how to reverse it (e.g., by identifying a specific factor in the credit profile that needs to be improved).

Our Goal. While the importance of algorithmic transparency is recognized, work on computational foundations for this research area has been limited. This paper initiates progress in that direction by focusing on a concrete algorithmic transparency question:

How can we measure the influence of inputs (or features) on decisions made by an algorithmic system about individuals or groups of individuals?

Our goal is to inform the design of transparency reports, which include answers to transparency queries of this form. To be concrete, let us consider a predictive policing system that forecasts future criminal activity based on historical data; individuals high on the list receive visits from the police. An individual who receives a visit from the police may seek a transparency report that provides answers to *personalized transparency queries* about the influence of various inputs (or features), such as race or recent criminal history, on the system's decision. An oversight agency or the public may desire a transparency report that provides answers to *aggregate transparency queries*, such as the influence of sensitive inputs (e.g., gender, race) on the system's decisions concerning the entire population or about systematic differences in decisions

DOI 10.1109/SP.2016.42

among groups of individuals (e.g., discrimination based on race or age). These reports can thus help identify harms and errors in input data, and provide guidance on what input features to work on to modify the decision.

Our Model. We focus on a setting where a *transparency report* is generated with black-box access to the decision-making system[1] and knowledge of the input dataset on which it operates. This setting models the kind of access available to a private or public sector entity that pro-actively publishes transparency reports. It also models a useful level of access required for internal or external oversight of such systems to identify harms introduced by them. For the former use case, our approach provides a basis for design of transparency mechanisms; for the latter, it provides a formal basis for testing. Returning to our predictive policing system, the law enforcement agency that employs it could proactively publish transparency reports, and test the system for early detection of harms like race-based discrimination. An oversight agency could also use transparency reports for post hoc identification of harms.

Our Approach. We formalize transparency reports by introducing a family of *Quantitative Input Influence (QII)* measures that capture the degree of influence of inputs on outputs of the system. Three desiderata drove the definitions of these measures.

First, we seek a formalization of a *general* class of transparency reports that allows us to answer many useful transparency queries related to input influence, including but not limited to the example forms described above about the system's decisions about individuals and groups.

Second, we seek input influence measures that appropriately account for *correlated inputs*—a common case for our target applications. For example, consider a system that assists in hiring decisions for a moving company. Gender and the ability to lift heavy weights are inputs to the system. They are positively correlated with each other and with the hiring decisions. Yet transparency into whether the system uses the weight lifting ability or the gender in making its decisions (and to what degree) has substantive implications for determining if it is engaging in discrimination (the business necessity defense could apply in the former case [7]). This observation makes us look beyond correlation coefficients and other associative measures.

Third, we seek measures that appropriately quantify input influence in settings where any input by itself does not have significant influence on outcomes but a set of inputs does. In such cases, we seek measures of *joint influence* of a set of inputs (e.g., age and income) on a system's decision (e.g., to serve a high-paying job ad). We also seek measures of *marginal influence* of an input within such a set (e.g., age) on the decision. This notion allows us to provide finer-grained transparency about the relative importance of individual inputs within the set (e.g., age vs. income) in the system's decision.

We achieve the first desideratum by formalizing a notion of a *quantity of interest*. A transparency query measures the influence of an input on a quantity of interest. A quantity of interest represents a property of the behavior of the system for a given input distribution. Our formalization supports a wide range of statistical properties including probabilities of various outcomes in the output distribution and probabilities of output distribution outcomes conditioned on input distribution events. Examples of quantities of interest include the conditional probability of an outcome for a particular individual or group, and the ratio of conditional probabilities for an outcome for two different groups (a metric used as evidence of disparate impact under discrimination law in the US [7]).

We achieve the second desideratum by formalizing *causal* QII measures. These measures (called *Unary QII*) model the difference in the quantity of interest when the system operates over two related input distributions—the real distribution and a hypothetical (or counterfactual) distribution that is constructed from the real distribution in a specific way to account for correlations among inputs. Specifically, if we are interested in measuring the influence of an input on a quantity of interest of the system behavior, we construct the hypothetical distribution by retaining the marginal distribution over all other inputs and sampling the input of interest from its prior distribution. This choice breaks the correlations between this input and all other inputs and thus lets us measure the influence of this input on the quantity of interest, independently of other correlated inputs. Revisiting our moving company hiring example, if the system makes decisions only using the weightlifting ability of applicants, the influence of gender will be zero on the ratio of conditional probabilities of being hired for males and females.

We achieve the third desideratum in two steps. First, we define a notion of joint influence of a set of inputs (called *Set QII*) via a natural generalization of the definition of the hypothetical distribution in the Unary QII definition. Second, we define a family of *Marginal QII* measures that model the difference on the quantity of interest as we consider sets with and without the specific input whose marginal influence we want to measure. Depending on the application, we may pick these sets in different ways, thus motivating several different measures. For example, we could fix a set of inputs and ask about the marginal influence of any given input in that set on the quantity of interest. Alternatively, we may be interested in the average marginal influence of an input when it belongs to one of several different sets that significantly affect the quantity of interest. We consider several marginal influence aggregation measures from cooperative game theory originally developed in the context of influence measurement in voting scenarios and discuss their applicability in our setting. We also build on that literature to present an efficient approximate algorithm for computing these measures.

Recognizing that different forms of transparency reports may be appropriate for different settings, we generalize our QII measures to be parametric in its key elements: the intervention

[1]By "black-box access to the decision-making system" we mean a typical setting of software testing with complete control of inputs to the system and full observability of the outputs.

used to construct the hypothetical input distribution; the quantity of interest; the difference measure used to quantify the distance in the quantity of interest when the system operates over the real and hypothetical input distributions; and the aggregation measure used to combine marginal QII measures across different sets. This generalized definition provides a structure for exploring the design space of transparency reports.

Since transparency reports released to an individual, regulatory agency, or the public might compromise individual privacy, we explore the possibility of answering transparency queries while protecting differential privacy [8]. We prove bounds on the sensitivity of a number of transparency queries and leverage prior results on privacy amplification via sampling [9] to accurately answer these queries.

We demonstrate the utility of the QII framework by developing two machine learning applications on real datasets: an income classification application based on the benchmark `adult` dataset [10], and a predictive policing application based on the National Longitudinal Survey of Youth [11]. Using these applications, we argue, in Section VII, the need for causal measurement by empirically demonstrating that in the presence of correlated inputs, observational measures are not informative in identifying input influence. Further, we analyze transparency reports of individuals in our dataset to demonstrate how Marginal QII can provide insights into individuals' classification outcomes. Finally, we demonstrate that under most circumstances, QII measures can be made differentially private with minimal addition of noise, and can be approximated efficiently.

In summary, this paper makes the following contributions:

- A formalization of a specific algorithmic transparency problem for decision-making systems. Specifically, we define a family of Quantitative Input Influence metrics that accounts for correlated inputs, and provides answers to a general class of transparency queries, including the absolute and marginal influence of inputs on various behavioral system properties. These metrics can inform the design of transparency mechanisms and guide proactive system testing and posthoc investigations.
- A formal treatment of privacy-transparency trade-offs, in particular, by construction of differentially private answers to transparency queries.
- An implementation and experimental evaluation of the metrics over two real data sets. The evaluation demonstrates that (a) the QII measures are *informative*; (b) they remain *accurate* while preserving differential privacy; and (c) can be *computed* quite quickly for standard machine learning systems applied to real data sets.

II. Unary QII

Consider the situation discussed in the introduction, where an automated system assists in hiring decisions for a moving company. The input features used by this classification system are : *Age*, *Gender*, *Weight Lifting Ability*, *Marital Status* and *Education*. Suppose that, as before, weight lifting ability is strongly correlated with gender (with men having better overall lifting ability than woman). One particular question that an analyst may want to ask is: "What is the influence of the input *Gender* on positive classification for women?". The analyst observes that 20% of women are approved according to his classifier. Then, he replaces every woman's field for gender with a random value, and notices that the number of women approved does not change. In other words, an *intervention* on the *Gender* variable does not cause a significant change in the classification outcome. Repeating this process with *Weight Lifting Ability* results in a 20% increase in women's hiring. Therefore, he concludes that for this classifier, *Weight Lifting Ability* has more influence on positive classification for women than *Gender*.

By breaking correlations between gender and weight lifting ability, we are able to establish a causal relationship between the outcome of the classifier and the inputs. We are able to identify that despite the strong correlation between a negative classification outcome for women, the feature gender was not a cause of this outcome. We formalize the intuition behind such causal experimentation in our definition of Quantitative Input Influence (QII).

We are given an algorithm $\mathcal{A}$. $\mathcal{A}$ operates on inputs (also referred to as *features* for ML systems), $N = \{1, \ldots, n\}$. Every $i \in N$, can take on various *states*, given by X_i. We let $\mathcal{X} = \prod_{i \in N} \mathcal{X}_i$ be the set of possible feature state vectors, let $\mathcal{Z}$ be the set of possible outputs of $\mathcal{A}$. For a vector $\mathbf{x} \in \mathcal{X}$ and set of inputs $S \subseteq N$, $\mathbf{x}|_S$ denotes the vector of inputs in S. We are also given a probability distribution π on $\mathcal{X}$, where $\pi(\mathbf{x})$ is the probability of the input vector $\mathbf{x}$. We can define a marginal probability of a set of inputs S in the standard way as follows:

$$\pi_S(\mathbf{x}|_S) = \sum_{\{\mathbf{x}' \in \mathcal{X} | \mathbf{x}'|_S = \mathbf{x}|_S\}} \pi(\mathbf{x}') \tag{1}$$

When S is a singleton set $\{i\}$, we write the marginal probability of the single input as $\pi_i(x)$.

Informally, to quantify the influence of an input i, we compute its effect on some *quantity of interest*; that is, we measure the difference in the quantity of interest, when the feature i is changed via an intervention. In the example above, the quantity of interest is the fraction of positive classification of women. In this paper, we employ a particular interpretation of "changing an input", where we replace the value of every input with a random independently chosen value. To describe the replacement operation for input i, we first define an expanded probability space on $\mathcal{X} \times \mathcal{X}$, with the following distribution:

$$\tilde{\pi}(\mathbf{x}, \mathbf{u}) = \pi(\mathbf{x})\pi(\mathbf{u}). \tag{2}$$

The first component of an expanded vector $(\mathbf{x}, \mathbf{u})$, is just the original input vector, whereas the second component represents an independent random vector drawn from the same distribution π. Over this expanded probability space, the random variable $X(\mathbf{x}, u_i) = \mathbf{x}$ represents the original feature vector.

The random variable $X_{-i}U_i(\mathbf{x},\mathbf{u}) = \mathbf{x}|_{N\setminus\{i\}}u_i$, represents the random variable with input i replaced with a random sample. Defining this expanded probability space allows us to switch between the original distribution, represented by the random variable X, and the *intervened distribution*, represented by $X_{-i}U_i$. Notice that both these random variables are defined from $\mathcal{X}\times\mathcal{X}$, the expanded probability space, to $\mathcal{X}$. We denote the set of random variables of the type $\mathcal{X}\times\mathcal{X}\to\mathcal{X}$ as $\mathfrak{R}(\mathcal{X})$.

We can now define probabilities over this expanded space. For example, the probability over X remains the same:

$$\begin{aligned}\Pr(X=\mathbf{x}) &= \sum_{\{(\mathbf{x}',\mathbf{u}')|\mathbf{x}'=\mathbf{x}\}} \tilde{\pi}(\mathbf{x}',\mathbf{u}') \\ &= \left(\sum_{\{\mathbf{x}'|\mathbf{x}'=\mathbf{x}\}} \pi(\mathbf{x}')\right)\left(\sum_{\mathbf{u}'}\pi(\mathbf{u}')\right) \\ &= \pi(\mathbf{x})\end{aligned}$$

Similarly, we can define more complex quantities. The following expression represents the expectation of a classifier c evaluating to 1, when i is randomly intervened on:

$$\mathbb{E}(c(X_{-i}U_i)=1) = \sum_{\{(\mathbf{x},\mathbf{u})|c(\mathbf{x}_{N\setminus i}u_i)=1\}} \tilde{\pi}(\mathbf{x},u_i).$$

Observe that the expression above computes the probability of the classifier c evaluating to 1, when input i is replaced with a random sample from its probability distribution $\pi_i(u_i)$.

$$\begin{aligned}&\sum_{\{(\mathbf{x},\mathbf{u})|c(\mathbf{x}_{N\setminus i}u_i)=1\}} \tilde{\pi}(\mathbf{x},u_i) \\ =&\sum_{\mathbf{x}}\pi(\mathbf{x}) \sum_{\{u_i'|c(\mathbf{x}_{N\setminus i}u_i')=1\}}\ \sum_{\{\mathbf{u}|u_i=u_i'\}}\pi(\mathbf{u}) \\ =&\sum_{\mathbf{x}}\pi(\mathbf{x}) \sum_{\{u_i'|c(\mathbf{x}_{N\setminus i}u_i')=1\}}\pi_i(u_i')\end{aligned}$$

We can also define conditional distributions in the usual way. The following represents the probability of the classifier evaluating to 1 under the randomized intervention on input i of X, given that X belongs to some subset $\mathcal{Y}\subseteq\mathcal{X}$:

$$\mathbb{E}(c(X_{-i}U_i)=1 \mid X\in\mathcal{Y}) = \frac{\mathbb{E}(c(X_{-i}U_i)=1\wedge X\in\mathcal{Y})}{\mathbb{E}(X\in\mathcal{Y})}.$$

Formally, for an algorithm $\mathcal{A}$, a *quantity of interest* $Q_{\mathcal{A}}(\cdot): \mathfrak{R}(\mathcal{X})\mapsto\mathbb{R}$ is a function of a random variable from $\mathfrak{R}(\mathcal{X})$.

Definition 1 (QII)**.** For a quantity of interest $Q_{\mathcal{A}}(\cdot)$, and an input i, the Quantitative Input Influence of i on $Q_{\mathcal{A}}(\cdot)$ is defined to be

$$\iota^{Q_{\mathcal{A}}}(i) = Q_{\mathcal{A}}(X) - Q_{\mathcal{A}}(X_{-i}U_i).$$

In the example above, for a classifier $\mathcal{A}$, the quantity of interest, the fraction of women (represented by the set $\mathcal{W}\subseteq\mathcal{X}$) with positive classification, can be expressed as follows:

$$Q_{\mathcal{A}}(\cdot) = \mathbb{E}(\mathcal{A}(\cdot)=1 \mid X\in\mathcal{W}),$$

and the influence of input i is:

$$\iota(i) = \mathbb{E}(\mathcal{A}(X)=1 \mid X\in\mathcal{W}) - \mathbb{E}(\mathcal{A}(X_{-i}U_i)=1 \mid X\in\mathcal{W}).$$

When $\mathcal{A}$ is clear from the context, we simply write Q rather than $Q_{\mathcal{A}}$. We now instantiate this definition with different quantities of interest to illustrate the above definition in three different scenarios.

A. QII for Individual Outcomes

One intended use of QII is to provide personalized transparency reports to users of data analytics systems. For example, if a person is denied a job application due to feedback from a machine learning algorithm, an explanation of which factors were most influential for that person's classification can provide valuable insight into the classification outcome.

For QII to quantify the use of an input for individual outcomes, we define the quantity of interest to be the classification outcome for a particular individual. Given a particular individual $\mathbf{x}$, we define $Q^{\mathbf{x}}_{\text{ind}}(\cdot)$ to be $\mathbb{E}(c(\cdot)=1 \mid X=\mathbf{x})$. The influence measure is therefore:

$$\iota^{\mathbf{x}}_{\text{ind}}(i) = \mathbb{E}(c(X)=1 \mid X=\mathbf{x}) - \mathbb{E}(c(X_{-i}U_i)=1 \mid X=\mathbf{x}) \tag{3}$$

When the quantity of interest is not the probability of positive classification but the classification that $\mathbf{x}$ actually received, a slight modification of the above QII measure is more appropriate:

$$\begin{aligned}\iota^{\mathbf{x}}_{\text{ind-act}}(i) &= \mathbb{E}(c(X)=c(\mathbf{x}) \mid X=\mathbf{x}) \\ &\quad -\mathbb{E}(c(X_{-i}U_i)=c(\mathbf{x}) \mid X=\mathbf{x}) \\ &= 1-\mathbb{E}(c(X_{-i}U_i)=c(\mathbf{x}) \mid X=\mathbf{x}) \\ &= \mathbb{E}(c(X_{-i}U_i)\neq c(\mathbf{x}) \mid X=\mathbf{x})\end{aligned} \tag{4}$$

The above probability can be interpreted as the probability that feature i is pivotal to the classification of $c(\mathbf{x})$. Computing the average of this quantity over X yields:

$$\begin{aligned}&\textstyle\sum_{\mathbf{x}\in\mathcal{X}}\Pr(X=\mathbf{x})\mathbb{E}(i \text{ is pivotal for } c(X) \mid X=\mathbf{x}) \\ =\ &\mathbb{E}(i \text{ is pivotal for } c(X)).\end{aligned} \tag{5}$$

We denote this average QII for individual outcomes as defined above, by $\iota_{\text{ind-avg}}(i)$, and use it as a measure for importance of an input towards classification outcomes.

B. QII for Group Outcomes

As in the running example, the quantity of interest may be the classification outcome for a set of individuals. Given a group of individuals $\mathcal{Y} \subseteq \mathcal{X}$, we define $Q^{\mathcal{Y}}_{\text{grp}}(\cdot)$ to be $\mathbb{E}(c(\cdot) = 1 \mid X \in \mathcal{Y})$. The influence measure is therefore:

$$\iota^{\mathcal{Y}}_{\text{grp}}(i) = \mathbb{E}(c(X) = 1 \mid X \in \mathcal{Y}) - \mathbb{E}(c(X_{-i}U_i) = 1 \mid X \in \mathcal{Y}) \tag{6}$$

C. QII for Group Disparity

Instead of simply classification outcomes, an analyst may be interested in more nuanced properties of data analytics systems. Recently, disparate impact has come to the fore as a measure of unfairness, which compares the rates of positive classification within protected groups defined by gender or race. The '80% rule' in employment which states that the rate of selection within a protected demographic should be at least 80% of the rate of selection within the unprotected demographic. The quantity of interest in such a scenario is the ratio in positive classification outcomes for a protected group $\mathcal{Y}$ from the rest of the population $\mathcal{X} \setminus \mathcal{Y}$.

$$\frac{\mathbb{E}(c(X) = 1 \mid X \in \mathcal{Y})}{\mathbb{E}(c(X) = 1 \mid X \notin \mathcal{Y})}$$

However, the ratio of classification rates is unstable at at low values of positive classification. Therefore, for the computations in this paper we use the difference in classification rates as our measure of group disparity.

$$Q^{\mathcal{Y}}_{\text{disp}}(\cdot) = |\mathbb{E}(c(\cdot) = 1 \mid X \in \mathcal{Y}) - \mathbb{E}(c(\cdot) = 1 \mid X \notin \mathcal{Y})| \tag{7}$$

The QII measure of an input group disparity, as a result is:

$$\iota^{\mathcal{Y}}_{\text{disp}}(i) = Q^{\mathcal{Y}}_{\text{disp}}(X) - Q^{\mathcal{Y}}_{\text{disp}}(X_{-i}U_i). \tag{8}$$

More generally, group disparity can be viewed as an association between classification outcomes and membership in a group. QII on a measure of such association (e.g., group disparity) identifies the variable that causes the association in the classifier. *Proxy variables* are variables that are associated with protected attributes. However, for concerns of discrimination such as *digital redlining*, it is important to identify which proxy variables actually introduce group disparity. It is straightforward to observe that features with high QII for group disparity are proxy variables, and also cause group disparity. Therefore, QII on group disparity is a useful diagnostic tool for determining discriminiation. The use of QII in identifying proxy variables is explored experimentally in Section VII-B. Note that because of such proxy variables, simply ensuring that protected attributes are not input to the classifier is not sufficient to avoid discrimination (see also [12]).

III. Set and Marginal QII

In many situations, intervention on a single input variable has no influence on the outcome of a system. Consider, for example, a two-feature setting where features are age (A) and income (I), and the classifier is $c(A, I) = (A = old) \wedge (I = high)$. In other words, the only datapoints that are labeled 1 are those of elderly persons with high income. Now, given a datapoint where $A = young, I = low$, an intervention on either age or income would result in the same classification. However, it would be misleading to say that neither age nor income have an influence over the outcome: changing both the states of income and age would result in a change in outcome.

Equating influence with the *individual* ability to affect the outcome is uninformative in real datasets as well: Figure 1 is a histogram of influences of features on outcomes of individuals for a classifier learnt from the adult dataset [13][2]. For most individuals, all features have zero influence: changing the state of one feature alone is not likely to change the outcome of a classifier. Of the 19537 datapoints we evaluate, more than half have $\iota^{\mathbf{x}}(i) = 0$ for all $i \in N$, Indeed, changes to outcome are more likely to occur if we intervene on *sets of features*. In order to get a better understanding of the influence of a feature $i \in N$, we should measure its effect when coupled with interventions on other features. We define the influence of a set of inputs as a straightforward extension of the influence of individual inputs. Essentially, we wish the influence of a set of inputs $S \subseteq N$ to be the same as when the set of inputs is considered to be a single input; when intervening on S, we draw the states of $i \in S$ based on the joint distribution of the states of features in S, $\pi_S(\mathbf{u}_S)$, as defined in Equation (1).

We can naturally define a distribution over $\mathcal{X} \times \prod_{i \in S} \mathcal{X}_i$, naturally extending (2) as:

$$\tilde{\pi}(\mathbf{x}, u_S) = \pi(\mathbf{x})\pi_S(\mathbf{u}_S). \tag{9}$$

We also define the random variable $X_{-S}U_S(\mathbf{x}, \mathbf{u}_S) = \mathbf{x}|_{N \setminus S}\mathbf{u}_S$; $X_{-S}(\mathbf{x}, \mathbf{u}_S)$ has the states of features in $N \setminus S$ fixed to their original values in $\mathbf{x}$, but features in S take on new values according to $\mathbf{u}_S$.

Definition 2 (Set QII)**.** For a quantity of interest Q, and an input i, the Quantitative Input Influence of set $S \subseteq N$ on Q is defined to be

$$\iota^{Q}(S) = Q(X) - Q(X_{-S}U_S).$$

Considering the influence of a set of inputs opens up a number of interesting questions due to the interaction between inputs. First among these is how does one measure the *individual effect* of a feature, given the measured effects of interventions on sets of features. One natural way of doing so is by measuring the *marginal effect* of a feature on a set.

[2]The adult dataset contains approximately 31k datapoints of users' personal attributes, and whether their income is more than $50k per annum; see Section VII for more details.

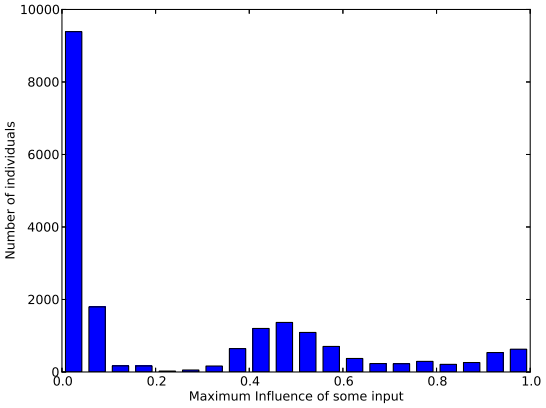


Fig. 1: A histogram of the highest specific causal influence for some feature across individuals in the adult dataset. Alone, most inputs alone have very low influence.

Definition 3 (Marginal QII)**.** For a quantity of interest Q, and an input i, the Quantitative Input Influence of input i over a set $S \subseteq N$ on Q is defined to be

$$\iota^Q(i, S) = Q(X_{-S}U_S) - Q(X_{-S\cup\{i\}}U_{S\cup\{i\}}).$$

Notice that marginal QII can also be viewed as a difference in set QIIs: $\iota^Q(S \cup \{i\}) - \iota^Q(S)$. Informally, the difference between $\iota^Q(S \cup \{i\})$ and $\iota^Q(S)$ measures the "added value" obtained by intervening on $S \cup \{i\}$, versus intervening on S alone.

The marginal contribution of i may vary significantly based on S. Thus, we are interested in the *aggregate marginal contribution* of i to S, where S is sampled from some natural distribution over subsets of $N \setminus \{i\}$. In what follows, we describe a few measures for aggregating the marginal contribution of a feature i to sets, based on different methods for sampling sets. The primary method of aggregating the marginal contribution is the Shapley value [14]. The less theoretically inclined reader can choose to proceed to Section V without a loss in continuity.

A. Cooperative Games and Causality

In this section, we discuss how measures from the theory of cooperative games define measures for aggregating marginal influence. In particular, we observe that the Shapley value [14] is characterized by axioms that are natural in our setting. However, other measures may be appropriate for certain input data generation processes.

Definition 2 measures the influence that an intervention on a set of features $S \subseteq N$ has on the outcome. One can naturally think of Set QII as a function $v : 2^N \to \mathbb{R}$, where $v(S)$ is the influence of S on the outcome. With this intuition in mind, one can naturally study influence measures using *cooperative game theory*, and in particular, prevalent influence measures in cooperative games such as the Shapley value, Banzhaf index and others. These measures can be thought of as *influence aggregation methods*, which, given an influence measure $v : 2^N \to \mathbb{R}$, output a vector $\phi \in \mathbb{R}^n$, whose i-th coordinate corresponds in some natural way to the aggregate influence, or aggregate causal effect, of feature i.

The original motivation for game-theoretic measures is *revenue division* [15, Chapter 18]: the function v describes the amount of money that each subset of players $S \subseteq N$ can generate; assuming that the set N generates a total revenue of $v(N)$, how should $v(N)$ be divided amongst the players? A special case of revenue division that has received significant attention is the measurement of voting power [16]. In voting systems with multiple agents with differing weights, voting power often does not directly correspond to the weights of the agents. For example, the US presidential election can roughly be modeled as a cooperative game where each state is an agent. The weight of a state is the number of electors in that state (i.e., the number of votes it brings to the presidential candidate who wins that state). Although states like California and Texas have higher weight, swing states like Pennsylvania and Ohio tend to have higher power in determining the outcome of elections.

A voting system is modeled as a cooperative game: players are voters, and the value of a coalition $S \subseteq N$ is 1 if S can make a decision (e.g. pass a bill, form a government, or perform a task), and is 0 otherwise. Note the similarity to classification, with players being replaced by features. The game-theoretic measures of revenue division are a measure of *voting power*: how much influence does player i have in the decision making process? Thus the notions of voting power and revenue division fit naturally with our goals when defining aggregate QII influence measures: in both settings, one is interested in measuring the aggregate effect that a single element has, given the actions of subsets.

A revenue division should ideally satisfy certain desiderata. Formally, we wish to find a function $\phi(N, v)$, whose input is N and $v : 2^N \to \mathbb{R}$, and whose output is a vector in $\mathbb{R}^n$, such that $\phi_i(N, v)$ measures some quantity describing the overall contribution of the i-th player. Research on fair revenue division in cooperative games traditionally follows an axiomatic approach: define a set of properties that a revenue division should satisfy, derive a function that outputs a value for each player, and argue that it is the unique function that satisfies these properties.

Several canonical fair cooperative solution concepts rely on the fundamental notion of *marginal contribution*. given a player i and a set $S \subseteq N \setminus \{i\}$, the marginal contribution of i to S is denoted $m_i(S, v) = v(S \cup \{i\}) - v(S)$ (we simply write $m_i(S)$ when v is clear from the context). Marginal QII, as defined above, can be viewed as an instance of a measure of marginal contribution. Given a permutation $\pi \in \Pi(N)$ of the elements in N, we define $P_i(\sigma) = \{j \in N \mid \sigma(j) < \sigma(i)\}$; this is the set of i's *predecessors* in σ. We can now similarly define the marginal contribution of i to a permutation $\sigma \in \Pi(N)$ as $m_i(\sigma) = m_i(P_i(\sigma))$. Intuitively, one can think of the players sequentially entering a room, according to some ordering σ; the value $m_i(\sigma)$ is the marginal contribution that i has to whoever is in the room when she enters it.

Generally speaking, game theoretic influence measures specify some reasonable way of aggregating the marginal contributions of i to sets $S \subseteq N$. That is, they measure a player's *expected marginal contribution* to sets sampled from some distribution $\mathcal{D}$ over 2^N, resulting in a payoff of

$$\mathbb{E}_{S\sim\mathcal{D}}[m_i(S)] = \sum_{S\subseteq N} \Pr_{\mathcal{D}}[S] m_i(S).$$

Thus, fair revenue division draws its appeal from the degree to which the distribution $\mathcal{D}$ is justifiable within the context where revenue is shared. In our setting, we argue for the use of the Shapley value. Introduced by the late Lloyd Shapley, the Shapley value is one of the most canonical methods of dividing revenue in cooperative games. It is defined as follows:

$$\varphi_i(N, v) = \mathbb{E}_\sigma[m_i(\sigma)] = \frac{1}{n!} \sum_{\sigma\in\Pi(N)} m_i(\sigma)$$

Intuitively, the Shapley value describes the following process: players are sequentially selected according to some randomly chosen order σ; each player receives a payment of $m_i(\sigma)$. The Shapley value is the expected payment to the players under this regime. The definition we use describes a distribution over permutations of N, not its subsets; however, it is easy to describe the Shapley value in terms of a distribution over subsets. If we define $p[S] = \frac{1}{n}\frac{1}{\binom{n-1}{|S|}}$, it is a simple exercise to show that

$$\varphi_i(N, v) = \sum_{S\subseteq N} p[S] m_i(S).$$

Intuitively, $p[S]$ describes the following process: first, choose a number $k \in [0, n-1]$ uniformly at random; next, choose a set of size k uniformly at random.

The Shapley value is one of many reasonable ways of measuring influence; we provide a detailed review of two others — the *Banzhaf index* [17], and the *Deegan-Packel index* [18] — in Appendix A.

B. Axiomatic Treatment of the Shapley Value

In this work, the Shapley value is our function of choice for aggregating marginal feature influence. The objective of this section is to justify our choice, and provide a brief exposition of axiomatic game-theoretic value theory. We present the axioms that define the Shapley value, and discuss how they apply in the QII setting. As we show, by requiring some desired properties, one arrives at a game-theoretic influence measure as the *unique* function for measuring information use in our setting.

The Shapley value satisfies the following properties:

Definition 4 (Symmetry (Sym))**.** We say that $i, j \in N$ are *symmetric* if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$. A value ϕ satisfies *symmetry* if $\phi_i = \phi_j$ whenever i and j are symmetric.

Definition 5 (Dummy (Dum))**.** We say that a player $i \in N$ is a *dummy* if $v(S \cup \{i\}) = v(S)$ for all $S \subseteq N$. A value ϕ satisfies the *dummy* property if $\phi_i = 0$ whenever i is a dummy.

Definition 6 (Efficiency (Eff))**.** A value satisfies the *efficiency* property if $\sum_{i\in N} \phi_i = v(N)$.

All of these axioms take on a natural interpretation in the QII setting. Indeed, if two features have the same probabilistic effect, no matter what other interventions are already in place, they should have the same influence. In our context, the dummy axiom says that a feature that never offers information with respect to an outcome should have no influence. In the case of specific causal influence, the efficiency axiom simply states that the total amount of influence should sum to

$$\begin{aligned}&\Pr(c(X) = c(\mathbf{x}) \mid X = \mathbf{x}) - \Pr(c(X_{-N}) = c(\mathbf{x}) \mid X = \mathbf{x})\\ =&1 - \Pr(c(X) = c(\mathbf{x})) = \Pr(c(X) \neq c(\mathbf{x})).\end{aligned}$$

That is, the total amount of influence possible is the likelihood of encountering elements whose evaluation is not $c(\mathbf{x})$. This is natural: if the vast majority of elements have a value of $c(\mathbf{x})$, it is quite unlikely that changes in features' state will have any effect on the outcome whatsoever; thus, the total amount of influence that can be assigned is $\Pr(c(X) \neq c(\mathbf{x}))$. Similarly, if the vast majority of points have a value different from $\mathbf{x}$, then it is likelier that a random intervention would result in a change in value, resulting in more influence to be assigned.

In the original paper by [14], it is shown that the Shapley value is the only function that satisfies (Sym), (Dum), (Eff), as well as the additivity (Add) axiom.

Definition 7 (Additivity (Add))**.** Given two games $\langle N, v_1\rangle, \langle N, v_2\rangle$, we write $\langle N, v_1 + v_2\rangle$ to denote the game $v'(S) = v_1(S) + v_2(S)$ for all $S \subseteq N$. A value ϕ satisfies the *additivity* property if $\phi_i(N, v_1) + \phi_i(N, v_2) = \phi_i(N, v_1 + v_2)$ for all $i \in N$.

In our setting, the additivity axiom makes little intuitive sense; it would imply, for example, that if we were to multiply Q by a constant c, the influence of i in the resulting game should be multiplied by c as well, which is difficult to justify.

[19] offers an alternative characterization of the Shapley value, based on the more natural *monotonicity* assumption, which is a strong generalization of the dummy axiom.

Definition 8 (Monotonicity (Mono))**.** Given two games $\langle N, v_1\rangle, \langle N, v_2\rangle$, a value ϕ satisfies *strong monotonicity* if $m_i(S, v_1) \geq m_i(S, v_2)$ for all S implies that $\phi_i(N, v_1) \geq \phi_i(N, v_2)$, where a strict inequality for some set $S \subseteq N$ implies a strict inequality for the values as well.

Monotonicity makes intuitive sense in the QII setting: if a feature has consistently higher influence on the outcome in one setting than another, its measure of influence should increase. For example, if a user receives two transparency reports (say, for two separate loan applications), and in one report gender had a consistently higher effect on the outcome than in the other, then the transparency report should reflect this.

Theorem 9 ([19])**.** *The Shapley value is the only function that satisfies (Sym), (Eff) and (Mono).*

To conclude, the Shapley value is a *unique* way of measuring aggregate influence in the QII setting, while satisfying a set of very natural axioms.

IV. Transparency Schemas

We now discuss two generalizations of the definitions presented in Section II, and then define a transparency schema that map the space of transparency reports based on QII.

a) Intervention Distribution: In this paper we only consider randomized interventions when the interventions are drawn independently from the priors of the given input. However, depending on the specific causal question at hand, we may use different interventions. Formally, this is achieved by allowing an arbitrary intervention distribution π^{inter} such that

$$\tilde{\pi}(\mathbf{x}, \mathbf{u}) = \pi(\mathbf{x})\pi^{\text{inter}}(\mathbf{u}).$$

The subsequent definitions remain unchanged. One example of an intervention different from the randomized intervention considered in the rest of the paper is one held constant at a vector $\mathbf{x}_0$:

$$\pi^{\text{inter}}_{\mathbf{x}_0}(\mathbf{u}) = \begin{cases} 1 & \text{for } \mathbf{u} = \mathbf{x}_0 \\ 0 & \text{o.w.} \end{cases}$$

A QII measure defined on the constant intervention as defined above, measures the influence of being different from a default, where the default is represented by $\mathbf{x}_0$.

b) Difference Measure: A second generalization allows us to consider quantities of interest which are not real numbers. Consider, for example, the situation where the quantity of interest is an output probability distribution, as in the case in a randomized classifier. In this setting, a suitable measure for quantifying the distance between distributions can be used as a difference measure between the two quantities of interest. Examples of such difference measures include the KL-divergence [20] between distribution or distance metrics between vectors.

c) Transparency Schema: We now present a transparency schema that maps the space of transparency reports based on QII measures. It consists of the following elements:

- A *quantity of interest*, which captures the aspect of the system we wish to gain transparency into.
- An *intervention distribution*, which defines how a counterfactual distribution is constructed from the true distribution.
- A *difference measure*, which quantifies the difference between two quantities of interest.
- An *aggregation technique*, which combines marginal QII measures across different subsets of inputs (features).

For a given application, one has to appropriately instantiate this schema. We have described several instances of each schema element. The choices of the schema elements are guided by the particular causal question being posed. For instance, when the question is: "Which features are most important for group disparity?", the natural quantity of interest is a measure of group disparity, and the natural intervention distribution is using the prior as the question does not suggest a particular bias. On the other hand, when the question is: "Which features are most influential for person A's classification as opposed to person B?", a natural quantity of interest is person A's classification, and a natural intervention distribution is the constant intervention using the features of person B. A thorough exploration of other points in this design space remains an important direction for future work.

V. Estimation

While the model we propose offers several appealing properties, it faces several technical implementation issues. Several elements of our work require significant computational effort; in particular, both the probability that a change in feature state would cause a change in outcome, and the game-theoretic influence measures are difficult to compute exactly. In the following sections we discuss these issues and our proposed solutions.

A. Computing Power Indices

Computing the Shapley or Banzhaf values exactly is generally computationally intractable (see [21, Chapter 4] for a general overview); however, their probabilistic nature means that they can be well-approximated via random sampling. More formally, given a random variable X, suppose that we are interested in estimating some determined quantity $q(X)$ (say, $q(X)$ is the mean of X); we say that a random variable q^* is an ε-δ approximation of $q(X)$ if

$$\Pr[|q^* - q(X)| \geq \varepsilon] < \delta;$$

in other words, it is extremely likely that the difference between $q(X)$ and q^* is no more than ε. An ε-δ *approximation scheme* for $q(X)$ is an algorithm that for any $\varepsilon, \delta \in (0, 1)$ is able to output a random variable q^* that is an ε-δ approximation of $q(X)$, and runs in time polynomial in $\frac{1}{\varepsilon}, \log 1\delta$.

[22] show that when $\langle N, v\rangle$ is a *simple* game (i.e. a game where $v(S) \in \{0, 1\}$ for all $S \subseteq N$), there exists an ε-δ approximation scheme for both the Banzhaf and Shapley values; that is, for $\phi \in \{\varphi, \beta\}$, we can guarantee that for any $\varepsilon, \delta > 0$, with probability $\geq 1 - \delta$, we output a value ϕ_i^* such that $|\phi_i^* - \phi_i| < \varepsilon$.

More generally, [23] observe that the number of i.i.d. samples needed in order to approximate the Shapley value and Banzhaf index is parametrized in $\Delta(v) = \max_{S \subseteq N} v(S) - \min_{S \subseteq N} v(S)$. Thus, if $\Delta(v)$ is a bounded value, then an ε-δ approximation exists. In our setting, coalitional values are always within the interval $[0, 1]$, which immediately implies the following theorem.

Theorem 10. *There exists an ε-δ approximation scheme for the Banzhaf and Shapley values in the QII setting.*

B. Estimating Q

Since we do not have access to the prior generating the data, we simply estimate it by observing the dataset itself. Recall that $\mathcal{X}$ is the set of all possible user profiles; in this

case, a dataset is simply a multiset (i.e. possibly containing multiple copies of user profiles) contained in $\mathcal{X}$. Let $\mathcal{D}$ be a finite multiset of $\mathcal{X}$, the input space. We estimate probabilities by computing sums over $\mathcal{D}$. For example, for a classifier c, the probability of $c(X) = 1$.

$$\hat{\mathbb{E}}_{\mathcal{D}}(c(X) = 1) = \frac{\sum_{\mathbf{x}\in\mathcal{D}} \mathbb{1}(c(\mathbf{x}) = 1)}{|\mathcal{D}|}. \tag{10}$$

Given a set of features $S \subseteq N$, let $\mathcal{D}|_S$ denote the elements of $\mathcal{D}$ truncated to only the features in S. Then, the intervened probability can be estimated as follows:

$$\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1) = \frac{\sum_{\mathbf{u}_S\in\mathcal{D}|_S} \sum_{\mathbf{x}\in\mathcal{D}} \mathbb{1}(c(\mathbf{x}|_{N\setminus S}\mathbf{u}_S) = 1)}{|\mathcal{D}|^2}. \tag{11}$$

Similarly, the intervened probability on individual outcomes can be estimated as follows:

$$\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1|X = \mathbf{x}) = \frac{\sum_{\mathbf{u}_S\in\mathcal{D}_S} \mathbb{1}(c(\mathbf{x}|_{N\setminus S}\mathbf{u}_S) = 1)}{|\mathcal{D}|}. \tag{12}$$

Finally, let us observe group disparity:

$$\left|\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1 \mid X \in \mathcal{Y}) - \hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1 \mid X \notin \mathcal{Y})\right|$$

The term $\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1 \mid X \in \mathcal{Y})$ equals

$$\frac{1}{|\mathcal{Y}|} \sum_{\mathbf{x}\in\mathcal{Y}} \sum_{\mathbf{u}_S\in\mathcal{D}_S} \mathbb{1}(c(\mathbf{x}|_{N\setminus S}\mathbf{u}_S) = 1),$$

Thus group disparity can be written as:

$$\Big|\frac{1}{|\mathcal{Y}|} \sum_{\mathbf{x}\in\mathcal{Y}} \sum_{\mathbf{u}_S\in\mathcal{D}_S} \mathbb{1}(c(\mathbf{x}|_{N\setminus S}\mathbf{u}_S) = 1)$$
$$-\frac{1}{|\mathcal{D}\setminus\mathcal{Y}|} \sum_{\mathbf{x}\in\mathcal{D}\setminus\mathcal{Y}} \sum_{\mathbf{u}_S\in\mathcal{D}_S} \mathbb{1}(c(\mathbf{x}|_{N\setminus S}\mathbf{u}_S) = 1)\Big|. \tag{13}$$

We write $\hat{Q}^{\mathcal{Y}}_{\text{disp}}(S)$ to denote (13).

If $\mathcal{D}$ is large, these sums cannot be computed efficiently. Therefore, we approximate the sums by sampling from the dataset $\mathcal{D}$. It is possible to show using the According to the Hoeffding bound [24], partial sums of n random variables X_i, within a bound Δ, can be well-approximated with the following probabilistic bound:

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2n\varepsilon^2}{\Delta}\right)$$

Since all the samples of measures discussed in the paper are bounded within the interval $[0, 1]$, we admit an ε-δ approximation scheme where the number of samples n can be chosen to be greater than $\log(2/\delta)/2\varepsilon^2$. Note that these bounds are independent of the size of the dataset. Therefore, given an efficient sampler, these quantities of interest can be approximated efficiently even for large datasets.

VI. Private Transparency Reports

One important concern is that releasing influence measures estimated from a dataset might leak information about individual users; our goal is providing accurate transparency reports, without compromising individual users' private data. To mitigate this concern, we add noise to make the measures differentially private. We show that the sensitivities of the QII measures considered in this paper are very low and therefore very little noise needs to be added to achieve differential privacy.

The *sensitivity* of a function is a key parameter in ensuring that it is differentially private; it is simply the worst-case change in its value, assuming that we change a single data point in our dataset. Given some function f over datasets, we define the sensitivity of a function f with respect to a dataset $\mathcal{D}$, denoted by $\Delta f(\mathcal{D})$ as

$$\max_{\mathcal{D}'} |f(\mathcal{D}) - f(\mathcal{D}')|$$

where $\mathcal{D}$ and $\mathcal{D}'$ differ by at most one instance. We use the shorthand Δf when $\mathcal{D}$ is clear from the context.

In order to not leak information about the users used to compute the influence of an input, we use the standard Laplace Mechanism [8] and make the influence measure differentially private. The amount of noise required depends on the sensitivity of the influence measure. We show that the influence measure has low sensitivity for the individuals used to sample inputs. Further, due to a result from [9] (and stated in [25]), sampling amplifies the privacy of the computed statistic, allowing us to achieve high privacy with minimal noise addition.

The standard technique for making any function differentially private is to add Laplace noise calibrated to the sensitivity of the function:

Theorem 11 ([8])**.** *For any function f from datasets to $\mathbb{R}$, the mechanism $\mathcal{K}_f$ that adds independently generated noise with distribution Lap$(\Delta f(\mathcal{D})/\epsilon)$ to the k output enjoys ϵ-differential privacy.*

Since each of the quantities of interest aggregate over a large number of instances, the sensitivity of each function is very low.

Theorem 12. *Given a dataset $\mathcal{D}$,*

1) $\Delta\hat{\mathbb{E}}_{\mathcal{D}}(c(X) = 1) = \frac{1}{|\mathcal{D}|}$
2) $\Delta\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1) \leq \frac{2}{|\mathcal{D}|}$
3) $\Delta\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1|X = \mathbf{x}) = \frac{1}{|\mathcal{D}|}$
4) $\hat{Q}^{\mathcal{Y}}_{disp}(S) \leq \max\left\{\frac{1}{|\mathcal{D}\cap\mathcal{Y}|}, \frac{1}{|\mathcal{D}\setminus\mathcal{Y}|}\right\}$

Proof. We examine some cases here. In Equation 10, if two datasets differ by one instance, then at most one term of the summation will differ. Since each term can only be either 0 or 1, the sensitivity of the function is

$$\Delta\hat{\mathbb{E}}_{\mathcal{D}}(c(X) = 1) = \left|\frac{0}{|\mathcal{D}|} - \frac{1}{|\mathcal{D}|}\right| = \frac{1}{|\mathcal{D}|}.$$

Similarly, in Equation 11, an instance appears $2|\mathcal{D}| - 1$ times, once each for the inner summation and the outer summation, and therefore, the sensitivity of the function is

$$\Delta\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1) = \frac{2|\mathcal{D}| - 1}{|\mathcal{D}|^2} \leq \frac{2}{|\mathcal{D}|}.$$

For individual outcomes (Equation (12)), similarly, only one term of the summation can differ. Therefore, the sensitivity of (12) is $1/|\mathcal{D}|$.

Finally, we observe that a change in a single element $\mathbf{x}'$ of $\mathcal{D}$ will cause a change of at most $\frac{1}{|\mathcal{D}\cap\mathcal{Y}|}$ if $\mathbf{x}' \in \mathcal{D} \cap \mathcal{Y}$, or of at most $\frac{1}{|\mathcal{D}\setminus\mathcal{Y}|}$ if $\mathbf{x}' \in \mathcal{D} \setminus \mathcal{Y}$. Thus, the maximal change to (13) is at most $\max\left\{\frac{1}{|\mathcal{Y}|}, \frac{1}{|\mathcal{D}\setminus\mathcal{Y}|}\right\}$.

□

While the sensitivity of most quantities of interest is low (at most a $\frac{2}{|\mathcal{D}|}$), $\hat{Q}^{\mathcal{Y}}_{\text{disp}}(S)$ can be quite high when $|\mathcal{Y}|$ is either very small or very large. This makes intuitive sense: if $\mathcal{Y}$ is a very small minority, then any changes to its members are easily detected; similarly, if $\mathcal{Y}$ is a vast majority, then changes to protected minorities may be easily detected.

We observe that the quantities of interest which exhibit low sensitivity will have low influence sensitivity as well: for example, the local influence of S is $\mathbb{1}(c(\mathbf{x}) = 1) - \hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1] \mid X = \mathbf{x})$; changing any $\mathbf{x}' \in \mathcal{D}$ (where $\mathbf{x}' \neq \mathbf{x}$ will result in a change of at most $\frac{1}{|\mathcal{D}|}$ to the local influence.

Finally, since the Shapley and Banzhaf indices are normalized sums of the differences of the set influence functions, we can show that if an influence function ι has sensitivity $\Delta\iota$, then the sensitivity of the indices are at most $2\Delta\iota$.

To conclude, all of the QII measures discussed above (except for group parity) have a sensitivity of $\frac{\alpha}{|\mathcal{D}|}$, with α being a small constant. To ensure differential privacy, we need only need add noise with a Laplacian distribution $\text{Lap}(k/|\mathcal{D}|)$ to achieve 1-differential privacy.

Further, it is known that sampling amplifies differential privacy.

Theorem 13 ([9], [25])**.** *If $\mathcal{A}$ is 1-differentially private, then for any $\epsilon \in (0, 1)$, $\mathcal{A}'(\epsilon)$ is 2ϵ-differentially private, where $\mathcal{A}'(\epsilon)$ is obtained by sampling an ϵ fraction of inputs and then running $\mathcal{A}$ on the sample.*

Therefore, our approach of sampling instances from $\mathcal{D}$ to speed up computation has the additional benefit of ensuring that our computation is private.

Table I contains a summary of all QII measures defined in this paper, and their sensitivity.

VII. Experimental Evaluation

We demonstrate the utility of the QII framework by developing two simple machine learning applications on real datasets. Using these applications, we first argue, in Section VII-A, the need for causal measurement by empirically demonstrating that in the presence of correlated inputs, observational measures are not informative in identifying which inputs were actually used. In Section VII-B, we illustrate the distinction between different quantities of interest on which Unary QII can be computed. We also illustrate the effect of discrimination on the QII measure. In Section VII-C, we analyze transparency reports of three individuals to demonstrate how Marginal QII can provide insights into individuals' classification outcomes. Finally, we analyze the loss in utility due to the use of differential privacy, and provide execution times for generating transparency reports using our prototype implementation.

We use the following datasets in our experiments:

- `adult` [10]: This standard machine learning benchmark dataset is a a subset of US census data that classifies the income of individuals, and contains factors such as age, race, gender, marital status and other socio-economic parameters. We use this dataset to train a classifier that predicts the income of individuals from other parameters. Such a classifier could potentially be used to assist credit decisions.
- `arrests` [11]: The National Longitudinal Surveys are a set of surveys conducted by the Bureau of Labor Statistics of the United States. In particular, we use the National Longitudinal Survey of Youth 1997 which is a survey of young men and women born in the years 1980-84. Respondents were ages 12-17 when first interviewed in 1997 and were subsequently interviewed every year till 2013. The survey covers various aspects of an individual's life such as medical history, criminal records and economic parameters. From this dataset, we extract the following features: age, gender, race, region, history of drug use, history of smoking, and history of arrests. We use this data to train a classifier that predicts history of arrests to aid in predictive policing, where socio-economic factors are used to decide whether individuals should receive a visit from the police. This application is inspired by a similar application in [26].

The two applications described above are hypothetical examples of decision-making aided by machine learning that use potentially sensitive socio-economic data about individuals, and not real systems that are currently in use. We use these classifiers to illustrate the subtle causal questions that our QII measures can answer.

We use the following standard machine learning classifiers in our dataset: Logistic Regression, SVM with a radial basis function kernel, Decision Tree, and Gradient Boosted Decision Trees. Bishop's machine learning text [27] is an excellent resource for an introduction to these classifiers. While Logistic Regression is a linear classifier, the other three are nonlinear and can potentially learn very complex models. All our experiments are implemented in Python with the numpy library, and the scikit-learn machine learning toolkit, and run on an Intel i7 computer with 4 GB of memory.

A. Comparison with Observational Measures

In the presence of correlated inputs, observational measures often cannot identify which inputs were causally influential. To illustrate this phenomena on real datasets, we train two

Name	Notation	Quantity of Interest	Sensitivity
QII on Individual Outcomes (3)	$\iota_{\text{ind}}(S)$	Positive Classification of an Individual	$1/\lvert\mathcal{D}\rvert$
QII on Actual Individual Outcomes (4)	$\iota_{\text{ind-act}}(S)$	Actual Classification of an Individual	$1/\lvert\mathcal{D}\rvert$
Average QII (5)	$\iota_{\text{ind-avg}}(S)$	Average Actual Classification	$2/\lvert\mathcal{D}\rvert$
QII on Group Outcomes (6)	$\iota^{\mathcal{Y}}_{\text{grp}}(S)$	Positive Classification for a Group	$2/\lvert\mathcal{D}\cap\mathcal{Y}\rvert$
QII on Group Disparity (8)	$\iota^{\mathcal{Y}}_{\text{disp}}(S)$	Difference in classification rates among groups	$2\max(1/\lvert\mathcal{D}\setminus\mathcal{Y}\rvert, 1/\lvert\mathcal{D}\cap\mathcal{Y}\rvert)$

TABLE I: A summary of the QII measures defined in the paper

classifiers: (A) where gender is provided as an actual input, and (B) where gender is not provided as an input. For classifier (B), clearly the input *Gender* has no effect and any correlation between the outcome and gender is caused via inference from other inputs. In Table II, for both the `adult` and the `arrests` dataset, we compute the following observational measures: Mutual Information (MI), Jaccard Index (Jaccard), Pearson Correlation (corr), and the Disparate Impact Ratio (disp) to measure the similarity between Gender and the classifiers outcome. We also measure the QII of Gender on outcome. We observe that in many scenarios the observational quantities do not change, or sometimes increase, from classifier A to classifier B, when gender is removed as an actual input to the classifier. On the other hand, if the outcome of the classifier does not depend on the input *Gender*, then the QII is guaranteed to be zero.

B. Unary QII Measures

In Figure 2, we illustrate the use of different Unary QII measures. Figures 2a, and 2b, show the Average QII measure (Equation 5) computed for features of a decision forest classifier. For the income classifier trained on the `adult` dataset, the feature with highest influence is *Marital Status*, followed by *Occupation*, *Relationship* and *Capital Gain*. Sensitive features such as *Gender* and *Race* have relatively lower influence. For the predictive policing classifier trained on the `arrests` dataset, the most influential input is *Drug History*, followed by *Gender*, and *Smoking History*. We observe that influence on outcomes may be different from influence on group disparity.

QII on group disparity: Figures 2c, 2d show influences of features on group disparity for two different settings. The figure on the left shows the influence of features on group disparity by Gender in the `adult` dataset; the figure on the right shows the influence of group disparity by Race in the `arrests` dataset. For the income classifier trained on the `adult` dataset, we observe that most inputs have negative influence on group disparity; randomly intervening on most inputs would lead to a reduction in group disparity. In other words, a classifier that did not use these inputs would be fairer. Interestingly, in this classifier, marital status and not sex has the highest influence on group disparity by sex.

For the `arrests` dataset, most inputs have the effect of increasing group disparity if randomly intervened on. In particular, *Drug history* has the highest positive influence on disparity in `arrests`. Although Drug history is correlated with race, using it reduces disparate impact by race, i.e. makes fairer decisions.

In both examples, features correlated with the sensitive attribute are the most influential for group disparity according to the sensitive attribute instead of the sensitive attribute itself. It is in this sense that QII measures can identify proxy variables that cause associations between outcomes and sensitive attributes.

QII with artificial discrimination: We simulate discrimination using an artificial experiment. We first randomly assign ZIP codes to individuals in our dataset. Then to simulate systematic bias, we make an f fraction of the ZIP codes discriminatory in the following sense: All individuals in the protected set are automatically assigned a negative classification outcome. We then study the change in the influence of features as we increase f. Figure 3a, shows that the influence of *Gender* increases almost linearly with f. Recall that *Marital Status* was the most influential feature for this classifier without any added discrimination. As f increases, the importance of *Marital Status* decreases as expected, since the number of individuals for whom *Marital Status* is pivotal decreases.

C. Personalized Transparency Reports

To illustrate the utility of personalized transparency reports, we study the classification of individuals who received potentially unexpected outcomes. For the personalized transparency reports, we use decision forests.

The influence measure that we employ is the Shapley value, with the underlying cooperative game defined over the local influence Q. In more detail, $v(S) = \iota^{Q_{\mathcal{A}}}(S)$, with $Q_{\mathcal{A}}$ being $\mathbb{E}[c(\cdot) = 1 \mid X = \mathbf{x}]$; that is, the marginal contribution of $i \in N$ to S is given by $m_i(S) = \mathbb{E}[c(X_{-S}) = 1 \mid X = \mathbf{x}] - \mathbb{E}[c(X_{-S\cup\{i\}}) = 1 \mid X = \mathbf{x}]$.

We emphasize that some features may have a negative Shapley value; this should be interpreted as follows: a feature with a high positive Shapley value often increases the certainty that the classification outcome is 1, whereas a feature whose Shapley value is negative is one that increases the certainty that the classification outcome would be zero.

Mr. X: The first example is of an individual from the `adult` dataset, who we refer to as Mr. X, and is described in Figure 4a. He is is deemed to be a low income individual, by an income classifier learned from the data. This result may be surprising to him: he reports high capital gains ($14k), and only 2.1% of people with capital gains higher than $10k are reported as low income. In fact, he might be led to believe that his classification may be a result of his ethnicity or country of origin. Examining his transparency report in Figure 4b, however, we find that the the most influential features that led

		logistic		kernel svm		decision tree		random forest	
		adult	arrests	adult	arrests	adult	arrests	adult	arrests
MI	A	0.045	0.049	0.046	0.047	0.043	0.054	0.044	0.053
MI	B	0.043	0.050	0.044	0.053	0.042	0.051	0.043	0.052
Jaccard	A	0.501	0.619	0.500	0.612	0.501	0.614	0.501	0.620
Jaccard	B	0.500	0.611	0.501	0.615	0.500	0.614	0.501	0.617
corr	A	0.218	0.265	0.220	0.247	0.213	0.262	0.218	0.262
corr	B	0.215	0.253	0.218	0.260	0.215	0.257	0.215	0.259
disp	A	0.286	0.298	0.377	0.033	0.302	0.335	0.315	0.223
disp	B	0.295	0.301	0.312	0.096	0.377	0.228	0.302	0.129
QII	A	0.036	0.135	0.044	0.149	0.023	0.116	0.012	0.109
QII	B	0	0	0	0	0	0	0	0

TABLE II: Comparison of QII with associative measures. For 4 different classifiers, we compute metrics such as Mutual Information (MI), Jaccard Index (JI), Pearson Correlation (corr), Group Disparity (disp) and Average QII between Gender and the outcome of the learned classifier. Each metric is computed in two situations: (A) when Gender is provided as an input to the classifier, and (B) when Gender is not provided as an input to the classifier.

to his negative classification were Marital Status, Relationship and Education.

Mr. Y: The second example, to whom we refer as Mr. Y (Figure 5), has even higher capital gains than Mr. X. Mr. Y is a 27 year old, with only Preschool education, and is engaged in fishing. Examination of the transparency report reveals that the most influential factor for negative classification for Mr. Y is his Occupation. Interestingly, his low level of education is not considered very important by this classifier.

Mr. Z: The third example, who we refer to as Mr. Z (Figure 6) is from the `arrests` dataset. History of drug use and smoking are both strong indicators of arrests. However, Mr. X received positive classification by this classifier even without any history of drug use or smoking. On examining his classifier, it appears that race, age and gender were most influential in determining his outcome. In other words, the classifier that we train for this dataset (a decision forest) has picked up on the correlations between race (Black), and age (born in 1984) to infer that this individual is likely to engage in criminal activity. Indeed, our interventional approach indicates that this is not a mere correlation effect: race is actively being used by this classifier to determine outcomes. Of course, in this instance, we have explicitly offered the race parameter to our classifier as a viable feature. However, our influence measure is able to pick up on this fact, and alert us of the problematic behavior of the underlying classifier. More generally, this example illustrates a concern with the black box use of machine learning which can lead to unfavorable outcomes for individuals.

D. Differential Privacy

Most QII measures considered in this paper have very low sensitivity, and therefore can be made differentially private with negligible loss in utility. However, recall that the sensitivity of influence measure on group disparity $\iota^{\mathcal{Y}}_{\text{disp}}$ depends on the size of the protected group in the dataset $\mathcal{D}$ as follows:

$$\iota^{\mathcal{Y}}_{\text{disp}} = 2\max\left(\frac{1}{|\mathcal{D} \setminus \mathcal{Y}|}, \frac{1}{|\mathcal{D} \cap \mathcal{Y}|}\right)$$

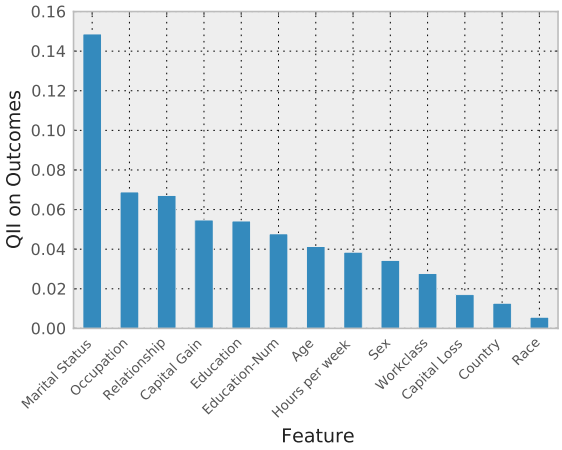
For sufficiently small minority groups, a large amount of noise might be required to ensure differential privacy, leading to a loss in utility of the QII measure. To estimate the loss in utility, we set a noise of 0.005 as the threshold of noise at which the measure is no longer useful, and then compute fraction of times noise crosses that threshold when Laplacian noise is added at $\epsilon = 1$. The results of this experiment are as follows:

$\mathcal{Y}$	Count	Loss in Utility
Race: White	27816	2.97×10^{-14}
Race: Black	3124	5.41×10^{-14}
Race: Asian-Pac-Islander	1039	6.14×10^{-05}
Race: Amer-Indian-Eskimo	311	0.08
Race: Other	271	0.13
Gender: Male	21790	3.3×10^{-47}
Gender: Female	10771	3.3×10^{-47}

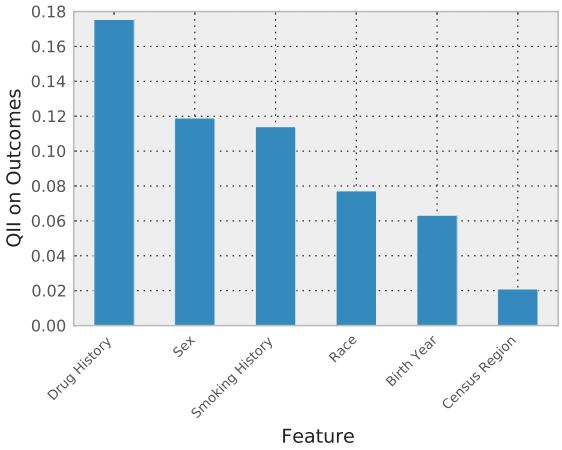
We note that for most reasonably sized groups, the loss in utility is negligible. However, the Asian-Pac-Islander, and the Amer-Indian-Eskimo racial groups are underrepresented in this dataset. For these groups, the QII on Group Disparity estimate needs to be very noisy to protect privacy.

E. Performance

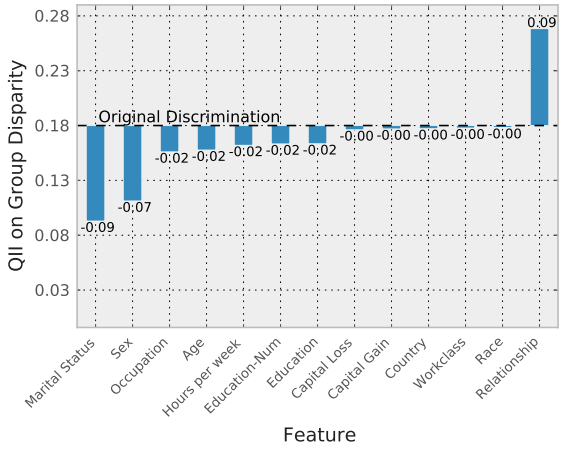
We report runtimes of our prototype for generating transparency reports on the `adult` dataset. Recall from Section VI that we approximate QII measures by computing sums over samples of the dataset. According to the Hoeffding bound to derive an (ϵ, δ) estimate of a QII measure, at $\epsilon = 0.01$, and $n = 37000$ samples, $\delta = 2\exp(-n\epsilon^2) < 0.05$ is an upper bound on the probability of the output being off by ϵ. Table III shows the runtimes of four different QII computations, for 37000 samples each. The runtimes of all algorithms except for kernel SVM are fast enough to allow real-time feedback for machine learning application developers. Evaluating QII metrics for Kernel SVMs is much slower than the other metrics because each call to the SVM classifier is very computationally intensive due to a large number of distance computations that it entails. We expect that these runtimes can be optimized significantly. We present them as proof of tractability.



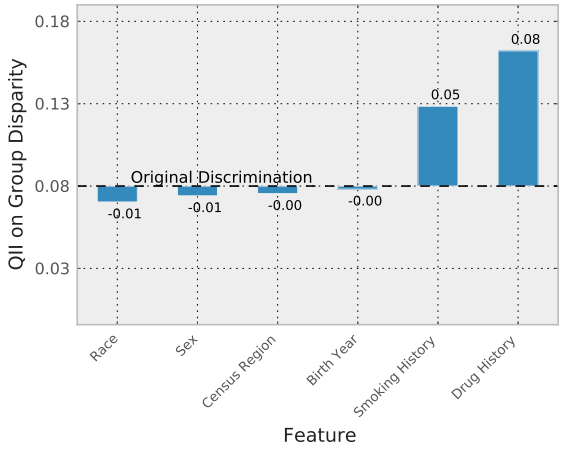
(a) QII of inputs on Outcomes for the `adult` dataset



(b) QII of inputs on Outcomes for the `arrests` dataset



(c) QII of Inputs on Group Disparity by Sex in the `adult` dataset



(d) Influence on Group Disparity by Race in the `arrests` dataset

Fig. 2: QII measures for the `adult` and `arrests` datasets

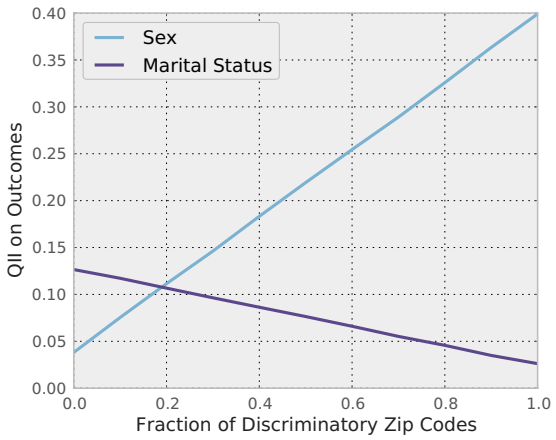
	logistic	kernel-svm	decision-tree	decision-forest
QII on Group Disparity	0.56	234.93	0.57	0.73
Average QII	0.85	322.82	0.77	1.12
QII on Individual Outcomes (Shapley)	6.85	2522.3	7.78	9.30
QII on Individual Outcomes (Banzhaf)	6.77	2413.3	7.64	10.34

TABLE III: Runtimes in seconds for transparency report computation

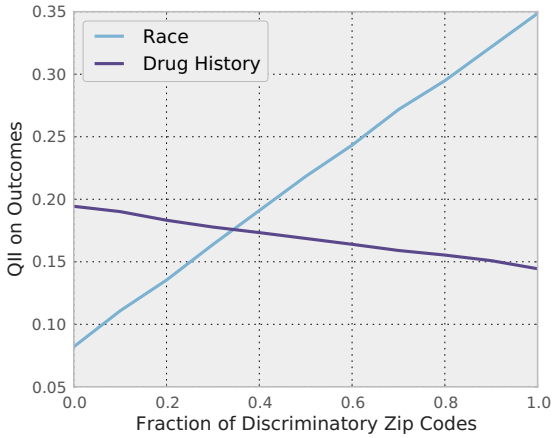
VIII. Discussion

A. Probabilistic Interpretation of Power Indices

In order to quantitatively measure the influence of data inputs on classification outcomes, we propose causal interventions on sets of features; as we argue in Section III, the aggregate marginal influence of i for different subsets of features is a natural quantity representing its influence. In order to aggregate the various influences i has on the outcome, it is natural to define some probability distribution over (or equivalently, a weighted sum of) subsets of $N \setminus \{i\}$, where $\Pr[S]$ represents the probability of measuring the marginal contribution of i to S; $\Pr[S]$ yields a value $\sum_{S \subseteq N \setminus \{i\}} m_i(S)$. For the Banzhaf index, we have $\Pr[S] = \frac{1}{2^{n-1}}$, the Shapley value has $\Pr[S] = \frac{k!(n-k-1)!}{n!}$ (here, $|S| = k$), and the Deegan-Packel Index selects minimal winning coalitions uniformly at random. These choices of values for $\Pr[S]$ are based on some natural assumptions on the way that players (features) interact, but they are by no means exhaustive. One can define other sampling methods that are more appropriate for the model at hand; for example, it is entirely possible that the only interventions that are possible in a certain setting are of size $\leq k+1$, it is reasonable to aggregate the marginal influence



(a) Change in QII of inputs as discrimination by Zip Code increases in the `adult` dataset



(b) Change in QII of inputs as discrimination by Zip Code increases in the `arrests` dataset

Fig. 3: The effect of discrimination on QII.

of i over sets of size $\leq k$, i.e.

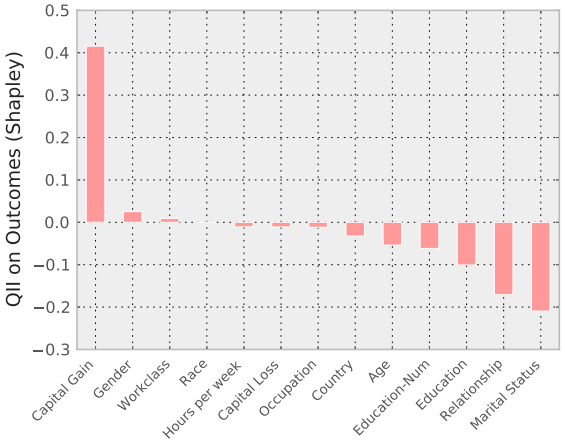
$$\Pr[S] = \begin{cases} \frac{1}{\binom{n-1}{|S|}} & \text{if } |S| \leq k \\ 0 & \text{otherwise.} \end{cases}$$

The key point here is that one must define *some* aggregation method, and that choice reflects some normative approach on how (and which) marginal contributions are considered. The Shapley and Banzhaf indices do have some highly desirable properties, but they are, first and foremost, *a-priori* measures of influence. That is, they do not factor in any assumptions on what interventions are possible or desirable.

One natural candidate for a probability distribution over S is some natural extension of the prior distribution over the dataset; for example, if all features are binary, one can identify a set with a feature vector (namely by identifying each $S \subseteq N$ with its indicator vector), and set $\Pr[S] = \pi(S)$ for all $S \subseteq N$.

Age	23
Workclass	Private
Education	11th
Education-Num	7
Marital Status	Never-married
Occupation	Craft-repair
Relationship	Own-child
Race	Asian-Pac-Islander
Gender	Male
Capital Gain	14344
Capital Loss	0
Hours per week	40
Country	Vietnam

(a) Mr. X's profile



(b) Transparency report for Mr. X's negative classification

Fig. 4: Mr. X

If features are not binary, then there is no canonical way to transition from the data prior to a prior over subsets of features.

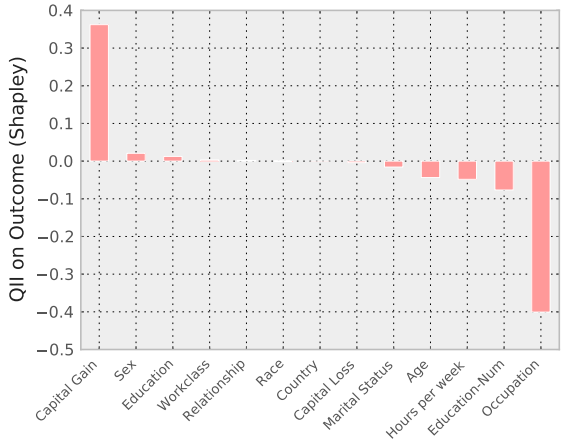
B. Fairness

Due to the widespread and black box use of machine learning in aiding decision making, there is a legitimate concern of algorithms introducing and perpetuating social harms such as racial discrimination [28], [6]. As a result, the algorithmic foundations of fairness in personal information processing systems have received significant attention recently [29], [30], [31], [12], [32]. While many of of the algorithmic approaches [29], [31], [32] have focused on group parity as a metric for achieving fairness in classification, Dwork et al. [12] argue that group parity is insufficient as a basis for fairness, and propose a similarity-based approach which prescribes that similar individuals should receive similar classification outcomes. However, this approach requires a similarity metric for individuals which is often subjective and difficult to construct.

QII does not suggest any normative definition of fairness. Instead, we view QII as a diagnostic tool to aid fine-grained fairness determinations. In fact, QII can be used in the spirit of the similarity based definition of [12]. By comparing the personalized privacy reports of individuals who are *perceived*

Age	27
Workclass	Private
Education	Preschool
Education-Num	1
Marital Status	Married-civ-spouse
Occupation	Farming-fishing
Relationship	Other-relative
Race	White
Gender	Male
Capital Gain	41310
Capital Loss	0
Hours per week	24
Country	Mexico

(a) Mr. Y's profile

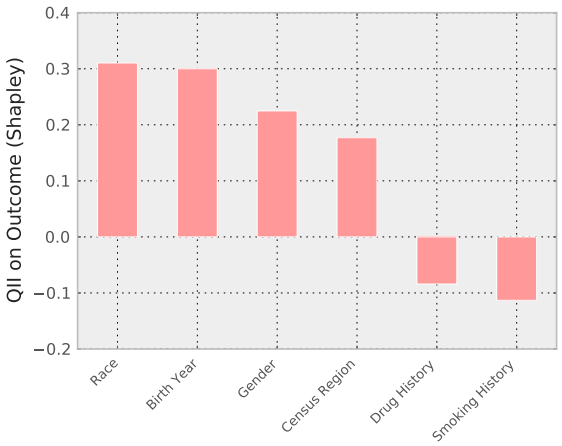


(b) Transparency report for Mr. Y's negative classification

Fig. 5: Mr. Y.

Birth Year	1984
Drug History	None
Smoking History	None
Census Region	West
Race	Black
Gender	Male

(a) Mr. Z's profile



(b) Transparency report for Mr. Z's positive classification

Fig. 6: Mr. Z.

to be similar but received different classification outcomes, and identifying the inputs which were used by the classifier to provide different outcomes. Additionally, when group parity is used as a criteria for fairness, QII can identify the features that lead to group disparity, thereby identifying features being used by a classifier as a proxy for sensitive attributes.

The determination of whether using certain proxies for sensitive attributes is discriminatory is often a task-specific normative judgment. For example, using standardized test scores (e.g., SAT scores) for admissions decisions is by and large accepted, although SAT scores may be a proxy for several protected attributes. In fact, several universities have recently announced that they will not use SAT scores for admissions citing this reason [33], [34]. Our goal is not to provide such normative judgments. Rather we seek to provide fine-grained transparency into input usage (e.g., what's the extent to which SAT scores influence decisions), which is useful to make determinations of discrimination from a specific normative position.

Finally, we note that an interesting question is whether providing a sensitive attribute as an input to a classifier is fundamentally discriminatory behavior, even if QII can show that the sensitive input has no significant impact on the outcome. Our view is that this is a policy question and different legal frameworks might take different viewpoints on it. At a technical level, from the standpoint of information use, the two situations are identical: the sensitive input is not really used although it is supplied. However, the very fact that it was supplied might be indicative of an intent to discriminate even if that intended goal was not achieved. No matter what the policy decision is on this question, QII remains a useful diagnostic tool for discrimination because of the presence of proxy variables as described earlier.

IX. Related Work

A. Quantitative Causal Measures

Causal models and probabilistic interventions have been used in a few other settings. While the form of the interventions in some of these settings may be very similar, our generalization to account for different quantities of interests enables us to reason about a large class of transparency queries for data analytics systems ranging from classification outcomes of individuals to disparity among groups. Further, the notion of marginal contribution which we use to compute responsibility does not appear in this line of prior work.

Janzing et al. [35] use interventions to assess the causal importance of relations between variables in causal graphs; in order to assess the causal effect of a relation between two variables, $X \rightarrow Y$ (assuming that both take on specific values $X = x$ and $Y = y$), a new causal model is constructed, where the value of X is replaced with a prior over the possible values of X. The influence of the causal relation is defined as the KL-

Divergence of the joint distribution of all the variables in the two causal models with and without the value of X replaced. The approach of the intervening with a random value from the prior is similar to our approach of constructing X_{-S}.

Independently, there has been considerable work in the machine learning community to define importance metrics for variables, mainly for the purpose of feature selection (see [36] for a comprehensive overview). One important metric is called Permutation Importance [37], which measures the importance of a feature towards classification by randomly permuting the values of the feature and then computing the difference of classification accuracies before and after the permutation. Replacing a feature with a random permutation can be viewed as a sampling the feature independently from the prior.

There exists extensive literature on establishing causal relations, as opposed to quantifying them. Prominently, Pearl's work [38] provides a mathematical foundation for causal reasoning and inference. In [39], Tian and Pearl discuss measures of causal strength for individual binary inputs and outputs in a probabilistic setting. Another thread of work by Halpern and Pearl discusses actual causation [40], which is extended in [41] to derive a measure of responsibility as degree of causality. In [41], Chockler and Halpern define the responsibility of a variable X to an outcome as the amount of change required in order to make X the counterfactual cause. As we discuss in Appendix A-B, the Deegan-Packel index is strongly related to causal responsibility.

B. Quantitative Information Flow

One can think of our results as a causal alternative to *quantitative information flow*. Quantitative information flow is a broad class of metrics that quantify the information leaked by a process by comparing the *information* contained before and after observing the outcome of the process. Quantitative Information Flow traces its information-theoretic roots to the work of Shannon [42] and Rényi [43]. Recent works have proposed measures for quantifying the security of information by measuring the amount of information leaked from inputs to outputs by certain variables; we point the reader to [44] for an overview, and to [45] for an exposition on information theory. Quantitative Information Flow is concerned with information leaks and therefore needs to account for correlations between inputs that may lead to leakage. The dual problem of transparency, on the other hand, requires us to destroy correlations while analyzing the outcomes of a system to identify the causal paths for information leakage.

C. Interpretable Machine Learning

An orthogonal approach to adding interpretability to machine learning is to constrain the choice of models to those that are interpretable by design. This can either proceed through regularization techniques such as Lasso [46] that attempt to pick a small subset of the most important features, or by using models that structurally match human reasoning such as Bayesian Rule Lists [47], Supersparse Linear Integer Models [48], or Probabilistic Scaling [49]. Since the choice of models in this approach is restricted, a loss in predictive accuracy is a concern, and therefore, the central focus in this line of work is the minimization of the loss in accuracy while maintaining interpretability. On the other hand, our approach to interpretability is forensic. We add interpretability to machine learning models after they have been learnt. As a result, our approach does not constrain the choice of models that can be used.

D. Experimentation on Web Services

There is an emerging body of work on systematic experimentation to enhance transparency into Web services such as targeted advertising [50], [51], [52], [53], [54]. The setting in this line of work is different since they have restricted access to the analytics systems through publicly available interfaces. As a result they only have partial control of inputs, partial observability of outputs, and little or no knowledge of input distributions. The intended use of these experiments is to enable external oversight into Web services without any cooperation. Our framework is more appropriate for a transparency mechanism where an entity proactively publishes transparency reports for individuals and groups. Our framework is also appropriate for use as an internal or external oversight tool with access to mechanisms with control and knowledge of input distributions, thereby forming a basis for testing.

E. Game-Theoretic Influence Measures

Recent years have seen game-theoretic influence measures used in various settings. Datta et al. [55] also define a measure for quantifying feature influence in classification tasks. Their measure does not account for the prior on the data, nor does it use interventions that break correlations between sets of features. In the terminology of this paper, the quantity of interest used by [55] is the ability of changing the outcome by changing the state of a feature. This work greatly extends and generalizes the concepts presented in [55], by both accounting for interventions on sets, and by generalizing the notion of influence to include a wide range of system behaviors, such as group disparity, group outcomes and individual outcomes.

Game theoretic measures have been used by various research disciplines to measure influence. Indeed, such measures are relevant whenever one is interested in measuring the marginal contribution of variables, and when sets of variables are able to cause some measurable effect. Lindelauf et al. [56] and Michalak et al. [57] use game theoretic influence measures on graph-based games in order to identify key members of terrorist networks. Del Pozo et al. [58] and Michalak et al. [59] use similar ideas for identifying important members of large social networks, providing scalable algorithms for influence computation. Bork et al. [60] use the Shapley value to assign importance to protein interactions in large, complex biological interaction networks; Keinan et al. [61] employ the Shapley value in order to measure causal effects in neurophysical models. The novelty in our use of the game theoretic power indices lies in the conception of a cooperative game via a valuation function $\iota(S)$, defined by an randomized intervention

on inputs S. Such an intervention breaks correlations and allows us to compute marginal causal influences on a wide range of system behaviors.

X. Conclusion & Future Work

In this paper, we present QII, a general family of metrics for quantifying the influence of inputs in systems that process personal information. In particular, QII lends insights into the behavior of opaque machine learning algorithms by allowing us to answer a wide class of transparency queries ranging from influence on individual causal outcomes to influence on disparate impact. To achieve this, QII breaks correlations between inputs to allow causal reasoning, and computes the marginal influence of inputs in situations where inputs cannot affect outcomes alone. Also, we demonstrate that QII can be efficiently approximated, and can be made differentially private with negligible noise addition in many cases.

An immediate next step in this line of work is to explore adoption strategies in the many areas that use personal information to aid decision making. Areas such as healthcare [3], predictive policing [1], education [4], and defense [5] all have a particularly acute need for transparency in their decision making. It is likely that specific applications will guide us in our choice of a QII metric that is appropriate for that scenario, which includes a choice for our game-theoretic power index.

We have not considered situations where inputs do not have well understood semantics. Such situations arise often in settings such as image or speech recognition, and automated video surveillance. With the proliferation of immense processing power, complex machine learning models such as deep neural networks have become ubiquitous in these domains. Defining transparency and developing analysis techniques in such settings is important future work.

References

[1] W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation, 2013.

[2] T. Alloway, "Big data: Credit where credits due," http://www.ft.com/cms/s/0/7933792e-a2e6-11e4-9c06-00144feab7de.html.

[3] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," http://jama.jamanetwork.com/article.aspx?articleid=1674245.

[4] "Big data in education," https://www.edx.org/course/big-data-education-teacherscollegex-bde1x.

[5] "Big data in government, defense and homeland security 2015 - 2020," http://www.prnewswire.com/news-releases/big-data-in-government-defense-and-homeland-security-2015---2020.html.

[6] J. Podesta, P. Pritzker, E. Moniz, J. Holdern, and J. Zients, "Big data: Seizing opportunities, preserving values," Executive Office of the President - the White House, Tech. Rep., May 2014.

[7] "E.G. Griggs v. Duke Power Co., 401 U.S. 424, 91 S. Ct. 849, 28 L. Ed. 2d 158 (1977)."

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Third Conference on Theory of Cryptography*, ser. TCC'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 265–284. [Online]. Available: http://dx.doi.org/10.1007/11681878_14

[9] S. Kasiviswanathan, H. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" in *Proceedings of the 49th IEEE Symposion on Foundations of Computer Science (FOCS 2008)*, Oct 2008, pp. 531–540.

[10] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[11] "National longitudinal surveys," http://www.bls.gov/nls/.

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012)*, 2012, pp. 214–226.

[13] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[14] L. Shapley, "A value for n-person games," in *Contributions to the Theory of Games, vol. 2*, ser. Annals of Mathematics Studies, no. 28. Princeton University Press, 1953, pp. 307–317.

[15] M. Maschler, E. Solan, and S. Zamir, *Game Theory*. Cambridge University Press, 2013.

[16] L. S. Shapley and M. Shubik, "A method for evaluating the distribution of power in a committee system," *The American Political Science Review*, vol. 48, no. 3, pp. 787–792, 1954.

[17] J. Banzhaf, "Weighted voting doesn't work: a mathematical analysis," *Rutgers Law Review*, vol. 19, pp. 317–343, 1965.

[18] J. Deegan and E. Packel, "A new index of power for simple n-person games," *International Journal of Game Theory*, vol. 7, pp. 113–123, 1978.

[19] H. Young, "Monotonic solutions of cooperative games," *International Journal of Game Theory*, vol. 14, no. 2, pp. 65–72, 1985.

[20] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[21] G. Chalkiadakis, E. Elkind, and M. Wooldridge, *Computational Aspects of Cooperative Game Theory*. Morgan and Claypool, 2011.

[22] Y. Bachrach, E. Markakis, E. Resnick, A. Procaccia, J. Rosenschein, and A. Saberi, "Approximating power indices: theoretical and empirical analysis," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 2, pp. 105–122, 2010.

[23] S. Maleki, L. Tran-Thanh, G. Hines, T. Rahwan, and A. Rogers, "Bounding the estimation error of sampling-based shapley value approximation with/without stratifying," *CoRR*, vol. abs/1306.4265, 2013.

[24] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963. [Online]. Available: http://www.jstor.org/stable/2282952?

[25] N. Li, W. H. Qardaji, and D. Su, "Provably private data anonymization: Or, k-anonymity meets differential privacy," *CoRR*, vol. abs/1101.2604, 2011. [Online]. Available: http://arxiv.org/abs/1101.2604

[26] Z. Jelveh and M. Luca, "Towards diagnosing accuracy loss in discrimination-aware classification: An application to predictive policing," *Fairness, Accountability and Transparency in Machine Learning*, vol. 26, no. 1, pp. 137–141, 2014.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[28] S. Barocas and H. Nissenbaum, "Big data's end run around procedural privacy protections," *Communications of the ACM*, vol. 57, no. 11, pp. 31–33, Oct. 2014.

[29] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10618-010-0190-x

[30] A. Datta, M. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," in *Proceedings on Privacy Enhancing Technologies (PoPETs 2015)*, 2015, pp. 92–112.

[31] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW 2011)*, 2011, pp. 643–650.

[32] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, 2013, pp. 325–333.

[33] G. W. University, "Standardized test scores will be optional for gw applicants," 2015. [Online]. Available: https://gwtoday.gwu.edu/standardized-test-scores-will-be-optional-gw-applicants

[34] The National Center for Fair and Open Testing, "850+ colleges and universities that do not use sat/act scores to admit substantial numbers of students into bachelor degree programs," 2015. [Online]. Available: http://www.fairtest.org/university/optional

[35] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf, "Quantifying causal influences," *Ann. Statist.*, vol. 41, no. 5, pp. 2324–2358, 10 2013.

[36] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944968

[37] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[38] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009.

[39] J. Tian and J. Pearl, "Probabilities of causation: Bounds and identification," *Annals of Mathematics and Artificial Intelligence*, vol. 28, no. 1-4, pp. 287–313, 2000.

[40] J. Halpern and J. Pearl, "Causes and explanations: A structural-model approach. part i: Causes," *The British journal for the philosophy of science*, vol. 56, no. 4, pp. 843–887, 2005.

[41] H. Chockler and J. Halpern, "Responsibility and blame: A structural-model approach," *Journal of Artificial Intelligence Research*, vol. 22, pp. 93–115, 2004.

[42] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. [Online]. Available: http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x

[43] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1961, pp. 547–561. [Online]. Available: http://projecteuclid.org/euclid.bsmsp/1200512181

[44] G. Smith, "Quantifying information flow using min-entropy," in *Proceedings of the 8th International Conference on Quantitative Evaluation of Systems (QEST 2011)*, 2011, pp. 159–167.

[45] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[46] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society Series B*, vol. 73, no. 3, pp. 273–282, 2011. [Online]. Available: http://EconPapers.repec.org/RePEc:bla:jorssb:v:73:y:2011:i:3:p:273-282

[47] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Stat.*, vol. 9, no. 3, pp. 1350–1371, 09 2015. [Online]. Available: http://dx.doi.org/10.1214/15-AOAS848

[48] B. Ustun, S. Trac, and C. Rudin, "Supersparse linear integer models for interpretable classification," *ArXiv e-prints*, 2013. [Online]. Available: http://arxiv.org/pdf/1306.5860v1

[49] S. Rping, "Learning interpretable models." Ph.D. dissertation, Dortmund University of Technology, 2006, http://d-nb.info/997491736.

[50] S. Guha, B. Cheng, and P. Francis, "Challenges in measuring online advertising systems," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '10. New York, NY, USA: ACM, 2010, pp. 81–87.

[51] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: ACM, 2014, pp. 597–608.

[52] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "Xray: Enhancing the web's transparency with differential correlation," in *Proceedings of the 23rd USENIX Conference on Security Symposium*, ser. SEC'14. Berkeley, CA, USA: USENIX Association, 2014, pp. 49–64.

[53] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings." *PoPETs*, vol. 2015, no. 1, pp. 92–112, 2015.

[54] M. Lecuyer, R. Spahn, Y. Spiliopolous, A. Chaintreau, R. Geambasu, and D. Hsu, "Sunlight: Fine-grained targeting detection at scale with statistical confidence," in *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15. New York, NY, USA: ACM, 2015, pp. 554–566.

[55] A. Datta, A. Datta, A. Procaccia, and Y. Zick, "Influence in classification via cooperative game theory," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015, pp. 511–517.

[56] R. Lindelauf, H. Hamers, and B. Husslage, "Cooperative game theoretic centrality analysis of terrorist networks: The cases of jemaah islamiyah and al qaeda." *European Journal of Operational Research*, vol. 229, no. 1, pp. 230–238, 2013.

[57] T. Michalak, T. Rahwan, P. Szczepanski, O. Skibski, R. Narayanam, M. Wooldridge, and N. Jennings, "Computational analysis of connectivity games with applications to the investigation of terrorist networks," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, 2013, pp. 293–301.

[58] M. del Pozo, C. Manuel, E. González-Arangüena, and G. Owen, "Centrality in directed social networks. a game theoretic approach," *Social Networks*, vol. 33, no. 3, pp. 191–200, 2011.

[59] T. Michalak, K. Aaditha, P. Szczepanski, B. Ravindran, and N. Jennings, "Efficient computation of the shapley value for game-theoretic network centrality," *Journal of Artificial Intelligence Research*, vol. 46, pp. 607–650, 2013.

[60] P. Bork, L. Jensen, C. von Mering, A. Ramani, I. Lee, and E. Marcott, "Protein interaction networks from yeast to human," *Current Opinions in Structural Biology*, vol. 14, no. 3, pp. 292–299, 2004.

[61] A. Keinan, B. Sandbank, C. Hilgetag, I. Meilijson, and E. Ruppin, "Fair attribution of functional contribution in artificial and biological networks," *Neural Computation*, vol. 16, no. 9, pp. 1887–1915, September 2004.

[62] M. Malawski, "Equal treatment, symmetry and banzhaf value axiomatizations," *International Journal of Game Theory*, vol. 31, no. 1, pp. 47–67, 2002.

APPENDIX A
ALTERNATIVE GAME-THEORETIC INFLUENCE MEASURES

In what follows, we describe two alternatives to the Shapley value used in this work. The Shapley value makes intuitive sense in our setting, as we argue in Section III-B. However, other measures may be appropriate for certain input data generation processes. In what follows we revisit the Banzhaf index, briefly discussed in Section III-A, and introduce the readers to the *Deegan-Packel index*, a game-theoretic influence measure with deep connections to a formal theory of responsibilty and blame [41].

A. The Banzhaf Index

Recall that the Banzhaf index, denoted $\beta_i(N, v)$ is defined as follows:

$$\beta_i(N, v) = \frac{1}{2^{n-1}} \sum_{S \subseteq N \setminus \{i\}} m_i(S).$$

The Banzhaf index can be thought of as follows: each $j \in N \setminus \{i\}$ will join a work effort with probability $\frac{1}{2}$ (or, equivalently, each $S \subseteq N \setminus \{i\}$ has an equal chance of forming); if i joins as well, then its expected marginal contribution to the set formed is exactly the Banzhaf index. Note the marked difference between the probabilistic models: under the Shapley value, we sample *permutations* uniformly at random, whereas under the regime of the Banzhaf index, we sample sets uniformly at random. The different sampling protocols reflect different normative assumptions. For one, the Banzhaf index is not guaranteed to be efficient; that is, $\sum_{i \in N} \beta_i(N, v)$ is not necessarily equal to $v(N)$, whereas it is always the case that $\sum_{i=1}^{n} \varphi_i(N, v) = v(N)$. Moreover, the Banzhaf index is more biased towards measuring the marginal contribution of i to sets of size $\frac{n}{2} \pm O(\sqrt{n})$; this is because the expected size of a randomly selected set follows a binomial distribution $B(n, \frac{1}{2})$. On the other hand, the Shapley value is equally likely to measure the marginal contribution of i to sets of any size $k \in \{0, \ldots, k\}$, as i is equally likely to be in any one position in a randomly selected permutation σ (and, in particular, the the set of i's predecessors in σ is equally likely to have any size $k \in \{0, \ldots, n-1\}$).

Going back to the QII setting, the difference in sampling procedure is not merely an interesting anecdote: it is a significant modeling choice. Intuitively, the Banzhaf index is more appropriate if we assume that large sets of features would have a significant influence on outcomes, whereas the Shapley value is more appropriate if we assume that even small sets of features might cause significant effects on the outcome. Indeed, as we mention in Section VIII, aggregating the marginal influence of i over sets is a significant modeling choice; while using the measures proposed here is perfectly reasonable in many settings, other aggregation methods may be applicable in others.

Unlike the Shapley value, the Banzhaf index is not guaranteed to be efficient (although it does satisfy the symmetry and dummy properties). Indeed, [62] shows that replacing the efficiency axiom with an alternative axiom, uniquely characterizes the Banzhaf index; the axiom, called *2-efficiency*, prescribes the behavior of an influence measure when two players merge. First, let us define a *merged game*; given a game $\langle N, v \rangle$, and two players $i, j \in N$, we write $T = \{i, j\}$. We define the game $\bar{v}$ on $N \setminus T \cup \{\bar{t}\}$ as follows: for every set $S \subseteq N \setminus \{i, j\}$, $\bar{v}(S) = v(S)$, and $\bar{v}(S \cup \{\bar{t}\}) = v(S \cup \{i, j\})$, note that the added player $\bar{t}$ represents the two players i and j who are now acting as one. The 2-Efficiency axiom states that influence should be invariant under merges.

Definition 14 (2-Efficiency (2-EFF))**.** Given two players $i, j \in N$, let $\bar{v}$ be the game resulting from the merge of i and j into a single player $\bar{t}$; an influence measure ϕ satisfies 2-Efficiency if $\phi_i(N, v) + \phi_j(N, v) = \phi_{\bar{t}}(N \setminus \{i, j\} \cup \{\bar{t}\}, \bar{v})$.

Theorem 15 ([62])**.** *The Banzhaf index is the only function to satisfy (Sym), (D), (Mono) and (2-EFF).*

In our context, 2-Efficiency can be interpreted as follows: suppose that we artificially treat two features i and j as one, keeping all other parameters fixed; in this setting, 2-efficiency means that the influence of merged features equals the influence they had as separate entities.

B. The Deegan-Packel Index

Finally, we discuss the *Deegan-Packel index* [18]. While the Shapley value and Banzhaf index are well-defined for any coalitional game, the Deegan-Packel index is only defined for *simple games*. A cooperative game is said to be simple if $v(S) \in \{0, 1\}$ for all $S \subseteq N$. In our setting, an influence measure would correspond to a simple game if it is binary (e.g. it measures some threshold behavior, or corresponds to a binary classifier). The binary requirement is rather strong; however, we wish to draw the reader's attention to the Deegan-Packel index, as it has an interesting connection to *causal responsibility* [41], a variant of the classic Pearl-Halpern causality model [40], which aims to measure the degree to which a single variable causes an outcome.

Given a simple game $v : 2^N \to \{0, 1\}$, let $\mathcal{M}(v)$ be the set of *minimal winning coalitions*; that is, for every $S \in \mathcal{M}(v)$, $v(S) = 1$, and $v(T) = 0$ for every strict subset of S. The Deegan-Packel index assigns a value of

$$\delta_i(N, v) = \frac{1}{|\mathcal{M}(v)|} \sum_{S \in \mathcal{M}(v) : i \in S} \frac{1}{|S|}.$$

The intuition behind the Deegan-Packel index is as follows: players will not form coalitions any larger than what they absolutely have to in order to win, so it does not make sense to measure their effect on non-minimal winning coalitions. Furthermore, when a minimal winning coalition is formed, the benefits from its formation are divided equally among its members; in particular, small coalitions confer a greater benefit for those forming them than large ones. The Deegan-Packel index measures the expected payment one receives, assuming that every minimal winning coalition is equally likely to form. Interestingly, the Deegan-Packel index corresponds nicely to the notion of responsibility and blame described in [41].

Suppose that we have a set of variables $X_1, \ldots, X_n$ set to $x_1, \ldots, x_n$, and some binary effect $f(x_1, \ldots, x_n)$ (written as $f(\mathbf{x})$) occurs (say, $f(\mathbf{x}) = 1$). To establish a causal relation between the setting of X_i to x_i and $f(\mathbf{x}) = 1$, [40] require that there is some set $S \subseteq N \setminus \{i\}$ and some values $(y_j)_{j \in S \cup \{i\}}$ such that $f(\mathbf{x}_{-S \cup \{i\}}, (y_j)_{j \in S \cup \{i\}}) = 0$, but $f(\mathbf{x}_{-S}, (y_j)_{j \in S}) = 1$. In words, an intervention on the values of both S and i may cause a change in the value of f, but performing the same intervention just on the variables in S would not cause such a change. This definition is at the heart of the marginal contribution approach to interventions that we describe in Section III-A. [41] define the responsibility of i for an outcome as $\frac{1}{k+1}$, where k is the size of the smallest set S for which the causality definition holds with respect to i. The Deegan-Packel index can thus be thought of as measuring a similar notion: instead of taking the overall minimal number of changes necessary in order to make i a direct, counterfactual cause, we observe all minimal sets that do so. Taking the average responsibility of i (referred to as *blame* in [41]) according to this variant, we obtain the Deegan-Packel index.

Example 16. Let us examine the following setup, based on Example 3.3 in [41]. There are $n = 2k + 1$ voters (n is an odd number) who must choose between two candidates, Mr. B and Mr. G ([41] describe the setting with $n = 11$). All voters elected Mr. B, resulting in an n-0 win. It is natural to ask: how responsible was voter i for the victory of Mr. B? According to [41], the degree of responsibility of each voter is $\frac{1}{k+1}$. It will require that i and k additional voters change their vote in order for the outcome to change. Modeling this setup as a cooperative game is quite natural: the voters are the players $N = \{1, \ldots, n\}$; for every subset $S \subseteq N$ we have

$$v(S) = \begin{cases} 1 & \text{if } |S| \geq k+1 \\ 0 & \text{otherwise.} \end{cases}$$

That is, $v(S) = 1$ if and only if the set S can change the outcome of the election. The minimal winning coalitions here are the subsets of N of size $k + 1$, thus the Deegan-Packel index of player i is

$$\begin{aligned} \delta_i(N, v) =& \frac{1}{|\mathcal{M}(v)|} \sum_{S \in \mathcal{M}(v): i \in S} \frac{1}{|S|} \\ =& \frac{1}{\binom{n}{k+1}} \binom{n}{k} \frac{1}{k+1} = \frac{1}{n-k} = \frac{1}{k+1} \end{aligned}$$

We note that if one assumes that all voters are equally likely to prefer Mr. B over Mr. G, then the blame of voter i would be computed in the exact manner as the Deegan-Packel index.