What makes you unique?[*]

Benjamin B. Seiler

e-mail: bbseiler@stanford.edu

Masayoshi Mase

e-mail: masayoshi.mase.mh@hitachi.com

and

Art B. Owen

e-mail: owen@stanford.edu

Abstract: This paper proposes a uniqueness Shapley measure to compare the extent to which different variables are able to identify a subject. Revealing the value of a variable on subject t shrinks the set of possible subjects that t could be. The extent of the shrinkage depends on which other variables have also been revealed. We use Shapley value to combine all of the reductions in log cardinality due to revealing a variable after some subset of the other variables has been revealed. This uniqueness Shapley measure can be aggregated over subjects where it becomes a weighted sum of conditional entropies. Aggregation over subsets of subjects can address questions like how identifying is age for people of a given zip code. Such aggregates have a corresponding expression in terms of cross entropies. We use uniqueness Shapley to investigate the differential effects of revealing variables from the North Carolina voter registration rolls and in identifying anomalous solar flares. An enormous speedup (approaching 2000 fold in one example) is obtained by using the all dimension trees of Moore and Lee (1998) to store the cardinalities we need.

Keywords and phrases: Conditional entropy, identifiability, Shapley value.

1. Introduction

An individual data point, such as one representing a person, can often be identified by specifying even a small subset of its variables. For instance, a large fraction of US residents are uniquely identified by just their date of birth, zip code, and gender (Sweeney, 2000, Golle, 2006). The website https://amiunique.org/ will examine some signature variables in your browser and report whether you are uniquely identified among the millions of participants. See Gómez-Boix, Laperdrix and Baudry (2018) for a description.

Variables are not equally powerful for the purpose of identifying an individual, and the variables that provide the most information for one person might not be very informative for another. It can also happen that two or more variables specified together can be much more identifying than we might surmise given their individual strengths for identifying people. The joint specification can also be less informative due to associations between variables such as near duplicates.

In this paper we propose a way to measure how important a variable is for identifying one specific subject in a set of data. Our definition of importance is based on Shapley value from economic game theory (Shapley, 1953). An important variable is one that, when revealed for a subject, greatly reduces the number of subjects who could match it. This measure takes account of which other variables might also have been revealed, and so it depends on the full joint distribution in the data set not just one or a few marginal distributions of the data. It does not assume that any set of variables necessarily provides a unique identification of the subject, as might fail to happen for twins.

The game theoretic formulation provides a principled way to aggregate subject-specific importances to variable importance measures for the whole data set or for subsets of special interest, using the additivity property of Shapley value. The measure is an extension of the cohort Shapley measure that Mase, Owen and Seiler (2019) use to quantify variable importance for black box functions.

Other things being equal, a more identifying variable is one that is more worth concealing for privacy purposes, or more valuable for personalization. That said, our measure is not designed for settings like differential privacy (Dwork, 2008) where one seeks privacy guarantees. We use it instead for exploratory purposes.

An outline of the paper is as follows. Section 2 introduces our notation, reviews Shapley value and defines the uniqueness Shapley values of each input variable for a given subject. Variables with greater Shapley value are more identifying. Section 3 shows that the uniqueness Shapley value can be related to an entropy measure. When aggregated to an entire data set, the uniqueness Shapley value for variable j is a weighted sum of the conditional empirical entropies of variable j given all subsets of variables not including j. When we aggregate only over a proper subset of subjects the resulting Shapley value expression replaces entropies by cross-entropies linking the empirical distribution on the subset to the full data set. A naive implementation of aggregated uniqueness Shapley value will have a cost that is quadratic in the number of subjects. Section 4 describes the all dimension trees of Moore and Lee (1998) that we have found to give an enormous speedup making the difference between feasible and infeasible computation in some of our examples. Section 5 explores a solar flare dataset from Dua and Graff (2017). We treat solar regions with the most extreme and potentially dangerous flares as anomalies and then, as a step towards explaining those anomalies, look at which variables most identify them. Section 6 looks at voter registration data from the state of North Carolina. We compare the extent to which race, age, gender and other variables serve to identify voters. Section 7 gives conclusions and discusses some further issues.

2. Notation and background

We suppose that there are d categorical variables measured on each of n subjects. Subject i is described by a vector $\boldsymbol{x}_i$ with components $x_{ij} \in \mathcal{X}_j$ for $j = 1, \ldots, d$. The set of all subjects is denoted $1{:}n$ and the set of all variables is $1{:}d$.

To begin, suppose that there is a target subject $t \in 1{:}n$, and we want to know what variables identify subject t. For every subset of variables $u \subseteq 1{:}d$ let $\boldsymbol{x}_{iu}$ be the tuple $(x_{ij})_{j \in u}$. Then we define the cohort

$$C(u) = C_t(u) = \{i \in 1{:}n \mid \boldsymbol{x}_{iu} = \boldsymbol{x}_{tu}\},$$

of all subjects i who match subject t on every variable in the set u. By convention, $C_t(\varnothing) = 1{:}n$ and no cohort is empty because they all include t. The size of a cohort is the cardinality

$$N(u) = N_t(u) = \sum_{i=1}^{n} 1\{i \in C_t(u)\}.$$

We consider variable j to be important for identifying subject t if $N_t(u \cup \{j\})$ is typically much smaller than $N_t(u)$ for $u \subseteq 1{:}d \setminus \{j\}$. In that case, knowledge of x_{tj} refines the cohort containing subject t by a large factor. There are 2^{d-1} different cohorts into which $\{j\}$ might be included, and we will use Shapley value to incorporate them all.

To simplify some of our expressions, we introduce some notational short forms. The set $1{:}d \setminus \{j\}$ is written as $-j$. When $j \notin u \subset 1{:}d$ then we write $u + j$ for $u \cup \{j\}$. The set of all subsets of $1{:}d$ is written $2^{1:d}$. When u is a finite set, then $|u|$ is its cardinality.

2.1. Shapley value

The Shapley value from game theory can be used to allocate value to the members of a team that produced the value. In our context, the team will be made up of variables whose values are specified, and the value will be defined by how much subject t is identified.

We work with a value function $\mathrm{val} \colon 2^{1:d} \to \mathbb{R}$ where $\mathrm{val}(u)$ is the value created by the team u, and we suppose that we are given $\mathrm{val}(u)$ for all $u \subseteq 2^{1:d}$. The total value created by the team is $\mathrm{val}(1{:}d)$, and the problem is to make a fair allocation to members $j = 1, \ldots, d$. The fair share for member j is denoted by ϕ_j. Shapley (1953) had these axioms:

1) (Efficiency) $\sum_{j=1}^{d} \phi_j = \mathrm{val}(1{:}d)$,
2) (Symmetry) if $\mathrm{val}(u + j) = \mathrm{val}(u + j')$ whenever $u \subseteq 1{:}d \setminus \{j, j'\}$ then $\phi_j = \phi_{j'}$,
3) (Dummy) if $\mathrm{val}(u + j) = \mathrm{val}(u)$ whenever $u \subseteq -j$ then $\phi_j = 0$ and
4) (Additivity) if two games have value functions val and val' and shares ϕ_j and ϕ_j' then the game with values $\mathrm{val}(u) + \mathrm{val}'(u)$ must have shares $\phi_j + \phi_j'$.

The fair shares depend strongly on the incremental value of variable j given that variables $u \subseteq -j$ are already included. It is convenient to use

$$\text{val}(j \,|\, u) \equiv \text{val}(u + j) - \text{val}(u) \tag{1}$$

to represent this incremental value. Shapley (1953) finds that there is a unique set of shares ϕ_j (called Shapley values) that satisfy his four axioms. They are

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \text{val}(j \,|\, u). \tag{2}$$

Another way to describe ϕ_j is to build a set from $\varnothing$ to 1:d by adding the variables $j \in 1{:}d$ in a random order. At some point variable j appears with a set $u \subseteq -j$ of previously introduced variables. Then ϕ_j is the average of $\text{val}(j \,|\, u)$ taken over all $d!$ variable orders.

We see from equation (2) that only value differences affect $\phi_1, \ldots, \phi_d$. It is often convenient to take $\text{val}(\varnothing) = 0$. If that does not hold we can replace every $\text{val}(u)$ by $\text{val}(u) - \text{val}(\varnothing)$ without changing any of the ϕ_j.

2.2. Uniqueness Shapley

The value function we choose for identifying subject t is

$$\text{val}(u) = -\log_2\Big(\frac{N_t(u)}{N_t(\varnothing)}\Big). \tag{3}$$

This definition satisfies $\text{val}(\varnothing) = 0$. The smaller the cardinality $N_t(u)$ of $C_t(u)$, the more $\boldsymbol{x}_{t,u}$ has served to identify subject t. One unit of $\text{val}(\cdot)$ corresponds to information that halves the size of the cohort containing subject t. We quantify the importance of x_{tj} to the identifiability of subject t via the Shapley value $\phi_j = \phi_{t,j}$ derived from the value function in (3). The extent to which revealing x_{tj} identifies subject t depends on any previously identified variables $\boldsymbol{x}_{tu}$ for $u \subseteq -j$. The uniqueness Shapley value combines all 2^{d-1} of these contributions in a way consistent with game theory. Those contributions take the form

$$\text{val}(j \,|\, u) = \log_2\Big(\frac{N_t(u)}{N_t(u + j)}\Big)$$

after cancellation of $\log_2(N_t(\varnothing))$.

The uniqueness Shapley value function is the cohort Shapley value function of Mase, Owen and Seiler (2019) after the within-cohort average of a response variable is replaced by the cardinality of the cohort.

Proposition 1. *The uniqueness Shapley value $\phi_{t,j}$ satisfies $\phi_{t,j} \geqslant 0$ with $\phi_{t,j} = 0$ if and only if $x_{ij} = x_{tj}$ for all $i = 1, \ldots, n$.*

Proof. If $j \notin u$, then $N_t(u + j) \leqslant N_t(u)$ and from this we find that $\phi_{t,j} \geqslant 0$. If $x_{ij} = x_{tj}$ for all i then $C_t(u + j) = C_t(u)$ for $j \notin u$ making $N_t(u) = N_t(u + j)$

and $\text{val}(u+j) - \text{val}(u) = 0$. Conversely, suppose that $\phi_{t,j} = 0$ but $x_{ij} \neq x_{tj}$ for some i. Then $N_t(\{j\}) > N_t(\varnothing)$ from which $\text{val}(\{j\} \mid \varnothing) > 0$. This provides a contradiction because there cannot be any compensating negative value differences to bring the Shapley value down to zero. $\qquad\square$

Now suppose that we want to quantify the importance of variable j to the whole set of subjects. The additivity axiom of Shapley value makes it natural to sum those values. For interpretability we scale that sum to an average over subjects taking

$$\phi_j^{1:n} = \frac{1}{n} \sum_{t=1}^{n} \phi_{t,j} \tag{4}$$

as our global measure of the cardinality importance of variable j.

For an arbitrary non-empty subset $v \subseteq 1{:}n$ of subjects we can also define

$$\phi_j^v = \frac{1}{|v|} \sum_{t \in v} \phi_{t,j}. \tag{5}$$

Suppose for instance that x_{ij} encodes the gender of subject i. We can then define a set v consisting of all the subjects with one of the genders in the data set and then ϕ_j^v describes the extent to which gender identifies people of the given gender. This is not necessarily zero even though gender is constant over $i \in v$ because the Shapley values $\phi_{t,j}$ are defined on the entire subject set $1{:}n$. For some other feature j' such as age in years we can then measure the extent to which j' identifies subjects of a given gender and see how this varies as we change the set v to each gender in turn.

3. Relationship to information theory

It is natural to consider entropy as a measure of how informative a feature is for identification. Gómez-Boix, Laperdrix and Baudry (2018) report entropy values for individual variables in their browser fingerprint data. Here we introduce some information theoretic quantities and show that uniqueness Shapley value aggregated over subjects is equivalent to entropy when the features are independent. More generally, aggregating the uniqueness Shapley measure yields a linear combination of conditional entropies. Aggregates over proper subsets of subjects involve cross-entropies.

For a categorical variable $\boldsymbol{x} \in \mathcal{X}$ with $\boldsymbol{x} \sim p$ we write the entropy of both $\boldsymbol{x}$ and p as $H(\boldsymbol{x}) = H(p) = -\sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) \log_2(p(\boldsymbol{x}))$. For disjoint $u, \tilde{u} \subset 1{:}d$ the conditional entropy of $\boldsymbol{x}_u$ given $\boldsymbol{x}_{\tilde{u}}$ is

$$H(\boldsymbol{x}_u \mid \boldsymbol{x}_{\tilde{u}}) = \sum_{\boldsymbol{z}_{\tilde{u}} \in \mathcal{X}_{\tilde{u}}} p(\boldsymbol{z}_{\tilde{u}}) H(\boldsymbol{x}_u \mid \boldsymbol{x}_{\tilde{u}} = \boldsymbol{z}_{\tilde{u}}), \quad \text{where}$$

$$H(\boldsymbol{x}_u \mid \boldsymbol{x}_{\tilde{u}} = \boldsymbol{z}_{\tilde{u}}) = - \sum_{\boldsymbol{x}_u \in \mathcal{X}_u} p(\boldsymbol{x}_u \mid \boldsymbol{x}_{\tilde{u}} = \boldsymbol{z}_{\tilde{u}}) \log_2(p(\boldsymbol{x}_u \mid \boldsymbol{x}_{\tilde{u}} = \boldsymbol{z}_{\tilde{u}})).$$

By the chain rule for entropy (Cover and Thomas, 2006, Theorem 2.2.1)

$$H(\boldsymbol{x}_u \mid \boldsymbol{x}_{\tilde{u}}) = H(\boldsymbol{x}_{u \cup \tilde{u}}) - H(\boldsymbol{x}_u).$$

We will work with the entropy of sub-vectors of $\boldsymbol{x}$ and for this we write $\mathcal{H}(u) = \mathcal{H}(u; \boldsymbol{x}) = H(\boldsymbol{x}_u)$ when the distribution of $\boldsymbol{x}_u$ is understood from context. Similarly, $\mathcal{H}(u \mid \tilde{u})$ denotes $H(\boldsymbol{x}_u \mid \boldsymbol{x}_{\tilde{u}})$. For $j \notin u$ we may abbreviate the conditional entropy $\mathcal{H}(\{j\} \mid u)$ to $\mathcal{H}(j \mid u)$.

3.1. Relationship to entropy

Let $p(\cdot)$ be the empirical distribution on $\mathcal{X}$ with

$$p(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} 1\{\boldsymbol{x} = \boldsymbol{x}_i\}. \tag{6}$$

For $j \in 1{:}d$, let p_j be the marginal distribution of $\boldsymbol{x}_j$ under (6) and for $u \subseteq 1{:}d$ let p_u be the marginal distribution of $\boldsymbol{x}_u$ under (6).

We say that two or more variables are independent in the data if they are independent random variables under (6). Exact independence is quite unlikely to occur but it provides an interpretable baseline via Shannon's entropy.

Proposition 2. *Suppose that $\boldsymbol{x}_{\{j\}}$ is independent of $\boldsymbol{x}_{-j}$ under* (6). *Then*

$$\phi_j^{1:n} = \mathcal{H}(\{j\}).$$

for $\boldsymbol{x}$ with the empirical distribution (6).

Proof. Consider subject t. Because $\boldsymbol{x}_{\{j\}}$ is independent of $\boldsymbol{x}_{-j}$ we find that $\boldsymbol{z}_{t,\{j\}}$ is independent of $\boldsymbol{z}_{t,-j}$. This means that for all $u \subseteq -j$,

$$\frac{N_t(u+j)}{N_t(u)} = p_j(x_{tj})$$

and then $\phi_{t,j} = -\log_2(p_j(x_{tj}))$ based on its expression as an average over permutations of incremental values. Now aggregating over subjects,

$$\phi_j^{1:n} = \frac{1}{n} \sum_{t=1}^{n} -\log_2(p_j(x_{tj})) = -\sum_{x \in \mathcal{X}_j} p_j(x) \log_2(p_j(x)). \qquad \square$$

As usual, the proper interpretation of $0\log_2(0)$ is zero. If all the variables are independent, then they all have a uniqueness Shapley value equal to their entropy. The connection to entropy goes further.

Proposition 3. *The global uniqueness Shapley value for variable j is*

$$\phi_j^{1:n} = \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \mathcal{H}(j \mid u) \tag{7}$$

where the conditional entropies are computed for a random vector $\boldsymbol{x}$ with the empirical distribution (6).

Proof. For $t \in 1{:}n$ we have $N_t(u)/N_t(\varnothing) = N_t(u)/n = p_u(\boldsymbol{x}_u)$. Then

$$\mathrm{val}^{1:n}(u) = \frac{1}{n} \sum_{t=1}^{n} -\log_2(N_t(u)/n) = \mathcal{H}(u) + \log_2(n)$$

and so $\mathrm{val}^{1:n}(j \mid u) = \mathrm{val}^{1:n}(u+j) - \mathrm{val}^{1:n}(u) = \mathcal{H}(u+j) - \mathcal{H}(u) = \mathcal{H}(j \mid u)$. $\square$

Because conditional entropies are non-negative and $\mathrm{val}(1{:}d)$ is the entropy of $\boldsymbol{x}$ under p, we have the bracketing inequality

$$\frac{\mathcal{H}(\{j\})}{d} \leqslant \phi_j^{1:n} \leqslant \mathcal{H}(1{:}d). \tag{8}$$

Noting that $\mathcal{H}(\varnothing) = 0$ we find for $d = 2$ that

$$\phi_1^{1:n} = \frac{1}{2}\mathcal{H}(\{1\}) + \frac{1}{2}\mathcal{H}(\{1\} \mid \{2\}) \tag{9}$$

with $\phi_2^{1:n}$ found by switching indices. For $d = 3$,

$$\phi_1^{1:n} = \frac{1}{3}\mathcal{H}(\{1\}) + \frac{1}{6}\mathcal{H}(\{1\} \mid \{2\}) + \frac{1}{6}\mathcal{H}(\{1\} \mid \{3\}) + \frac{1}{3}\mathcal{H}(\{1\} \mid \{2,3\}).$$

It may seem counterintuitive that larger entropy corresponds to greater power to identify subjects. If a variable takes two levels, say 0 and 1, then the distribution with greatest entropy is the one that gives them each probability 0.5. Revealing that variable provides '1 bit' of cohort reduction. If instead there is a 90:10 split for some variable then 10% of the population find that their cohort size is greatly reduced by 10-fold ($\log_2(10) \approx 3.3$) but 90% find their cohort size reduced by the much lower amount, $1/0.9$. This is about 11% and $\log_2(1/0.9) \approx 0.15$, so the average number of bits is $0.1 \times 3.3 + 0.9 \times 0.15 \approx 0.46$.

The largest possible global uniqueness Shapley value for a binary predictor variable x_j is $\phi_j^{1:n} = -\log_2(1/2) = 1$. In the solar flare example of Section 5 we will see $\phi_j^v > 1$ for a binary predictor x_j and a set v of anomalous observations.

3.2. Relationship to cross entropy

For two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ on a discrete set $\mathcal{X}$ the relative entropy (Kullback-Leibler distance) from p to q is

$$D(q \parallel p) = \sum_{x \in \mathcal{X}} q(\boldsymbol{x}) \log_2 \Big(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\Big).$$

Similarly, the cross entropy of p relative to q is

$$H(p, q) = -\sum_{\boldsymbol{x} \in \mathcal{X}} q(\boldsymbol{x}) \log_2(p(\boldsymbol{x})) = H(q) + D(q \parallel p).$$

It is very common to write these expressions with symbols p and q reversed, but in our setting, the second argument needs to be the empirical distribution from (6) that we have labeled p. For distribution q, we use marginal distributions q_j, q_u and q_{u+j} analogous to the quantities that we have used previously for p.

Proposition 4. *The uniqueness Shapley values for subset v of subjects at* (5) *can be written*

$$\phi_j^v = \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \left(H(p_{u+j}, q_{u+j}) - H(p_u, q_u) \right) \tag{10}$$

where q is the uniform distribution on $\boldsymbol{x}_t$ for $t \in v$.

Proof. First, by definition

$$\phi_j^v = \frac{1}{|v|} \sum_{t \in v} \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \mathrm{val}(j \,|\, u)$$

$$= \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \frac{1}{|v|} \sum_{t \in v} \log_2 \left(\frac{N_t(u)}{n_t(u+j)} \right).$$

Next, for a set $w \subseteq 1{:}d$,

$$H(p_w, q_w) = - \sum_{\boldsymbol{x} \in \mathcal{X}} q_w(\boldsymbol{x}) \log_2(p_w(\boldsymbol{x})).$$

This sum can be rewritten over $t \in v$ as long as we divide out the multiplicity for each $\boldsymbol{x}$ among the $\boldsymbol{x}_t$ for $t \in v$ yielding

$$H(p_w, q_w) = - \sum_{t \in v} \frac{1}{N_t^v(w)} q_w(\boldsymbol{x}_t) \log_2(p_w(\boldsymbol{x}_t)).$$

Here $N_t^v(w) = \sum_{i \in v} 1\{i \in C_t(w)\}$ is the number of subjects in v that match $\boldsymbol{x}_t$ on the features in w. As q_w is the uniform distribution over $\boldsymbol{x}_t$ for $t \in v$, $q_w(\boldsymbol{x}_t) = (N_t^v(w)/|v|)$, i.e., the proportion of subjects in v that match $\boldsymbol{x}_t$ on the features in w. Therefore

$$H(p_w, q_w) = - \sum_{t \in v} \frac{1}{|v|} \log_2 \left(\frac{N_t(w)}{n} \right).$$

With this formulation we can see that

$$\frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \left(H(p_{u+j}, q_{u+j}) - H(p_u, q_u) \right)$$

$$= \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \left(\left(-\sum_{t \in v} \frac{1}{|v|} \log_2 \left(\frac{N_t(u+j)}{n} \right) \right) - \left(-\sum_{t \in v} \frac{1}{|v|} \log_2 \left(\frac{N_t(u)}{n} \right) \right) \right)$$

$$= \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \frac{1}{|v|} \sum_{t \in v} \log_2 \left(\frac{N_t(u)}{n_t(u+j)} \right) = \phi_j^v. \qquad \square$$

3.3. Duplicate and redundant variables

If some other variable j' is equivalent to variable j, then it must have the same uniqueness Shapley value. The extreme version of this is that the second variable could have been copy-pasted from the first one by accident. More plausibly there could be two variables like a person's email address and cell phone number that have very nearly a one to one relationship in a data set of transactions.

Let's look at $\phi_1^{1:3}$ in the special case where all $x_{i3} = x_{i1}$. Then $\mathcal{H}(\{1,3\}) = \mathcal{H}(\{1\}) = \mathcal{H}(\{3\})$ and $\mathcal{H}(\{1,2\}) = \mathcal{H}(\{2,3\}) = \mathcal{H}(\{1,2,3\})$ and we find that

$$\phi_1^{1:s} = \frac{1}{3}\mathcal{H}(\{1\}) + \frac{1}{6}\mathcal{H}(\{1\}\,|\,\{2\}).$$

This is strictly smaller than the value in (9) unless $\boldsymbol{x}_{\{1\}}$ is constant in the data in which case both are zero.

Next we consider the effect of introducing this duplicate on $\phi_2^{1:n}$. We get

$$\phi_2^{1:n} = \frac{1}{3}\mathcal{H}(\{2\}) + \frac{1}{6}\mathcal{H}(\{2\}\,|\,\{1\}) + \frac{1}{6}\mathcal{H}(\{2\}\,|\,\{3\}) + \frac{1}{3}\mathcal{H}(\{2\}\,|\,\{1,3\})$$
$$= \frac{1}{3}\mathcal{H}(\{2\}) + \frac{2}{3}\mathcal{H}(\{2\}\,|\,\{1\}).$$

The coefficient of $\mathcal{H}(\{2\})$ has decreased from $1/2$ to $1/3$ while the coefficient of $\mathcal{H}(\{2\}\,|\,\{1\})$ has increased from $1/2$ to $2/3$.

A redundant variable is one that can be perfectly identified based on some subset of other variables. Like a duplicated variable, the redundant ones do not get a uniqueness Shapley value of zero.

3.4. Database key

Suppose that one of the variables uniquely identifies each subject. That is, we never have $x_{ij} = x_{tj}$ unless $i = t$. We can think of this as the key variable in a database or even the row number in a data frame. The presence of this variable forces $\mathrm{val}(1{:}d) = \log_2(n)$ (its largest possible value) for every subject. For each subject, $N_t(u) = 1$ whenever the key is in u.

We can look at $\mathrm{val}(j\,|\,u)$ where $u \subseteq 1{:}d$ is a set of variables and now suppose that we introduce a database key with index $j = 0$. For each of the prior $d!$ orders in which variable $j \geqslant 1$ could have been included there are $d + 1$ positions at which the new variable $j = 0$ could be introduced. If variable 0 is introduced after variable j then the incremental value is unchanged. If variable 0 is introduced before variable j then the new incremental value for variable j is 0. Therefore introducing the key changes the uniqueness Shapley value for one subject to

$$\phi_j = \frac{1}{d} \sum_{u \subseteq -j} \binom{d-1}{|u|}^{-1} \left(\frac{d-|u|}{d+1}\right) \mathrm{val}(j\,|\,u)$$

$$= \frac{1}{d+1} \sum_{u \subseteq -j} \binom{d}{|u|}^{-1} \mathrm{val}(j \,|\, u) \tag{11}$$

because $d - |u|$ of the $d + 1$ possible insertion points preserve the incremental value while the others remove it. Equation (11) holds for $v \subseteq 1{:}n$ with $\phi_{t,j}$ corresponding to $v = \{t\}$. The sum in (11) is taken over subsets u of the original d variables exclusive of both variable j and the posited key variable 0.

After introducing the key variable, the contribution of $\mathrm{val}(j \mid u)$ is down-weighted by a factor of $(d - |u|)/(d+1)$, which ranges from $d/(d+1)$ for $u = \varnothing$ to $1/(d+1)$ for $u = -j$. The average value of this factor over sets $u \subseteq -j$ is

$$\frac{1}{d} \sum_{r=0}^{d-1} \frac{d-r}{d+1} = \frac{1}{2}.$$

Other things being equal we might expect that introducing the database key will halve the other variables' uniqueness Shapley values. Variables that get their importance mostly from $\mathrm{val}(j \,|\, u)$ for small $|u|$ are less affected by the key, and variables that get their importance mostly from $\mathrm{val}(j \,|\, u)$ for large $|u|$ will lose more than half of their uniqueness Shapley values.

4. All dimension trees

It would require $O(n|u|)$ time to compute $N_t(u)$ by naively checking which subjects match the target t on a particular subset of features u. For a naive calculation of the uniqueness Shapley, we would therefore need to compute $N_t(u)$ $O(2^d)$ times for each combination of feature and subjects to calculate each ϕ_j^t. Therefore, the full run time of the naive implementation of uniqueness Shapley would be $O(n^2 d^2 2^d)$. For a reasonably large number of subjects, the n^2 factor can be computationally prohibitive even when d is not large. To improve upon the naive implementation, we employ a more suitable data structure: the all dimension tree of Moore and Lee (1998).

The all dimension tree is optimized for tasks similar to calculating $N_t(u)$, i.e., generating contingency tables. The basic structure takes categorical data and constructs a tree where each branch corresponds to a particular "feature equals value" query, and the subsequent node stores the count of all subjects for whom that query and all preceding queries in the tree are true. To compute $N_t(u)$ for a given t, u pair, you start at the root of the tree and follow the branches corresponding to queries $x_j = x_{tj}$ for all $j \in u$, and the count stored in the resulting node would be $N_t(u)$. Therefore, with such a tree structure, we only require $O(|u|)$ time to compute $N_t(u)$ which no longer scales with the number of subjects. Note that the same tree can be used for all subjects, so it need not be constructed more than once for a single data set.

A naive version of this structure would generally require a prohibitive amount of memory, thereby rendering the computational savings moot even for relatively small d. The innovation of the all dimension tree comes from its techniques to

---

Algorithm 1 Uniqueness Shapley Pseudo Code

1: input: feature matrix X
2: T=ADTree(X)
3: Shap=zeros(n,d)
4: **for** subject $i = 1, 2, \ldots, n$ **do**
5:     **for** feature $j = 1, 2, \ldots, d$ **do**
6:         **for** $u \subseteq -j$ **do**
7:             $\gamma = \frac{1}{d\binom{d-1}{|u|}}$
8:             N_1=T.query(i,u)
9:             N_2=T.query(i,u+j)
10:            Shap[i,j]$+ = \gamma \log(N_1/N_2)$
11:         **end for**
12:     **end for**
13: **end for**
14: Return Shap

TABLE 1
Run times in seconds for uniqueness Shapley for the solar flare data, the full North Carolina census data, and two specific counties of North Carolina census data.

Data	n	d	ADTree	Standard
Solar Flare	1,066	9	5.30	54.55
Dare County Census	30,921	5	1.91	2,522.40
Durham County Census	253,563	5	16.46	32,426.91
North Carolina Census	7,538,125	5	362.51	n/a

reduce memory requirements especially in the common cases of sparsity and correlated features. They use several different methods to achieve this goal which are not relevant to the scope of this discussion except that they succeed in reducing the memory cost. For example, for binary features, the memory requirement is $O(2^d)$ in the worst case compared to $O(3^d)$ in the dense naive implementation, and it can achieve $O(d)$ in the best case. Performance closer to the best case is achieved when features are distributed more unevenly and are more correlated, resulting in a sparser distribution. The time to initially construct the all dimension tree is also linear in n and while worst-case exponential in d, it is not worse than our dependence to d in the Shapley calculation. This makes our overall implementation of uniqueness Shapley using all dimension trees $O(nd^2 2^d)$. A speed up by a factor of n. We use the all dimension tree Python implementation developed by Ding (2018).

For a reasonably large number of features d, this implementation is no longer computationally feasible. In those cases, a Monte Carlo approximation must be employed as in Maleki et al. (2013). The all dimension tree structure can also be adapted for large d to reduce its memory dependence at the cost of only approximately calculating $N_t(u)$ which is also discussed in Moore and Lee (1998). Our present examples did not have such very large d. An implementation of uniqueness Shapley can be found on our GitHub: https://github.com/cohortshapley/uniquenessshapley.

Nine solar flare predictor variables. The 'this pass' variable answers the question: Did the region become historically complex on this pass across the sun's disk? Source: https:// archive.ics.uci.edu/ml/datasets/Solar+Flare

Variable	Levels
Modified Zurich class	(A, B, C, D, E, F, H)
Largest spot size	(X, R, S, A, H, K)
Spot distribution	(X, O, I, C)
Activity	(1 = reduced, 2 = unchanged)
Evolution	(1 = decay, 2 = no growth, 3 = growth)
Prior 24 hr activity	(1 = no $\geqslant$ M1s, 2 = one $\geqslant$ M1, 3 = multiple $\geqslant$ M1s)
Historically-complex	(1 = yes, 2 = no)
This pass	(1 = yes, 2 = no)
Area	(1 = small, 2 = large)

5. Solar flare data example

Many data sets have a few entries that are anomalies, such as outliers. There have been many efforts to detect anomalies and others to explain them. For a survey of anomaly detection, see Chandola, Banerjee and Kumar (2009). Jacob et al. (2020) provide the Exathlon benchmark for anomaly explanation methods, aimed at time series. They include a method based on marginal entropies.

In this section we look at the solar flare data set from the UC Irvine repository (Dua and Graff, 2017). Some of the solar flares have been marked as unusual (anomaly detection). We consider uniqueness Shapley as a way to understand which variables make the anomalous data most unique. We must add that finding an identifying variable for an anomaly is a kind of association and is not necessarily causal. For instance, an unusual person's social security number is very identifying but is unlikely to be causal.

We use the second solar flare data set from the UC Irvine repository because it is said to be more reliable. It describes $n = 1066$ regions on the surface of the sun. There are 10 categorical predictors and 3 responses indicating the number of common (C class), moderate (M class) and severe (X class) solar flares in each region over a 24 hour period. Severe flares are 100 times as strong as moderate ones which in turn are 10 times as strong as common ones. The M class flares can cause radio blackouts or endanger astronauts. There are also numerical gradations within these classes. See https://www.nasa.gov/mission_pages/ sunearth/news/X-class-flares.html.

Some of those solar regions are much more interesting than others. All but five of them had no severe flares. Of those five, one had two severe flares. Four of the regions had three or more flares rated moderate or severe, so we consider those too. We will use uniqueness Shapley to study these anomalies in terms of the categorical predictors of the data set. The first nine categorical predictor variables are described in Table 2. A tenth variable, about the area of the largest spot, was constant for all 1066 regions. We omit that variable and work with $d = 9$ others.

The first two columns of Table 3 show entropy and uniqueness Shapley val-

TABLE 3

The first two columns give entropy and uniqueness Shapley for the nine solar flare predictors. The next three columns cover three types of anomalies in increasing order of severity as described in the text. The 'Area' variable takes on increasing importance as severity increases while 'This Pass' decreases.

Variable	Entropy	Shapley	Common	Moderate	Severe
Zurich	1.64	1.37	1.72	1.62	1.81
Large Spot	1.52	1.55	1.69	1.62	1.23
Spot Dist	1.16	0.92	1.17	1.34	1.49
Activity	0.43	0.45	0.73	0.67	0.86
Evolution	0.90	1.17	0.98	0.98	0.85
Prev Activ	0.18	0.16	0.33	0.43	0.93
Complex	0.67	0.77	0.69	0.69	0.29
This Pass	0.38	0.35	0.15	0.11	0.03
Area	0.12	0.07	0.22	0.43	1.36

ues for the 9 predictors in the solar flare data. Zurich and large spot are the most identifying. Area is least identifying. In this data, many of the uniqueness Shapley values are below their corresponding marginal entropy.

The last three columns of Table 3 shows cardinality Shapley for some subsets of solar regions of increasingly anomalous nature. They are those with at least one common flare, at least one moderate flare, and, finally, at least one severe flare. The area variable which is so unimportant globally becomes ever more identifying for these anomalies. So does 'previous activity'. These variables are associated with anomalous solar behavior in that the more extreme the behavior, the more identifying these become.

6. North Carolina voter registration data

Here we consider a real world demographic example in the vein of Sweeney's original work (Sweeney, 2000). The state of North Carolina publishes voting and demographic information about all of its registered voters each election. Some information is withheld to maintain privacy such as exact birth dates, exact addresses, and social security numbers. From the available features, we looked at zip code, age, race, gender, and political party affiliation for $n \approx 7.5 \times 10^6$ registered voters in the state. Summary uniqueness Shapley values and baseline marginal entropy values can be found in Table 4. There we can see that the relative ordering of the uniqueness Shapley values is consistent with the entropy values, but that we have large positive deviations for zip code and age. All variables had Shapley value greater than their entropy. Recall that the lower bound on the Shapley value is one fourth of the entropy, since $d = 4$.

In Table 5, we can see the average uniqueness Shapley values for various sub-populations of the state. Some patterns follow logically from the relative class sizes, as for example, members of races that make up a smaller percentage of the population have larger uniqueness Shapley values for race on average. Clearly membership in a less common class should help to uniquely identify someone. Other noticeable patterns highlight the effects of the correlation structure not

TABLE 4

North Carolina voter registration average uniqueness Shapley and marginal entropy. Source:
https://www.ncsbe.gov/results-data/voter-registration-data

	Zip Code	Race	Party	Gender	Age
Shapley	8.58	1.22	1.48	1.17	5.39
Entropy	6.08	1.08	1.14	0.92	4.24

TABLE 5

North Carolina voter registration data's average uniqueness Shapley values for variables
aggregated over voter subgroups. Final column is percentage of total population.

Subgroup	Zip Code	Race	Party	Gender	Age	% Pop.
Democratic	8.44	1.54	1.31	1.14	5.47	36
Libertarian	8.15	1.16	6.65	1.14	4.54	1
Republican	8.76	0.72	1.51	1.15	5.50	30
None	8.57	1.32	1.51	1.22	5.23	33
Female	8.58	1.18	1.46	0.97	5.46	49
Male	8.61	1.15	1.49	1.20	5.40	42
No Answer	8.42	1.76	1.53	2.24	4.91	8
White	8.75	0.53	1.58	1.09	5.52	63
Black	8.28	1.70	1.14	1.10	5.35	21
Asian	7.56	5.44	1.44	1.09	5.01	1
Native American	7.64	5.14	1.40	1.01	5.13	1
No Answer	8.45	2.27	1.54	1.91	4.94	10
Age < 32	8.49	1.52	1.52	1.26	4.63	25
32–48	8.48	1.33	1.50	1.20	5.22	25
48–63	8.61	1.13	1.46	1.15	5.33	25
> 63	8.71	0.90	1.44	1.09	6.34	25

captured by the marginal entropy measures. For example, the average cardinality value for race decreases for older age cohorts.

To further investigate the relationship between class imbalance and the uniqueness Shapley value, we look at zip code defined subpopulations in Figure 1. We plot the Shapley value for race versus the proportion of the population that does not identify as white. There are a handful of large positive outliers to this trend. Upon further inspection, we can note that these are all zip codes with a very high proportion of American Indian voters.

We can also measure the effect of feature granularity on the uniqueness Shapley values. Instead of recording location by zip code, we can coarsen it to the county level or even the whole state. The results for this adjusted dataset are in Table 6. The uniqueness Shapley values for the other variables are not sensitive to this coarsening. Similarly, instead of considering each age by year, we can coarsen the grouping level to five or ten or 25 years and again we see no meaningful changes outside of the expected reduction in the age feature, even if we lump all ages into one bucket.

Beyond summary tables, we can visualize individual values in plots such as Figure 2. Each vertical line corresponds to a single voter with the height of each colored segment representing the uniqueness Shapley value for the respective
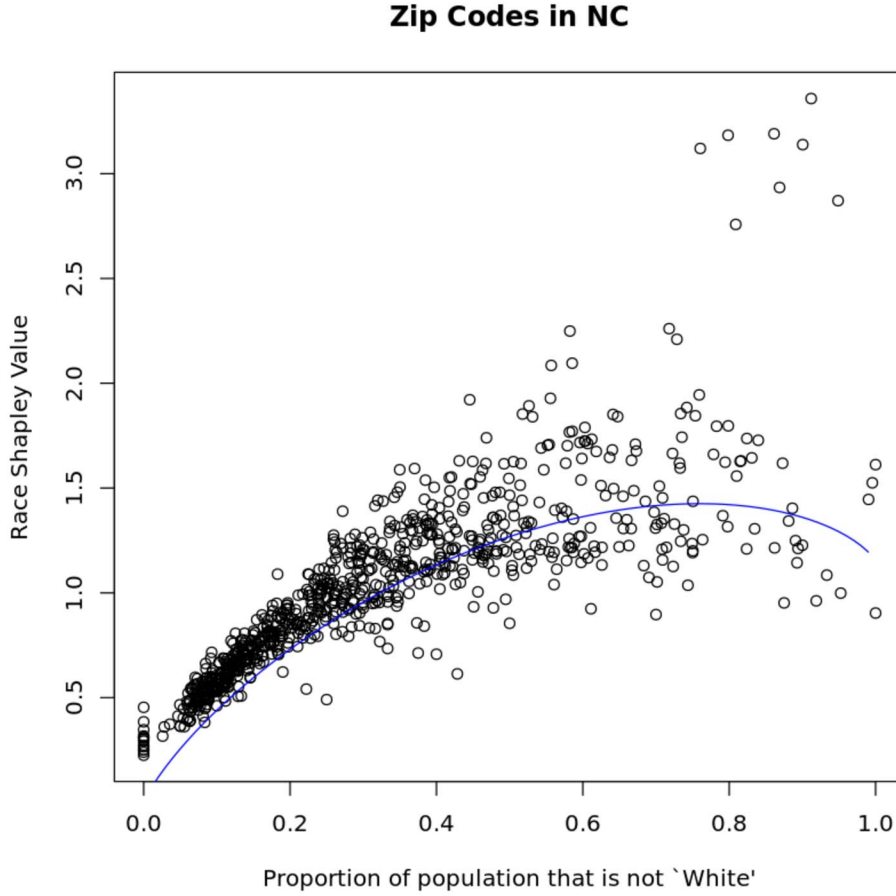
Zip Codes in NC

FIG 1. *Average uniqueness Shapley for race in a zip code versus proportion of the population in that zip code that is not 'White'. The reference shows entropy of race for a hypothetical population sampled independently from North Carolina except that it has the implied proportion 'White'.*

feature. They are ordered by their overall uniqueness, and we only display every hundredth voter so they fit in the figure.

7. Discussion

We have used Shapley value to quantify and compare the power that different categorical variables have to identify a subject. When we use the additive property of Shapley value to average this measure over a population we get an expression that equals a weighted sum of conditional entropies. Variables that when revealed increase entropy are the ones that most identify subjects. When we average over a different distribution, such as a sub-population of interest, the entropies are replaced by cross entropies.

North Carolina voter registration average uniqueness Shapley. The first block coarsens location from zip code to county to state. The second block coarsens age, starting with the original single year granularity.

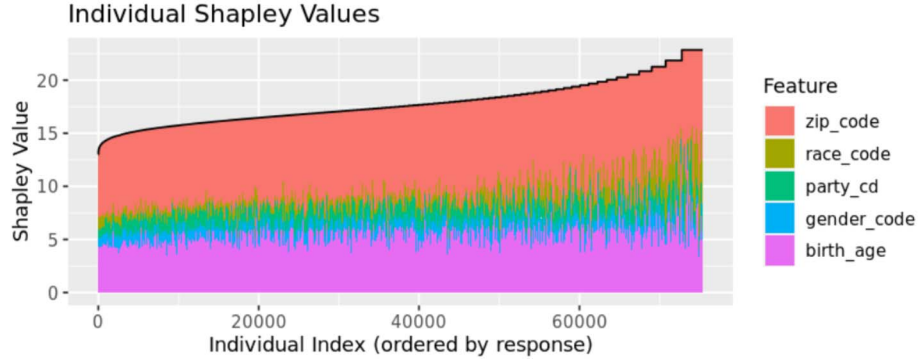|	Location	Race	Party	Gender	Age
Zip Code	8.58	1.22	1.48	1.17	5.39
County	5.55	1.27	1.50	1.19	5.43
State	0.00	1.32	1.53	1.19	5.45
	Zip code	Race	Party	Gender	Age
Single year	8.58	1.22	1.48	1.17	5.39
5 year age buckets	8.61	1.24	1.50	1.18	3.34
10 year	8.62	1.24	1.50	1.19	2.39
25 year	8.63	1.25	1.51	1.19	1.19
One bucket	8.64	1.25	1.51	1.19	0.00

FIG 2. *Individual Shapley values stacked vertically (subsampled every 100).*

The extent to which a variable makes you unique depends on the order in which it and other variables are revealed. Some variables might, if revealed last, be very identifying. Other variables might be redundant if revealed last but very informative if revealed early due to associations among the variables. The Shapley formulation combines all of the orders in which a variable might be revealed.

In this work we have kept to data sets with a modest number d of variables because computation of Shapley value can include a cost that grows proportionally to $d2^d$, the number of cohorts a subject might belong to. There are Monte Carlo sampling algorithms for Shapley value that allow larger d.

We have focused on categorical variables. Continuous variables can be coarsened into categorical ones by setting ranges. The finer the range the greater the Shapley value is. This is appropriate because finer classifications really are more revealing.

It is also possible to use asymmetric notions of symmetry for continuous variables as considered in Mase, Owen and Seiler (2019). For instance if we

declare x_{ij} to be similar to x_{tj} whenever $|x_{ij} - x_{tj}| \leqslant \delta_j |x_{tj}|$ we might find that x_{ij} is similar to x_{tj} but not the converse. The proper way to account for a continuous variable when quantifying uniqueness depends on how we expect that variable might be revealed. We can compare the effects of revealing age in 1 or 5 or 10 year windows and can also measure how the effect of revealing another variable such as race or gender depends on the granularity with which age has been revealed.

Acknowledgments

References

[1] Chandola, V., Banerjee, A. and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)* **41** 1–58.

[2] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley and Sons, New York.

[3] Ding, F. (2018). Sparse AD-Tree package in Python. https://github.com/uraplutonium/adtree-py.

[4] Dua, D. and Graff, C. (2017). UCI Machine Learning Repository.

[5] Dwork, C. (2008). Differential privacy: A survey of results. In *International conference on theory and applications of models of computation* 1–19. Springer.

[6] Golle, P. (2006). Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society* 77–80.

[7] Gómez-Boix, A., Laperdrix, P. and Baudry, B. (2018). Hiding in the crowd: an analysis of the effectiveness of browser fingerprinting at large scale. In *Proceedings of the 2018 world wide web conference* 309–318.

[8] Jacob, V., Song, F., Stiegler, A., Diao, Y. and Tatbul, N. (2020). AnomalyBench: An Open Benchmark for Explainable Anomaly Detection Technical Report, arXiv:2010.05073.

[9] Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T. and Rogers, A. (2013). Bounding the estimation error of sampling-based Shapley value approximation Technical Report, arXiv:1306.4265.

[10] Mase, M., Owen, A. B. and Seiler, B. B. (2019). Explaining black box decisions by Shapley cohort refinement Technical Report, arXiv:1911.00467.

[11] Moore, A. and Lee, M. S. (1998). Cached sufficient statistics for efficient machine learning with large datasets. *Journal of Artificial Intelligence Research* **8** 67–91.

[12] Shapley, L. S. (1953). A value for n-person games. In *Contribution to the Theory of Games II (Annals of Mathematics Studies 28)* (H. W. Kuhn and A. W. Tucker, eds.) 307–317. Princeton University Press, Princeton, NJ.

[13] Sweeney, L. (2000). Simple demographics often identify people uniquely Privacy Working Paper No. 3, Carnegie Mellon University.