Counterfactual Explanation of Shapley Value in Data Coalitions

Michelle Si
Duke University
Durham, NC, USA
michelle.si@duke.edu

Jian Pei
Duke University
Durham, NC, USA
j.pei@duke.edu

ABSTRACT

The Shapley value is widely used for data valuation in data markets. However, explaining the Shapley value of an owner in a data coalition is an unexplored and challenging task. To tackle this, we formulate the problem of finding the counterfactual explanation of Shapley value in data coalitions. Essentially, given two data owners A and B such that A has a higher Shapley value than B, a counterfactual explanation is a smallest subset of data entries in A such that transferring the subset from A to B makes the Shapley value of A less than that of B. We show that counterfactual explanations always exist, but finding an exact counterfactual explanation is NP-hard. Using Monte Carlo estimation to approximate counterfactual explanations directly according to the definition is still very costly, since we have to estimate the Shapley values of owners A and B after each possible subset shift. We develop a series of heuristic techniques to speed up computation by estimating differential Shapley values, computing the power of singular data entries, and shifting subsets greedily, culminating in the SV-Exp algorithm. Our experimental results on real datasets clearly demonstrate the efficiency of our method and the effectiveness of counterfactuals in interpreting the Shapley value of an owner.

PVLDB Reference Format:
Michelle Si and Jian Pei. Counterfactual Explanation of Shapley Value in Data Coalitions. PVLDB, 17(11): 3332 - 3345, 2024.
doi:10.14778/3681954.3682004

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/michelleeesi/shapley_counterfactual.

1 INTRODUCTION

The power of big data largely comes from many secondary uses of data, such as enabling numerous machine learning and AI models [24, 29, 35], recommender systems [3, 49], causal inference [26, 45, 46], and data-driven decision-making applications [48]. However, incentivizing and facilitating data sharing and collaboration at scale remains a great challenge. Data markets [11, 17, 31, 44, 51] are emerging as a promising way to enable and facilitate data sharing among many potential owners and consumers. Essentially, a **data market** is an online platform where various parties with demands can acquire datasets or data services; at the same time, data owners can exchange data and data services for revenue or compensations in one way or another. There are already many active data markets, such as AWS Data Exchange, Windows Azure Marketplace, Dawex, Datarade, dmi.io, WorldQuant, Xignite, and BlueTalon [31].

At the core of every data market, there is a data valuation module. In a data market, a group of data owners collaborate to produce a target dataset or complete a target task that a buyer would like to acquire, such as assembling a dataset for data analytics or machine learning. We call collaboration among data owners a **data coalition**, where multiple datasets owned by different parties are used to produce target datasets. In a nutshell, **data valuation** assigns a score to a data owner to reflect the data owner's contribution towards a task achieved by a data coalition.

Data valuation plays a central role in ensuring fairness, effectiveness, and efficiency of data markets. There are various requirements on data valuation in different data markets [47], such as truthfulness [2], revenue maximization, fairness [2, 52], arbitrage-freeness [36], privacy-preservation [1, 4, 5, 10, 13, 16, 18, 19, 21, 22, 27, 42, 43, 57, 61, 62], and computational efficiency [2, 9, 23]. The rich diversity in requirements poses many technical challenges for data valuation solutions.

The Shapley value [52] is often used as a measure in data valuation [32, 33, 60], which is the expectation of marginal utility gain that a data owner can bring into coalitions. We will review the mathematical details in Section 2. The Shapley value is the only valuation measure that provably satisfies efficiency, symmetry, zero element, and additivity.

While numerous studies have investigated the fast computation of Shapley values by either designing cost-saving sampling strategies [28, 39] or tackling the Shapley computation in specific settings [15], one important question remains unexplored: **how can one understand and explain the Shapley value of a data owner in a data coalition?**

Let us consider a concrete example. Suppose two data owners, Alice and Brittany, participate in a data coalition $\mathbb{O}$ and obtain their Shapley values $\psi_{\mathbb{O}}(Alice)$ and $\psi_{\mathbb{O}}(Brittany)$, respectively. Without loss of generality, let us assume $\psi_{\mathbb{O}}(Alice) > \psi_{\mathbb{O}}(Brittany)$. Then, one may ask how we can explain Alice's advantage over Brittany according to their Shapley values. To answer this question, one intuitive approach is to look for a counterfactual explanation, which is a minimal subset S of data owned by Alice such that transferring S from Alice to Brittany can flip the direction of the inequality, that is, making $\psi_{\mathbb{O}}(Alice) < \psi_{\mathbb{O}}(Brittany)$.

Using counterfactuals as an explanation tactic is a well established approach in philosophy and has enjoyed numerous applications in many domains [54]. A counterfactual explanation S provides some interesting insights. For example, by checking the data entries in S, one may understand which data entries are the most crucial for Alice's advantage—what really makes Alice be able to

Proceedings of the VLDB Endowment, Vol. 17, No. 11 ISSN 2150-8097.
doi:10.14778/3681954.3682004

contribute more and thus be more valuable than Brittany in the coalition? The counterfactual explanation allows us to detail the root of differences in Shapley values between data owners in a way that the Shapley value itself does not illuminate. For example, if Alice owns much more data than Alice but the size of the counterfactual explanation is small, it means that Alice's advantage is driven by a few powerful rows of data–the elements of the counterfactual. To use another analogy, using a counterfactual explanation would allow us to see if, given two sports teams, if the success of one over the other on average is driven by one star player (a small counterfactual) or an overall better playing team (a large counterfactual). In Section 4, we report two interesting case studies. The first one demonstrates that counterfactual explanations can help select features from one subset to enhance the features in another subset. The second case illustrates that counterfactual explanations can disclose the data distribution differences among different data owners.

Motivated by this insight, we tackle the problem of finding the counterfactual explanation for the Shapley value. To the best of our knowledge, we are the first to model this problem. We show that the counterfactual explanation always exists, but finding the counterfactual explanation of the Shapley value is NP-hard.

Even if we use Monte Carlo estimation to approximate counterfactual explanations according to the definition, the algorithm is still very costly–we have to estimate the Shapley values of data owners after every change. To address the computational challenges, we develop a series of heuristic techniques to improve computation. Firstly, we provide techniques to estimate the differential Shapley (the difference between the Shapley values of two data owners) directly. The differential Shapley not only avoids the complication of estimating the Shapley values of two data owners and then calculating the difference, but also improves the estimation quality when using Monte Carlo sampling approaches. Secondly, estimating differential Shapley values is still costly when there are many data entries. For this, we develop an iterative greedy search approach. In each iteration, we find a data entry such that moving the data entry from one data owner to the other causes an estimated maximal change in differential Shapley value, which is measured using the notion of the *power* of a data entry. The iteration continues until an approximation of the counterfactual explanation is obtained.

We also conduct experiments and highlight case studies on real datasets to examine the efficiency of our method and the effectiveness of using counterfactuals to explain the Shapley value.

The rest of the paper is organized as follows. We formulate the problem and present an exact algorithm in Section 2. In Section 3 we develop the heuristic approximation methods. We report the experimental results in Section 4, discuss related work in Section 5, and address the limitations and possible extensions of our method in Section 6. Section 7 concludes the paper.

2 PROBLEM FORMULATION AND AN EXACT ALGORITHM

Assume a set of data entries $\mathbb{D} = \{x_1, \ldots, x_m\}$ of interest. Let $\mathbb{O}$ be a set of data owners who achieve a task in a coalition. For each data owner $O \in \mathbb{O}$, we overload symbol O to also denote the dataset that O owns, that is, $O \subseteq \mathbb{D}$ is the subset of data entries that O owns. A (data) **coalition** $\mathcal{S} \subseteq \mathbb{O}$ is a subset of data owners and their datasets. Correspondingly, the union of datasets owned by a set of data owners $\mathcal{S}$, that is, $\cup_{O\in\mathcal{S}} O \subseteq \mathbb{D}$, is called a **composed dataset**. $\mathbb{O}$ itself as a coalition is called the **grant coalition**.

Given a set of data owners $\mathbb{O}$ and a data owner $O \in \mathbb{O}$, denote by $\psi_{\mathbb{O}}(O)$ the Shapley value [52] of O, that is,

$$\begin{aligned}\psi_{\mathbb{O}}(O) &= \frac{1}{|\mathbb{O}|}\sum_{\mathcal{S}\subseteq\mathbb{O}\setminus\{O\}} \frac{U(\mathcal{S}\cup\{O\}) - U(\mathcal{S})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|}} \\ &= \frac{1}{|\mathbb{O}|!}\sum_{\pi\in\Pi(\mathbb{O})} (U(P_O^\pi \cup \{O\}) - U(P_O^\pi))\end{aligned} \tag{1}$$

where $U : 2^{\mathbb{O}} \to \mathbb{R}$ is a **utility function** that returns the utility of a coalition by a set of data owners, $\mathbb{R}$ is the set of real numbers, $U(\emptyset) = 0$, $\Pi(\mathbb{O})$ is the set of all possible permutations of data owners, and P_O^π is the set of data owners preceding O in permutation π.

In many applications, the more data, the better the utility. In other words, a utility function is often monotonic. Even with a non-monotonic utility function $U(\cdot)$, since a user often has the incentive to try every possible way to extract the best value from a set of data, the attempts lead to a utility function $U^*(D) = \max_{D'\subseteq D}\{U(D')\}$, which is monotonic. Based on this rationale, we from now on assume that the utility function is **monotonic**. That is, for any two subsets of data entries $D_1, D_2 \subseteq \mathbb{D}$, if $D_1 \subseteq D_2$, then $U(D_1) \leq U(D_2)$.

Consider two data owners $A, B \in \mathbb{O}$ such that $\psi_{\mathbb{O}}(A) > \psi_{\mathbb{O}}(B)$. We ask the following question: *which data entries in A can explain the higher Shapley value of A compared to that of B?* Particularly, we are interested in the **counterfactual explanation**, that is, a subset $\Delta A \subseteq A$ of the minimum size such that, if ΔA is transferred from A to B, the Shapley value of B will be larger than that of A.

Formally, let $\mathbb{O}[A \xrightarrow{\Delta A} B] = \mathbb{O} \setminus \{A, B\} \cup \{A \setminus \Delta A, B \cup \Delta A\}$. We want to solve the following **counterfactual explanation problem of the Shapley value** as an optimization problem.

$$\begin{aligned}&\min\{|\Delta A|\} \\ \text{s.t. } &\Delta A \subseteq A \\ &\psi_{\mathbb{O}[A\xrightarrow{\Delta A}B]}(A \setminus \Delta A) < \psi_{\mathbb{O}[A\xrightarrow{\Delta A}B]}(B \cup \Delta A)\end{aligned} \tag{2}$$

We can show that the problem of counterfactual explanation is NP-hard.

Theorem 1 (Complexity). *The problem of counterfactual explanation is NP-hard.*

Proof sketch. We prove by constructing a reduction from the set cover problem, whose search version is known to be NP-hard [30]. Given a set of elements $\mathbb{D} = \{x_1, \ldots, x_n\}$ (called the universe) and a collection $\mathbb{S} = \{S_1, \ldots, S_m\}$ of m subsets whose union equals the universe, that is, $S_i \subseteq \mathbb{D}$ and $\cup_{i=1}^m S_i = \mathbb{D}$, the set cover problem is to find the smallest sub-collection of $\mathbb{S}$ whose union equals the universe $\mathbb{D}$.

For each sub-collection $\mathcal{S} = \{S_{i_1}, \ldots, S_{i_k}\} \subseteq \mathbb{S}$, we define an encoding function $f(\mathcal{S}) = \frac{\sum_{j=1}^k 2^{i_j}}{2^{m+1}}$. Clearly, $0 < f(\mathcal{S}) < 1$ as long as $\mathcal{S} \neq \emptyset$.

We construct a game as follows. We treat each subset S_i as a data entry. There are only two data owners, $O_0 = \emptyset$ and $O_1 = \mathbb{S}$, that is, O_0 does not have any data and O_1 has all the subsets. We define a utility function $U : 2^{\mathbb{S}} \to [0, m+1)$ such that for each sub-collection $\mathcal{S} \subseteq \mathbb{S}$, $U(\mathcal{S}) = 0$ if $\cup_{S_i \in \mathcal{S}} S_i \neq \mathbb{D}$; and otherwise $U(\mathcal{S}) = m - |\mathcal{S}| + f(\mathcal{S})$.

Clearly, $\psi(O_1) > 0$ and $\psi(O_0) = 0$, and thus $\psi(O_1) > \psi(O_0)$. Let $\Delta O \subseteq O_1$ be a counterfactual explanation of the Shapley value. Then, one of the following three cases may happen. Firstly, if $\mathbb{S}$ has only one cover $\mathcal{S} \subseteq \mathbb{S}$, then $\Delta O = \mathcal{S}$. Secondly, if $\mathbb{S}$ has two or more covers and the minimal cover $\mathcal{S}$ satisfies $|\mathcal{S}| < |\mathbb{S} \setminus \mathcal{S}|$, then $\Delta O = \mathcal{S}$ (note that if the size of a cover S is strictly greater than $|\mathbb{S} \setminus \mathcal{S}|$, then $\mathcal{S}$ cannot be minimal). Otherwise, we have the third case, where there are two covers and the minimal cover $\mathcal{S}$ satisfies $|\mathcal{S}| = |\mathbb{S} \setminus \mathcal{S}|$. Thus, $\mathbb{S}$ has two disjoint covers $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathbb{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ and $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, and then we have that $\Delta O = \arg\max_{\mathcal{S} \in \{\mathcal{S}_1, \mathcal{S}_2\}} U(\mathcal{S})$. Therefore, the subsets in ΔO is a set cover in $\mathbb{D}$. □

It is important to note the feasibility of this problem.

Proposition 2. *The counterfactual explanation problem of Shapley value always has a feasible solution.*

Proof sketch. Since $\psi_{\mathbb{O}}(A) > \psi_{\mathbb{O}}(B)$ and the utility function is monotonic, the Shapley value is non-negative, that is, $\psi_{\mathbb{O}}(A) > 0$. Therefore, $A \neq \emptyset$. Let $\mathbb{O}' = \mathbb{O}[A \xrightarrow{\Delta A} B]$. In the trivial scenario where $\Delta A = A$, due to the monotonicity of the utility function, we have $\psi_{\mathbb{O}'}(A \setminus \Delta A) = \psi_{\mathbb{O}'}(\emptyset) = 0$. Moreover, $B \cup \Delta A \supseteq \Delta A = A \neq \emptyset$. Thus, $\psi_{\mathbb{O}'}(B \cup \Delta A) > 0$. Then, $\psi_{\mathbb{O}'}(A \setminus \Delta A) < \psi_{\mathbb{O}'}(B \cup \Delta A)$. The feasibility of the problem follows immediately. □

We now introduce the notion of the **differential Shapley value**. For two data owners $A, B \in \mathbb{O}$, we define the differential Shapley value between A and B as $\Psi_{\mathbb{O}}(A, B) = \psi_{\mathbb{O}}(A) - \psi_{\mathbb{O}}(B)$. We have the following useful result, a variation of Lemma 1 by Jia et al. [28] on efficient data valuation.

Theorem 3 (Differential Shapley Value). *For two data owners $A, B \in \mathbb{O}$,*

$$\Psi_{\mathbb{O}}(A, B) = \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{1}{(|\mathcal{S}|+1)\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}} (U(\mathcal{S} \cup \{A\}) - U(\mathcal{S} \cup \{B\}))$$

Proof. According to the definition of the Shapley value (Equation 1), we have

$$\begin{aligned}
\psi_{\mathbb{O}}(A) =& \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A\}} \frac{U(\mathcal{S} \cup \{A\}) - U(\mathcal{S})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|}} \\
=& \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{U(\mathcal{S} \cup \{A\}) - U(\mathcal{S})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|}} \\
&+ \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{U((\mathcal{S} \cup \{B\}) \cup \{A\}) - U(\mathcal{S} \cup \{B\})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}}
\end{aligned}$$

Algorithm 1: The Brute-force Exact Algorithm

Input: A set of data owners $\mathbb{O}$ and two data owners $A, B \in \mathbb{O}$ such that $\psi_{\mathbb{O}}(A) > \psi_{\mathbb{O}}(B)$
Output: Solution to Equation 2
for $i = 1$ **to** $|A| - 1$ **do**
  for $\Delta A \subset A$ *s.t.* $|\Delta A| = i$ **do**
    Let $\mathbb{O}' = \mathbb{O} \setminus \{A, B\} \cup \{A \setminus \Delta A, B \cup \Delta A\}$;
    if $\psi_{\mathbb{O}'}(A \setminus \Delta A) < \psi_{\mathbb{O}'}(B \cup \Delta A)$ **then**
      return ΔA

Similarly,

$$\begin{aligned}
\psi_{\mathbb{O}}(B) =& \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{U(\mathcal{S} \cup \{B\}) - U(\mathcal{S})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|}} \\
&+ \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{U((\mathcal{S} \cup \{A\}) \cup \{B\}) - U(\mathcal{S} \cup \{A\})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\Psi_{\mathbb{O}}(A, B) = \psi_{\mathbb{O}}(A) - \psi_{\mathbb{O}}(B) \\
=& \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{U(\mathcal{S} \cup \{A\})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|}} - \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{U(\mathcal{S} \cup \{B\})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}} \\
&- \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{U(\mathcal{S} \cup \{B\})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|}} + \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{U(\mathcal{S} \cup \{A\})}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}} \\
=& \frac{1}{|\mathbb{O}|} \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{\left(\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1} + \binom{|\mathbb{O}|-1}{|\mathcal{S}|}\right)}{\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}\binom{|\mathbb{O}|-1}{|\mathcal{S}|}} (U(\mathcal{S} \cup \{A\}) - U(\mathcal{S} \cup \{B\})) \\
=& \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{1}{(|\mathcal{S}|+1)\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}} (U(\mathcal{S} \cup \{A\}) - U(\mathcal{S} \cup \{B\})) \quad \square
\end{aligned}$$

Since computing the exact Shapley value is #P-hard [14], Theorem 3 allows us to work on the difference between the Shapley values of two data owners directly without estimating the individual values.

To obtain the exact counterfactual explanation of Shapley value, a straightforward approach uses Theorem 3, enumerates all possible non-empty subsets ΔA of A, and checks whether the resulting datasets satisfy Equation 2. Among all those ΔA satisfying Equation 2, we pick the one with the smallest size. Since we are searching for the smallest subset, we can enumerate the subsets of A in the size-ascending order. The algorithm stops the first time Equation 2 is satisfied, guaranteeing that ΔA is minimum in size. The pseudocode is given in Algorithm 1. Clearly, a straightforward implementation of Algorithm 1 generates the powerset of A in the worst case, and thus its time complexity is exponential.

3 HEURISTIC APPROXIMATION METHODS

Computing the exact answer to the counterfactual explanation problem is costly and does not scale up for the scenarios where there are many data owners. Thus, in this section we explore heuristic methods in two ways.

Firstly, we approximate the difference between the Shapley values of two data owners through Monte Carlo sampling. Secondly, we explore how to estimate Shapley value changes when we move

some data entries from one data owner to the other and then greedily search for a counterfactual explanation.

3.1 Estimating Differential Shapley Value

A straightforward approach to estimate the difference between the Shapley values of two data owners is to first estimate the individual Shapley values and then infer the difference. This approach has two drawbacks. Firstly, we need to draw samples to estimate the two individual Shapley values, and each has independent estimation error. Secondly, the estimation of the difference using the estimated individual Shapley values introduces a new estimation error. Can we reduce the estimation error and improve efficiency by estimating the difference directly?

According to Theorem 3, the differential Shapley value between two data owners depends only on the difference between the marginal utility $U(\mathcal{S} \cup \{A\})$ and $U(\mathcal{S} \cup \{B\})$ on all coalitions $\mathcal{S}$ in which neither A nor B participates. This insight can be used in estimating the differential Shapley value directly using Monte Carlo approximation.

Corollary 1 (Differential Shapley Value by Permutation). *For two data owners $A, B \in \mathbb{O}$,*

$$\Psi_{\mathbb{O}}(A,B) = \frac{1}{2(|\mathbb{O}|-1)!} \sum_{\pi \in \Pi(\mathbb{O})} \frac{1}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} (U(P^{\pi}_{\{A,B\}} \cup \{A\}) - U(P^{\pi}_{\{A,B\}} \cup \{B\}))$$

where $P^{\pi}_{\{A,B\}}$ is the set of data owners preceding A and B in permutation π.

Proof. For each subset of data owners $\mathcal{S} \subseteq \mathbb{O} \setminus \{A,B\}$, there are $|\mathcal{S}|! \times 2 \times (|\mathbb{O} \setminus \mathcal{S}| - 1)! = 2|\mathcal{S}|!(|\mathbb{O}| - |\mathcal{S}| - 1)!$ permutations π in $\Pi(\mathbb{O})$ such that $P^{\pi}_{\{A,B\}} = \mathcal{S}$. Following Theorem 3, we have

$$\begin{aligned}
\Psi_{\mathbb{O}}(A,B) &= \sum_{\mathcal{S} \subseteq \mathbb{O} \setminus \{A \cup B\}} \frac{1}{(|\mathcal{S}|+1)\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}} (U(\mathcal{S} \cup \{A\}) - U(\mathcal{S} \cup \{B\})) \\
&= \sum_{\pi \in \Pi(\mathbb{O})} \frac{1}{(|P^{\pi}_{\{A,B\}}|+1)\binom{|\mathbb{O}|-1}{|P^{\pi}_{\{A,B\}}|+1}} \\
&\quad \times \frac{(U(P^{\pi}_{\{A,B\}} \cup \{A\}) - U(P^{\pi}_{\{A,B\}} \cup \{B\}))}{2|P^{\pi}_{\{A,B\}}|!(|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1)!} \\
&= \frac{1}{2(|\mathbb{O}|-1)!} \sum_{\pi \in \Pi(\mathbb{O})} \frac{1}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} (U(P^{\pi}_{\{A,B\}} \cup \{A\}) \\
&\quad - U(P^{\pi}_{\{A,B\}} \cup \{B\})) \quad \square
\end{aligned}$$

The corollary immediately leads to a Monte Carlo estimator, which takes a uniform sample of permutations $X \subseteq \Pi(\mathbb{O})$ to estimate the differential Shapley value $\Psi_{\mathbb{O}}(A,B)$. We can show that this estimation is unbiased.

Corollary 2 (Unbiased Estimate). *For a uniform sample of permutations $X \subseteq \Pi(\mathbb{O})$,*

$$\Delta\psi^{A-B}_{\mathbb{O}} = \frac{|\mathbb{O}|}{2|X|} \sum_{\pi \in X} \frac{1}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} (U(P^{\pi}_{\{A,B\}} \cup \{A\}) - U(P^{\pi}_{\{A,B\}} \cup \{B\}))$$

is an unbiased estimate of $\Psi_{\mathbb{O}}(A,B)$.

Algorithm 2: The Monte Carlo Baseline Algorithm

Input: the same as Algorithm 1
Output: An approximation to solution to Equation 2
for $i = 1$ **to** $|A| - 1$ **do**
  for $\Delta A \subset A$ *s.t.* $|\Delta A| = i$ **do**
    Let $\mathbb{O}' = \mathbb{O} \setminus \{A,B\} \cup \{A \setminus \Delta A, B \cup \Delta A\}$;
    Estimate $\Delta\psi^{A\setminus\Delta A - B\cup\Delta A}_{\mathbb{O}'}$ as an unbiased estimate of $\Psi_{\mathbb{O}'}(A \setminus \Delta A, B \cup \Delta A)$ using Monte Carlo based on Corollary 2
    if $\Delta\psi^{A\setminus\Delta A - B\cup\Delta A}_{\mathbb{O}'} < 0$ **then**
      return ΔA
return A

Proof. By linearity of expectation and uniformity of samples,

$$\begin{aligned}
&\mathbb{E}[\Delta\psi^{A-B}_{\mathbb{O}}] \\
=&\mathbb{E}[\frac{|\mathbb{O}|}{2|X|} \sum_{\pi \in X} \frac{1}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} (U(P^{\pi}_{\{A,B\}} \cup \{A\}) \\
&- U(P^{\pi}_{\{A,B\}} \cup \{B\}))] \\
=&\frac{|\mathbb{O}|}{2|X|} |X| \times \mathbb{E}[\frac{1}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} (U(P^{\pi}_{\{A,B\}} \cup \{A\}) \\
&- U(P^{\pi}_{\{A,B\}} \cup \{B\}))] \\
=&\frac{|\mathbb{O}|}{2} \sum_{\pi \in \Pi(\mathbb{O})} \frac{1}{|\mathbb{O}|!} \times (\frac{1}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} (U(P^{\pi}_{\{A,B\}} \cup \{A\}) \\
&- U(P^{\pi}_{\{A,B\}} \cup \{B\}))) \\
=&\frac{1}{2(|\mathbb{O}|-1)!} \sum_{\pi \in \Pi(\mathbb{O})} \frac{1}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} (U(P^{\pi}_{\{A,B\}} \cup \{A\}) \\
&- U(P^{\pi}_{\{A,B\}} \cup \{B\})) \\
=&\Psi_{\mathbb{O}}(A,B) \quad \square
\end{aligned}$$

Using Corollary 2 in Algorithm 1, we can obtain a Monte Carlo baseline algorithm for the counterfactual explanation problem. The pseudo-code is given in Algorithm 2.

3.2 A Greedy Approach: the Framework

The Monte Carlo baseline algorithm (Algorithm 2) still needs to search across many subsets of the data sets in ascending order and thus needs to use Monte Carlo estimation many times.

To tackle the computational cost, we explore a greedy approach. We iteratively identify the best data entry owned by data owner A such that, if moved to B, reduces the difference of the Shapley values between A and B most. As the goal is to bring down the differential Shapley $\Psi_{\mathbb{O}}(A,B)$ from Theorem 3 to a negative value as quickly as possible, we move these "powerful" data entries one by one from A to B until $\Psi_{\mathbb{O}}(A,B) < 0$.

Specifically, for a data entry $x \in A$, the change of the differential Shapley value caused by moving x from A to B is

$$\begin{aligned}
\Lambda^{A \xrightarrow{x} B}_{\mathbb{O}} &= \Psi_{\mathbb{O}}(A,B) - \Psi_{\mathbb{O}'}(A \setminus \{x\}, B \cup \{x\}) \\
&= [\psi_{\mathbb{O}}(A) - \psi_{\mathbb{O}}(B)] - [\psi_{\mathbb{O}'}(A \setminus \{x\}) - \psi_{\mathbb{O}'}(B \cup \{x\})],
\end{aligned} \tag{3}$$

where $\mathbb{O}' = \mathbb{O}[A \xrightarrow{\{x\}} B]$. The larger the value of $\Lambda_{\mathbb{O}}^{A \xrightarrow{x} B}$, the more significant the change for counterfactual explanation.

Given data owners A and B, $\Psi_{\mathbb{O}}(A, B)$ in Equation 3 is a constant. Thus, we define the **power** of x with respect to A and B, denoted by $power_{A\to B}^{\mathbb{O}}(x)$, as the change of the differential Shapley value when x is moved from A to B, that is,

$$power_{A\to B}^{\mathbb{O}}(x) = \psi_{\mathbb{O}'}(B \cup \{x\}) - \psi_{\mathbb{O}'}(A \setminus \{x\})$$

Heuristically, the larger the power of a data entry, the more the data entry contributes to a counterfactual explanation of the Shapley values.

The framework of our greedy approach works as follows. We find the data entry x that has the largest power and move it from A to B. If the Shapley value of B becomes larger than that of A after the move, that is, $\psi_{\mathbb{O}'}(A \setminus \{x\}) < \psi_{\mathbb{O}'}(B \cup \{x\})$ and thus $power_{A\to B}^{\mathbb{O}}(x) > 0$, then x is a counterfactual explanation. If not, then we iteratively find the next entry with the largest power, append this entry to the set containing the previously computed best entries, and conduct the move. The iteration continues until the Shapley value of B becomes larger than that of A after the moves. The set of all entries moved form a greedy approximation of the counterfactual explanation.

3.3 Computing the Power of a Data Entry

In our greedy approach, the key operation that is performed again and again is computing the power of a data entry. Now, let us consider how to compute the power of a data entry efficiently.

First of all, if data owner A has only one data entry, that is, $A = \{x\}$, then after moving x to B, A becomes empty and thus the Shapley value is 0. x is the trivial counterfactual explanation. Thus, in the rest of the discussion, we assume $|A| > 1$.

A data entry $x \in A$ must be in one of the two cases, $x \in A \cap B$ and $x \in A \setminus B$. We consider the two situations one by one.

3.3.1 The power of a Common Data Entry. Suppose $x \in A \cap B$, that is, x is a **common data entry** to A and B. If we move x from A to B, what change would happen to the Shapley values of the two data owners?

Since $x \in A \cap B \subseteq B$, $B \cup \{x\} = B$. Thus, using Theorem 3, we have

$$power_{A\to B}^{\mathbb{O}}(x) = \sum_{\mathcal{S}\subseteq\mathbb{O}'\setminus\{B,A\setminus\{x\}\}} \frac{U(\mathcal{S}\cup\{B\}) - U(\mathcal{S}\cup\{A\setminus\{x\}\}}{(|\mathcal{S}|+1)\binom{|\mathbb{O}'|-1}{|\mathcal{S}|+1}} \tag{4}$$

Since $|A| > 1, A\setminus\{x\} \neq \emptyset$, moving a data entry x from A to B does not make A empty. Thus, the number of non-empty data owners remains the same after the moving, that is, $|\mathbb{O}| = |\mathbb{O}'|$. Moreover, for every coalition $\mathcal{S} \subseteq \mathbb{O} \setminus \{A, B\}$, there exists a unique coalition $\mathcal{S}' \subseteq \mathbb{O}' \setminus \{A \setminus \{x\}, B\}$ such that $\mathcal{S} = \mathcal{S}'$, and vice versa. Based on these two observations, Equation 4 can be further rewritten to

$$\begin{aligned} &power_{A\to B}^{\mathbb{O}}(x) \\ =&\sum_{\mathcal{S}\subseteq\mathbb{O}\setminus\{B,A\setminus\{x\}\}} \frac{U(\mathcal{S}\cup\{B\})}{(|\mathcal{S}|+1)\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}} - \sum_{\mathcal{S}\subseteq\mathbb{O}\setminus\{B,A\setminus\{x\}\}} \frac{U(\mathcal{S}\cup\{A\setminus\{x\}\}}{(|\mathcal{S}|+1)\binom{|\mathbb{O}|-1}{|\mathcal{S}|+1}} \end{aligned} \tag{5}$$

To enable Monte Carlo approximation, using Corollary 1, we have:

$$\begin{aligned} &power_{A\to B}^{\mathbb{O}}(x) = \psi_{\mathbb{O}'}(B\cup\{x\}) - \psi_{\mathbb{O}'}(A\setminus\{x\}) \\ =&\psi_{\mathbb{O}'}(B) - \psi_{\mathbb{O}'}(A\setminus\{x\}) \\ =&\frac{1}{2(|\mathbb{O}'|-1)!}\sum_{\pi\in\Pi(\mathbb{O}')} \frac{1}{|\mathbb{O}'| - |P^{\pi}_{\{A\setminus\{x\},B\}}| - 1}\Big(U(P^{\pi}_{\{A\setminus\{x\},B\}}\cup\{B\}) \\ &- U(P^{\pi}_{\{A\setminus\{x\},B\}}\cup\{A\setminus\{x\}\})\Big) \end{aligned} \tag{6}$$

Notice that, for any permutation $\pi' \in \Pi(\mathbb{O}')$, there exists a unique permutation $\pi \in \Pi(\mathbb{O})$ such that $P^{\pi'}_{\{A\setminus\{x\},B\}} = P^{\pi}_{\{A,B\}}$ and vice versa. Thus, Equation 6 can be further rewritten to

$$\begin{aligned} &power_{A\to B}^{\mathbb{O}}(x) \\ =&\frac{1}{2(|\mathbb{O}|-1)!}\sum_{\pi\in\Pi(\mathbb{O})} \frac{U(P^{\pi}_{\{A,B\}}\cup\{B\}) - U(P^{\pi}_{\{A,B\}}\cup\{A\setminus\{x\}\}}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} \end{aligned} \tag{7}$$

Based on Corollary 2 and Equation 7, we have the following Monte Carlo estimation.

Theorem 4 (MC-Common Entry). *For $x \in A \cap B$ and a uniform sample of permutations $X \subseteq \Pi(\mathbb{O})$,*

$$\widehat{power_{A\to B}^{\mathbb{O}}}(x) = \frac{|\mathbb{O}|}{2|X|}\sum_{\pi\in X} \frac{U(P^{\pi}_{\{A,B\}}\cup\{B\}) - U(P^{\pi}_{\{A,B\}}\cup\{A\setminus\{x\}\}}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1}$$

is an unbiased estimation of $power_{A\to B}^{\mathbb{O}}(x)$. □

3.3.2 The power of a Differential Data Entry. Suppose $x \in A \setminus B$, that is, x is a data entry that A has but B does not. In such a case, we call x a **differential data entry**. Moving x from A to B may not only reduce the contributions from A to coalitions but may also improve those from B. Using the exact same logic and theorems as above, we get the following Monte Carlo estimation:

Theorem 5 (MC-Differential Item). *For $x \in A \setminus B$ and a uniform sample of permutations $X \subseteq \Pi(\mathbb{O})$,*

$$\begin{aligned} \widehat{power_{A\to B}^{\mathbb{O}}}(x) =&\frac{|\mathbb{O}|}{2|X|}\sum_{\pi\in X} \frac{1}{|\mathbb{O}| - |P^{\pi}_{\{A,B\}}| - 1} \\ &\cdot\Big(U(P^{\pi}_{\{A,B\}}\cup\{B\cup\{x\}\}) - U(P^{\pi}_{\{A,B\}}\cup\{A\setminus\{x\}\})\Big) \end{aligned}$$

is an unbiased estimation of $power_{A\to B}^{\mathbb{O}}(x)$. □

3.4 The SV-Exp Algorithm

After carefully assembling all the necessary components, we are now ready to introduce our greedy algorithm for a counterfactual Shapley value explanation, SV-Exp.

3.4.1 Framework. SV-Exp works in two phases. In the first phase, we conduct Monte Carlo estimation to find the top-1 data entry in A that has the largest power. Then, in the second phase, we move the top-1 data entry from A to B. After moving the top-1 data entry from A to B, if $\psi(A) > \psi(B)$ still holds, we repeat the process, that is, finding the next top data entry and moving it from A to B, until the Shapley value relationship is reversed. The pseudocode is shown in Algorithm 3.

Algorithm 3: The SV-Exp Algorithm

Input: the same as Algorithm 1 and, in addition, a confidence threshold $0 < \delta < 1$, and a threshold for the top-1 data entry's confidence interval ϵ
Output: An approximation solution to Equation 2
Counterfactual explanation $\Delta A \leftarrow \emptyset$;
stimate $d = \psi_{\mathbb{O}'}(A \setminus \Delta A) - \psi_{\mathbb{O}'}(B \cup \Delta A)$ using Monte Carlo based on Corollary 2
while *d is not converged and $d > 0$* **do**
  // Phase 1: finding the best data entry in A
  for $x \in A$ **do**
    $power_x \leftarrow 0$ and δ-confidence interval $interval_x \leftarrow [0, \infty]$;
  while *the δ-confidence interval of the top-1 entry in A has size larger than ϵ* **do**
    use Thompson sampling to find the approximate best data entry x_{best}, let p be the estimated probability that x_{best} is indeed the best entry
    $x \leftarrow x_{best}$ with probability p and a random data entry in A other than x_{best} with probability $(1-p)$
    draw a uniform sample of permutations from $\Pi(\mathbb{O})$;
    estimate and update $power^{\mathbb{O}}_{A \to B}(x)$ and the δ-confidence interval using Theorems 4 and 5;
  $x \leftarrow$ top-1 data entry in A
  // Phase 2: producing counterfactual explanation in a greedy manner
  $\Delta A \leftarrow \Delta A \cup \{x\}$
  $\mathbb{O} \leftarrow (\mathbb{O} \setminus \{A, B\}) \cup \{A \setminus \{x\}, B \cup \{x\}\}$;
  $A \leftarrow A \setminus \{x\}, B \leftarrow B \cup \{x\}$
  estimate $d = \psi_{\mathbb{O}'}(A \setminus \Delta A) - \psi_{\mathbb{O}'}(B \cup \Delta A)$ using Monte Carlo based on Corollary 2
  if *d converged and $d < 0$* **then**
    return ΔA

3.4.2 Phase 1. In Phase 1, SV-Exp iteratively draws a uniform sample of permutations and then uses the Monte Carlo approach to estimate the power value of a selected data entry in A so that the δ-confidence interval is no more than ϵ (e.g., $\delta = 95\%$ and $\epsilon = 0.01$).

We use **Thompson sampling** [50, 55, 56] to accommodate the explore-exploit nature of the problem. The explore part of the problem arises from the fact that we must sample and estimate the powers of different data entries to find the "true" best data entry. The exploit part of the problem arises from our desire to actually confirm that the data entry currently ranked first is *indeed* the best, which we can only find out by continuing to sample for that data entry and becoming more confident about its power. Thus, there is a trade-off between sampling heavily for our current-ranked-first data entry and sampling widely for other data entries in A's data in the case another entry happens to be a better top-1 item. Thompson sampling addresses both these concerns, making it an effective algorithmic choice for this problem.

Thus, we begin by sampling a small amount of permutations and estimating the differential Shapley value for each entry to get our prior. Then, we sort the data entries by power and pick a data entry for further sampling. We pick this data entry by approximating a normal distribution over the power of each data entry using the current means and standard deviations of each draw, drawing one random power value from each distribution, sorting the randomly drawn power values (one associated with each data entry $x \in A$), and selecting the data entry associated with the best power.

For the selected data entry, we draw a uniform batch of samples of its powers using Theorems 4 and 5, update the differential Shapley value, sort, and check if the power of the first-place data entry has δ-confidence interval no more than ϵ. This Bayesian approach to ranking the data entries is efficient due to our objective of finding the top-1 data entry in every round. This means that we are unlikely to have to fully estimate the power for every single data entry. This is a huge advantage of Thompson sampling and is extremely time efficient compared to the frequentist approach, which would require sampling the powers of every single data entry the same amount of times until the first-place and second-place data entries had significantly different powers. With Thompson sampling, we utilize our posterior beliefs about the best data entry every single time we sample.

At the end of Phase 1, we have our estimated best data entry to shift.

3.4.3 Phase 2. In this phase, we move the top-1 data entry from A to B. After the shift, we check whether the Shapley value relationship is reversed. We use Corollary 2 to check whether with δ-confidence, $\psi(A') < \psi(B')$ holds. If so, we have successfully obtained a counterfactual explanation, and the algorithm terminates. Otherwise, we repeat phase 1. We continue this iteration between Phase 1 and Phase 2 until a valid counterfactual is found.

4 EXPERIMENTS

In this section, we report experimental results on three real datasets to examine the effectiveness of the counterfactual explanation of Shapley values and the efficiency of our method.

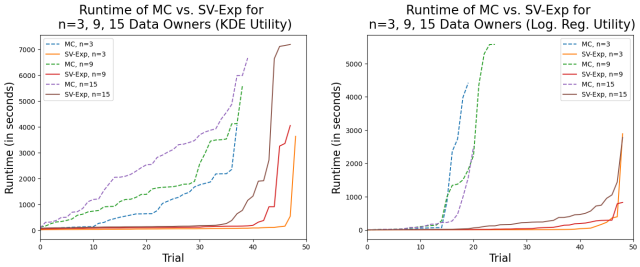
4.1 Experimental Setup

The experiments were run on the Duke Computing Cluster (DCC) using Slurm. We used nodes from the 10x TensorEX TS2-673917-DPN Intel Xeon Gold 6226 Processor, 2.7Ghz (768GB RAM 48 cores). Each of these machines has 2 Nvidia GeForce 2080 RTX Tis.

We implemented and compare two methods, the Monte Carlo baseline (Algorithm 2, denoted by MC) and our SV-Exp algorithm (Algorithm 3). Both were implemented in Python.

We utilized three different datasets. The Breast Cancer Wisconsin Dataset from the UCI Machine Learning Repository [63] has 455 records for training and 63 records for testing, both in 32 attributes. The Boston Housing Prices dataset accessed through Kaggle [25] has 354 records for training and 152 records for testing, both in 14 attributes. The Hotel Reservations Dataset, also accessed through Kaggle [7], has 800 records for training and 200 records for testing in 19 attributes. This 1,000-record dataset was a uniform sample without replacement of the original 36,275 records, which was then split using the train-test-split function from Sklearn (https://scikit-learn.org).

The utility of a set of data entries is as follows: given some task (kernel density estimation, logistic regression, etc.), we use the set of data entries as a training set for the specified task. The utility is exactly the measured performance of the task on the test set. We



(a) KDE (b) Logistic Regression

Figure 1: Runtime for $n = 3, 9, 15$ data owners for different utility functions. For each method, trials were sorted in runtime ascending order.

used four different tasks and the associated utility functions in our experiments: kernel density estimation (η−sum-of-absolute errors), logistic regression prediction (η−log loss), random forest regression (η−MSE), and linear regression (η−MSE), where η is a sufficiently large number so that the utility value is non-negative.

4.2 Efficiency

We test the runtime of the MC and the SV-Exp algorithms using the Breast Cancer Wisconsin dataset. We set the number n of data owners to 3, 6, 9, 12, and 15, respectively. For each set of parameters in the Breast Cancer dataset, 50 trials were run. We use KDE and logistic regression as our two utility functions, as KDE is a very local function relative to logistic regression. We diversify our utility functions to show the generalizability of our methods.

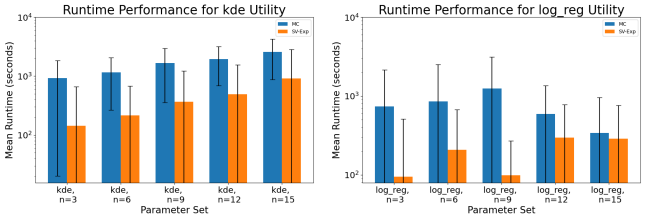
Each data owner's data was drawn uniformly without replacement from the whole training set. The size of each data owner's dataset is distributed uniformly over the range $[1, 455]$. We allow each run to take up to 7,200 seconds. If a method in a test cannot finish in 7,200 seconds, then the trail is marked as timeout.

Limited by space, we only show the results of $n = 3, 9, 15$ of KDE and logistic regression in Figure 1. SV-Exp is much faster than MC for every set of parameters, as MC times out in many trials.

The general trend is that the runtime increases as the number of data owners n increases. When there are many data owners with uniformly distributed size and uniformly distributed data entries, each owner's marginal contribution lessens as the number of data owners increases. Thus, we expect counterfactual sizes to be small in the presence of many data owners. When there are a small number of data owners, each owner's marginal contribution is highly correlated with the size of their dataset, so we expect a large counterfactual between owners.

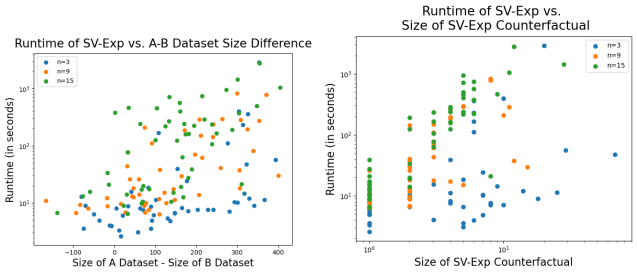
Figure 2 shows the average runtime with respect to the number of data owners in all 5 settings. SV-Exp consistently outperforms MC in mean runtime. The standard deviations of both methods are big, as expected, due to the exponential nature of the problem and the randomness in how data owners A and B are chosen and the sampling of data coalitions.

Figure 3 shows the runtime with respect to the size difference between A and B, and the size of counterfactual. Only the results of SV-Exp were shown since the majority of MC trials timed out.



(a) KDE (b) Logistic Regression

Figure 2: Runtime statistics, where the Y-axis is in logarithmic scale.



(a) Size of $|A| - |B|$ (b) Size of counterfactual exp

Figure 3: Runtime with respect to size difference between A and B, and size of counterfactual. Only the results of SV-Exp were shown since the majority of MC trials timed out.

Figure 3(a) clearly shows that the runtime is roughly positively correlated to the size difference between A and B. The bigger the difference, likely the longer the runtime. Note that B may have more data entries than A and still may have a lower Shapley value (negative x-values), but these situations are often flipped very easily because A may own one or two very powerful data entries that are driving its higher Shapley value. This is a perfect example of the motivation for finding a counterfactual explanation, which can help us glean scenarios where there are "star player" data entries. Figure 3(b) demonstrates the power of the SV-Exp algorithm. Smaller counterfactuals take shorter runtime. The SV-Exp algorithm has no problem in handling counterfactuals as large as 90 data entries within the timeout limit 7,200 seconds, even with a large number of data owners.

4.3 Effect of Number of Data Owners

How does the size of the counterfactual explanation change as the number of data owners increases? Since MC times out in many trials, we only report the results achieved by SV-Exp.

Table 1 shows the average size of the counterfactual explanation with respect to the number of data owners in uniformly distributed data. As the number of data owners increases, each owner's data most likely becomes less critical, and thus the counterfactual explanation between two data owners will contains less data entries on average. Because each data owner's size is uniformly redrawn for each trial, the standard deviation of the counterfactual explanation is large especially when n is small. Note that the standard deviation

Table 1: The average size of counterfactual explanations with respect to number of data owners, when data entries are assigned to data owners randomly in uniform distribution. Two different utility functions KDE and Logistic Regression (LG) are used. Standard deviations are reported in parentheses.

n	SV-Exp-KDE	SV-Exp-LG
3	2.63 (6.31)	7.02 (10.80)
6	2.92 (6.48)	5.40 (8.83)
9	2.37 (4.34)	5.22 (13.31)
12	1.93 (2.09)	3.55 (2.73)
15	1.68 (1.74)	4.06 (4.38)

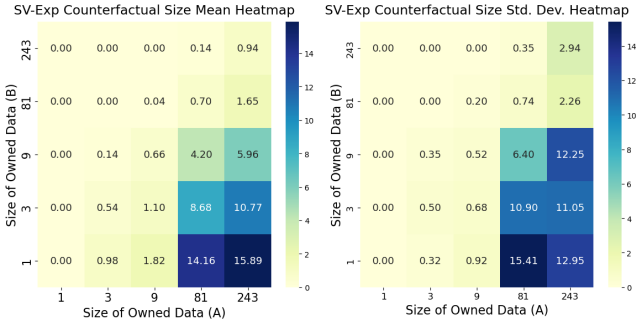


Figure 4: Size of counterfactual explanations when data entries are assigned to data owners following the Zipfian Distribution. In each subfigure, the left shows the mean and the right shows the standard deviations.

also tends to decrease as n increases—the more data owners, the higher the probability that a pair of randomly chosen owners have similar Shapley values.

4.4 Effect of Allocation Among Data Owners

To examine the effect of various allocations among data owners, we use the Zipfian distribution, where the size of each data owner's dataset is drawn from this distribution. This setting simulates real-world scenarios where a few large data owners hold most of the data, while many smaller players have much less.

We report the results on two data owners in experiments, one with a dataset size n_1 such that $log_a n_1 = k_1$ and another with a size n_2 such that $\log_a n_2 = k_2$. To match the scale of the Breast Cancer Wisconsin training dataset, we set $a = 3$ and let $k \in \{0, \ldots, 4\}$. We fix the number of data owners to 9 (the median value from the uniform distribution experiments) and test the 25 possible pairings of data owner sizes for all values of k over 50 trials. The results are shown in Figure 4. Note that the left upper triangle of the matrices have counterfactual explanations of size 0 because A and B fail to satisfy the precondition that $\psi(A) > \psi(B)$. The counterfactual explanations of the largest sizes happen when A owns a large portion of the data and B owns a small portion. We see from Figure 4 that in a case where the whole data set has 455 training records and 63 testing records, the counterfactual explanation computed by SV-Exp between a data owner A who owns more than half of the dataset and a data owner B who owns only one entry comes down to almost 16 data entries on average. Those 16 estimated "best" data entries provide insight on the data records that are contributing most to the difference in the predictive power of data owner A versus data owner B. Similarly to the cases in the uniform distribution case, the standard deviation of the size of counterfactual explanation increases as the size of the counterfactual explanation increases, demonstrating the uncertainty in the difference of Shapley values at high levels of disparity between data owners.

We observe an interesting case where the size of counterfactual explanation between a data owner A having 81 data entries against another data owner B having only one data entry is larger than that in another case where $|A| = 243$ and $|B| = 3$. This speaks to some notion of decreasing returns in the Shapley value (3 items has a lot more power than 1 item, but 243 items does not have much more power than 81). Please note that the size of counterfactual explanations is very sensitive to the detailed data assignments.

4.5 Accuracy in Finding Counterfactuals

We use a small random subset of the Breast Cancer dataset to compare the counterfactuals given by the Brute-Force algorithm (Algorithm 1) with exact Shapley value computation, which are indeed the ground truth. We compare with the results from the two approximation algorithms, MC (Algorithm 2) and SV-Exp (Algorithm 3). To allow the Brute Force algorithm to finish, we extract a uniform random sample of size 40 as the training data set and another disjoint uniform random sample of size 10 as the test set. We run the experiments with 3 data owners. The data owners are assigned the data from a uniform distribution, two utility functions (KDE and Logistic Regression) were used, and 50 trials were run for each utility function. The results are shown in Tables 2 and 3.

As expected, when dataset sizes are tiny, MC approximates both a similar size and similar data records as the Brute Force algorithm because it evaluates subsets of data owner A in the same order as the brute force method (size-ascending). MC even tends to underestimate the size (finishes early) compared to SV-Exp, as there is some level of error when checking if the differential Shapley is negative. Due to SV-Exp's greedy nature, it is also expected that it will overestimate the sizes of counterfactuals.

The average counterfactual length and Jaccard similarity are not agnostic to utility. Using Logistic Regression as the utility function leads to more similarity between the counterfactuals from the two approximation algorithms compared to the exact algorithm than using KDE.

The results show that on tiny datasets with a small number of data owners, MC provides a better approximation of the ground-truth. However, when the dataset size and the number of data owners increase, MC quickly loses the edge due to its weak scalability. Moreover, MC tends to underestimate the size of counterfactuals. In the context of obtaining data in data markets, a slight overestimation of the counterfactual size is most likely more beneficial than an underestimation—-if flipping the Shapley value is the goal, it is better to err on the side of being sure that the amount of data bought will successfully flip the Shapley value than to not buy enough data and still lose to the opposing data owner.

Table 2: Average counterfactual lengths across the three algorithms: BF (Brute-Force), MC (Monte-Carlo), SV-Exp. Standard deviations are given in parentheses.

Utility Function	BF	MC	SV-Exp
KDE	2.50 (1.73)	2.14 (1.19)	4.20 (4.68)
Log. Reg.	2.83 (1.62)	2.60 (1.30)	3.36 (2.68)

Table 3: Average Jaccard similarity indices of counterfactuals between BF (Brute-Force) and MC (Monte-Carlo) as well as BF and SV-Exp. Standard deviations are given in parentheses.

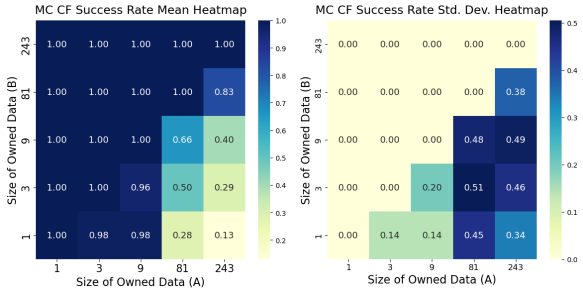
Utility Function	(BF, MC)	(BF, SV-Exp)
KDE	0.89 (0.18)	0.32 (0.20)
Log. Reg.	0.92 (0.13)	0.41 (0.18)

Table 4: Success rates of finding counterfactual explanations under uniformly distributed data owners in KDE and Logistic Regression (LG) tasks. Standard deviations are given in parentheses.

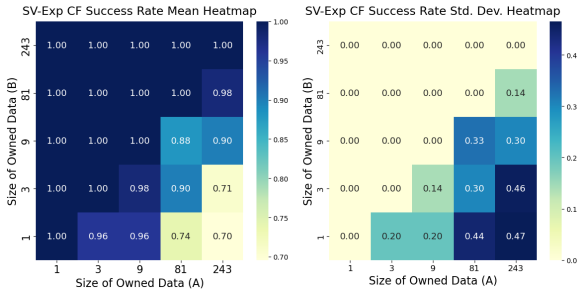
n	MC-KDE	SV-Exp-KDE	MC-LG	SV-Exp-LG
3	0.95 (0.23)	0.98 (0.14)	0.90 (0.31)	0.88 (0.33)
6	0.98 (0.16)	1.00 (0.00)	0.88 (0.34)	0.98 (0.15)
9	1.00 (0.00)	1.00 (0.00)	0.88 (0.33)	0.92 (0.28)
12	0.94 (0.24)	0.96 (0.21)	0.81 (0.40)	0.86 (0.35)
15	0.93 (0.27)	0.98 (0.15)	0.67 (0.48)	0.92 (0.28)

Let us now compare using datasets whose sizes are not tiny whether the approximated counterfactuals by MC and SV-Exp actually reverse the Shapley value relation successfully. Specifically, after each trial, $\psi(A \setminus \Delta A) - \psi(B \cup \Delta A)$ was estimated using Monte Carlo, where ΔA is the answer generated by MC or SV-Exp. If the difference was negative, the trial was marked a success and a failure otherwise. Both MC and SV-Exp were set to time out at 7,200 seconds. Those that timed out were not counted in the comparison: only successful trials were used in calculating the average success rate of the methods so as to not doubly penalize MC for timeouts. Note that our measure for "success" is also an approximation, as we cannot check the exact Shapley value difference when the datasets are not tiny. However, checking the approximate differential Shapley value after the approximate ΔA's are produced still gives a good sense about whether our resulting counterfactuals reverse the Shapley value relation, since the Monte Carlo estimation is the state-of-the-art approach in practice.

Table 4 shows the success rate with respect to the number of data owners under uniform distribution. For every set of parameters except for 3 owners on the Logistic Regression task, SV-Exp outperformed MC. SV-Exp achieves not only higher success rates but also smaller standard deviations. Additionally, it is clear that when the number of data owners increases, MC becomes less and less successful (its success rate with 15 data owners on the Logistic Regression task is only 67%). The comparison clearly shows the practical value of SV-Exp.



(a) MC Success Rate



(b) SV-Exp Success Rate

Figure 5: Statistics of the success rate under the Zipfian distribution of data owners. In each subfigure, the mean is shown in the left and the standard deviation is shown in the right.

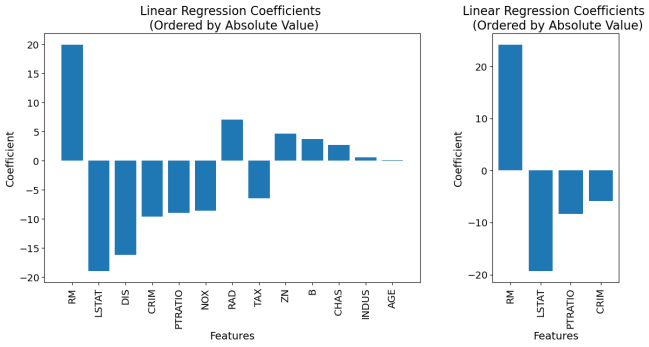
The difference in performance between the two methods follows from the fact that SV-Exp approximates the *most differential* subset to shift–the returned counterfactual reduces $\Psi(A \setminus \Delta A, B \cup \Delta A')$ as much as possible. The high success rates and low standard deviations in Table 4 demonstrate SV-Exp's efficiency and consistency in finding a successful counterfactual.

Figure 5 shows the success rate with respect to data allocation among data owners using the Zipfian distribution. Again, the number of data owners is set to 9. When the size difference between two data owners becomes large, the problem becomes challenging for both methods as the estimation error may accumulate after many entries are moved to a counterfactual. SV-Exp is still able to achieve a much higher success rate than MC.

4.6 Case Study 1: Counterfactual Explanations and Feature Selection on the Boston Housing Prices Dataset

In this section, we conduct an interesting case study showing counterfactual explanation can be used as feature selection. In this case study, a dataset is partitioned *vertically* using the Boston Housing Prices dataset–different data owners own different subsets of attributes in a data set. We will conduct another case study where a dataset is partitioned horizontally in Section 4.7.

The features of the Boston Housing Dataset include CRIM (per capita crime rate by town), ZN (proportion of residential land zoned



(a) Full feature set **(b) Four features**

Figure 6: Coefficients of a linear regression on the Boston Housing Prices Dataset.

for lots over 25,000 sq.ft.), INDUS (proportion of non-retail business acres per town), CHAS (Charles River dummy variable: 1 if tract bounds river, 0 otherwise), NOX (nitric oxides concentration in parts per 10m), RM (average number of rooms per dwelling), AGE (proportion of owner-occupied units built prior to 1940), DIS (weighted distances to five Boston employment centres), RAD (index of accessibility to radial highways), TAX (full-value property-tax rate per $10k), PTRATIO (pupil-teacher ratio by town), B ($1000(Bk - 0.63)^2$) where Bk is the proportion of blacks by town, LSTAT (% lower status of the population), and MEDV (Median value of owner-occupied homes in thousands of dollars). MEDV serves as our target variable.

We assign the features to three owners according to their meaning (semantically): (1) property features RM and AGE to owner P; (2) geographic features CHAS, NOX, DIS, and RAD to owner G; and (3) social features CRIM, B, ZN, INDUS, TAX, PTRATIO, LSTAT to owner S.

In this experiment, to keep the discussion easy to understand, we use the entire dataset as our training set and simply use the training error to define our utility function. Figure 6(a) shows the coefficients of the attributes in the linear regression on all attributes. We sort all attributes in the coefficient descending order in the figure.

Owner S has a higher Shapley value than owner P in the task of linear regression. Which features are in the counterfactual explanation? Since among the social features owned by S, LSTAT and CRIM have the largest absolute values of the coefficients in linear regression, an intuitive guess is that LSTAT and CRIM may be the first two features selected for the counterfactual explanation. Surprisingly, {LSTAT, PTRATIO} is the counterfactual explanation returned by SV-Exp and does revert the Shapley value relation.

We further investigate the correlation among several features as shown in Table 5. Interestingly, CRIM is more correlated with LSTAT than PTRATIO is. This means, after LSTAT is chosen in the counterfactual explanation, choosing CRIM is not as effective as choosing PTRATIO in further reducing the utility of owner S and enhancing the utility of owner P. Indeed, if we conduct a linear regression using only four features, RM, which is the dominating feature owned by P, LSTAT, PTRATIO, and CRIM, the absolute value

Table 5: Correlation matrix between four selected features in the Boston Housing Prices dataset.

	LSTAT	CRIM	RM	PTRATIO
LSTAT	1.00	0.46	-0.61	0.37
CRIM	0.46	1.00	-0.22	0.29
RM	-0.61	-0.22	1.00	-0.36
PTRATIO	0.37	0.29	-0.36	1.00

Table 6: Shapley values of the month's data estimated using Monte Carlo Sampling, where utility = η−log-loss, $\eta = 20$, and the task is Logistic Regression.

Jan	Feb	Mar	Apr	May	Jun
0.71	1.58	1.67	1.78	1.72	1.87
Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
1.86	1.84	1.79	1.80	1.83	1.79

of coefficient of PTRATIO is larger than that of CRIM, as shown in Figure 6(b).

Essentially, SV-Exp automatically feature selects one of the variables when there are different levels of correlation, simplifying how we deal with multicollinearity and understand the features of datasets with one algorithm. We make a case for the Shapley counterfactual as a tool for feature selection.

In this illustrative case study, the counterfactual explanation of Shapley values not only shows the comparative advantages between two feature sets but also facilitates the identification and enhancement of features within one set based on those of the other.

4.7 Case Study 2: Counterfactual Explanations and Differences of Distributions on the Hotel Reservations Dataset

In this case study, we partition the Hotel Reservation dataset *horizontally*. In this experiment, we use logistic regression to predict the probability of whether a client will cancel the reservation. We assign to each data owner all data of one month and thus we have in total 12 data owners. The research question aims to determine which months' data contribute more in modeling reservation cancellations. By computing the counterfactual explanations we are interested in understanding the differences in contribution between two months. We perform pairwise experiments, testing all 144 possible pairs of months and thus data owners over 10 trials per pairing. The diagonals represented the cases where A and B are the same data owner and were thus moot experimentally. The sizes of the data subsets from January to December owned by the 12 data owners are 25, 37, 52, 45, 59, 61, 65, 79, 105, 117, 73, and 82, respectively. The Shapley value of each month's data is detailed in Table 6.

Figure 7 shows the initial difference in Shapley value between different months. Theoretically the matrix should be symmetric. However, due to the estimation errors in practice, the matrix is not perfectly so.

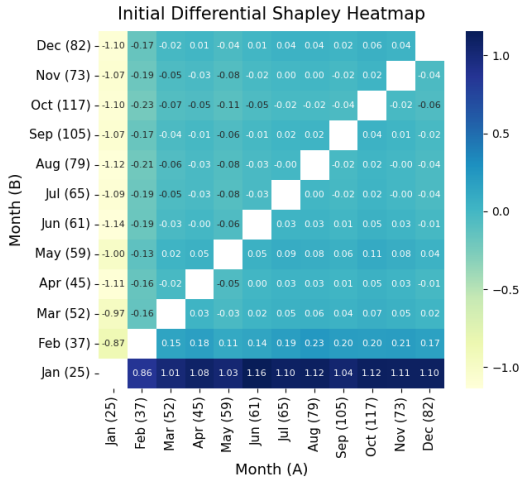


Figure 7: Initial Differential Shapley ($\Psi_{\mathbb{O}}(A, B)$) Between Data Owners.

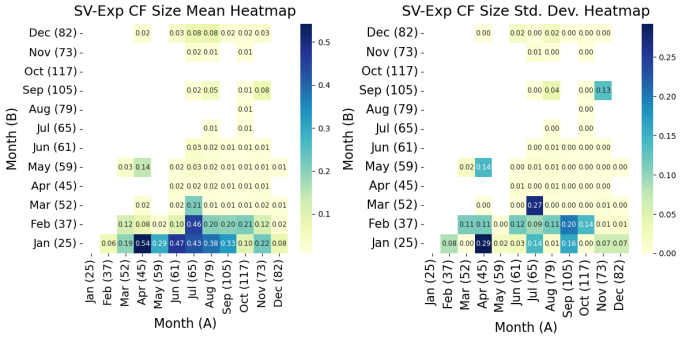


Figure 8: Average size of counterfactuals (from the SV-Exp algorithm) between data owners in percentage of A's original data. Zero values are left blank for more visual clarity, and sizes of each month's data are in parentheses on the axes labels.

The differential Shapley values between all months and January are similar, as shown in Figure 7, but the sizes of the Shapley counterfactuals are vastly different (Figure 8). This is a clear example showing that the similar differential Shapley values may mask important differences about how advantages between pairs of data owners may be established. Take March and April as a concrete example. Both of their Shapley values are about 1.0 larger than the Shapley value of January. Moreover, their dataset sizes are also close, March having 52 data entries and April 45. But the counterfactual explanation of March against January takes about 19% of the data in March, while the counterfactual explanation of April against January takes 54% of the data in April. This indicates that the data in April is much more "robust" than that in March with respect to the data in January—most likely, the data distribution of April is more different than that of January, while the data distribution in March is more similar. To verify if this insight holds, we use the Wasserstein distance between two months to understand their

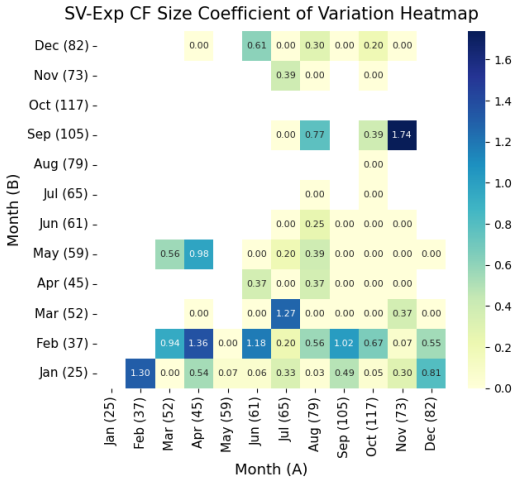


Figure 9: Coefficient of Variation on the average size of counterfactuals produced by SV-Exp. Coefficient of Variation is the ratio of the standard deviation to the mean.

relative differences in distribution. The Wasserstein distance, or the earth mover's distance, measures the estimated "cost" of turning one probability distribution into the other. The Wasserstein distance between March and January is 3.24, while the Wasserstein distance between April and January is 6.52. The much larger Wasserstein distance between April and January confirms our conjecture. Just for the reader's reference, the Wasserstein distance between March and April is 5.59.

This case study shows another interesting use of the counterfactual explanation of Shapley values, capturing and understanding the differences of data distributions between pairs of data owners.

4.8 Stability of Approximate Counterfactual Explanations in SV-Exp

SV-Exp approximates counterfactual explanations of Shapley values. We now explore one last question: given that SV-Exp draws random samples of coalitions, how stable are the approximate counterfactuals? In this question, we only consider those cases where the differential Shapley values are large, since a small differential Shapley value can be reverted easily using any small counterfactual. For example, if the initial differential Shapley is only very slightly positive and any random item from A could flip the Shapley relation, we do not get much more information about A and B from the counterfactual. Thus, since we will be studying the content of the outputted SV-Exp counterfactuals, we focus on those that are larger—in these cases, the data entries making up the counterfactuals will have some level of significance. We still use the Hotel Reservations dataset as in Section 4.7.

We look at the *coefficient of variation*, also known as normalized root-mean-square deviation (NRMSD), percent RMS, and relative standard deviation (RSD), which is defined as the ratio of the standard deviation over the mean. Interestingly, in Figure 9, there are many cases where the coefficient of variation is small, such as March, May, June, August, and October against January. In those

cases, SV-Exp often chooses the same subset of data entries in many trials. As a concrete example, in the case of the counterfactual of May against January, the subset of 18 data entries in May, {162, 628, 711, 651, 69, 63, 678, 379, 281, 524, 74, 355, 406, 439, 758, 662, 615, 609} (the numbers of the record-ids in the dataset) is chosen every time by SV-Exp. This shows that SV-Exp demonstrates surprising consistency on some datasets.

There are also some cases where the coefficient of variation is large. For example, April has a similar Shapley value as March (Table 6 and Figure 7), but its coefficient of variation is high as it has a large standard deviation. Looking at the detailed counterfactual results, we find that counterfactual explanations of April against January can be as small as size 13 ({528, 289, 312, 218, 204, 725, 398, 559, 671, 301, 710, 774, 572}) and as large as size 40 ({572, 33, 462, 312, 588, 692, 792, 475, 559, 344, 192, 785, 301, 155, 134, 473, 360, 725, 671, 501, 218, 502, 362, 184, 248, 142, 141, 292, 461, 185, 510, 289, 81, 528, 95, 774, 346, 204, 398, 710}). Moreover, SV-Exp implicitly ranks the data entries from most to least power in the subset, the ranking of the entries is also very different from run to run. This indicates that entries in the data of April may have utilities that are much more sensitive to what coalition the samples drawn.

5 RELATED WORK

In Section 1, we already discuss a series of related work on data markets, the Shapley value, and computation. To the best of our knowledge, we are the first to formulate the problem of counterfactual explanation of the Shapley value. In this section, we focus on briefly reviewing the related work on counterfactual explanations.

Counterfactual explanations [6, 20, 34, 41, 53, 59] have been widely used in interpreting and understanding algorithmic decisions made in many real world applications [6, 20, 34]. Those methods [6, 20, 41, 58] often explain a prediction on a given case using small and interpretable perturbations on the case such that the prediction is changed [41]. For example, Fong and Vedaldi [20] interpret image prediction by identifying the smallest pixel-deletion mask that causes the most significant drop of the prediction score. Akula et al. [6] find image patches that need to be added to or deleted from an input image in order to change the prediction. Van Looveren and Klaise [58] use class prototypes to produce counterfactual explanations that are close to the classifier's training data distribution. Moore et al. [40] propose a method to generate counterfactual explanations from adversarial examples with gradient constraints. Le et al. [34] propose an entropy-based feature selection method to limit the features to be perturbed. Cong et al. [12] compute understandable counterfactual explanations for Kolmogorov-Smirnov Test results. Bajaj et al. [8] provide counterfactual explanations for graph nueral networks.

Surprisingly, the problem of a counterfactual explanation of the Shapley value has not been studied in literature. Moreover, the existing counterfactual explanation methods cannot be applied to explain the Shapley value directly.

6 LIMITATIONS AND EXTENSIONS

Counterfactual explanations of the Shapley value present a novel and significant challenge. This study represents an initial foray into this promising problem. Despite our encouraging advancements, our SV-Exp method remains subject to certain limitations. Chief among these constraints is its scalability. Although SV-Exp demonstrates superior scalability compared to brute force and Monte Carlo methods, it encounters difficulties when confronted with very large datasets. This challenge arises primarily due to the extensive training requirements inherent in SV-Exp. Notably, the utility function, utilized twice for every sample, necessitates the training of a new model for each iteration. Consequently, when handling substantial datasets or data necessitating convolutional neural networks (CNNs) or more intricate training algorithms, SV-Exp may prove inefficient in identifying counterfactual explanations.

Our SV-Exp method can also account for variances of the counterfactual explanation problem for Shapley values. For example, instead of transferring items from A to B, what is the minimal number of items we need to simply delete from A in order for $\phi(A) < \phi(B)$? Essentially, what is the minimal set of entries in A that make up the difference in Shapley value between A and B? This is still an NP-hard problem with a feasible solution (worst case, delete all of A), and our algorithm can easily be extended to approximate a solution for this problem by simply taking away the step where we transfer data entries to B.

7 CONCLUSION

Data valuation is a fundamental mechanism within a data market. While there are more and more studies on efficient data valuation, how we understand and explain data valuation remains an open problem. In this paper, we formulate the problem of counterfactual explanation of the Shapley value in data coalitions, which, to the best of our knowledge, is the first study tackling this important issue. We show the complexity of the problem, propose a series of techniques, and develop a greedy approximation method. Our experiments on real datasets clearly show the efficiency of our approach and the effectiveness of a counterfactual Shapley explanation in interpreting data value, feature selection, and detecting data distribution differences.

Our study illuminates a methodical way to interpret how individual data entries contribute to the overall value of a data owner's dataset in a game theoretic setting. This paper opens a new direction for promising future work. For example, building on the general framework developed in this study, it is interesting to explore more effective and efficient approaches to Shapley value interpretation in specific types of data collaboration, such as data assemblage [37, 38]. Moreover, in some applications, one may be interested in finding coherent subsets of data as counterfactual explanations. It will also be important to explore the explanation of other types of data valuation metrics for data markets.

ACKNOWLEDGMENTS

We thank Kiran Dwivedi and Phillip Si for their invaluable comments and feedback, as well as Eric Zelikman and Kate Donahue for useful discussions. This research is supported in part by a startup grant and a Beyond the Horizon grant by Duke University. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

REFERENCES

[1] Alessandro Acquisti, Curtis Taylor, and Liad Wagman. 2016. The Economics of Privacy. *Journal of Economic Literature* 54, 2 (June 2016), 442–92. https://doi.org/10.1257/jel.54.2.442

[2] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation* (Phoenix, AZ, USA) *(EC'19)*. Association for Computing Machinery, New York, NY, USA, 701–726. https://doi.org/10.1145/3328526.3329589

[3] Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1st ed.). Springer Publishing Company, Incorporated.

[4] Charu C. Aggarwal and Philip S. Yu. 2008. Privacy-Preserving Data Mining: A Survey. In *Handbook of Database Security: Applications and Trends*, Michael Gertz and Sushil Jajodia (Eds.). Springer US, Boston, MA, 431–460. https://doi.org/10.1007/978-0-387-48533-1_18

[5] William Aiello, Yuval Ishai, and Omer Reingold. 2001. Priced Oblivious Transfer: How to Sell Digital Goods. In *Advances in Cryptology - EUROCRYPT 2001, International Conference on the Theory and Application of Cryptographic Techniques, Innsbruck, Austria, May 6-10, 2001, Proceeding (Lecture Notes in Computer Science)*, Vol. 2045. Springer, 119–135. https://doi.org/10.1007/3-540-44987-6_8

[6] Arjun R Akula, Shuai Wang, and Song-Chun Zhu. 2020. CoCoX: Generating Conceptual and Counterfactual Explanations via Fault-Lines.. In *AAAI*. 2594–2601.

[7] Nuno Antonio, Ana de Almeida, and Luis Nunes. 2019. Hotel booking demand datasets. *Data in Brief* 22 (2019), 41–49. https://doi.org/10.1016/j.dib.2018.11.126

[8] Mohit Bajaj, Lingyang Chu, Zi Yu Xue, Jian Pei, Lanjun Wang, Peter Cho-Ho Lam, and Yong Zhang. 2021. Robust Counterfactual Explanations on Graph Neural Networks. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 5644–5655. https://proceedings.neurips.cc/paper/2021/file/2c8c3a57383c63caef6724343eb62257-Paper.pdf

[9] Magdalena Balazinska, Bill Howe, and Dan Suciu. 2011. Data Markets in the Cloud: An Opportunity for the Database Community. *PVLDB* 4, 12 (2011), 1482–1485. http://dblp.uni-trier.de/db/journals/pvldb/pvldb4.html#BalazinskaHS11

[10] Elisa Bertino, Dan Lin, and Wei Jiang. 2008. A Survey of Quantification of Privacy Preserving Data Mining Algorithms. In *Privacy-Preserving Data Mining: Models and Algorithms*, Charu C. Aggarwal and Philip S. Yu (Eds.). Springer US, Boston, MA, 183–205. https://doi.org/10.1007/978-0-387-70992-5_8

[11] Lingjiao Chen, Paraschos Koutris, and Arun Kumar. 2019. Towards Model-based Pricing for Machine Learning in a Data Marketplace. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska (Eds.). ACM, 1535–1552. https://doi.org/10.1145/3299869.3300078

[12] Zicun Cong, Lingyang Chu, Yu Yang, and Jian Pei. 2021. Comprehensible Counterfactual Explanation on Kolmogorov-Smirnov Test. *Proc. VLDB Endow.* 14, 9 (oct 2021), 1583?1596. https://doi.org/10.14778/3461535.3461546

[13] David Dao, Dan Alistarh, Claudiu Musat, and Ce Zhang. 2018. DataBright: Towards a Global Exchange for Decentralized Data Ownership and Trusted Computation. *CoRR* abs/1802.04780 (2018). arXiv:1802.04780 http://arxiv.org/abs/1802.04780

[14] Xiaotie Deng and Christos H. Papadimitriou. 1994. On the Complexity of Cooperative Solution Concepts. *Mathematics of Operations Research* 19, 2 (1994), 257–266. https://doi.org/10.1287/moor.19.2.257 arXiv:https://doi.org/10.1287/moor.19.2.257

[15] Daniel Deutch, Nave Frost, Benny Kimelfeld, and Mikaël Monet. 2022. Computing the Shapley Value of Facts in Query Answering. In *Proceedings of the 2022 International Conference on Management of Data* (Philadelphia, PA, USA) *(SIGMOD '22)*. Association for Computing Machinery, New York, NY, USA, 1570–1583. https://doi.org/10.1145/3514221.3517912

[16] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation*, Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–19.

[17] Raul Castro Fernandez, Pranav Subramaniam, and Michael J. Franklin. 2020. Data Market Platforms: Trading Data Assets to Solve Data Problems. *Proc. VLDB Endow.* 13, 12 (sep 2020), 1933–1947. https://doi.org/10.14778/3407790.3407800

[18] M. A. Ferrag, L. Maglaras, and A. Ahmim. 2017. Privacy-Preserving Schemes for Ad Hoc Social Networks: A Survey. *IEEE Communications Surveys Tutorials* 19, 4 (2017), 3015–3045.

[19] Lisa K. Fleischer and Yu-Han Lyu. 2012. Approximately Optimal Auctions for Selling Privacy When Costs Are Correlated with Data. In *Proceedings of the 13th ACM Conference on Electronic Commerce* (Valencia, Spain) *(EC'12)*. Association for Computing Machinery, New York, NY, USA, 568–585. https://doi.org/10.1145/2229012.2229054

[20] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3429–3437.

[21] Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. 2010. Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Comput. Surv.* 42, 4, Article 14 (June 2010), 53 pages. https://doi.org/10.1145/1749603.1749605

[22] Arpita Ghosh, Katrina Ligett, Aaron Roth, and Grant Schoenebeck. 2014. Buying Private Data without Verification. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation* (Palo Alto, California, USA) *(EC'14)*. Association for Computing Machinery, New York, NY, USA, 931–948. https://doi.org/10.1145/2600057.2602902

[23] Andrew V. Goldberg, Jason D. Hartline, and Andrew Wright. 2001. Competitive Auctions and Digital Goods. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms* (Washington, D.C., USA) *(SODA'01)*. Society for Industrial and Applied Mathematics, USA, 735–744.

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*. MIT press.

[25] David Harrison and Daniel L Rubinfeld. 1978. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* 5, 1 (1978), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

[26] Miguel A Hernán and James M Robins. 2010. Causal inference.

[27] Nick Hynes, David Dao, David Yan, Raymond Cheng, and Dawn Song. 2018. A Demonstration of Sterling: A Privacy-Preserving Data Marketplace. *Proc. VLDB Endow.* 11, 12 (Aug. 2018), 2086–2089. https://doi.org/10.14778/3229863.3236266

[28] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. 2019. Towards Efficient Data Valuation Based on the Shapley Value. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.), Vol. 89. PMLR, 1167–1176. https://proceedings.mlr.press/v89/jia19a.html

[29] Michael I Jordan and Tom M Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260.

[30] Richard M. Karp. 1972. *Reducibility among Combinatorial Problems*. Springer US, Boston, MA, 85–103. https://doi.org/10.1007/978-1-4684-2001-2_9

[31] Javen Kennedy, Pranav Subramaniam, Sainyam Galhotra, and Raul Castro Fernandez. 2022. Revisiting Online Data Markets in 2022: A Seller and Buyer Perspective. *SIGMOD Rec.* 51, 3 (nov 2022), 30–37. https://doi.org/10.1145/3572751.3572757

[32] Yongchan Kwon and James Zou. 2022. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.), Vol. 151. PMLR, 8780–8802. https://proceedings.mlr.press/v151/kwon22a.html

[33] Yongchan Kwon and James Zou. 2023. Data-OOB: Out-of-Bag Estimate as a Simple and Efficient Data Value. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 749, 18 pages.

[34] Thai Le, Suhang Wang, and Dongwon Lee. 2020. GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 238–248. https://doi.org/10.1145/3394486.3403066

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[36] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. 2015. A Theory of Pricing Private Data. *ACM Trans. Database Syst.* 39, 4, Article 34 (Dec. 2015), 28 pages. https://doi.org/10.1145/2691190.2691191

[37] Xuan Luo, Jian Pei, Zicun Cong, and Cheng Xu. 2022. On shapley value in data assemblage under independent utility. *Proc. VLDB Endow.* 15, 11 (jul 2022), 2761–2773. https://doi.org/10.14778/3551793.3551829

[38] Xuan Luo, Jian Pei, Cheng Xu, Wenjie Zhang, and Jianliang Xu. 2024. Fast Shapley Value Computation in Data Assemblage Tasks as Cooperative Simple Games. In *Proceedings of the 2024 ACM SIGMOD International Conference on Management of Data*. Santiago, Chile. https://doi.org/10.1145/3639311

[39] Sasan Maleki, Long Tran-Thanh, Greg Hines, Talal Rahwan, and Alex Rogers. 2013. Bounding the Estimation Error of Sampling-based Shapley Value Approximation. arXiv:1306.4265 [cs.GT]

[40] Jonathan Moore, Nils Hammerla, and Chris Watkins. 2019. Explaining deep learning models with constrained adversarial examples. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 43–56.

[41] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal Interpretability for Machine Learning-Problems, Methods and Evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.

[42] Kobbi Nissim, Salil Vadhan, and David Xiao. 2014. Redrawing the Boundaries on Purchasing Data from Privacy-Sensitive Individuals. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science* (Princeton, New Jersey, USA) *(ITCS'14)*. Association for Computing Machinery, New York, NY, USA, 411–422. https://doi.org/10.1145/2554797.2554835

[43] Chaoyue Niu, Zhenzhe Zheng, Fan Wu, Shaojie Tang, Xiaofeng Gao, and Guihai Chen. 2018. Unlocking the Value of Privacy: Trading Aggregate Statistics over Private Correlated Data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (London, United Kingdom) *(KDD'18)*. Association for Computing Machinery, New York, NY, USA, 2031–2040. https://doi.org/10.1145/3219819.3220013

[44] K. Pantelis and L. Aija. 2013. Understanding the value of (big) data. In *2013 IEEE International Conference on Big Data*. 38–42.

[45] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.

[46] Judea Pearl. 2010. Causal inference. *Causality: objectives and assessment* (2010), 39–58.

[47] J. Pei. 2022. A Survey on Data Pricing: From Economics to Data Science. *IEEE Transactions on Knowledge and Data Engineering* 34, 10 (oct 2022), 4586–4608. https://doi.org/10.1109/TKDE.2020.3045927

[48] Foster Provost and Tom Fawcett. 2013. Data science and its relationship to big data and data-driven decision making. *Big data* 1, 1 (2013), 51–59.

[49] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–58.

[50] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. 2020. A Tutorial on Thompson Sampling. arXiv:1707.02038 [cs.LG]

[51] Fabian Schomm, Florian Stahl, and Gottfried Vossen. 2013. Marketplaces for Data: An Initial Survey. *SIGMOD Rec.* 42, 1 (May 2013), 15–26. https://doi.org/10.1145/2481528.2481532

[52] Lloyd S. Shapley. 1952. *A Value for n-Person Games.* Technical Report P-295. RAND Corporation, Santa Monica, CA. https://www.rand.org/pubs/papers/P0295.html

[53] Kacper Sokol and Peter A Flach. 2019. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. In *SafeAI@ AAAI*.

[54] W. Starr. 2022. Counterfactuals. In *The Stanford Encyclopedia of Philosophy* (Winter 2022 ed.), Edward N. Zalta and Uri Nodelman (Eds.). Metaphysics Research Lab, Stanford University.

[55] William R Thompson. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25, 3-4 (12 1933), 285–294. https://doi.org/10.1093/biomet/25.3-4.285 arXiv:https://academic.oup.com/biomet/article-pdf/25/3-4/285/513725/25-3-4-285.pdf

[56] William R. Thompson. 1935. On the Theory of Apportionment. *American Journal of Mathematics* 57, 2 (1935), 450–456. http://www.jstor.org/stable/2371219

[57] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing Machine Learning Models via Prediction APIs. In *Proceedings of the 25th USENIX Conference on Security Symposium* (Austin, TX, USA) *(SEC'16)*. USENIX Association, USA, 601–618.

[58] Arnaud Van Looveren and Janis Klaise. 2019. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584* (2019).

[59] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[60] Jiachen T Wang, Yuqing Zhu, Yu-Xiang Wang, Ruoxi Jia, and Prateek Mittal. 2023. Threshold KNN-Shapley: A Linear-Time and Privacy-Friendly Approach to Data Valuation. *arXiv preprint arXiv:2308.15709* (2023).

[61] Xintao Wu, Xiaowei Ying, Kun Liu, and Lei Chen. 2010. A Survey of Privacy-Preservation of Graphs and Social Networks. In *Managing and Mining Graph Data*, Charu C. Aggarwal and Haixun Wang (Eds.). Springer US, Boston, MA, 421–453. https://doi.org/10.1007/978-1-4419-6045-0_14

[62] Bin Zhou, Jian Pei, and WoShun Luk. 2008. A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data. *SIGKDD Explor. Newsl.* 10, 2 (Dec. 2008), 12–22. https://doi.org/10.1145/1540276.1540279

[63] Matjaz Zwitter and Milan Soklic. 1988. Breast Cancer. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C51P4M.