Collaborative Causal Inference with Fair Incentives

Rui Qiao [1] **Xinyi Xu** [1 2] **Bryan Kian Hsiang Low** [1]

Abstract

Collaborative causal inference (CCI) aims to improve the estimation of the causal effect of treatment variables by utilizing data aggregated from multiple self-interested parties. Since their source data are valuable proprietary assets that can be costly or tedious to obtain, every party has to be incentivized to be willing to contribute to the collaboration, such as with a guaranteed fair and sufficiently valuable reward (than performing causal inference on its own). This paper presents a reward scheme designed using the unique statistical properties that are required by causal inference to guarantee certain desirable incentive criteria (e.g., fairness, benefit) for the parties based on their contributions. To achieve this, we propose a data valuation function to value parties' data for CCI based on the distributional closeness of its resulting treatment effect estimate to that utilizing the aggregated data from all parties. Then, we show how to value the parties' rewards fairly based on a modified variant of the Shapley value arising from our proposed data valuation for CCI. Finally, the Shapley fair rewards to the parties are realized in the form of improved, stochastically perturbed treatment effect estimates. We empirically demonstrate the effectiveness of our reward scheme using simulated and real-world datasets.

1. Introduction

Causal inference estimates the causal effect of treatment variables for some target population(s) and is widely adopted across various fields. In healthcare, hospitals perform causal inference of the efficacy of drugs (Glass et al., 2013; Hernán et al., 2002; Wendling et al., 2018). In agriculture, causal inference is used to determine the achievable growth from applying a particular nutrient (Rubin, 1990; 2005). Various causal inference methods have been designed for interventional data (from experimental trials) and observational data (from historical records). However, the collected data may be of low quality: Data sparsity and non-representativeness for the population of interest are the major obstacles to accurately estimating the *treatment effect* (e.g., efficacy of drugs). For example, patients prefer to visit the hospitals nearby, which results in each hospital having few and demographically biased records that give rise to data sparsity and non-representativeness, respectively. If the hospitals perform causal inference individually, then they are likely to get inaccurate treatment effect estimates and potentially fail to prescribe the most efficacious medications (Masic et al., 2008).

Collaborative causal inference (CCI) uses the aggregation of shared data from participating parties (e.g., company, organization, or individual) to overcome the issues of data sparsity and non-representativeness. Consequently, they obtain more accurate and statistically significant treatment effect estimates. Such estimates for medical treatments help doctors improve their prescriptions, while those for nutrients help farmers determine appropriate amounts of nutrients and the associated costs to improve their overall profits from crop yields. By using the data from all parties, simple aggregation or multi-source causal inference (Bareinboim & Pearl, 2016; Dersimonian & Laird, 1986; Guo et al., 2021; Xiong et al., 2021) competitively yield the most accurate and statistically significant estimate (i.e., the most *valuable* estimate whose exact value is to be defined later) which is accessible to every party. However, such classes of methods rely on the willingness of all parties to share their data, which is not always the case. In practice, parties are often *self-interested* (Chalkiadakis et al., 2011; Sim et al., 2020; 2021; Smith, 2018; Tay et al., 2022; Thibaut, 1960) and unwilling to share their valuable and proprietary source data in the collaboration because the process of data collection is costly. So, some parties may consider it *unfair* if the others with less valuable data can benefit equally from the same most valuable estimate as themselves. Without active participation, the amount of data available for causal inference is insufficient to be effective for existing solutions. This motivates the need to promote collaboration with guaranteed *benefit* and *fairness* (Adams, 1963; Konow, 2000; Tabibnia & Lieberman, 2007) for self-interested parties; Appendix A

[1]Department of Computer Science, National University of Singapore, Republic of Singapore. [2]Institute for Infocomm Research, A*STAR, Republic of Singapore. Correspondence to: Xinyi Xu <xinyi.xu@u.nus.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

addresses potential ethical concerns related to fairness.

To establish a fair collaborative framework for causal inference, an important ingredient is a quantitative measure of the value of data called the data valuation function v that can be used to compare the usefulness/utility of different datasets towards estimating the treatment effect of the target population. For instance, when two parties with datasets A and B collaborate, such a measure quantifies their value of data in the form of $v(A)$, $v(B)$, and $v(A \cup B)$. Several existing data valuation functions (Ghorbani & Zou, 2019; Sim et al., 2020; 2022; Wu et al., 2022; Xu et al., 2021b) tend to be highly correlated with validation accuracy. However, they cannot be applied to CCI as ground truth is not available and statistical significance is an emphasis. Performing data valuation in CCI has two unique challenges: Firstly, since the accuracy of the treatment effect and the confidence interval are both important statistical properties of causal inference, we need to formalize the notion of quality of a dataset w.r.t. these properties. Secondly, a proper surrogate to the ground truth is required because the ground truth treatment effect is the quantity to be derived but usually unknown in practice. ① *How then can the value of data be measured in CCI?*

A reward is often used to incentivize each party to collaborate and we adopt a quantitative proxy of it called the reward value. Given a data valuation function for CCI, the reward values need to satisfy certain desirable incentive criteria to encourage participation, which include (a) numerical validity: actual rewards can be realized from the reward values, (b) benefit: parties are guaranteed to perform causal inference at least as well as when without collaboration, otherwise they will not participate due to receiving worse estimates, (c) fairness: parties contributing more valuable datasets should receive more valuable rewards to avoid the free-rider problem (Sim et al., 2020; Tay et al., 2022), and (d) efficiency and group welfare: reward values should be maximized as much as possible such that at least some party can be rewarded with an estimate with the best achievable quality. ② *How can a fair reward scheme be designed to satisfy the above incentive criteria in CCI?*

The reward values determined by the reward scheme will then be used to realize the actual rewards to the parties, each of which corresponds to a treatment effect estimate with a confidence interval. Such an estimate may be of a low fidelity and hence inaccurately indicate that a truly effective treatment effect is not, which is undesirable and can be avoided by imposing explicit constraints. Moreover, since the treatment effect estimates are simple, some parties with less valuable datasets may be able to exploit knowledge of the reward scheme to be unfairly rewarded with more valuable estimates, which can be prevented by using strategies like random perturbation. It is thus challenging to simultaneously preserve the fidelity of the estimate and prevent the reward scheme from being exploited during reward realization. ③ *In practice, how can the rewards to the parties be realized in CCI?*

This paper presents a novel game-theoretic reward scheme to incentivize the collaboration of multiple self-interested parties for causal inference by fairly rewarding them with more valuable treatment effect estimates. In particular, for ①, we use the estimate obtained by utilizing data aggregated from all collaborating parties as the surrogate to the "ground truth". Then, we propose to value each dataset by the distributional divergence between its resulting treatment effect estimate vs. the ground truth surrogate. For ②, we propose a set of desirable incentive criteria based on the divergence-based data valuation, and prove that a variant of the Shapley value (Shapley, 1953) satisfies all criteria and determines the reward value fairly for each party. For ③, we realize the reward to each party with a new treatment effect estimate and the corresponding confidence interval according to its fair reward value. Specifically, we propose a stochastic reward realization strategy with rejection sampling that perturbs the ground truth estimate according to the reward value and additional desirable criteria such as fidelity and information obscurity for causal inference. Our reward scheme is applicable to a broad range of causal inference estimators including observational estimators (Imbens & Rubin, 2015; Robins et al., 1994) and *randomized control trial* (RCT). Our contributions of the work in this paper are summarized as follows:

- We propose to value a party's data using the negative reverse *Kullback–Leibler* (KL) divergence between the distribution of its resulting treatment effect estimate vs. that utilizing the data from all parties (Sec. 4).
- We propose a novel reward scheme using a modified ρ-Shapley fair reward value which satisfies desirable incentive criteria like numerical validity, efficiency, individual rationality, fairness, and group welfare (Sec. 5).
- We propose to realize the reward as a treatment effect estimate using rejection sampling such that the estimate preserves fidelity and obscures the ground truth (Sec. 5.3).
- We empirically demonstrate using simulated and real-world datasets that our CCI framework can fairly reward parties with more valuable treatment effect estimates (Sec. 6).

2. Preliminaries

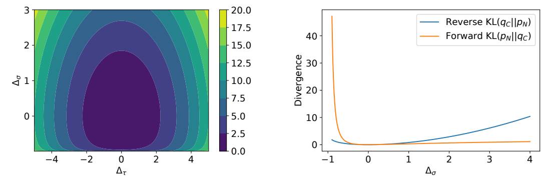
For simplicity, we illustrate our framework based on Neyman-Rubin potential outcome model (Imbens & Rubin, 2015) for single binary treatment. Let $X \in \mathbb{R}^d$ denote the covariates and $Y \in \mathbb{R}$ denote the observed outcome. Let $W \in \{0, 1\}$ denote the treatment variable. For each subject, k in population $\mathcal{P}$, k is in the treatment group if $W_k = 1$

and in control if $W_k = 0$. Let $M \coloneqq M_0 + M_1$ be the total number of samples, where M_1, M_0 are sample sizes for the treatment and the control. Denote $Y_k(1)$ as the potential outcome for sample k being under treatment, and $Y_k(0)$ for the case under control.

We are primarily interested in the *average treatment effect* (ATE) τ for the target population $\mathcal{P}$, which is defined as the expected difference in potential outcomes: $\tau \coloneqq \mathbb{E}[Y_k(1) - Y_k(0)]$. With experimental data from RCT under standard assumptions (Appendix B), the sample estimate of ATE is the difference in average sample outcome between the treatment and the control: $\hat{\tau} \coloneqq \hat{\mathbb{E}}[Y|W = 1] - \hat{\mathbb{E}}[Y|W = 0]$. If only observational data are available, with additional identifiability assumptions (Appendix B), we can perform observational causal inference using *potential outcome regression* (POR) (Imbens & Rubin, 2015): $\hat{\tau} \coloneqq (1/M)\sum_k[\hat{Y}_k(1) - \hat{Y}_k(0)]$ where $\hat{Y}_k(1)$ and $\hat{Y}_k(0)$ are estimated potential outcomes. Other alternative approaches include inverse propensity weighting (Rosenbaum & Rubin, 1983) and augmented inverse propensity weighting (Robins et al., 1994). Furthermore, the standard error $\hat{\sigma}$ of the sample estimate is important to quantify statistical significance and confidence interval. Analytical expressions (Appendix C) or bootstrapping can be used to obtain $\hat{\sigma}$ depending on the estimators.

3. Collaborative Causal Inference

We consider n self-interested parties $N \coloneqq \{1, \dots, n\}$. We assume that the parties are non-malicious and they may acquire data from local and potentially biased populations, but these data collectively form the common target population of interest. This assumption can also be interpreted as: individual datasets may have limited representativeness, but when combined, they provide a more comprehensive view of the target population. A subset $C \subseteq N$ is a coalition formed by several parties, and N is often referred to as the grand coalition. For all possible coalitions $C \subseteq N$, let $D_C \coloneqq \{X^{(C)}, Y^{(C)}, W^{(C)}\}$ denote the dataset where $X^{(C)}$, $Y^{(C)}$, and $W^{(C)}$ are covariates, the outcome, and the treatment variable in the data of coalition C, respectively. Let $\hat{\tau}_C$ and $\hat{\sigma}_C$ denote the sample estimate of ATE and the standard error of the estimate obtained from D_C respectively. To simplify notation, the sub/superscript C is replaced by party index i when $C = \{i\}$. Thus, $D_C = \bigcup_{i \in C} D_i$. Let $D = \bigcup_{i \in N} D_i$. Let $v : 2^N \to \mathbb{R}$ be the valuation function for coalitions and let $v_C = v(C)$ be the value of the dataset D_C obtained from all parties in C. The reward value for each party $i \in N$ is denoted as r_i. Thereafter, a reward R_i with value r_i will be realized for each party i. Specifically, $R_i \coloneqq \{\tau_{r,i}, \sigma_{r,i}\}$ consists of both the treatment effect estimate and its standard error. The goal is to design valuation function v, determine the reward value r, and produce the realization of reward R to encourage the collaboration of parties for causal inference. We assume a trusted coordinator who can access the data from all parties and produce the reward according to the framework.



(a) Reverse KL contours (b) Reverse vs. forward KL

Figure 1: Illustration on reverse KL divergence ($\Delta_\tau \coloneqq \hat{\tau}_C - \hat{\tau}_N, \Delta_\sigma \coloneqq \hat{\sigma}_C - \hat{\sigma}_N,\ \hat{\sigma}_N = 1$) .

4. Data Valuation for Causal Inference

Parties are often interested in obtaining both an accurate sample ATE estimate $\hat{\tau}$ and its standard error $\hat{\sigma}$. Hence, the data valuation function for the causal inference dataset should take both $\hat{\tau}$ and $\hat{\sigma}$ produced by the dataset into consideration. In practice, the ground truth treatment effect is an unknown quantity and needs to be inferred. If we assume that all contributed data are non-malicious and come from the same general population, the best available estimate is the sample ATE computed using all data in the grand coalition N due to the asymptotic consistency of the estimators. Thus, we treat the grand coalition estimate $\hat{\tau}_N$ as the surrogate for the ground truth population ATE τ, similarly for its standard error $\hat{\sigma}_N$. We justify the validity of the surrogate in Appendix I.6.

For each dataset, $\hat{\tau}$ itself is a point estimate, but in the form of sample mean. With sufficient sample size m (≥ 30), $\hat{\tau}$ approximately follows the normal distribution $p \sim \mathcal{N}(\tau, \sigma^2)$ by the central limit theorem. Since both τ and σ are unknown true population-level statistics, we approximate p using sample estimates through $q \coloneqq \mathcal{N}(\hat{\tau}, \hat{\sigma}^2)$. This distribution can be useful for scenarios when different levels of confidence interval are required.

The parties value the accuracy of both $\hat{\tau}$ and $\hat{\sigma}$. The sample ATE $\hat{\tau}$ is the fundamental goal of causal inference. The sample standard error $\hat{\sigma}$ serves the purpose of quantifying statistical significance for $\hat{\tau}$. A high $\hat{\sigma}$ indicates low confidence of $\hat{\tau}$ and that the result is less reliable. Consequently, the valuation function v should assign higher values to the more accurate ATE and standard error estimates.

We adopt the *negative reverse Kullback-Leibler* (KL) divergence between the normal distribution $q_C \coloneqq \mathcal{N}(\hat{\tau}_C, \hat{\sigma}_C^2)$ obtained by dataset D_C of coalition C vs. the distribution $p_N \coloneqq \mathcal{N}(\hat{\tau}_N, \hat{\sigma}_N^2)$ by D_N of the grand coalition N as the valuation function for coalition C:

$$v(C) := -\mathrm{KL}(q_C || p_N) = \log \hat{\sigma}_C - \log \hat{\sigma}_N - \frac{\hat{\sigma}_C^2 + (\hat{\tau}_C - \hat{\tau}_N)^2}{2\hat{\sigma}_N^2} + \frac{1}{2} \; . \quad (1)$$

The smaller the distributional divergence from the grand coalition estimate $(\hat{\tau}_N, \hat{\sigma}_N)$, the more valuable the dataset of a coalition. The reverse KL considers the accuracy of both the ATE estimate and its standard error. Moreover, reverse KL is convex as shown in Fig. 1a. When $\hat{\sigma}_C$ is fixed, the reverse KL is minimized when $\hat{\tau}_C = \hat{\tau}_N$. Similarly for a fixed $\hat{\tau}_C$, the minimum is achieved when $\hat{\sigma}_C = \hat{\sigma}_N$. This allows us to get a closed-form upper bound whenever one is known, which is particularly useful when we sample rewards in Sec. 5.3. Reverse KL also has a nice property that describes the asymmetry between overconfidence and underconfidence. Suppose that $\hat{\tau}_C = \hat{\tau}_N$: When $\hat{\sigma}_C > \hat{\sigma}_N$ (underconfident), the reverse KL grows sub-quadratically w.r.t. $\hat{\sigma}_C$. When $\hat{\sigma}_C < \hat{\sigma}_N$ (overconfident), it grows sub-linearly w.r.t. $\hat{\sigma}_C^{-1}$. This asymmetric behavior is desirable because practically it is the underconfidence that prevents parties from using the estimates. Moreover, the value of reverse KL is more stable as in Fig. 1b. We highlight that the direction of KL divergence matters in our case. The forward $\mathrm{KL}(p_N||q_C)$, in contrast, behaves undesirably due to its zero-avoiding property (Bishop, 2006) which punishes insufficient coverage of q_C on the target distribution p_N. This yields overwhelmingly large divergence when q_C is overconfident at super-quadratic rate w.r.t. $\hat{\sigma}_C^{-1}$, as shown in Fig. 1b. The divergence also goes close to 0 with larger $\hat{\sigma}_C$, which is undesirable because overly large $\hat{\sigma}_C$ indicates low confidence and should be thus less valuable.

5. Reward Scheme

This section formally discusses the reward scheme with desirable incentive criteria to ensure a fair and more valuable outcome for each party and encourage their participation. To be compatible with the data valuation function, we use a scalar reward value as a quantitative proxy for the actual reward before its distribution to each party and design the scheme to satisfy the desirable incentive criteria. We first discuss the incentive criteria considering the distinct statistical properties of the causal estimates.

5.1. CCI Incentive Criteria

Let $v_\emptyset := \min_{i \in N} v_i$ for the empty coalition $C = \emptyset$.

(R1) CCI Lower Bound. The reward value is lower bounded by the worst standalone estimate of a party: $\forall i \in N \;\; r_i \geq v_\emptyset$.

(R2) CCI Feasibility. No estimate can be more valuable than that derived from the grand coalition N with 0 divergence: $\forall i \in N \;\; r_i \leq 0$.

(R3) CCI Efficiency. At least one party should be rewarded an estimate with the best achievable quality, i.e., the grand coalition estimate $(\hat{\tau}_N, \hat{\sigma}_N)$: $\exists i \in N \;\; r_i = 0$.

R1 and R2 together constitute the numerical validity of the reward values. We choose to lower bound the negative divergence with $v_\emptyset$ because otherwise, the reward can be arbitrarily bad, and the reward value computation (Sec. 5.2) requires a minimum value. Note that $v_\emptyset$ is only a lower bound for all reward values and it does *not* imply party i with the lowest standalone value ($v_i = \min_{j \in N} v_j$) will receive the lowest reward value. The feasibility in R2 is naturally satisfied due to the non-positivity of negative KL divergence. R3 ensures the efficiency of reward distribution and avoids a wastage of resources, which is also necessary for optimal group welfare (R6) to be defined later.

The work of Sim et al. (2020) has adopted an axiomatic approach for the reward scheme based on *cooperative game theory* (CGT) and proposed several desirable incentive criteria according to the characteristics of ML models. We adapt some of the criteria to suit causal inference:

(R4) Individual Rationality. Higher valued reward is guaranteed: $\forall i \in N \;\; r_i \geq v_i$.

(R5) Fairness. Fairness consists of four components:

- **(F1) Uselessness.** The party i should receive a valueless reward if its data does not improve the treatment effect estimation of any other coalitions.
- **(F2) Symmetry.** If two parties yield the same improvement for all other coalitions, then they should receive equally valuable estimates as rewards.
- **(F3) Strict Desirability.** If the data from party i strictly improves the estimate for at least one coalition more than that of party j, but the reverse is not true, then party i should receive a more valuable reward than party j.
- **(F4) Monotonicity.** For a party i, if its dataset D_i strictly improves the estimate for at least one coalition more compared to that of its other dataset D'_i, but the reverse is not true, then sharing D_i should give party i more valuable reward than sharing D'_i.

(R6) Group Welfare. The group welfare $U := \sum_i r_i$ should be maximized as much as possible.

R4 is fundamental to the cooperative framework because it can encourage participation by ensuring the parties can obtain more valuable estimates than without participation. R5 is an important criterion to ensure the parties are "fairly" rewarded in CCI. In particular, F1 defines the notion of useless datasets, F2 defines the equality in reward for identically contributing datasets, F3 requires the reward values to be proportional to the contribution of the datasets, and F4 encourages parties to share more valuable information by guaranteeing more valuable estimates in return.

5.2. Modified Shapley Fair Reward Value

Previously in Sec. 4, we have only discussed how to compute the standalone value of a dataset. The in-collaboration value of data may vary considerably depending on the composition of the parties. Some data may be more valuable when they are unique in the coalition compared to the scenario when other parties already have similar data. Thus, building on the standalone data valuation function v, we consider the marginal contribution $m_i(T)$ of a dataset i to a coalition T, which is formally defined as

$$m_i(T) \coloneqq v(T \cup \{i\}) - v(T)\ . \tag{2}$$

Recall that v is the data valuation function (a.k.a. characteristic function in CGT). Moreover, in R5, F1-F4 also emphasize the importance of marginal contribution to other parties when comparing party i against another party j. This motivates the design of the in-collaboration reward value function to be based on marginal contributions because only using the standalone value potentially ignores the interaction among parties in the collaboration and can violate other desirable properties in R5. For example, when using POR for observational causal inference, a dataset may produce an ATE estimate that is far away from the ground truth because it lacks samples in the treatment group, but the whole dataset is representative to the target population. Then, including this dataset can reduce the absolute error significantly for the ATE estimates of other coalitions, making the dataset extremely valuable to the coalition even though it does not provide a very accurate ATE estimate on its own. Similarly, a dataset with extremely accurate ATE may not be able to reduce the error for other datasets if this dataset has noisy measurements. Therefore, we use the Shapley value (Shapley, 1953) to carefully account for such interaction among the parties in the collaboration to determine the reward value and to satisfy R5.

Definition 1 (Shapley Value (SV) (Shapley, 1953))**.** Given the marginal contribution function $m_i(\cdot)$ in (2), the SV for a dataset i in grand coalition N is defined as

$$\phi_i \coloneqq (1/n!) \textstyle\sum_{T\subseteq N\setminus\{i\}} |T|!\ (n - |T| - 1)!\ m_i(T)\ . \tag{3}$$

The Shapley value ϕ_i is the expected marginal contribution from i to the coalitions $T \subseteq N\backslash\{i\}$ and satisfies the desirable criteria for our collaborative context, especially fairness (R5).

Proposition 1 (Shapley Fairness)**.** *If $r_i = \alpha\phi_i$ for all $i \in N, \alpha > 0$, then fairness (R5) is satisfied.*

This is modified from Definitions 1 in (Sim et al., 2020). Shapley value also satisfies hard efficiency (Chalkiadakis et al., 2011) in cooperative games constrained by $\sum_i r_i = v_N$. The equality constraint is valid when dividing a finite pool of goods such as monetary compensation and conference votes, but it is not necessary under the CCI context since rewarding a party with the estimate does not "consume" v_N, and hence we have more flexibility. Directly using the Shapley value unnecessarily reduces the group welfare (R6). Therefore, the more relaxed weak efficiency (R3) criterion is followed. To improve group welfare while satisfying other desirable criteria, we propose to determine the reward value of parties in CCI based on a modified version of Shapley value:

Definition 2 (Modified ρ-Shapley fair reward value)**.**

$$r_i \coloneqq \max\{v_i - v_\emptyset, (-v_\emptyset)(\phi_i/\phi^*)^\rho\} + v_\emptyset \tag{4}$$

where $\rho \in (0, 1]$ is the crucial scaling factor and $\phi^* \coloneqq \max_{i\in N} \phi_i$ is the maximum Shapley value of a party in the coalition. The offset by $v_\emptyset$ is necessary to ensure positivity within the max operator because $v_i \leq 0$. This reward is similar in concept to those in the setting of collaborative ML (Sim et al., 2020; Tay et al., 2022).

If $\rho \to 0$, then $r_i \to 0$ for all $i \in N$ and all parties receive the same optimal estimate. This solution achieves maximum group welfare and satisfies individual rationality (R4), but it violates the fairness defined by F3 in R5, discouraging self-interested parties with more valuable data to participate because they can lose their advantages compared to parties with even *useless* or *worthless* data. Supposing $\rho = 1$ and only the right-side value $(-v_\emptyset)(\phi_i/\phi^*)^\rho$ is used for the max-operator, we recover a Shapley value ϕ_i linearly scaled by a factor of $-v_\emptyset/\phi^*$. This satisfies R5 but potentially violates R4 for parties with low Shapley value but high standalone value, i.e., $v_i - v_\emptyset > (-v_\emptyset)(\phi_i/\phi^*)^\rho$. Moreover, supposing $\phi_i > 0$ for all $i \in N$, $\phi_i/\phi^* \in [0, 1]$. Thus, ρ is inversely proportional to r_i. By increasing ρ from 0 to 1, the gap of reward value between parties with different contributions is gradually enlarged for better fairness,[1] but at the cost of reduced group welfare. We show the trade-off caused by ρ in Appendix I.2. When $\rho > 1$, it starts to punish parties with non-maximal Shapley value ($\{i : \phi_i < \phi^*\}$) while *not* satisfying any of the criteria R1 to R6. Thus, we constrain $\rho \leq 1$. By introducing the max over $v_i - v_\emptyset$, we are guaranteed to satisfy R4 but may violate F3 in R5. Under mild assumptions, we can adjust ρ according to the datasets to satisfy all incentive criteria:

Proposition 2 (Main result)**.** *Suppose that the data valuation function v is monotonic at the coalition level, i.e., $v(T \cup \{i\}) - v(T) \geq 0$ for all $i \in N, T \subseteq N$. Then, the Shapley value ϕ_i is non-negative. Furthermore, the modified ρ-Shapley fair reward scheme (Definition 2) satisfies R1 to R4. It also satisfies CCI Fairness (R5) if $\rho \leq \min_{i\in N} \log(1 - v_i/v_\emptyset)/\log(\phi_i/\phi^*)$.*

The proof is in Appendix F. However, we should take note

[1] Informally, "better fairness" means that the rewards to parties with higher contributions receive are (considerably) higher than those to parties with lower contributions.

that the monotonic assumption on valuation function v made by Proposition 2 might not hold (for less $1 < \%$ of the cases, shown in Appendix I.5 on realistic datasets when some partitions consist of mostly non-representative data points and produce too inaccurate estimates). The violation of monotonicity can cause the marginal contribution to another party to be negative and potentially results in negative Shapley value ϕ, which triggers the $\max$-operator to reach the value $v_i - v_\emptyset$ since it is non-negative. As a result, all parties with non-positive Shapley values all have the reward $r_i = v_i$ regardless of the ranking of their Shapley values, which causes a violation of Shapley fairness. Subsequently, R5 is no longer guaranteed. Fortunately, in our empirical investigation, we find this situation is rare, and usually only the party with the lowest standalone value sometimes has a negative Shapley value, which does not violate F3 because of the consistency in ranking between the reward value and the Shapley value. We regard the monotonicity assumption as a theoretical limitation. To overcome this in implementation, we threshold negative Shapley values to 0 and reward the corresponding party minimally with v_i.

The group welfare $U := \sum_i r_i$ is inversely proportional to $\rho \in (0, 1]$ and is maximized as $\rho \to 0$ while satisfying the other criteria. However, the reward gap between parties will also become negligible, which decreases fairness (R5) and causes parties with large and valuable datasets unwilling to participate. On the other hand, when $\rho \to \min_{i \in N} \log(1 - v_i/v_\emptyset)/\log(\phi_i/\phi^*)$, all CGT criteria defined in R1-R5 are satisfied while also encouraging the participation of parties with more valuable data, at the cost of lower group welfare, when compared to a smaller $\rho \to 0$. It is up to the coordinator and the participating parties to set the value according to the particular problem since different use cases require different levels of precision and normalization on the ATE estimation. For example, increasing the recovery rate by 0.1 is not comparable to increasing the farm yield by 0.1kg. Practically, parties can decide a minimally acceptable threshold ρ' according to the problem. Thereafter, we can let the final $\rho := \min\{\rho', \min_{i \in N : \phi_i > 0} \log(1 - v_i/v_\emptyset)/\log(\phi_i/\phi^*)\}$. By introducing ρ', group welfare can be further maximized if ρ can be further decreased after satisfying the inequality in Proposition 2. This parameter allows a more explicit trade-off between fairness and group welfare. As an exception to the rule, parties with negative Shapley values are ignored when determining ρ because it is impossible to always satisfy R5 for them, and their reward values are set to the standalone value v_i.

5.3. Reward Realization

To translate the reward value r into an actual reward $R = \{\tau_r, \sigma_r\}$ as a meaningful incentive in practice, we propose additional criteria when rewarding ATE estimates other than it being consistent with R1 to R6 defined for r previously.

Definition 3 (**(R7) CCI Fidelity**). The rewarded ATE estimate should not provide wrong information about the basic question on whether the treatment is effective. The signs of the estimate τ_r and the grand coalition ATE estimate $\hat{\tau}_N$ must agree: $\text{sign}(\hat{\tau}_N) = \text{sign}(\tau_r)$.

By this criterion, we can perturb the grand-coalition ATE estimate and its standard error $(\hat{\tau}_N, \hat{\sigma}_N)$ according to the value of the dataset of a party as the realization of reward, but up to a degree that the fidelity (sign) is preserved. Otherwise, unfortunate consequences such as clinics mistakenly prescribing ineffective drugs can occur.

Definition 4 (**(R8) CCI Information Obscurity**). The reward R should be sufficiently obscured such that no party i can infer more valuable ATE estimates from R_i and their own data D_i.

To promote collaboration, our proposed CCI framework relies on fairness, and the reward scheme must be transparent to all parties. However, knowing the scheme should not enable any party i to infer better estimates closer to the true ATE estimate $\hat{\tau}_N$ using the dataset D_i and the reward R_i, which overturns the consistency between reward value and the actual reward, defeating the purpose of fair reward scheme. At first, any deterministic reward realization strategy will more likely disclose the grand coalition estimate $\hat{\tau}_N$. For example, if the strategy always moves the estimate of every party closer to $\hat{\tau}_N$ without changing its side (i.e., $\tau_r \leq \hat{\tau}_N$ *only* or $\tau_r \geq \hat{\tau}_N$ *only*), parties who think their data are less valuable can slightly shift $\tau_{r,i}$ further from their own reward $\hat{\tau}_i$ to obtain new ATE estimates closer to $\hat{\tau}_N$ with higher value. Moreover, if we always reward the grand coalition ATE estimate $\hat{\tau}_N$ as τ_r and only perturb the standard error σ_r according to the reward value r, it is almost equivalent to rewarding parties with equally valuable estimates, because parties can just blindly trust the estimate from the grand coalition. Knowing the significance level by σ_r is no longer as useful.

5.3.1. Stochastic Reward Sampling

We propose to inject noise when realizing the reward to obscure the ground truth surrogate. We first randomly sample $\tau_{r,i}$ from a distribution q_i centered on $\hat{\tau}_N$ whose variance is inversely proportional to r_i and bounded by κ^2 whose value is related to $\hat{\sigma}_N$:

$$\tau_{r,i} \sim \mathcal{N}\left(\hat{\tau}_N, \kappa^2(r^* - r_i)/(r^* - v_\emptyset)\right)$$

where $r^* := \max_i r_i$ and κ is set to $2\hat{\sigma}_N$. We use rejection sampling whenever fidelity (R7) is violated. Then we solve for $\sigma_{r,i}$ by equating the value of the reward R_i computed by negative reverse KL divergence defined in (1) to

r_i. Formally,

$$r_i = \log \sigma_{r,i} - \log \hat{\sigma}_N - \frac{\sigma_{r,i}^2 + (\tau_{r,i} - \hat{\tau}_N)^2}{2\hat{\sigma}_N^2} + \frac{1}{2} \,. \quad (5)$$

The only unknown in the equation is $\sigma_{r,i}$, with respect to which the expression on the *right-hand side* (RHS) is concave. The RHS is upper bounded by the Euclidean distance $-(\tau_{r,i} - \hat{\tau}_N)^2/(2\hat{\sigma}_N^2)$ and drawing overly perturbed samples may cause (5) to have no feasible solution. Therefore, a second rejection sampling is used to make sure $\tau_{r,i} \in [\hat{\tau}_N - \hat{\sigma}_N\sqrt{-2r_i}, \hat{\tau}_N + \hat{\sigma}_N\sqrt{-2r_i}]$ for feasibility (Appendix G.1). Provided with $\tau_{r,i}$, we can efficiently solve (5) using root-finding algorithms, prioritizing the solution with a larger variance (underconfidence). In this way, we have guaranteed correspondence between the reward value and the value of the rewarded estimate. For a more contributing party with a higher reward value, it benefits not only from a more valuable reward estimate but also from a lower degree of perturbation when sampling $\tau_{r,i}$.

Using random sampling, our reward realization strategy addresses both R7 and R8 by preserving the fidelity and obscuring $\hat{\tau}_N, \hat{\sigma}_N$. Nonetheless, we still observe some limitations, because the absolute error of $\tau_{r,i}$ is inversely proportional to the error of $\sigma_{r,i}$ under fixed reward value. Overly large $\sigma_{r,i}$ may still indicate more accurate $\tau_{r,i}$. Fortunately, that is not absolutely true since parties with low-value data will also get wider confidence intervals even with less accurate $\tau_{r,i}$. By increasing κ during the sampling of $\tau_{r,i}$, we decrease the cumulative probability of getting more accurate $\tau_{r,i}$ for party i to reduce the chance of getting overly large values for $\sigma_{r,i}$ that may disclose $\hat{\tau}_N$.

6. Experiments

We perform simulated CCI on three datasets based on real-life data distribution to demonstrate the fairness and gain in group welfare for our algorithm. Our implementation can be found at `https://github.com/qiaoruiyt/CollabCausalInference`.

6.1. Datasets

TCGA (Weinstein et al., 2013) is a modified large-scale dataset collected from a public cancer genomics program named The Cancer Genome Atlas (TCGA), on the effectiveness of different treatments in curing cancer. Similar to (Schwab et al., 2018), we focus on the effect of binary treatment (either chemotherapy or surgery) on the binary outcome (recovery) with continuously valued RNA gene expressions as covariates. There are 9659 observed patients while 4130 of them are treated. For the demonstration purpose, we randomly choose 50 covariates out of the 20531-dimensional RNA features. In particular, TCGA dataset has the closest resemblance to healthcare datasets, which are rarely publicly available due to proprietorship. We use TCGA to simulate a real-life instance of CCI where private hospitals collaboratively improve their cancer treatment effect estimates by sharing the data.

JOBS (Lalonde, 1984) consists of experimental samples originating from National Supported Work Demonstration (NSW), a US-based job training program to help disadvantaged individuals. The dataset has 8 covariates such as education, demography, and previous earnings. The outcome is a continuous variable about earnings 2 years after the training. We follow the split by (Louizos et al., 2017; Shalit et al., 2017) with 297 samples in treatment and 425 in control. Performing CCI on JOBS can be interpreted as few employment training organizations are interested in the treatment effect of a common strategy used in their program.

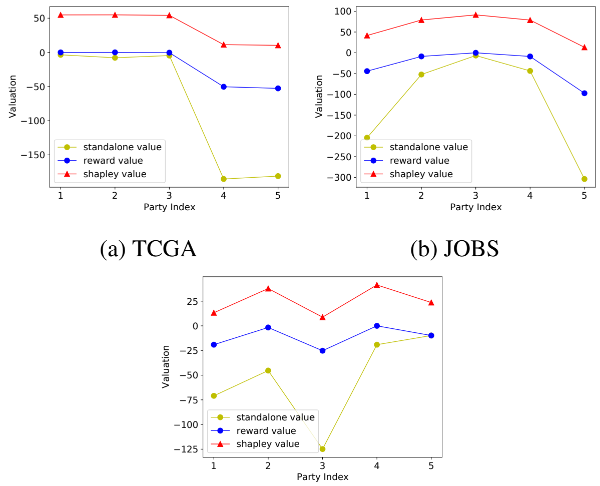
IHDP (Hill, 2011) is a simulated dataset based on a real randomized experiment named Infant Health and Development Program (IHDP), which aims to evaluate the treatment effect of high-quality child care provided by specialists on premature infants. There are 25 covariates (e.g., family condition) and 1 continuous outcome on the cognitive test scores. The original experimental dataset is converted to observational samples by leaving out a nonrandom portion of the group to create bias, resulting in new treatment group (139 samples) and control group (608 samples). We use IHDP to simulate CCI among childcare companies to improve their product through better treatment effect estimation.

6.2. Setups and Results

6.2.1. Simulated CCI

For each of the three causal inference datasets, we randomly create 5 disjoint equal-sized partitions indexed by $j = 1, \ldots, 5$ to simulate an instance of CCI with 5 parties. Causal inference can be performed on each partition or a coalition of partitions for data valuation. We perform all experiments using POR with linear models for simplicity. Our framework is also applicable to other estimators and models (Appendices I.3 and I.4).

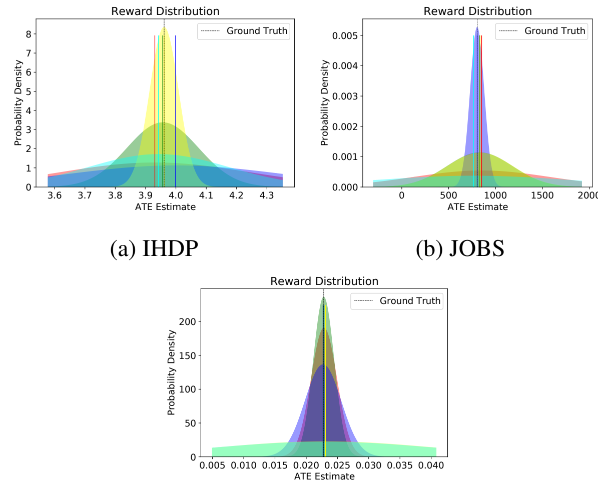
We demonstrate the intermediate and final results of running the simulated collaborative causal inference framework on the three datasets partitioned in a disjoint manner in Fig. 2. The reward value always upper bounds the standalone value, satisfying individual rationality (R4) and demonstrating that participating in the collaboration improves the existing estimation for any party. Even though the three plotted values are correlated, still notably, the ranking of reward value is determined by Shapley value instead of standalone value, because the contribution on improving the estimates of other parties is more important to the collaboration. For example, party 4 is rewarded more than party 5 in Fig. 2c despite having a lower standalone valuation (also party 2 vs. party



(a) TCGA (b) JOBS

(c) IHDP

Figure 2: Simulating CCI framework on three datasets with disjoint partitioning.



(a) IHDP (b) JOBS

(c) TCGA

Figure 3: Reward estimates for simulated CCI on three datasets with disjoint partitioning. The dashed black line is the ground truth (surrogate). Other colored lines represent the sampled reward ATE estimates for the five parties. The shaded area represents the normal distribution parameterized by the ATE and its standard error.

1 and 3 in Fig. 2a), indicating that the proposed modified ρ-Shapley fair reward provides better fairness. Moreover, TCGA with 9659 data points is at a much larger scale compared to IHDP and TCGA (both are less than 1000). Our experiments demonstrate that CCI is beneficial to parties regardless of the sample size. Parties $4, 5$ in Fig. 2a, parties $1, 5$ in Fig. 2b, and parties $1, 2, 3$ in Fig. 2c receive considerable improvements due to sufficient contribution to the collaboration with non-negative Shapley values.

We also visualize the actual sampled reward of the three simulated CCI experiments in Fig. 3, where most of the parties will receive estimates that are quite close to the ground truth, but they may appear on either side of the ground truth with certain distances according to our stochastic reward sampling strategy.

6.2.2. Average Improvement

Table 1 shows that our CCI framework achieves significant improvement over the non-collaborative case by column "Gain", which is defined as the average difference between the value of the estimate obtained by participating in the collaboration vs. the standalone value for the parties, i.e., $(1/n)\sum_i(r_i - v_i)$. We observe that the minimum gain is positive, showing that our scheme is always beneficial. Moreover, we show that the fairness guaranteed by our framework only requires a limited trade-off in group welfare (R6) compared to the naive case where all parties share the same best estimate in column "Cost", which is defined as the difference in value between the grand coalition estimate v_N vs. the reward estimate, i.e., $(1/n)\sum_i(r^* - r_i)$.

Dataset	Gain (min)	Cost (min)
TCGA	73.3 (2.6)	14.4 (0.2)
JOBS	81.5 (4.5)	13.0 (0.4)
IHDP	164.9 (4.5)	52.6 (0.9)

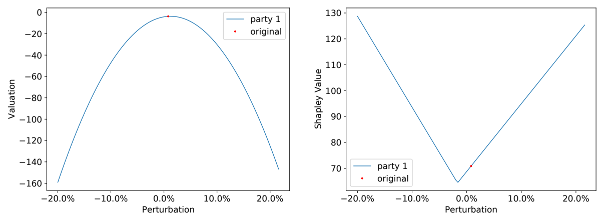
Table 1: Average improvement/cost (and their min) in group welfare. The results are obtained from 1000 independent runs of 5 parties.

7. Sensitivity Analysis

We perform local sensitivity analysis for the data valuation function with respect to the estimate of the ground truth surrogate. For illustration purposes, we slightly abuse the notation on the data valuation function $v(C)$ by including the ground truth surrogate $\hat{\tau}_N$ as part of the functional input since now it is being varied. Then we compute the partial derivative for local sensitivity analysis:

$$\frac{\partial v(C, \hat{\tau}_N)}{\partial \hat{\tau}_N} = \frac{\hat{\tau}_C - \hat{\tau}_N}{\hat{\sigma}_N^2} = \frac{\Delta_C}{\hat{\sigma}_N^2}$$

Suppose that the standard error $\hat{\sigma}_N$ is fixed, the sensitivity of the data valuation function is locally determined by the accuracy of the estimate Δ_C produced by the coalition C. The more inaccurate the estimate is, the more sensitive it is to the changes in the ground truth surrogate.



(a) IHDP Standalone Value (b) IHDP Shapley Value

Figure 4: Sensitivity analysis by perturbing the value of the ground truth surrogate. Without loss of generality, we only plot the change in valuation for the first party.

For the Shapley value of party i, the partial derivative is:

$$\frac{\partial \phi_i}{\partial \hat{\tau}_N} = \frac{1}{n!} \sum_{T \subseteq N \setminus \{i\}} |T|!\,(n - |T| - 1)!\, \frac{\hat{\tau}_{T\cup\{i\}} - \hat{\tau}_T}{\hat{\sigma}_N^2} \,.$$

This is in fact a constant with respect to $\hat{\tau}_N$. Thus, any change of the ground truth surrogate $\hat{\tau}_N$ will have a linear effect on the Shapley value for each coalition (including each party). With respect to the coalitions, the sensitivity gets larger for a party's data if including it causes a larger magnitude for the weighted net change in the ATE estimate.

We empirically study the sensitivity of the quantities proposed in our work with respect to the value of the ground truth surrogate. We perturb $\hat{\tau}_N$ uniformly in the range $[-20\%, +20\%]$ and plot the corresponding values in Fig. 4. Empirically, our approach is quite sensitive with respect to the accuracy of the ground truth surrogate, which is intuitive since the ground truth surrogate is important in our method. However, we note that the grand coalition estimate is the best estimate we can obtain in practice, *without* making and exploiting additional assumptions, which can be improved further by including more parties in the collaboration.

8. Related Work

To leverage the effectiveness of big data, various approaches are being proposed to take advantage of the data from multiple sources or parties. For instance, for machine learning, there are approaches that consider federated learning (Kairouz et al., 2021), collaborative machine learning (Sim et al., 2020; Xu et al., 2021a; Nguyen et al., 2022; Lin et al., 2023), unsupervised learning (Tay et al., 2022), parametric learning (Agussurja et al., 2022), (personalized) model fusion (Lam et al., 2021; Hoang et al., 2021), active learning (Xu et al., 2023) or reinforcement learning (Fan et al., 2021). As many of works require assigning scalar values to the datasets of the parties, the so-called *data valuation* functions (Sim et al., 2022) are often leveraged, such as (Ghorbani & Zou, 2019; Xu et al., 2021b; Wu et al., 2022). These data valuation works exploit certain structures in ML (e.g., the accuracy on a validation dataset (Ghorbani & Zou, 2019)). In contrast, our work differs from them in exploiting the unique statistical perspective of causal inference.

Similarly in causal inference, federated causal inference (Vo et al., 2021; Xiong et al., 2021) and causal data fusion (Bareinboim & Pearl, 2016; Li et al., 2020) are also active research areas. However, these works all implicitly assume that all parties are altruistic and willing to contribute their valuable data regardless of the cost-effectiveness. Our work generalizes to the case with self-interested parties and incentivizes the parties to collaborate, thus helping to meet the assumption made in those cited works.

9. Conclusion and Future Work

We propose a novel collaborative causal inference framework that incentivizes the collaboration of self-interested parties for causal inference by fairly rewarding them with more valuable treatment effect estimates. The framework consists of (a) a causal inference data valuation function using the negative reverse KL divergence towards the target estimate, (b) a reward scheme based on ρ-Shapley fair reward value to satisfy desirable incentive criteria, and (c) a stochastic reward realization strategy based on rejection sampling. We empirically demonstrate the effectiveness of the framework.

We aim to encourage practical collaboration in causal inference by addressing the fairness aspect via the Shapley value and it is interesting to explore whether the Shapley fairness can still be satisfied when the number of parties is large (Zhou et al., 2023). Our work focuses on the case where parties share a common population of interest and pursue homogeneous ATE estimates, but there are scenarios where conditional ATE for heterogeneous populations is also important. Such extension requires non-trivial effort on data valuation and incentive mechanism design, which we leave for future work. Another assumption is that the parties are honest and non-malicious, which may not be guaranteed in practice, as some parties can be untruthful and and try to exploit the transparent framework to achieve selfish outcomes or even cause harm to other parties. Strategyproofness (Chalkiadakis et al., 2011) in CGT is an interesting future research direction for encouraging truthfulness.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD/2021-08-017[T]). Xinyi Xu is supported by the Institute for Infocomm Research of Agency for Science, Technology and Research (A*STAR).

References

Adams, J. S. Towards an understanding of inequity. *The Journal of Abnormal and Social Psychology*, 67(5):422–436, 1963.

Agussurja, L., Xu, X., and Low, B. K. H. On the convergence of the Shapley value in parametric Bayesian learning games. In *Proc. ICML*, pp. 180–196, 2022.

Bareinboim, E. and Pearl, J. Controlling selection bias in causal inference. In *Proc. AISTATS*, pp. 100–108, 2012.

Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. *PNAS*, 113(27):7345–7352, 2016.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2006.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd, 2009.

Chalkiadakis, G., Elkind, E., and Wooldridge, M. Computational aspects of cooperative game theory. In Brachman, R. J., Cohen, W. W., and Dietterich, T. G. (eds.), *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool Publishers, 2011.

Delgado-Rodríguez, M. and Sillero-Arenas, M. Systematic review and meta-analysis. *Medicina Intensiva*, 42(7): 444–453, 2018.

Dersimonian, R. and Laird, N. M. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, 1986.

Fan, F. X., Ma, Y., Dai, Z., Jing, W., Tan, C., and Low, B. K. H. Fault-tolerant federated reinforcement learning with theoretical guarantee. In *Proc. NeurIPS*, pp. 1007–1021, 2021.

Ghorbani, A. and Zou, J. Y. Data Shapley: Equitable valuation of data for machine learning. In *Proc. ICML*, 2019.

Glass, T. A., Goodman, S. N., Hernán, M. A., and Samet, J. M. Causal inference in public health. *Annual Review of Public Health*, 34:61–75, 2013.

Guo, W., Wang, S., Ding, P., Wang, Y., and Jordan, M. I. Multi-source causal inference using control variates. arXiv:2103.16689, 2021.

Hernán, M. A., Brumback, B. A., and Robins, J. M. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*, 21(12):1689–1709, 2002.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *J. Computational and Graphical Statistics*, 20 (1):217–240, 2011.

Hoang, T. N., Hong, S., Xiao, C., Low, B. K. H., and Sun, J. AID: Active distillation machine to leverage pre-trained black-box models in private data settings. In *Proc. WWW*, pp. 3569–3581, 2021.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, 2015.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z. B., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. Y. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, O., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D. X., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1-2):1–210, 2021.

Konow, J. Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review*, 90(4):1072–1091, 2000.

Lalonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4):604–620, 1984.

Lam, C. T., Hoang, N., Low, B. K. H., and Jaillet, P. Model fusion for personalized learning. In *Proc. ICML*, pp. 5948–5958, 2021.

Li, H., Miao, W., Cai, Z., Liu, X., Zhang, T., Xue, F., and Geng, Z. Causal data fusion methods using summary-level statistics for a continuous outcome. *Statistics in Medicine*, 39(8):1054–1067, 2020.

Lin, X., Xu, X., Ng, S.-K., Foo, C.-S., and Low, B. K. H. Fair yet asymptotically equal collaborative learning. In *Proc. ICML*, 2023.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D. A., Zemel, R. S., and Welling, M. Causal effect inference with deep latent-variable models. In *Proc. NeurIPS*, pp. 6449–6459, 2017.

Masic, I., Miokovic, M., and Muhamedagić, B. Evidence based medicine – new approaches and challenges. *Acta Informatica Medica*, 16:219–225, 2008.

Nguyen, Q. P., Low, B. K. H., and Jaillet, P. Trade-off between payoff and model rewards in Shapley-fair collaborative machine learning. In *Proc. NeurIPS*, pp. 30542–30553, 2022.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *JASA*, 89(427):846–866, 1994.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.*, 5(4): 472–480, 1990.

Rubin, D. B. Causal inference using potential outcomes. *JASA*, 100(469):322–331, 2005.

Schwab, P., Linhardt, L., and Karlen, W. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. arXiv:1810.00656, 2018.

Shalit, U., Johansson, F. D., and Sontag, D. A. Estimating individual treatment effect: generalization bounds and algorithms. In *Proc. ICML*, 2017.

Shapley, L. S. A value for n-person games. In Kuhn, H. W. and Tucker, A. W. (eds.), *Contributions to the Theory of Games*, volume 2, pp. 307–317. Princeton Univ. Press, 1953.

Sim, R. H. L., Zhang, Y., Chan, M. C., and Low, B. K. H. Collaborative machine learning with incentive-aware model rewards. In *Proc. ICML*, pp. 8927–8936, 2020.

Sim, R. H. L., Zhang, Y., Low, B. K. H., and Jaillet, P. Collaborative Bayesian optimization with fair regret. In *Proc. ICML*, pp. 9691–9701, 2021.

Sim, R. H. L., Xu, X., and Low, B. K. H. Data valuation in machine learning: "ingredients", strategies, and open challenges. In *Proc. IJCAI*, pp. 5607–5614, 2022.

Smith, A. The wealth of nations. In Cohen, M. (ed.), *Princeton Readings in Political Thought: Essential Texts since Plato - Revised and Expanded Edition*, pp. 298–315. 2018.

Tabibnia, G. and Lieberman, M. D. Fairness and cooperation are rewarding: evidence from social cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1118(1):90–101, 2007.

Tay, S. S., Xu, X., Foo, C. S., and Low, B. K. H. Incentivizing collaboration in machine learning via synthetic data rewards. In *Proc. AAAI*, pp. 9448–9456, 2022.

Thibaut, J. W. *The Social Psychology of Groups*. Routledge, 1960.

Vo, T. V., Hoang, T. N., Lee, Y., and Leong, T.-Y. Federated estimation of causal effects from observational data. arXiv:2106.00456, 2021.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45:1113–1120, 2013.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., and Gallego, B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 37(23):3309–3324, 2018.

Wu, Z., Shu, Y., and Low, B. K. H. Davinz: Data valuation using deep neural networks at initialization. In *Proc. ICML*, pp. 24150–24176, 2022.

Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., and Athey, S. Federated causal inference in heterogeneous observational data. arXiv:2107.11732, 2021.

Xu, X., Lyu, L., Ma, X., Miao, C., Foo, C.-S., and Low, B. K. H. Gradient driven rewards to guarantee fairness in collaborative machine learning. In *Proc. NeurIPS*, pp. 16104–16117, 2021a.

Xu, X., Wu, Z., Foo, C. S., and Low, B. K. H. Validation free and replication robust volume-based data valuation. In *Proc. NeurIPS*, pp. 10837–10848, 2021b.

Xu, X., Wu, Z., Verma, A., Foo, C. S., and Low, B. K. H. FAIR: Fair collaborative active learning with individual rationality for scientific discovery. In *Proc. AISTATS*, pp. 4033–4057, 2023.

Yang, S. and Ding, P. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115(531):1540–1554, 2020.

Zhou, Z., Xu, X., Sim, R. H. L., Foo, C. S., and Low, B. K. H. Probably approximate Shapley fairness with applications in machine learning. In *Proc. AAAI*, 2023.

A. Ethics statement

We would like to highlight that our goal is to encourage collaboration and benefit society rather than limiting knowledge discovery. In the ideal world, it is good to share complete causal knowledge among every party (e.g., scientific research). However, our work focuses on the realistic scenario where self-interested parties who care about fairness are common across industries (e.g., private hospitals, pharmaceutical firms, agricultural farms), but the collaboration is missing/rare prior to incentivization. Without fairness, these parties are not willing to collaborate, thus further limiting the discovery of knowledge and their welfare. Our fairness-based framework removes one important roadblock to collaboration. Comparing the ideal case of equally sharing the causal knowledge vs. the practical case of proportional but fair sharing, which view is correct? Our opinion is that they both have their own use cases. In our case, proportional knowledge sharing is actually more ethical than complete knowledge sharing.

B. Assumptions

B.1. Causal Inference

We make the following assumptions for the identifiability of ATE under Neyman-Rubin potential outcome framework.

1. **Stable Unit Treatment Value Assumption (SUTVA)** (Imbens & Rubin, 2015): The treatment for one unit does not change the effect of treatment for other units, i.e., $\forall j, k \in \mathcal{P}$ such that $j \neq k$, $Y_j \perp\!\!\!\perp W_k$.

2. **Consistency**: The potential outcome agrees with the observed outcome in the dataset, i.e., $\forall j \in \mathcal{P}, Y_j = Y_j(0)(1 - W_j) + Y_j(1)W_j$.

3. **Unconfoundedness** (Rosenbaum & Rubin, 1983): The potential outcomes are independent of the treatment given the covariates, i.e., $\forall j \in \mathcal{P}, (Y_j(0), Y_j(1)) \perp\!\!\!\perp W_j | X_j$.

4. **IID**: All units $j \in \mathcal{P}$ are independently and identically distributed (IID) samples from the general population of interest.

B.2. Collaborative Causal Inference

In addition, for the theoretical properties of the collaborative scheme, we assume that the parties are self-interested but non-malicious. Self-interestedness means that parties are not altruistically sharing their data. Being non-malicious is a different concept such that parties are not performing harmful actions (e.g., deliberately providing wrongly labeled data) to degrade the estimates of other parties.

We argue that the assumption of being self-interested and non-malicious is valid in practice. Firstly, in our motivating example, hospitals are self-interested entities that are mostly self-funded and profit-seeking, but they still share the principle of helping the community, and any malicious act that causes inaccurate ATE estimate is not aligned with their objective. Thus, the non-malicious assumption is valid. Secondly, a similar assumption is either explicitly or implicitly adopted in a variety of works in multi-source causal inference (Bareinboim & Pearl, 2016; 2012; Yang & Ding, 2020) and collaborative machine learning (Sim et al., 2020; Tay et al., 2022). To our knowledge, very little work in the causal inference setting explicitly discusses malicious data sources whilst not knowing the true causal effect. Furthermore, to rigorously relax this assumption would require a suitable and well-motivated definition of malicious parties which presents a challenging future research direction.

C. Standard Error Estimation

For RCT, $\hat{\sigma} = (\hat{\sigma}_{Y|W=1}/M_1 + \hat{\sigma}_{Y|W=0}/M_0)^{1/2}$, where $\hat{\sigma}_{Y|W}$ is the standard error of the mean for the treatment ($W = 1$) or the control ($W = 0$).

For POR, $\hat{\sigma} = ((\hat{\sigma}_{Y(1)} + \hat{\sigma}_{Y(0)})/M)^{1/2}$, where $\hat{\sigma}_{Y(1)}$ is the standard error of the mean for the potential outcome of the treatment and $\hat{\sigma}_{Y(0)}$ is for that of the control.

For other estimators which may not have analytical expressions for standard error, bootstrapping can be used.

D. Additional Data Valuation Functions and Why Not Choose Them

D.1. Discrepancy

One natural choice of the data valuation function v is the negative discrepancy between the estimated ATE from the subset S and from the grand coalition N:

$$v_d(S) = -d(\tau_S, \tau_N) \tag{6}$$

where d is an arbitrary metric distance such as squared difference $d = (\tau_S - \tau_N)^2$ or absolute difference $d = |\tau_S - \tau_N|$.

Why not? This measure does not consider the uncertainty - standard error. In practice, the standard error is a required justification to show that the estimate is statistically significant.

D.2. Inverse Variance Weighting

Meta-analysis (Borenstein et al., 2009; Dersimonian & Laird, 1986) is a statistical technique proposed to perform systematic review (Delgado-Rodríguez & Sillero-Arenas, 2018), which aggregates the treatment effect estimates from multiple independent studies. For example, different researchers may have performed randomized control trial to test the treatment effect for different demographics across the globe. A team of systematic reviewers may use meta analysis to draw conclusions for the entire human population or discover hidden heterogeneity that may indicate fundamental inconsistency within the problem and potential future research direction.

The statistical result of meta analysis is usually in the form of a weighted average of the treatment effect from existing studies based on the inverse variance weighting (IVW). If we assume a simple (fixed effect) model, the weights of the study, which can be used as a valuation metric for the dataset S in our case, is computed simply as follows:

$$v_{ivw,f} = W_i = \frac{1}{\sigma_S^2} \; . \tag{7}$$

Why not? Note that IVW does not explicitly account for any divergence from a ground truth estimation. It assumes no data samples are bad as long as it reduces the variance which ties to the statistical significance. This does not account for the reality such that if working independently, parties have to use their own estimation without knowing the “ground truth” and bare the consequence of having a discrepancy. A data valuation function has to consider the discrepancy from the “ground truth”.

D.3. Information Gain

Many causal inference estimators rely on ML techniques. A validation-free information-theoretic approach is proposed in collaborative ML (Sim et al., 2020) for ML models. The greater the reduction in uncertainty of the model parameters θ, the more valuable the data is. A proper measure of uncertainty is information gain (IG) $\mathbb{I}(\theta; D)$ and the corresponding valuation function is defined as:

$$v_e(S) = \mathbb{I}(\theta; D_S) = \mathbb{H}(\theta) - \mathbb{H}(\theta|D_S), \tag{8}$$

where $\mathbb{H}$ is the entropy function. If we only take the ML component of causal inference into account, then by using the Bayesian version of the regressor and classifier (e.g., Gaussian process), the IG for a causal inference dataset D can be efficiently computed analytically in closed form:

$$\mathbb{I}(\theta; D) = \frac{1}{2} \log(\mathbf{I} + K\sigma^{-2}), \tag{9}$$

where σ^2 is the variance for the target of prediction and K is a $|D| \times |D|$ gram matrix defined over kernel function $k(x, x')$ for the covariates. We have $K = X^\top X$ when the kernel function is linear, i.e., $k(x, x') = x^\top x'$.

This valuation function possesses many desirable properties (e.g., monotonicity, submodularity) motivated by cooperative game theory, which are crucial to the proof of the propositions in collaborative ML (Sim et al., 2020). In fact, many ATE estimators contain ML as a sub-problem of causal inference. Specifically, *potential outcome regression* (POR) is based on regression, *inverse propensity weighting* (IPW) is based on binary classification, and *augmented IPW* (AIPW) can be viewed as a combination of regression and classification.

Why not? Unfortunately, the uncertainty of ML model used by the subsystem has a relatively low correlation with the final ATE accuracy in causal inference because it's fundamentally a density estimation problem, which is different from prediction. Moreover, not all ML models have efficiently computable Bayesian counterparts, for example, neural networks.

D.4. Volume

Alternatively to IG, volume and robust volume (Xu et al., 2021b) are proposed as a simpler validation-free data valuation function. The approach is based on the gram matrix similar to that of IG (i.e., $X \top X$), but has the advantage of having model-agnostic property, fewer hyperparameters, and computational efficiency. It has been formally proven that a larger volume corresponds to a lower mean square error (MSE) in predictive performance.

Why not? Volume is only guaranteed to work with low-dimensional datasets. In practice, volume is dependent on the scale of variables and suffers from extremely large values if no normalization is applied. Furthermore, the predictive MSE is not necessarily correlated with the ATE accuracy.

E. Full Axioms for Collaborative Causal Inference

(R1) CCI Lower Bound. The reward value is lower bounded by the worst standalone estimate of a party: $\forall i \in N \;\; r_i \geq v_\emptyset = \min_{i \in N} v_i$.

(R2) CCI Feasibility. No estimate can be more valuable than that derived from the grand coalition N with 0 divergence: $\forall i \in N \;\; r_i \leq 0$.

(R3) CCI Weak Efficiency. At least one party is rewarded an estimate with the best achievable quality, i.e., the grand coalition estimate: $\exists i \in N \;\; r_i = 0$.

(R4) Individual Rationality. Each party i should receive an estimate with value that is at least as good as the standalone estimate produced by itself: $\forall i \in N \;\; r_i \geq v_i$.

(R5) Fairness. CCI Fairness includes the following four components:

- **(F1) Uselessness.** The party i should receive valueless reward if its data does not improve the estimation of any other coalition: $\forall i \in N \;\; (\forall C \subseteq N \backslash \{i\} \;\; v_{C \cup \{i\}} \leq v_C) \Rightarrow r_i = v_\emptyset$.
- **(F2) Symmetry.** If including the data of party i yields the same improvement as that of another party j in the quality of an estimator using the aggregated data of any coalition, then they should receive equally valuable estimator rewards: $\forall i, j \in N$ s.t. $i \neq j \;\; (\forall C \subseteq N \backslash \{i, j\} \;\; v_{C \cup \{i\}} = v_{C \cup \{j\}}) \Rightarrow r_i = r_j$.
- **(F3) Strict Desirability.** If the data from party i improves the estimator for at least one coalition more comparing to that of party j, but the reverse is not true, then party i should receive a more valuable reward than party j: $\forall i, j \in N$ s.t. $i \neq j \;\; (\forall B \subseteq N \backslash \{i, j\} \;\; v_{B \cup \{i\}} > v_{B \cup \{j\}}) \wedge (\forall C \subseteq N \backslash \{i, j\} \;\; v_{C \cup \{i\}} \geq v_{C \cup \{j\}}) \Rightarrow r_i > r_j$.
- **(F4) Monotonicity.** Consider the case where only party i improves its dataset from D_i to D'_i (e.g., having lower noise or better samples) and results in an updated set of values v' in coalition. Let r_i and r'_i denote the reward for party i under the respective situation. If at least one coalition strictly benefits more from D'_i than D_i (with more accurate or confident estimate), *ceteris paribus*, then i should receive more reward than before. $\forall i \in N \;\; (\forall B \subseteq N \backslash \{i\} \;\; v'_{B \cup \{i\}} > v_{B \cup \{i\}}) \wedge (\forall C \subseteq N \backslash \{i\} \;\; v'_{C \cup \{i\}} \geq v_{C \cup \{i\}}) \wedge (\forall A \subseteq N \backslash \{i\} \;\; v'_A = v_A) \wedge (v'_N \geq r_i) \Rightarrow r'_i > r_i$.

(R6) Group Welfare. The group welfare $U := \sum_i r_i$ should be maximized while satisfying R1 to R5.

(R7) CCI Fidelity. The rewarded ATE estimate should not provide wrong information about the basic question on whether the treatment is effective. The following relationship must hold between the gold ATE estimate $\hat{\tau}_N$ and the reward estimate τ_r: $\text{sign}(\hat{\tau}_N) = \text{sign}(\tau_r)$.

By this criterion, we can perturb the grand-coalition ATE estimate and its standard error $(\hat{\tau}_N, \hat{\sigma}_N)$ according to the value of the dataset of a party as the realization of reward, but up to a degree that the fidelity (sign) is preserved. Otherwise, unfortunate consequences such as clinics mistakenly prescribing ineffective drugs can occur.

(R8) CCI Information Obscurity. The reward R should be sufficiently obscured such that no party i can infer more valuable ATE estimates from R_i and their own data D_i.

E.1. General versions of R7 and R8

We define fidelity for the case of non-binary but discrete treatment.

(R7.1) CCI Fidelity for Ranking Preservation. The rewarded ATE estimate should not provide wrong information about the basic question of which treatment is more effective, meaning that the ranking of the treatment effects is preserved when the estimates are distributed to the parties. Let W be the set of treatments and $\tau(w)$ be a function that returns the treatment effect estimate of $w \in W$. The following relationship must hold between the ground truth surrogate ATE estimate $\hat{\tau}_N$ and any reward estimate τ_r:

$$\forall w, w' \in W \text{ s.t. } w \neq w' \;\; \hat{\tau}_N(w) \geq \hat{\tau}_N(w') \Rightarrow \tau_r(w) \geq \tau_r(w') \; .$$

(R7.2) Fidelity. The reward R should meet a minimum performance bar by preserving the most essential information of the inference problem.

(R8.1) Information Obscurity. The reward R should be sufficiently obscured such that no party i can infer more valuable estimates from R_i and their own data D_i.

F. Proof of Propositions

We restate Proposition 2:

Proposition. *Assume that the data valuation function v is monotonic at the dataset level, i.e., adding more data never hurts. The modified ρ-Shapley fair reward scheme described in Definition 2 satisfies R1 to R4. Moreover, it satisfies CCI Fairness (R5) if $\rho \leq \min_{i \in N} \log(1 - v_i/v_\emptyset)/\log(\phi_i/\phi^*)$.*

Proof. The proof resembles the case of CGM (Tay et al., 2022), which is also based on collaborative ML (Sim et al., 2020). Recall the definition of modified ρ-Shapley fair reward:

$$r_i = \max\left\{v_i - v_\emptyset, (-v_\emptyset)\left(\frac{\phi_i}{\phi^*}\right)^\rho\right\} + v_\emptyset$$

(R1) CCI Lower Bound. $\forall i \in N \;\; r_i \geq (v_i - v_\emptyset) + v_\emptyset \geq v_i \geq v_\emptyset$.
(R2) CCI Feasibility. At first, if $r_i = v_i$ for all $i \in N$, $v_i \leq 0$ since $\hat{\tau}_N$ is the ground truth. Otherwise, if $r_i = (1 - (\phi_i/\phi^*)^\rho)v_\emptyset$, then $(\phi_i/\phi^*)^\rho \leq 1$ since $\rho \in (0, 1]$ and $\phi_i \geq 0$ for all $i \in N$. Therefore, r_i equals to $v_\emptyset$ multiplied by a coefficient in $[0, 1]$. As $v_\emptyset \leq 0$, $r_i \leq 0$.
(R3) CCI Weak Efficiency. Since $v_i \leq 0$, for $j = \arg\max_j \phi_j$, $r_j = \max\{v_j - v_\emptyset, (\phi^*/\phi^*)^\rho \times (-v_\emptyset)\} + v_\emptyset = \max\{v_j - v_\emptyset, 0 - v_\emptyset\} + v_\emptyset = 0$.
(R4) CCI Individual Rationality. $r_i \geq v_i - v_\emptyset + v_\emptyset = v_i$.
(R5) Since $\phi_i/\phi^* \in [0, 1]$ and $\log(\phi_i/\phi^*) < 0$, setting ρ to that particular value is equivalent of saying $v_i - v_\emptyset \leq (-v_\emptyset)(\phi_i/\phi^*)^\rho$ for all $i \in N$. Thus, $r_i = (1 - (\phi_i/\phi^*)^\rho)v_\emptyset \geq v_i$, the situation exactly matches the required condition for fairness in Theorem 1 of collaborative ML (Sim et al., 2020). □

G. Derivations

G.1. Solution Bound Derivation for Sec. 5.3.1

Recall the equation to be solved:

$$r_i = -\log \hat{\sigma}_N + \log \sigma_{r,i} - \frac{\sigma_{r,i}^2 + (\tau_{r,i} - \hat{\tau}_N)^2}{2\hat{\sigma}_N^2} + \frac{1}{2} \; . \tag{10}$$

We use v_i' to denote the right-hand side of the equation and it is a concave function with respect to $\sigma_{r,i}$. Its first-order derivative:

$$\frac{\partial v_i'}{\partial \sigma_{r,i}} = \frac{1}{\sigma_{r,i}} - \frac{\sigma_{r,i}}{\hat{\sigma}_N^2} \; . \tag{11}$$

Setting it to 0 gives $\sigma_{r,i} = \hat{\sigma}_N$. Thus, the maximum of v_i' is achieved when $\sigma_{r,i} = \hat{\sigma}_N$ and $v_{i,max}' = -(\tau_{r,i} - \hat{\tau}_N)^2/(2\hat{\sigma}_N^2)$. In order to make sure that valid solution of $\sigma_{r,i}$ exists when sampling $\tau_{r,i}$, we just need to have $r_i \leq \max_{\sigma_{r,i}} v_i'$ and satisfy the following inequality:

$$
\begin{aligned}
& & r_i &\le -\frac{(\tau_{r,i} - \hat{\tau}_N)^2}{2\hat{\sigma}_N^2} \\
&\Rightarrow & 2r_i\hat{\sigma}_N^2 &\le -(\tau_{r,i} - \hat{\tau}_N)^2 \\
&\Rightarrow & |\tau_{r,i} - \hat{\tau}_N| &\le \hat{\sigma}_N\sqrt{-2r_i} \ .
\end{aligned}
$$

Thus, $\tau_{r,i} \in [\hat{\tau}_N - \hat{\sigma}_N\sqrt{-2r_i}, \hat{\tau}_N + \hat{\sigma}_N\sqrt{-2r_i}]$.

G.2. Gradient Derivation for Sensitivity Analysis

We compute the partial derivative for the data valuation function $v(C)$ with respect to the ground truth surrogate $\hat{\tau}_N$:

$$
\begin{aligned}
\frac{\partial v(C, \hat{\tau}_N)}{\partial \hat{\tau}_N} &= \frac{\partial[-\mathrm{KL}(q_C||p_N)]}{\partial \hat{\tau}_N} \\
&= \frac{\partial\left[\log\hat{\sigma}_C - \log\hat{\sigma}_N - [\hat{\sigma}_C^2 + (\hat{\tau}_C - \hat{\tau}_N)^2]/(2\hat{\sigma}_N^2) + 1/2\ \right]}{\partial \hat{\tau}_N} \\
&= \frac{\hat{\tau}_C - \hat{\tau}_N}{\hat{\sigma}_N^2} = \frac{\Delta_C}{\hat{\sigma}_N^2} \ .
\end{aligned}
$$

For the Shapley value of party i, the partial derivative is:

$$
\begin{aligned}
\frac{\partial \phi_i}{\partial \hat{\tau}_N} &= \frac{\partial (1/n!) \sum_{T \subseteq N\setminus\{i\}} |T|!\,(n - |T| - 1)!\, m_i(T)}{\partial \hat{\tau}_N} \\
&= (1/n!) \sum_{T \subseteq N\setminus\{i\}} |T|!\,(n - |T| - 1)!\, \frac{\partial[v(T \cup \{i\}, \hat{\tau}_N) - v(T, \hat{\tau}_N)]}{\partial \hat{\tau}_N} \\
&= (1/n!) \sum_{T \subseteq N\setminus\{i\}} |T|!\,(n - |T| - 1)!\, \frac{\hat{\tau}_{T\cup\{i\}} - \hat{\tau}_N - (\hat{\tau}_T - \hat{\tau}_N)}{\hat{\sigma}_N^2} \\
&= (1/n!) \sum_{T \subseteq N\setminus\{i\}} |T|!\,(n - |T| - 1)!\, \frac{\hat{\tau}_{T\cup\{i\}} - \hat{\tau}_T}{\hat{\sigma}_N^2} \ .
\end{aligned}
\tag{12}
$$

H. Discussion

H.1. Comparison with Existing Collaborative Frameworks

The proposed framework is similar to the seminal work in collaborative ML (Sim et al., 2020) and CGM (Tay et al., 2022). We highlight the distinctive contributions of our work in comparison with the previous two.

First, we consider the problem of causal inference whose essential goal is to produce an accurate estimate with confidence interval. Despite the fact that the treatment effect estimate utilizes machine learning model to facilitate its computation, the predictive performance differs fundamentally from and may not have correlation with the accuracy of the estimate. Thus, we have proposed a novel data valuation function for causal inference datasets based on negative KL divergence, which considers both the accuracy and uncertainty of the estimate with closed-form expression. This specially designed data valuation function is based on treating the estimate as a form of distribution, then value the dataset by the distributional divergence between the estimate obtained by the dataset vs. that of the grand coalition (ground truth). This distinguishes our work from collaborative ML (Sim et al., 2020) and CGM (Tay et al., 2022) which both focuses on machine learning.

Second, we propose modified incentive criteria to incorporate the new data valuation function. Moreover, we propose modified ρ-Shapley fair reward value as the core reward scheme for CCI, such that all desirable incentive criteria of collaboration can be satisfied. In particular, our reward scheme differs from the previous works by considering the unique problem of causal inference and tackling the data valuation function without non-negativity assumption. Moreover, we do not use the stability criterion from R7 in collaborative ML (Sim et al., 2020; Tay et al., 2022). Stability ensures parties cannot strictly benefit more by forming another coalition (e.g., in a 5-party collaboration, 4 parties can strictly gain more by abandoning the other one party). However, deviating from the grand coalition and forming another coalition is a strategy that requires the result from the grand coalition. Since it is very difficult for parties to try and compare different composition of coalitions and choose their partners, we argue that this criterion is not necessary.

Third, we propose additional two criteria for the reward in consideration of the unique properties of the cooperative game in causal inference to guarantee the usefulness and fairness, which has not been investigated in prior collaborative ML (Sim et al., 2020; Tay et al., 2022). In particular, the treatment effect estimate represents a form of knowledge (e.g., whether the treatment is effective) and is usually in the form of a scalar. Without randomness, parties may be able to exploit the reward scheme and their rewarded treatment effect estimates to infer more valuable estimates (as in Sec. 5.3. This is undesirable because these parties can get more valuable rewards than what they deserve according to their contribution to the collaboration. Introducing randomness is one way to prevent exploitation, but the estimate should not be perturbed to the extent that parties have to bear with the excessively wrong knowledge. For instance, the treatment is effective but the perturbed estimate shows ineffective. We design practical and efficient stochastic reward sampling strategies according to the two criteria. These problems were not considered in collaborative ML (Sim et al., 2020; Tay et al., 2022), because ML models are less intuitive and "safer". Therefore, parties cannot easily infer better ML models even when our reward scheme is transparent.

H.2. Extension to Heterogeneous ATE

We hope that our work can initiate a novel research direction that motivates collaboration for causal inference in a fair way. Therefore, we begin our research with a more approachable setting. We acknowledge that some parties may want conditional ATE (CATE) to their own demographic distribution in practice. However, how to perform multi-source causal inference using heterogeneous datasets is still an active research area (Bareinboim & Pearl, 2016; 2012; Guo et al., 2021; Yang & Ding, 2020). These solutions for heterogeneous datasets often require additional assumptions (e.g., knowing more complex causal diagrams (Bareinboim & Pearl, 2016)) and non-trivial procedures to obtain the estimate for the target population (Yang & Ding, 2020). Adopting them can complicate the setting and miss out on the collaborative component which we target, and also would require much more extensive research and discussion. Nonetheless, our work is extendable to the case of CATE where different parties require different ATE estimates for their own interested distributions, as long as collaboration can improve such estimates. In particular, adaptations are required on the definition of valuable data and the amount of contribution for this collaboration. Subsequently, the reward scheme needs to be modified accordingly since different parties may want different ATEs. We are keen to contribute and explore these options as future work.

I. Experiments

I.1. Hardware

All experiments are run on Intel Xeon Gold 6226R CPU only. Typically, 8-cores are used for more efficient parallel computing.

I.2. Impact of ρ

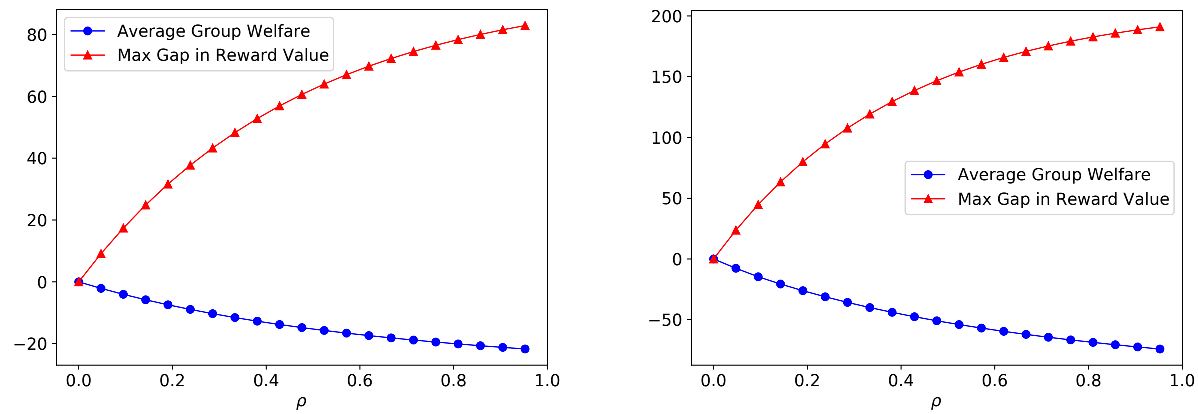
We show the effect of hyperparameter ρ on group welfare and fairness on JOBS and IHDP with the same 5 random partitions. We plot the average group welfare and the maximum difference between reward values in Fig. 5. Increasing ρ monotonically enlarges the gap between the reward value of the datasets for better fairness at the cost of reducing the (average) group welfare.

I.3. Using Other Causal Inference Estimators

In Fig. 6, we show that our framework also works for other causal inference estimators such as inverse propensity weighting (IPW) and doubly robust augmented IPW (AIPW), since bootstrapping can be used to estimate the standard error. All parties enjoy strict improvements in terms of the value of the estimate, and fairness is guaranteed since reward value is proportional to Shapley value.

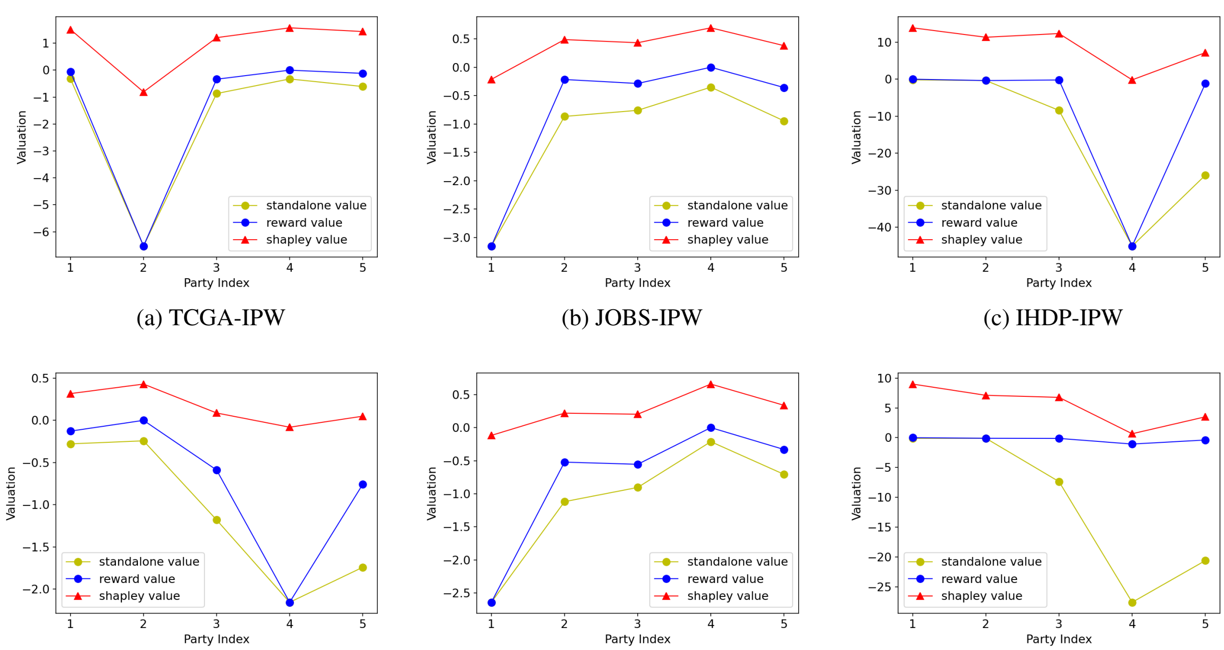
I.4. Using Other Regressors for Causal Inference

In this section, we empirically demonstrate that our framework can work seamlessly when other regressors are applied for potential outcome regression. Due to this flexibility, our framework can handle both continuous and categorical inputs.



(a) JOBS (b) IHDP

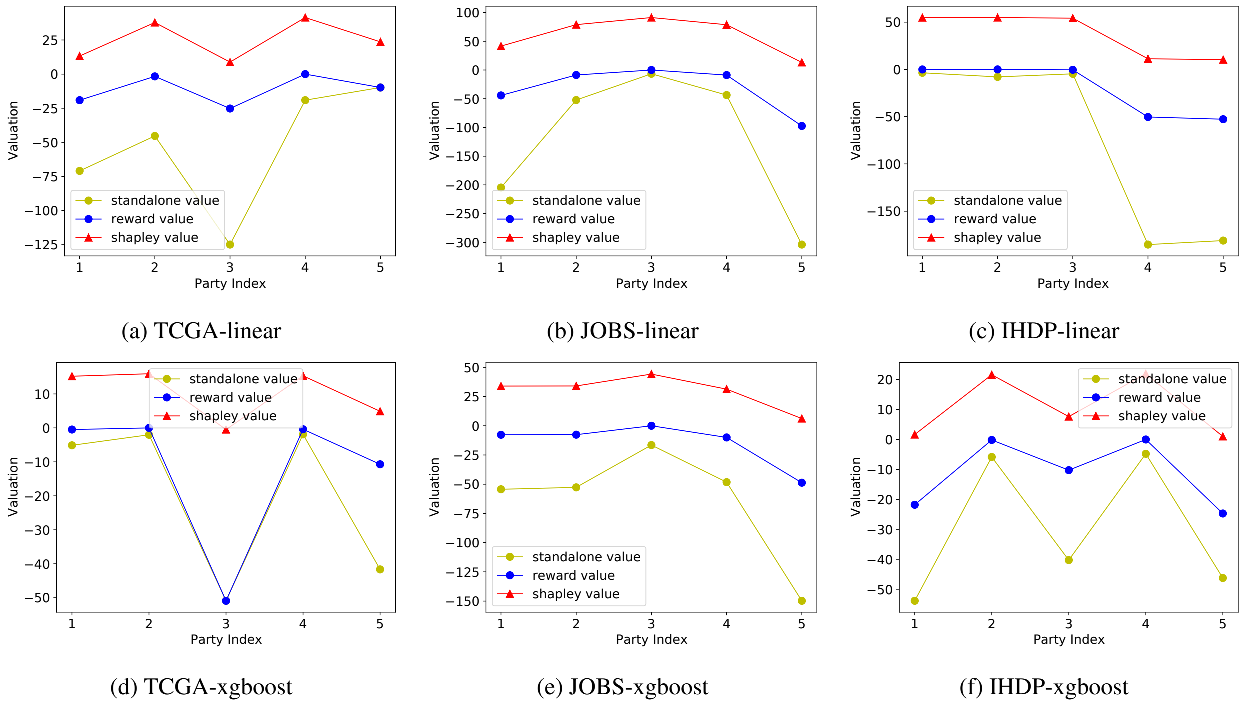
Figure 5: The effect of hyperparameter ρ



(a) TCGA-IPW (b) JOBS-IPW (c) IHDP-IPW

(d) TCGA-AIPW (e) JOBS-AIPW (f) IHDP-AIPW

Figure 6: Simulating CCI framework on three datasets with disjoint partitioning using IPW or AIPW estimator.



(a) TCGA-linear (b) JOBS-linear (c) IHDP-linear

(d) TCGA-xgboost (e) JOBS-xgboost (f) IHDP-xgboost

Figure 7: Simulating CCI framework on three datasets with disjoint partitioning using linear and xgboost regressors.

The results are shown in Figure 7 and the valuations resemble each other quite well with slight variations when different regression models are applied, which is expected because different models will have different biases and "prefer" different datasets.

I.5. Monotonicity

We empirically test the probability of violating monotonicity assumption TCGA, JOBS, IHDP datasets. The statistics are obtained from running 1000 experiments with 5 equal-sized partitions. As shown in Table 2, the probability of getting non-positive *marginal contribution* (MC) is less than 30%, and the probability of getting non-positive Shapley value is even lower at less than 1%. Since the Shapley is positive more than 99% of the time, the reward value is almost always strictly higher than the standalone value of the dataset. Thus, all parties are likely to be rewarded with more valuable treatment effect estimates than not participating, which is a strong incentive with fairness guaranteed.

Dataset	Prob Positive Shapley (%)	Prob Positive MC (%)
TCGA	99.36	73.28
JOBS	99.86	73.08
IHDP	99.54	81.26

Table 2: Empirical probability of violating monotonicity assumption.

I.6. Analysis for Surrogate of Ground Truth ATE

I.6.1. Surrogate is more accurate than the individual estimate of each party.

As discussed in Sec. 4, we use the estimate $(\hat{\tau}, \hat{\sigma})$ of the grand coalition N as the surrogate to the unavailable ground truth. We first empirically demonstrate that the surrogate obtained by collaboration is much more superior compared to the estimate obtained by each party working individually. We use IHDP and JOBS datasets since they have actual ground truth available. The comparison is done between the grand coalition estimate $\hat{\tau}_N$ vs. the individual ATE estimate of each party

$\hat{\tau}_i$. We report the *absolute error* (ABSE) of the estimate along with the standard error (SE) across 1000 runs. As shown in Table 3, the error of the grand coalition estimate is much smaller than the average error of the individual estimates.

Dataset	ABSE of $\hat{\tau}_N$ (SE)	ABSE of $\hat{\tau}_i$ (SE)
IHDP	0.05 (0)	1.62 (0.76)
JOBS	82.5 (0)	811.23 (8.61)

Table 3: Empirical result on comparing the most accurate estimate obtained by collaboration vs. no collaboration. We report the *absolute error* (ABSE) of the estimate along with the *standard error* (SE) across 1000 runs.

I.6.2. Surrogate performs well for data valuation and reward scheme.

We present further empirical analysis on how accurate the surrogate is and how that affects our data valuation and reward computation. In this experiment, we first compute the true data value $v'(C)$ with respect to the ground truth ATE τ, assuming it is available:

$$v'(C) = -\mathrm{KL}(q_C||p) = \log \hat{\sigma}_C - \log \sigma - [\hat{\sigma}_C^2 + (\hat{\tau}_C - \tau)^2]/(2\sigma) + 1/2 \tag{13}$$

where we define $\sigma = 0.1$ with a small value since the ground truth ATE has no variance. This is necessary because if we discard σ, we can no longer consider the uncertainty as part of the data valuation process and the proposed reverse KL divergence no longer works. The corresponding reward value will be denoted as r'. Then, we compare two sets of values:

1. The ranking of the standalone value of the parties under $v(C)$ (1) vs. under $v'(C)$; and
2. The ranking of the reward of the parties under r (5) vs. under r'.

We choose to compare the rankings because the absolute value of those quantities may differ a lot in value, and what matters more is which party has more valuable data in comparison to other parties. To compare the rankings, we use Kendall rank correlation coefficient $\kappa \in [-1, 1]$ to measure the similarity between two ordered indices of parties. We denote the Kendall correlation between the standalone values as κ_v and denote that between the reward values as κ_r. We report our result across 1000 runs with standard error (SE). As shown in Table 4, the correlation between the rankings is pretty high, indicating that an imperfect ground truth surrogate can still capture the correct value of the datasets most of the time.

Dataset	κ_v (SE)	κ_r (SE)
IHDP	0.78 (0.01)	0.70 (0.01)
JOBS	0.74 (0.01)	0.55 (0.01)

Table 4: Comparison between our $v(C)$ and $v'(C)$ w.r.t. ground truth ATE.

I.7. Effect of Malicious Party

We conduct an additional empirical study by converting one of the 5 simulated parties into a malicious party with large noise on IHDP dataset. In particular, we add Gaussian noise with mean 0 and variance 5. We compute the *average welfare loss* as the difference in value between the original reward value and **the value of the return** under malicious attack for all parties and *self loss* as that for the malicious party. Note that we consider the value of the return, which is the divergence between the returned estimate (sampled under the malicious setting) and the ground truth surrogate (non-malicious setting), since a malicious attack causes deviation in the surrogate too. We run the simulation 1000 times and report the error bars. As shown in Table 5, having a malicious party exhibits significant damage to the actual value of the return with respect to the ground truth surrogate, but that party cannot gain from the collaboration either.

I.8. Rewards for Edge Cases

Consider the edge case where the ground truth ATE $\tau = \epsilon \to 0$ and is non-significant. As shown in Figure 8, with extremely high probability, the perturbed estimates will consistently overestimate the ATE by having larger values depending on the

Dataset	Average Welfare Loss (SE)	Self Loss (SE)
IHDP	30.92 (5.61)	23.56 (17.26)

Table 5: Effect of Having one Malicious Party

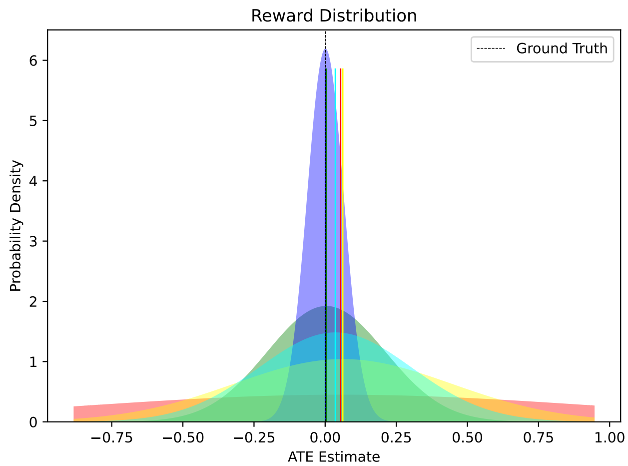


Figure 8: Reward estimates for the edge case with ground truth ATE $\tau \to 0$ $(\tau > 0)$ on synthetic dataset. The dashed black line is the ground truth (surrogate). Other lines represent the means of ATE estimates.

reward level. This behavior is still expected because the collaboration suggests that the treatment has negligible effect. Some parties will obtain the knowledge that ATE is negligible, and even parties with less valuable data will enjoy less overestimation compared to not joining the collaboration.