A Practical Data Repository for Causal Learning with Big Data

Lu Cheng[1], Ruocheng Guo*[1], Raha Moraffah*[1], K. Selçuk Candan[1], Adrienne Raglin[2], and Huan Liu[1]

[1] Arizona State University, Tempe, AZ, USA
{lcheng35,rguo12,rmoraffa,candan,huanliu}@asu.edu
[2] Army Research Laboratory, Adelphi, MD, USA
adrienne.raglin2.civ@mail.mil

Abstract. The recent success in machine learning (ML) has led to a massive emergence of AI applications and the increases in expectations for AI systems to achieve human-level intelligence. Nevertheless, these expectations have met with multi-faceted obstacles. One major obstacle is ML aims to predict future observations given real-world data dependencies while human-level intelligence AI is often beyond prediction and seeks the underlying causal mechanism. Another major obstacle is that the availability of large-scale datasets has significantly influenced causal study in various disciplines. It is crucial to leverage effective ML techniques to advance causal learning with *big data*. Existing benchmark datasets for causal inference have limited use as they are too "ideal", i.e., small, clean, homogeneous, low-dimensional, to describe real-world scenarios where data is often large, noisy, heterogeneous and high-dimensional. It, therefore, severely hinders the successful marriage of causal inference and ML. In this paper, we formally address this issue by systematically investigating existing datasets for two fundamental tasks in causal inference: *causal discovery* and *causal effect estimation*. We also review the datasets for two ML tasks naturally connected to causal inference. We then provide hindsight regarding the *advantages*, *disadvantages* and the *limitations* of these datasets. Please refer to our github repository[3] for all the discussed datasets in this work.

Keywords: Causal Learning · Treatment Effect Estimation · Causal Discovery · Datasets · Big Data · Benchmarking.

1 Introduction

The goal of a myriad of scientific research is to understand the *causal mechanisms* that reveal outcomes of interventions and counterfactuals [10]. Compared to the extensive literature on causal inference in statistics, econometrics, biostatistics and epidemiology, the interest in discovering causal relations and estimating causal effects within computer science (data science especially) has been

* Equal contribution.
[3] https://github.com/rguo12/awesome-causality-data

rapidly growing recently. On one hand, as big data has significantly influenced causal study in various disciplines, it is important to leverage machine learning (ML) techniques to enhance our capability of modeling complex and large-scale data; On the other hand, ML seeks correlations among data to predict future observations. The discovered patterns have limited use when the goal is, instead, to understand the the underlying causal mechanisms. One needs to go beyond *correlations* to assay causal structures underlying statistical dependencies.

A major challenge of studying causal inference with big data is the lack of benchmark datasets. Although growing computer power enables us to easily collect massive amount of data, it is extremely challenging to obtain the *groundtruth* from *observational data*. This is due to the fundamental question in causal inference that we can not observe the *counterfactuals*. Most existing benchmark datasets for learning causality are therefore, synthetic or semi-synthetic. They are often clean, small-scale, homogeneous and low-dimensional while real-world data is noisy, large-scale, heterogeneous and high-dimensional. Additionally, as there is no unified principle to regulate the data simulation processes, it is hard to evaluate the models and interpret the empirical results. To address these issues, we first summarize existing datasets for the two fundamental tasks in causal inference: *causal discovery*, problem of discovering the underlying causal structure of data; and *causal effect estimation*, problem of estimating causal effect of a certain set of variable on others. We seek to answer two research questions: i) What are the *advantages* and *disadvantages* of these datasets? ii) What are the *limitations* in existing datasets? In addition, we investigate datasets for two ML problems that are naturally connected to causal inference, i.e., off policy evaluation and debiasing recommender system. The main contributions are:

- We formally address an urgent but almost untouched problem that hinders the marriage of causal inference and ML. That is, the lack of benchmark datasets for causal learning with big data.
- We investigate existing datasets for two fundamental causal inference tasks, i.e., *causal discovery* and *causal effect estimation*, as well as two ML tasks that have been recently studied from the causal inference perspective.
- We answer important research questions regarding the advantages, disadvantages and limitations of these datasets. We aim to offer some crude remarks that can draw attentions from researchers to together create and share new benchmark datasets for causal learning with big data.

2 Causal Discovery

Causal discovery from empirical data is a fundamental problem in many scientific domains. Causal discovery addresses the problem of learning the underlying causal mechanisms and the causal relationships amongst variables in the data. Datasets for this task are collected from either pure observational data or with both observational data and experimental data in hand. Papers in this area can be divided into three major categories:

- Learning causal direction (causal or anti causal relations) between two variables. Specifically, given the observations $\{(x_i, y_i)\}_{i=1}^n$ of random variables, the goal is to infer the causal direction, i.e. whether $x \to y$ or $y \to x$.
- Learning the trio-relationships (V-structures) and directions among three variables.
- Learning the underlying Causal Bayesian Network (CBN) of the data which is used to show the relationships between all the variables in the data.

2.1 Datasets

Common datasets for learning causal direction between two variables are:

- *Tübingen Cause-Effect Pairs (TCEP)* [27]: This dataset consists of real-world cause-effect samples which are collected across various subject areas. The groundtruths are true causal direction provided by human experts. This dataset is expected to contain diverse functional dependencies due to the fact that pairs are collected from diverse origins.
- *AntiCD3/CD28* [31]: A dataset with 853 observational data points corresponding to general perturbations without specific interventions. This dataset is used in protein network problem.
- *Note* [26]: One innovative way of testing causal/anti-causal learning algorithms is to test the model on causal time series datasets to infer the direction of the arrow. To achieve this, [26] used a dataset containing quarterly growth rates of the real gross domestic product (GDP) of the UK, Canada and USA from 1980 to 2011.
- *Pittsburgh Bridges* [2]: There are 108 bridges in this dataset. The following 4 cause-effect pairs are known as groundtruth in this dataset. They are 1) Erected (Crafts, Emerging, Mature, Modern) → Span (Long, Medium, Short), 2) Material (Steel, Iron, Wood) → Span (Long, Medium, Short); 3) Material (Steel, Iron, Wood) → Lanes (1, 2, 4, 6); 4) Purpose (Walk, Aqueduct, RR, Highway) → type (Wood, Suspen, Simple-T, Arch, Cantilev, CONT-T).
- *Abalone* [2]: This dataset contains 4,177 samples and each sample has 4 different properties. Sex, Length, Diameter and Height. The property sex has three values, male, female and infant. The length, diameter, and height are measured in mm and treated as discrete values, similar to [Peters et al., 2010]. The groundtruth contains three cause-effect pairs.

In order to evaluate the performance of a model for distinguishing cause from effect on the above-mentioned benchmark datasets, the accuracy of the model on the datasets is calculated and reported. Next we introduce the datasets used in learning the CBN. As real-world datasets are often not available, we describe the benchmark synthetic datasets below:

- *Lung Cancer Simple Set (LUCAS)* is a synthetic dataset which was made publicly available through the causality workbench [12]. The true causal

DAG consists of 12 binary variables: 1) Smoking, 2) Yellow Fingers, 3) Anxiety, 4) Peer Pressure, 5) Genetics, 6) Attention Disorder, 7) Born on Even Day, 8) Car Accident, 9) Fatigue, 10) Allergy, 11) Coughing and 12) Lung Cancer. The true causal graph consists of causal edges between variables.
- A common approach to generate synthetic data in learning CBN is to use a random generation of chordal graphs approach [18, 36].

Moreover, there is a line of research which focuses on causal discovery problems from both observational and interventional data. In this task, we can assume that an intervention on every node of the underlying Bayesian Network is allowed. Below is the dataset designed and used in this task :
Gene perturbation data: Usually some yeast genes are selected from the data. Some observations from this data are as follows: the gene YFL044C reaches 2 genes directly and has an indirect influence on all 11 remaining genes; finally, the genes YML081W and YNR063W are reached by almost all other genes. One common way of evaluating Causal Bayesian Networks and in general structural learning problems on the above-mentioned datasets is to measure structural Hamming distance (SHD). The SHD is defined as the minimum number of edge insertions, deletions, and changes required to transform one model into another [40].

2.2 Advantages, Disadvantages, and Limitations

Advantages. There exists a number of real-world datasets for the task of learning the causal direction between two variables that can be used in future research. These datasets are collected for real world scenarios and are annotated by the experts in corresponding fields, which make these desirable and useful for research in this field.
Disadvantages. There exists no large-scale data for the task of finding the underlying Bayesian Network of the data, which is one of the most important tasks in causal inference. Moreover, no real-world data is available for the task of learning V-stucture (i.e. trio-relationships among variables), which makes it difficult to verify the proposed methods, and therefore, researchers often evaluate their proposed methods on only the datasets available for causal direction discovery and fail to show the effectiveness of them on finding the relationships between three variables.
Limitations. Many machine learning algorithms require huge number of samples to be trained on. However, for the task of causal discovery, the only real-world dataset available is LUCAS data which contains only 12 variables. Therefore, it is hard for the researchers to leverage the available dataset in big data scenarios and train a machine learning model on it. Moreover, collecting datasets with groundtruth for underlying CBN of all variables available in the data is a tremendously difficult task due to the lack of availability of human experts and resources to annotate the data and come up with the groundthruth. Another limitation is that there exists no real-world dataset for the problem of detecting V-Structure from the data, which also requires human resources and can be costly and time consuming.

3 Causal Effect Estimation

The task of causal effect estimation is to investigate to what extent manipulating the value of a potential cause would change the value of the outcome variable. Following the literature [17, 24, 33, 35] , the variable that we seek to manipulate is the *treatment* and the corresponding variable that we observe from measuring the effect of that manipulation is *outcome*. In this task, the treatment can be a single variable taking binary values, discrete values or continuous values, or multiple treatment variables that take various values. *Potential Outcomes* framework is widely used in the literature of causal effect estimation [28,30]. Potential outcomes are defined as:

Potential Outcome. Given an instance i and the treatment t, the potential outcome of i under treatment t, denoted by y_i^t, is the value that y would have taken if the treatment of instance i had been set to t.

With this definition, the individual treatment effect (ITE) is:

$$\tau_i = \mathbb{E}[y_i^t] - \mathbb{E}[y_i^c], \tag{1}$$

where y_i^c (y_i^t) denotes the potential outcome of the i-th instance under control (treatment). Intuitively, ITE is referred to as the expected difference between the two potential outcomes. Average treatment effect, or ATE, is then the average of ITE over the whole population. It is defined as: $\bar{\tau} = \mathbb{E}_i[\tau_i]$. Based on these definitions, we introduce two widely used evaluation metrics. Given the ground truth of ATE ($\bar{\tau}$) and the inferred ATE $\hat{\bar{\tau}}$, the mean absolute error (MAE) on ATE is widely adopted. It is defined as:

$$\epsilon_{MAE\_ATE} = |\bar{\tau} - \hat{\bar{\tau}}|. \tag{2}$$

In addition, the inferred ITEs can be evaluated by the precision in estimation of heterogeneous effect (PEHE). Formally, PEHE is defined as:

$$\epsilon_{PEHE} = \frac{1}{n}\sum_{i=1}^{n}(\tau_i - \hat{\tau}(\mathbf{x}_i))^2, \tag{3}$$

where τ_i denotes the ground truth ITE of the instance i and $\hat{\tau}(\mathbf{x}_i)$ signifies the corresponding estimate.

3.1 Datasets with Binary Treatment

- *Jobs.* The dataset consists of two parts. The first part is from the randomized trial study by LaLonde [19] (297 treated and 425 control). The second part is the PSID comparison group (2,490 control) [37]. The features are the same as those used in [6]. In addition, this dataset has groundtruth of ATT. One common metric used for evaluation on this dataset is policy risk (PR) [35].
- *IHDP.* This is a dataset with simulated treatments and outcomes, which is initially complied by [14]. The most widely used simulation setting is the

setting "A" in the NPCI package[4]. This dataset comprises 747 instances (139 treated and 608 control). There are 25 features describing the children and their mothers from the original IHDP data [8]. We can study the problem of estimating ITE and ATE from observational data using this dataset.

- *ACIC Benchmark.* ACIC benchmark is from ACIC data analysis challenge 2017 [13]. The features of ACIC benchmark are also from the original IHDP data [8]. Various settings have been adopted to synthesize treatments and outcomes. The ACIC dataset contains 58 features and 4,302 instances.
- *Twins.* The Twins dataset in [1] is used to study the individual treatment effect of twins' weights on their mortality in the first year of lives. In [24], the authors focused on the twins with weights less than 2kg to get a more balanced dataset in terms of the outcome. This results in a dataset consisting of 11,984 such twins. Each twin-pair is represented by 46 features relating to the parents, the pregnancy and birth. As both potential outcomes are considered as available in the dataset, to simulate an observational study, one of the two treatments need to be sampled for each twin-pair. To generate confounding bias, Louizos et al. [24] sampled treatments from the inferred propensity scores.
- *News.* The News dataset is introduced in [17]. In this dataset, each instance is a news item. The features are originally word counts. The treatment is defined as whether the news is consumed on a mobile device or on desktop. The outcome is the readers' experience. In addition, we need to assume that users prefer to read some news items on mobile devices. To model this, a topic model is trained on a large set of documents and two centroids are defined in topic space. Then, the treatment is simulated as a function of the similarity between the topic distribution of the news item and the two centroids. Finally, the potential outcomes of a news item are defined as a function of (1) the similarity between the topic distribution of the news item and the two centroids (2) and the treatment. The dataset consists of 5,000 new items and the topic model is a LDA model with 50 topics trained from the NY Times corpus[5].

3.2 Datasets with Binary Treatment and Network Information

- *BlogCatalog* is an online social network service where users can post blogs. Each instance is a blogger. Each edge signifies the friendship between two bloggers. This dataset comes with 5,196 instances, 173,468 edges and 8,189 observed features. Guo et al. [11] extended the original BlogCatalog dataset [21] for the task of causal effect estimation. In particular, treatments and outcomes are synthesized based on the observed features, the social network structure and the Homophily phenomenon [34].
- *Amazon* [29] is an extension of the original dataset [25]. The goal is to estimate the causal effect of positive (or negative) reviews on the sales of

[4] https://github.com/vdorie/npci
[5] Downloaded from the UCI repository [7]

products. Each instance is a product. Each edge represents the co-purchase relationship. The observed features are bag-of-word representation of the product description. Two datasets are created, one for positive and one for negative reviews. For the positive (negative) case, we say a product is under treatment iff (1) receives more than three reviews and (2) is rated higher (lower) than three stars. The counterfactual outcome is set as the observed sales of the most similar product with an opposite treatment status. The positive (negative) dataset contains 50,000 positive (20,000 negative) instances, 10,000 (5,000) controlled instances and 96,132 (28,136) edges.

3.3 Datasets with Multiple Treatments

- *Twins-Mult.* Yoon et al. [42] extended the Twins dataset to 4 treatments by considering the combination of the original treatment and the sex of the infant. The method to sample treatments are adapted accordingly.
- *News-Mult.* Schwab et al. [33] adapted the News dataset to multiple treatments. Instead of using two centroids, $k+1$ centroids are randomly picked in the topic space where k is the number of treatments and the rest represents the control group. Then the treatment is sampled from a Bernoulli distribution $t|x \sim Bern(softmax(\kappa \bar{y}_j))$ where $\kappa \in \{10, 7\}$ and the unscaled outcome is calculated as $\bar{y}_j = \tilde{y}_j * [D(z(X), z_j) + D(z(X), z_c)]$. $z(X)$ denotes the topic distribution of the news item with bag-of-word features X, z_j signifies the centroid of the instances receiving treatment j, z_c represents the centroid for the control, and $\tilde{y}_j \sim \mathcal{N}(\mu_j, \sigma_j) + \epsilon$ where $\mu_j \sim \mathcal{N}(0.45, 0.15)$, $\sigma_j \sim \mathcal{N}(0.1, 0.05)$ and $\epsilon \sim \mathcal{N}(0, 0.15)$. D is the Euclidean distance function. Then the true outcome of the j-th treatment is $y_j = C\bar{y}_j$, where $C = 50$.
- *TCGA.* In [33], the authors introduced the TCGA dataset which is a collection of gene expression data from types of cancers in 9,659 individuals [41]. There are four possible clinical treatments: medication, chemotherapy, surgery or both surgery and chemotherapy. The outcome is the risk of recurrence of cancer. Similar to the News dataset, $k+1$ points in the original feature space (gene expression features) are selected as centroids. Treatments and outcomes are simulated accordingly.

3.4 Datasets with Continuous Treatment

The treatment can also take continuous values. Here, we introduce a dataset for the study of causal effect estimation with continuous variable.

NMES. The National Medical Expenditures Survey (NMES) dataset is complied by [9]. We study the problem of estimating the treatment effect of the amount of smoking on the medical expenditure. Both the treatment and the outcome variables are continuous. The dataset consists of 10 features describing each of the 9,708 individuals.

3.5 Advantages, Disadvantages, and Limitations

Advantages. Most of the existing datasets are collected to solve treatment effect estimation problems. For example, the Jobs dataset is collected to answer the causal question: Does job training help people to get employed? Moreover, studying these datasets can provide insights for decision making in real-world scenarios. For example, an employer can decide whether it is necessary to participate the job training program based on the individual treatment effect.
Disadvantages. It is often impossible to collect data with ground truth for counterfactual outcomes – outcomes could have been observed iff another treatment had been assigned. Instead, researchers mainly rely on semi-synthetic datasets, where treatments and outcomes are synthesized based on certain data-generating process. Therefore, developing high-quality data simulation models can be a time-consuming and labor-intensive task.
Limitations. Existing benchmark datasets are not suitable in estimating causal effects in many real-world applications due to the unavailability of counterfactual outcomes. For example, it is convenient to collect climate data from Google earth engine and user behavior data from Twitter in order to develop ML models to predict user behavior from climate statistics. Nevertheless, to understand how climate changes influence user behavior, we need to collect data from the same user under exactly the same conditions with different climate. This is often impossible in real-world scenarios.

In terms of estimating average treatment effects, the challenges arise from how to design cheap, easy-to-implement, reliable and ethical experiments. In addition, the importance of reducing the sample size and time in need for a statistically significant randomized trial is still underestimated in the data mining and machine learning community. Another limitation of current datasets for causal effect estimation is the missing of the underlying structure between instances. The potential types of structure include (but are not limited to) networks and temporal dependencies.

4 Causal Inference in ML

4.1 Off-policy Evaluation

Given that an existing policy h_0 selects actions based on item features and observes corresponding rewards (e.g., online Q&A communities [22], recommender systems [32]). This process generates *log* data with the form (x, y, δ, p) where $x \in \mathcal{X}$ is the context (feature vector), $y \in \mathcal{Y}$ is the selected action. $\mathcal{X}$ and $\mathcal{Y}$ are the input space and the output space respectively. p is the probability of y being selected given x and $\delta(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ denotes the feedback/reward received. The goal of off-policy evaluation is to exam the performance of a new policy h on future observations using the log data generated from h_0.

First, we give a formal problem definition. Given the input features $x \in \mathcal{X}$, the output prediction of selected action $y \in \mathcal{Y}$ and a hypothesis space $\mathcal{H}$ of *stochastic policies* [38], which is calculated from the observed data. Additionally, the inputs

are assumed drawn from a fixed but unknown distribution $Pr(\mathcal{X}), x \overset{i.i.d}{\sim} Pr(\mathcal{X})$. A hypothesis $h(\mathcal{Y}|x) \in \mathcal{H}$ makes predictions by sampling $y \sim h(\mathcal{Y}|x)$. In an interactive learning system, we can only observe the feedback $\delta(x, y)$ for y sampled from $h(\mathcal{Y}|x)$. For instance, in a recommender system, $\mathcal{X}$ are the attributes of the items, $\mathcal{Y}$ is set of items recommended by the system, and δ denotes the user feedback, e.g., whether a user clicks on the item or not. In precision medicine, $\mathcal{X}$ denotes the patients' attributes, $\mathcal{Y}$ is the set of received treatments. We then collect the outcomes δ from patients. A large δ indicates high user's satisfaction with y for x. The expected rewards of a hypothesis $R(h)$ is defined as [38]

$$R(h) = \mathrm{E}_{x \sim Pr(\mathcal{X})} \mathrm{E}_{y \sim h(\mathcal{Y}|x)}[\delta(x, y)]. \tag{4}$$

Then, the goal is to maximize the reward with policy $h(\mathcal{Y}|x)$ given data $\mathcal{D} = \{(x_1, y_1, \delta_1), (x_2, y_2, \delta_2), ..., (x_n, y_n, \delta_n)\}$ collected from the system using policy $h_0(\mathcal{Y}|x)$, i.e., $y_i \sim h_0(\mathcal{Y}|x), \delta_i = \delta(x_i, y_i)$. Evaluation of the proposed policy is extremely hard due to sample selection bias and partial information.

Dataset from Real World.

Music Streaming Sessions Dataset (MSSD). This dataset from Spotify[6] consists of over 160 million listening sessions with user interaction information. It has metadata for approximately 3.7 million unique tracks referred to in the logs, making it the largest collection of such track data currently available to the public [5]. In particular, it consists of music streaming sessions with corresponding user interactions, audio features and metadata describing the tracks streamed during the sessions, and snapshots of the playlists listened to during the sessions [5]. The log data contains rich information such as session id, timestamp, contextual information about the stream, and the timing and type of user interactions within the stream. A subset of MSSD is crawled and labelled by a uniformly random shuffle to satisfy the conditions of RCT.

Semi-simulated Datasets.

Bandit Data Generation. Despite log data is ubiquitous in the real world, it is often hard to gather for researchers in academia. In search of alternatives, synthetic or semi-synthetic data is often used for off-policy evaluation. Here, we present a widely used bandit data generating approach proposed in [3]. This approach converts the training partition of a full-information multi-class classification dataset $D^* = \{[x_i, y_i^*]\}_{i=1,...,n}$ with $y_i^* \in \{0, 1\}^k$ into a partial-information bandit dataset for training off-policy learning methods while the test dataset remains intact to evaluate the new policy. To this end, the optimal policy is known because $\delta(x_i, y_i^*) > \delta(x_i, \neg\, y_i)$ where $\neg\, y_i$ is any of the items/treatments other than y_i^*. Therefore, given x_i, the optimal policy selects action y_i^*. Then we simulate a bandit feedback dataset from a logging policy h_0 by sampling $y_i \sim h_0(\mathcal{Y}|x_i)$ and collecting feedback $\triangle(y_i^*, y_i)$, which is the loss between groundtruth and the recommended item. h_0 can be logistic regression and is often trained with a small portion (e.g. 5%) of the training set. $\triangle(y^*, y)$ is then the Hamming loss or Jaccard index between the label y^* and the sampled label y for input x. This completes the procedure of generating a bandit

[6] https://www.spotify.com/

dataset $\mathcal{D} = \{[x_i, y_i, \triangle(y_i^*, y_i), h_0(y_i|x_i)]\}_{i\in\{1,\ldots,n\}}$. One thing to note is that the propensity score function $h_0(y_i|x_i)$ is usually estimated from data directly, which may introduce undesired biases. A large-scale real-world dataset[7] containing accurately logged propensities is introduced in [20].

Limitations. While this data generating method has been adopted wildly in off-policy evaluation in contextual bandits [16, 38, 39], it has several limitations:

- It might not be clear how it can be used in other applications of off-line evaluations. Take the medical study for an example, mapping the concept of binary multi-label $\in \{0, 1\}$ to treatments indicates that several drugs may be assigned to the same patient simultaneously. This might be detrimental to the patients' health due to the interactions between drugs. In addition, estimation of propensity score function using a small portion of the supervised training set is not appropriate in medical study as the underlying mechanism of treatment selection is often not fully understood.
- The predefined hypothesis h_0 can largely affect the performance of the new policy. By using the above mentioned method, we can obtain h_0 with nearly 100% accuracy, i.e., $y = y^*$ for all x in the training set. Nevertheless, it is often impossible for a real-world system to have an optimal policy. Consequently, how many training data should be used to estimate h_0? What is the desirable accuracy that h_0 should achieve? Answering these questions is critical for the evaluation.
- The mismatch of synthetic data and the observed data from true environment is often unavoidable in practice, resulting in policies that do not generalize to the real environment [15].

4.2 Causal Inference for Recommendation

Causal inference is also particularly useful in learning de-biased recommender policies. Consider a recommender system that takes as input a user $u_i \in \mathcal{U}$ from the user population $\mathcal{U}$ and outputs the prediction of possible products $p_j \in \mathcal{P}$. The recommendation policy decides what products the recommender system shows to its users. Most existing "de-biased" recommendation systems aim to find the optimal treatment recommendation policy that maximizes the reward with respect to the control recommendation policy for each user, i.e., individual treatment effect. Traditional recommender systems are biased as they use the click data (or ratings data) to infer the user preferences. These data encode users' selection bias, i.e., users do not consider each product independently.

The input data of learning a recommendation policy consists of products each user decided to look at and those each user liked/clicked. The treatment is the recommended products and the outcome is whether this user clicks this product. Standard datasets for recommender systems are not applicable in the evaluation of the deconfounded recommender systems due to the lack of outcomes for counterfactuals. Consequently, simulated or semi-simulated datasets are often the preferred alternatives. The core idea of generating an eligible dataset to

[7] http://www.cs.cornell.edu/∼adith/Criteo/

evaluate a recommendation policy is to ensure the distributions of the training and test set are different, that is, to exam if the deconfounded recommendation policy is generalizable. A more generalizable policy indicates a less-biased recommender system. Next, we introduce several datasets that have been used in recent publications [4, 23, 32]. Based on the different data collection/generation mechanisms, we divide the data into three categories: data collected from RCT, semi-simulated datasets and simulated datasets.

Randomized Control Trial (RCT).

Yahoo-R3. Music ratings collected from Yahoo! Music services. This dataset contains ratings for 1,000 songs collected from 15,400 users with two different sources. One of the sources consist of ratings for randomly selected songs collected using an online survey conducted by Yahoo! Research. The other source consists of ratings supplied by users during normal interaction with Yahoo! Music services. The rating data includes at least ten ratings collected for each user during the normal use of Yahoo! Music services and exactly ten ratings for randomly selected songs for each of the first 5,400 users in the dataset. The dataset includes approximately 300,000 user-supplied ratings, and exactly 54,000 ratings for randomly selected songs[8].

Semi-synthetic Datasets.

- *MovieLens10M*. User-movie ratings collected from a movie recommendation service. It has 71,567 unique users and 10,677 unique products. The ratings are on a 1–5 scale [4]. The treatment is binary indicating if a user has rated an item, the outcome is if rating is greater or equal to 3.
- *Netflix*. This dataset includes 480,189 unique users and 17,770 unique products. The treatment is if a user has rated an item, the outcome is if rating is greater or equal to 3.
- *ArXiv*. User-paper clicks from the 2012 log-data of the arXiv pre-print server. The data are binarized: multiple clicks by the same user on the same paper are considered to be a single click. The treatment in this dataset is if a user has viewed the abstract of a paper, outcome is if she downloaded the paper.

Now the question is how to generate new datasets from existing datasets to evaluate de-biased recommender systems. One common approach is to ensure the different distributions between the training/validation sets and the test set. Previous work [4,23] has tried to create two test splits from the standard datasets – regular and skewed. The regular split is generated by randomly selecting the exposed items for each user into training/validation/test sets with proportions 70/20/10, i.e., the standard method that researchers use to evaluate recommendation models. The skewed split re-balances the splits to better approximate an intervention. In particular, it first samples a test set with roughly 20% of the total exposures such that each item has uniform probability. Training and validation sets are then sampled from the remaining data (as in a regular split) with 70/10 proportions. The test set then has a different exposure distribution from the training and validation sets. Experimental results have shown that

[8] https://webscope.sandbox.yahoo.com/catalog.php?datatype=r

causality-embedded recommender systems can largely improve the performance on the skewed split while present similar performance compared to baseline models on the regular split.

Simulated Datasets.

Coat Shopping Dataset [32]. This is a synthetic dataset that simulates customers shopping for a coat in an online store. The training data was generated by giving Amazon Mechanical Turkers a simple web-shop interface with facets and paging. Users were asked to find the coat in the store that they wanted to buy the most. Afterwards, they had to rate 24 of the coats they explored (self-selected) and 16 randomly picked ones on a five-point scale. The dataset contains ratings from 290 Turkers on an inventory of 300 items. The self-selected ratings are the training set and the uniformly selected ratings are the test set.

Limitations. RCT for a recommender system is often not an option in real-world applications. For example, a recommender system that randomly recommends songs to its users can largely degrade user experience. Leveraging simulated/semi-simulated datasets to show the generalizability of a de-biased recommender system is technically sound, but the mismatch of synthetic data and the observed data from the true environment is often unavoidable.

Humans are biased in nature. A desired recommender systems should be able to capture idiosyncratic user preferences in order to make personalized recommendations. Therefore, debiasing recommender system may not necessarily make better recommendations than a biased one. A more intriguing question to ask may be what causes a recommendation system to make certain suggestions and how to quantify their causal effects. Such systems are causally interpretable and can help identify the underlying causal relations between users and items. As a result, another limitation of current datasets is the lack of formal definitions of elements for causal studies such as treatments that indicate user's characteristics, features of recommendable items, and the corresponding potential outcomes.

5 Conclusions and Future work

In this paper, we discuss the **advantages, disadvantages** and **limitations** of existing benchmark datasets for the two fundamental tasks in causal inference. We then present applications of causal inference in two standard ML tasks and investigate how to leverage existing datasets to evaluate the causality-embedded ML models. Our goal is to provide easier access to researchers who share similar research interests in causal learning and more importantly, to draw attentions and seek contributions from research communities to together create and share new benchmark datasets for causal learning with big data.

Acknowledgement

This material is based upon work supported by ARO/ARL and the National Science Foundation (NSF) Grant #1610282, NSF #1909555.

References

1. Almond, D., Chay, K.Y., Lee, D.S.: The costs of low birth weight. The Quarterly Journal of Economics **120**(3), 1031–1083 (2005)
2. Bache, K., Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
3. Beygelzimer, A., Langford, J.: The offset tree for learning with partial labels. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 129–138. ACM (2009)
4. Bonner, S., Vasile, F.: Causal embeddings for recommendation. In: Proceedings of the 12th ACM Conference on Recommender Systems. pp. 104–112. ACM (2018)
5. Brost, B., Mehrotra, R., Jehan, T.: The music streaming sessions dataset. In: The World Wide Web Conference. pp. 2594–2600. ACM (2019)
6. Dehejia, R.H., Wahba, S.: Propensity score-matching methods for nonexperimental causal studies. Review of Economics and statistics **84**(1), 151–161 (2002)
7. Dua, D., Graff, C.: UCI machine learning repository (2017), `http://archive.ics.uci.edu/ml`
8. Duncan, G.J., Brooks-Gunn, J., Klebanov, P.K.: Economic deprivation and early childhood development. Child development **65**(2), 296–318 (1994)
9. Galagate, D., Schafer, J., Galagate, M.D.: Package 'causaldrf' (2015)
10. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: Problems and methods. arXiv preprint arXiv:1809.09337 (2018)
11. Guo, R., Li, J., Liu, H.: Learning individual treatment effects from networked observational data. arXiv preprint arXiv:1906.03485 (2019)
12. Guyon, I., Aliferis, C., Cooper, G., Elisseeff, A., Pellet, J.P., Spirtes, P., Statnikov, A.: Design and analysis of the causation and prediction challenge. In: Guyon, I., Aliferis, C., Cooper, G., Elisseeff, A., Pellet, J.P., Spirtes, P., Statnikov, A. (eds.) Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008. Proceedings of Machine Learning Research, vol. 3, pp. 1–33. PMLR, Hong Kong (03–04 Jun 2008), `http://proceedings.mlr.press/v3/guyon08a.html`
13. Hahn, P.R., Dorie, V., Murray, J.S.: Atlantic causal inference conference (acic) data analysis challenge 2017. Tech. rep., Tech. rep (2018)
14. Hill, J.L.: Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics **20**(1), 217–240 (2011)
15. Jiang, N., Li, L.: Doubly robust off-policy value evaluation for reinforcement learning. arXiv preprint arXiv:1511.03722 (2015)
16. Joachims, T., Swaminathan, A., de Rijke, M.: Deep learning with logged bandit feedback (2018)
17. Johansson, F., Shalit, U., Sontag, D.: Learning representations for counterfactual inference. In: International conference on machine learning. pp. 3020–3029 (2016)
18. Kocaoglu, M., Dimakis, A., Vishwanath, S.: Cost-optimal learning of causal graphs. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1875–1884. JMLR. org (2017)
19. LaLonde, R.J.: Evaluating the econometric evaluations of training programs with experimental data. The American economic review pp. 604–620 (1986)
20. Lefortier, D., Swaminathan, A., Gu, X., Joachims, T., de Rijke, M.: Large-scale validation of counterfactual learning methods: A test-bed. arXiv preprint arXiv:1612.00367 (2016)
21. Li, J., Guo, R., Liu, C., Liu, H.: Adaptive unsupervised feature selection on attributed networks. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 92–100. ACM (2019)

22. Li, Y., Guo, R., Wang, W., Huan, L.: Causal learning in question qualityimprovement. In: 2019 BenchCouncil International Symposium on Benchmarking, Measuring and Optimizing (Bench'19) (2019)
23. Liang, D., Charlin, L., Blei, D.: Causal inference for recommdendation (2016)
24. Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. In: Advances in Neural Information Processing Systems. pp. 6446–6456 (2017)
25. McAuley, J., Pandey, R., Leskovec, J.: Inferring networks of substitutable and complementary products. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 785–794. ACM (2015)
26. Mitrovic, J., Sejdinovic, D., Teh, Y.W.: Causal inference via kernel deviance measures. CoRR **abs/1804.04622** (2018), `http://arxiv.org/abs/1804.04622`
27. Mooij, J.M., Peters, J., Janzing, D., Zscheischler, J., Schölkopf, B.: Distinguishing cause from effect using observational data: methods and benchmarks. CoRR **abs/1412.3773** (2014), `http://arxiv.org/abs/1412.3773`
28. Neyman, J.S.: On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). Annals of Agricultural Sciences **10**, 1–51 (1923)
29. Rakesh, V., Guo, R., Moraffah, R., Agarwal, N., Liu, H.: Linked causal variational autoencoder for inferring paired spillover effects. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management. pp. 1679–1682. ACM (2018)
30. Rubin, D.B.: Bayesian inference for causal effects: The role of randomization. The Annals of statistics pp. 34–58 (1978)
31. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A., Nolan, G.P.: Causal protein-signaling networks derived from multiparameter single-cell data. Science **308**(5721), 523–529 (2005). https://doi.org/10.1126/science.1105809, `https://science.sciencemag.org/content/308/5721/523`
32. Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., Joachims, T.: Recommendations as treatments: Debiasing learning and evaluation. arXiv preprint arXiv:1602.05352 (2016)
33. Schwab, P., Linhardt, L., Karlen, W.: Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks. arXiv preprint arXiv:1810.00656 (2018)
34. Shakarian, P., Bhatnagar, A., Aleali, A., Shaabani, E., Guo, R.: Diffusion in social networks. Springer (2015)
35. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 3076–3085. JMLR. org (2017)
36. Shanmugam, K., Kocaoglu, M., Dimakis, A.G., Vishwanath, S.: Learning causal graphs with small interventions. In: Advances in Neural Information Processing Systems. pp. 3195–3203 (2015)
37. Smith, J.A., Todd, P.E.: Does matching overcome lalonde's critique of nonexperimental estimators? Journal of econometrics **125**(1-2), 305–353 (2005)
38. Swaminathan, A., Joachims, T.: Counterfactual risk minimization: Learning from logged bandit feedback. In: International Conference on Machine Learning. pp. 814–823 (2015)
39. Swaminathan, A., Joachims, T.: The self-normalized estimator for counterfactual learning. In: advances in neural information processing systems. pp. 3231–3239 (2015)

40. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Mach. Learn. **65**(1), 31–78 (Oct 2006). https://doi.org/10.1007/s10994-006-6889-7, http://dx.doi.org/10.1007/s10994-006-6889-7
41. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., et al.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113 (2013)
42. Yoon, J., Jordon, J., van der Schaar, M.: Ganite: Estimation of individualized treatment effects using generative adversarial nets (2018)