A Practical Data Repository for Causal Learning with Big Data

Lu Cheng
Department of Computer Science
Arizona State University
Tempe, USA
lcheng35@asu.edu

Raha Moraffah
Department of Computer Science
Arizona State University
Tempe, USA
Raha.Moraffah@asu.edu

Ruocheng Guo
Department of Computer Science
Arizona State University
Tempe, USA
rguo12@asu.edu

Kasim Selcuk Candan
Department of Computer Science
Arizona State University
Tempe, USA
candan@asu.edu

Huan Liu
Department of Computer Science
Arizona State University
Tempe, Arizona
huanliu@asu.edu

***Abstract*—The recent success in machine learning (ML) has led to an explosive emergence of AI applications and the increasing expectations for AI systems to achieve human level intelligence. Nevertheless, these expectations have met with multi-faceted obstacles. One such obstacle is that ML has focused on predictions of future observations given the real-world data dependencies while human-level intelligence AI is often beyond prediction and seeks the underlying causal mechanism. Another obstacle is the era of big data has significantly influenced the causality analysis on various disciplines. Therefore, it is necessary to leverage effective ML techniques to boost causal learning with *big data*. Existing benchmark datasets for causal inference are not appropriate in a sense that they are mostly too "ideal" (e.g., small, clean, homogeneous, low-dimensional) to be practical in real-world scenarios with large, noisy, heterogeneous and high-dimensional data. It, therefore, hinders the successful marriage of causal inference and ML. In this paper, we formally address this issue by systematically investigating the existing datasets for two fundamental tasks in causal inference: *causal discovery* and *causal effect estimation*, as well as those for two ML tasks that are naturally connected to causal inference. We then provide hindsight of the pros, cons and the limitations of the current datasets. The purpose is to reveal our efforts, more importantly, seek attentions from all the research communities to together contribute to creating and sharing new benchmark datasets for causal learning in the near future.**

***Index Terms*—Causal Learning, Datasets, Big Data, Benchmarking**

I. Introduction

The goal of many sciences is to understand the *causal mechanism* that demands intervention and retrospective thinking. For example, researchers in biology gather data about the gene activation levels in a standard system in order to understand the causal relation among genes, and to predict the effects of turning some genes on or off would be [1]. Similarly, answering fundamental questions in health, social and behavioral sciences often needs data interventions/manipulations to acquire some knowledge of the data-generating process [2]. Understanding the data-generating process, or estimating the effects of variables after some other variables have been intervened, is generically *causal inference*. Therefore, causality manifests generic relationship between an effect and the cause that gives rise to it [3].

Compared to the large literature on causal inference in statistics, econometrics, biostatistics and epidemiology, the interest in discovering or estimation of causal relations within computer science (data science especially) is rapidly spread recently and in part by technological developments, especially the advanced data collection methods, storage techniques and the unprecedented power of modern computers [1]. The benefits of the marriage between causal inference and data science unfold in two directions: i) On one hand, the age of big data has significantly influenced the causality analysis on various disciplines as the identification of causal relation among big data has dramatically increased. Consequently, it can be particularly useful to leverage data mining and machine learning techniques to enhance our capability of modeling complex and large-scaled data, and therefore, boosting causal inference with big data. ii) On the other hand, ML explores data to seek dependencies (correlations) in the world with the goal of predicting future observations. The uncovered patterns can be less useful when the goal is instead to understand the causal relation. However, one can go beyond *correlations*, assaying causal structures underlying statistical dependencies to build more robust, adaptive and interpretable machine learning models.

Nevertheless, learning causal inference with big data presents the challenge of lacking proper benchmark data that is representative for *big data*. Although the advanced techniques and growing computer power enable us to easily collect massive amount of data from various sources, it is extremely challenging to gather reliable *ground truth* for *observational data* for causal studies. The difficulty arises from many reasons. One reason is the fact that, often times, we can only observe the *factual outcome* from observational data,

but not *counterfactual outcome*, a major component in the studies of causal learning. For instance, to estimate the causal effect of a given drug, ideally, we need to observe the outcome of the same group of patients under the same condition having and not having the drug. This is clearly impossible in reality. An alternative, which is often unethical, is to randomly assign the drug to patients and conduct *randomized control trial* [4] to reduce certain sources of bias. Another example is randomly recommending songs to online users to reduce the bias from various sources such as users' selection bias, however, the results can cause large damages to the reputation of a product in a company.

The second reason is the lack of reliable domain knowledge about the data-generating process and the underlying causal mechanism. As real-world data contains a sea of noise and presents complex formats, it is often impossible for domain experts to identify the ground truth, e.g., causes and causal relations/directions, for causal studies. Consequently, most existing benchmark datasets for causal studies are synthetic data collected from simulated experiments. These datasets are often not appropriate for ML models, which are designed for real-world data that is noisy, large-scale, heterogeneous and high-dimensional. Therefore, the training and evaluation of ML-based causal learning models can be problematic. Another drawback with the synthetic datasets is the lack of unified principles to regulate various data simulation processes.

As a result, the goal of this paper is to show our efforts, more importantly, seek attentions and contributions from research communities to create and share new benchmark datasets for *causal learning with big data*. To achieve this, we first systematically summarize existing datasets for the two fundamental tasks in causal inference: *causal discovery*, problem of discovering the underlying causal structure of the data; and *causal effect estimation*, problem of estimating causal effect of one variable on another. On top of that, we then seek to answer three important research questions: i) What are the *pros* and *cons* of these datasets? ii) What are the potential approaches to explore these datasets? iii) What are the missing parts in existing datasets? In addition, we depict the datasets for two well-studied ML problems–Off policy evaluation and recommender system–that are naturally connected to causal inference. The main contributions of this paper are:

- We formally address an urgent but not well-noticed problem that hinders the marriage of causal inference and ML, that is, the lack of new benchmark datasets for causal learning in the era of big data. Without representative and high-quality data, the scientific progress can be slow and limited.
- We systematically summarize the existing datasets for two fundamental causal inference tasks, i.e., *causal discovery* and *causal effect estimation*, as well as two ML tasks – Off-policy evaluation and recommender system. Based on that, we propose and answer three important research questions to show our efforts, and seek attentions from research communities to create new benchmark datasets.

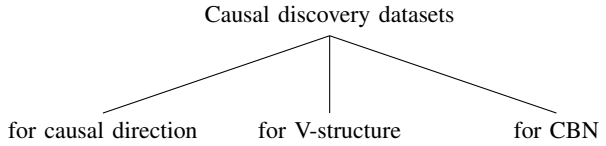

Fig. 1. Datasets categories for causal discovery tasks.

The remainder of this paper is organized as follows: we summarize and analyze the existing datasets for causal discovery and causal effect estimation in Sec. 2 and Sec. 3 respectively. Datasets for two ML tasks are introduced in Sec. 4 and Sec. 5. We conclude the paper with future key avenue in Sec. 6.

II. Causal Discovery

Causal discovery from empirical data is a fundamental problem in many scientific domains. Causal discovery address 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. To be more specific, given the observations $\{(x_i, y_i)\}_{i=1}^{n}$ of random variables, the goal is to infer the causal direction, i.e. whether $x \rightarrow y$ or $y \rightarrow x$.
- Learning the trio-relationships (v-structures) and directions among three variables
- Learning the underlying CBN which is used to show the relationships between all the variables in the data.

A. Datasets

In this section, we explain the available benchmark datasets for different causal discovery tasks. Figure 1 shows an overview of categorization of datasets for different tasks in causal discovery.

Common datasets for the first causal discovery task (i.e. learning causal direction between two variables) are as follows:

- Tbingen Cause-Effect Pairs (TCEP) [5]: This dataset is comprised of real-world cause-effect samples that are collected across very diverse subject areas with the true causal direction provided by human experts. Due to the heterogenous origins of the data pairs, many diverse functional dependencies are expected to be present in TCEP.
- AntiCD3/CD28 [6]: A dataset with 853 observational data points corresponding to general perturbations without specific interventions. This dataset is used in Protein network problem.
- Note: 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. In order to do so [7] 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 dataset [8]: There are 108 bridges in this dataset. The following 4 cause-effect pairs are known as ground truth in this experiment. 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 Data Set [8]: This dataset contains 4177 samples and each sample has 4 different properties. The ground truth contains three cause-effect pairs, Sex $\rightarrow$ Length, Diameter, 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].

Below is a list of datasets used in learning the Causal Bayesian Network of the data:

- Lung Cancer Simple Set (LUCAS) is a synthetic dataset which was made publicly available through the causality workbench [9]. 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.
- Usually, a random generation of chordal graphs approach is used to generate the Causal Bayesian Network.

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 datasets 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.

B. Pros and Cons, What is Missing?

Pros. There exists a number of real-world datasets for the task of learning the causal direction between two variables that can be leveraged in future research in this direction. 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.

Cons. Even though there are various numbers of datasets available for inferring the causal direction task, 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 if the proposed methods can be leveraged for this task, 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.

What is Missing? 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. This makes it 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 causal bayesian network 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. As mentioned before also, 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.

III. 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 we are interested in. Following the literature [10]–[13] , the variable that we imagine to manipulate is the treatment and another variable that we observe measure the effect of that manipulation is the outcome. In this task, we can have either binary treatment variables which take 0 or 1, or we can have multiple treatment variables. To give a formal definition of individual treatment effect (ITE), we first define *potential outcomes* framework which is widely used in the literature of causal effect estimation [14], [15]:

Definition 1: **Potential Outcomes.** 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.

Then we can formulate ITE for the i-th instance as:

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

where y_i^c denotes the potential outcome of the i-th instance under control. Intuitively, ITE is referred to as the expected outcome under treatment subtracted by the expected outcome under control, which reflects how much improvement of the outcome is caused by the treatment. With ITE defined, the formulation of average treatment effect (average treatment effect for the treated), or ATE (ATT), can be provided through taking the average of ITE over the whole population (the instances under a certain treatment) as: $ATE = \mathbb{E}_i[\tau_i]$ ($ATT^t = \mathbb{E}_{i:t_i=t}[\tau_i]$).

A. Datasets with Binary Treatment

Below is a list of datasets used for estimating the effect of binary treatments:

Jobs. The dataset consists of two parts. The first part is from the randomized trial study by LaLonde [16] (297 treated and 425 control). The second part is the PSID comparison group

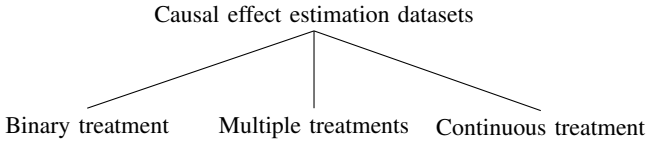

Fig. 2. Datasets for Causal Effect Estimation

(2490 control) [17]. The features are the same as those used in [18]. This dataset has ground truth of ATT. As in [11], the metric called policy risk (PR) can also be applied to evaluation on this dataset as it has a randomized trial subset.

IHDP. With the IHDP dataset, we can study the problem of estimating individual treatment effects and average treatment effect from observational data. This is a dataset with simulated treatments and outcomes, which is initially complied by [19]. The most widely used simulation setting is the setting "A" in the NPCI package[1]. This dataset comprises 747 instances (139 treated and 608 control). There are 25 features describing children and their mothers collected from the original IHDP data [20].

ACIC Benchmark. ACIC benchmark is the dataset from ACIC data analysis challenge 2017 [21]. It is similar to the IHDP dataset in terms of the features, treatment and outcome. The features are also from the original IHDP data [20]. The ACIC dataset includes 58 features and 4,302 instances.

Twins. The Twins dataset is a subset of the study [22]. In this dataset, we study the individual treatment effect of being the heavier (lighter) twin on the mortality of the twin in the first year of life. In [13], the authors focused on the twins with weights less than 2kg to get a more balanced dataset in terms of the outcome. Thus, the dataset consists of 11,984 such twins. Each twin-pair comes with 46 features relating to the parents, the pregnancy and birth. As both potential outcomes are available in the dataset, to simulate an observational study, one of the two treatments need to be sampled for each twin-pair. To create confounding bias, authors of [13] sample treatments by a function of the features.

News. The News dataset is introduced by the work [10]. In this dataset, each instance is a new 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. The assumption is that users may 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 similar between the topic distribution of the new item and the two centroids. Finally, the potential outcomes of a news item are simulated as a function of (1) the similar between the topic distribution of the new 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[2].

1) Datasets with Multiple Treatments: As mentioned, we can have multiple treatments variables. Datasets used in these problems are listed below:

Twins-Mult. In [24], the authors extend the Twins dataset to 4 treatments by combining being heavier (lighter) with the sex of the infant. The method to sample treatments are adapted accordingly.

News-Mult. In [12], the authors adapted the News dataset to multiple treatments. Instead of two centroids, $k+1$ centroids are picked in the topic space, k of them are for k treatments and the last one is for 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 news item X, z_j signifies the centroid for the 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 real outcome of the j-th treatment is $y_j = C\bar{y}_j$, where $C = 50$.

TCGA. In [12], the authors introduced the TCGA dataset which is a collection of gene expression data from types of cancers in 9,659 individuals [25]. There are four possible clinical treatments: medication, chemotherapy, surgery or both surgery and chemotherapy. The outcome is risk of the 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.

B. Datasets with Multiple Treatments

Here, we introduce a dataset as an example for those with multiple treatments. **NMES.** The National Medical Expenditures Survey (NMES) dataset is complied by [26]. 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.

C. Pros and Cons, What is Missing?

Pros. Most of the existing datasets are collected based on interesting treatment effect estimation problems. For example, the Jobs dataset is collected to answer to the question: does job training help people to get employed? Studying these datasets can provide insights for decision making in real-world scenarios. For example, an individual can rely on the inferred individual treatment effect based on her features to decide whether it is worth participating the job training program.

Cons. It is often notoriously difficult to collect data with ground truth for more than one potential outcomes, especially for applications with multiple treatments or continuous treatment. First, in a vast majority of problems, we are not able

[1]https://github.com/vdorie/npci

[2]downloaded from the UCI repository [23]

to obtain the counterfactual outcomes – the outcomes could have been observed iff an other treatment had been assigned. Therefore, researchers heavily rely on semi-synthetic datasets, where treatments and outcomes have to be synthesized based on strong assumptions of the data-generating process. This can become a time and labor consuming task in terms of the development of reasonable data simulation models.

What is Missing? In terms of the real-world applications, the research area is still in its early stage to obtain datasets that can be used to study interesting causal effects in many tasks because of the unavailability of counterfactual outcomes. For example, it is not difficult to obtain climate data from Google earth engine and user behavior data from Twitter. So it is painless to develop predictive models to predict user behavior from climate statistics. But for causal effect estimation, to understand how climate influences user behavior, we need data from the same user, collected in the exactly same situation except with different climate, which is literally not possible.

In terms of estimating average treatment effects, the challenges arise from how to design cheap, easy-to-implement, reliable and ethical experiments. When it comes to the research problem of reducing the sample size and time needed for a statistically significant randomized trial, its importance is still underestimated in the data mining and machine learning community.

In addition, if we focus on what is missing from the existing datasetsincomplete sentence. One of the most important component is the missing underlying structure between instances. The potential types of structure include but are not limited to networks and temporal dependencies.

IV. Off-Policy Evaluation in Contextual Bandits

So far, we have discussed the details about the existing benchmark datasets for causal discovery and causal effect estimation. As we may observe that the core to answer a causal problem is knowing the unknown, i.e., the *counterfactuals*. But why is this important to ML and what are the resulting benefits? To answer these questions, we present two specific ML tasks where counterfactual reasoning is a natural and powerful tool to leverage - off-policy evaluation and recommender systems. We start with the first task in this section. The learning setting of off-policy evaluation we consider here is an *offline* variant of the *contextual* k-armed bandit problem[3]. Suppose an existing policy π_0 can choose actions based on item features and observes rewards (e.g., search engines, recommender systems). This process generates *log* data with the form (x, y, δ, p) where x is the context (feature vector), y is the selected action, p is the probability of y being chosen given x and δ denotes the reward/feedback received. The goal of off-policy evaluation is to exam if a new policy π will perform desirably on future observations using the log data from π_0. This research problem is interesting due to its ubiquity in real-world applications:

- **Medical Studies.** In a medical study, doctors usually have to determine which treatment (e.g., surgery, drug therapy) is assigned to each patient. It is then often necessary to answer counterfactual questions such as: *Would this patient lived longer had she been assigned an alternative treatment?* Using the off-policy evaluation described here, we may approximate the causal effects from off-line datasets [27] to help doctors make better decisions.
- **Content Recommendation.** Log data is ubiquitous and often generated from many internet sites that can recommend ads, webpages, or news based on user history/behavior, search engine queries and many other observable quantities, at little cost. Studies of the interaction logs of such systems can help exam if the new recommendation policies can improve user satisfaction.

Now, let us take a closer look on the definition of the problem and the standard formats of the log data. Consider the input $x \in \mathcal{X}$, the output prediction $y \in \mathcal{Y}$ and a hypothesis space $\mathcal{H}$ of *stochastic policies* [28], which is calculated from the observed data. A hypothesis $h(\mathcal{Y}|x) \in \mathcal{H}$ makes predictions by sampling $y \sim h(\mathcal{Y}|x)$. In an interactive system, we can observe the feedback $\delta : \mathcal{X} \times \mathcal{Y} \to \mathrm{R}$ for the y sampled from $h(x)$. For instance, in a recommender system, $\mathcal{X}$ can be 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, $\mathcal{X}$ describing the patients' attributes, $\mathcal{Y}$ is the set of received treatments. We then observe the outcome δ from the patients. A small δ indicates user/patient satisfaction with y for x, while large values indicate dissatisfaction. The expected loss/risk of a hypothesis $R(h)$ is defined as [28]

$$R(h) = \mathrm{E}_{x\sim Pr(\mathcal{X})}\mathrm{E}_{y\sim h(x)}[\delta(x, y)]. \tag{2}$$

The goal of is then given the 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, to minimize the risk.

Evaluation of the proposed policy is extremely hard due to the following challenges:

- Sample selection bias. The recommendations are based on user history, thus the predictions favored by the historical algorithm will be over-emphasized.
- Partial information. Feedback/outcome for other predictions cannot be observed.

Bandit Data Generation. Despite log data is ubiquitous in the real world, it is often hard to gather for scientific researchers. In search of alternatives, synthetic or semi-synthetic data is often generated for off-policy evaluation. Here, we present a widely used bandit data generating approach proposed in [29]. 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, that is the feedback $\delta(x_i, y_i)$ is defined such that $\delta(x_i, y_i^*) > \delta(x_i, \neg\, y_i)$ where

[3]There are other approaches such as regression and reinforcement learning, but we focus on contextual bandits setting for simplicity.

$\rightharpoonup y_i$ is any of the items/treatments other than y_i^*. Then we simulate a bandit feedback dataset from a logging policy h_0 by sampling $y_i \sim h_0(x_i)$ and collecting feedback $\triangle(y_i^*, y_i)$. h_0 can be as simple as the logistic regression and is often trained with a small portion (e.g. 5%) of the supervised training set. $\triangle(y^*, y)$ is then the Hamming loss or Jaccard index between the supervised label y^* and the sampled label y for input x. This completes the procedure of generating a bandit dataset $\mathcal{D} = \{[x_i, y_y, \delta(x_i, y_i), h_0(y_i|x_i)]\}_{i\in\{1,\dots,n\}}$.

One thing to note is that the propensity score is usually estimated from data directly, which may introduce undesired biases. A large-scale real-world dataset[4] that contains accurately logged propensities can be seen in [30].

Limitations. While this data generating method has been adopted by several notable work in off-policy evaluation in contextual bandits [28], [31], [32], there are 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. This might be detrimental to the patients' health due to the interactions between drugs. The estimation of propensity score function using a small portion of the supervised training set will not work either because the underlying mechanism of treatment selection in medical study is often not fully understood.
- The selection of the hypothesis h_0 seems trivial (e.g., logistic regression), but it largely decides the measured performance of the new policy. Imagine the worst case that all y sampled from h_0 happens to be y^*, then the generated bandit dataset is essentially not eligible to be used as the new policy would have "seen" the groundtruth, i.e., items that users click. Consequently, how much portion of the multi-label training set should be used to estimate h_0 and what is the desirable accuracy that h_0 should have? These questions need to be addressed and clarified further.
- The mismatch of synthesized data and the observed data from true environment is often unavoidable in practice, resulting in policies that do not generalize to the real environment [33].

V. Causal Inference for Recommendation

Another ML task where causal inference starts taking off recently is 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 how the recommender system selects and shows the products to its users. From the causal inference perspective, most current "de-biased" recommendation systems are modeled as finding 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 which encodes users' selection bias, that is, users do not consider each product independently at random.

Input data to learn a recommendation policy often 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 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 or not. A more generalizable policy indicates a less-biased recommender system. Next, we introduce several datasets that have been used in recent publications [34]–[36]. Based on the different data collection/generation mechanisms, we divide the data into three categories: data collected from randomized control trial, semi-simulated datasets and simulated datasets.

Randomize Control Trial

Yahoo-R3. Music ratings collected from Yahoo! Music services. This dataset contains ratings for 1000 songs collected from 15,400 users in two different ways. One of the sources consists 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 during normal use of Yahoo! Music services for each user and exactly ten ratings for randomly selected songs for each of the first 5400 users in the dataset. The dataset includes approximately 300,000 user-supplied ratings, and exactly 54,000 ratings for randomly selected songs [5].

Semi-simulated Datasets

1. *MovieLens10M* (with 71567 unique users and 10677 unique products), User-movie ratings collected from a movie recommendation service. The ratings are on a 1–5 scale [36]. The treatment is if a user has rated an item, the outcome is if rating is greater or equal to 3.
2. *Netflix* (with 480189 unique users and 17770 unique products), the treatment is if a user has rated an item, the outcome is if rating is greater or equal to 3.
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. This data contains information of which paper a user downloaded and which she only read the abstract. The treatment in this dataset is if a user has viewed the abstract of a paper, outcome is if she downloaded the paper.

Listed are part of standard datasets often used in the evaluation of a recommender system. To ensure the different

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

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

data distributions, a common approach is to create two training/validation/test splits from the standard datasets: regular (REG) and skewed (SKEW). The regular split is generated by randomly selecting the exposed items for each user into training/validation/test sets with proportions 70/20/10. 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. Consequently, the test set will have a different exposure distribution from the training and validation sets.

Simulated Datasets

Coat Shopping Dataset [34]. 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. Randomized Control Trial for a recommender system is not practical in the real-world setting as it randomly recommends songs and overlooks information that is important to characterize user preference. This may largely influences user experience and brings large damages to companies. Leveraging simulated/semi-simulated datasets to show the generalizability of a de-biased recommender system is indeed technically sound, but one may give a second thought about its motivation to apply causal inference in recommender systems. Humans are biased in nature and recommender systems should be able to capture different user preferences in order to make personalized recommendations. Therefore a de-biased/generalizable recommender system may not necessarily make better recommendations than a biased recommender system. Instead, it might be more important to investigate the causes that a recommendation system makes certain recommendations. Such causally interpretable systems can identify underlying causal relation between users and items that might be overlooked before. To this end, the missing part in the current datasets is the careful design of a set of treatments that describe the user characteristics, the features of recommendable items and the corresponding potential outcome. Then we can leverage causal-effect estimation or counterfactual reasoning to find out the personalized "causes" for each user clicking a specific item.

VI. Conclusions and Future work

In this paper, we provide a systematic view of common benchmark datasets used in *causal discovery* and *causal effect estimation*, and extensively analyze the **pros, cons** and **limitations** of these datasets. Given the recent success in ML and its natural connection to causal inference, we also investigate the datasets for two specific ML tasks where the marriage of causal inference and ML starts taking off. With the comprehensive summarization and hindsight of these datasets, we hope to provide easy access to researchers who are interested in causal learning and more importantly, to call for the contributions to creating/sharing new benchmark datasets which can certainly benefit us all.

References

[1] P. Spirtes, "Introduction to causal inference," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1643–1662, 2010.

[2] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.

[3] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *arXiv preprint arXiv:1809.09337*, 2018.

[4] T. C. Chalmers, H. Smith Jr, B. Blackburn, B. Silverman, B. Schroeder, D. Reitman, and A. Ambroz, "A method for assessing the quality of a randomized control trial," *Controlled clinical trials*, vol. 2, no. 1, pp. 31–49, 1981.

[5] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: methods and benchmarks," *CoRR*, vol. abs/1412.3773, 2014. [Online]. Available: http://arxiv.org/abs/1412.3773

[6] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005. [Online]. Available: https://science.sciencemag.org/content/308/5721/523

[7] J. Mitrovic, D. Sejdinovic, and Y. W. Teh, "Causal inference via kernel deviance measures," *CoRR*, vol. abs/1804.04622, 2018. [Online]. Available: http://arxiv.org/abs/1804.04622

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

[9] I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, "Design and analysis of the causation and prediction challenge," in *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, ser. Proceedings of Machine Learning Research, I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, Eds., vol. 3. Hong Kong: PMLR, 03–04 Jun 2008, pp. 1–33. [Online]. Available: http://proceedings.mlr.press/v3/guyon08a.html

[10] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*, 2016, pp. 3020–3029.

[11] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3076–3085.

[12] P. Schwab, L. Linhardt, and W. Karlen, "Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks," *arXiv preprint arXiv:1810.00656*, 2018.

[13] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Advances in Neural Information Processing Systems*, 2017, pp. 6446–6456.

[14] J. S. Neyman, "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*, vol. 10, pp. 1–51, 1923.

[15] D. B. Rubin, "Bayesian inference for causal effects: The role of randomization," *The Annals of statistics*, pp. 34–58, 1978.

[16] R. J. LaLonde, "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, pp. 604–620, 1986.

[17] J. A. Smith and P. E. Todd, "Does matching overcome lalonde's critique of nonexperimental estimators?" *Journal of econometrics*, vol. 125, no. 1-2, pp. 305–353, 2005.

[18] R. H. Dehejia and S. Wahba, "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and statistics*, vol. 84, no. 1, pp. 151–161, 2002.

[19] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

[20] G. J. Duncan, J. Brooks-Gunn, and P. K. Klebanov, "Economic deprivation and early childhood development," *Child development*, vol. 65, no. 2, pp. 296–318, 1994.

[21] P. R. Hahn, V. Dorie, and J. S. Murray, "Atlantic causal inference conference (acic) data analysis challenge 2017," Tech. rep, Tech. Rep., 2018.

[22] D. Almond, K. Y. Chay, and D. S. Lee, "The costs of low birth weight," *The Quarterly Journal of Economics*, vol. 120, no. 3, pp. 1031–1083, 2005.

[23] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[24] J. Yoon, J. Jordon, and M. van der Schaar, "Ganite: Estimation of individualized treatment effects using generative adversarial nets," 2018.

[25] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, p. 1113, 2013.

[26] D. Galagate, J. Schafer, and M. D. Galagate, "Package causaldrf," 2015.

[27] G. W. Imbens and D. B. Rubin, *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[28] A. Swaminathan and T. Joachims, "Counterfactual risk minimization: Learning from logged bandit feedback," in *International Conference on Machine Learning*, 2015, pp. 814–823.

[29] A. Beygelzimer and J. Langford, "The offset tree for learning with partial labels," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 129–138.

[30] D. Lefortier, A. Swaminathan, X. Gu, T. Joachims, and M. de Rijke, "Large-scale validation of counterfactual learning methods: A test-bed," *arXiv preprint arXiv:1612.00367*, 2016.

[31] A. Swaminathan and T. Joachims, "The self-normalized estimator for counterfactual learning," in *advances in neural information processing systems*, 2015, pp. 3231–3239.

[32] T. Joachims, A. Swaminathan, and M. de Rijke, "Deep learning with logged bandit feedback," 2018.

[33] N. Jiang and L. Li, "Doubly robust off-policy value evaluation for reinforcement learning," *arXiv preprint arXiv:1511.03722*, 2015.

[34] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims, "Recommendations as treatments: Debiasing learning and evaluation," *arXiv preprint arXiv:1602.05352*, 2016.

[35] D. Liang, L. Charlin, and D. Blei, "Causal inference for recommdendation," 2016.

[36] S. Bonner and F. Vasile, "Causal embeddings for recommendation," in *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 2018, pp. 104–112.