Supplementary Material

1 Additional Details on Software Data

We introduce a source of empirical data where interventions are possible: large-scale software systems. We performed experiments on three large computational systems: Postgres, the Java Development Kit, and HTTP processing. These systems have many desirable properties for the purposes of empirical evaluation: (1) They are pre-existing systems created by people other than the researchers for a purpose other than evaluating algorithms for causal discovery; (2) They produce non-deterministic experimental results due to latent variables and natural stochasticity; (3) System parameters provide natural treatment variables; and (4) Each experiment is recoverable, allowing the same experiment to be performed multiple times with different combinations of interventions.

Within each computational system, we measure three classes of variables: outcomes, treatments, and subject covariates. Here, outcomes are measurements of the result of a computational process, treatments correspond to system configurations and are selected such that they could plausibly induce changes in outcomes, and subject covariates logically exist prior to treatment and are invariant with respect to treatment. Using these variables, we can apply all combinations of treatments to all subjects, and we can use these results to estimate actual interventional distributions for the effects of each treatment variable on each outcome variable. We can also then sub-sample these experimental data sets in a manner which simulates observational bias to produce observational-style data sets, allowing us to evaluate an algorithm's performance on pseudo-observational data and evaluate it using actual interventional effects. These data sets will be made available after publication.

We had a number of goals in mind when gathering data from our real domains:

- **Causal Sufficiency:** The algorithms we studied require that no pair of variables in the model are both caused by a latent variable. We can guarantee this is true for pairs of treatments and outcomes (since treatments have no parents in the original data set), but needed to employ domain knowledge to limit sources of causal sufficiency violations with regard to other pairs of variables.
- **Acyclicity:** Each of the systems can be described by a "single-shot" computational process which starts and finishes without the possibility for feedback.
- **Instance Independence:** We took efforts to ensure that each execution of the computational process was independent of previous executions. In most cases, this required clearing caches and resetting other aspects of system state.
- **Plausible Dependence:** We selected variables that we believed would be causally related.

Each domain is characterized by three classes of variables: subject covariates, treatments, and outcomes. Under the factorial experiment design, outcomes were measured for every combination of subjects and treatments. This yields a data set with many records for the same subject, as in the example in Table 1. To permit greater opportunities for observational sampling, we performed multiple trials of each factorial experiment. Given the difficulty associated with modeling highly complicated outcomes such as runtime, we employed a normalization scheme for each data set, dividing outcome values by a "baseline" value—the median control-case outcome value. Thus, we ultimately recorded outcomes which represent a deviation from this baseline. In this regard, our experimental results resemble a within-subjects design Greenwald [1976], although without many of the pitfalls that plague experiments on humans, such as non-independence of outcome measurements. In the original data

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Subject ID	Covariate	Treatment	Outcome
1	A	0	1.33
1	A	1	0.96
2	B	0	1.89
2	B	1	0.54
3	A	0	1.02
3	A	1	0.99
4	A	0	1.35
4	A	1	1.12

Table 1: An Example of a Factorial Experiment with Four Subjects and a Binary Treatment

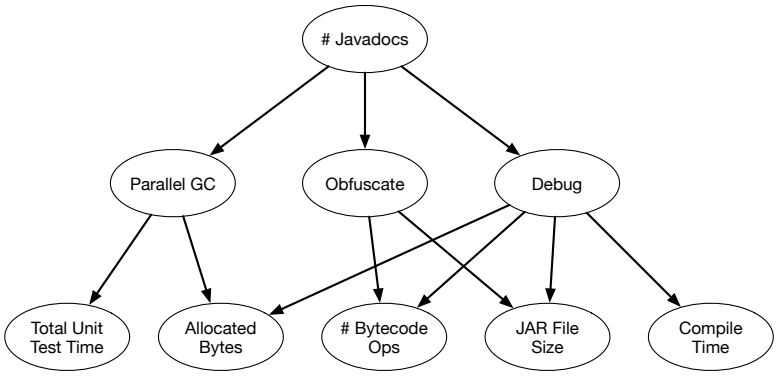

Figure 1: Consistent Model for the JDK Domain

from each domain, subject covariates are either discrete, continuous, or binary; treatments are binary; and outcomes are continuous. We converted each of the variables to a discrete representation to make parametrization and inference more robust.

1.1 Java Development Kit

Our experiments on the Java Development Kit (version `1.7.0_60`) used 2,500 Java projects obtained from GitHub as the subjects under study. We retrieved only projects which use the Maven build tool to facilitate automated compilation and execution. Additionally, we constrained our search to include only projects which had unit tests. This may introduce selection bias in our data collection processes, but this is acceptable. It is not important that our conclusions generalize to some population of computational systems, only that there are causal dependencies which hold on the sub-population under investigation. Of those, 473 compiled and ran without intervention. This group yielded a total of 7,568 subject-treatment combinations. For each combination, we compile and execute the unit tests of the Java project. In order to obtain full state recovery between each trial, any compiled project files were cleared between executions. Thirty-five CPU days were required to collect this data using several Amazon EC2 instances.

1.1.1 Treatments

- **Aggressive Compiler Optimization**: Disabling this option (enabled by default) prevents some compiler optimizations from running, potentially slowing down execution time but perhaps reducing compilation time. This option is disabled with the `javac` option `-XX:+AggressiveOpts`.
- **Emission of Debugging Symbols**: Debugging symbols are used to provide a map through the compiled source code that can be used for interactive debugging and diagnostics. Inclusion of these symbols may require some time during the compilation phase, increase the size of the compiled program, and could possibly impact runtime. This corresponds to the `-g` flag of `javac`.

- **Garbage Collection Methodology**: The Java Development Kit supports several garbage collection schemes. Two were considered: parallel and serial. These schemes are activated with the `-XX:-UseParallelGC` or `-XX:-UseSerialGC` arguments.
- **Code Obfuscation**: Several third-party tools are capable of obfuscating compiled code, making reverse-engineering difficult. This process could also affect the size of the compiled project files. The yGuard[1] tool was used for this purpose.

1.1.2 Outcomes

- **Number of Bytecode Instructions**: Before execution, Java code is compiled to an intermediate language referred to as bytecode. We measured the number of atomic instructions, or operations, in this compiled code to form this outcome using a custom-built bytecode analysis tool based on Javassist[2].
- **Total Unit Test Time**: Each project we gathered contains one or more unit tests. To capture the runtime of the full unit test workload, we computed the sum of runtimes of all unit tests for a given project.
- **Allocated Bytes**: The Java Virtual Machine supports a profiling option (`-agentlib:hprof=heap=sites`) which can be used to track heap statistics throughout a program's execution. We utilized this feature to obtain the total number of bytes allocated during unit test execution.
- **Compiled Code Size**: Java programs are often packaged in an format known as a JAR (Java ARchive). To characterize the size of the compiled code, we recorded the size in bytes of the associated JAR file.
- **Compilation Time**: In order to execute unit tests, the entire project needs to be compiled. This outcome represents the time used to convert all source files to their bytecode equivalents.

1.1.3 Subject Covariates

All subject covariates were obtained using the JavaNCSS tool[3].

- **# NCSS (non-comment source statements) in Project Source**: This covariate is highly predictive of compiled code size. Conceivably, in observational settings, large projects could also be associated with more liberal use of advanced compilation settings and tools, such as a code obfuscator.
- **# NCSS, Functions, and Classes in Unit Test Source**: These covariates are somewhat representative of the unit test workload. Projects with many lengthy unit tests may also have longer total unit test runtime.
- **# "Javadoc" comments in Unit Test Source**: This covariate could be indicative of code quality. Well-commented code is perhaps more likely to be found in high-quality projects. This code may be more likely to be used in production environments, and thus could be less likely to be observed with debugging symbols. This feature is used in the treatment-biasing procedure for construction of observational data sets.

1.2 Postgres

Consistent with a data warehousing scenario, we employ a fixed database for our Postgres (version 9.2.2) experiments: a sample of the data from Stack Overflow, drawn from the Stack Exchange Data Explorer[4]. The data explorer also houses many user-generated queries. We collected 29,375 of the most popular queries to use as subjects for this study. Stack Exchange's data warehouse uses Microsoft SQL Server, which does not completely overlap with Postgres in supported features and syntax. Some queries use only ANSI-compliant syntax and run successfully on either SQL Server or Postgres. To obtain as large a set of subjects as possible, we employed a semantics-preserving

[1] http://www.yworks.com/en/products_yguard_about.html
[2] http://www.csg.ci.i.u-tokyo.ac.jp/ chiba/javassist/
[3] http://javancss.codehaus.org/
[4] http://data.stackexchange.com/

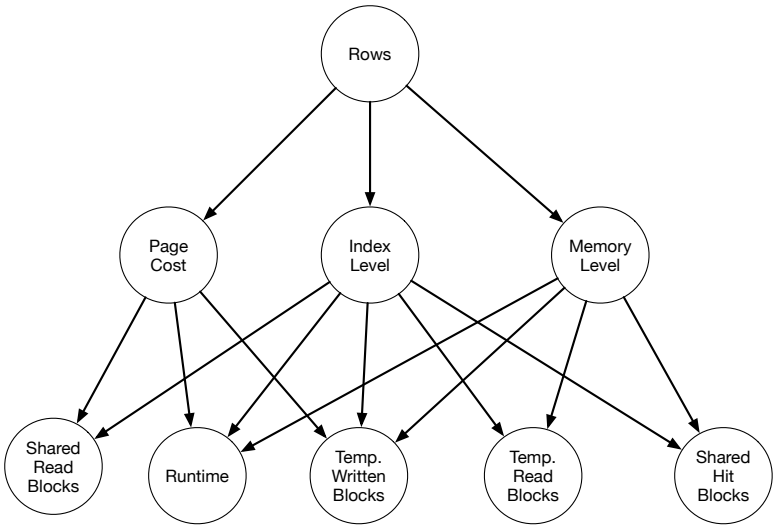

Figure 2: Consistent Model for the Postgres Domain

query rewriting scheme to adapt queries into Postgres-compliant syntax wherever possible. This yielded a set of 11,252 user-generated queries which executed successfully within Postgres for a total of 90,016 subject-treatment combinations. In order to recover system state between trials, the shared memory setting (specifying how much main memory Postgres can use for caching) was set to 128 kilobytes, limiting caching significantly. Any queries which required more than 30 seconds to execute were marked as "failures" in order to prevent long-running queries from holding up other queries, which typically required one second to execute. As with the JDK data set, this may induce sampling bias, but we are not aiming for our experimental findings to generalize to the broader population of database queries.

1.2.1 Treatments

- **Indexing**: A common administration task is to identify indices that can be used to accelerate lookup of commonly-referenced columns with a particular value or falling within a range. For our experiments, we employed two indexing settings: no indexing, and indexing on primary key/foreign key fields. Domain knowledge suggests that that the latter approach would dramatically reduce runtime of some queries. In all cases, the default B-tree index was employed.
- **Page Cost Estimates**: In order to determine if an index should be used, the database employs estimates of the relative cost of sequentially accessing disk pages and randomly accessing disk pages. We utilized two extremes for this setting: one scheme in which random page access is estimated to be fast, relative to the sequential page access, and one scheme in which the opposite relation holds. The corresponding database settings we adjusted were `random_page_cost` and `seq_page_cost`.
- **Working Memory Allocation**: The database engine can make use of fast random-access memory, if available, to store intermediate query results. The amount of working memory that is allocated to the system can be controlled with a configuration option. For our investigation, we employed a low-memory setting and a high-memory setting, with background knowledge suggesting that the latter would result in faster-executing queries. This treatment was instrumented with the `work_mem` and `temp_buffers` options.

1.2.2 Outcomes

- **Blocks Read from Shared and Temporary Memory**: These two outcomes identify the number of blocks, or memory regions, that were read during query execution. Shared memory is persistent (disk) and is accessed during normal table-retrieval procedures. Temporary memory is volatile (main memory) and is used for staging ordering or joining operations.

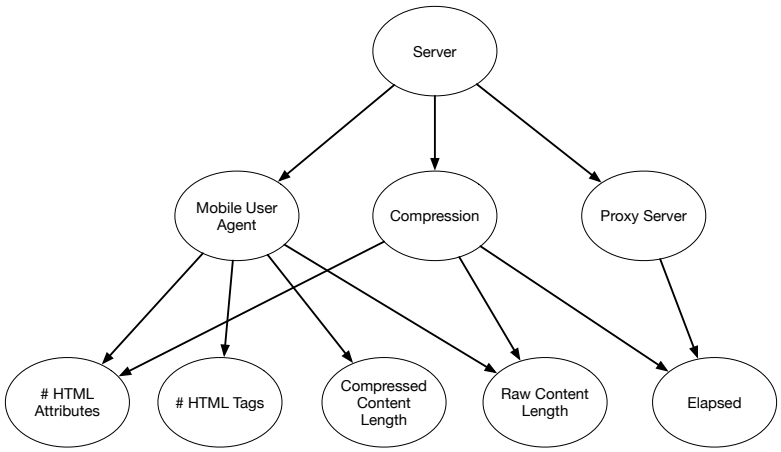

Figure 3: Consistent Model for the HTTP Domain

- **Blocks Hit in Shared Memory Cache**: This outcome represents the number of memory reads that were to be performed against shared memory, but were identified instead in a main memory cache.
- **Runtime**: The total time to execute the query.

1.2.3 Subject Covariates

- **Year of Query Creation**: The year that the query was entered on the Stack Exchange data explorer.
- **Number of Referenced Tables**: The number of distinct tables that are referenced in the query.
- **Total Number of Rows in Referenced Tables**: The sum of cardinalities of tables referenced in the query.
- **Number of Join Operators**: The number of join operators employed in the query, requiring merging data from two tables.
- **Number of Grouping Operators**: The number of grouping operators employed in the query, requiring reduction and possibly summarization of the data.
- **Number of Other Queries Created by the Same User**: The total number of queries that the Stack Exchange user has created.
- **Length of the Query in Characters**: The length of the query after application of relevant rewrite rules.
- **Number of Rows Retrieved**: The number of rows that are returned by the query. Logically, this value exists prior to application of any treatment and is invariant with respect to treatment (since the database is fixed), even though we can only measure it after query execution.

1.3 Hypertext Transfer Protocol

For our experiment on HTTP & networking infrastructure, we used requests to specific web sites as subjects. We identified a number of target sites through a breadth-first web crawl initiated at `dmoz.org`. We ended the crawl after retrieving 5,472 sites. For 4,350 of those sites, we were able to issue successful web requests with all combinations treatments, yielding 34,800 subject-treatment combinations. We employed numerous techniques to ensure that content would not be cached, which could induce carryover across treatment regimes.

1.3.1 Treatments

- **Use of a Mobile User Agent**: Web browsers supply a *user agent* to identify themselves to the web servers that they request pages from. Some sites have different versions for mobile applications. We artificially adjusted the user agent from a standard user agent to a mobile user agent to explore this phenomena. This is accomplished with the HTTP `User-Agent` header.
- **Proxy Server**: Web requests can be routed through a *proxy*, a server which issues web requests on behalf of a client. The additional time required to route the request to and from the proxy server can increase the elapsed time of the request. Our experiments were executed with Amazon EC2. Our "client" computers were making web requests from the east cost of the United States, and a proxy server was set up on the west coast.
- **Compression**: Applications can use the HTTP protocol to request that content be delivered with or without compression, possibly reducing the cross-network transmission time. In one compression configuration, the client requests `identity` compression, indicating that the content should be transmitted at face value. In another compression scheme, the client requests `gzip`, a common and effective scheme for HTTP content compression.

1.3.2 Outcomes

- **# of HTML Attributes and Tags**: These two outcomes describe the logical structure of the page. They may vary with respect to "mobile user agent".
- **Elapsed Time**: The time between issuance of the request and receipt of a response. This could be affected by network characteristics, which are determined in part by the time at which the request is issued and whether a proxy server is employed. Requests containing smaller payloads (influenced by compression) may also be faster to service.
- **Decompressed and Raw Content Length**: Two outcomes representing the size of a web page before and after content decompression, if applicable.

1.3.3 Subject Covariates

Only one subject covariate was identified for the HTTP domain, the web server reported via the `Server` header. This variable was coarsened into a version with 7 levels: Apache/2, Other Apache, Microsoft-IIS, nginx, Other, and Unknown.

2 Identifying Consistent DAGs

To identify DAGs that can consistently estimate the all interventional distributions $P(O|do(T))$, we need to ensure that (1) the parent set of T is a valid adjustment set with respect to O, and (2) if T has a causal effect on O, there is a chain connecting T and O in the DAG model. The first condition is straightforward to satisfy since we know the only parent of any treatment to be the covariate used to introduce observational bias. The second condition requires identification of which pairs of treatments and outcomes are causally related. These d-connection properties were identified for each domain using the full interventional data set using the Friedman test for blocked difference in means, allowing for correction of subject variability Friedman [1937]. An edge was introduced between any causally related pair to satisfy condition (2). Then, ground truth interventional distributions $P(O|do(T = t))$ were produced by applying the do-Calculus model adjustment rules, and answering probability queries $P(P|T = t)$ on the resulting model using belief propagation.

3 Pseudo-Observational Configurations

We can transform the factorial experiments on our real domains into pseudo-observational data by sub-sampling the experimental data in a way that is correlated with a "subject covariate". This mirrors the process of treatment self-selection common to observational data. This transformation is outlined in Algorithm 1.

Input: Interventional data set I, biasing strength $\beta \geq 0$, biasing covariate C
Output: Observationally biased data set O, $|O| = nd$
$l \leftarrow$ The number of distinct values of C
foreach *Subject* $e \in I$ **do**
  Let $C_e \in \{1..l\}$ represent the C value of subject e
  $Assign \leftarrow \{\}$
  foreach *Treatment* T_j **do**
    $s_{ej} \leftarrow \begin{cases} 1 & \text{if } C_e \times j \text{ is even} \\ -1 & \text{if } C_e \times j \text{ is odd} \end{cases}$
    $p \leftarrow \text{logit}^{-1}(s_{ej}\beta)$
    $t_j \leftarrow \text{Bernoulli}(p)$
    $Assign \leftarrow Assign \cup \{T_j = t_j\}$
  end
  $M \leftarrow$ Record in I corresponding to $(e, Assign)$
  $O \leftarrow O \cup M$
end

Algorithm 1: Logistic Sampling of Observational Data

4 Limitations of Empirical Data

In the paper, we discuss popular sources of empirical data that is suitable for evaluation. These data sets differ significantly in many ways, including level of realism and data quality, and they each have different benefits and limitations.

The cause-effect pairs challenge [Mooij et al., 2016] provides observational data on pairs of variables where the direction of causality is known from domain knowledge. This data set is useful for evaluating bivariate orientation algorithms, but the lack of any additional measured covariates limits its utility for evaluating multivariate structure learning algorithms.

The 2016 Atlantic Causal Inference Conference Competition data [Dorie et al., 2019] and the IBM Causal Inference Benchmarking Framework Shimoni et al. [2018] use covariates taken from a real-world data set, allowing for potentially complicated interactions between them. Treatment and outcome functions were then generated synthetically, using a variety of data generating processes to allow for the construction of many data sets with different features. This allows algorithms to be tested on many data sets, providing a more robust evaluation. However, the need to construct synthetic treatment and outcome functions limits the level of realism.

The software data we collected contains measurements of covariates, treatments, and outcomes from three real-world systems. While the treatment function is generated synthetically, the outcome function is not, lending the ground truth causal effects from treatment to outcome a high degree of realism. However, as with the above ACIC and IBM data sets, the treatment function still needs to be synthetically defined.

The flow cytometry data provided by Sachs et al. [2005] contains measurements of protein signaling pathways, where multiple activating and inhibitory interventions were performed. However, the ground truth is not clearly obtainable and most analysis using this dataset relies on structural measures.

Partially randomized experiments, where a population is split into randomized and an observational groups, are another useful source of empirical data [Shadish et al., 2008]. The collection of randomized data drawn from the same base population as observational data creates a convenient ground truth for causal effect estimation. However, due the nature of these experiments, they require careful experimental design to make sure the populations are equvalient and the treatments are correctly assigned and measured.

The DREAM in silico data sets [Schaffter et al., 2011] are taken from a sophisticated simulation derived from multiple known gene regulatory network structures, which, while non-empirical, is intended to be complex enough to approximate empirical data. However, realism is limited due to the use of a simulator.

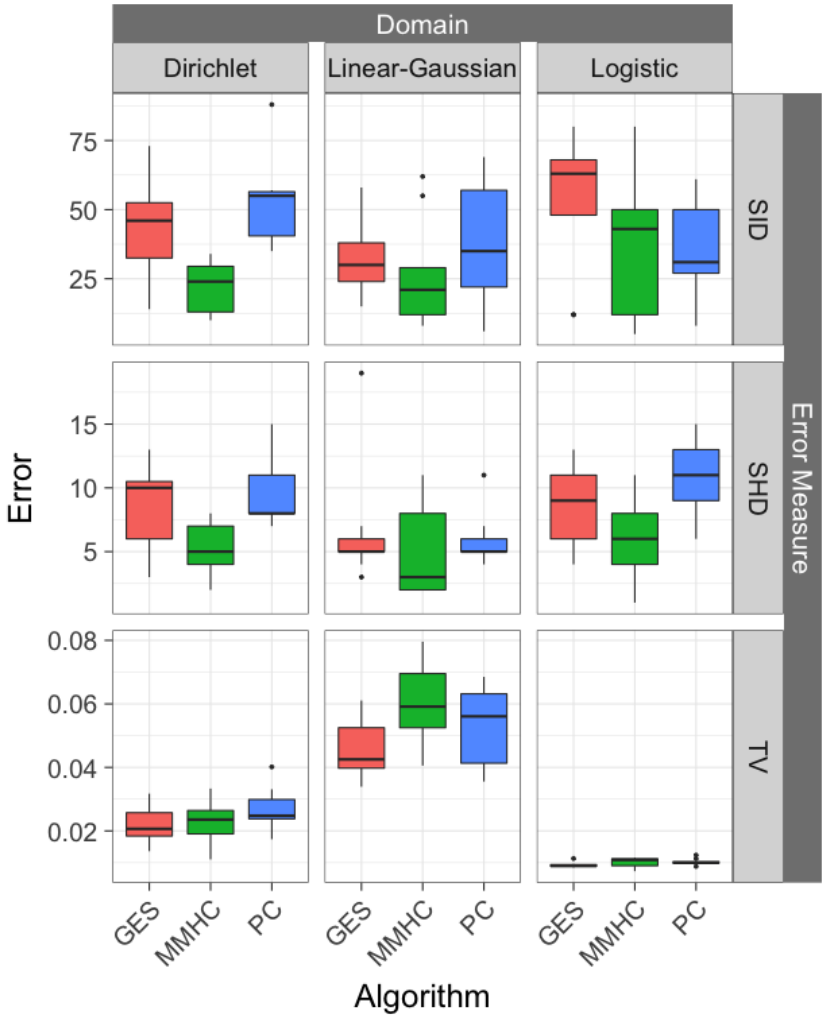

Figure 4: Relative Performance of Causal Discovery Algorithms on Synthetic Data Sets

5 Additional Experiments

In the paper, we provided experiments that demonstrate that TVD and structural measures provide different information and that information is relevant for over and under specification. To expand on these results, we performed an additional experiment to evaluate if different types of measures would lead to different conclusions about the relative performance of causal modeling algorithms. Figure 4 shows results on synthetic data that demonstrate that TVD does, inf act, imply a very different ordering of the relative performance of different learning algorithms than that implied by SHD and SID. We began by constructing 30 random DAGs with 14 variables and $E[N] = 2$. We generated parameters on those DAGs using each of the synthetic data techniques and sampled 5,000 data points from each DAG. Then, we applied PC, MMHC, and GES to the resulting data sets and measured the SID, SHD, and sum of pairwise total variations. As shown in Figure 4, some of the findings that would be reached with SID and SHD are not supported by a TV evaluation. The structural measures suggest that MMHC outperforms PC on the Dirichlet domain. However, the performance of the two algorithms is statistically indistinguishable as measured by TV. When measured with SID or SHD, GES does not outperform either MMHC or PC. However, GES is consistently the best performing algorithm in terms of interventional distribution accuracy.

Experiments in the paper demonstrate that TVD can, at least in some cases, provide information that structural measures cannot. However, that does not mean that the additional information is

Table 2: Metric Comparison on Real Domains with Over-specification and Under-specification

Domain	Subjects	Model Type	SID: Min, Median, Max			SHD: Min, Median, Max			TVD: Min, Median, Max		
JDK	473	Over-specify	0	0	0	1	3	3	0.04	0.17	0.21
		Under-specify	4	5	9	2	2	4	0.22	0.41	0.58
Postgres	5,000	Over-specify	0	0	0	0	1	2	0.00	0.06	0.09
		Under-specify	4	6	8	3	4	5	0.17	0.35	0.61
HTTP	2,599	Over-specify	0	0	0	1	2	4	0.06	0.06	0.09
		Under-specify	2	6	10	1	3	4	0.22	0.25	0.30

useful. To address this concern, we sought to measure how TVD responds to specific types of errors in learned structure. Specifically, we evaluate the effects of over-specification (extraneous edges) and under-specification (omitted edges) on model performance. We used our three empirical data sets drawn from large-scale computational systems (JDK, Postgres, and HTTP) to perform this analysis. From the original exhaustive experiments, we can identify which treatment-outcome pairs are causally related. We construct a partial DAG, consisting only of edges between treatment and outcome, by introducing an edge between each pair of causally related treatment and outcome. Then, a pseudo-observational data set can be constructed by sub-sampling treatment assignments according to a biasing covariate (details in Supplemental Materials). The resulting DAG model (illustrated for the JDK data set in Figure 1) consistently estimates distributions $P(O|do(T = t))$ for all treatment-outcome pairs.

We altered the consistent models of each data set to induce over-specification and under-specification. To quantify the effects of over-specification, we produced models in which one of the treatment variables had a directed edge into every outcome, regardless of the causal relationships in the true model. To quantify the effects of under-specification, we produced models in which one of the treatment variables had no outgoing edges. This process was repeated for each of our three domains and each treatment variable within that domain. For each model, a sum of pairwise total variations was computed as $\sum_{T,O} TV_{P,\hat{P},T=1}(O)$, where P represents the reference distribution given by the consistent model (as in Figure 1) and $\hat{P}$ represents the distribution induced by the altered model. A comparison of TVD, SHD, and SID on these experiments is shown in Table 2.

Two properties are apparent. First, over-specification is penalized differently by different evaluation measures. For small data sets, such as the JDK domain, over-specified models have zero SID but significant TVD values due to loss of statistical efficiency. Second, penalizing over-specification and under-specification with equal cost, as in SHD, is inconsistent with interventional distribution quality. In these domains, model under-specification has 2-5x the distributional impact of under-specification as measured by total variation.

6 Additional Details on Presented Experiments

Figures 5 and 6 show the results of comparing synthetic and interventional measures on synthetic data for both MMHC and PC. (results for GES were presented in the paper) Interestingly, while the correlation between SID and SHD is relatively consistent for all three structure learning algorithms, the correlation between TVD and SHD varies substantially, from seemingly completely uncorrelated (GES) to very clearly correlated (PC). This suggests that, in some cases, structural measures can provide a decent proxy for interventional measures. However, it is unlikely that the researcher knows this to be the case ahead of time, and the comparative difference in TVD between the three algorithms suggests the value of using TVD when comparing multiple causal learning algorithms.

We also provide additional results for experiments discussed in the paper that created synthetic data sets by learning their structure from empirical data. While we reported results using GES and PC, here we show results for MMHC. Figure 7 shows the performance of three learning algorithms (GES, MMHC, and PC). MMHC was used to infer a causal model from empirical data, and that model was then used to generate the synthetic data. Compared with the results in the paper, the relative performance of different algorithms looks somewhat similar to the results using GES, though there are some differences (e.g., PC is clearly the worst on all data sets in Figure 7, while this is not the case for GES in Figure 1 in the paper).

Sample sizes for some of the software system data sets are small, so in Figure 7 and Figure 1 in the paper, we report results as distributions over 30 trials for each algorithm and data set.

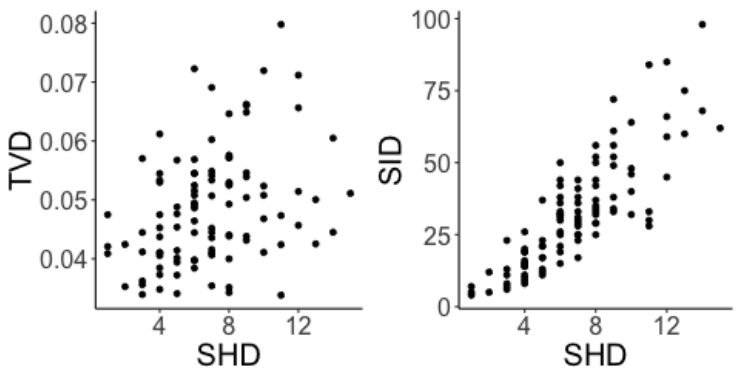

Figure 5: Structural and Interventional Measures Compared on Synthetic Data with MMHC.

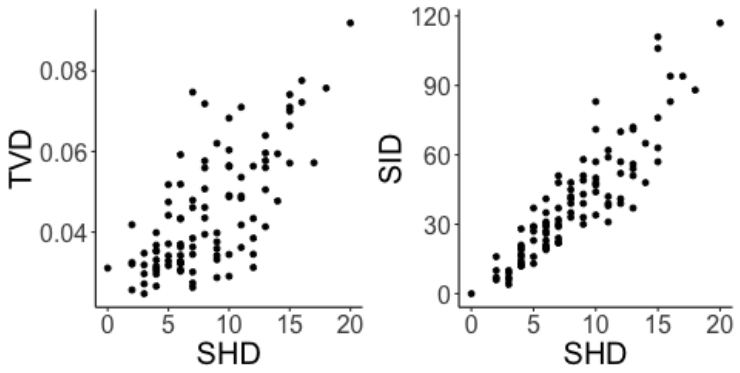

Figure 6: Structural and Interventional Measures Compared on Synthetic Data with PC.

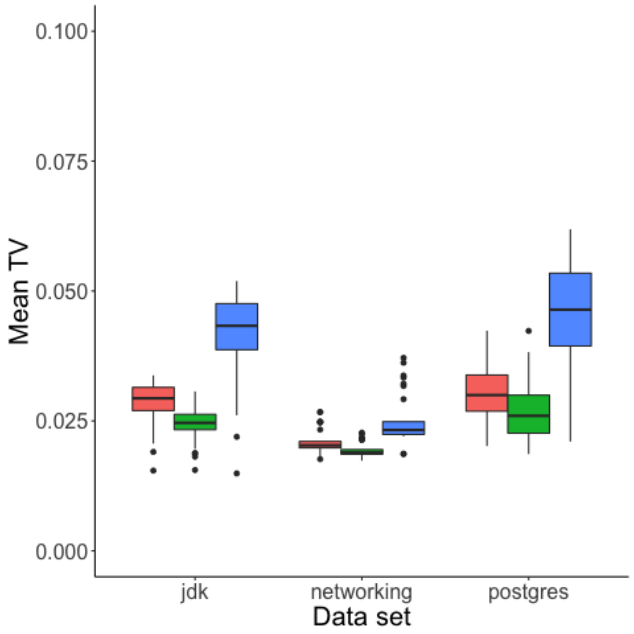

Figure 7: Results for MMHC for the Experiments Described in the Paper, using Synthetic Data that has been Created to Look like Empirical Data

References

Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate hawkes integrated cumulants. *Journal of Machine Learning Research*, 18(1):6998–7025, 2017.

Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *Advances in Neural Information Processing Systems*, pages 9469–9481, 2018.

Séverine Affeldt and Hervé Isambert. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In *Proceedings of the 31st international conference on Uncertainty in Artificial Intelligence*, pages 1–29, 2015.

Raj Agrawal, Tamara Broderick, and Caroline Uhler. Minimal i-map mcmc for scalable structure discovery in causal dag models. *International Conference on Machine Learning*, 2018.

Dalal Alrajeh, Hana Chockler, and Joseph Y Halpern. Combining experts causal judgments. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

Luca Ambrogioni, Max Hinne, Marcel Van Gerven, and Eric Maris. Gp cake: Effective brain connectivity with causal kernels. In *Advances in Neural Information Processing Systems*, pages 950–959, 2017.

David Arbour, Katerina Marazopoulou, and David Jensen. Inferring causal direction from relational data. In *Proceedings of the 32nd international conference on Uncertainty in Artificial Intelligence*, pages 12–21, 2016.

David T Arbour, Katerina Marazopoulou, Dan Garant, and David D Jensen. Propensity score matching for causal inference with relational data. In *Causality workshop at the 30th international conference on Uncertainty in Artificial Intelligence*, pages 25–34, 2014.

Angelos P. Armen and Robin J. Evans. Towards characterising bayesian network models under selection. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Nuaman Asbeh and Boaz Lerner. Pairwise cluster comparison for learning latent variable models. In *Causality workshop at the 32nd international conference on Uncertainty in Artificial Intelligence*, 2016.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3475–3481, 2015.

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 2410–2416, 2014.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, pages 1342–1350, 2015.

Kevin Bello and Jean Honorio. Computationally and statistically efficient learning of causal bayes nets using path queries. In *Advances in Neural Information Processing Systems*, pages 10954–10964, 2018.

Eli Ben-Michael and Avi Feller. Matrix constraints and multi-task learning for covariate balance. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Tineke Blom and Joris Mooij. Generalized structural causal models. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Tineke Blom, Anna Klimovskaia, Sara Magliacane, and Joris M. Mooij. Causal discovery in the presence of measurement error. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 570–579, 2018.

Stephan Bongers and Joris Mooij. Bridging the gap between random differential equations and structural causal models. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Giorgos Borboudakis and Ioannis Tsamardinos. Towards robust and versatile causal discovery for business applications. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 1435–1444, 2016.

Ruichu Cai, Jie Qiao, Kun Zhang, Zhenjie Zhang, and Zhifeng Hao. Causal discovery from discrete data using hidden compact representation. In *Advances in Neural Information Processing Systems*, pages 2671–2679, 2018a.

Ruichu Cai, Jie Qiao, Zhenjie Zhang, and Zhifeng Hao. Self: Structural equation likelihood framework for causal discovery. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018b.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In *Proceedings of the 31st international conference on Uncertainty in Artificial Intelligence*, pages 181–190, 2015.

Krzysztof Chalupka, Tobias Bischoff, Pietro Perona, and Frederick Eberhardt. Unsupervised discovery of el nino using causal feature learning on microlevel climate data. In *Proceedings of the 32nd international conference on Uncertainty in Artificial Intelligence*, pages 72–81, 2016.

Aditya Chaudhry, Pan Xu, and Quanquan Gu. Uncertainty assessment and false discovery rate control in high-dimensional granger causal inference. In *International Conference on Machine Learning*, pages 684–693, 2017.

Wei Cheng, Kai Zhang, Haifeng Chen, Guofei Jiang, Zhengzhang Chen, and Wei Wang. Ranking causal anomalies via temporal and dynamical analysis on vanishing correlations. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 805–814, 2016.

Belkacem Chikhaoui, Mauricio Chiazzaro, and Shengrui Wang. A new granger causal model for influence evolution in dynamic social networks: The case of dblp. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 51–57, 2015.

Juan D Correa and Elias Bareinboim. Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3740–3746, 2017.

Ruifei Cui, Perry Groot, Moritz Schauer, and Tom Heskes. Learning the causal structure of copula models with latent variables. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 188–197, 2018.

Vanessa Didelez. Causal reasoning for events in continuous time: a decisiontheoretic approach. In *Proceedings of the 31st international conference on Uncertainty in Artificial Intelligence*, pages 40–45, 2015.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34:43–68, 2019.

Robert W. Spekkens Elie Wolfe and Tobias Fritz. The inflation technique for causal inference with latent variables. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Flavio Figueiredo, Guilherme Resende Borges, Pedro O. S. Vaz de Melo, and Renato Assunção. Fast estimation of causal interactions using wold processes. In *Advances in Neural Information Processing Systems*, pages 2975–2986, 2018.

Patrick Forré and Joris M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 269–278, 2018.

Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

Tian Gao and Qiang Ji. Local causal discovery of direct causes and effects. In *Advances in Neural Information Processing Systems*, pages 2512–2520, 2015.

P. Geiger, D. Janzing, and B. Schölkopf. Estimating causal effects by bounding confounding. In *Proceedings of the 30th international conference on Uncertainty in Artificial Intelligence*, pages 240–249, Oregon, 2014.

Philipp Geiger, Kun Zhang, Bernhard Schoelkopf, Mingming Gong, and Dominik Janzing. Causal inference by identification of vector autoregressive processes with hidden components. In *International Conference on Machine Learning*, pages 1917–1925, 2015.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. *International Conference on Machine Learning*, 2017a.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Learning causal structures using regression invariance. In *Advances in Neural Information Processing Systems*, pages 3011–3021, 2017b.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. In *Advances in Neural Information Processing Systems*, pages 6269–6279, 2018.

Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *International Conference on Machine Learning*, pages 1898–1906, 2015.

Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. Causal discovery from temporally aggregated time series. In *Proceedings of the 33rd international conference on Uncertainty in Artificial Intelligence*, volume 2017, 2017.

Anthony G Greenwald. Within-subjects designs: To use or not to use? *Psychological Bulletin*, 83 (2):314, 1976.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. Generating event causality hypotheses through semantic relations. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2396–2403, 2015.

Daniel N Hill, Robert Moakler, Alan E Hubbard, Vadim Tsemekhman, Foster Provost, and Kiril Tsemekhman. Measuring causal impact of online actions via natural experiments: Application to display advertising. In *Proceedings of the 21th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 1839–1847, 2015.

Huining Hu, Zhentao Li, and Adrian R Vetta. Randomized experimental design for causal graph discovery. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2014.

Shoubo Hu, Zhitang Chen, Vahid Partovi Nia, Lai-Wan Chan, and Yanhui Geng. Causal inference and mechanism clustering of A mixture of additive noise models. In *Advances in Neural Information Processing Systems*, pages 5212–5222, 2018.

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 340–349, 2014.

Mohammad T Irfan and Luis E Ortiz. Causal strategic inference in networked microfinance economies. In *Advances in Neural Information Processing Systems*, pages 1161–1169, 2014.

Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under markov equivalence. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 978–987, 2018.

Dominik Janzing and Bernhard Schoelkopf. Detecting non-causal artifacts in multivariate linear regression models. *International Conference on Machine Learning*, 2018.

Mohammad Javidian and Marco Valtorta. Finding minimal separators in ancestral graphs. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Nathan Kallus. Causal inference by minimizing the dual norm of bias: Kernel matching & weighting estimators for causal effects. In *Causality workshop at the 32nd international conference on Uncertainty in Artificial Intelligence*, 2016.

Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6921–6932, 2018.

Ken Kansky, Tom Silver, David A Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In *International Conference on Machine Learning*, pages 1809–1818, 2017.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

Murat Kocaoglu, Alexandros G Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *Proceedings of the 31nd AAAI Conference on Artificial Intelligence*, pages 1156–1162, 2017a.

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems*, pages 7018–7028, 2017b.

Samory Kpotufe, Eleni Sgouritsa, Dominik Janzing, and Bernhard Schölkopf. Consistency of causal inference under the additive noise model. In *International Conference on Machine Learning*, pages 478–486, 2014.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3466–3473, 2017.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 1655–1664, 2016.

Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart J. Russell, and Pieter Abbeel. Learning plannable representations with causal infogan. In *Advances in Neural Information Processing Systems*, pages 8747–8758, 2018.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 1181–1189, 2016.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Maximum causal tsallis entropy imitation learning. In *Advances in Neural Information Processing Systems*, pages 4408–4418, 2018.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2573–2583, 2018.

Sanghack Lee and Vasant Honavar. A characterization of markov equivalence classes of relational causal models under path semantics. In *Proceedings of the 32nd international conference on Uncertainty in Artificial Intelligence*, pages 387–396, 2016a.

Sanghack Lee and Vasant Honavar. On learning causal models from relational data. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 3263–3270, 2016b.

Erik M. Lindgren, Murat Kocaoglu, Alexandros G. Dimakis, and Sriram Vishwanath. Experimental design for cost-aware learning of causal graphs. In *Advances in Neural Information Processing Systems*, pages 5284–5294, 2018.

David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

Sara Magliacane, Tom Claassen, and Joris M Mooij. Ancestral causal inference. In *Advances in Neural Information Processing Systems*, pages 4466–4474, 2016.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, pages 10869–10879, 2018.

Molly Lucas Manjari Narayan and Amit Etkin. Learning time-varying bi-variate causal structure using interventional neuroimaging. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Katerina Marazopoulou, Marc Maier, and David Jensen. Learning the structure of causal models with relational and temporal dependence. In *Proceedings of the 31st international conference on Uncertainty in Artificial Intelligence*, pages 66–75, 2015.

Christopher Meek. Toward learning graphical and causal process models. In *Causality workshop at the 30th international conference on Uncertainty in Artificial Intelligence*, page 43, 2014.

Christopher A Merck and Samantha Kleinberg. Causal explanation under indeterminism: A sampling approach. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 1037–1043, 2016.

Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. In *Advances in Neural Information Processing Systems*, pages 6986–6994, 2018.

Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 350–360, 2018.

Joris M. Mooij and Jerome Cremers. An empirical study of one of the simplest causal prediction algorithms. In *Causality workshop at the 31st international conference on Uncertainty in Artificial Intelligence*, 2015.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Razieh Nabi and Ilya Shpitser. Semi-parametric causal sufficient dimension reduction of high dimensional treatment. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Razieh Nabi, Phyllis Kanki, and Ilya Shpitser. Estimation of personalized effects associated with causal pathways. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 673–682, 2018.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Jose M Peña. Alternative markov and causal properties for acyclic directed mixed graphs. In *Proceedings of the 32nd international conference on Uncertainty in Artificial Intelligence*, pages 577–586, 2016.

Jose M. Pea. Identifiability of gaussian structural equation models with dependent errors having equal variances. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Sergey Plis, David Danks, Cynthia Freeman, and Vince Calhoun. Rate-agnostic (causal) structure learning. In *Advances in Neural Information Processing Systems*, pages 3303–3311, 2015.

Guillaume Pouliot. Modern methods for spatial econometrics. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Daniel Malinsky Razieh Nabi and Ilya Shpitser. Learning optimal fair policies. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, pages 1513–1521, 2015.

Anna Roumpelaki, Giorgos Borboudakis, Sofia Triantafillou, and Ioannis Tsamardinos. Marginal causal consistency in constraint-based causal learning. In *Causality workshop at the 32nd international conference on Uncertainty in Artificial Intelligence*, 2016.

Paul K Rubenstein, Ilya Tolstikhin, Philipp Hennig, and Bernhard Schölkopf. Probabilistic active learning of functions in structural causal models. *arXiv preprint arXiv:1706.10234*, 2017.

Paul K. Rubenstein, Stephan Bongers, Joris M. Mooij, and Bernhard Schölkopf. From deterministic odes to dynamic structural causal models. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 114–123, 2018.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, April 2005.

Thomas Schaffter, Daniel Marbach, and Dario Floreano. GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16): 2263–2270, 2011.

Bernhard Schölkopf, David W Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Removing systematic errors for exoplanet search via latent causes. In *Proceedings of The 32nd international conference on Machine Learning*, volume 37 of *Journal of Machine Learning Research Workshop and Conference Proceedings*, page 22182226. JMLR, 2015.

William R. Shadish, M. H. Clark, and Peter M. Steiner. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1344, 2008.

Naji Shajarisales, Dominik Janzing, Bernhard Schölkopf, and Michel Besserve. Telling cause from effect in deterministic linear dynamical systems. In *International Conference on Machine Learning*, pages 285–294, 2015.

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, pages 3195–3203, 2015.

Shilpa Mahatma Italo Buleje Yanyan Han Sharon Hensley-Alford, Piyush Madan and Fang Lu. Effect of secular trend in drug effectiveness study in real world data. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Eli Sherman and Ilya Shpitser. Identification and estimation of causal effects from dependent data. In *Advances in Neural Information Processing Systems*, pages 9446–9457, 2018.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018.

Ilya Shpitser and Eli Sherman. Identification of personalized effects associated with causal pathways. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 530–539, 2018.

Ricardo Silva and Robin Evans. Causal inference through a witness protection program. In *Advances in Neural Information Processing Systems*, pages 298–306, 2014.

Hossein Soleimani, Adrash Subbaswamy, and Suchi Saria. Treatment-response models for counterfactual reasoning with countinuous-time, continuous-valued interventions. In *Proceedings of the 33rd international conference on Uncertainty in Artificial Intelligence*, volume 2017, 2017.

Andrew Stanton, Amanda Thart, Ashish Jain, Priyank Vyas, Arpan Chatterjee, and Paulo Shakarian. Mining for causal relationships: A data-driven study of the islamic state. In *Proceedings of the 21th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 2137–2146, 2015.

Jon McAuliffe Steven Howard, Aaditya Ramdas and Jasjeet Sekhon. Uniform nonasymptotic confidence sequences for sequential treatment effect estimation. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, pages –, 2018.

Adarsh Subbaswamy and Suchi Saria. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 947–957, 2018.

Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. Causal inference via sparse additive models with application to online advertising. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 297–303, 2015.

Panagiotis Toulis and David C Parkes. Long-term causal effects via behavioral game theory. In *Advances in Neural Information Processing Systems*, pages 2604–2612, 2016.

Sofia Triantafillou. Score-based vs constraint-based causal learning in the presence of confounders. In *Causality workshop at the 32nd international conference on Uncertainty in Artificial Intelligence*, 2016.

Ran Wang. Identify heterogeneous effect and confounding effect via l_1-regularized soft decision tree. In *Causality workshop at the 34th international conference on Uncertainty in Artificial Intelligence*, 2018.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5822–5831, 2017.

Yuhao Wang, Chandler Squires, Anastasiya Belyaeva, and Caroline Uhler. Direct estimation of differences in causal graphs. In *Advances in Neural Information Processing Systems*, pages 3774–3785, 2018.

Yongkai Wu, Lu Zhang, and Xintao Wu. On discrimination discovery and removal in ranked data using causal graph. *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, 2018.

Chang Xu, Dacheng Tao, and Chao Xu. Large-margin multi-label causal feature learning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 1924–1930, 2015.

Hongteng Xu, Mehrdad Farajtabar, and Hongyuan Zha. Learning granger causality for hawkes processes. In *International Conference on Machine Learning*, pages 1717–1726, 2016.

Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. Causal bandits with propagating inference. *International Conference on Machine Learning*, 2018.

Karren D Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. *International Conference on Machine Learning*, 2018.

Jinyoung Yeo, Geungyu Wang, Hyunsouk Cho, Seungtaek Choi, and Seung-won Hwang. Machine-translated knowledge transfer for commonsense causal reasoning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

Zhalama, Jiji Zhang, Frederick Eberhardt, and Wolfgang Mayer. Sat-based causal discovery under weaker assumptions. In *Proceedings of the 33rd international conference on Uncertainty in Artificial Intelligence*, volume 2017, 2017.

Hao Zhang, Shuigeng Zhou, Kun Zhang, and Jihong Guan. Causal discovery using regression-based conditional independence tests. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 1250–1256, 2017a.

Hao Zhang, Shuigeng Zhou, and Jihong Guan. Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018a.

Junzhe Zhang and Elias Bareinboim. Equality of opportunity in classification: A causal approach. In *Advances in Neural Information Processing Systems*, pages 3675–3685, 2018a.

Junzhe Zhang and Elias Bareinboim. Non-parametric path analysis in structural causal models. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 653–662, 2018b.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making–the causal explanation formula. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018c.

Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 3150–3157, 2015.

Kun Zhang, Jiji Zhang, Biwei Huang, Bernhard Schölkopf, and Clark Glymour. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *Proceedings of the 32nd international conference on Uncertainty in Artificial Intelligence*, pages 825–834, 2016.

Kun Zhang, Mingming Gong, Joseph Ramsey, Kayhan Batmanghelich, Peter Spirtes, and Clark Glymour. Causal discovery in the presence of measurement error: Identifiability conditions. *arXiv preprint arXiv:1706.03768*, 2017b.

Kun Zhang, Mingming Gong, Joseph Ramsey, Kayhan Batmanghelich, Peter Spirtes, and Clark Glymour. Causal discovery with linear non-gaussian models under measurement error: Structural identifiability results. In *Proceedings of the 34th International Conference on Uncertainty in Artificial Intelligence*, pages 1063–1072, 2018b.

Yuxun Zhou and Costas J Spanos. Causal meets submodular: Subset selection with directed information. In *Advances in Neural Information Processing Systems*, pages 2649–2657, 2016.