Computational and Mathematical Organization Theory
https://doi.org/10.1007/s10588-021-09351-y

S.I. : GROUND TRUTH: IN SILICO SOCIAL SCIENCE (GTIS3)

Explaining and predicting human behavior and social dynamics in simulated virtual worlds: reproducibility, generalizability, and robustness of causal discovery methods

Svitlana Volkova, et al. ***[full author details at the end of the article]***

Accepted: 23 October 2021

Abstract
Ground Truth program was designed to evaluate social science modeling approaches using simulation test beds with ground truth intentionally and systematically embedded to understand and model complex Human Domain systems and their dynamics Lazer et al. (Science 369:1060–1062, 2020). Our multidisciplinary team of data scientists, statisticians, experts in Artificial Intelligence (AI) and visual analytics had a unique role on the program to investigate accuracy, reproducibility, generalizability, and robustness of the state-of-the-art (SOTA) causal structure learning approaches applied to fully observed and sampled simulated data across virtual worlds. In addition, we analyzed the feasibility of using machine learning models to predict future social behavior with and without causal knowledge explicitly embedded. In this paper, we first present our causal modeling approach to discover the causal structure of four virtual worlds produced by the simulation teams—Urban Life, Financial Governance, Disaster and Geopolitical Conflict. Our approach adapts the state-of-the-art causal discovery (including ensemble models), machine learning, data analytics, and visualization techniques to allow a human-machine team to reverse-engineer the true causal relations from sampled and fully observed data. We next present our reproducibility analysis of two research methods team's performance using a range of causal discovery models applied to both sampled and fully observed data, and analyze their effectiveness and limitations. We further investigate the generalizability and robustness to sampling of the SOTA causal discovery approaches on additional simulated datasets with known ground truth. Our results reveal the limitations of existing causal modeling approaches when applied to large-scale, noisy, high-dimensional data with unobserved variables and unknown relationships between them. We show that the SOTA causal models explored in our experiments are not designed to take advantage from vasts amounts of data and have difficulty recovering ground truth when latent confounders are present; they do not generalize well across simulation scenarios and are not robust to sampling; they are vulnerable to data and modeling assumptions, and therefore, the results are hard to reproduce. Finally, when we outline lessons learned and provide recommendations to improve models for causal discovery and prediction of human social behavior from

Published online: 18 November 2021

observational data, we highlight the importance of learning data to knowledge representations or transformations to improve causal discovery and describe the benefit of causal feature selection for predictive and prescriptive modeling.

Keywords Causal discovery · Causal structure learning · Ensemble models · Reproducibility · Generalizability · Robustness · Predictive modeling · Machine learning · Data science

1 Introduction

The ability to learn causal relationships from observational data is considered a significant component of human-level intelligence and can serve as one of the foundations of artificial intelligence (AI) (Bengio 2019; Chollet 2020; Pearl 2019; Lake et al. 2017). Understanding how latent properties of the data, including various sources of bias effect causal discovery accuracy, generalizability (Yarkoni 2019), reproducibility (Munafò et al. 2017), and robustness (Kummerfeld and Rix 2019; Olteanu et al. 2019) is essential to make progress and improve the existing approaches for causal discovery across many domains such as earth sciences, biology, economy (Runge et al. 2019; Glymour et al. 2019; Athey 2015), and social sciences (Lazer et al. 2020; Watts et al. 2018; Hofman et al. 2017).

Many different algorithms for causal discovery (aka causal structure learning) have been developed over the last twenty years (Guo et al. 2020; Pearl 2009). Existing approaches broadly fall into two categories: constraint-based (Spirtes et al. 2000; Yu et al. 2016 and score-based Chickering 2002).

- *Constraint-based methods* subject causal relationships to a set of constraints, for example conditional dependencies among the variables.
- *Score-based methods* discover causal relationships by optimizing a scoring function.

While each causal structure learning algorithm often relies on assumptions about the data generation process and underlying causal structure Greenland and Mansournia 2015 as shown in Table 1, it cannot be known from the data alone whether these assumptions are satisfied. Some causal discovery methods may tend to perform better on data from specific domains with different complexity and data generated from certain types of causal graph structures (e.g., sparser graphs) but such properties are obviously unknown a priori. Therefore, given a large number of possible causal modeling approaches, it is not clear which one to use in any given situation and whether a single approach will generalize across datasets and tasks with different complexities (Yarkoni 2019), which is especially important for the Human Domain Lazer et al. (2020). It is also important to investigate the relationship between causal model accuracy and robustness to sampling, and study the reproducibility of the SOTA causal modeling techniques (Stodden et al. 2016).

Table 1 Table of assumptions defined in Greenland and Mansournia (2015) for example causal discovery algorithms ordered by score-based versus constraint-based approaches

	Algorithm	Causal suffi-ciency	Causal faithful-ness	Causal markov	Gaussian data	Non-gauss. data	Multi-nomial data	Linear relations.
Score	CAM	×		×				
	CCDr	×		×	×			×
	GES	×	×	×	∘		∘	
	LiNGAM	×		×		×		×
Constr.	GS	×	×	×	∘		∘	
	MMPC	×	×	×	∘		∘	
	PC	×	×	×	∘		∘	
	IAMB	×	×	×	∘		∘	

×—indicates necessary assumption. ∘—indicates sufficient assumption

Our contributions to the Ground Truth program are presented below. We first outline our causal discovery workflow and discuss scenario-specific representation learning (node discovery) and modeling (link discovery) experimental decisions in Sect. 2. Then, acting as a reasonable upper bound, in addition to being a reproducibility control to other research methods teams' approaches, we recover the ground truth signal from the full simulation output to determine whether it was not possible to uncover the ground truth due to methodological failures or due to the absence of usable ground truth signal in the sampled simulation output and describe our findings in Sects. 2.3 and 2.4. Next we investigate robustness and generalizability of individual causal discovery algorithms and our causal ensemble approach on a range of simulated datasets (Saldanha et al. 2020) in additional to four virtual worlds produced by the simulation teams in Sect. 3. We further present and evaluate our predictive approach that takes advantage of machine learning and deep learning models to anticipate human behavior and social dynamics in the Human Domain using sampled data collected by research methods teams from four virtual worlds and additional simulated datasets produced by our team in Sect. 4. Finally, we conclude by summarizing our key results on reproducibility, generalizability, and robustness analysis of causal discovery approaches and data-driven research methods to explain, and predict human behavior and social dynamics in the Human Domain.

2 Causal discovery in the human domain: selected methods and limitations

This section presents our approach to causal structure learning of causal structure learning from fully observable and sampled data across four simulation scenarios provided under the GT program (Urban, Power, Disaster, and Conflict), that served as proxies for the real world. Our main objective for the causal structure learning (aka the explain task) was to analyze the existing causal discovery

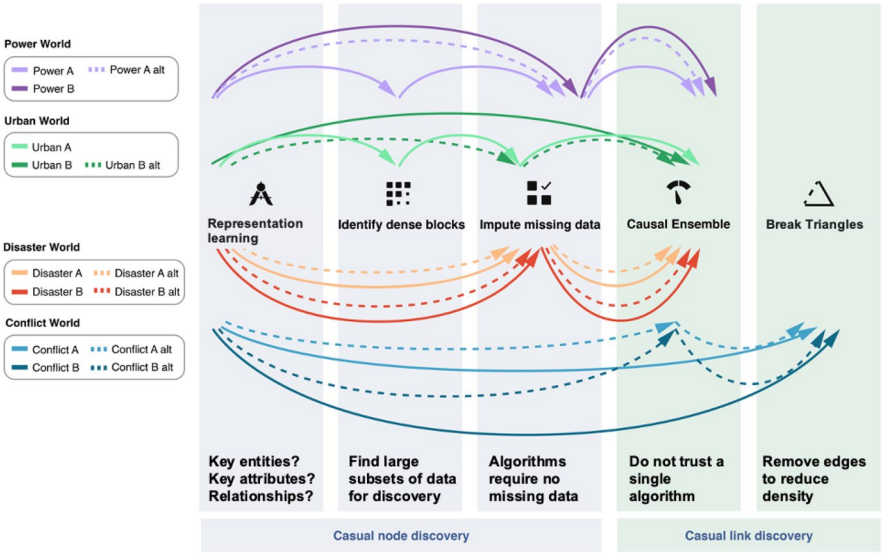


Fig. 1 Causal discovery workflow for four simulated virtual worlds (as defined in details in other chapters) when we rely on sampled data collected by two research methods teams (A and B). Alternative (Alt) experiments were performed on sampled data after we performed causal discovery on the full dataset to measure the effect of additional variables and modeling assumptions on causal discovery performance (RQ1)

approaches' limitations when applied to large-scale, noisy, high-dimensional data with unobserved variables (aka unknown unknowns), mixed data types, and unknown statistical dependencies between them that describe complex social dynamics. More specifically, we focused on answering research questions below.

RQ1: Is it possible to design generalizable workflows for causal discovery of complex social behavior and social dynamics *(generalizability analysis)*?

RQ2: Are other research method teams' reproducible using state-of-the-art causal discovery approaches when applied to the same sampled data *(reproducibility analysis)*?

RQ3: In case it is impossible to uncover the ground truth using sampled data, is it because of research method failures or simply the absence of usable ground truth signal in the sampled simulation output *(robustness analysis)*?

Figure 1 presents our causal discovery workflow with the human-in-the-loop evaluation (Cottam et al. 2021), taking specific steps for individual scenarios. For example, it shows that for the Urban scenario with the sampled data A we performed representation learning, dense block identification, and SOTA data imputation steps before applying our causal ensemble approach to the data. Note, SOTA imputation algorithms assume a Missing At Random mechanism which may bias downstream causal discovery. Thus, approaches like (Strobl 2019; Tu

et al. 2019; Gain and Shpitser 2018) are designed for causal discovery in the presence of missing not at random mechanisms.

2.1 Causal node discovery

As shown in the workflow diagram, causal node discovery steps focused on learning variable representations at multiple levels of granularity by performing data fusion, construct building (aka feature extraction), aggregation, data imputation, and normalization steps. For that we used a range of data science and statistical approaches including but not limited to regression, correlation analysis, statistical tests, social network analysis, data visualization, and machine learning.

The most time consuming step during causal node discovery was to understand the complexity of each scenario. Processing sampled (aka research request data A and B) was scenario-specific. Each scenario required learning *customized data representations* and perform *scenario-specific data manipulations* as reflected in the workflow diagram in Fig. 1.

In all scenarios we worked with missing and extremely sparse sampled data with limited temporal overlap across variables, for example data sparsity for samples A and B in the Urban scenario was 60% and 77%, and in the Power scenario 79% and 63%, respectively. *Data sparsity* and *the granularity of variable representations* could constrain causal discovery results.

However, our additional analysis of causal discovery performance and data sparsity demonstrated that the final results are not only constrained by sparsity. We observed no correlation between data sparsity and node discovery F1 score, but we found that lower density leads to higher edge F1 score. Thus, it is important to note that causal discovery performance also depends on scenario complexity, data size, and data quality—the presence of the signal in the data and feature representations (e.g., constructs), observed versus unobserved variables constructed by subject matter experts.

2.2 Causal link discovery

For causal link discovery, we developed an ensemble approach that combines several commonly used causal discovery approaches in order to produce one optimal causal link prediction model as presented in Fig. 2. The output of our causal ensemble pipeline is a causal model that formally consists of two sets of variables U (exogenous variables that are external to the model) and V (endogenous variables that are descendants of exogenous variables), and a set of functions f that assign each variable in V a value based on the values of the other variables in the model. To expand this definition: a variable X is a direct cause of a variable Y if X appears in the function that assigns Y value.

As expected, there was *no universal causal discovery model that generalized across all scenarios*, but some algorithms worked consistently (the algorithm finished running and returned a causal graph)—Greedy Equivalence Search (GES) and Max-Min Parents and Children (MMPC)—with full or sampled data A and

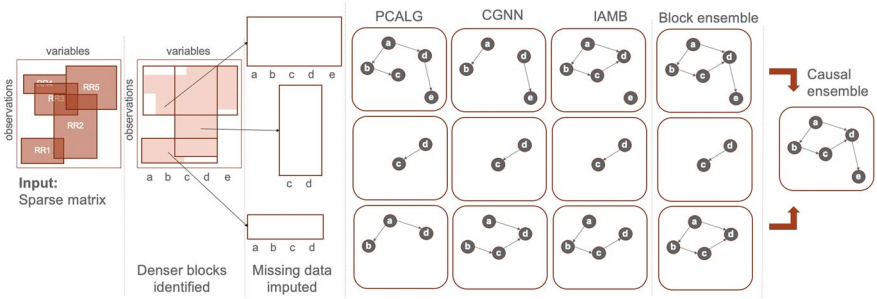


Fig. 2 Our ensemble approach to discover the causal structure of simulated human behavior and social dynamics from observational data (RQ1)

B as demonstrated in Table 2. We evaluated causal discovery ensemble performance using an in-house-developed the human-in-the-loop visual analytics tool (Cottam et al. 2021).

We observed that early assumptions (e.g., in the data fusion or representation learning steps) hurt the resulting causal discovery performance. Moreover, testing algorithm-specific data and modeling assumptions outlined in Table 1 was non-trivial and, sometimes, impossible.

2.3 Reproducibility of causal discovery

Figure 3 presents reproducibility analysis of causal discovery results with data samples A and B using our causal ensemble approach applied to the same research request data (aka data samples A and B). We observe that even when using the same sample data as other performers, changes in modeling assumptions and data manipulations created big discrepancies across inferred causal graphs. Our causal pipeline with the state-of-the-art causal discovery approaches was able to demonstrate improvement over TA2 results only in the Urban scenario in terms of node discovery F1 score, and in the Disaster and Power scenarios in terms of edge discovery F1 score. We can also see that it was more difficult to outperform causal discovery approaches applied to sampled data B than sampled data A.

It is important to note that our ability to discover nodes from sampled data was limited because our team did not collect sampled data compared to other teams, which in turn bounded downstream causal link discovery. Finally, our team made different data and modeling assumptions compared to other teams, for example our causal structure learning approach did not use any social theory and was data-driven which could explain our inability to fully reproduce other teams' causal discovery results across all simulated scenarios. Our modeling assumptions about how agents make decisions, interact with each other and with the environment, and the interaction between the environmental factors drove or constrained the final causal discovery performance.

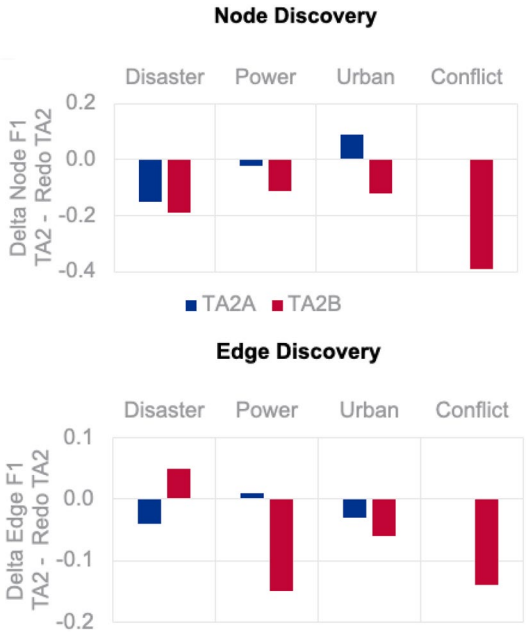


Fig. 3 Reproducibility analysis presented as differences in causal node and link discovery performance (measured as F1 score) between our causal ensemble approach and other causal discovery approaches using data samples A and B (RQ2)

Table 2 An overview of which causal discovery algorithms executed without errors and returned the causal graph when they were directly applied to sampled data (A and B collected by other performers) and full simulated data across four simulated worlds (RQ3)

	Urban			Power			Disaster			Conflict		
Algorithm	A	B	Full	A	B	Full	A	B	Full	A	B	Full
PC	+	−	+	+	−	+	+	+	+	+	+	+
MMPC	+	+	+	+	+	+	+	+	+	+	+	+
GS	−	+	+	+	−	−	+	+	−	+	+	+
IAMB	−	+	+	+	−	−	−	−	−	−	−	−
GES	+	+	+	+	+	+	+	+	+	+	+	+
GEIS	+	−	−	−	+	−	−	−	−	+	+	+
LiNGAM	−	−	−	−	−	+	−	−	+	−	−	−
CCDr	+	+	−	+	−	+	−	+	+	+	+	+

2.4 Causal discovery with sampled versus full data

Figure 4 presents causal discovery performance using our ensemble approach applied across four simulation scenarios with full versus sampled data. As expected, node discovery performance for the full data was much higher compared to sampled data (aka research request data sampled by teams A and B). Depending on the scenario, node discovery F1 score ranged between 0.13 and 0.53 for the sampled data and between 0.3 and 0.8 for the full data. Edge discovery F1 was significantly lower. The highest F1 of 0.3 was obtained for the Disaster scenario on both sampled and full data.

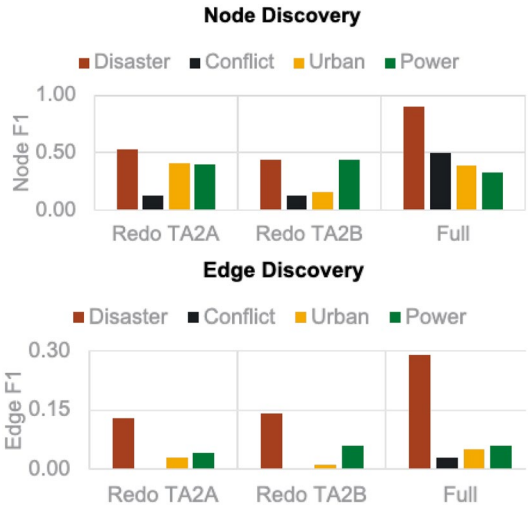


Fig. 4 Causal discovery results (measured as F1 score) across four simulated worlds using our causal ensemble approach on sampled (Redo TA2A and Redo TA2B) and fully observed data (RQ3)

Our full versus sampled data results further demonstrate that causal discovery is not about having lots of data like e.g., deep learning, it is about having the signal in the data, learning the right representations and encoding the complexity of the scenario. As we can see from Fig. 4, Urban scenario has 2TB of data, but causal discovery performance is much higher for the Disaster scenario with 300Mb of data.

Knowledge representations are important for both node and edge discovery with full or sampled data as shown in Fig. 4. Extracting knowledge from data through transformations (e.g., aggregation, construct building, fusion, imputation, and normalization) effects the final node discovery performance, which in turn effects edge discovery results. We found that the full data performance exceeds sampled data performance only for the Disaster scenario and it is equal for other scenarios. This could be explained by strategic and targeted sampling by subject matter experts from teams A and B during research request data collection.

Finally, our results demonstrate that SOTA causal discovery approaches are vulnerable to data and modeling assumptions. We found that only half of the algorithms worked per scenario as shown in Table 2. GES and MMPC were the most generalizable across four simulation scenarios, then Peter-Clark (PC), Concave penalized Coordinate Descent with reparameterization (CCDr), and Grow-Shrink (GS) approaches.[1]

[1] References to causal discovery approaches are provided at https://github.com/FenTechSolutions/CausalDiscoveryToolbox.

3 Robustness evaluation of causal discovery

In this section we perform additional analysis of causal discovery algorithm robustness—specifically robustness to sampling—which is extremely important in the real-world setting when it is not possible to observe the full data. We aim to answer two research questions below and present an extended analysis in Saldanha et al. (2020).

RQ5: How sensitive are the individual causal discovery algorithms and the ensemble approach to sampling in terms of variability of predictions?
RQ6: Does robustness depend on properties of the underlying causal graph or the observational data?

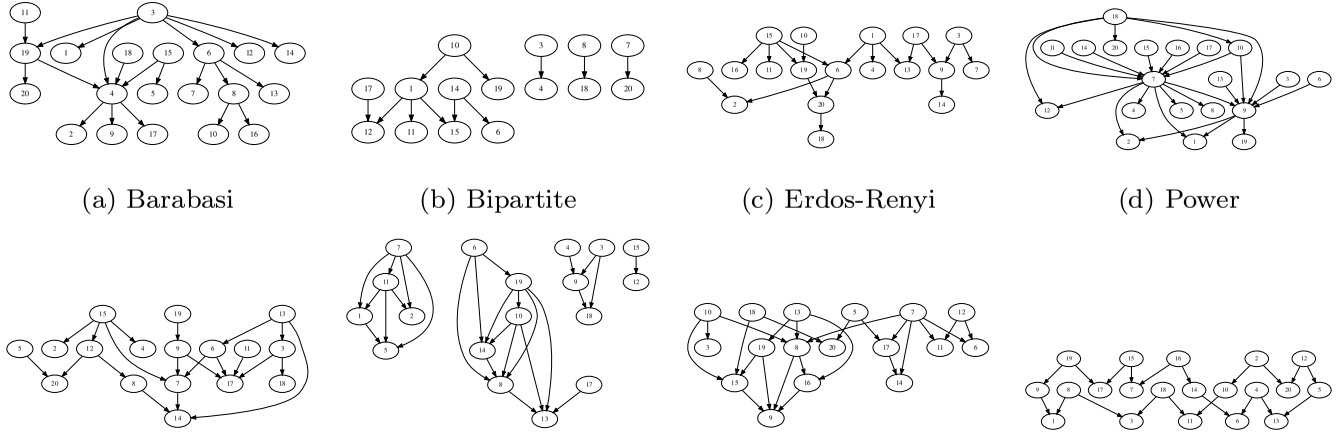
3.1 PCALG causal graphs

For our additional experiments, we generated 1140 random directed acyclic graphs (DAGs) with different properties using the randDAG function of the R pcalg library.[2] We used DAGs of size 20, 40, and 60 nodes, with 1 through 5 expected edges per node, and 8 different generation methods designed to target different graph topological properties. These generation methods were *regular*—a graph where every node has exactly *d* incident edges, *er*—an Erdos-Renyi graph where every edge is present independently, *watts*—an interpolation between regular graph and Erdoes-Renyi graph, *power*—a graph with power-law degree distribution, *bipartite*—a bipartite graph, *barabasi* —a graph with power-law degree distribution and preferential attachment, *geometric*—a geometric random graph, and *interEr*—a graph with two islands of Erdoes-Renyi graphs connected by a small number of edges. For each combination of DAG properties, we randomly generated 10 graphs. We used each generated DAG to simulate data that follows the given causal structure using linear Gaussian models with the edge weight and noise parameters drawn from uniform distributions. Example graphs can be seen in Fig. 5.

3.2 BNLEARN causal graphs

In addition to PCALG data, we leverage eight public datasets provided by the Bayesian Network Repository[3] to perform generalization tests of our results on datasets outside the Human Domain that have varied complexity, more data types, and different relationships between variable. The properties of the data are described in Table 3.

[2] https://www.rdocumentation.org/packages/pcalg/versions/2.6-8/topics/randDAG.
[3] https://www.bnlearn.com/bnrepository/.



(a) Barabasi (b) Bipartite (c) Erdos-Renyi (d) Power

(e) Watts (f) Geometric (g) Inter-Er (h) Regular

Fig. 5 Examples of causal graphs with $N = 20$ nodes and $E = 2$ expected edges per node, generated using each method of the PCALG random DAG function. We generated 1140 graphs total with $N = 20$, $N = 40$, $N = 60$ nodes and $E = 1$, $E = 2$, $E = 3$, $E = 4$, and $E = 5$ expected edges per node

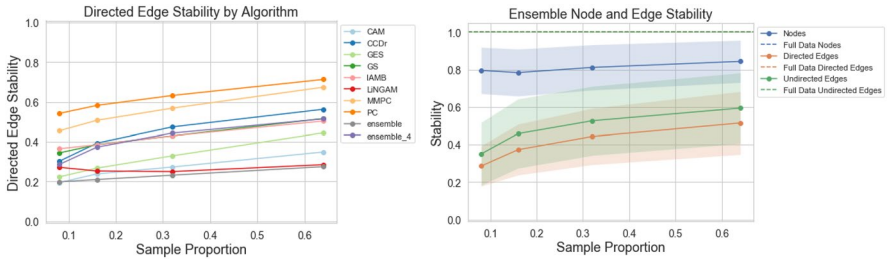


Fig. 6 Robustness of different approaches as a function of the fraction of the data sampled. (Left) The directed edge robustness of the individual algorithms and ensembles. (Right) The node, directed edge, and undirected edges robustness of the four-algorithm ensemble—GES, PC, GS, IAMB algorithms

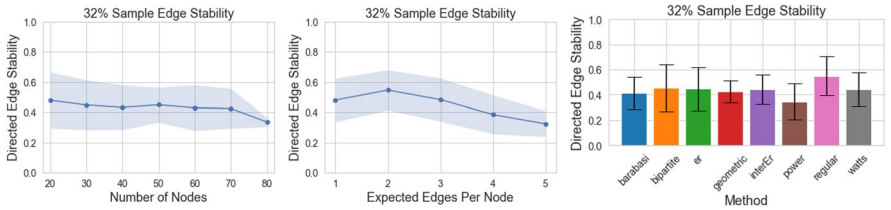


Fig. 7 The mean and standard deviation of the directed edge robustness of the four-algorithm ensemble with a 32% sample of the data across different graph structure properties including the number of nodes (left), the expected number of edges per node (middle), and the graph generation method (right)

Table 3 Properties of the BNLEARN datasets

Dataset	Nodes	Edges	Data type	Data size
Coronary	6	9	Binary	1841
Asia	8	8	Binary	5000
Sachs	11	17	Continuous	853
Child	20	25	Categorical	10,000
Insurance	27	52	Categorical	20,000
Alarm	37	46	Categorical	20,000
Water	32	66	Continuous	10,000
Andes	223	338	Binary	10,000

3.3 Robustness analysis

To measure robustness of the causal discovery approach, we repeated the causal discovery 10 times and calculated the average proportion of these repetitions that each node or edge is present. For example, if $A \rightarrow B$ appears in 8 out of 10 graphs, $A \rightarrow C$ appears in 6 out of 10, and $B \rightarrow D$ appears in 4 out of 10, the directed edge robustness of the graph would be 0.6. We evaluated the robustness of both directed edges, counting $A \rightarrow B$ different than $B \rightarrow A$, and undirected edges, where we evaluated robustness of the pairs of variables that are causally

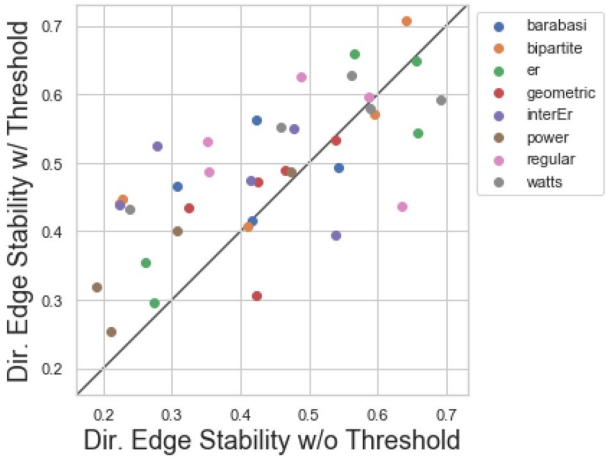


Fig. 8 The robustness of predicted edges when applying an edge weight threshold of 0.65 to the ensemble prediction versus without applying a threshold. Each point is an individual graph

related in either direction. We also evaluated node robustness because when certain edges fail to be discovered it can cause nodes to drop out.

We measure this robustness starting from a very small sample size of 8% of the data and double the sample size to 16%, 32%, and 64% to evaluate the sensitivity of the algorithm to the sampling proportion. Figure 6 (left) shows the robustness of directed edges for each algorithm and ensemble methods without edge weight thresholding as a function of the size of the data.

We find that the all-algorithm ensemble approach is less stable than each of the individual algorithms at each sample size. Ensembles with the top four performing algorithms have better robustness, but are still hindered by the least stable algorithms. This indicates that the algorithms are sensitive to data variability unless a very large fraction of the data is included.

In Fig. 6 (right) we examine the robustness of all graph components (nodes, undirected edges, and directed edges) of the top four ensemble method without edge weight thresholding as a function of sample size. As we increase with sample size, the robustness of ensemble algorithms also increases. With access to the full data sample (dashed lines), we find the four-algorithm ensemble to be highly stable across multiple runs.

In addition to studying the robustness across the full population of test datasets, we also explore whether the robustness varies based on how the graph structure was generated. In Fig. 7, we examine the directed edge stability for the 32% sample in comparison to several graph properties from all 10 runs of the PCALG data. For data generated from graphs with many nodes (e.g., 80 nodes), the robustness is lower on average than for smaller graphs with fewer nodes (e.g., 20 nodes). A similar trend exists when we examine the expected number of edges per node. We see increasingly more instability as the number of edges per node rises.

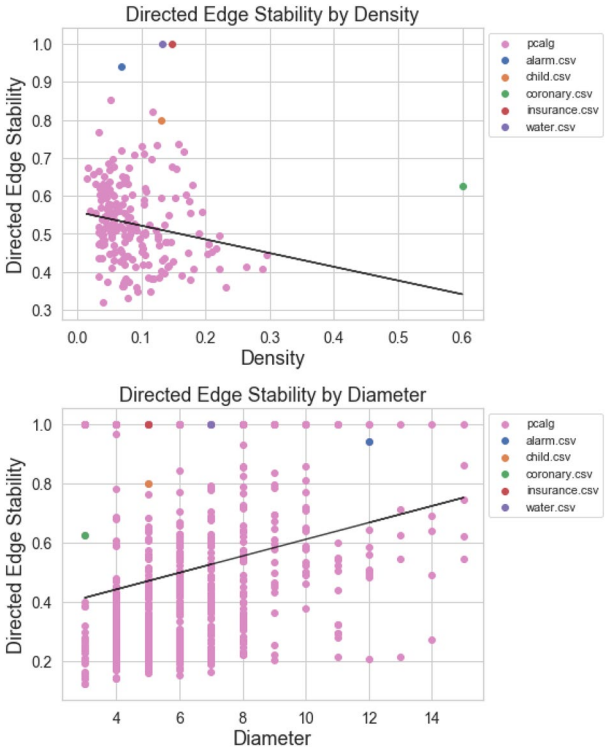


Fig. 9 Robustness of the four-algorithm ensemble as a function of two graph structure properties—the graph density (top) and the graph diameter (bottom). Each pink point is an individual PCALG graph, while other colors are the BNLEARN graphs. The line of best fit is plotted in black

Finally, we compare the directed edge robustness of data generated from each PCALG graph generation method. The most stable are *regular* graphs and the least stable are *power* graphs. These results are presented for the ensemble method with edge filtering, which may include some low-confidence edges. To study whether filtering to high-confidence edges impacts the robustness of the predictions, we compare the robustness with and without edge filtering for a subset of the 40-nodes graphs in Fig. 8. We find that filtering the edges increases the robustness of the predictions by about 6% on average.

3.4 Robustness and graph properties

We compare the performance of the four-algorithm ensemble across graphs with different structural properties. In Fig. 9, we show how robustness varies with the density and diameter of the ground truth causal graph. Because the causal graph may not be fully connected, we consider the largest diameter among the graph components rather than the diameter of the full graph. We find that *robustness decreases for denser graphs, while increasing for causal graphs with larger diameters*.

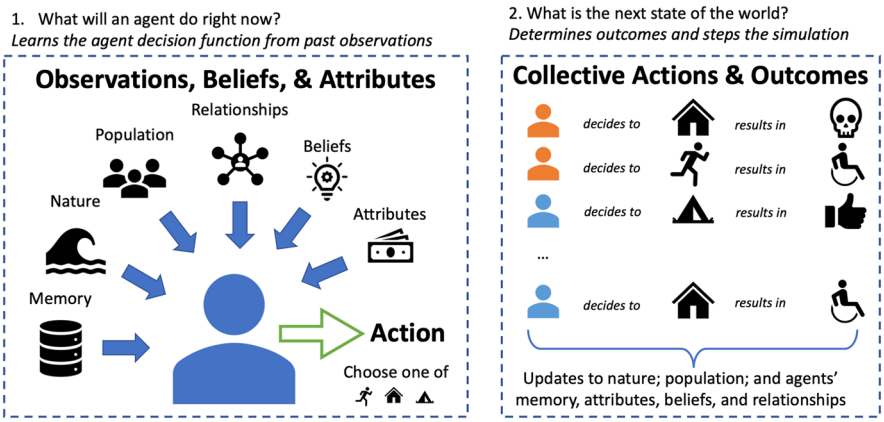


Fig. 10 An overview of our modeling approach to predict human behavior and social dynamics in simulated virtual words Shmueli 2010

When we compare the BNLEARN results to the PCALG results in these plots, we see that the F1 scores for the BNLEARN data are typically somewhat lower than average given their graph properties while their robustness values are significantly higher than those observed for the PCALG graphs. This indicates that the data generation process of the BNLEARN data is overall more challenging for the causal discovery algorithms, but that interestingly the predictions of the ensemble are more consistent across samples.

4 Predictive modeling of human behavior and social dynamics

We implemented an agent-based approach outlined in Fig. 10 to answer predict questions for four simulated scenarios, for example "How many people will evacuate at least once during the new hurricane?" Agents are modeled as having an internal state that consists of relationships, beliefs, and attributes. Agents can observe the population (e.g., the current total number of casualties) and the state of nature (e.g., current hurricane severity). Agents can remember their past, such as how many times an agent experienced a severe hurricane.

We experimented with SOTA machine learning models—Random Forest (RF), k-Nearest Neighbors (KNN), Logistic Regression (LR), Deep Neural Network (DNN)—to model decisions that agents make during individual time steps. We use data-driven models to fit a function from observational data, for example sampled data collected by research methods teams, that predicts what action an agent will take and how an agent will change as a result of their observation's current state (Zhang et al. 2016). Thus, our mechanistic simulation "stepper" then considers the agents collectively to determine outcomes and updates the agent states appropriately.

We apply our causal ensemble approach as described in the previous section to determine the causal relationships between agent observations, beliefs, attributes and actions. This produces a causal graph we use to perform feature selection in the

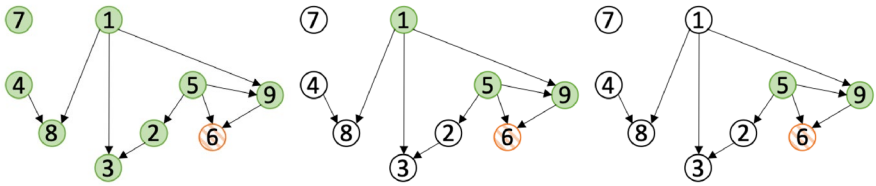
predictive model. When there is a chain between inputs and actions, all ancestors in that chain are included as features to train the machine learning model. The advantage is that this reduces the dimensionality of the problem and removes inputs that are spuriously correlated with the agent's decision. As a baseline, we simply do not perform feature selection and train ML model using all features.

Different scenarios produced vastly different training data for the agent decision models, with the smallest training data coming from the Disaster scenario and the largest coming from the Conflict scenario. Though we attempted to use a standard set of machine learning models, not all models were practical or effective across all scenarios. We found that LR was typically not fast enough to apply to most scenarios. In some scenarios RF was also too time consuming to apply. Most models we trained had similar single-step accuracy, they could predict what an agent will do next. Interestingly KNN models tended to exhibit better end-to-end accuracy on our held-out validation set. Generally, causal discovery did not produce better end-to-end results on the held-out validation set. Across all four scenarios, the benefit of causal feature selection was only shown for the Disaster and Power scenarios. Using TA2A versus TA2B sampled (research request) data led to equal predictive performance. For the Power scenario DNN demonstrated the highest performance followed by KNN and RF models; however, for the Disaster scenario KNN outperformed DNN and RF models. The performance was comparable when we experimented with different modeling decision for the Disaster scenario, for example model at the agent versus population-level, deterministic versus stochastic modeling, agent making multiple or one-choice decisions, etc.

To summarize our findings, our *predict performance was influenced by how compatible our ML-based simulation architecture matched the original simulation approach*. It is important to note that predict questions were of different complexity (Mitchell and Newman 2002; Ladyman et al. 2013) across and within simulation scenarios, which explains varied performance and the fact that no universal predictive model could be applied across four simulation scenarios. Predict answers with sampled data A versus B were comparable. Running predict analysis on full data would have helped our understanding of the effect of sampling on predict performance. Thus, additional experiments needed to fully explain our predict results and determine whether there were incorrect modeling assumptions made, the causal graphs were too noisy, the causal knowledge was not incorporated properly, there was not sufficient data for models to generalize on, or the predict questions were beyond a forecasting horizon (Martin et al. 2016; Abeliuk et al. 2020; Salganik et al. 2020).

4.1 Incorporating causal knowledge into predictive models

In addition to evaluating predictive models with and without causal knowledge systematically embedded on sampled data A and B across four simulation scenarios, we performed an extensive evaluation on internally simulated data with known ground truth (instead of inferred ground truth). We experimented with continuous binary



(a) All Inputs (b) Ancestors Causal (c) Direct Causal

Fig. 11 Each graph illustrates an example of the input nodes and output node defined by the experimental setting under which we trained ML models (Tsamardinos et al. 2003; Aliferis et al. 2010). Green signifies an input node; striped orange indicates the output node. In setting 1, all nodes excluding the output node are inputs to the model. In setting 2, all nodes in the ancestry of the output node are model inputs. In setting 3, only direct parents of the output node are model inputs

and mixed data types on non-intervened datasets and demonstrated that embedding causal knowledge improved predictive performance in several experimental settings. *Binary output variables (both causal parents and ancestors of mixed and binary inputs) and continuous output variables (causal parents of mixed and continuous inputs) demonstrated the benefit of relying on causal knowledge for predictive modeling.*

The data for our predict experiments were simulated using the R pcalg library as described in Sect. 3. We generated 1140 random DAGs of various sizes to represent varied causal structures and presented the example graphs in Fig. 5. For every simulated graph, we predicted the value of each node under three experimental setting as illustrated in Fig. 11. In the first setting, input information from all nodes is available (excluding the node we are predicting). In the second setting, information from the nodes in the causal ancestry of the predicted node is available. Finally, in the third setting, only nodes that are direct parents of the predicted node are available as inputs to the model.

With the generated datasets, we trained a DNN and two baseline ML models, RF and LR for prediction, classification (for binary outputs), and regression (for continuous outputs). Distinct models were trained for each graph with 70% of the samples used for training, 10% for validation, and 20% held out for testing. Our DNN for both regression and classification consists of three layers with 64 units and a dropout rate of 0.25. We used the Adam optimizer with a learning rate of 0.005 and early stopping. Mixed datasets made use of all models depending on the data type of the output node. A mix of binary and continuous data was given as input.

Training three types of models for each node as the output in every graph under all three experimental settings is computationally expensive. Therefore, we randomly select 255 unique graphs that covered each generation method at each graph size. In total, 6,468 models were trained and evaluated.

We investigated the relationship between the predictive power of each model and the inclusion of the causal knowledge. For that we measured the differences in the mean performance scores between the non-causal and the causally informed models, and evaluated the statistical significance of the comparisons using 1-tailed *t*-test. Table 4 shows the *p*-values and significance for the pairwise *t*-tests.

Table 4 Predictive model performance on non-interventional simulated data

Input type	Output type	Model	All nodes avg F1	Indirect avg F1	Direct avg F1
Mixed	Binary	DNN	**0.874****	0.845	0.852**
		Random Forest	**0.855*****	0.816	0.821
		Logistic Reg.	**0.865*****	0.845	0.849*
Binary	Binary	DNN	0.702	0.750***	**0.763****
		Random Forest	0.706	0.751***	**0.762*****
		Logistic Reg.	0.731	0.744***	**0.748****
Input type	Output type	Model	All nodes avg RMSE	Indirect avg RMSE	Direct avg RMSE
Mixed	Continuous	DNN	1.574	1.420**	**1.278****
		Random Forest	1.091**	1.421	**1.270*****
		Linear Reg.	**1.607*****	1.938	1.800
Continuous	Continuous	DNN	0.177	0.145***	**0.141****
		Random Forest	**0.148***	0.151	0.149**
		Linear Reg.	**0.079*****	0.092	0.092

Significance is denoted with * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$

In datasets with a mix of continuous and binary data types, a non-causal model (all available variables as input) outperformed any causal model in most instances. *For strictly binary datasets, a model leveraging causal feature selection had shown to always improve F1 scores.* In particular, a model trained with only direct causal parents of the prediction node yielded the best performance (similar to Aliferis et al. 2010). In a continuous setting, our results were more varied. The root-mean-square deviation (RMSE) values from a causal DNN model are statistically lower than the non-causal DNN; however, this result was reversed when using a linear model. The RF model showed little differences in the RMSE values.

It is important to note that unlike predict experiments with four simulated worlds, our additional experiments and analyses rely on having the true causal graph for a dataset. In practice, access to the ground truth graph is exceedingly rare. Most likely, researchers will have a learned causal graph produced from one of the many causal discovery algorithms or other methods. Errors in the inferred causal relationships are likely to lead to reduced performance of causal feature selection methods. In future work, we will perform similar predict experiments with (a) interventional simulated data and (b) learned causal graphs in order to quantify the impact of such errors in the causal structure. In combination with our current results, such analysis will provide practical evidence to researchers about the importance of causal feature selection and the potential need for improved methods of determining the underlying causal structure as discussed in earlier e (Aliferis et al. 2010).

5 Conclusions and future work

In this work we evaluated multiple approaches to discover the causal mechanisms of human behavior and social dynamics from observational data using four simulated worlds. In addition, we performed an additional evaluation on simulated datasets frequently used for benchmarking outside the human domain. We validated generalizability, reproducibility and robustness of these approaches for causal discovery (aka causal structure learning) and outlined their strength, weaknesses and limitations. We demonstrated that the existing methods are not generalizable across use cases and datasets, and are not robust to sampling. Specifically, we showed that causal ensembles with the top four performing algorithms are more robust to sampling, but are still hindered by the least stable algorithms. As expected, as we increase the sample size, the stability of ensemble algorithms also increases. Both explain and predict methods are vulnerable to data and modeling assumptions e.g., how agents make decisions, interact with each other and with the environment, and how interactions occur across the environmental factors. We also measured how causal discovery performance depends on the task complexity, data size, and the signal in the data. We demonstrated the importance of data to knowledge representation learning for causal discovery (Schölkopf et al. 2021) by empirically evaluating how knowledge extraction from data effects model performance.

When explicitly incorporating inferred causal knowledge into predictive models, we demonstrated the benefit of causal feature selection for two out of four simulation scenarios. However, it is important to note that the causal knowledge were inferred with high uncertainty. Therefore, it is necessary but not sufficient to improve causal discovery methods in order to boost predictive modeling of complex systems including but not limited to human behavior. The causes of uncertainty were compatibility of our simulation approach with virtual worlds' simulation approaches, not having sufficient data for models to learn from, or task complexity and the forecasting horizon. Our additional predict experiments where we incorporated known ground truth into machine learning models (rather than the inferred ground truth) showed the benefit of including causal knowledge for predictive modeling for multiple output variable type -- binary and continuous.

Our causal discovery and modeling results to explain and predict human behavior and social dynamics raise a number of interesting questions and directions for future work. However, as of now, traditional causal discovery approaches are limited and are insufficient to explain and anticipate human social dynamics. First, accounting for individual differences can significantly increase dimensionality of the data and confound estimates of causal effects when the structure of the causal model is not known a priori. Second, relatively little research exists on designing personalized interventions. Finally, little is known about how to enable contextualized reasoning, when changes in the individual and the environment inform interventions.

Mining real-world human behavioral data to discover natural experiments (King et al. 2011; Alipourfard et al. 2018) could be an alternative to inferring

the causal mechanisms from human behavioral data to study complex social phenomena in the Human Domain like social inequality, perception and susceptibility to disinformation or the spread infectious diseases e.g., Haushofer and Metcalf, 2020. But it presents major computational challenges for causal discovery and inference, and other multidisciplinary computational social science approaches. The major challenge is explicitly measuring the effects, which is difficult as treatment may itself be correlated with some aspects of human behavioral data, confounding analysis. Additional challenges include continuous treatments, fair causal inference, high-dimensional feature spaces (Feder et al. 2021) etc. Addressing national security challenges relevant to the Human Domain by discovering natural experiments or by using other computational methods, to save lives or to strengthen the democracy, will require extensive validation of the existing computational methods, as well as rethinking ethical usage of sharing data and strong multidisciplinary collaborations (Lazer et al. 2020; Watts 2011; Kahneman 2011).

Acknowledgements The project was conducted at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy and was sponsored by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreements HR0011939944 and HR0011044266. The content of the information does not necessarily reflect the position or the policy of the government and no official endorsement should be inferred. We would like to thank Robin Cosbey, Sannisth Soni, Zhuanyi H Shaw, Austin Golding, and Ryan Rabello for their contributions to this work. We are also grateful to Asmeret Naugle and Adam Russell for their guidance and recommendations.

References

Abeliuk A, Huang Z, Ferrara E, Lerman K (2020) Predictability limit of partially observed systems. Scientific Rep 10(1):1–10

Aliferis CF, Statnikov A, Tsamardinos I, Mani S, Koutsoukos XD (2010) Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. J Mach Learn Res 11(1):171–234

Alipourfard N, Fennell PG, Lerman K (2018) Using Simpson's paradox to discover interesting patterns in behavioral data. Preprint at arXiv:1805.03094

Athey S (2015) Machine learning and causal inference for policy evaluation. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 5–6

Bengio Y (2019) From system 1 deep learning to system 2 deep learning. http://www.iro.umontreal.ca/bengioy/NeurIPS-11dec2019.pdfAccessed 11 Nov 2021

Chickering DM (2002) Optimal structure identification with greedy search. J Mach Learn Res 3:507–554

Chollet F (2020) A definition of intelligence for the real world. J Artif Gen Intell 11(2):27–30

Cottam J, Glenski M, Shaw Y, Rabello R, Golding A, Volkova S, Arendt D (2021) Graph comparison for causal discovery. Visualization in data science

Feder A, Keith KA, Manzoor E, Pryzant R, Sridhar D, Wood-Doughty Z, Eisenstein J, Grimmer J, Reichart R, Roberts ME, et al (2021) Causal inference in natural language processing: estimation, prediction, interpretation and beyond. Preprint at arXiv:2109.00725

Gain A, Shpitser I (2018) Structure learning under missing data. In: International conference on probabilistic graphical models, PMLR, pp 121–132

Glymour C, Zhang K, Spirtes P (2019) Review of causal discovery methods based on graphical models. Front Genet 10:524

Greenland S, Mansournia MA (2015) Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. Eur J Epidemiol 30(10):1101–1110

Guo R, Cheng L, Li J, Hahn PR, Liu H (2020) A survey of learning causality with data: problems and methods. ACM Comput Surv (CSUR) 53(4):1–37

Haushofer J, Metcalf CJE (2020) Which interventions work best in a pandemic? Science 368(6495):1063–1065

Hofman JM, Sharma A, Watts DJ (2017) Prediction and explanation in social systems. Science 355(6324):486–488

Kahneman D (2011) Thinking, fast and slow. Macmillan, London

King G, Nielsen R, Coberley C, Pope JE, Wells A (2011) Comparative effectiveness of matching methods for causal inference. 15(1):41–67

Kummerfeld E, Rix A (2019) Simulations evaluating resampling methods for causal discovery: ensemble performance and calibration. Preprint at arXiv:1910.02047

Ladyman J, Lambert J, Wiesner K (2013) What is a complex system? Eur J Philos Sci 3(1):33–67

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. Behav Brain Sci. https://doi.org/10.1017/S0140525X16001837

Lazer DM, Pentland A, Watts DJ, Aral S, Athey S, Contractor N, Freelon D, Gonzalez-Bailon S, King G, Margetts H et al (2020) Computational social science: obstacles and opportunities. Science 369(6507):1060–1062

Martin T, Hofman JM, Sharma A, Anderson A, Watts DJ (2016) Exploring limits to prediction in complex social systems. In: Proceedings of the 25th international conference on world wide web, pp. 683–694

Mitchell M, Newman M (2002) Complex systems theory and evolution. Encycl Evol 1:1–5

Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Du Sert NP, Simonsohn U, Wagenmakers EJ, Ware JJ, Ioannidis JP (2017) A manifesto for reproducible science. Nat Hum Behav 1(1):1–9

Olteanu A, Castillo C, Diaz F, Kiciman E (2019) Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data 2:13

Pearl J (2009) Causality. Cambridge University Press, Cambridge

Pearl J (2019) The seven tools of causal inference, with reflections on machine learning. Commun ACM 62(3):54–60

Runge J, Bathiany S, Bollt E, Camps-Valls G, Coumou D, Deyle E, Glymour C, Kretschmer M, Mahecha M, Muñoz J, Nes E, Peters J, Quax R, Reichstein M, Scheffer M, Schölkopf B, Spirtes P, Sugihara G, Sun J, Zscheischler J (2019) Inferring causation from time series in earth system sciences. Nat Commun. https://doi.org/10.1038/s41467-019-10105-3

Saldanha E, Cosbey R, Ayton E, Glenski M, Cottam J, Shivaram K, Jefferson B, Hutchinson B, Arendt D, Volkova S (2020) Evaluation of algorithm selection and ensemble methods for causal discovery

Salganik MJ, Lundberg I, Kindel AT, Ahearn CE, Al-Ghoneim K, Almaatouq A, Altschul DM, Brand JE, Carnegie NB, Compton RJ et al (2020) Measuring the predictability of life outcomes with a scientific mass collaboration. Proc Natl Acad Sci 117(15):8398–8403

Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, Bengio Y (2021) Toward causal representation learning. Proc IEEE 109(5):612–634

Shmueli G et al (2010) To explain or to predict? Stat Sci 25(3):289–310

Spirtes P, Glymour CN, Scheines R, Heckerman D (2000) Causation, prediction, and search. MIT press, Cambridge

Stodden V, McNutt M, Bailey DH, Deelman E, Gil Y, Hanson B, Heroux MA, Ioannidis JP, Taufer M (2016) Enhancing reproducibility for computational methods. Science 354(6317):1240–1241

Strobl EV (2019) Improved causal discovery from longitudinal data using a mixture of dags. In: The 2019 ACM SIGKDD workshop on causal discovery, PMLR, pp 100–133

Tsamardinos I, Aliferis CF, Statnikov AR, Statnikov E (2003) Algorithms for large scale Markov blanket discovery. FLAIRS conference 2:376–380

Tu R, Zhang C, Ackermann P, Mohan K, Kjellstrm H, Zhang K (2019) Causal discovery in the presence of missing data. In: The 22nd international conference on artificial intelligence and statistics, PMLR, pp 1762–1770

Watts, Duncan J (2011) Everything is obvious:* Once you know the answer. Currency

Watts DJ, Beck ED, Bienenstock EJ, Bowers J, Frank A, Grubesic A, Hofman J, Rohrer JM, Salganik M (2018) Explanation, prediction, and causality: three sides of the same coin?

Yarkoni T (2019) The generalizability crisis. Behav Brain Sci. https://doi.org/10.1017/S0140525X20001685

Yu K, Li J, Liu L (2016) A review on algorithms for constraint-based causal discovery. Preprint at arXiv:1611.03977

Zhang H, Vorobeychik Y, Letchford J, Lakkaraju K (2016) Data-driven agent-based modeling, with application to rooftop solar adoption. Auton Agents Multi-Agent Syst 30(6):1023–1049

Dr. Svitlana Volkova is a recognized leader in the field of computational linguistics, machine learning and computational social science. She leads the development of human-centered analytics to explain, predict and prescribe human social systems and behaviors as they relate to national security challenges in the human domain. Solutions developed by Svitlana and her team advance understanding, analysis, and effective reasoning about extreme volumes of dynamic, multilingual, multimodal real-world social data. Since joining PNNL in 2015, Dr. Volkova has led more than ten projects, including two DARPA efforts. She authored more than 50 peer-previewed conference and journal publications. Svitlana was a Vice Chair of the ACM Future of Computing Academy between 2017 and 2019. She received her PhD in Computer Science in 2015 from Johns Hopkins University where she was affiliated with the Center for Language and Speech Processing and the Human Language Technology Center of Excellence.

Dr. Dustin Arendt is a Senior Research Scientist and Team Lead in the Visual Analytics Group at Pacific Northwest National Laboratory, joining the lab in October 2014. He received his Ph.D. from Virginia Tech in 2012 and completed a postdoc at the Air Force Research Laboratory from 2012 to 2014 where he researched graph visualization and applied machine learning. Since joining PNNL he has worked in several domains including visualization for cybersecurity, streaming data visual analytics, visual abstraction, dynamic graph visualization, visualization for natural language processing, interactive machine learning, explainable machine learning, and visualization for machine learning model validation and comparison. Currently, his interests are at the intersection of human-computer interaction, data science, and visual analytics, with a focus on validating machine learning models through explanations and exploratory data analysis. His research involves rapid prototyping and empirical evaluation of tools that blend machine learning, data science, and visualization.

Dr. Emily Saldanha is a research scientist in the area of data science at PNNL where her work focuses on applying machine learning and deep learning techniques to identify and understand patterns in complex data. She has specific interests in the development of robust methods for application areas ranging from energy technologies to computational social science. She received her Ph.D. in physics from Princeton University in 2016, where her work focused on the development and application of calibration algorithms for microwave sensors for cosmological observations.

Dr. Maria Glenski is a Data Scientist in the Data Science and Analytics Group, National Security Directorate at the Pacific Northwest National Laboratory. Her research focuses include computational social science approaches to behavior analysis, characterization, and modeling of complex social behavior in diverse online social environments and explainable artificial intelligence (XAI) evaluating the impacts of algorithmic biases in machine learning models. Dr. Glenski received her Ph.D. in Computer Science from the University of Notre Dame where she was an Arthur J Schmitt Leadership in Science and Engineering Fellow. Dr. Glenski's research has been published in top tier venues including WWW, ACL, ACM TIST, and CSCW and she regularly serves on the program committee of several international conferences and journals.

Ellyn Ayton is a Data Scientist at PNNL. She received her Master's in Computer Science from Western Washington University. Her research areas of interest include deep learning and its many applications, such as detecting digital deception. She contributes to projects that use NLP to extract predictive signals from open source data, evaluates the effectiveness of causal mechanisms in machine learning models, and develops methods of interpretability and explainability of black-box deep learning models.

Dr. Joseph Cottam works on visualization and data analysis frameworks. He is interested in involving expert knowledge in analytic processes and applications where data comparison s are major part of exploration. Currently he is working with biological data, networks that change over time and causal analysis. He has significant experience with graph analytics, data analysis in out-of-core and streaming scenarios and HPC systems. He is also interested in programming language design and distributed computing.

Dr. Sinan G. Aksoy is a discrete mathematician and data scientist specializing in graph theory and network science at Pacific Northwest National Laboratory. In addition to his theoretical research in spectral and extremal graph theory, he is interested in applying graph theoretic methods to study wide-ranging complex systems, including those arising from social networks, power and communication systems, supercomputing, and cyber security. In these domains, his work focuses on developing methodologies that draw from probability, statistics and combinatorics. Sinan's research spans over 20 publications and has appeared in journals such as the SIAM Journal on Discrete Mathematics, SIAM Journal on Mathematics of Data Science, EPJ Data Science, Advances in Applied Mathematics, and the Journal of Supercomputing. Sinan holds a PhD in mathematics from the University of California, San Diego, and a BA in economics and mathematics from the University of Chicago.

Dr. Brett Jefferson is a data scientist at Pacific Northwest National Laboratory, where he's conducted research for the past two years. Prior to joining the laboratory, Brett completed a Ph.D. program in mathematical psychology and a M.A. degree program in mathematics at Indiana University. Brett's research focuses on three main areas, mathematical modeling using topology, robustness assessment of machine learning classifiers, and quantitative assessments of human and machine systems. Dr. Jefferson is a committee member for the National Association of Mathematicians, member of the Cognitive Science Society and member of the Society for Mathematical Psychology.

Karthik Shivaram is PhD Student at Tulane University, where his research focuses on Computational Social Science, Machine Learning and Natural Language Processing. Currently he is working on developing semi-supervised learning methods to aid Stance Classification in Social Networks. He has previously interned at PNNL where he worked on applying Graphical Causal Discovery Models to determine causal relationships for computer simulations and investigated the use of Representational Learning methods to aid Causal Inference Performance. Previous to this he has worked as a Data Scientist and as a Software Developer at Accenture. He also holds a master's degree in computer science from Illinois Institute of Technology and a bachelor's degree in Mechanical Engineering from BMS Institute of Technology.

Authors and Affiliations

Svitlana Volkova[1] · Dustin Arendt[1] · Emily Saldanha[1] · Maria Glenski[1] · Ellyn Ayton[1] · Joseph Cottam[1] · Sinan Aksoy[1] · Brett Jefferson[1] · Karthnik Shrivaram[1]

✉ Svitlana Volkova
svitlana.volkova@pnnl.gov

[1] Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99354, USA