



Causal Data Integration

Brit Youngmann
CSAIL MIT
brity@mit.edu

Michael Cafarella
CSAIL, MIT
michjc@csail.mit.edu

Babak Salimi
University of California,
San Diego
bsalimi@ucsd.edu

Anna Zeng
CSAIL, MIT
annazeng@mit.edu

ABSTRACT

Causal inference is fundamental to empirical scientific discoveries in natural and social sciences; however, in the process of conducting causal inference, data management problems can lead to false discoveries. Two such problems are (i) not having all attributes required for analysis, and (ii) misidentifying which attributes are to be included in the analysis. Analysts often only have access to partial data, and they critically rely on (often unavailable or incomplete) domain knowledge to identify attributes to include for analysis, which is often given in the form of a causal DAG. We argue that data management techniques can surmount both of these challenges. In this work, we introduce the Causal Data Integration (CDI) problem, in which unobserved attributes are mined from external sources and a corresponding causal DAG is automatically built. We identify key challenges and research opportunities in designing a CDI system, and present a system architecture for solving the CDI problem. Our preliminary experimental results demonstrate that solving CDI is achievable and pave the way for future research.

PVLDB Reference Format:

Brit Youngmann, Michael Cafarella, Babak Salimi, and Anna Zeng. Causal Data Integration. PVLDB, 16(10): 2659 - 2665, 2023.
doi:10.14778/3603581.3603602

PVLDB Artifact Availability:

The source code, data, and/or other artifacts have been made available at <https://github.com/brityoungmann/CDI/blob/main/README.md>.

1 INTRODUCTION

Causal inference lies in the heart of empirical research in natural and social sciences and is commonly used in multiple disciplines, including sociology, medicine, and economics [26, 33]. It aims to answer causal queries, such as, “Does being overweight cause coronary heart disease – independent of cholesterol, and diabetes?” or “Do tobacco advertisements entice adolescents to buy more cigarettes regardless of whether their parents smoke?” Causal inference enables analysts to answer causal questions about attributes from a dataset, ultimately enabling them to make real-world discoveries. Pearl’s framework [40], which we adapt in this work, provides a principled way to causal inference using structural causal models.

Cause-effect questions are designed to determine whether an *exposure* variable (a pain reliever) causes or affects an *outcome* variable (pain). To correctly estimate causal effects, one must consider

Table 1: Example Input Dataset.

State	Mask Policy	Confirmed Cases	New Cases	Recovered	Death cases
MA	yes	121046	2740	4980	109
FL	yes	640978	24349	25140	55
CA	no	735235	31150	42170	34
SD	no	15300	1791	2083	49

confounding variables—variables that influence both the exposure and outcome and thus might distort the association between them. Otherwise, one might draw perplexing conclusions due to *confounding bias* [41] that can lead to overestimation or underestimation of the association between exposure and outcome. The internal validity of a study heavily relies on how well confounding bias has been addressed. However, to determine a sufficient set of confounding variables to account for confounding bias, *background knowledge* about the data generative process is required. This knowledge is often given in the form of a *causal DAG* depicting causal relationships between the attributes [40] (and is wildly used in econometric and social sciences [26, 33]). Pearl [40] presented sufficient and necessary conditions for identifying¹ the *adjustment set* of variables to include in the analysis, which can be checked against a causal DAG. However, analysts often lack such a DAG [22]. While associational assumptions are testable from data, causal relations cannot, in general, be fully recovered from data [41].

We identify two fundamental challenges for conducting valid causal inference: First, critical confounding variables may not be included in the data. Second, even if they are included, the background knowledge required to identify the correct adjustment set of variables might be missing. We argue that data management techniques and ideas can surmount both of these challenges. We illustrate that via the following example:

Mary, a data analyst in the WHO organization, aims to estimate the (treatment) effect of a mask policy on the coronavirus mortality rate. To do this, she examines a dataset containing Covid-19-related facts in multiple states in the US (an illustration of this dataset is given in Table 1), and uses a Causal analysis tool (e.g., [49, 53]). To correctly estimate the effectiveness of face masks, Mary must consider confounding variables. However, there are confounding variables that are not included in her data. For example, the weather affects people’s willingness to wear masks and Covid-19 death rate [47]. Moreover, to determine a sufficient set of confounding variables, Mary must use background knowledge of causal relationships she may lack.

In this example, variables missing from the dataset are critical for the analysis to avoid confounding bias. Thus, Mary must *integrate her data* with external data sources. This is a laborious task, typically done using data discovery and integration tools (e.g., [36, 59]) by skilled programmers—which is often not the case for social or

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.

Proceedings of the VLDB Endowment, Vol. 16, No. 10 ISSN 2150-8097.
doi:10.14778/3603581.3603602

¹In causal inference, identifiability typically refers to the conditions that permit measuring causal effect from observed data. Here, we discuss the ability to identify a sufficient adjustment set of variables for accurate causal inference.

natural scientists. We argue this is the case for many real-life scenarios. After augmenting the data to include unobserved variables, the next step is determining a sufficient adjustment set of variables to account for confounding bias. However, Mary does not have the required background knowledge to do that.

Our Vision: Our goal is to mine unobserved confounding variables and missing background knowledge needed to answer causal questions. To this end, we introduce the Causal Data Integration (CDI) problem. CDI is an effort to augment an input dataset with unobserved variables and background knowledge required for causal analysis. To address the challenge of finding unobserved variables, the system mines variables from external sources. To overcome the challenge of identifying the right adjustment set of confounders, the system then builds a corresponding causal DAG, ensuring the adjustment set is sufficient for controlling the confounding bias. A CDI system may greatly reduce the manual effort devoted to performing causal analysis, allowing analysts to get all required data and background knowledge quickly and effectively.

CDI may also be beneficial for the data management community. Causal inference has been shown to be useful for various data management tasks, including query result and classifier explanation [8, 42, 56], and hypothetical reasoning [21]. However, some of these studies make strong assumptions about the underlying causal DAG that may not hold in practice [26, 33]. For example, [21] assumes that the DAG is known, and [42] assumes that all confounding variables are present in the input dataset. To address these issues, a CDI system can be utilized to construct a causal DAG for a given dataset while integrating relevant missing attributes. This not only addresses the limitations of existing studies but also enables the use of causal inference in more complex and realistic scenarios.

In this work, we show how a CDI system can be accomplished. **Measure of Success:** Distinct from the classical data integration task, where success is defined by its ability to combine data from different sources and provide users with a unified materialized view of them, the success of a CDI system is defined by the ability to discover all relevant unobserved variables and recover the correct set of confounders for a causal question. The CDI task is potentially easier than data integration. However, its impact is greater, as it enables analysts to conduct causal analysis more easily. A CDI system enables analysts to include all confounding variables in the analysis; failing to do so may lead to false discoveries, incorrect medical diagnoses, and erroneous conclusions [26, 33]. In our example, a CDI system mines, among others, attributes describing the weather and population density per state, and infers causal relations among the attributes in the augmented dataset. A successful CDI system would assist Mary in identifying the right adjustment set of attributes to account for confounding bias, allowing her to accurately estimate the effect of an enforced mask policy on the mortality rate.

Challenges & Opportunities: The first challenge is to augment an input dataset with unobserved variables, ensuring the *output causal DAG is complete* (i.e., there are no unobserved confounding variables). Data discovery methods have been extensively researched to identify relevant data sources that can be integrated with an input dataset for attribute extraction [12, 14]. However, these methods are not specifically designed to discover unobserved attributes that are relevant for cause-effect estimation. This presents a unique

opportunity to develop dedicated data discovery tools tailored for identifying unobserved confounding variables.

Extracted attributes may contain data quality issues, such as outliers and missing values. These issues may lead to erroneous conclusions, compromising the reliability of causal inference. In particular, missing values pose a significant challenge in causal inference as they may result in *selection bias* [10], where certain samples are preferentially excluded from the data. Other data quality issues, such as inconsistent data formats and data duplication, can also impact the accuracy of causal inference analysis. Therefore, in addition to handling missing values, it is crucial to ensure that the system is robust to various data quality issues to ensure the validity of causal inference results.

The challenge of identifying the correct adjustment set and constructing a causal DAG is particularly difficult when dealing with hundreds of extracted attributes. A human-driven causal DAG construction (from domain expertise) is infeasible in this scenario, and automatic methods for building a full causal DAG may not be feasible either [22]. To this end, we propose the construction of a cluster causal DAG (C-DAG) [9], which groups related attributes and only specifies causal relationships between clusters of attributes. By doing so, the correct adjustment set of attributes can still be identifiable from this C-DAG. This approach reduces cognitive overhead for analysts and makes causal analysis possible for high-dimensional data, which is currently not feasible for existing tools (e.g., [49, 53]). However, a main challenge is how to maintain the necessary level of granularity to ensure accurate causal inference. The limitations of existing solutions and challenges in constructing the desired C-DAG are discussed, and a limited prototype implementation for solving the CDI problem is presented with preliminary experimental results demonstrating that solving CDI is achievable.

2 THE CDI PROBLEM

2.1 Data Model and Background

The input dataset \mathcal{D} contains the exposure (T) and outcome (O) attributes. A *causal DAG* for \mathcal{D} is a DAG whose nodes are the attributes and whose edges capture all causal relationships between the attributes [40]. Generally, because causes must precede effects, the literature considers causal DAGs rather than graphs [40]. Causal DAGs provide a simple way of graphically representing causal relationships and are particularly helpful in understanding potential sources of bias in causal estimations. It is well-known that background knowledge is required to determine a causal DAG for a given dataset [41]. A causal DAG can only be as good as the background information used to create it [38]; a DAG is complete and therefore has a causal interpretation only if it contains all common causes of any two variables. In Section 3.3, we review existing approaches for building a causal DAG for a given dataset and discuss their limitations. We assume the analyst is interested in investigating a causal relationship between T and O . She may wish to estimate, e.g., *the direct, indirect, or total effect* [40]. The total effect of T on O is the total extent to which O is changed by the T . It is equal to the sum of the direct and indirect effects. Causal questions about total and (in)direct effect could be answered by identifying the right set of *confounding* (that influence both T and O) and *mediating* (that

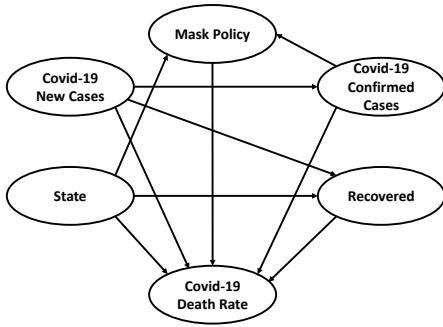


Figure 1: Example causal DAG for Table 1

mediate the relationship between T and O variables. Such variables could be identified using graphical criteria (e.g., the backdoor criterion [40]) that can be checked against a causal DAG.

Example 2.1. In our example, T is mask policy, and O is death cases. A corresponding causal DAG is shown in Figure 1. To obtain a reliable estimate of the total effect of T on O , it is crucial to adjust for all relevant confounding variables. However, mistakenly assuming that the DAG contains all relevant information, the adjustment set only includes the attributes `state` and `confirmed cases`. As is common in causal inference, missing confounders can lead to biased estimates. Thus, it is necessary to consider external data to incorporate missing variables.

2.2 Problem Formulation & Challenges

Given an input dataset, the system mines unobserved confounding attributes relevant for a given causal estimation, yielding an augmented dataset $\hat{\mathcal{D}}$. It then maps the attributes in $\hat{\mathcal{D}}$ into a causal DAG, to enable analysts to determine the adjustment set. We next outline key challenges in designing such a system:

(1) Completeness: Unobserved variables can dramatically affect the quality of causal analysis [32, 41]. Thus, the first challenge is to *ensure the generated DAG is complete*, i.e., there are no unobserved confounding variables w.r.t. the provided data sources. Our goal is to ensure that all relevant information can be automatically extracted from the provided sources. However, completeness, in general, cannot be guaranteed since relevant variables may exist outside of the provided data sources.

(2) Robustness: The robustness of a CDI system is determined by its ability to handle various data quality issues that may arise during the integration process. As in general data integration, there are multiple data quality issues that a CDI system has to face, including missing values and outliers. The impact of such issues on causal inference is particularly significant. Unlike in machine learning, where data quality issues may only result in a slight drop in accuracy, these issues can lead to false discoveries in causal inference if not handled properly. This is because if these issues are not randomly distributed across the exposure groups, they can significantly impact the validity of the resulting conclusions. One common issue that can significantly impact causal inference is missing data. Extracted attributes may contain missing values, which can lead to selection bias and erroneous conclusions. Thus, a key challenge is to ensure the system is robust to various data quality issues.

Table 2: Extracted attributes from different sources.

Extracted Attributes from US Open Data				
State	Population size	Population density		
MA	6,981,974	901		
FL	22,244,823	402		
CA	39,029,342	254		
SD	909,824	12		

Extracted Attributes from DBpedia				
State	Governor	snow inch	Avg temp.	Min temp.
MA	Maura Healey	51.05	48.14	23
FL	Ron DeSantis	-	71.8	70
CA	Gavin Newsom	-	61.17	54
SD	Kristi Noem	37.43	45.54	34

(3) Conciseness: We can potentially extract hundreds of attributes from given data sources, which can lead to a complex, high-dimensional causal DAG. To overcome this, the CDI system should aim to reduce users’ cognitive overhead in interpreting the DAG. It is also crucial to ensure the scalability of causal analysis, as existing tools are not feasible for high-dimensional data [49, 53]. One approach is to ensure that related attributes are grouped into clusters while still ensuring that the correct set of confounders is identifiable. Thus, a third challenge is to ensure the output DAG is concise. The CDI system should be able to construct a causal DAG that effectively captures the causal relationships between attributes while maintaining the necessary level of granularity to ensure accurate causal inference. This will enable analysts to efficiently handle high-dimensional data while facilitating information elicitation about the validity of the assumptions used for drawing causal inferences.

3 THE CDI SYSTEM

Our system operates as a 3-steps pipeline. First, the *Knowledge Extractor*, extracts candidate attributes to handle the *completeness* challenge. The *Data Organizer* then handles data quality issues to account for the *robustness* challenge. Last, the *Causal DAG Builder* builds a clustered casual DAG to handle the *conciseness* challenge. We next present the key challenges and research opportunities for each system component.

3.1 The Knowledge Extractor

This module receives a table \mathcal{D} (containing T and O) and sources of knowledge. From each source, it extracts attributes representing additional properties of entities from \mathcal{D} . A successful Knowledge Extractor mines all unobserved variables that should be included in the analysis. The primary goal of the *Knowledge Extractor* is to extract all relevant variables from the provided data sources to enable reliable causal inference. In causal inference, the *unconfoundedness assumption* [39] (the assumption that all variables affecting both the exposure T and the outcome Y are observed and can be controlled for) is a key consideration. This assumption is similar to the closed-world assumption in database in that it assumes that the database is complete. However, since complete knowledge of all variables affecting both T and O is often unattainable, the focus should be on extracting all relevant variables from the provided data sources. The module is designed to ensure that as much relevant information as possible can be automatically extracted from the provided sources, and its capabilities can be expanded over time to incorporate new sources or refine the extraction and integration process. We next provide an example of how it works with specific data sources.

Data Lakes: There has been unprecedented growth in the volume of publicly available data (e.g., the US Federal Open Data [6] and the

Federal Reserve Economic Data [4]). We can rely on the rich body of work on data discovery [12, 14] to find relevant data sources that can be aligned with the input table for attribute extraction. However, existing methods use notions of relevance for discovery [36, 59, 60], whereas we search for unobserved confounding variables relevant to cause-effect estimations. Recent work has proposed solutions that can discover joinable datasets with column(s) correlated with a given target column and an input table [19, 44]. However, future research is required to extend these techniques for extracting candidate confounding attributes from tabular data, which can further improve the completeness of the extracted information.

Knowledge Graphs: There are multiple general-purpose (e.g., DBpedia [3]) or domain-specific (e.g., for medical proteomics [45], or protein discovery [35]) KGs that act as central storage for data. To extract attributes from a KG, we may map values that appear in \mathcal{D} to their corresponding entities in the KG \mathcal{G} , using named entity disambiguation tools (e.g., [37]). Next, given an entity, one can extract all of its properties, building the universal relation [20] out of all derived relations. One of the strengths of a KG is that most of the attributes are already reconciled. To extract more attributes, one may "follow" links in \mathcal{G} (i.e., extract the properties of values which are entities in \mathcal{G} as well). Future work will identify which links in a KG are worth following to extract relevant confounders.

Unstructured Data: Unstructured data like text and images are valuable for identifying missing confounding variables. Medical records often include unstructured text, such as physician notes, which offer insights into patients' health status and potential confounders. Medical images also contain important features that can influence treatment outcomes. However, analyzing unstructured data for causal analysis poses challenges due to their lack of organization. Existing techniques can be applied to extract meaningful information [57]. For example, sentiment analysis can be applied to text to extract relevant features, and vision methods can be used to extract image parts. This can improve the accuracy of causal analysis. Future work could develop methods for extracting interpretable features from unstructured data tailored for causal inference.

Example 3.1. The *Knowledge Extractor* extracts potential confounding variables from multiple knowledge sources. Specifically, it extracts the population density and size by state from the US Open data lake by joining the input table with a table containing statistics on population (based on state names). From DBpedia, it extracts available properties of each state (by aligning the state name with its entity in DBpedia), such as min temperature, amount of snow, and governor. A visualization of extracted attributes is given in Table 2.

Challenges & Opportunities: The *Knowledge Extractor* faces the challenge of identifying relevant unobserved attributes to be added for any arbitrary input dataset. The module must operate correctly even in cases of value mismatches, such as when a state name is given explicitly or as a shortcut. To tackle this challenge, researchers can rely on a rich body of work on entity linking and disambiguation [17, 18, 31]. Another challenge is to ensure that only relevant attributes are extracted and to avoid the curse of dimensionality. To achieve this, we plan to draw inspiration from first principles, such as an information-theoretic approach or approaches based on the maximum likelihood principle.

3.2 The Data Organizer

The *Data Organizer* receives the extracted attributes and outputs an augmented dataset. Extracted attributes may contain data quality issues, such as missing values and duplicated data, which can compromise the reliability of causal analysis. A successful implementation of this module would handle all data quality issues and ensure the system is robust to data quality issues. We next discuss several example failure modes the *Data Organizer* have to face:

Functional Dependencies: We search for attributes affecting both T and O . The presence of logical dependencies can hinder this process, as they can obscure the true relationships between them. In particular, causal inference literature assumes that the underlying distribution is *strictly positive* [40]. This assumption breaks down in the presence of logical dependencies: Assume we have an attribute E with the FD $E \Rightarrow T$. This means that when conditioning on E , there is no association between T and O , regardless of O . Namely, the existence of E violates the strict positivity assumption. A straightforward solution is to discard such attributes [42]. Another solution is to group attributes with FDs, treating them as a single node. We plan to develop a principled approach to discovering causal relations with the presence of FDs. **Missing Values:** Inferring causal relations requires recovering the probability distributions of extracted attributes [22]. But since extracted attributes may contain missing values, we must ensure that those probabilities are *recoverable* to overcome potential selection bias. Inverse Probability Weighting (IPW) is a commonly used technique to overcome selection bias [48]. In IPW, we restrict the attention only to complete tuples, but more weight is given to some tuples. A main challenge is to define sufficient conditions to detect cases where selection bias may occur. In such cases, there is a need for a principled approach to assigning weights to complete tuples (following the IPW approach) to be used to discover causal relations correctly. This can be done by extending the ideas proposed in [56]. **Complex Table Relations:** Another challenge is to handle complex table relations. Extracted attributes may have one/many-to-many relations with values in the input table. We aim to generate a single, augmented table. A straightforward solution is to aggregate multi-valued attributes; another possible direction is to develop representation learning techniques tailored for causal inference [11]. To accommodate many-to-many relations, one may use similar ideas as proposed in [43].

Challenges & Opportunities: Other data quality issues, such as inconsistencies, and outliers, can also compromise the reliability of causal inference and lead to erroneous conclusions if not handled properly. This is because if these issues are not randomly distributed across the exposure groups, they can lead to bias that may significantly impact the validity of the resulting conclusions. Merely using existing data-cleaning techniques (e.g., [16, 27]) may not be sufficient in the context of causal inference, as they are not designed to validate that bias is resolved. Thus, there is a need for innovative techniques that can effectively address these data quality issues to ensure accurate and reliable causal analysis. This is particularly important in high-stakes domains, such as healthcare, where erroneous conclusions can have severe consequences.

Example 3.2. The *Data Organizer* identifies the functional dependency between state and governor and drops the governor attribute. For attributes containing missing values (e.g., snow inch),

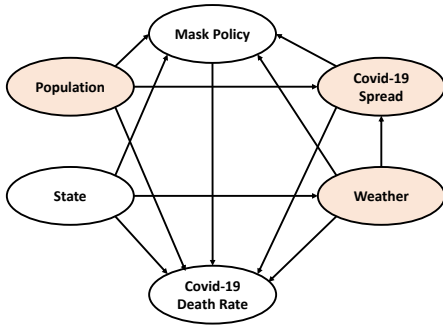


Figure 2: The generated C-DAG. Nodes representing a cluster of nodes are colored pink.

it checks if selection bias might occur. If so, it computes the weights for existing values required for the analysis. It then joins the input dataset with extracted attributes, yielding an augmented dataset.

3.3 The Causal DAG Builder

This module receives the augmented dataset and outputs a *concise causal DAG*. Any successful *Causal DAG Builder* should output a comprehensible and concise causal DAG while ensuring that the correct adjustment set of attributes is identifiable from that DAG.

Desiderata: The desiderata for a desired causal DAG is summarized as follows. First, there is a need to limit the number of nodes in the output DAG. This is critical to ensure that causal analysis is possible for high-dimensional data (which is not feasible for existing tools [49, 53]) and will reduce analysts’ cognitive overhead in interpreting the DAG. Second, to ensure the output DAG is comprehensible, only semantically-related attributes (e.g., avg temp, snow inch) should be grouped and represented as a single node in the DAG. Last, we must ensure that the output DAG enables identification of the right adjustment set of confounders for accurate causal inference.

Clustered Causal DAG: To address the above desiderata, we propose to build a *cluster causal DAG* (C-DAG) [9]. A C-DAG provides only a partial specification of causal relationships among attributes, alleviating the requirement of specifying all causal relationships. In a C-DAG, some attributes are grouped, and only the causal relationships between clusters of attributes are specified. Previous work [9] showed that cause-effect computations can be done directly from a C-DAG. In the output C-DAG, only semantically-related attributes should be clustered, and the number of nodes should be limited. A remaining open question is how to ensure that the correct adjustment set of confounding attributes is identifiable since, by grouping attributes, we lose information about their inter-relations. We next discuss the main challenges in building a desired C-DAG.

Identifying Causal Relationships: Causal discovery is a well-studied problem [22, 50, 51, 55, 61], whose goal is to infer causal relations among a set of attributes. Existing solutions can be split into two main approaches. Data-centric methods infer a causal DAG based on data properties [22]. Though it is well-known that background knowledge is required to determine causal relations [41], a causal DAG can be inferred from the data under some assumptions [15, 22] (e.g., sufficiency, faithfulness). However, such assumptions may not hold in practice and therefore limit the applicability of these algorithms. Further, such algorithms may not capture causal

relations that are not present in the data and ignore available semantic information. Text-mining approaches (e.g., [23–25]) infer causal relations among concepts by extracting claims from text documents. Other promising approaches are language models such as ChatGPT [13] that are trained to discover causal relations. The strength of such methods is in identifying cause-effect relations between seemingly-independent attributes or relations that otherwise could not be found in the data. However, they lack any concrete data on which to validate causal relations and may fail to distinguish between direct and indirect effects, resulting in graphs containing redundant edges. Further, such methods are also sensitive to the quality of attribute names and cannot be fully trusted.

We argue that a hybrid approach that merges the two approaches is required to overcome their shortcomings. Such an approach uses both data-centric and text-mining solutions to infer causal relations. However, achieving high accuracy in determining causal relationships is still a challenge for CDI; it requires substantial in-domain knowledge. A possible improvement is to use human-in-the-loop tasks, minimizing human effort while maximizing accuracy.

Grouping Attributes: We aim to group semantically related attributes and lift low-level information (e.g., temp, snow) into higher-level concepts (weather). Semantic similarity between attributes can be measured using embedding techniques (e.g., [34]), or ontological relationships (e.g., [28]). To allow analysts to reason about the C-DAG, we must further assign meaningful topics to the obtained clusters. This can be done using recent advances in topic modeling or zero-shot topic classification methods (e.g., [54]).

Identifiability: Besides causal relationships, a causal DAG specifies conditional independencies among the attributes [40]. To ensure the right adjustment set of attributes is identifiable from the C-DAG, the set of conditional independencies among attributes induced by the full causal DAG (where attributes are not clustered) should hold in the C-DAG (and vice versa). But since we group attributes, some of the conditional independencies may not be recoverable. While there is a rich body of work on graph summarization [29, 30, 52, 58], a tailored causal solution is required to ensure that all relevant conditional independencies are recoverable from the output C-DAG. Another open question is whether a single C-DAG is sufficient to identify the adjustment sets for multiple cause-effect estimations. In Section 4 we propose a best-effort algorithm to build a C-DAG.

Example 3.3. In our example (see Figure 2), the *C-DAG builder* groups low-level related attributes (e.g., pop density, pop size) into clusters associated with relevant topics (e.g., population), and infers causal relations among the attribute clusters.

4 PROOF OF CONCEPT IMPLEMENTATION

We have built CATER, a limited, proof-of-concept prototype of our CDI system. Our code is available at [7]. We next review some early experimental results. Though our experiments are limited, our results are highly promising, showing that solving CDI is achievable.

Implementation Details: In CATER, the *Knowledge Extractor* extracts attributes from DBpedia [3]. The *Data Organizer* discards attributes that have functional dependencies with the exposure or outcome, following [42]. We detect and handle selection bias in extracted attributes using similar ideas as proposed in [56]. The *C-DAG Builder* first groups attributes using VARCLUS [46] and

Table 3: Quality Evaluation.

Dataset	Baseline	E	Inclusion of Directed Edges			Absence of Edges			Direct Effect
			Precision	Recall	F1	Precision	Recall	F1	
FLIGHTS (V = 9, E = 17)	CATER	25	0.94	0.64	0.74	0.62	0.72	0.66	0.04
	GES	33	0.94	0.48	0.64	0.60	0.43	0.50	0.2
	LiNGAM	23	0.70	0.52	0.60	0.65	0.41	0.50	0.2
	PC	27	0.70	0.44	0.54	0.62	0.42	0.49	0.2
	GPT-3 Only	63	1.0	0.27	0.43	0.16	1.0	0.28	0.07
	FCI	39	0.70	0.30	0.43	0.40	0.30	0.35	0.15
COVID-19 (V = 11, E = 23)	CATER	27	0.74	0.63	0.68	0.59	0.61	0.6	0.01
	GPT-3 Only	82	0.91	0.26	0.4	0.25	0.78	0.38	0.02
	GES	16	0.26	0.37	0.31	0.49	0.31	0.38	0.05
	PC	13	0.17	0.31	0.22	0.49	0.3	0.37	0.05
	FCI	13	0.13	0.23	0.17	0.26	0.24	0.25	0.05
	LiNGAM	1	0	0	0	0.63	0.42	0.5	0.05

assigns topics to clusters using GPT-3 [13]. It then populates the graph edges according to templated query responses from GPT-3, and prunes redundant edges according to the PC algorithm [51]. This algorithm may result in a graph containing cycles; however, we report that CATER yields DAGs in examined cases.

Datasets: We examine two datasets: COVID-19 [1], which includes information such as number of confirmed/death/recovered Covid-19 cases worldwide; and FLIGHTS [2], which contains transportation statistics of domestic flights in the USA. Our pipeline was executed end-to-end in 645 and 304 seconds for FLIGHTS and COVID-19, resp.

Examined Scenarios: We use CATER to estimate the direct effect of an exposure variable on an outcome (when not mediated through mediator variables)—a common casual analysis task [40]. In COVID-19, we estimate the direct effect of a country on Covid death rate, and in FLIGHTS, we estimate the effect of departure city on flight delays. The goal is to identify mediator variables that should be included in the analysis, as many such variables are not included in the datasets (e.g., weather, pop. density). *In both cases, the direct effect should be equal to zero (there is no direct effect). Failing to identify mediators correctly will result in incorrectly estimating the direct effect and falsely concluding that there is one.*

Baseline Methods: We picked our current best configurations for the Knowledge Extractor and the Data Organizer. The node clusters and their assigned topics are the same across all baselines. We evaluate different approaches to infer causal relations, demonstrating the superiority of a hybrid approach over existing solutions. Our baselines include data-centric algorithms: PC [51], FCI [51], GES [15] and LiNGAM [50], and a text-mining approach (asking templated queries to GPT-3 [13]). To serve as **ground truth**, we gather domain expert knowledge C-DAGs. For instance, airline carriers and weather are causes for flight delays [5].

Quality Metrics: We measure the quality of generated C-DAGs w.r.t. ground truth, in two ways: (1) directed edge presence or absence, and (2) established effect between T and O . The precision, recall, and F1 scores for presence/absence edges are reported; High F1 scores indicate alignment with the ground truth. We report the effect between T and O described by each C-DAG. An effect close to 0 indicates that all mediator attributes are identifiable.

Results Summary: As illustrated in Table 3, CATER achieved for both datasets the highest F1 scores and the lowest direct effect values. Further, in both scenarios, CATER identified the same sets of mediator variables as those identified in the ground truth. This promising result illustrates that even a naive hybrid approach can improve upon existing causal discovery solutions.

Performance of GPT-3-Only closely trails CATER’s performance, especially in COVID-19. Along with the ground truth and CATER, GPT-3-Only captured many causal relations which data-centric methods were not able to, such as climate \rightarrow spread of Covid-19. However, it outputs graphs with notably more edges than the ground truth. Consequently, these graphs are far from being DAGs (in COVID-19, there is 2-cycle between economy and population size). It is also unable to distinguish between direct and indirect effect. For instance, GPT-3 and ground truth agree that population size \rightarrow spread of Covid-19 \rightarrow Covid-19 death rate, but GPT-3 also includes the edge population size \rightarrow Covid-19 death rate).

Of the remaining baselines, GES yielded the highest F1 scores, since it is tolerant to relaxation of the sufficiency assumption. Data-centric baselines were able to capture causal relationships that CATER missed. For example, in FLIGHTS, these baselines recognized that origin airport does not directly cause departure delay. However, these methods fell behind in discerning mediator variables. All failed to identify any mediator in COVID-19. Even with high F1 scores on FLIGHTS, all failed to find any mediator. This implies that not all causal relations can be discerned from data alone.

5 LIMITATIONS

Our proposed CDI system has the following limitations: First, the quality of the generated C-DAG may be affected by factors such as incorrect values in the extracted data and the performance of black-box components like the entity linker used for connecting values to entities in a KG. Second, the generated C-DAG may not capture all real-life confounding attributes, violating the unconfoundedness assumption. Common confounding variables like age or ethnicity may be absent in datasets containing information about people. Lastly, while methods for identifying confounding variables through causal DAGs are well understood, they typically rely on having access to complete causal DAGs. However, in practice, obtaining full causal DAGs can be challenging. This is particularly true in our dynamic framework, where automatically discovering a full causal DAG may not always be feasible due to the need for substantial domain knowledge in determining causal relationships.

ACKNOWLEDGMENTS

We appreciate the support of NSF Convergence Accelerator Program award 2132318, the Israeli Science Foundation (grant No. 2121/22), and the Schmidt Futures initiative.

REFERENCES

- [1] 2020. Kaggle Datasets: Covid-Data. <https://www.kaggle.com/imdevskp/coronavirus-report>.
- [2] 2020. Kaggle Datasets: Flights Delay. <https://www.kaggle.com/usdot/flight-delays>.
- [3] 2022. DBpedia. <https://www.dbpedia.org/>.
- [4] 2022. The Federal Reserve Economic Data. <https://fred.stlouisfed.org/#>.
- [5] 2022. Flights Statistics. <https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations>.
- [6] 2022. The US Federal Open Data. <https://data.gov/>.
- [7] 2023. Prototype Implementation. <https://github.com/briyoungmann/CDI/blob/main/README.md>.
- [8] Kamran Alipour, Aditya Lahiri, Ehsan Adeli, Babak Salimi, and Michael Paz-zani. 2022. Explaining Image Classifiers Using Contrastive Counterfactuals in Generative Latent Spaces. *arXiv preprint arXiv:2206.05257* (2022).
- [9] Tara V Anand, Adèle H Ribeiro, Jin Tian, and Elias Bareinboim. 2022. Effect Identification in Cluster Causal Diagrams. *arXiv preprint arXiv:2202.12263* (2022).
- [10] E. Bareinboim and J. Pearl. 2012. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*. PMLR, 100–108.
- [11] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.
- [12] Alex Bogatu, Alvaro AA Fernandes, Norman W Paton, and Nikolaos Konstantinou. 2020. Dataset discovery in data lakes. In *ICDE*. IEEE, 709–720.
- [13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [14] Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A dataset search engine for data discovery and augmentation. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2791–2794.
- [15] D.M Chickering. 2002. Optimal structure identification with greedy search. *JMLR* 3, Nov (2002), 507–554.
- [16] Xu Chu, Ihab F Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 international conference on management of data*. 2201–2206.
- [17] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2013. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal* 22, 5 (2013), 665–687.
- [18] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, Tim Finin, et al. 2010. Entity disambiguation for knowledge base population. In *Proceedings of the 23rd International Conference on Computational Linguistics*.
- [19] Mahdi Esmailoghli, Jorge-Arnulfo Quiané-Ruiz, and Ziawasch Abedjan. 2021. COCOA: COrelation COefficient-Aware Data Augmentation. In *EDBT*. 331–336.
- [20] Ronald Fagin, Alberto O Mendelzon, and Jeffrey D Ullman. 1982. A simplified universal relation assumption and its properties. *TODS* 7, 3 (1982), 343–360.
- [21] Sainyam Galhotra, Amir Gilad, Sudeepa Roy, and Babak Salimi. 2022. HypeR: Hypothetical Reasoning With What-If and How-To Queries Using a Probabilistic Causal Approach. *arXiv preprint arXiv:2203.14692* (2022).
- [22] Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics* 10 (2019), 524.
- [23] Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from Wikipedia. In *Proc. EMNLP-IJCNLP*. 2988–2999.
- [24] Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering Binary Causal Questions Through Large-Scale Text Mining: An Evaluation Using Cause-Effect Pairs from Human Experts. In *IJCAI*. 5003–5009.
- [25] Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Pottthast. 2020. Causenet: Towards a causality graph extracted from the web. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3023–3030.
- [26] V. Henderson and J.F Thisse. 2004. *Handbook of regional and urban economics: cities and geography*. Vol. 4. Elsevier.
- [27] Ihab F Ilyas and Xu Chu. 2019. *Data cleaning*. Morgan & Claypool.
- [28] Vipul Kashyap and Amit Sheth. 1996. Semantic and schematic similarities between database objects: a context-based approach. *The VLDB journal* 5, 4 (1996), 276–304.
- [29] Jihoon Ko, Yunbum Kook, and Kijung Shin. 2020. Incremental lossless graph summarization. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 317–327.
- [30] K Ashwin Kumar and Petros Efsthathopoulos. 2018. Utility-driven graph summarization. *Proceedings of the VLDB Endowment* 12, 4 (2018), 335–347.
- [31] Xueling Lin, Haoyang Li, Hao Xin, Zijian Li, and Lei Chen. 2020. KBPearl: a knowledge base population system supported by joint entity and relation linking. *Proceedings of the VLDB Endowment* 13, 7 (2020), 1035–1049.
- [32] A. Lindmark. 2021. Sensitivity analysis for unobserved confounding in causal mediation analysis allowing for effect modification, censoring and truncation. *Statistical Methods & Applications* (2021), 1–30.
- [33] R. McNamee. 2003. Confounding and confounders. *Occupational and environmental medicine* 60, 3 (2003), 227–234.
- [34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
- [35] Sameh K Mohamed, Vit Nováček, and Aayah Nounu. 2020. Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics* 36, 2 (2020), 603–610.
- [36] Fatemeh Nargesian, Erkang Zhu, Ken Q Pu, and Renée J Miller. 2018. Table union search on open data. *Proceedings of the VLDB Endowment* 11, 7 (2018), 813–825.
- [37] Alberto Parravicini, Rhicheek Patra, Davide B Bartolini, and Marco D Santambrogio. 2019. Fast and accurate entity linking via graph embedding. In *Proceedings of the 2nd Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*. 1–9.
- [38] Neil Pearce and Debbie A Lawlor. 2016. Causal inference—so much more than statistics. *International journal of epidemiology* 45, 6 (2016), 1895–1903.
- [39] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.
- [40] Judea Pearl et al. 2000. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press* 19, 2 (2000).
- [41] J. Pearl and D. Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
- [42] Babak Salimi, Johannes Gehrke, and Dan Suciu. 2018. Bias in olap queries: Detection, explanation, and removal. In *SIGMOD*. 1021–1035.
- [43] Babak Salimi, Harsh Parikh, Moe Kayali, Lise Getoor, Sudeepa Roy, and Dan Suciu. 2020. Causal relational learning. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 241–256.
- [44] Aécio Santos, Aline Bessa, Fernando Chirigati, Christopher Musco, and Juliana Freire. 2021. Correlation sketches for approximate join-correlation queries. In *Proceedings of the 2021 International Conference on Management of Data*. 1531–1544.
- [45] Alberto Santos, Ana R Colaço, Annelaura B Nielsen, Lili Niu, Maximilian Strauss, Philipp E Geyer, Fabian Coscia, Nicolai J Wewer Albrechtsen, Filip Mundt, Lars Juhl Jensen, et al. 2022. A knowledge graph to interpret clinical proteomics data. *Nature biotechnology* 40, 5 (2022), 692–702.
- [46] WS Sarle. 1990. SAS/STAT user’s guide: The VARCLUS procedure. *SAS Institute, Inc., Cary, NC, USA*, (1990), 134.
- [47] Steven G Schauer, Jason F Naylor, Michael D April, Brandon M Carius, and Ian L Hudson. 2021. Analysis of the effects of COVID-19 mask mandates on hospital resource consumption and mortality at the county level. *Southern medical journal* 114, 9 (2021), 597.
- [48] S.R Seaman and I.R White. 2013. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research* 22, 3 (2013), 278–295.
- [49] A. Sharma and E. Kiciman. 2020. DoWhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216* (2020).
- [50] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7, 10 (2006).
- [51] P. Spirtes et al. 2000. *Causation, prediction, and search*. MIT press.
- [52] Yuanyuan Tian, Richard A Hankins, and Jignesh M Patel. 2008. Efficient aggregation for graph summarization. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 567–580.
- [53] Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. Mediation: R package for causal mediation analysis. (2014).
- [54] I. Vayansky and S. Kumar. 2020. A review of topic modeling methods. *Information Systems* 94 (2020), 101582.
- [55] Marco A Wiering et al. 2002. Evolving causal neural networks. In *Benelearn’02: Proceedings of the Twelfth Belgian-Dutch Conference on Machine Learning*. 103–108.
- [56] Brit Youngmann, Michael Cafarella, Yuval Moskovitch, and Babak Salimi. 2022. On Explaining Confounding Bias. *arXiv preprint arXiv:2210.02943* (2022).
- [57] Jiaming Zeng, Michael F Gensheimer, Daniel L Rubin, Susan Athey, and Ross D Schachter. 2022. Uncovering interpretable potential confounders in electronic medical records. *Nature Communications* 13, 1 (2022), 1014.
- [58] Ning Zhang, Yuanyuan Tian, and Jignesh M Patel. 2010. Discovery-driven graph summarization. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)*. IEEE, 880–891.
- [59] Y Zhang and Z. Ives. [n.d.]. Finding related tables in data lakes for interactive data science. In *SIGMOD*. 1951–1966.
- [60] Erkang Zhu, Dong Deng, Fatemeh Nargesian, and Renée J Miller. 2019. Josie: Overlap set similarity search for finding joinable tables in data lakes. In *Proceedings of the 2019 International Conference on Management of Data*. 847–864.
- [61] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. 2019. Causal discovery with reinforcement learning. *arXiv preprint arXiv:1906.04477* (2019).