Causal Relationship over Knowledge Graphs

Hao Huang
Hao.Huang@tib.eu
supervised by Maria-Esther Vidal
Leibniz University of Hannover
TIB Leibniz Information Centre for Science and Technology
Germany

ABSTRACT

Causality has been discussed for centuries, and the theory of causal inference over tabular data has been broadly studied and utilized in multiple disciplines. However, only a few works attempt to infer the causality while exploiting the meaning of the data represented in a data structure like knowledge graph. These works offer a glance at the possibilities of causal inference over knowledge graphs, but do not yet consider the metadata, e.g., cardinalities, class subsumption and overlap, and integrity constraints. We propose CareKG, a new formalism to express causal relationships among concepts, i.e., classes and relations, and enable causal queries over knowledge graphs using semantics of metadata. We empirically evaluate the expressiveness of CareKG in a synthetic knowledge graph concerning cardinalities, class subsumption and overlap, integrity constraints. Our initial results indicate that CareKG can represent and measure causal relations with some semantics which are uncovered by state-of-the-art approaches.

CCS CONCEPTS

• **Information systems** → **Query languages**; **Data analytics**; • **Theory of computation** → **Semantics and reasoning**.

KEYWORDS

Causal inference, Knowledge graphs, Semantic data models

ACM Reference Format:
Hao Huang. 2022. Causal Relationship over Knowledge Graphs. In *Proceedings of the 31st ACM Int'l Conference on Information and Knowledge Management (CIKM '22), Oct. 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3511808.3557818

1 INTRODUCTION

Knowledge graphs (KGs) are flexible and expressive data structures that represent the convergence of data and knowledge. They include entities, attributes, and relations in a triple-based form and metadata based on an ontology, modeling the KG schema and integrity constraints. By exploiting knowledge encoded in KGs, the accuracy of multiple tasks, such as prediction, classification, and reasoning have benefited greatly [2]. However, existing approaches discover associated patterns which do not necessarily imply causality.

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10.
https://doi.org/10.1145/3511808.3557818

Both Rubin's Potential Outcome Framework [6] or Pearl's Structural Causal Model [4] are widely used for causal inference but assume (1) a pre-processed flat table where each row referred to attributes of a unit, i.e., an object subjected to a treatment and responding to the treatment with an outcome, and (2) the Stable Unit Treatment Value Assumption (SUTVA) [5] implying that one unit's attributes cannot impact attributes of others. However, constructing a flat table from KGs is not an easy task. First, the treatment and outcome may lie in different classes or relations, making it difficult to recognize the units of interest. Second, each unit can be exposed under multiple treatments or/and results in multiple outcomes due to multi-valued attributes or complex relations among entities. In addition, interference among units (violating the SUTVA) usually happens in KGs, because entities interact with each other through different relations. As a result, the outcome of one unit can be impacted by the treatments of others.

To the best of our knowledge, [1, 7, 8] are the closest works to this thesis. However, these approaches do not exploit rich semantics encoded in KGs (e.g., cardinality constraint, class subsumption and overlap, and integrity constraint) and ignore their impact on causal inference. This thesis is targeted to provide a new formalism, named causal relationship over knowledge graphs (CareKG), to represent causal relations and enable causal queries over KGs with a focus on semantics. We demonstrate the expressiveness of CareKG by comparing with [7] in a synthetic KG.

2 MOTIVATION

Consider a scenario depicted in Figure 1, which represents an ontology and a causal graph modeling the causal dependencies between attributes of the ontology classes. They include: *Movie* (with attribute *Success*), *Company* (with attribute *Revenue*), *Director* (with attribute *Experience*), *Actor* (with attribute *Fame*), and *Person* which is a super-class of *Director* and *Actor*; the overlap (noted by "*o*" in an ellipse shape) is allowed between these two sub-classes ("⊂" denotes "sub-class of"), which means that a director can also be an actor. In addition, there are three binary relations: *ActIn* (between *Actor* and *Movie*), *Direct* (between *Director* and *Movie*), and *Invest* (between *Company* and *Movie*). Moreover, a 3-ary relation *Recruit* connects three classes (*Director*, *Actor*, and *Movie*). Lastly, there are cardinality constraints, where each is formulated as $Card(C, R, (min, max))$ specifies that each entity of a class C can participate in (min, max) instances of the relation R; for example, $Card(Actor, ActIn, (1, N))$ means each actor should act in at least one movie. The *causal graph* is a directed acyclic graph where the nodes represent attributes, and the directed edges model causal relations (the red edges in Figure 1) among attributes. In our example, the causal graph represents that the *Movie*'s *Success* is affected by the *Director*'s *Experience* and

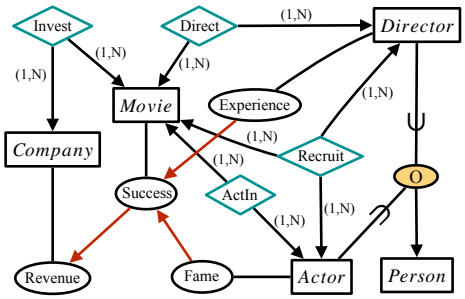

Figure 1: Ontology and Causal Model. Rectangles represent classes; ellipses represent attributes; green diamond shapes represent relations; the orange ellipse with "O" inside represents the "overlap"; the ⊂ denotes "sub-class of".

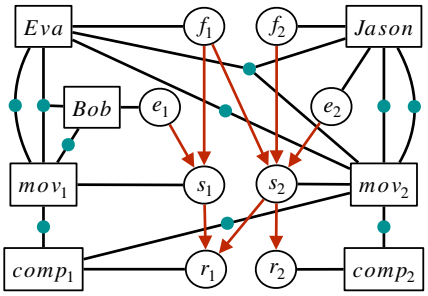

Figure 2: Knowledge Graph with causal relation over properties. Green dots represent relations among entities; red edges model the causal relations between entities' attributes.

Actor's *Fame*, and the *Revenue* of *Company* is causally depended on the *Success* of *Movie*.

Figure 2 depicts a KG using the ontology in Figure 1 as template; it represents causal relations (the red edges) in an instance-level conforming with the causal graph in Figure 1. In our example, the KG comprises two companies: $comp_1$ and $comp_2$ with revenue r_1, and r_2, respectively; two movies: mov_1 (invested by $comp_1$) and mov_2 (invested by $comp_1$ and $comp_2$) with success value, s_1 and s_2, respectively. The revenue of $comp_1$ (r_1) is causally dependent on the success of mov_1 (s_1) and mov_2 (s_2), and $comp_2$'s revenue r_2 is impact by the success of mov_2 (s_2). Moreover, the KG contains three people: *Eva* who is an actor acting in mov_1 with fame value f_1, *Bob* who is a director of mov_2 with experience value e_1, and *Jason* who is both actor and director of mov_2 with fame f_2 and experience e_2. mov_1's success (s_1) is influenced by *Bob*'s experience e_1 and *Eva*'s fame f_1, while mov_2's success (s_2) is impacted by *Eva*'s fame f_1, and *Jason*'s fame f_2 and experience e_2.

Preliminary Notations. In a KG, given a pair of cause-effect attributes X and Y, the class (or relation) of them I_X and I_Y, respectively. The *relational path* $P = [I_X, ..., I_Y]$ between I_X and I_Y is a sequence of classes and relations beginning at I_X and ending at I_Y. The *treated units* U_X are the instances of I_X, and the *response units* U_Y are the instances of I_Y. Heterogeneity between U_X and U_Y happens when U_X and U_Y are of different types. For example, considering the cause effect of *Fame* on *Success*, where $U_X = Actor$ and $U_Y = Movie$, therefore, U_X and U_Y are heterogeneous to each other. However, the traditional causal inference frameworks [4, 6] assume the *homogeneous unit* U_{XY} where $U_{XY} \equiv U_X \equiv U_Y$. To apply these causal inference frameworks over KGs, we need a perspective $\mathcal{P}$ to specify the way of constructing the unit (or object), so that the treatments from $u \in U_X$ and outcomes from $u \in U_Y$ can be recognized as features of a homogeneous unit $u \in U_{XY}$. For example, in Figure 2, we would like to know the causal impact of $X = Fame$ on $Y = Success$. Given a *relational path* $P = [Actor, ActIn, Movie]$ where $I_X = Actor$ and $I_Y = Movie$, if the perspective $\mathcal{P} = Actor$ (class), the $U_{XY} = \{Eva, Jason\}$ with two units.

Why traditional causal inference framework fails in our example? The traditional causal inference frameworks [4, 6] cannot be applied over this type of data described in Figure 2, because the heterogeneity between U_X and U_Y. For example, considering the causal impact of *Experience* on *Success* where the object of treatment is the entity of *Director*, and the object of outcome is the entity of *Movie*. In addition, the interference existed in KG violates the SUTVA, e.g., *Jason*'s movie's success (s_2) is caused by not only e_2 (*Jason*'s experience) but also *Eva*'s experience e_1.

Why the state-of-the-art approaches cannot express our example? Conquering the previous challenges, Salimi et al. [7] propose a Causal Relational Learning framework; it allows causal inference over relational data by relaxing the unit homogeneity assumption and the SUTVA assumption. They offer a formalism called CaRL (Causal Relational Language), which defines a *relational causal model* to represent the causal dependencies among attributes, and a *relational causal graph* to model the causal relation at an instance level. In addition, CaRL supports various types of causal queries over relational data.

A. Shortcomings in the *relational causal model*. The *relational causal model* is made of a set of *relational causal rules*, where each one represents a causal assumption using the form as $A[\mathbf{Y}] \Leftarrow A_1[\mathbf{X}_1], ..., A_k[\mathbf{X}_k]$ *WHERE* $Q(\mathbf{W})$, where $A, A_i (i \in [1, k])$ are attribute functions for obtaining attributes of entities or relationships, Q is a conjunctive query over a relational schema, $\mathbf{Y}$, $\mathbf{X}_i (i \in [1, k])$, $\mathbf{W}$ are sets of variables and/or constants, and $\mathbf{Y} \cup \bigcup_i \mathbf{X}_i \subseteq \mathbf{W}$. The $A[\mathbf{Y}]$, $A_1[\mathbf{A}_1]$, ..., $A_k[\mathbf{X}_k]$, and $Q(W)$ are named, respectively, the head, the body, and the condition of the rule. In our example of Figure 1, CaRL can represent the following *relational causal rules*:

$$Success[M] \Leftarrow Experience[D] \ WHERE \ Direct(D, M) \quad (1)$$

$$Success[M] \Leftarrow Experience[D] \ WHERE \ Recruit(M, D, A) \quad (2)$$

However, the condition $Q(W)$ cannot express the semantics of metadata in KGs, which will impact the way of formulating causal queries and the causal inference. Here, we summarize three types of metadata that CaRL fails to express. (1) *Class subsumption and overlap*. For example, CaRL cannot represent the causal relation between *Experience* and *Success* on a special group of directors $Director - Actor$, or $Director \cap Actor$. (2) *Cardinality constraints*. CaRL cannot specify the cardinality constraints. For example, if there is a $Card(Director, Direct, (1, 1))$ between *Direct* and *Director* instead of $Card(Director, Direct, (1, N))$ in Figure 1. Therefore, there is no interference between directors, and we cannot execute relational or isolated effect queries. (3) *Integrity constraints*. For instance, if a director is not allowed to recruit him/ herself as an actor. Therefore, when estimating the cause effect of *Experience* on *Success* via relational path $[Director, Recruit, Movie]$, the director *Jason* should not be included as a unit.

B. Shortcoming in causal query language. The CaRL causal query language supports three types of causal queries, i.e., a) the average treatment effect query; b) the aggregated response query;

and c) the causal queries under interference. However, there are several limitations considering its expressiveness when doing causal queries on KGs. Firstly, CaRL is unable to specify the perspective and the causal effect is always calculated under the perspective of U_X. As a result, CaRL supports only the causal queries from the perspective of *Director* (U_X) rather than *Movie* (U_Y) when we want to know the causal impact of *Director*'s *Experience* on *Movie*'s *Success*. Without a perspective specification, CaRL cannot differentiate the causal query $AVG\_Success[M] \Leftarrow Experience[D]?$ from the causal query $Success[M] \Leftarrow Experience[D]?$, and it answers both queries by doing causal inference over the same table (similar situation in Table 1 of [7]). In addition, CaRL cannot specify to which relational path the causal query refers. For example, the causal query $Success[M] \Leftarrow Experience[D]?$ is formulated in CaRL to request the causal effect of *Experience* on *Success*. However, in our case (Figure 1), there are two relational paths [*Director*, *Direct*, *Movie*] and [*Director*, *Recruit*, *Movie*]. Therefore, CaRL cannot distinguish the cause effect led by different relational paths. Lastly, CaRL cannot represent the cause effect considering class overlap or disjoint. For example, considering the same cause-effect pair (e.i., *Experience* and *Success*), CaRL cannot tell the difference between the causal impact from those directors who are not actors in their movies (denoted by $Director - Author$) versus the cause effect from those who act in their movies (denoted by $Director \cap Author$).

3 RELATED WORK

Causality over Relational Data. Maier et al. [3] introduce a *relational d-separation* (extended from *d-separation*) deriving conditional independence in relational data. They provide an *abstract ground graph* for representing the causal dependencies among relational variables (i.e., attributes connected by a relational path) concerning a given perspective (i.e., an entity or relation type). Relaxing two assumptions of traditional causal inferences, Salimi et al. [7] propose the CaRL language to support different causal queries on relational data. However, CaRL never mentions and lacks of specification of the perspective in its query language. As a result, the causal answering is limited to the perspective of the U_X. In addition, CaRL cannot express relational paths in their queries. Therefore, it cannot consider the causal impact from different relational paths. Finally, CaRL is not yet take into account the meaning of data, which is known as the semantics expressed by axioms in KGs (e.g., subclasses and subproperties).

Causality over Knowledge Graphs. Jaimini and Sheth [1] present CausalKG which represent causal concepts on KGs (i.e., natural direct, nature indirect, total causal effect, and mediator). CausalKG relies on a *causal ontology* and a *causal Bayesian network* (CBN) to support causal reasoning. Moreover, RDF*, an extension of the resource description framework (RDF), helps to represent causal relations in KG. However, CausalKG cannot support counterfactual reasoning which is over-promised by them, because a CBN can reach maximum the second rung (i.e., intervention) of Pearl's causal hierarchy [4]. Simonne et al. [8] present a framework for mining gradual or categorical differential causal rules, where strata describe the context of a causal rule. By comparing pairs of units, they apply a causal ratio to measure to what extent the causal relation exists. As a result, they discover causal rules indicating that the difference on treatments explains the difference on outcomes. However, this framework is limited to the causality under perspective of the target class, and cannot deal with the situation where one unit (i.e., the instance of the target class) has multiple treatments or outcomes.

In this work, we aim to propose a new formalism, named CareKG, which enables causal inference over KGs with special concern over the rich semantics from KG. These features allow CareKG to express causal relations against KGs which is out of scope of CaRL [7], because it is limited to relational data. In addition, this thesis is different from [1] by considering richer causal concepts for causal knowledge representation and introducing causal queries with attention to semantics of KGs. Finally, this thesis aims at defining causal knowledge representations and inference methods, but using more expressive formalisms than the ones extracted by [8].

4 PROBLEM DEFINITION

Problem Statement. Given (1) an ontology $\mathcal{O} = \{\mathcal{C}, \mathcal{R}, \mathcal{A}\}$, where $\mathcal{C}$ is the class set; $\mathcal{R}$ is the relation set, each $R \in \mathcal{R}$ is a relation which connects n classes, R can be a self-relation ($n = 1$), a binary relation ($n = 2$), an n-ary relation ($n > 2$); $\mathcal{A}(I)$ is the attribute set of I, where $I \in \mathcal{C} \cup \mathcal{R}$; (2) a knowledge graph $KG = \{V, E, T\}$, where V is a set of constants, including entities and literals; E is a set of properties; and T is a set of triples with form of (s, p, o) where $p \in E$ and $s, o \in V$; and (3) A set of axioms ζ include logical rules that express concepts' definitions (e.g., subclass or subproperty) and integrity constraints (e.g., cardinality).

Research Questions.

Q1: How to represent causal relations and knowledge in KGs?

Q2: How to formulate a query language to support various causal queries over KGs with a special concern of semantics?

Q3: How to estimate the cause effect efficiently in a large-scale KG?

Solutions. We currently offer sub-solutions (i) and (ii) responding to the research questions **Q1** and **Q2**, respectively. Further investigation is needed to answer **Q3**.

(i) We propose an *ontological causal model* (OCM) to encode causal assumptions (i.e., causal relations among attributes) at a concept level. It is composed of a set of *ontological causal rules* with a form of $Y[I_Y] \Leftarrow X[I_X]$ $WHERE$ $P(I_X, I_Y)$, $CTX(I_X, I_Y)$, where X and Y are the attributes regarded as treatment and outcome; I_X and I_Y are, respectively, the class or relation of X and Y. The expression $P(I_X, I_Y) = [I_X, ..., I_Y]$ specifies a relational path between I_X and I_Y with length of m; $CTX(I_X, I_Y)$ is an optional conjunction of conditions over property paths rooted from I_X and I_Y, offering the context where the causal relation holds. Additionally, we propose a *grounded causal KG* to represent causal knowledge at an instance level, which includes all facts from KG, and the causal relations between attributes of instances (entities or relations). All these causal relations are encoded by a set of *grounded causal rules*, where each is formulated using an *ontological causal rule* as a template without the condition clause and replaces I_X and I_Y with instances.

(ii) We provide a query language CareKG allowing different causal queries considering the perspective and semantics of KGs. A query is formalized as FUN($Y[I_Y]$) $\Leftarrow$ FUN($X[I_X]$)? FROM PATH P, UNDER PERSPECTIVE $\mathcal{P}$, SUBJECT TO $\langle axiom \rangle$, WHEN $\langle cnd \rangle$, where P is a relational path from I_X to I_Y; $\mathcal{P} \subseteq_s P$ ($\subseteq_s$ denotes "subsequence of") specifying the way of unit constructing (each instance

(value %)	Path	Perspective		
		Director	Movie	Direct / Recruit
CareKG	P_1	44.12 (± 5.47)	78.70 (± 4.77)	41.81 (± 4.22)
	P_2	46.97 (± 5.55)	71.42 (± 5.82)	45.02 (± 0.68)
CaRL	P_1	44.07 (± 5.47)	-	-
	P_2	47.25 (± 5.56)	-	-

Table 1: ACE of *Experience* on *Success* lead by relational path P_1 and P_2 represented by "mean (± confidence interval)". "-" means ACE of the relevant causal query cannot be answered.

of $\mathcal{P}$ is a unit), which allows a single perspective (when $len(\mathcal{P}) = 1$) and multi perspectives (when $len(\mathcal{P}) > 1$); $\langle axiom \rangle$ are optional axioms specifying a KG's semantics over the selected units using logical rules. Multiple axioms are considered, such as cardinality constraint, class disjoint, and integrity constraints. The optional condition $\langle cnd \rangle$ defines the peers of a unit and specifies a causal query when interference happens, for example, peer, isolated causal effect. FUN(.) is an optional function for processing attributes, e.g., aggregating multiple attribute values of a unit into one; redesigning the attribute values (i.e., discretization or replacement).

5 RESULTS SO FAR

We demonstrate the expressiveness of CareKG compared with CaRL on a synthetic KG, which includes 300 directors, 582 movies, 3,043 actors. Each director is allowed to direct randomly one to 20 movies, and recruit at least one actor for a movie; one movie is directed randomly by one to 5 directors, and has a random number of actors (from 20 to 100); one actor can be recruited by multiple directors of a movie, and acts in one to 20 movies. The *Success* of a movie M is generated by a function $Success[M] = 0.5 \times Mean(Experience[D]) + 0.5 \times Mean(Experience[D']) + 0.5 \times Mean(Fame[A])$ where director D directs movie M, director D' directs movie M also acts in movie M as a actor, and actor A acts in movie M.

We experience on average cause effect (ACE) between *Experience* and *Success*. There are two relational paths $P_1 = [Director, Direct, Movie]$, $P_2 = [Director, Recruit, Movie]$ between *Director* and *Actor* (*Director* and *Actor* share some entities due to overlap). We show limitations of CaRL compared with CareKG in two experiments. From experiment one, we want to know the cause effect from different relational paths and under different perspectives. From experiment two, the relational path and the perspective is chosen to be P_1 and *Director*, and we want to know the cause effect from different groups of *Director*. CaRL formulates the *relational causal rules*, rule (1) and rule (2), and the relevant causal query $Success[M] \Leftarrow Experience[D]$? for both experiments. In contrast, CareKG formulates two *ontological causal rules*: $Success[M] \Leftarrow Experience[D]$ $WHERE$ $[Director(D), Direct(D, M), Movie(M)]$ and $Success[M] \Leftarrow Experience[D]$ $WHERE$ $[Director(D), Recruit(D, M, A), Movie(M)]$; and the causal query can be $Success[M] \Leftarrow Experience[D]$? $FROM\ PATH\ P, UNDER\ PERSPECTIVE\ \mathcal{P}, SUBJECT\ TO\ \langle axiom \rangle$, where P can be P_1 or P_2; $\mathcal{P}$ can be *Director*, *Movie*, *Direct* (for P_1), or *Recruit* (for P_2); $\langle axiom \rangle$ can be I_X *is Director* (all directors), I_X *is Director and Actor* (directors who are also actor), or I_X *is Director not Actor* (i.e., directors who are not actor).

Table 1 reports the ACE (values represented under percentage, "value %") results from relational paths P_1 and P_2 under perspective of *Director*, *Movie*, *Direct* (for P_1), or *Recruit* (for P_2). CareKG can differentiate causal queries considering different relational paths, and different perspectives. However, CaRL supports only the default perspective U_X (i.e., *Director*). Therefore, it cannot answer causal queries from other perspectives (denoted by "-"). In addition, CaRL is not able to specify the relational path in its causal query. Thus, it needs extra effort to materialize in two different attributes the values of *Success* according to the relational paths P_1 and P_2. Results of experiments two (Table 2) show that CareKG can support semantics of class subsumption, disjoint, and overlap expressed in $\langle axiom \rangle$ (demonstration of other axioms is left for further works). Consequently, CareKG can distinguish causal impact from different director groups. However, CaRL can only answer the causal query regarding all directors.

(value %)	SUBJECT TO $\langle axiom \rangle$: I_X is		
	Director (default)	Director ∩ Actor	Director - Actor
CareKG	44.12 (± 5.47)	67.75 (± 10.15)	39.73 (± 6.11)
CaRL	44.07 (± 5.47)	-	-

Table 2: ACE of *Experience* on *Success* in perspective of *Director* and lead by relational path P_1. ACE from different director groups *Director*, $Director \cap Actor$, and $Director - Actor$.

6 CONCLUSIONS AND FUTURE DIRECTIONS

In this work, we propose a new formalism CareKG, enabling causal relation learning and causal inference by exploiting the rich semantics of KGs (e.g., integrity constraints, class overlap and disjoint, and cardinality constraints). We demonstrate that the expressiveness of CareKG is more powerful than the state of the art in a synthetic KG. We consider extending CareKG to perform causal inference over multiple types of attributes, and improving efficiency of causal inference over large-scale KGs as future works.

ACKNOWLEDGMENTS

This work is funded by the Leibniz Association in the program "Leibniz Best Minds: Programme for Women Professors", project TrustKG-Transforming Data in Trustable Insights with grant P99/2020.

REFERENCES

[1] Utkarshani Jaimini and Amit Sheth. 2022. CausalKG: Causal Knowledge Graph Explainability using interventional and counterfactual reasoning. *arXiv preprint arXiv:2201.03647* (2022).
[2] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems* (2021).
[3] Marc Maier, Katerina Marazopoulou, and David Jensen. 2013. Reasoning about independence in probabilistic models of relational data. *arXiv preprint arXiv:1302.4381* (2013).
[4] Judea Pearl. 2009. *Causality.* Cambridge university press.
[5] Donald B Rubin. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American statistical association* 75, 371 (1980), 591–593.
[6] Donald B Rubin. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* 100, 469 (2005), 322–331.
[7] 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.
[8] Lucas Simonne, Nathalie Pernelle, Fatiha Saïs, and Rallou Thomopoulos. 2021. Differential Causal Rules Mining in Knowledge Graphs. In *Proceedings of the 11th on Knowledge Capture Conference*. 105–112.