Thesis

Konstantin Kueffner

August 6, 2020

Contents

Chapter 1

Introduction

1.1 Preliminaries

1.1.1 General

Exogenous and Endogenous Variables, denoted by $\mathcal{W}$, can be used for describing the state of the world. However, often it is useful to partition those variables into two groups, i.e. exogenous and endogenous variables. *Exogenous* variables, denoted by $\mathcal{U}$, are variables whose values are determined by factors outside of the model. Whereas *endogenous*, denoted by $\mathcal{V}$ variables are determined by the values of exogenous variables based on the rules relating variables within the model.

1.1.2 Graph Theory

A directed graph G is the pair (V, E) with V being a set of vertices and $E \subset V \times V$ being a set of edges. For $v, w \in V$ v is a *direct predecessor* of w iff there exists $(v, w) \in E$; y is a direct successor of w iff $(w, v) \in E$. For $v \in V$, $\mathcal{N}^+(v)$ is the set of direct predecessors of v; $\mathcal{N}^-(v)$ is the set of direct successor of v; $\deg^+(v) := |\mathcal{N}^+(v)|$ is the in-degree of v; $\deg^-(v) := |\mathcal{N}^-(v)|$ is the out-degree of v. A vertex $v \in V$ is called a sink *sink* if it has no successors, i.e. $\mathcal{N}^-(v) = \varnothing$ and a *source* if it has no predecessors, i.e. $\mathcal{N}^-(v) = \varnothing$.

Centralities

1.1.3 Neuron Diagrams

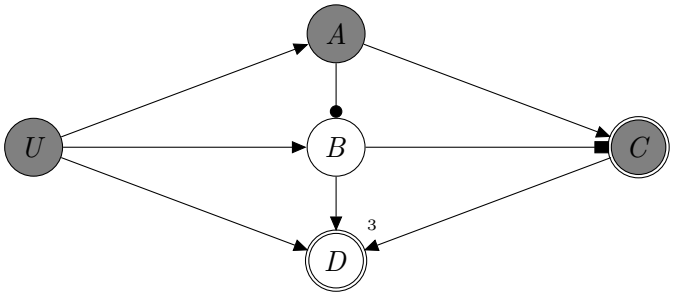
In their simplest form a neuron diagram can be understood as a directed acyclic graph, where each vertex, called neuron, can either fire or not, often indicated by its color. A neuron can be either *exogenous*, i.e. it has no incoming edges, or *endogenous*, i.e. it has at least one incoming edge. Moreover, edges between neurons can also be separated into two categories, i.e.

stimulating edges and *inhibiting* edges, often distinguished through having a triangle and respectively a circle as arrow head. In its simplest form such neuron diagrams follow a fairly straight forward semantic. That is, while it is externally specified whether an exogenous neuron fires of not, an endogenous neuron fires if and only if it is stimulated by at least one firing neuron, and inhibited by zero firing neurons Hitchcock 2009; Erwig and Walkingshaw 2010; Baumgartner 2013.
Unfortunately using only one kind of neuron is insufficient to capture many examples, e.g. encoding a conjunction is already difficult. Hence, as for example demonstrated in Hitchcock 2009 and Baumgartner 2013, alternative, more complicated neurons can be introduced. For example, one could consider a stubborn neuron that only fires, if all or some of its predecessors fire as well, or a neuron that only fires if the number of stimulating inputs is greater that the number of inhibiting inputs.
Using neuron diagrams as a formalism to encode causal structures is particularly ubiquitous in the philosophical literature. Due to their graphical nature, they provide a rather intuitive method of representation of causal dependencies for the small scale examples common in the literature. Although, providing a grater vocabulary as causal diagrams, this simplicity naturally restricts this language in its expressivity, e.g. in their common form they can not encode causation by omission (see Subsection **??**) Their use was criticised by Hitchcock Hitchcock 2009 on similar ground. That is, it is their failure to encode complex relationships between variables, that makes him an advocate for the use of structural equation, commonly used in causal models, as the default method of representation. Although, acknowledged in Erwig and Walkingshaw 2010 they justify the use of neuron diagrams by citing their simplicity. As already mentioned in Section **??**, in (Hall 2007) Hall criticises the structural equations approach, especially due to its failure to distinguish between default and deviant behaviour. Moreover, while acknowledging their value, he perceives their status as inflated, favouring neuron diagrams instead. In particular, he endorses them not only due to their simplicity, but also due to their ability to encode a default/deviant distinction. (Baumgartner 2013; Erwig and Walkingshaw 2010; Beckers and Vennekens 2016).
This simplicity, the restriction to binary accounts of causation only and the fact that none of the discussed definitions uses neuron diagrams as their modelling language, is the motivation behind the use of a neuron diagram variant as the universal modelling tool for introducing examples in Chapter **??**. As some of the examples require special neurons, it seems sensible to introduce a simple graphical definition of neuron diagrams that differs slightly from the versions commonly used in the literature.

Definition 1.1.1. A *neuron graph* is a directed acyclic graph $\mathcal{G} \coloneqq (\mathcal{W}, \mathcal{E})$. $\mathcal{W}$ is a set of neurons. Every neuron is associated/labelled with a variable.

An exogenous neuron is a neuron with no incoming edges, while an endogenous neuron must have at least one incoming edge.
Neuron can be either active or not. If a neuron is active, it will be coloured gray. By default all endogenous neurons are considered to be inactive.
Every endogenous neuron has a trigger threshold that indicates how many signals are required for the neuron to activate. An endogenous neuron with a single (double) border requires a single (two) signals, any neuron with a higher inertia has a double border and is annotated with a number.
There are three kinds of relations. Firstly, stimulating edges, indicated by an arrow head, stimulate the target neuron, if the source neuron is active. Secondly, inhibiting edges, indicated by a circle, prevent the target neuron to fire, if the source neuron is active. Thirdly, negating edges, indicated by a square, stimulate the target neuron, if the source neuron is not active.
A *neuron diagram* is a neuron graph with an assignment specifying the status of all exogenous variables.

The most unorthodox choice made in Definition 1.1.1 was to add the negating edge to the language. While not necessary, it allows for a cleaner modelling of some examples discussed in Chapter **??**. However, to get accustom to this informal definition consider the following structure.



The structure depicted below is a neuron diagram. U is an exogenous neuron, while all other are endogenous. A and B fire on a single stimuli, C requires two and D needs 3 stimuli. Since U is active, and is connected with A through a stimulating edge A receives a single stimuli, which in this case is sufficient for A to fire. Now, with A being active, B being connected to A via an inhibiting edge can not be active, even though it received a stimuli from U. C receives one stimuli from A and another from B, as it is connected via a stimulating edge to the prior and a negating one to latter. Finally, with U and C being the only stimulants for D the necessary threshold is not reached. Hence, D is not active.

Although not necessary in this work it may be of interest that Erwig and Walkingshaw 2010 properly formalised (and generalised) neuron diagrams. To do so they destination between *neuron graphs* capturing an abstract neuron structure and *neuron diagrams* representing the execution of neuron

graphs for some set of inputs. For the subsequent definitions, let $\mathbb{F}_{\mathbb{B}}^{k,m} := \{f \mid f : \mathbb{B}^k \to \mathbb{B}^m\}$ be the set of functions mapping k boolean inputs to m boolean outputs. Let $\mathbb{F}_{\mathbb{B}} := \bigcup_{k>0} \mathbb{F}_{\mathbb{B}}^{k,1}$ be the set of all boolean functions.

Definition 1.1.2. A *neuron graph* is a directed acyclic graph $\mathcal{G} := (\mathcal{W}, \mathcal{E}, \mathcal{F}, \pi)$, where $\mathcal{W}$ is the set of neurons and can be partitioned into endogenous variables $\mathcal{V}$ and exogenous variables $\mathcal{U}$; $\mathcal{E}$ is a set of edges; $\mathcal{F} := \{F_v \mid \forall v \in \mathcal{V}\ F_v : \mathbb{F}_{\mathbb{B}}^{\deg^+(v)} \to \mathbb{B}\}$ is a set of boolean functions; for $v \in \mathcal{W}$ let $\pi(v) := (v_1, \ldots, v_n)$ an ordering of $\mathcal{N}^+(v)$.

Moreover, from this one can build a neuron diagram by taking a neuron graph and providing an assignment of the exogenous variables.

Definition 1.1.3. Let $\mathcal{G} := (\mathcal{W}, \mathcal{E}, \mathcal{F}, \pi)$ be a neuron graph, then $\mathcal{D} := (\mathcal{G}, \sigma)$ is a *neuron diagram* and $\sigma : \mathcal{U} \to \mathbb{B}$ being an assignment mapping exogenous variables to binary values.

To evaluate such a neuron diagram, a firing semantic is required. Intuitively, an endogenous neuron v fires, if and only if the assigned boolean function evaluates to 1 given the values of its predecessors ordered based on the ordering $\pi(v)$. Being assigned a constant function and having no predecessors, this recursive process ends at an exogenous node.

Definition 1.1.4. Let $\mathcal{D} := (\mathcal{G}, \sigma)$ be a neuron diagram, then $\mathcal{I}$ is defined such that $\forall v \in \mathcal{W}$ and for $\pi(v) = (v_1, \ldots, v_n)$

$$\mathcal{I}(v) := \begin{cases} \sigma(v) & \text{if } v \in \mathcal{U} \\ F_v(\mathcal{I}(v_1), \ldots, \mathcal{I}(v_n)) & \text{if } v \in \mathcal{V} \end{cases}$$

(As a shorthand let $\mathcal{I}(v) := \sigma_v(\mathcal{I}(v_1), \ldots, \mathcal{I}(v_n))$.)

This formalisation is of note, as it was used in (Erwig and Walkingshaw 2010) to create of yet another definition of token causality, which unfortunately was not captured by the methodology outlined in Chapter 5.

1.2 Approaches to Causation

According to Beebee, Hitchcock, and Menzies 2009 there are standard and alternative approaches to causation. The prior including regularity theories, counter-factual theories, probabilistic theories, causal process theories and agency interventionist theories, while the latter includes theories about causal power and capacities, an anti-reductionist approach, the field of causal modelling, an approach requiring the existence of causal mechanism and one that embraces pluralism. Moreover, in Joseph Y Halpern 2016a distinguishes between two separate notions of causality, i.e. type (or general) causality and token (or actual) causality.

1.2.1 Type and token causality

The classification of causality into type and token causality is rooted in the metaphysical distinction between types and tokens, which is used to differentiate a general sort of thing and its particular occurrence Wetzel 2018.

Example 1.2.1 (Hausman 2005)**.** Consider the statement

> Rose is a rose is a rose is a rose.

How many different words does this sentence have. Depending on what one may understand as word, this sentence contains three or ten different words. In the prior, the word-types of the sentence are counted, while in the latter the word-tokens are counted.

In a similar vain one can distinguish two (possibly distinct) notions of causality. *Type causality*, is concerned with forward looking statements such as "smoking causes lung cancer", granting their wielder some predictive capabilities. Hence, establishing type causality is often the pursuit of scientific enquiry. However, often type-causal relations do not establish a strong causal connection, but rather a causal tendency. Meaning, while smoking may cause lung cancer, it is not necessary the case that a smoker will develop lung cancer, thus one may be better suited with the statement "smoking tends cause lung cancer". By contrast, if one wants to establish that the act of smoking caused lung cancer in a particular person, one speaks of token causality. Meaning *token causality* tries to establish a connection between events that explains a certain outcome arose, thus it tends to be backwards looking. Unfortunately, it does not seem clear whether those two notion of causation are distinct; whether type causation is merely a generalisation token causal relations, which are assumed to be fundamental, whether token causation is merely an instantiation of type level laws, which are considered as the fundamental element; whether type and token causality are simple different expression of a singular causal relation. For example, as will be observed in Chapter **??**, Halpern defines token causality in terms of type causal statements Hausman 2005; Joseph Y Halpern 2016a.
The debate of what constitutes token or type causality can be extended to variables. Within this context, there is a clear difference in considering relationships between variables and relationships between variable values. Although debated, one can classify the prior as a type-level relation, due to the fact that variables are not particulars but properties. An argument which attributes the latter as the only token-level relation Hausman 2005. For example, many of the formalisms using structural equations to token capture causality according to counterfactual tradition internalise this distinction as follows. For a given causal model of the world (see causal models XXXXX), a variable X is a type-level cause of a variable Y, if and only if there is some

state of the model for which an intervention on X would change the value of Y. An intervention on X, being an external change to the value of X ceteris paribus all other variables. Whereas, the value x of X is an actual cause for the actual value y of Y (wrt. the model), if and only if an intervention setting X to x' ($x \neq x$), with other variables in the model held fixed by interventions at some combination of permissible possible values, would result in Y being y' ($y \neq y'$). What constitutes permissible being subject of discussion and particular for each formalism **Weslake2015partialtheory**.

1.2.2 Traditions

1.3 Applications

It is vital to build an adequate repository of papers upon which this survey shall draw upon. Inherent in this construction is the tension between scope and depth, as all else equal, one compromises the other. Positioning a survey on this spectrum in a justifiable manner is rather difficult. To alleviate this tension, this section is concerned with providing a broad overview over a wide arrange of formalisms, while Chapter **??** discusses and analysis a narrow selection of those formalisms. The objectives of this section are accomplished by systematically constructing a literature database. This database is analysed to obtain a structural overview of the literature, allowing for an educated selection of relevant formalisms for the subsequent discussion in Chapter **??**.

Chapter 2

The Causation Literature Landscape: A rough Picture

The primary objective of this chapter is to draw a rough picture of the literature landscape surrounding the concept of causality in the context of computer science and logic. This includes a heuristic grouping of authors into research communities, pointers to particularly relevant researchers and most importantly a collection of publications that provide significant contributions in the quest of illuminating and formalising the enigma that is causality. Intersecting with a wide range of subjects, e.g. philosophy, statistics, computer science, law, natural science, social sciences and more, as well as stretching over centuries, e.g. being already discussed by Hume in the 18th century, inquiries into causality are have produced an incredible the wealth of literature. Hence, to remain within a reasonable scope, it is of utmost importance to rely on a properly defined methodology and on a suitable set heuristics, to navigate this vast ocean of literature. That is, Section 2.1 will provide a detailed characterisation of the publication collection process, as well as introduce and justify the methods used to identify relevant literature and their authors among the collected publications. The results of this voyage are then discussed in Section 2.2.

2.1 Methodology

This section provides a detailed description of the methodology used to identify, collect and analyse the computer science, logic and philosophy literature surrounding causality. Firstly, data collection. The data, i.e. publications, are collected using a snowball search strategy. This requires a detailed characterisation of the set of publications from which snowball steps are conducted, as well as an adequate description of how and when those steps are employed. This information is provided in Subsection 2.1.1. Moreover, this subsection provides a detailed description of the publicly available database

used to store the meta information of the collected publications. The purpose of which is to provide other researches access to the constructed snapshot of the literature and to enhance transparency. Lastly, the methods used for the analysis of the collected data are discussed in Subsection 2.1.2.

2.1.1 Data Collection

The methodology underlying this systematic literature review employs a snowball search strategy. In general, according to (Wohlin 2014) any snowball search strategy should start by characterising an appropriate initial set of publications, i.e. the start set, which is then iteratively expanded by either forward or backward snowballing until a desirable final set of publications is obtained. Furthermore, according to (Wohlin 2014) this start set should satisfy the following criteria.

- The start set should cover a diversity of communities.
- The number of papers in the start set should not be too small.
- The number of papers in the start set should not be too big.
- The start set should cover several different publishers, years and authors.
- The start set ought to be formulated from keywords (and their synonyms) in the research question.

Moreover, they state that any snowball step on a given set of publications consists of both forward and backward snowballing. The latter, adds all relevant references from all unprocessed publications to the set of publications. By contrast, the prior leverages modern technologies, such as Google Scholar to identify every relevant publication that references any unprocessed publication in the provided set (Wohlin 2014).
Using this as a template, the actual methodology is constructed as follows. Firstly, the objectives that ought to be satisfied by the snowball search strategy are made explicit. In this particular case those objectives are to

- focus on token causality publications;
- focus primarily on publications related to computer science and artificial intelligence;
- focus secondarily on publications related to philosophy or law;
- focus on publication that approach causality with sufficient formality;
- focus on logic and rule based approaches to causality;

- focus on the recent literature, i.e. publications between 2010 and (early) 2020.

Being a snowball search, the growth rate of the publications to consider is exponential. Hence, to serve the outlined objectives it is vital to construct a starting set that provides a sufficient strong directive. Since the primary focus is to remain within the greater context of computer science, logic and (symbolic) artificial intelligence, the start set construction is initiated by considering all articles from

- Journal Knowledge-Bases Systems (KBS)
- Journal Artificial Intelligence (AI)
- Journal Artificial Intelligence and Law (AI&Law)
- International Joint Conferences on Artificial Intelligence Organization (IJCAI)

that where published between 01.2017 and 3.2020. Focusing on such recent publications, should serve the recency bias established in the methodology's objectives. The collected publications are subsequently preprocessed using a simple key word search. That is the first necessary condition for a publication to be in the start set is

- that its title contains a string starting with the character sequence *"caus"* or
- its abstract contains a string starting with the character sequence *"causal"*.

Let $\mathcal{S}_0$ be the subset of all collected publications, that satisfy this criteria. To focus on logic and rule based approaches, all publications that are deemed irrelevant under closer inspection or are inaccessible will be removed. The classification as relevant is done based on a list of soft criteria. By satisfying positive criteria the publication increases its chance of being deemed relevant, satisfying negative ones decreases its chance, and criteria marked by "$*$" are necessary.

$*$ Does the publication discuss causality or any related concept, e.g. counterfactuals?

$+$ Does the publication engage with the philosophical aspects of causality?

$+$ Does the publication try to formalise causality using logic (or another formal language)?

$+$ Does the publication's title explicitly mention logic and/or causality?

+ Does the publication is discuss token causality (see Section XXX)?

– Does the publication discuss causality in the context of machine learning?

– Does the publication discuss causality in a highly informal manner?

– Is the publication a book?

To explain the snowballing step, some general notation must be introduced. Let $\mathcal{X}$ be some set of publications. Then $\mathcal{X}^c$ is the set of publication deemed relevant by the previously stated criteria. Furthermore, let $\mathcal{X}^r$ be the set of publication deemed relevant by the previously stated criteria, which are published after (and including 2010). Additionally, let $\pi^-(\mathcal{X})$ be the set of accessible publications, not contained in $\mathcal{X}$, that are obtained by backward snowballing on $\mathcal{X}$. Moreover, let $\pi^+(\mathcal{X})$ be the set of accessible publications, are not contained in $\mathcal{X}$, which are obtained by forward snowballing on $\mathcal{X}$ using Google Scholar between 16.04.2020 until 20.04.2020.
Utilising this notation, let $\mathcal{S}_0^r$ be the start set of this snowball search. From there, a variation of forwards- and backwards-snowballing steps are applied to construct the set $\mathcal{S}$. That is,

- $\mathcal{S}_{-1} := \pi^-(\mathcal{S}_0^r)$;
- $\mathcal{S}_{-2} := \pi^-(\mathcal{S}_{-1}^r) = \pi^-(\pi^-(\mathcal{S}_0^r)^r)$;
- $\mathcal{S}_{+1} := \pi^+(\mathcal{S}_0^r)$;
- $\mathcal{S}_{+1-1} := \pi^-(\mathcal{S}_{+1}^r) = \pi^-(\pi^+(\mathcal{S}_0^r)^r)$;
- $\mathcal{S}_{+2} := \pi^+(\mathcal{S}_{+1}^r) = \pi^+(\pi^+(\mathcal{S}_0^r)^r)$;
- $\mathcal{S}_{+2-1} := \pi^-(\mathcal{S}_{+2}^r) = \pi^-(\pi^+(\pi^+(\mathcal{S}_0^r)^r)^r)$;

and finally, $\mathcal{S} := \mathcal{S}_0 \cup \mathcal{S}_{-1} \cup \mathcal{S}_{-2} \cup \mathcal{S}_{+1} \cup \mathcal{S}_{+1-1} \cup \mathcal{S}_{+2} \cup \mathcal{S}_{+2-1}$.
Before moving on, a small discussion about the single most undesirable property of this process. That is, in an ideal world the methodology would provide a perfectly reproducible algorithm that reliably and deterministically produces the same set of publications on each execution. Unfortunately, this property can not be satisfied by the constructed methodology, as any employment of forward snowballing introduces variability into the system. Considering the importance of forward snowballing to identify the most recent literature, foregoing the application of this tool in the construction of $\mathcal{S}$ would not have been feasible. Having already parted with this desirable property, conditions that further prohibit reproducibility, i.e. the soft categorisation of relevance and the removal of inaccessible publications, could be added without concerns.

The dataset constructed using this methodology is publicly available (including a Python script for easy access). It is stored in a SQLite-database and was constructed using SQLAlchemy. This database can store publications, authors, venues and tags. Its structure is depicted in Figure 2.1 as an ER-diagram using Chen notation (P. P.-S. Chen 1976).

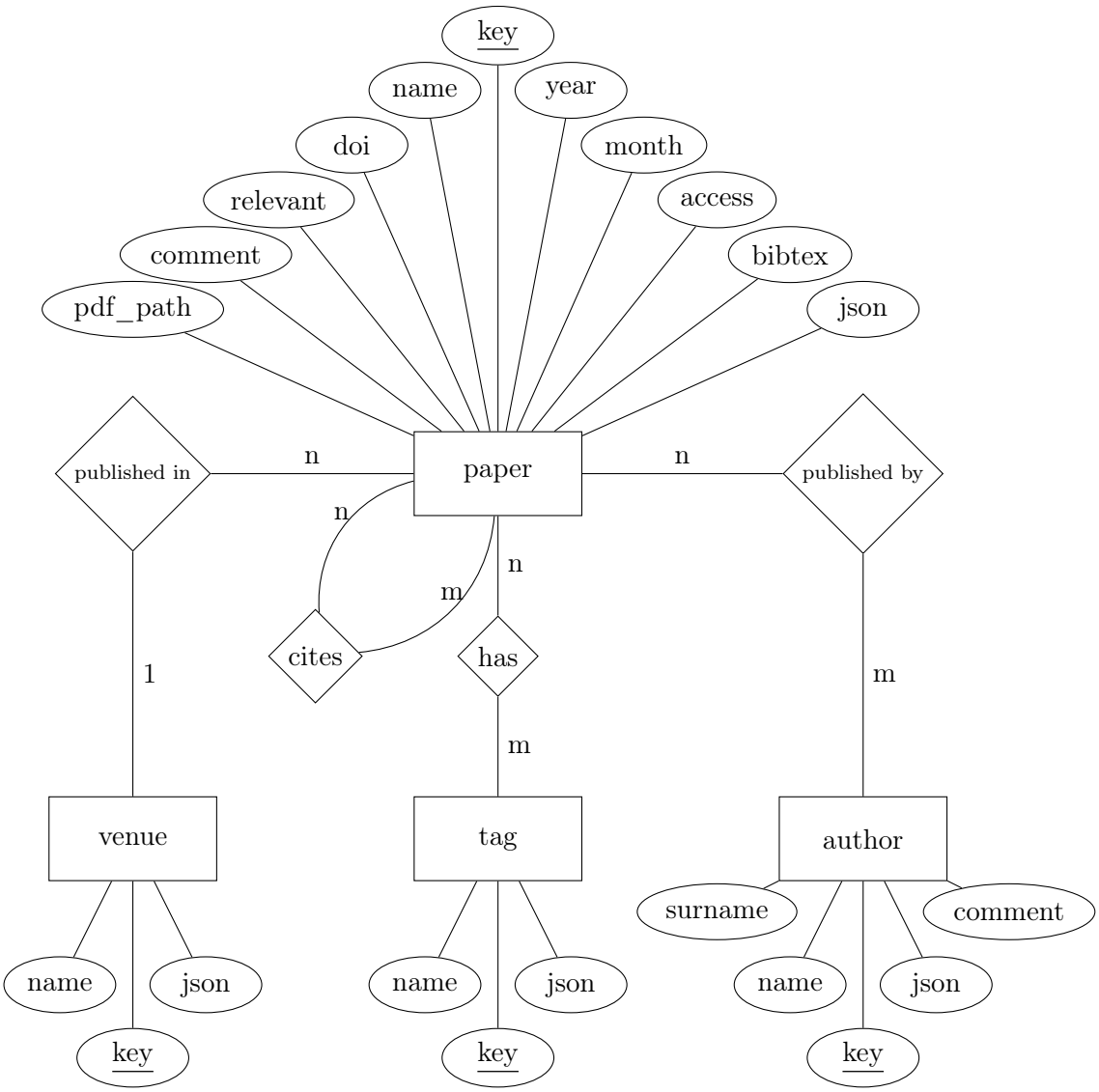


Figure 2.1: ER-Diagram of the database.

The table "Venue" stores platforms that publish research, e.g. journals such as AI or conferences such as IJCAI. This table is rather sparse, having only a column for the names of the venues and a column reserved for a JSON-string

in case additional data must be stored. Similar in structure is the "Tag" table. It's intended purpose is to store tags that can be used to further categorise publications, e.g. they are used to identify which publications were added to the database at which snowballing step. Slightly more complex is the "Author" table, which contains an additional column for surnames, as well as a column reserved for comments. The last column allows one to differentiate authors with identical name. By far the most extensive one is the Table "Paper". Firstly, it provides columns for basic information such as the title, the DOI, the publication year and month, as well as its BibTeX-string. Moreover, it contains columns to track whether a publication was accessible and whether it is deemed relevant. Furthermore, for easy access there exists also a column storing the local path to the pdf-file of the publication. Lastly, columns for comments and for storing additional data in JSON-format are provided as well.
Those table relate to each other as follows. Firstly, the table "Venue" is connected to the table "Paper" via an 1:n-relation. Secondly, the "Author"- and the "Paper"-table are in an m:n-relation. The same holds true for the "Tag"- and the "Paper"-table. Thirdly, to store the citation relations between publications, the table "Paper" is in an m:n-relation with itself.
During the data collection, publications will be assigned one of seven tags. Those are `0`, `-1`, `-2`, `+1`, `+1-1` `+2` and `+2-1` and indicate membership to the respective sets constructed during the snowballing process.

2.1.2 Data Analysis Methods

The data analysis will proceed according to the following steps. The first, is the construction of several graphs using the stored citation relation and the co-authorship relation. This is followed by a heuristic detection of research communities obtained through the application of a community detection algorithm. Following this, a combination of centrality measures and publication counts will be used to identify the most prominent authors in this field. Lastly, a similar approach will be employed to identify it's most significant publications.
Considering the collected data it is possible to construct several graph structures. The first two are constructed such that their vertices represent publications and their edges correspond to citations, i.e. each edge (A, B) implies that there exists a publication A that references another publication B. The only difference between those two graphs, called $\mathcal{G}_t$ and $\mathcal{G}_p$, is that the vertex set of $\mathcal{G}_t$ is $\mathcal{S}$ and that of $\mathcal{G}_p$ is $\mathcal{S}^r$, i.e. it contains only those publication that were deemed relevant and that have been published from 2010 onwards. By construction, both of them are directed[1] Of particular relevance is $\mathcal{G}_p$, also called "publication graph", because it serves as the bedrock of all subsequent

[1] The graph in question does contain cycles. This is due to the fact that sometimes not yet publicised works are cited, which contain a reference back to the original publication.

analysis. That is, it is used to construct a set of important publications from which all formal languages, definition and examples discussed in subsequent chapters are extracted. While it is tempting to use $\mathcal{G}_t$ for such a task, succumbing it is ill advised. Because, by construction, only the publications in S that were published after (and including) 2010 contain a complete mapping of their citations. Hence, using $\mathcal{G}_t$ for such an important task is undesirable, as this would obviously distorts the relative relevance between the analysed publications. However, a quick analysis of $\mathcal{G}_t$ will be presented to provide a rough understanding of the literature published before 2010. The other three graphs, encode information about the authors and the relationships between them. Starting with the "author graph" $\mathcal{G}_a$. It encodes the citations between authors, rather than between publications. Every vertex in $\mathcal{G}_a$ represents an author, who has at least one publication represented in $\mathcal{G}_p$. Every edge in $\mathcal{G}_a$ between an author A and an author B indicates that there exists at least one publication of A that cites at least one publication of B (with respect to $\mathcal{G}_p$), while its weight represents the frequency of this occurrence. To account for multiple citations, as well as self-referential behaviour, $\mathcal{G}_a$ must a be a directed, weighted graph containing loops. The second, $\mathcal{G}_c$ or "collaboration graph", encodes the collaborations between authors. That is, while its vertex set is identical to $\mathcal{G}_a$'s vertex set, an edge in $\mathcal{G}_c$ between an author A and an author B, indicates that A and B co-authored a publication in $\mathcal{G}_p$, thus the weight of an edge represents how often they collaborated on publications in $\mathcal{G}_p$. The last, $\mathcal{G}_m$ or "merged graph" is simply a merger of $\mathcal{G}_a$ and $\mathcal{G}_c$, where each undirected edge is replaced by two opposing directed edges (in the case of duplicate edges, their weights are summed up). To summarise, $\mathcal{G}_t$ and $\mathcal{G}_p$ are a directed graphs; $\mathcal{G}_a$ and $\mathcal{G}_m$ are directed, weighted graphs containing loops and $\mathcal{G}_c$ is an undirected, weighted graph.
The as a whole the constructed graphs are subjected to two different kinds of information extraction processes. The first uses the community detection algorithm to identify research communities, while the second uses a variety of centrality measures to identify important publications and authors. Approaches leveraging such techniques fall under the term "citation analysis". A area of research concerned with the discovery and management of literature by analysing its references to evaluate scholarly contributions, track the flow of knowledge, study the structure of research field, etc. (D. Zhao and Strotmann 2015, p. 1-5). While useful, this requires the acceptance of several assumptions. Namely,

- citation of a document implies use of that document by the citing author;
- citation of a document (author, journal, etc.) reflects the merit (quality, significance, impact);
- citations are made to the best possible works;

- a cited document is related in content to the citing document;
- all citations are equal.

While accepting those rather strong assumptions is problematic, many of which were already violated by this text, additional concerns with this technique arise when one also considers that there can be various problems in the data, e.g. errors, self-citations or multiple authors. However, due to the fact that the those techniques are only used in a rudimentary manner to provide a starting point for the subsequent research, a proper justification of the applicability of those assumptions with respect to the given data will be omitted. However to satiate possible curiosity, please consider (Smith 1981) for detailed discussion about the validity of those assumption.
Firstly, the detection of communities. For the purposes of this work, a groups of researches is classified as a community (with respect to their work on causation), if the group is of size greater than two and if the group produced more than two relevant publications. This analysis approach should provide a rough estimate of the research clusters in the literature based on the information encoded in $\mathcal{G}_m$. The community detection itself is accomplished using an algorithm published in (Rosvall and Bergstrom 2008). This particular algorithm uses an information theoretic approach to lay bare the community structure in an directed weighted graph. Hence, if one excludes self-referential behaviour, the same algorithm can be used on any graph up for analysis, thus it is well suited for $\mathcal{G}_m$. Furthermore, (Rosvall and Bergstrom 2008) introduces their algorithm by studying a citation graph. Demonstrating the suitability of the algorithm for such tasks in the process. As the results obtain through community detection are less significant for the subsequent chapters, please consult (Rosvall and Bergstrom 2008) for further discussion and additional details. Nevertheless, to identify relevant communities the following proceeder will be used. Firstly, $\mathcal{G}_m$ will be cleared from all loops. Secondly, all authors that are only cited or that only cite, with respect to the collected data, are removed. Thirdly, the community detection algorithm is applied to the graph. Lastly, all communities are ranked based on the average number of relevant publications per author. Hence, providing the possibility for smaller communities to get some spotlight as well.
The primary technique used in this work to assess the importance of a vertex in a graph, relies on the use of centrality measures. Being significant for the work in subsequent chapters, the discussion of such measures warrens a more detailed discussion as compared to community detection part. In general, those centrality measures are used to rank vertices based on some notion of importance. In particular, they can be used to understand diffusion processes, assess an individuals risk of infection or explain the influence of a person in a social network (Bloch, Jackson, and Tebaldi 2019). According to (Pozo et al. 2011) degree, closeness and betweenness centralities are the

most popular ones. Additionally, there exists a family of centralities that is closely tied to field of spectral graph theory (see (Spielman 2012)), as those centralities use eigenvalues and eigenvectors in their computation. This includes measures such as the eigenvector centrality, alpha centrality, page rank and Katz-Bonacich centrality (Bloch, Jackson, and Tebaldi 2019).
The following, provides a brief intuition about some of the mentioned centrality measures and is compiled from information found in (Segarra and Ribeiro 2015; Pozo et al. 2011; Bloch, Jackson, and Tebaldi 2019; Bonacich and Lloyd 2001; Page et al. 1999). The degree centrality is a local measure of importance based on the degree of a vertex. In the case of a weighted graph, the weighted degree of a vertex is taken for this measurement, thus implying that the weight of an edge must reflect some notion of similarity. That is, a higher weight implies a stronger connection, e.g. number of interactions. If one is confronted with an directed graph, the degree centrality dissolves into an in- and out-degree centrality. Although it can be used to assess the "popularity" of a vertex, due to its locality it neglects the remaining structure of the graph. The closeness centrality is computed using the sum of all shortest path lengths. Hence, it is a measure for assessing the importance of a vertex based on how quick such a vertex can reach every other vertex. Moreover, this implies that edge weights must indicated a notion of dissimilarity, e.g. distance between vertices. Unfortunately, this measure requires the graph to be strongly connected. Hence, it is not suitable for directed graphs in general. The betweenness centrality gives higher values to vertices that are part of many shortest paths between pairs of vertices. Meaning it attempts to assess the importance of a vertex based how vital a vertex is for the flow of information between the other vertices in the graph. As this centrality builds upon the notion of a shortest path, it requires the weights of a graph to represent a dissimilarity between vertices. However, a benefit of this centrality is that it is suitable for both directed and undirected graphs. The eigenvector centrality is similar to the degree centrality, as it assess the importance of a vertex based on the number of neighbours, thus it requires weights to denote similarities. However, it differs in the evaluation of those neighbours, determining the importance of a vertex based in the importance of the vertices in its neighbourhood. That is, the eigenvector centrality is computed by assuming that the centrality of a vertex is proportional to the sum of eigenvector centralities of the vertex's neighbours. Hence, it is a self-referential process. Unfortunately, common implementations requires graphs to be undirected and connected. The alpha centrality is a generalisation of the Eigenvector centrality for directed graphs. The idea behind this centrality is that it assumes that a vertex has some exogenously defined start value. The Katz-Bonacich centrality generalises the Eigenvector centrality by reducing the importance of distant vertices. Page Rank relativises the centrality score passed on by a vertex, based on the number if neighbours. Hence, a vertex having a directed edge to an important vertex must not nec-

essarily have a high importance itself, e.g. a webpage linking to an important webpage must not necessarily be important itself. Furthermore, to ensure sensible results in directed graphs, dead ends are avoided by jumping to a random vertex instead.

Each of those measures allow for a separate ranking of publications and authors. However, a blind application of those methods would neglect the structure of the graphs and thus could lead to erroneous results. Hence, some additional care must be given and some slight adjustments to the graphs are required. In the case of $\mathcal{G}_p$, one is faced with an directed (and not strongly connected) graph. Therefore, the closeness or eigenvector centrality are ill suited for application on this graph. Furthermore, the degree centrality is obviously applicable. However, it decomposes into two separate measures. Additionally, the regular degree centrality will be used as well. That is, in the context of this particular dataset, the regular degree distribution actually provides a rough compromise between the recency bias of the out-degree, as well as the conservative tendencies associated with the in-degree measure (this can be observed in Figure 6.1). Hence, the undirected degree measure will be used as well. Unfortunately, due to their locality degree centralities provide a rather limiting picture, thus in an attempt to compensate for this shortcoming an alternative to the eigenvector centrality, namely page rank, is used. Although it is somewhat unusual to use this algorithm for citation graphs, this approach is not unheard of, see (Ding et al. 2009; Ma, Guan, and Y. Zhao 2008; P. Chen et al. 2007; Maslov and Redner 2008; Nykl et al. 2014) for an in-depth discussions. One particular benefit of determining the importance of a publication in such a manner, is that under Page Rank simply referencing an important publication, does not indicate a publications own importance. The last remaining common centrality measure, the Betweenness centrality, can and thus will be applied. Providing yet another dimension for selecting publications.

In the case of $\mathcal{G}_a$, one is faced with a directed, weighted graph containing loops. Hence, modifications to the graph are required. Firstly, while included for the sake of completion in the graph $\mathcal{G}_a$, it seems sensible to discount self-referential behaviour for the ranking. Secondly, all centrality measures used in the ranking of publications can accommodate weights in their assessment. However, the Betweenness Centrality requires the weight of an edge to express dissimilarity, thus it is necessary to convert the weights such that they express dissimilarity rather than similarity between vertices (see Runkler 2012, p. 13). Moreover, in addition to the centralities it is reasonable to include the number of publications (in the examined field) as an additional measure.

2.2 Analysis

This section starts by discussing the data collection process and by describing the collected data through the performance of a rudimentary quantitative analysis. All of which is discussed in Subsection 2.2.1. The subsequent subsection, i.e. Subsection 2.2.2, attempts to draw a rough picture of the literature landscape, by utilising the techniques outlined in Subsection 2.1.2.

2.2.1 Data Preparation and Basic Analysis

This subsection quantitatively describes the data collection process, while highlighting detected mistakes as well as deviations from the constructed methodology. Moreover, this is followed by a brief and offensively basic investigation into the collected data. Am analysis that is subsequently expanded upon, by presenting some key properties of the graphs $\mathcal{G}_t$, $\mathcal{G}_p$, $\mathcal{G}_a$, $\mathcal{G}_c$ and $\mathcal{G}_m$.
Using the outlined methodology the following literature database was constructed. By collecting all publications from the venues Journal Artificial Intelligence (AI), Journal Artificial Intelligence and Law (AI&Law), International Joint Conferences on Artificial Intelligence Organization (IJCAI) and Knowledge-Bases Systems (KBS) that where published between 01.2017 and 3.2020 one obtains a set containing 4223 publications. To be precise, AI contributed 267 publications, by contrast AI&Law provided only 60. Furthermore, from KBS a total of 1281 publications could be obtain, while the majority of publications, i.e. 2615, was sourced from IJCAI. After applying the keyword based filter, the set of publications, i.e. $\mathcal{S}_0$ up for consideration is reduced to a meagre 37.[2]
After closer investigation, the publications deemed relevant according to the specified criteria are

- Proof with and without probabilities (Verheij 2017);
- Characterizing causal action theories and their implementations in answer set programming[3]; (Haodi Zhang and Lin 2017);

[2](Zander, Liśkiewicz, and Textor 2019; Verheij 2017; D. L. Chen 2019; Neil et al. 2019; Li et al. 2019; J. Lu et al. 2018; Z. Zhang et al. 2018; Constantinou and Fenton 2017; Liang, Wang, and Hongyu Zhang 2017; Haodi Zhang and Lin 2017; K. Mu 2018; Kronegger, Ordyniak, and Pfandler 2019; Hyttinen, Saikko, and Järvisalo 2017; Junzhe Zhang and Bareinboim 2017; K. Zhang et al. 2017; Liu et al. 2017; Summerville, Osborn, and Mateas 2017; L. Zhang, Y. Wu, and X. Wu 2016; Albrecht and Ramamoorthy 2016; Chai et al. 2018; Bochman 2018a; Ibeling and T. Icard 2018; Laurent, J. Yang, and Fontana 2018; Chikahara and Fujino 2018; L. Zhang, Y. Wu, and X. Wu 2018; Bäckström, Jonsson, and Ordyniak 2018; Jaber, Jiji Zhang, and Bareinboim 2018; Sridhar, Pujara, and Getoor 2018; Wenjuan, F. Lu, and Chunchen 2018; Xu et al. 2019; Zhalama et al. 2019; Cai et al. 2019; Sridhar and Getoor 2019; Xie and F. Mu 2019; Hassanzadeh et al. 2019; Shankar et al. 2019; Liepina, Sartor, and A. Z. Wyner 2019)

[3]Since (Haodi Zhang and Lin 2017) was not accessible “Characterizing causal action

- Actual Causality in a Logical Setting (Bochman 2018a);
- On the conditional logic of simulation models (Ibeling and T. Icard 2018);
- Counterfactual Resimulation for Causal Analysis of Rule-Based Models (Laurent, J. Yang, and Fontana 2018);
- Scalable Probabilistic Causal Structure Discovery (Sridhar, Pujara, and Getoor 2018);
- ASP-based discovery of semi-Markovian causal models under weaker assumptions (Zhalama et al. 2019);
- Arguing about causes in law: a semi-formal framework for causal arguments (Liepina, Sartor, and A. Z. Wyner 2019).

Hence, $\mathcal{S}_0^c$ contains only 8 publications. Executing the described snowballing steps on the start set, one obtains a total of 872 publications. Out of which only 294 (around 34%) are categorised as relevant. As depicted in Figure 2.2, this can be made more precise. Meaning that $\mathcal{S}_{-1}$ obtained by performing backwards snowballing on $\mathcal{S}_0^r$, contains 204 publications out of which only 79 are relevant. The second backwards snowballing step, i.e. $\pi^-(\mathcal{S}_{-1}^r)$, provided a total of 486 publications with 165 being relevant. The set S_{+1} contains 30 publications collected by forward snowballing on $\mathcal{S}_0^r$. Performing a backward snowballing step on S_{+1}^r generates S_{+1-1} which contains 63 publications from which 25 are deemed relevant. The second forward snowballing step, i.e. $\pi^+(\mathcal{S}_{+1}^r)$, produces $\mathcal{S}_{+2}$ resulting in an additional 7 publications with only 3 relevant ones among them. Lastly, by performing a final backward snowballing step on $\mathcal{S}_{+2}^r$, 45 new publications are discovered increasing the number of relevant publications by another 7.
Although great effort was taken to make the data collection sufficiently accurate. Some mistakes were discovered the data collection was already completed. Particularly of note is that there are two publications titled "Causes and explanations: a structural-model approach: part i: causes" by the same authors, i.e. (Joseph Y. Halpern and Pearl 2001) and (Joseph Y Halpern and Pearl 2005). During the data collection process those two publication were unfortunately conflated. Fortunately, the similarities between those two publications, the latter provides an updated to definition presented in the prior, relativise this mistake.
From the meta-data of the publications alone, one is able to observe the contributions to this field over the years. That is, given the publication dates of the literature collected in $\mathcal{S}^c$ it is possible to construct Figure 2.3, which

theories and their implementations in answer set programming: Action languages b, c, and beyond" (Haodi Zhang and Lin 2015) will be used for the snowballing step. This departure from the methodology, is justified due to its initially high reprieved relevancy.

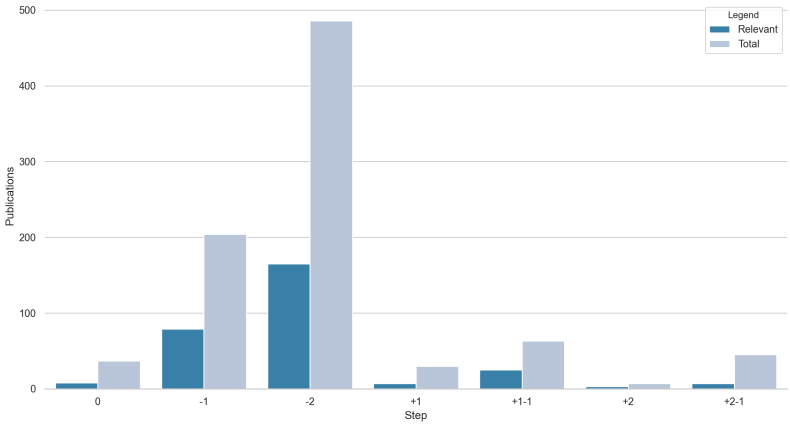


Figure 2.2: Number of publications and relevant publications added during each snowballing step. (From left to right $\mathcal{S}_x^c/\mathcal{S}_x$: 8/37, 79/204, 165/486, 7/30, 25/63, 3/7, 7/45)

depicts the distribution of number of publications (in $\mathcal{S}^c$) per year across the past 50 years. Furthermore, according to the data collected, the decade between 2000 and 2010 was the most productive period, i.e. S contains 73 publications before 2000, 114 between 2000 and 2010 and 107 publications from 2010 onwards. Additionally, it can be observed that 2004, 2007 and 2009 were the most productive years over all. Containing notable publications such as "Nonmonotonic Causal Theories" (Giunchiglia et al. 2004), "Causes and Norms" (Hitchcock and Knobe 2009), "Prevention, Preemption, and the Principle of Sufficient Reason" (Hitchcock 2007), "Two Concepts of Causation" (Hall 2004) and "Structural Equations and Causation" (Hall 2007) [4].

Further information can be extracted by encoding the collected data as graphs. As discussed earlier, the set of publications $\mathcal{S}$ and their references naturally induce a directed graph containing 872 vertices and 2052 edges. This graph, i.e. $\mathcal{G}_t$ can be observed in Figure 2.4. Induced by the set $\mathcal{S}^r$, containing publications that are both relevant and are published after (and including) 2010, one obtains $\mathcal{G}_p$ as a sub-graph of $\mathcal{G}_t$. $\mathcal{G}_p$ contains only 107 and 326 edges and can be observed in Figure 2.5. Using $\mathcal{G}_p$ one can then compute $\mathcal{G}_a$, visible in Figure 2.6, which contains a total of 130 vertices and 462 edges. As discussed this graph encodes the citations between authors and not the one between publications. Hence, it requires that its directed edges are weighted. To analyse the co-authorship relation one can create $\mathcal{G}_c$.

[4] discussing and using formalisms such as Neuron Diagrams, Structural Equations and some variant of McCain and Turner's Causal Logic

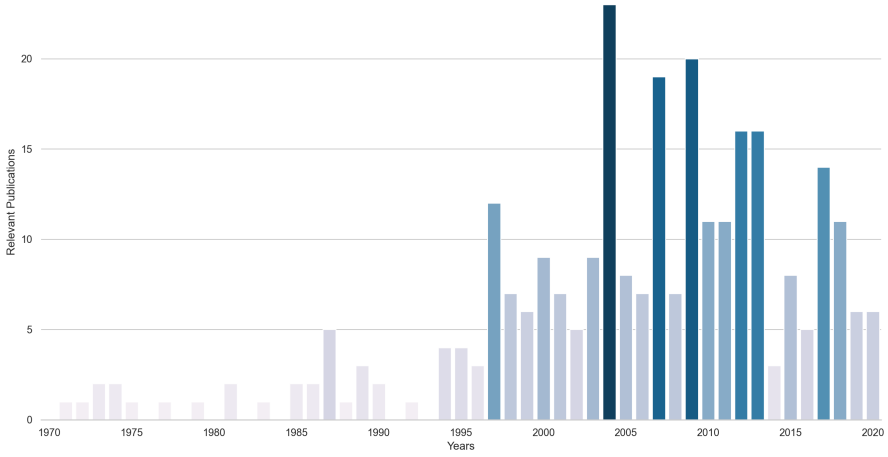


Figure 2.3: Number of relevant publications per year (a negligible amount publications occur before 1970)

	Vertices	Edges	Density	Clustering Coefficient
$\mathcal{G}_p$	107	326	0.0287	0.2791
$\mathcal{G}_a$	130	462	0.0275	0.3144
$\mathcal{G}_c$	130	192	0.0229	0.7843
$\mathcal{G}_a$	130	755	0.045	0.5141

Table 2.1: General properties of the discussed graphs. Other common measures such as average path length, radius and diameter, as well as vertex- and edge connectivity are omitted as all graphs in question are disconnected.

It is depicted in Figure 2.7 and contains 130 vertices and 192 undirected, weighted, edges. Lastly, $\mathcal{G}_m$, depicted in Figure 2.8, is the merger of $\mathcal{G}_a$ and $\mathcal{G}_c$, thus it contains 130 vertices and 755 edges. For a quick overview of some of their basic properties please consult Table 2.1 and Table 2.2, as well as Figure 2.9, Figure 2.10 and Figure 2.12 respectively.

	Minimum	Maximum	Average	Median
$\mathcal{G}_p$				
Degree	1	22	6.09346	5
In-Degree	0	18	3.04673	2
Out-Degree	0	13	3.04673	2
$\mathcal{G}_a$				
Degree	0	50	7.10769	4
In-Degree	0	29	3.55385	3
Out-Degree	0	23	3.55385	0
Weighted Degree	0	85	10.2615	5
Weighted In-Degree	0	53	5.13077	3
Weighted Out-Degree	0	32	5.13077	0
$\mathcal{G}_c$				
Degree	0	13	2.95385	2
Weighted Degree	0	17	3.67692	2.5
$\mathcal{G}_m$				
Degree	0	53	11.6154	10
In-Degree	0	29	5.80769	5
Out-Degree	0	24	5.80769	4
Weighted Degree	0	97	17.6154	13
Weighted In-Degree	0	59	8.80769	5.5
Weighted Out-Degree	0	32	8.80769	5.5

Table 2.2: Degree Statistic of $\mathcal{G}_p$, $\mathcal{G}_a$ and $\mathcal{G}_c$.

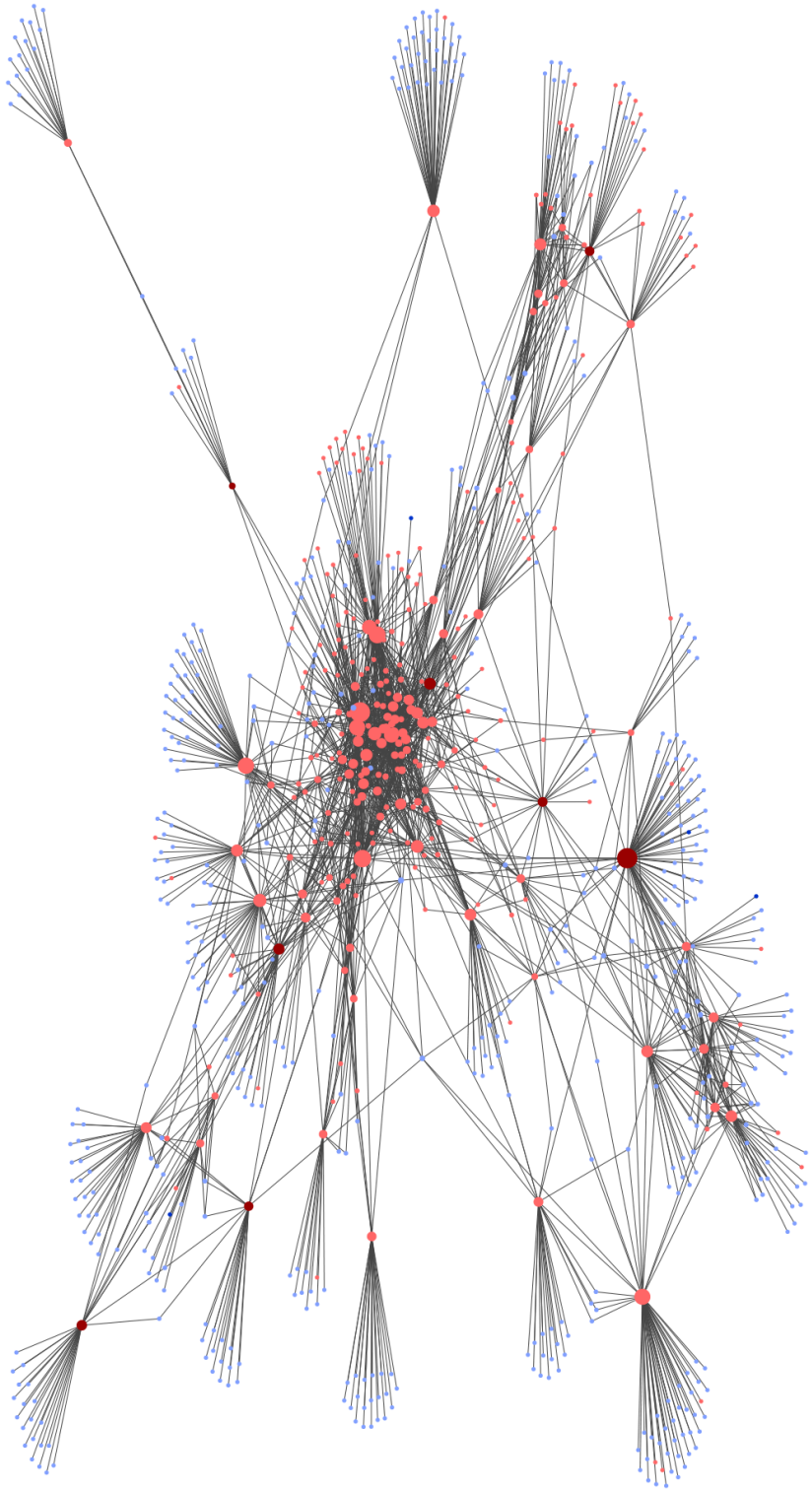


Figure 2.4: $\mathcal{G}_t$ where dark red (dark blue) indicates a(n) (ir)relevant publication in $\mathcal{S}_0$ and light red (light blue) indicates a(n) (ir)relevant publication in $\mathcal{S} \setminus \mathcal{S}_0$. (Isolated vertices are not depicted in this graph and edge direction is suppressed.)

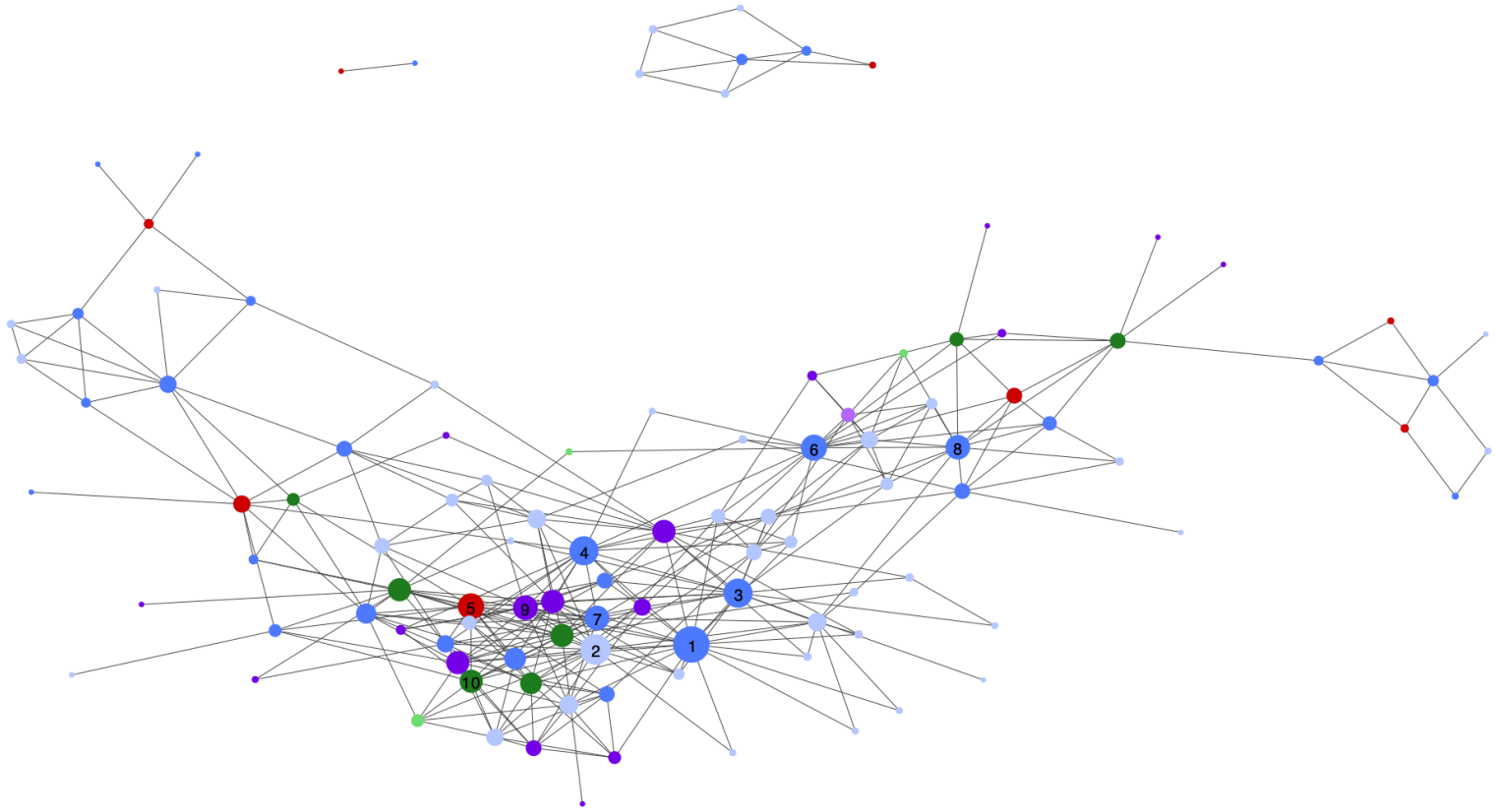


Figure 2.5: The publication graph $\mathcal{G}_p$, containing all relevant publications published after (and including) 2010. It consists out of 107 vertices (8 are in $\mathcal{S}^r_0$; 29 are in $\mathcal{S}^r_{-1}$; 42 are in $\mathcal{S}^r_{-2}$; 7 are in $\mathcal{S}^r_{+1}$; 17 are in $\mathcal{S}^r_{+2}$; 3 are in $\mathcal{S}^r_{+1-1}$; 1 are in $\mathcal{S}^r_{+2-1}$) and 326 edges. The vertex size correlates with vertex degree. 1: graded causation and defaults; 2: actual causation: a stone soup essay; 3: cause without default; 4: actual causation and the art of modeling; 5: actual causality in a logical setting; 6: counterfactuals; 7: a partial theory of actual causation; 8: from programs to causal models; 9:: a modification of the halpern-pearl definition of causality: 10: explaining actual causation in terms of possible causal processes. (Isolated vertices are not depicted in this graph and edge direction is suppressed.)

Figure 2.6: The Author Graph $\mathcal{G}_a$ based on $\mathcal{G}_p$. It consists of 130 vertices and 462 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree; edge width correlates with edge weight. 1: halpern; 2: lagnado; 3: vennekens; 4: gerstenberg; 5: hitchcock; 6: bex; 7: beckers; 8: verheij; 9: bochman; 10: icard (Isolated vertices are not depicted in this graph.)

Figure 2.7: The collaboration graph $\mathcal{G}_c$, based on $\mathcal{G}_p$. It consists of 130 vertices and 192 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree. Edge width correlates with edge weight. An edge with color red has weight greater than 1. (1: lagnado; 2: eberhardt; 3: gerstenberg; 4: vennekens; 5: zhang; 6: goodman; 7: tenenbaum; 8: fontana; 9: lee; 10: lifschitz.)

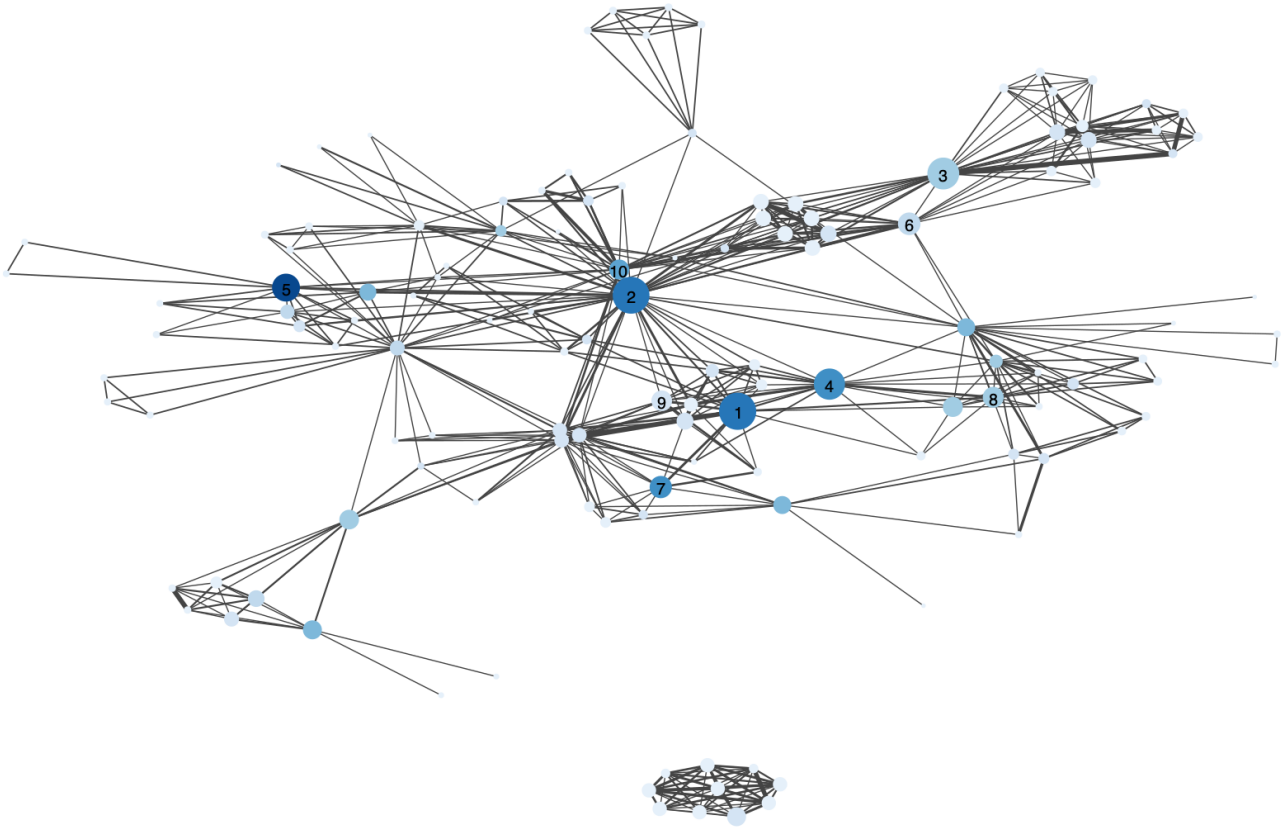


Figure 2.8: The merged graph $\mathcal{G}_m$ based on $\mathcal{G}_p$. It consists of 130 vertices and 755 edges. Darker colors indicate a higher number of publications in $\mathcal{G}_p$. Vertex size correlates with the weighted vertex degree; edge width correlates with edge weight. 1: lagnado; 2: halpern; 3: eberhardt; 4: gerstenberg; 5: vennekens; 6: zhang; 7: bex; 8: tenenbaum; 9: chockler; 10: hitchcock. (Isolated vertices are not depicted in this graph and edge direction is suppressed.)

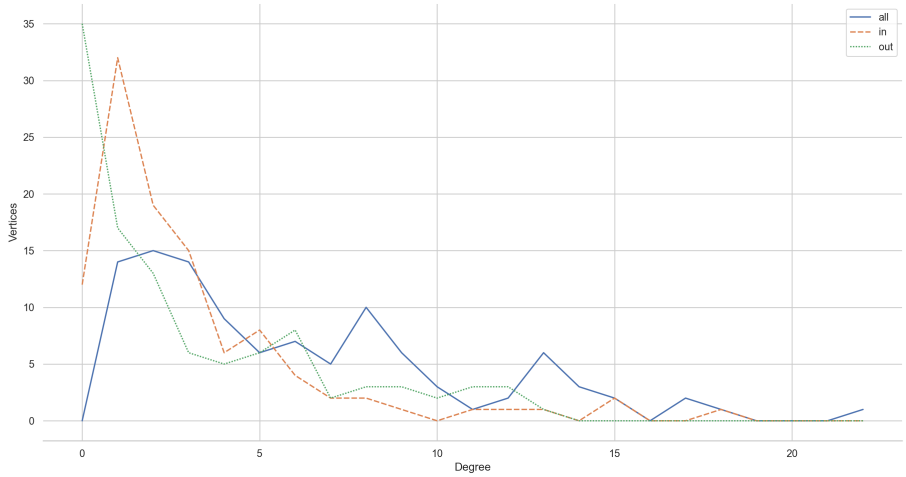


Figure 2.9: A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_p$

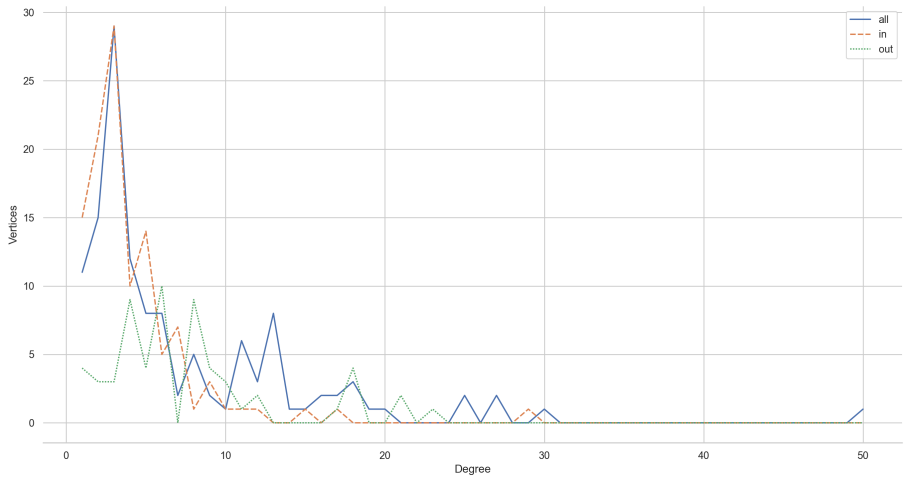


Figure 2.10: A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_a$

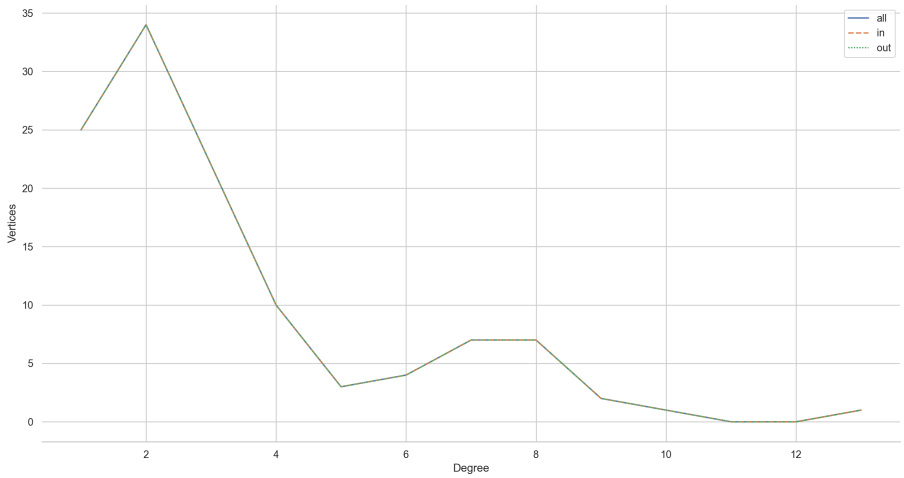


Figure 2.11: A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_c$

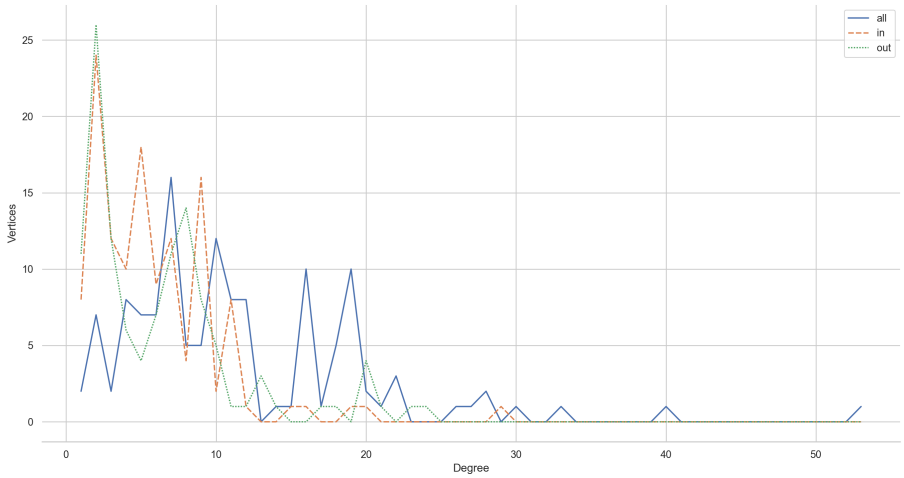


Figure 2.12: A line graph depicting the in-degree/out-degree/degree distribution of $\mathcal{G}_m$

2.2.2 Communities, Authors and Publications

This subsection presents the results obtained by using the tools discussed in Subsection 2.1.2 to analyse the data discussed in Subsection 2.2.1, which was collected according to the methodology presented in Subsection 2.1.1. The discussion starts by presenting the results of the community detection algorithm, is followed by providing a selection of authors deemed important and concludes by building the set of important publications which is subjected to further investigations in the subsequent chapters.

Applying (Rosvall and Bergstrom 2008)'s community detection algorithm to a subgraph of $\mathcal{G}_m$, which was obtained by removing all vertices with zero in- or zero out-degree, the following communities could be identified. Recall that a grouping of researches are classified as a community only if there group has a size greater than two and if the sum of their relevant publications exceeds two. As listed in Table 2.3 , the algorithm detected eight communities. As a whole they can be viewed in Figure 2.13 while the connections within each individual group can be viewed in Figure 6.2-6.9.

When ranked based on the amount of relevant publications per author, there are three groups, namely Group 4, 7 and 8, that set themselves apart from the remaining groups by having a disproportionally high relevancy. Starting with the lowest of those three, Group 8. This group consists of 4 people, Ibeling, Icard, Kominsky and Knobe, with Icard the most prominent author with respect to this group. They published 6 relevant publications, most of those publications discuss to some extend or another a language called simulation models, which can be used to encode causal relationships. Additionally, they discuss other formal languages such as causal models and Bayesian networks. The second research community is Group 7 and consists of five people, i.e. Verheij, Bex, Walton, van Koppen and Prakken. They contributed a total of eight relevant publications, all of which are placed within the context of causality and law. Furthermore, applications of Bayesian networks and a heavy emphasis on stories can be detected. The first one, Group 4, consisting of a total of 16 people which together are responsible for 31 relevant publications. One unifying aspect exhibited by many of the publications in this community, is the emphasis on logic, as well as their attempts to formalise token causality from an inductive example first approach. However, being of a considerable size this group can thematically be further segmented. In particular, one cluster seems to emerge around Vennekens, discussing a variety of approaches to causation with the most notable ones being based on a formal language called CP-Logic. Another, is thematically grouped around causal models, where Halpern seems to be the most dominant influence. While those are the two main areas discussed in Group 4, this is by no means exhaustive. For example, the work of Bochman, while being subject wise in closer proximity to the community around Halpern, can not be placed in either of the two groups with absolute certainty. A

Nr.	Size	Relevancy	Relevant Publications	Authors
1	15	0.53	8	zhang jiji, eberhardt frederick, mayer wolfgang, li mark junjie, baumgartner michael, hyttinen antti, hoyer patrik o, jarvisalo matti, glymour clark, danks david, glymour bruce, ramsey joseph, scheines richard, spirtes peter, teng choh man
2	15	0.8	12	goodman noah d, tenenbaum joshua b, gerstenberg tobias, chockler hana, fenton norman, keppens jeroen, lagnado david a, neil martin, tenenbaum josh, ullman tomer d, aleksandrowicz gadi, ivrii alexander, zultan ro'i, lake brenden m, gershman samuel j
3	5	0.6	3	livengood jonathan, alicke mark d, rose david, bloom dori, sytsma justin
4	16	1.94	31	bochman alexander, beckers sander, vennekens joost, blanchard thomas, schaffer jonathan, halpern joseph y, hitchcock christopher, bruynooghe maurice, denecker marc, weslake brad, huber franz, bogaerts bart, cabalar pedro, fandinno jorge, leblanc emily, balduccini marcello
5	9	0.89	8	zhang haodi, lin fangzhen, ferraris paolo, lee joohyung, lierler yuliya, lifschitz vladimir, yang fangkai, casolary michael, bartholomew michael
6	6	0.5	3	santorio paolo, romoli jacopo, wittenberg eva, ciardelli ivano, zhang linmin, champollion lucas
7	5	1.6	8	verheij bart, bex floris, walton douglas, van koppen peter j, prakken henry
8	4	1.5	6	ibeling duligur, icard thomas, kominsky jonathan f, knobe joshua

Table 2.3: Communities Overview

more significant failure of the employed heuristic can be observed with the work of Cabalar and Fandinno, would is thematically closer to the research conducted by Group 5, which investigates causality in the context of logic programming.

To identify the most important publications, six different rankings are established. That is, the three rankings rely on the weighted degree of a vertex, with the second and third being established using the weighted in- and out-degree respectively. The fourth, ranks the vertices according to the results provided by the betweenness centrality, the fifth relies on the values computed by the page rank algorithm and the last measures the importance of an author using their publication count. By aggregating the top 15 authors, see Table 2.4, across all 6 rankings into a single set, 33 important authors can be identified. Among those authors such as Lifschitz, Icard, Bochman, Eberhardt, Hitchcock, Gerstenberg, Lagnado and Halpern consistently score

high across each ranking and are thus of particularly of note. Using $\mathcal{G}_c$ one can observe there have been collaborations between Bochman and Lifschitz; Halpern and Hitchcock; Gerstenberg and Icard. Those collaboration are also reflected in the kinds of approaches those authors ascribe to when studying causation. That is, Both Lifschitz and Bochman focus on variants of the causal theory put forward in McCain, Turner, et al. 1997 and tend to approach causality from a regularity theoretic point of view. By contrast, Halpern and Hichcock strongly adhere to the structural equation framework. Their investigations into causality, while emerging from the counterfactual tradition, recently incorporate some regularity theoretic tools, e.g. extending causal models with normality rankings. As opposed to all other authors mentioned, who tend to approach causality from a more theoretical angle, Gerstenberg and Icard, set themselves apart, by conducting empirical studies investigating how humans form their causal judgements and what role the attribution of responsibility play in those judgements. Additionally, they investigate the role of causation in the legal domain. Moreover, similar to Halpern they tend to follow the counterfactual approach to causation and sometimes use structural equations for their modelling. Eberhardt, who cooperated with Clark Glymour on (Glymour et al. 2010), focuses on the discovery of causal structures. This includes formalisms such as causal Bayesian networks and seems to align closer with the part of the literature centring around machine learning. Lastly, Icard seems to argue for the need of an expressive formalism that emphasises the for him apparent procedural character of causation, thus his latest publications introduce and discuss simulations models, which share a close relationship to Turing machines. Having the highest number of relevant publications, an honourable mention must be given to Vennekens Joost, who worked with structural equations, CP-logic and action languages.

Moving on to the identification of publications deemed important by the outlined methodology using the data encoded in $\mathcal{G}_p$. For each of the five proposed orderings (i.e. degree centrality, in-degree centrality, out-degree centrality, betweenness centrality, page rank), the 15 publication deemed most important, see Table 2.5, are selected and aggregated into a single set, resulting in a total of 36 unique publications.
Particularly of note are the articles Weslake 2015, Blanchard and Schaffer 2017, Joseph Y Halpern and Hitchcock 2011, Glymour et al. 2010 and Joseph Y Halpern and Hitchcock 2015, all of which are ranked highly across all measures. All publications use causal models as their preferred method of encoding causal relations. While all of those publications take on a counterfactual perspective, (Joseph Y Halpern and Hitchcock 2011), (Weslake 2015) and (Joseph Y Halpern and Hitchcock 2015) expand the causal model framework to incorporating some aspects of the regularity theory of in the causal model approach. For example, this is accomplished by extending

	Degree Centrality	In-Degree Centrality	Out-Degree Centrality	Betweenness Centrality	Page Rank	Publications
1	halpern joseph y	halpern joseph y	icard thomas	halpern joseph y	claassen tom	vennekens joost
2	hitchcock christopher	lagnado david a	bochman alexander	lagnado david a	heskes tom	halpern joseph y
3	bochman alexander	gerstenberg tobias	halpern joseph y	eberhardt frederick	halpern joseph y	lagnado david a
4	lagnado david a	hitchcock christopher	liepina ruta	gerstenberg tobias	lagnado david a	bex floris
5	icard thomas	lifschitz vladimir	sartor giovanni	bochman alexander	gerstenberg tobias	gerstenberg tobias
6	gerstenberg tobias	eberhardt frederick	wyner adam	hitchcock christopher	lee joohyung	verheij bart
7	eberhardt frederick	claassen tom	hitchcock christopher	bex floris	lifschitz vladimir	icard thomas
8	liepina ruta	heskes tom	ibeling duligur	zhang jiji	lierler yuliya	lee joohyung
9	sartor giovanni	zultan ro'i	blanchard thomas	fenton norman	yang fangkai	beckers sander
10	wyner adam	hyttinen antti	baumgartner michael	schaffer jonathan	zultan ro'i	hitchcock christopher
11	ibeling duligur	hoyer patrik o	schaffer jonathan	lifschitz vladimir	hitchcock christopher	ibeling duligur
12	schaffer jonathan	jarvisalo matti	eberhardt frederick	lee joohyung	eberhardt frederick	eberhardt frederick
13	fenton norman	fenton norman	gerstenberg tobias	verheij bart	hyttinen antti	lifschitz vladimir
14	lifschitz vladimir	lee joohyung	keppens jeroen	icard thomas	hoyer patrik o	schaffer jonathan
15	chockler hana	lierler yuliya	lagnado david a	blanchard thomas	jarvisalo matti	goodman noah d

Table 2.4: Top 15 authors according to the Degree Centrality, the In- and Out-Degree Centrality, the Betweenness Centrality, the Page Rank algorithm and the number of publication.

	Degree Centrality	In-Degree Centrality	Out-Degree Centrality	Betweenness Centrality	Page Rank
1	graded causation and defaults	actual causation: a stone soup essay	necessary and sufficient conditions for actual root causes	graded causation and defaults	counterfactuals
2	actual causation: a stone soup essay	counterfactuals	explaining actual causation in terms of possible causal processes	cause without default	actual causation: a stone soup essay
3	cause without default	actual causation and the art of modeling	causal reasoning in a logic with possible causal process semantics	a modification of the halpern-pearl definition of causality	actual causation and the art of modeling
4	actual causation and the art of modeling	a partial theory of actual causation	causation in legal and moral reasoning	from programs to causal models	a hybrid formal theory of arguments, stories and criminal evidence
5	actual causality in a logical setting.	graded causation and defaults	cause without default	a principled approach to defining actual causation	representing synonymity in causal logic and in logic programming
6	counterfactuals	actual causality	on laws and counterfactuals in causal reasoning	a general framework for defining and extending actual causation using cp-logic	embracing events in causal modelling: interventions and counterfactuals in cp-logic
7	a partial theory of actual causation	a hybrid formal theory of arguments, stories and criminal evidence	situation calculus semantics for actual causality	appropriate causal models and the stability of causation	trumping and contrastive causation
8	from programs to causal models	causation: a user's guide	actual causality in a logical setting.	actual causality in a logical setting.	spreading the blame: the allocation of responsibility amongst multiple agents
9	a modification of the halpern-pearl definition of causality	embracing events in causal modelling: interventions and counterfactuals in cp-logic	graded causation and defaults	the computational complexity of structure-based causality.	causal discovery in multiple models from different experiments
10	explaining actual causation in terms of possible causal processes	a regularity theoretic approach to actual causation	from programs to causal models	grounding in the image of causation	discovering cyclic causal models with latent variables: a general sat-based procedure
11	necessary and sufficient conditions for actual root causes	actual causation in cp-logic	normality and actual causal strength	normality and actual causal strength	graded causation and defaults
12	on laws and counterfactuals in causal reasoning	cause without default	a modification of the halpern-pearl definition of causality	causal analysis for attributing responsibility in legal cases	a partial theory of actual causation
13	appropriate causal models and the stability of causation	interventionist counterfactuals	arguing about causes in law: a semi-formal framework for causal arguments	actual causation and the art of modeling	"if you'd wiggled a, then b would've changed": causality and counterfactual conditionals."
14	situation calculus semantics for actual causality	"if you'd wiggled a, then b would've changed": causality and counterfactual conditionals."	probabilistic reasoning across the causal hierarchy	on laws and counterfactuals in causal reasoning	translating first-order causal theories into answer set programming
15	causation in legal and moral reasoning	spreading the blame: the allocation of responsibility amongst multiple agents	appropriate causal models and the stability of causation	a proposed probabilistic extension of the halpern and pearl definition of 'actual cause'	a regularity theoretic approach to actual causation

Table 2.5: Top 15 publications according to the Degree Centrality, the In- and Out-Degree Centrality, the Betweenness Centrality and the Page Rank algorithm.

causal models such that they allow for the expression of normality. By contrast, (Blanchard and Schaffer 2017) argues that being more conservative when selecting appropriate causal models, should be preferred over incorporating additional widgets into structure of causal models itself. As allowing for defaults in causal models does not only increases their complexity, but also provides to much flexibility. (Glymour et al. 2010) does not engage with the debate about normality in causal inference. Instead it heavily criticises the attempt of inductively defining causal models from small examples alone, a strategy employed throughout most of the literature.

Moreover, this set of publications will be increased by adding all publication from $\mathcal{S}_0^r$ and by including all publication with 0 in- or out-degree (wrt. to $\mathcal{G}_p$) that have a higher that average (with respect to their respective cohort) degree. This result in a set of 44 publications. Meaning that the publications"Causal Reasoning in a Logic with Possible Causal Process Semantics", "On the Conditional Logic of Simulation Models", "Evaluation of Causal Arguments in Law: The Case of Overdetermination", "Explaining Actual Causation via Reasoning about Actions and Change", "Probabilistic Reasoning across the Causal Hierarchy" and "Arguing about Causes in Law: A Semi-formal Framework for Causal Arguments" are added to the set. Finally, after removing books from this set, i.e. removing "Counterfactuals", "Causation: A User's Guide" and "Actual Causality", as well as removing older publications from authors having more than two important publications, i.e. removing older publications from "Denecker", "Halpern", "Hitchcock", "Icard", "Lagnado" and "Vennekens", it contains only 36 publications. [5]. Let this set be called $\mathcal{S}^f$. The graph induced from $\mathcal{S}^f$ can be can be observed in Figure 2.14.

To conclude, some literature suggestions based on the whole graph $\mathcal{G}_t$ irrespective of the relevancy marker. That is, by analysing $\mathcal{G}_t$ it is possible to provide some literature recommendations based on the number of citation a publications has received. Starting with the five articles with the greatest amount of citations, i.e. "Causes and explanations: A structural-model approach. Part I: Causes" (Joseph Y Halpern and Pearl 2005), "Causation" (Lewis 1974), "The intransitivity of causation revealed in equations

[5](Vennekens, Bruynooghe, and Denecker 2010; Bex et al. 2010; Lee et al. 2010; Lifschitz and F. Yang 2010; Glymour et al. 2010; Claassen and Heskes 2010; Gerstenberg and Lagnado 2010; Joseph Y Halpern and Hitchcock 2011; Shulz 2011; Briggs 2012; Baumgartner 2013; Hyttinen, Hoyer, et al. 2013; Joseph Y Halpern and Hitchcock 2015; Weslake 2015; Chockler et al. 2015; Beckers and Vennekens 2016; Schaffer 2016; Joseph Y Halpern 2016b; Blanchard and Schaffer 2017; Wright and Goldberg 2017; T. F. Icard, Kominsky, and Knobe 2017; Aleksandrowicz et al. 2017; Fenton-Glynn 2017; Lagnado and Gerstenberg 2017; Bochman 2018a; Ibeling and T. Icard 2018; Beckers and Vennekens 2018; Bochman 2018b; Denecker, Bogaerts, and Vennekens 2018; Batusov and Soutchanski 2018; Denecker, Bogaerts, and Vennekens 2019; Liepiņa, Sartor, and A. Wyner 2019; LeBlanc, Balduccini, and Vennekens 2019; Liepiņa, Sartor, and A. Wyner 2020; Khan and Soutchanski n.d.; Ibeling and T. Icard 2020)

and graphs" (Hitchcock 2001), "Structural Equations and Causation" (Hall 2007) and "Two Concepts of Causation" (Hall 2004). Particularly of note is (Lewis 1974), as it is is one of the foundational publication responsible for the current surge of interest in the counterfactual approach to causation (Beebee, Hitchcock, and Menzies 2009). It's perceived influence is further supported by the fact that all authors represented in the list above build upon Lewis'es legacy by discussing causation from a counterfactual point of view. However, they differ in their preferred language to represent causal dependencies and in their specific definition of token causality.
Moreover, some book recommendations can be given as well. The top five most cited books on the topic of causation are, "Causality: Models, Reasoninig and Inference" (Pearl 2009), "Making things happen: A theory of causal explanation" (Woodward 2005), "Causation, prediction, and search" (Spirtes et al. 2000), "Causation, prediction, and search" (Spirtes et al. 2000), "Causation in the Law" (Hart and Honoré 1959) and "Counterfactuals" (Lewis 2013). An honourable mentions should be given to the sixth place "Actual Causality" (Joseph Y Halpern 2016a), which provides a great summery over the vast amount of work put forward by Halpern on the topic of causation.

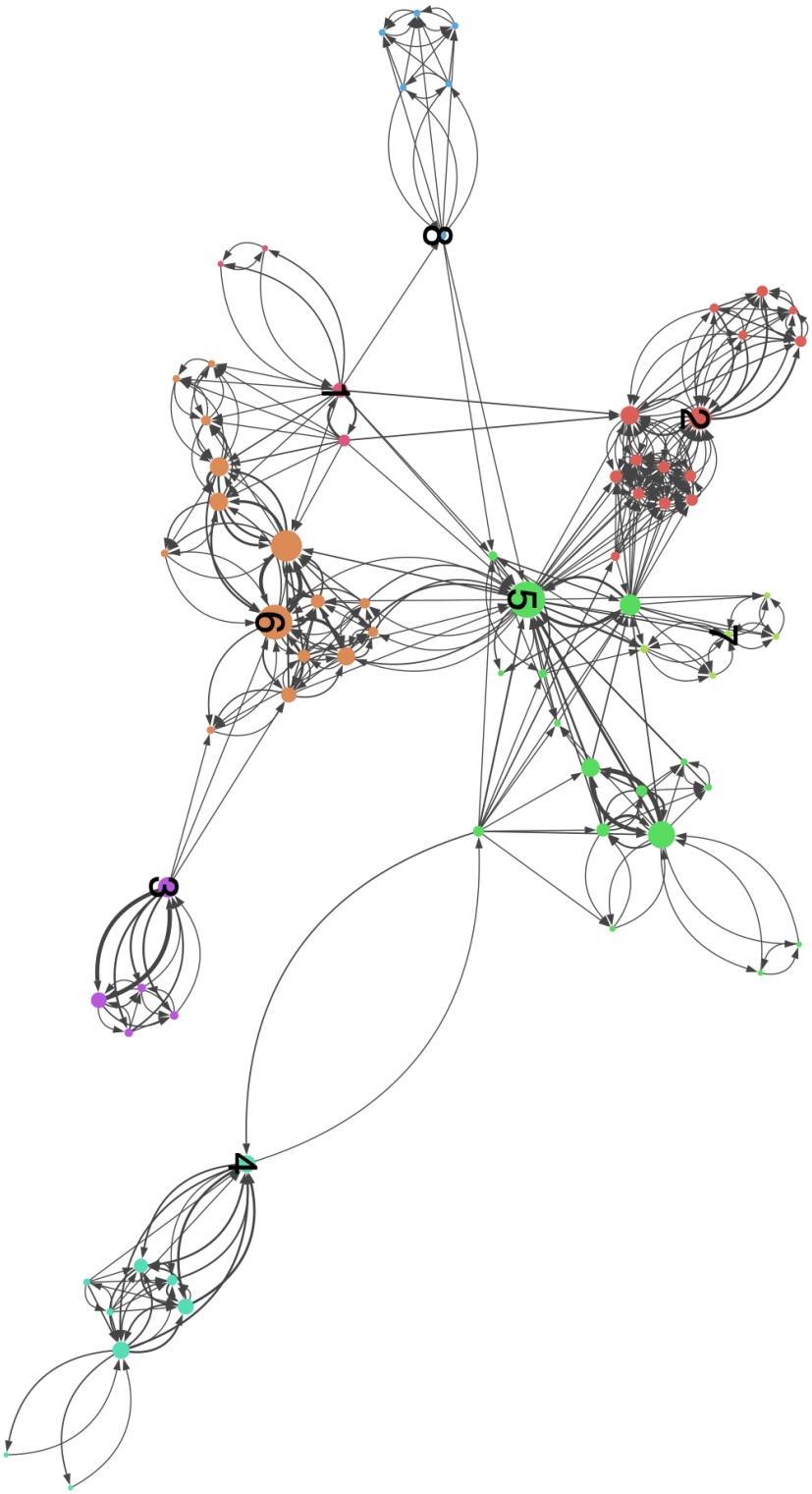


Figure 2.13: A subgraph of $\mathcal{G}_m$, where the colours indicate community affiliation.

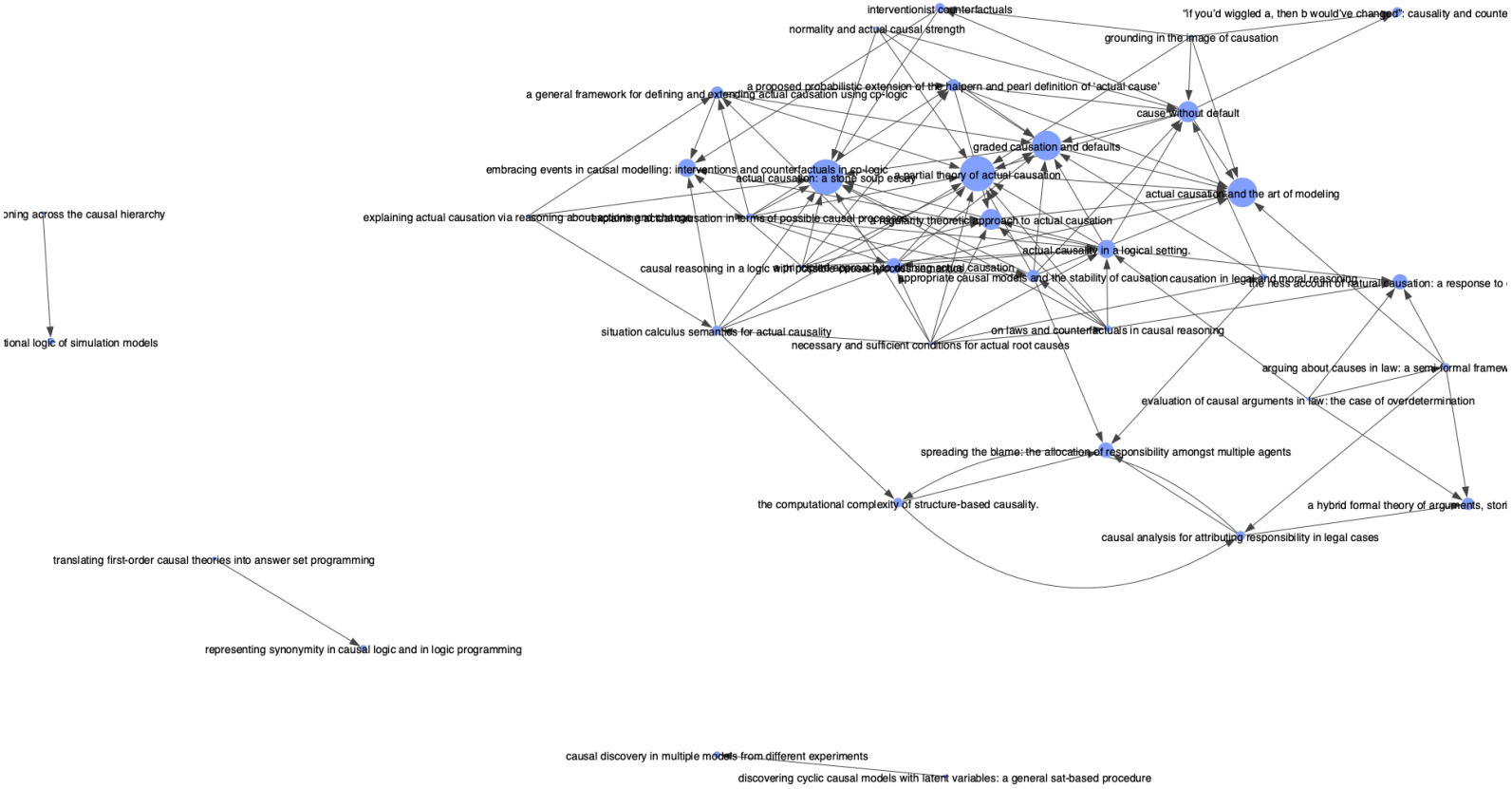


Figure 2.14: The subgraph from $\mathcal{G}_p$ induced by S_A

Chapter 3

Approaches to Causation: An Overview

3.1 Modelling Languages

The important publications as contained in S_A (see Subsection **??**) rely on a variety of languages to encode causal relations. Table 5.1 provides insight into which publications discuss which language family. However, it must be remarked that some liberty with respect to aggregation was taken, e.g. causal models with and without defaults are considered part of the same language family. Here it is important to point out that by far the most discussed framework are causal models, with CP-Logic and Causal Logic tying for a distant second place.

3.1.1 Causal Models

Spear headed by Pearl, see (Pearl 1995), causal models are the most common method for encoding the causal structures. This is, especially true in the context of token causality. The idea underlying causal models is that the causal mechanisms governing the world, can be described by a set of random variables, which can be partitioned into exogenous and endogenous variables, and a set of deterministic structural equations. And that by specifying the context, i.e. an assignment of specifying the values of all exogenous variables, one has sufficient expressivity to detect token causes for most scenarios (Joseph Y Halpern 2015). A structural equation is a method for conveniently expressing all type-causal relations for a variable in a single equation. Those equations are not algebraic in nature and are thus best understood as assignments. Meaning that the value of the dependent variable are best read as assignments rather than as a algebraic equation. A rather sensible choice, as a cause influences its effect, while an effect does not necessarily impact its cause. For example, a volcanic eruption may cause one to reconsider

Articles	Causal Models	CP-Logic	Situation Calculus	Non-Monotonic Causal Theory	FOL	Neuron Diagrams	Conditional Logic	$\mathcal{AL}$	SFCA	Abductive Causal Theory
Vennekens, Bruynooghe, and Denecker 2010		X								
Bex et al. 2010										X
Lee et al. 2010				X						
Lifschitz and F. Yang 2010				X						
Glymour et al. 2010	X					X				
Claassen and Heskes 2010	X									
Gerstenberg and Lagnado 2010	X									
Joseph Y Halpern and Hitchcock 2011	X									
Shulz 2011	X									
Briggs 2012	X									
Baumgartner 2013					X	X				
Hyttinen, Hoyer, et al. 2013	X									
Joseph Y Halpern and Hitchcock 2015	X									
Weslake 2015	X									
Chockler et al. 2015	X									
Beckers and Vennekens 2016		X				X				
Schaffer 2016	X									
Joseph Y Halpern 2016b	X									
Blanchard and Schaffer 2017	X									
Wright and Goldberg 2017										
T. F. Icard, Kominsky, and Knobe 2017										
Aleksandrowicz et al. 2017	X									
Fenton-Glynn 2017										
Lagnado and Gerstenberg 2017	X									
Bochman 2018a	X			X						
Ibeling and T. Icard 2018							X			
Beckers and Vennekens 2018	X									
Bochman 2018b	X			X						
Denecker, Bogaerts, and Vennekens 2018		X								
Batusov and Soutchanski 2018	X		X							
Denecker, Bogaerts, and Vennekens 2019	X	X								
Liepiņa, Sartor, and A. Wyner 2019									X	
LeBlanc, Balduccini, and Vennekens 2019								X		
Liepiņa, Sartor, and A. Wyner 2020	X								X	
Khan and Soutchanski n.d.	X		X							
Ibeling and T. Icard 2020	X						X			

Table 3.1: Depicts which publication discuss which languages families

their plans of going on vacation in Pompeii, yet not taking a vacation in Pompeii is (most likely) not a cause of said volcanic eruption. In general, such causal models could have cyclic relationships among their variables. However, so called *acyclic* causal models tend to be the primary subject of investigation, which considering the context of token causality, seems to be a sensible decision. Moreover, due to the fact that in an acyclic model the value of an endogenous variable is uniquely determined given the values of the exogenous variables, it allows for simpler reasoning. Intuitively, a causal model is considered acyclic, if one can order the endogenous variables such that a variable can not be influenced be the values of the variables above Joseph Y Halpern 2015. Relying on its acyclicity one can easily compute the values of the variables in a recursive fashion, by starting with the context, then iteratively assigning those variables their values that have structural equations relying only on already computed values. This can be continued until all variables are assigned a value, i.e. until a fixed point is reached.
While those relationships are allowed to be cyclic, they are deterministic. Initially this may strike as a rather strange decision, as it is rather tempting to declare the relationship between cause and effect to be probabilistic, e.g. a lightning strike has a 60% chance of causing a forest fire. However, a deliberate choice was made to keep structural equations deterministic. By contrast, approaches such as Causal Bayesian Networks are inherently probabilistic. In (Pearl 2009) compares those two approaches by drawing an analogy to the Laplacian and the quantum mechanical conception of physics. That is, the prior considers natures laws as deterministic and uncertainty only emerges due to ignorance, while the latter understands determinism as a mere approximation of inherently probabilistic laws. However, as the goal of this endeavour is to capture the understanding of causality intuitive to humans, the focus on the prior is apt (Pearl 2009). Similarly to Pearl, Haplern advises against such relations, suggesting instead the expansion of the model such that this uncertainty can be described, e.g. adding variables such as dryness or altitude. However, as this is not always possible, one can easily push probability out of the equations by putting a probability distribution of the exogenous variable in a causal model (Joseph Y Halpern 2015; Joseph Y Halpern 2016a, p. 13).
There seem to be concerns about the capabilities of causal models. Neglecting criticism about the lack of generalisation to first order logic (Batusov and Soutchanski 2018), as an approach causal models, are criticised and/ or extended many fronts. Firstly, to overcome the confinement to deterministic structural equations, (Fenton-Glynn 2017) proposed a probabilistic extension of causal models `CM+P`. Secondly, (Beckers and Vennekens 2018) advocate an extension that allows one to incorporate temporal information directly into the causal model. This extension will be called `CM+T`. Classically, this was achieved using standard causal models by adding timestamps to the variables in a causal model, an approach that is not without its merits

(Ibeling and T. Icard 2018). Finally, the main thrust behind the desire of increasing are a group of examples (see Example 4.1.12 and Example 4.1.14) that suggest that isomorphic causal models can have differing causal intuitions. Those examples are called non-structural counterexamples. Although Although contested, see (Blanchard and Schaffer 2017), it is the perceived failure of causal models on that front, that motivated the extension of causal models with some theory of normality, e.g. `CM+D` and `CM+N`. On the one hand, `CM+D` only distinguishes between default and deviant values of variables, while on the other `CM+N` relies on a ranking of multiple world, deeming some more normal than others. The latter provides a high degree of modelling power. Such power generates certain advantages, such as allowing one to capture the distinction between conditions and causes present in the legal tradition of causality, with conditions being simply values of variables with a higher degree of normality. However, it also provides the modeller with the ability to render many causal semantics irrelevant, thus further exacerbating Hall's complaint[1] about the structural equation approach (Blanchard and Schaffer 2017; Joseph Y Halpern and Hitchcock 2015; Weslake 2015).

3.1.2 CP-Logic

The family of CP-Logic is closely related to logic programming, sharing overlapping themes in both syntax and semantics. Using the methodology outlined in XXXX, two members of this family were discovered. Let them be called `CP` and `CP2` (Denecker, Bogaerts, and Vennekens 2019).
The prior, i.e. `CP`, seems to be the first member of the family. It was developed in (Vennekens, Denecker, and Bruynooghe 2009) as a probabilistic logic programming language with a informal semantics, independent from the epistemic agent based semantics of deterministic logic programs and the frequentist interpretation in the probability calculus, capable of expressing probabilistic causal laws. The view probabilistic causal laws as follows. Each law connects a cause with their possible effects. That is, an event[2] can be the cause of multiple effects; if this event occurs there can be at most one effect; the probability of the effect taking place, given the occurrence of the event is dictated by the probability assigned to each effect in the rule. On the semantic side, the took inspiration form action languages and used an approach championed by Shafer to create an appropriate semantic for this language. In particular, the underlying idea is that causal and probabilistic concepts should be considered as a dynamic context, i.e. they should be

[1] The structural equations approach places a much greater emphasis on problem modelling, amounting to little more than building the solution into the model Erwig and Walkingshaw 2010

[2] Due to the fact that this language operates in the intersection of both logic programs and probability theory, they have to distinguish, between events that cause transitions between states and events that are set a collection of possible outcomes. Hence, they follow Shafer and call the prior Humean event and the latter Demoivrian event

understood as a story explaining how the domain evolves. Shafer formalises this intuition by relying of probability trees. In such a tree, vertices represent states, edges represent events, each having a probability of occurrence, that lead to state transitions. This can be conceptualised as an event causes the system to move from the parent node to one of its children. The child to which one transitions to, is determined by the probability of the event's occurrence.

3.1.3 Situation Calculus

3.1.4 Non-Monotonic Causal Theory

3.1.5 Neuron Diagrams

3.1.6 Conditional Logic

3.1.7 Semi-Formal Framework for Causal Arguments

Starting with the definition provided in (Liepiņa, Sartor, and A. Wyner 2020), which is categorised by its authors as semi-formal. That is, rather than being concerned with capturing the notion of causation within a single definition, they propose a framework with enough modularity to substitute one definition of token causality with another. This choice is a reflection of the fact that the purpose of this framework is to bridge the gap between the formal literature on causation and legal analysis. Moving fully into the informal realm

3.1.8 Other

3.2 Token Causality Definition

Right after causal models, which clearly are the dominating formalism used in the context of token causality, is the language family called CP-Logic.
The formalisms summarised in Table 3.1 can be used to express causal dependence in one way or another. Unfortunately, this alone is not sufficient to make inferences about token causality. For example, the structural equations in an acyclic causal models allow one to encode causal relationships on the type level. Using those equations and a description of the world it is possible to determine the precise value of each variable in the model (Joseph Y Halpern 2015). However, when it comes to token causality one is interested in the question of what caused variable X to have its particular value x. Therefore, a good part of the literature is concerned in finding the appropriate conditions to construct a set of variables explain why $X = x$. Finding an appropriate definition of this kind of causality, proofs to be rather difficult. Hence, given the 9 formal languages used, a total of 32^3 definition are discussed in the articles from S_A.
To do justice to all of the discovered formalisms, the subsequent paragraphs will provide a brief summary. However, for a quick overview consult Table **??**.

3.2.1 Overview

The chronologically first definition, referenced by `HP-01`, is due to Haplern and Pearl (HP). It was originally formulated in (Joseph Y. Halpern and Pearl 2001), inspired by Pearl's causal beam notion (see (Pearl 1998)) and uses counterfactuals to identify token causes. Being seminal it became quickly apparent that further improvements are not only necessary but also possible. In particular, (Hopkins and Pearl 2003) proposed an example, that seemingly demonstrated that this definition is insufficient. Leading to the formulation of the updated HP-definition, abbreviated as `HP-05`. It originated in (Joseph Y Halpern and Pearl 2005) and is by far the most popular, i.e. the most widely used and discussed, formalism yet. Being merely an update of the original, `HP-05` remains firmly rooted in the counterfactual tradition. Unfortunately, it was demonstrated in that the computational complexity of finding causes with `HP-05` is D_2^P-complete[4] for both binary and general causal models (Aleksandrowicz et al. 2017). By contrast, (Eiter and Lukasiewicz 2002) demonstrated that `HP-01` is merely NP-complete in

[3] Four of them But-For, INUS, NESS, "causally relevant factor" tend to be defined using natural language.

[4] D_k^P is defined in (Aleksandrowicz et al. 2017) as the set of all languages L_3 such that there exists a language $L_1 \in \Sigma_k^P$ and a language $L_2 \in \Pi_k^P$ such that $L_3 = L_1 \cap L_2$. Σ_k^P and Π_k^P are simply levels on the polynomial hierarchy (see (Arora and Barak 2009, p. 97-99)). This complexity class is a generalised version of the $k = 1$ case coined in (Papadimitriou and Yannakakis 1982)

the binary and Σ_2^P-complete in the general case. Fortunately, the necessity of HP-05 was challenged in (Joseph Y Halpern 2016b), where it is argued that the model used to discredit HP-01 neglected to properly formalise the provided scenario, thus that a small, but unfortunately not always natural, expansion of said model is sufficient for HP-01 to produce the desired judgement. In (J. Halpern 2015) an new variant of this family was formulated. This definition is referred to as the modified HP definition and is therefore abbreviated here with HP-15. According to Halpern this definition is not only conceptually and computationally simpler, but also provides the more preferable answers. That is, it deals with the critique raised in (Hopkins and Pearl 2003) and it handles various examples better that HP-05, e.g. Hall's non-existent threat example found in(Hall 2007) (J. Halpern 2015; Joseph Y Halpern 2016a, p. 27). With respect to computational complexity, HP-15 is only NP-complete in the binary and D_1^P-complete in the general case, making it the most efficient HP-definition yet (Joseph Y Halpern 2016a, p. 153-154). Another extension of the HP-definition, was presented in (Fenton-Glynn 2017). The purpose of which was to adjust HP-05 to make contrastive causal judgement. This extension is based on the view that causation is contrastive in nature, thus is not a binary, but a tertiary relation. For example, administering one does of medicine saves the patients live and administering a second dose is absolutely redundant, i.e. same outcome as giving only a single dose. Hence, giving two doses instead of zero caused the patient to survive, while giving two doses instead of one is immaterial to the persons survival. For a detailed discussion see Section XXXX.
There are also independent definitions that use only the basic causal model variant. Firstly, there is Hitch-01 which was provided in (Hitchcock 2001). Secondly, there is Wood-03 formulated in (Woodward 2005). Thirdly, in (Glymour et al. 2010) two simplified versions of HP-05 called, *Simple* and *SimpleJ*, or abbreviated Simple and SimpleJ, are proposed. However, all of the above disagree with HP-05 on some of the traditional examples found in the literature. Lastly and particularly of note is the "Partial Theory of Actual Causation", or abbreviated PTC, put forward by Weslake in (Weslake 2015), as he claims that his version improves upon the HP-05 definition. This is partially accomplished by incorporating some tools from the regularity theoretic tradition without extending the causal model by some additional structure. Although their necessity is subject of contention, e.g. see (Blanchard and Schaffer 2017), during there years HP-05 was extended by some form of default reasoning, i.e. they incorporated a normality ordering over various contexts. This extension was motivated by the discovery of several so called non-structural examples, i.e. examples that have the same causal model and yet exhibit different intuitive answers. Hence, it using such extended causal models provides the necessary flexibility, possibly an excessive amount of flexibility, to resolve the issues put forward by those examples. Additionally, this notion of normality brings forth the possibility of introducing normative

reasoning into causal judgements. In the selected literature only HP-05d , i.e. the extension of HP-05 with default reasoning was mentioned. However, it is possible to extend HP-15 in a similar fashion. Moreover, the discussed approach is not the only method of extension (]p. 97-103]halpern2016actual.
Apart from the HP-definition, there are other definitions that use causal models and incorporate some notion of normality. Given characterised literature, both, HmM and HP-15 can be found in (Blanchard and Schaffer 2017). The first one is called *Hitchcock-meets-Menzies* or HmM, and is a modified version of a definition found in (Hitchcock 2007). It incorporates normality assumptions by simply partitioning the range of each variable in a causal model into default and deviant values. The second one is called *Menzies-by-Menzies* or MbM, it builds on the ideas proposed in (Menzies et al. 2007), requires one to rank the possible states of the world based on their normality.
Naturally, there are definitions that require causal models to be extended and generalised to capture other notions than normality. Firstly, BV-CM proposed in (Beckers and Vennekens 2018), distinguishes itself by extending causal models with a timing function. Thereby, allowing it to deal with a large range of controversial examples with time critical scenarios. Moreover, rather than just building on the Halpern and Perl definitions, this approach builds heavily on Hall's separation of causality into production and counterfactual dependence. From there they construct their definition according to an collection necessary and sufficient conditions derived from common examples that are allegedly inherent to causation. Secondly, (Fenton-Glynn 2017) generalises HP-05 , which is solely concerned with deterministic cases, to the probabilistic case. While not demonstrated in the article in question, they conjecture that their natural probabilistic extension is set up to deal with a wide range of examples circulating in the literature. Furthermore, they claim that they improved upon a previous attempt by (Twardy and Korb 2011) where they tried to generalise causal models to the probabilistic level using "Probabilistic Active Paths", thus their approach is abbreviated here with PAP .
Before discussing definitions that use languages other than causal models, it must be noted that there are additional, but rather rudimentary definitions of token causality. For reference see (Joseph Y Halpern and Hitchcock 2011), (Joseph Y Halpern and Hitchcock 2015), (Schaffer 2016) and (Weslake 2015).
Among those token causality definition that do not utilise causal models, there exists one block of formalisms using some kind of action language and another one that utilises CP-Logic. Apart from that there are two additional formal definitions and a multitude of (semi-)formal definitions.
Starting with the definitions that utilise action languages. In particular, there are two apparently equivalent definitions, SC-ACC and SC-CF introduced in (Batusov and Soutchanski 2018) and (Khan and Soutchanski n.d.) respectively, that leverage the expressive capabilities of situation calculus. Hence, both of them approach causality from a more procedural point of view. They

define the notion of achievement causal chains to identify token causes. With respect to the set S_A this definition is one of the few that can identify causes expressed in first-order logic, as many of the other approaches are, as of now, limited by the expressibility of their underlying language, which for the most part is bound to propositional logic with no natural path for generalisation in sight, e.g. causal models. In (Khan and Soutchanski n.d.) the authors remain faithful, relying on situation calculus once more. However, they introduce a revised definition that is while build in the shadows of the counterfactual tradition, still capable of of illustrating some of the accounts found in the regularity tradition, e.g. INUS. The approach presented in (LeBlanc, Balduccini, and Vennekens 2019), which shall be abbreviated with `AT` uses the action language $\mathcal{AL}$ and is therefore related to the approaches relying on situation calculus. However, when contrasted against those approaches $\mathcal{AL}$ is, according to (LeBlanc, Balduccini, and Vennekens 2019) not only better equipped for representing indirect effects, but also commands an (arguably) simple semantic.

Within the family of CP-Logic there are three definitions to be found. The first one is a definition of token causality in the original CP-Logic, which was defined in (Vennekens 2011) and will be called `BV-11`. Later the same authors presented a modified version of said definition, called here `BV-12` in the article (Beckers and Vennekens 2012). Those two definitions are take an inherent probabilistic view on causation. Moreover, given the semantics of CP-Logic it can be argued that those definitions contain a procedural element as well. Akin to the definitions form Halpern and Pearl they later extended `BV-12` to keep on par with `HP-05` after its normality extension, thus creating `HH-CP` in the process. Moving away from a probabilistic conception of causality, and more towards a process orientated view. Another definition in the wider context of CP-Logic can be found in (Denecker, Bogaerts, and Vennekens 2018; Denecker, Bogaerts, and Vennekens 2019), their definition is according to its creators constructed with a regularity theoretical perspective in mind. Furthermore, another benefit of this language is that the CP-Logic variant used allows one to reference causal processes within the language. Since they call the semantic underpinning their approach the possible causal process semantics, thus their definition will be abbreviated with `PCPS`.

The definitions `BCI` formulated in (Bochman 2018a) and `BReg` originating in (Baumgartner 2013) can not be categories as relying on either one of the above mentioned modelling languages, i.e. Causal Models, Action Languages and CP-Logic. Firstly, with `BCI`, Bochman coined a definition emerging out of the regularity theoretic tradition that uses two separate logics, namely causal theories introduced in (McCain, Turner, et al. 1997) and his logic of causal rules introduced in (Bochman 2004), for reasoning about causality. This is in stark contrast with `BReg`. Rather than relying a specialised language, in `BCI` cases even two, with a complicated and heavy semantic machinery, Baumgartner took great effort in requiring only fairly common

logical concepts. That is, his definition, which clearly follows the regularity tradition, relies only on material implication and some minimality constraint, thus allowing him to construct his definition using first-order logic only.
Before moving on to the more informal definitions, there are also some that are difficult to place. Among the publications in the S_A those definitions are primarily discussed in (Beckers and Vennekens 2016). All three of which were put forward by Hall, one termed `Hall-07` was coined in (Hall 2007) and the other two, called here `Hall-04p` and `Hall-04d`, are taken from (Hall 2004). Those definition were originally defined using neuron diagrams and structural equations. The two latter definitions reflect the Hall's view in (Hall 2004) that causality is separated into different relations, namely production and dependence.
Additionally, there are some definitions of token causality that are not completely formalised. Starting with the definition provided in (Liepiņa, Sartor, and A. Wyner 2019), using the language FCA, is categorised by its authors as semi-formal. Their definition called causal argument evaluation criteria or `CAEC` is strongly inspired by the NESS account. Moving fully into the informal realm, arguably the simplest definition in this category is the `But-For` or *sine qua non* test. Heavily used in the legal profession, it captures a highly simplified form of counterfactual reasoning. An improvement on the but-for test is the definition of Hart and Honore's called *causally relevant factors*, here abbreviated with `CRF`, which according to (Wright and Goldberg 2017) has its origins in (Hart and Honoré 1959). Later on, John Mackie introduced the so called `INUS`-condition (Insufficient but Necessary part of an Unnecessary but Sufficient condition) in (Mackie 1965). Wright, inspired by both of those accounts, formulates his `NESS`-account (Necessary Element of a Sufficient Set) in (Wright 1987). The latter two are both considered to be contributions to the regularity theoretic literature (Baumgartner 2013).
A honourable mention must be given to a definition from (Bex et al. 2010). This definition exists in the intersection of logic and law and is formulated using abductive causal theories. Its core idea is to check whether a given story (a chronological sequence of events) to explains a designated set of proposition using the provided abductive causal theory. Then they use an abstract argumentation framework to rank the stories based on their compliance with evidence. It was not considered a token causality definition, as it is not intended for extracting causally significant events from the given story.

3.2.2 Categorisation

In this subsection, the introduced definitions are categorised based on several aspects. Firstly, their popularity (wrt. S_A) will be investigated, secondly they are categorised based on the languages used and finally there will be an attempt to roughly identify tradition from which each of those definitions

emerged.
The popularity of a definition will be assessed based on the how many publications in S_A are mentioning the definition in question. To that extend Table 3.2 and Table 3.3 were constructed. Those tables depict which publication from S_A mentions which definition. On further analysis of the fully formalised definitions, it becomes quite apparent that `HP-05`, which is mentioned by 15 publications, is the most popular of the considered formalism by a considerable margin. As the distant second place is held by `HP-05d` with 4 mentions. However, with `HP-05d` being merely an extension of `HP-05`, this further solidifies its position as the benchmark formalism across the literature. Furthermore, with another definition from Hapern and Pearl `HP-01` being the third most popular, the dominance Halpern's ideas throughout the field becomes quite apparent. However, on closer look this assessment may change slightly. It is clear that this assessment favours old formalisms that were consistently discussed, investigated and refined. Since, this could be (and most likely is) done by the same author one can recalculate this ranking while ignoring self-references. When corrected for this, i.e. when considering publications form other authors only, `HP-05` is still referenced 11 times. However, the second place with 2 references is now shared by `HP-15`, `Wood-03` and `Hitch-01`.

	But For	CRF	INUS	NESS	Hitch-01	HP-01	Wood-03	Hall-04p	Hall-04d	HP-05	Hall-07	HP-08d	S	SJ
Vennekens, Bruynooghe, and Denecker 2010														
Bex et al. 2010														
Lee et al. 2010														
Lifschitz and F. Yang 2010														
Glymour et al. 2010							X			X			X	X
Claassen and Heskes 2010														
Gerstenberg and Lagnado 2010	X									X				
Joseph Y Halpern and Hitchcock 2011	X									X		X		
Shulz 2011														
Briggs 2012														
Baumgartner 2013			X	X										
Hyttinen, Hoyer, et al. 2013														
Joseph Y Halpern and Hitchcock 2015										X		X		
Weslake 2015					X	X	X			X				
Chockler et al. 2015										X				
Beckers and Vennekens 2016								X	X	X	X	X		
Schaffer 2016														
Joseph Y Halpern 2016b						X				X		X		
Blanchard and Schaffer 2017					X									
Wright and Goldberg 2017		X	X	X										
T. F. Icard, Kominsky, and Knobe 2017														
Aleksandrowicz et al. 2017						X				X				
Fenton-Glynn 2017										X				
Lagnado and Gerstenberg 2017	X													
Bochman 2018a				X										
Ibeling and T. Icard 2018														
Beckers and Vennekens 2018										X				
Bochman 2018b				X										
Denecker, Bogaerts, and Vennekens 2018										X				
Batusov and Soutchanski 2018										X				
Denecker, Bogaerts, and Vennekens 2019														
Liepiņa, Sartor, and A. Wyner 2019														
LeBlanc, Balduccini, and Vennekens 2019														
Liepiņa, Sartor, and A. Wyner 2020	X	X								X				
Khan and Soutchanski n.d.			X							X				
Ibeling and T. Icard 2020														

Table 3.2: Depicts which publication discuss which token causality definition (formulated before 2011).

	PAP	BV-11	BV12	B-Reg	PTC	HP-15	HH-CP	HP-05c	HP-05p	HmM	MbM	BV-CM	BCI	SC-ACC	PCPS	AT	SC-CF	CAEC
Vennekens, Bruynooghe, and Denecker 2010																		
Bex et al. 2010																		
Lee et al. 2010																		
Lifschitz and F. Yang 2010																		
Glymour et al. 2010																		
Claassen and Heskes 2010																		
Gerstenberg and Lagnado 2010																		
Joseph Y Halpern and Hitchcock 2011																		
Shulz 2011																		
Briggs 2012																		
Baumgartner 2013				X														
Hyttinen, Hoyer, et al. 2013																		
Joseph Y Halpern and Hitchcock 2015																		
Weslake 2015					X													
Chockler et al. 2015																		
Beckers and Vennekens 2016		X	X				X											
Schaffer 2016																		
Joseph Y Halpern 2016b																		
Blanchard and Schaffer 2017										X	X							
Wright and Goldberg 2017																		
T. F. Icard, Kominsky, and Knobe 2017																		
Aleksandrowicz et al. 2017																		
Fenton-Glynn 2017	X							X	X									
Lagnado and Gerstenberg 2017																		
Bochman 2018a													X					
Ibeling and T. Icard 2018																		
Beckers and Vennekens 2018												X						
Bochman 2018b													X					
Denecker, Bogaerts, and Vennekens 2018															X			
Batusov and Soutchanski 2018						X								X				
Denecker, Bogaerts, and Vennekens 2019															X			
Liepiņa, Sartor, and A. Wyner 2019																		X
LeBlanc, Balduccini, and Vennekens 2019																X		
Liepiņa, Sartor, and A. Wyner 2020																		X
Khan and Soutchanski n.d.						X											X	
Ibeling and T. Icard 2020																		

Table 3.3: Depicts which publication discuss which token causality definition (formulated from 2011 onwards).

The definitions can also be differentiated based on the original language used to formulate them. The data collected to do so can be viewed in Table 3.4. Here, it must be remarked that the depicted data are already aggregated based on language families. That is, rather that listing every variant in a particular language family, the family itself is used in the categorisation process. In particular this effects definitions using one of the CP-Logics, and definitions using one of the many extensions and generalisations of causal models. A mere glance at the data suffices, to further strengthen the claim that the ideas put forward by Halpern and Pearl shape the literature around causality. Namely, a total of 16 definitions rely on causal models in one form or another. Removing all definitions formulated by either Halpern or Pearl, still provides us with a total of 12 definitions. That is in stark contrast with the second most popular modelling language, CP-Logic. This language is only used by four definitions, all of which were either formulated by its creator Vennekens. Among those languages with zero definitions under their belt, particular mention must be given to neuron diagrams. Firstly, although no definition found in S_A uses this language, it is often applied as a modelling tool for conveying the type causal relations of an example in an intuitive manner. Secondly, as already mentioned (Erwig and Walkingshaw 2010) formulated a definition for token causes using neuron diagrams. Within the language families the definitions can be distinguished further. In the CP-Logic family, `BV-11` , `BV-12` use the original formulation, `HH-CP` uses a slight extension of the original language that allows for the expression of norms and `PCPS` relies on the deterministic second language in this family. With respect to causal models the following differentiations can be made. `HP-05d` , `MbM` (and arguably `HP-15` as it can be extended to a definition that incorporates normality) rely on causal models extended by a normality ranking. `HmM` used causal models where variables have default values. Both `PAP` and `HP-05p` use some form of probabilistic causal models and `BV-CM` require their causal models to be extended by a timing function.

Another categorisation of the discovered definitions can be made based on the traditions they follow. That is, whether the take a counterfactual, regularity, probability or process orientated view of causality. The assessment is primarily done by relying on assessments made in the literature. Nevertheless, the reasoning behind each of the judgement will be given.

Firstly, all those that belong to the counterfactual tradition

`But-For` was classified as belonging to the counterfactual Due to the fact `CRF` served as a source of inspiration for `NESS` , and thus shares strong similarities with it, the this definition was judged as being part of the regularity tradition by analogy.

Finally, Table 3.6 provides a summery of the token causal definition and their categorisation based in the dimension discussed in this section. Additionally, it provides information about their age and provides a reference to the publication of origin.

	Causal Model	CP-Logic	Situation Calculus	Non-Monotonic Causal Theory	First-order Logic	Neuron Diagrams	Conditional Logic	AL	SFCA	Abductive Causal Theory
But-For										
CRF										
INUS										
NESS										
Hitch-01	X									
HP-01	X									
Wood-03	X									
Hall-04p										
Hall-04d										
HP-05	X									
Hall-07	X									
HP-05d	X									
Simple	X									
SimpleJ	X									
PAP	X									
BV-11		X								
BV-12		X								
BReg					X					
PTC	X									
HP-15	X									
HH-CP		X								
HP-05c	X									
HP-05p	X									
HmM	X									
MbM	X									
BV-CM	X									
BCI				X						
SC-ACC			X							
PCPS		X						X		
AT										
SC-CF			X							
CAEC									X	

Table 3.4: This table depicts which definitions rely on which language family

	Counterfactual	Probability	Process	Regularity
But-For	X			
CRF				X
INUS				X
NESS				X
Hitch-01	X			
HP-01	X			
Wood-03	X			
Hall-04p	X			
Hall-04d	X			
HP-05	X			
Hall-07	X			
HP-05d	X			X
Simple	X			
SimpleJ	X			
PAP	X	X		
BV-11	X	X		
BV-12	X	X		
BReg				X
PTC	X			X
HP-15	X			
HH-CP	X	X		X
HP-05c	X			
HP-05p	X	X		
HmM	X			X
MbM	X			X
BV-CM	X		X	
BCI				X
SC-ACC			X	
PCPS			X	X
AT			X	
SC-CF	X		X	
CAEC				X

Table 3.5: This table depicts which definitions follow which tradition.

	Year	References	Language	Approach	Source
But-For	?	4	NL	CF	?
CRF	1959	2	NL	RE	(Hart and Honoré 1959)
INUS	1965	3	NL	RE	(Mackie 1965)
NESS	1987	4	NL	RE	(Wright 1987)
Hitch-01	2001	2	CM	CF	(Hitchcock 2001)
HP-01	2001	3	CM	CF	(Joseph Y Halpern and Pearl 2005)*
Wood-03	2003	2	CM	CF	(Woodward 2005)°
Hall-04p	2004	1	?	CF	(Hall 2004)
Hall-04d	2004	1	?	CF	(Hall 2004)
HP-05	2005	15	CM	CF	(Joseph Y Halpern and Pearl 2005)
Hall-07	2007	1	CM	CF	(Hall 2007)
HP-05d	2008	4	CM+N	CF,RE	(Joseph Y Halpern 2008)
Simple	2010	1	CM	CF	(Glymour et al. 2010)
SimpleJ	2010	1	CM	CF	(Glymour et al. 2010)
PAP	2011	1	CM+P	CF, PR	(Twardy and Korb 2011)
BV-11	2011	1	CP	CF, PR	(Vennekens 2011)
BV-12	2012	1	CP	CF, PR	(Beckers and Vennekens 2012)
BReg	2013	1	FOL	RE	(Baumgartner 2013)
PTC	2015	1	CM	CF,RE	(Weslake 2015)
HP-15	2015	2	CM	CF	(J. Halpern 2015)
HH-CP	2016	1	CP+N	CF, PR, RE	(Beckers and Vennekens 2016)
HP-05c	2017	1	CM	CF	(Fenton-Glynn 2017)
HP-05p	2017	1	CM+P	CF, PR	(Fenton-Glynn 2017)
HmM	2017	1	CM+D	CF, RE	(Blanchard and Schaffer 2017)
MbM	2017	1	CM+N	CF, RE	(Blanchard and Schaffer 2017)
BV-CM	2018	1	CM+T	CF, PO	(Beckers and Vennekens 2018)
BCI	2018	2	CT	RE	(Bochman 2018a)
SC-ACC	2018	1	SC	PO	(Batusov and Soutchanski 2018)
PCPS	2018	2	CP2	PO, RE	(Denecker, Bogaerts, and Vennekens 2018)
AT	2019	1	AL	PO	(LeBlanc, Balduccini, and Vennekens 2019)
SC-CF	2020	1	SC	CF, PO	(Khan and Soutchanski n.d.)
CAEC	2020	2	FCA	RE	(Liepiņa, Sartor, and A. Wyner 2020)

Table 3.6: Summary of the token causality definitions discussed. The language used are Action Language $\mathcal{AL}$ (AL); Causal Models (CM); Causal Models with Default/Deviant distinction (CM+D); Causal Models with Normality ranking (CM+N); Probabilistic Causal Models (CM+P); Causal Models with Timing (CM+T); CP-Logic (CP); CP-Logic with Norms (CP+N); CP-Logic: Causal Logic (CP2); Causal Theories (CT); Natural Language (NL); Situation Calculus (SC). The approaches considered are the Counterfactual approach (CF); a Process Oriented approach (PO); the Regularity Theoretic approach (RE); a (explicit) Probabilistic approach (PR). (*: Original in 2001; °: Original in 2003)

Chapter 4

Properties of Causation: A Collection of Examples

Currently amidst the quest of formally capturing causation, the literature concerned with assessing the capabilities of certain formalisms against a diverse array of examples. Those examples, increasingly complex, attempt to capture fragments of causality, as intuitively understood by humans. With many authors proposing new examples to highlight points of failures for previously established formalisms, the literature concerning causation has amassed a wealth of such examples. Unfortunately, as far as I am aware, no attempt was made to collect and classify those examples. Hence, the subsequent section shall be an approximation of precisely that. However, being fully aware that a mere list of possibly redundant examples is of little use, thus only prominent examples will be discussed while the remaining examples will be relegated to the database accompanying this work.

4.1 Examples

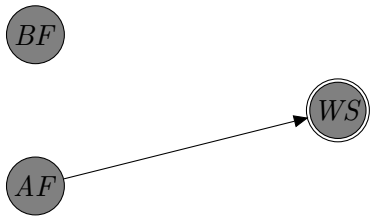
4.1.1 Basic Examples

Example 4.1.1 was used by (Beckers and Vennekens 2018) to introduce their first fundamental principle of causation, counterfactual dependence. Which intuitively can be characterised as follows. Assuming that both A and B occurred, if A had not occurred, the B would not have occurred either. Hence, in this case B is counterfactually dependent of B.

Example 4.1.1. Alice fires a bullet at a window (AF). Regardless of whether Bob fires his bullet or not (BF), he will not hit the window. Alice's bullet hits the window, at which point it shatters (WS).

	Basic	Sym. Overdet.	Switching	Late Preemp.	Early Preemp.	Double Preemp.	Bogus preemp.	Short Circuit
Glymour et al. 2010			X	X		X		
Gerstenberg and Lagnado 2010								
Joseph Y Halpern and Hitchcock 2011	X	X	X	X			X	
Baumgartner 2013		X	X	X	X		X	X
Joseph Y Halpern and Hitchcock 2015	X	X			X		X	X
Weslake 2015		X	X	X	X		X	X
Chockler et al. 2015		X		X			X	
Beckers and Vennekens 2016				X	X			
Joseph Y Halpern 2016b				X				
Blanchard and Schaffer 2017	X	X			X		X	X
Wright and Goldberg 2017		X		X	X			
Aleksandrowicz et al. 2017				X				
Fenton-Glynn 2017					X			
Lagnado and Gerstenberg 2017								
Bochman 2018a		X	X	X	X		X	
Beckers and Vennekens 2018	X	X	X	X	X	X	X	X
Bochman 2018b				X				
Denecker, Bogaerts, and Vennekens 2018		X	X	X		X	X	
Batusov and Soutchanski 2018	X	X	X		X			
Denecker, Bogaerts, and Vennekens 2019		X	X	X	X	X	X	
Liepiņa, Sartor, and A. Wyner 2019								
LeBlanc, Balduccini, and Vennekens 2019								
Liepiņa, Sartor, and A. Wyner 2020		X		X				
Khan and Soutchanski n.d.				X				

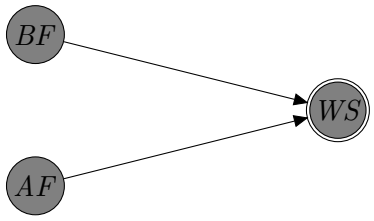
Table 4.1: Depicts which publication discuss which token causality definition (formulated from 2011 onwards).



Intuitively one would declare AF to be the cause of WS. This can be established by using basic counterfactual reasoning. That is, if Alice would not have fired the bullet, the window would not have shattered. Moreover, as BF is not considered to be a cause as the value of BF is immaterial in determining the value of WS. While this example is fairly trivial, even small modifications suffice to create disagreement among definitions and intuitions alike (see Example 4.1.3).
The following example, originally given in the context of forest fires, is frequently used by Halpern, e.g. (Joseph Y Halpern and Hitchcock 2011; Joseph Y Halpern and Hitchcock 2015), as a benign introductory example. Especially in the context of his definitions for token causality (`HP-01`, `HP-05` and `HP-15`), the impact of minimality conditions within can be observed.

Example 4.1.2. Alice (AF) and Bob (BF) each fire a bullet at a window, simultaneously striking the window. The window only shatters (WS), if it is hit by two bullets. What caused the window to shatter?

In the neuron diagram below the only thing of note is the use of a stubborn neuron that fires only if more than two stimuli are received.



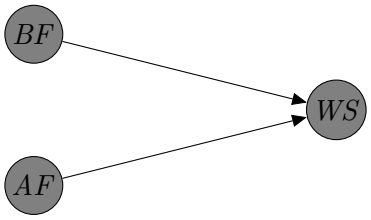
Already in such a small example, it becomes difficult to asses what a token cause should be. The first possibility would be to consider AS, BS and the conjunct of AS and BS as causes for WS. The second possibility is the rejection of the conjunct as cause for WS. That is, only AS and BS are declared as such. The third possibility contrasts the previous one by declaring the conjunct as the sole cause of WS. For example, The minimality constraint present in all of Halpern's definitions forces those definitions to reject the conjunct (Joseph Y Halpern 2016a, p. 28).

4.1.2 Symmetric Overdetermination

Symmetric Overdetermination is one of the problems that provides some difficulties to the counterfactual approach, leading to development of fairly

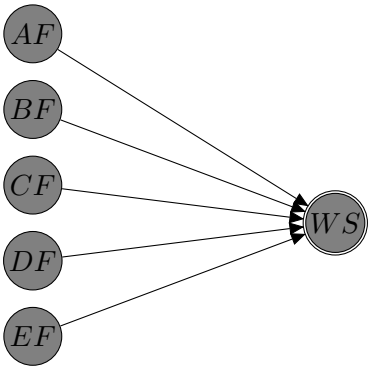
complicated analytical tools. Intuitively, an outcome can be considered overdetermined, if there are multiple processes, which produce said outcome, terminating at the same time. The subsequent example is a variant of the forest fire example found in several of Halpern's publications, e.g. (Joseph Y Halpern and Hitchcock 2011; Joseph Y Halpern and Hitchcock 2015). Moreover, this variants of this examples are heavily used discussed, e.g. (Glymour et al. 2010; Joseph Y Halpern and Hitchcock 2011; Baumgartner 2013; Joseph Y Halpern and Hitchcock 2015; Weslake 2015; Blanchard and Schaffer 2017; Wright and Goldberg 2017; Bochman 2018a; Beckers and Vennekens 2018; Denecker, Bogaerts, and Vennekens 2018; Batusov and Soutchanski 2018; Denecker, Bogaerts, and Vennekens 2019; Liepiņa, Sartor, and A. Wyner 2020).

Example 4.1.3. Alice (AF) and Bob (BF) each fire a bullet at a window, simultaneously striking the window, shattering it (WS). What caused the window to shatter?



According to Hiddleston 2005 the correct solution for Example 4.1.3 for this answer seems to be subject of contention. That is, it is unclear whether AF or BF individually should be considered a cause or whether the conjunct of AF and BF are the sole cause of WS.
By expanding Example 4.1.3 slightly, one obtains Example 4.1.4. Variants of this example can be found in (Glymour et al. 2010; Chockler et al. 2015).

Example 4.1.4. Alice (AF), Bob (BF), Carol (CF), Dave (DF) and Eve (EF) all fire at a window. The window shatters after three hits (WS). What is the cause of the window shattering?

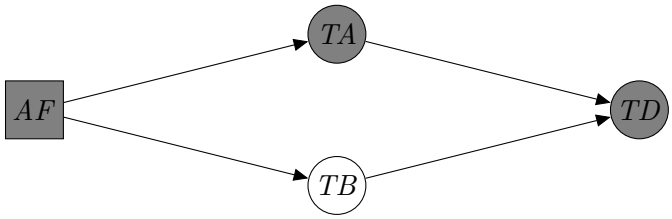


Given the story presented in Example 4.1.3, most accounts, i.e. `HP-05`, `PTC`, `BV-CM`, `BCI` and `PCPS` conclude that AF and BF individually are considered to be the cause of WS (Beckers and Vennekens 2018; Bochman 2018a; Denecker, Bogaerts, and Vennekens 2018; Weslake 2015; Joseph Y Halpern 2016a). However, notes are to be taken. `HP-05` considers AF and BF as the sole cause of WS. By contrast, `HP-15` only considers the conjunct $AF \wedge BF$ as the cause of WS. This somewhat counter-intuitive inference is justified, by arguing that AF and BF individually should be considered as parts of causes rather than a cause. Alternatively, he proposes that it would be best to view $AF \vee BF$ as the cause, which would hold in `HP-05` and `HP-15` (Joseph Y Halpern 2016a, p. 29). Mirroring the intuition given for `HP-15`'s result, (Beckers and Vennekens 2018) argue that while WS is not dependent on either AF and BF, both contribute to WS and thus both should be considered a contributing cause. Their account, i.e. `PCPS` differentiates between counterfactually irrelevant and strongly counterfactually irrelevant. They wager that strongly counterfactually irrelevant variables should not be considered causes, while counterfactually irrelevant could still be considered causes. Since they classify AF and BF only counterfactually irrelevant

4.1.3 Switching

Examples discussing switching are usually build as follows. There is a variable representing some form of action, e.g. the flicking of a switch, irrespective of the variable's value a causal process is triggered. Each of those processes produce the same outcome. Hence, the original action was immaterial in the occurrence of said outcome. For the binary case, this effect can be observed in the following example, variants of which can be found in (Glymour et al. 2010; Joseph Y Halpern and Hitchcock 2011; Baumgartner 2013; Weslake 2015; Bochman 2018a; Beckers and Vennekens 2018; Denecker, Bogaerts, and Vennekens 2018; Batusov and Soutchanski 2018; Denecker, Bogaerts, and Vennekens 2019).

Example 4.1.5. Alice flicks a switch (AF). The train travels on track A (TA), otherwise the train would have travelled on track B (TB). In both cases the train arrives at its destination (TD). Was AF the cause of TD?



With the flicking of the switch being immaterial, it proclaimed in (Beckers and Vennekens 2018) that most people would reject calling AF a cause of

TD. Although this view is not uncontroversial, especially as embracing it requires one to accept that causation is not transitive. That is, it is clear that AF is the cause of TA, yet it is not the cause of TD. More about this discussion can be found in XXXXX. Moreover, the formalisation presented in Example 4.1.5 is also subject of contention. In particular, in the given model AT and BT being logically independent. However, this would allow for the possibility of the train being on two tracks at the same time. One possible method of mitigating this inaccuracy is to replace the variables AT and BT with the exogenous variables BA and BB. Those variables indicate whether a track is blocked or not. For example, if BA and AF, then TD would not hold, while in the case of $\neg AF$ the train would still arrive (Joseph Y Halpern and Hitchcock 2011).
In the context of switching there are examples were the intuition is less clear. That is, consider Example 4.1.6, which is discussed in (Weslake 2015; Bochman 2018a)

Example 4.1.6. Alice pushes Bob. Therefore, Bob is hit by a truck. Bob is dies. Otherwise, Bob would have been hit by a bus, which would have killed him as well.

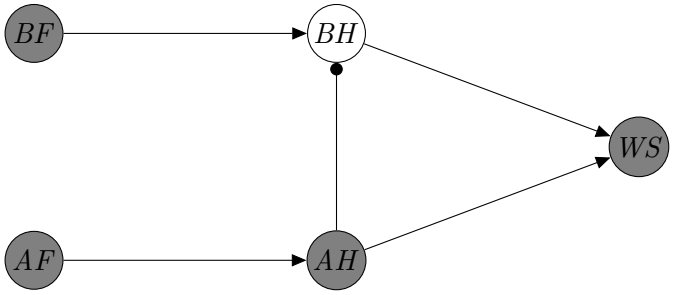
In Example 4.1.6 one is clearly faced with an instance of switching. However, (McDermott 1995) claims that intuition would dictate that Alice did in fact kill Bob. (Weslake 2015) argues that this intuition is a product of hidden assumption, namely that the assumption (or hope) there would be another option, like "Push Bob to safety" would be available. He therefore, claims that Example 4.1.6 is underspecified. To retain the intuition regarding switching he suggests that one should declaring the available option as exhaustive and only omit those options that result in the same outcome. As otherwise, i.e. the case where one mentions that the modelled options are not exhaustive and omits some options with a different outcome, one should consider the switch, in this case Alice's action, a cause.
Given the story of Example 4.1.5, the definitions `HP-05` and `HP-15` claim AF to be the cause of TD. While the definitions `PTC`, `BV-CM`, `BCI`, `SC-ACC`, `SC-CF` and `PCPS`, do not (Beckers and Vennekens 2018; Bochman 2018a; Denecker, Bogaerts, and Vennekens 2018; Weslake 2015; J. Halpern 2015; Batusov and Soutchanski 2018). To bridge the gap between `HP-05`, `HP-15` and the rest of the definitions, (J. Halpern 2015) resolves this issue by appealing to normality. Furthermore, It must be noted that `HP-15` does not declare AF a cause of TD, if one chooses the formalisation using BA and BB instead. By contrast, `HP-05` still declares AF to be a cause (J. Halpern 2015).

4.1.4 Late Preemption

Late preemption, is on the surface similar to symmetric overdetermination. In fact, it is sometimes called asymmetric overdetermination (Erwig and

Walkingshaw 2010). In particular, one can refer to late preemption, if there are two causal processes running in parallel. Each of them would produce the same outcome. However, as one process terminates before the other does. Thereby, bringing forth the outcome and rendering the second process irrelevant. Hence, late preemption distinguishes itself from symmetric overdetermination, based on the fact that the two running processes are temporally not aligned (Beckers and Vennekens 2018).

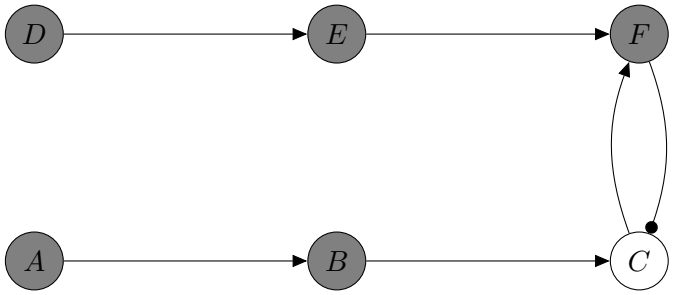
Example 4.1.7. Alice (AF) and Bob (BF) each fire a bullet at a window. Alice's bullet hits the window first (AH). The window shatters (WS). Bob's bullet arrives second and does not hit the window (BH). What caused the window to shatter?



The square neuron represents a switch. That is, depending on the status of the switch either the first or the second direct successor neuron will fire.

The intuition with this example is clear. AF is the cause of WS. Because, Alice's bullet prevents Bob's bullet to hit the widow, by hitting it earlier. Hence, BF can not be a cause of WS.
The similarities to symmetric overdetermination, demonstrate serve as a good example for highlighting the importance of proper modelling. As neglecting to model the relationship between Alice's and Bob's actions, results in a case of symmetric overdetermination. The model presented avoids this by adding additional variables, i.e. AH and BH. That is, there must be a variable that exhibits a different value depending the actual cause (Joseph Y Halpern and Hitchcock 2011). An alternative approach suggested in (Joseph Y Halpern 2016a, p. 34) is to encode temporal information into the model by introducing time indexed variables. Irrespectively of the additional difficulty of constructing a suitable model, the alternative approach is arguably the more intuitive one. In particular, (Beckers and Vennekens 2018) criticise the formalisation presented in Example 4.1.7 on the grounds that that AF and BF trigger entirely different mechanisms. Hence, constructing a model that incorporates such a relationship is conceptually wrong. Especially, as it is obvious that Bob failed to hit the window, considering that Alice hit the window before him, i.e. Bob was too late. Hence, the addition of AH and BH simply hide the temporal aspect of the story, by implicitly encoding

the order at which the bullets would hit the window, without explicitly assigning a timing to the events. The state that the principle of "causes come before - or at most simultaneous with - effects" is accepted across the board. Therefore, they extend causal models with a timing function.
Abstracting away from the story (Baumgartner 2013) provides a different late preemption example using solely neuron diagrams.



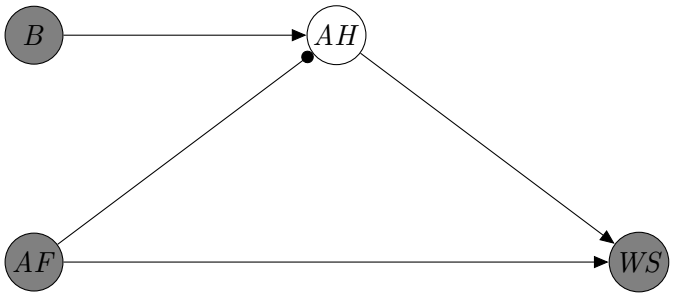
Mapping this formalisation onto the story Example 4.1.7. F would be WS, E would be AF, B would be BF and D would be AF (A has no counterpart). Meaning that given this structure, it was the shattering of the window that prevented Bob from hitting the window. Clearly, a sensible formalisation of the problem. This, arguably more intuitive, formalisation, however, requires circularity. Given the given context having such a circular dependence is innocuous. Unfortunately, a slight change of the exogenous variables, i.e. setting AF to false, results in the shattering of the window, without Bob's bullet hitting it. XXX Hence, while (Baumgartner 2013) claims that this is a canonical representation of late preemption, it seems to differ from the common formalisation of Example 4.1.7.
The accounts `HP-05` , `HP-15` , `PTC` , `BV-CM` , `BCI` , `SC-ACC` , `SC-CF` and `PCPS` satisfy the provided intuition for the story presented in Example 4.1.7. (Beckers and Vennekens 2018; Bochman 2018a; Denecker, Bogaerts, and Vennekens 2018; Weslake 2015; Khan and Soutchanski n.d.; Joseph Y Halpern 2016a, p. 33) `BV-CM` being equipped with a timing function, it can serve the provided intuition wile using a model similar to the symmetric overdetermination one (Beckers and Vennekens 2018).

4.1.5 Early Preemption

Many authors consider early and late preemption to be the same (or at least similar), thus they resolve examples discussing early preemption in a similar fashion. The difference between those two is, that in early preemption the outcome of the two processes occurred before the second process was in motion (Beckers and Vennekens 2018). Example 4.1.8 seems to be a canonical example for this effect, variants of it can be found in (Baumgartner 2013; Joseph Y Halpern and Hitchcock 2015; Weslake 2015; Beckers and

Vennekens 2016; Blanchard and Schaffer 2017; Wright and Goldberg 2017; Fenton-Glynn 2017; Bochman 2018a; Beckers and Vennekens 2018; Batusov and Soutchanski 2018; Denecker, Bogaerts, and Vennekens 2019).

Example 4.1.8. (Early Preemption) Alice fires a bullet at the window (AF). If Alice hits the window (AH), the window shatters (WS). If Alice does not hit the window, Bob fires a bullet at the window (BF), hitting it (BH) leading to its shattering. What caused the window to shatter?
In the neuron diagram the neuron B is used to ensure that BF fires per default. That is, it allows BF to fire if A does not fire, i.e. they together encode the behaviour of a negation.



As eluded to earlier, the intuition is similar to the late preemption one. That is, AF is attributed to be the cause of WS, while BF is naturally not a cause of WS. However, this straight forward analysis, is deceptive. (Beckers and Vennekens 2018) noticed that early preemption has a close relationship with switching. Alice, assuming that she is certain that Bob will shoot at the window if she neglects to do so, is faced with a choice. Either she shoots the window and it shatters or Bob will shoot at the window, shattering it in the process. Regardless, the status of the window is independent of her decision. In fact, Alice can only choose the causal path responsible for shattering the window, i.e. she can decide how and not that the window shatters. Hence, it is a case of switching. To further strengthen the similarity (Denecker, Bogaerts, and Vennekens 2019) add an additional variable to the model, representing the bullet leafing the gun. In this case let it be AH representing that Alice hit the window. This produces a model that is isomorphic to Example 4.1.5. They claim, not uncontested, see (Weslake 2015), that adding this variable should not influence the intuition about what is the cause of what.
Excepting that, one must explain the conflicting intuitions. Some argue for the necessity of probability or some other method of expressing uncertainty in the success of the causal process (Beckers and Vennekens 2018; Hall 2007). For example, one could argue that people assume the Bob firing and hitting the window may not succeed, while the arrival of the train will always succeed. This view seems to be supported by the fact that if one

attaches probabilities of arrival to the respective railway tracks found in Example 4.1.5, some causal attribution is warranted. For example, if on track A the train has a 99% chance of arrival and on track B the train has a 1% chance, then Alice's flicking of the switch contributed in the trains arrival. Such an asymmetry can be observed in an example taken from (Beckers and Vennekens 2018)

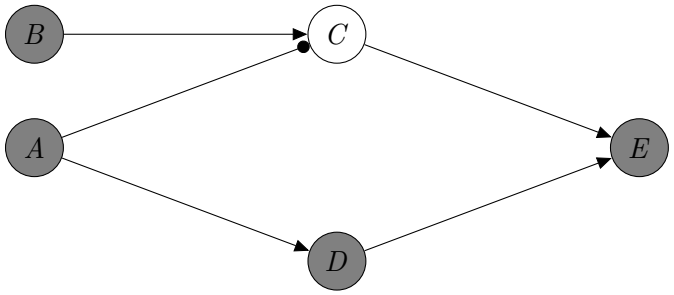
Example 4.1.9 (Beckers and Vennekens 2018)**.** Suppose Alice reach out and catch a passing cricket ball. The next thing along in the ball's direction of motion was a solid brick wall. Beyond that was a window.

People tend to classify this example as an instance of switch. That is, catching the ball is immaterial for the status of the window. However, by replacing the wall with another person Bob, this intuition shifts, declaring Alice's action to be the causal for the well being of the window. The presumption is that, this asymmetry arises due to the fact the prospect of the wall failing to stop the ball is not taken seriously (Beckers and Vennekens 2018; Blanchard and Schaffer 2017).
(Denecker, Bogaerts, and Vennekens 2019) resolve this issue by distinguishing between enabling and triggering conditions. The prior preempts the causal mechanism if not present, while the latter setts the mechanism in operation. Considering the discussed cases. Alice's flicking of the switch would be merely an enabling condition for the train taking track A or track B. By contrast, Alice's firing of the bullet is clearly a triggering condition for the bullet hitting the window.
Another example that highlight the similarities and differences between switching and early preemption is Example 4.1.10.

Example 4.1.10 (Weslake 2015)**.** Two two-state switches are wired to an electrode. The switches are controlled by A and B respectively, and the electrode is attached to C. A has the first option to flip her switch. B has the second option to flip her switch. The electrode is activated and shocks C if both switches are in the same position. B wants to shock C, and so flips her switch iff A does.

This example shares similarity with switch and early preemption, it can be found in (Weslake 2015; Bochman 2018a). While structurally similar to early preemption, one could argue that the action of A does not trigger the shocking of C and thus it should rather be considered to be a case of switching. Intuition would suggest that interpreting this example as a species of switch, would be more apt. That is, A having no choice in the matter should not be considered a cause of C.
Furthermore, similar to late preemption, (Baumgartner 2013) attributes early preemption a structure that is (slightly) different to the one presented in Example 4.1.8. However, in this case the structure actually corresponds

with the model extension used by (Denecker, Bogaerts, and Vennekens 2019) to further highlight the similarities to switching. That is, A is AF; C is BF; D is AH; R is WS; B is is used as mentioned in Example 4.1.8



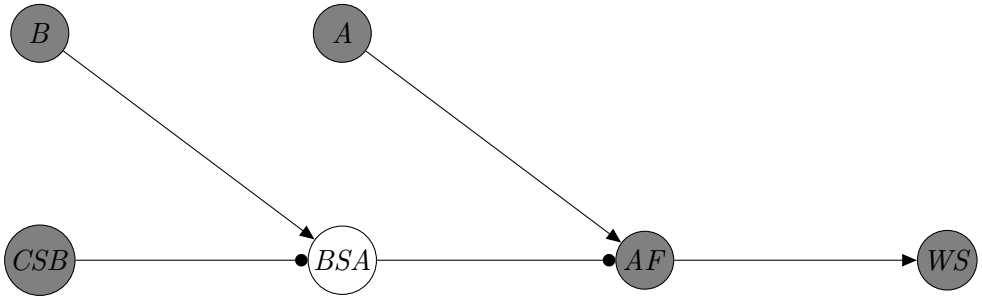
By comparing the neuron diagrams of Example 4.1.7 and Example 4.1.8 it is hard to ignore their striking similarity. This would explain the fact that they are sometimes treated alike. So what is exactly the difference between those two phenomena? When observing the neuron diagrams provided in (Baumgartner 2013) the distinction can be characterised as follows. One speaks of early preemption, if the backup process, here BF, is made redundant before the effect occurs. Whereas, in late preemption the backup process is interrupted by the effect itself. An alternative distinction, claims that the characteristic feature of early preemption is that the process in question is actually interrupted by another process, while in late preemption the process is never interrupted, it simple never has the opportunity to terminate (Baumgartner 2013).
The intuition presented through the story in Example 4.1.8 is satisfied by `HP-05`, `HP-15`, `PTC`, `BCI`, `SC-ACC`, `SC-CF` and `PCPS` (Bochman 2018a; Denecker, Bogaerts, and Vennekens 2018; Weslake 2015; Batusov and Soutchanski 2018). This result is deliberately not shared by `BV-CM`. As eluded to earlier, (Beckers and Vennekens 2018) understands the usual formalisation of this example as switch. Hence, `BV-CM` does not declare AF to be the cause of WS. However, by properly extending the model with variables representing the accuracy of either Alice or Bob, AF becomes a cause. In the case of `BCI`, Alice is not only the cause of the window's shattering when she fires her bullet, but also when she does not. This result is particularly interesting in the context of the discussion in (Beckers and Vennekens 2018). That is, in (Bochman 2018a) AF could be considered as a variable that switches between two processes with the same outcome. However, rather than intuition declaring the switch to be irrelevant for the occurrence of the outcome, it is always considered a cause of the outcome.

4.1.6 Double Preemption

One speaks of double preemption if there is an process instigated to interrupt some process is prevented by yet another process. Hence, the potential

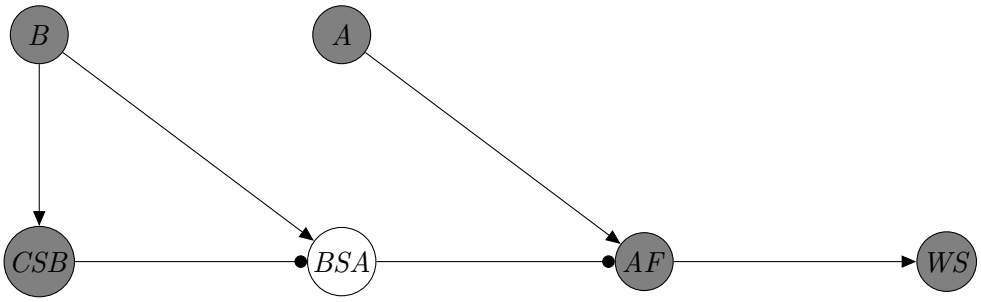
preempter is preempted. Put differently, the binary variable A would not hold if the binary variable B holds. However, B does not hold because of C (Denecker, Bogaerts, and Vennekens 2019). Variants of Example 4.1.11 can be found in (Glymour et al. 2010; Beckers and Vennekens 2018; Denecker, Bogaerts, and Vennekens 2018; Denecker, Bogaerts, and Vennekens 2019).

Example 4.1.11. Alice intends to fire a bullet at a window (A). Bob is intends to prevent Alice from hitting the window (B). Bob tries to stop Alice (BSA). Bob is stopped by Carol (CSB). Alice fires a bullet (AF), hits the window (AH) and shatters it (WS). The window shatters (WS). What caused the window to shatter?



The inclusion of A ensures that Alice fires and hits the window by default. That is, one could remove A, however, this would require the specification of a default value for AF. As otherwise, regardless of the specified context AF, being an endogenous variable, would never fire. The same holds for B.

According to (Joseph Y Halpern 2016a, p. 35), the intuition for this example is to attribute not only AF, but also CSB with being the causes of WS. This intuition is satisfied by all of its definitions, thus in particular `HP-05` and `HP-15` satisfy it. By contrast, `PCPS` does not deem CSB to be the cause of WS. However, they argue that their definition can easily adapted into a sate of compliance with Halpern's proposed intuition (Denecker, Bogaerts, and Vennekens 2019; Joseph Y Halpern 2016a, p. 36).
Moreover, an issue arises in the case were Bob never intends to stop Alice, i.e. where B never fires. Here intuition would dictate that clearly CSB can not be a cause of WS. This slight change of context produces counterintuitive inferences in some formalisms. Halpern suggest that this issue is the result of a to simplistic model. That is, Carol can only stop Bob if Bob actually tries to stop Alice, a causal dependence clearly not present within the model of Example 4.1.11 (Joseph Y Halpern 2016a, p. 36).

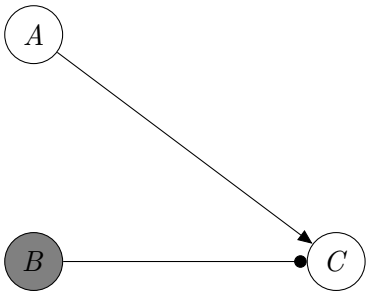


Lastly, it should be noted that it does not stop with double preemption. Meaning, one could extend the causal chain by adding a fourth party preventing Carol stopping Bob (Denecker, Bogaerts, and Vennekens 2018).

4.1.7 Bogus Preemption

In essence, bogus preemption or bogus prevention occurs when an action is taken to interrupt an inactive process. Examples discussing bogus preemption can be found in (Joseph Y Halpern and Hitchcock 2011; Baumgartner 2013; Joseph Y Halpern and Hitchcock 2015; Weslake 2015; Chockler et al. 2015; Blanchard and Schaffer 2017; Bochman 2018a; Beckers and Vennekens 2018; Denecker, Bogaerts, and Vennekens 2018; Denecker, Bogaerts, and Vennekens 2019). The canonical example, called "Careful Antidote", revolves around a poisoned water.

Example 4.1.12. Alice is in possession of a lethal poison, but has a last-minute change of heart and refrains from putting it in Carol's water (A - A is true if Alice does *not* poison the water). Bob puts antidote in the water (B), which would have neutralized the poison. Carol drinks the water and survives (C).



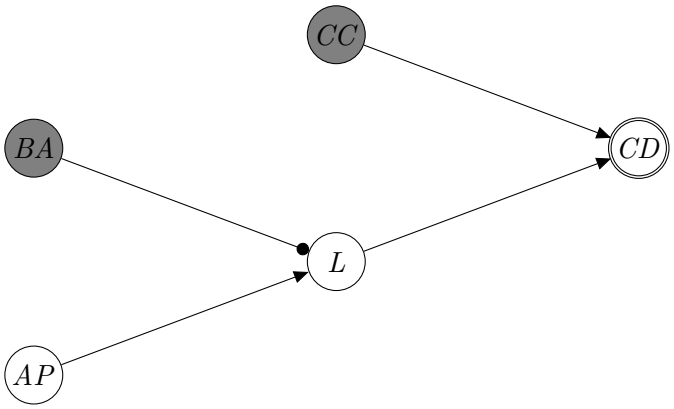
The formalisation in Example 4.1.12 is used to demonstrate the limitation of structural equations, because while being isomorphic to the example of symmetric overdetermination the intuition underlying both phenomena are vastly different. That is, Carol survives if Alice refrains from poisoning the water or if Bob adds the antidote to the water. Hence, C fires if either A or B fires. Hence, one is confronted with symmetric overdetermination, indicating

that A and B or their conjunct should be considered a cause of C. Yet, in the context of the given story, neither A nor B should be considered a cause. (Baumgartner 2013) elegantly observes that symmetric overdetermination discusses the overdetermination of occurrences, while bogus preemption is concerned with overdetermined absence. (Joseph Y Halpern and Hitchcock 2011; Weslake 2015; Joseph Y Halpern and Hitchcock 2015).

One suggested solution to this problem is to appeal to some notion of normality. That is, in addition of formalising the causal structure, it is necessary provide the inference system with a theory of normality in order to derive suitable causes. The idea behind this approach is that one can use the notion of normality to exclude certain unreasonable contingencies. In this particular case, one could add a statement "typically, people do not put poison in the water." to the model. Thus, the scenario where Alice actually poisons the water is less "normal" than the actual scenario. Hence, given this normality assumption, one can classify Bob's action as completely redundant. Thereby, excluding it from being a cause. (Joseph Y Halpern and Hitchcock 2011; Joseph Y Halpern and Hitchcock 2015).

Another suggestion to resolve this issue is to adapt the model used to represent the scenario. The aptness of the presented model is directly criticised in (Blanchard and Schaffer 2017), where it is called it impoverish. The source of such harsh words, lies in the fact the coarse nature of this formalisation ignores vital information. To compensate this deficit (Blanchard and Schaffer 2017) suggest the inclusion of a variable indicating the toxicity of the water. In (Joseph Y Halpern and Hitchcock 2015) notes, that such an extension is arguably more preferable than introducing normality. While not explicitly criticising the approach from Example 4.1.12, another frequent formalisation extends the model by a variable encoding the drinking of the water. The prior is also found in (Bochman 2018a; Joseph Y Halpern and Hitchcock 2015), the latter is discussed in (Denecker, Bogaerts, and Vennekens 2018; Denecker, Bogaerts, and Vennekens 2019) and a combination of both can be found in (Beckers and Vennekens 2018). Being the most detailed, a variant of the last is presented in 4.1.13.

Example 4.1.13. Alice intents to put lethal poison into Carol's water (AP). However, Alice does not put lethal poison into Carol's water. Bob puts an antidote into Carol's water (BA). The water is lethal (L), if the poison is added without the addition of an antidote If Carol would consumes the lethal water she would die (CD). Carol consumes her water (CC). Carol does not die.



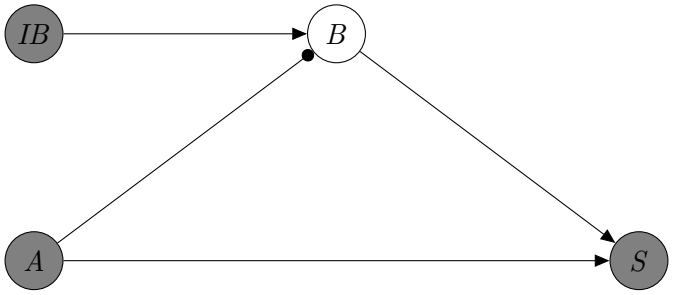
In (Weslake 2015), they use the simplistic formalisation found in Example 4.1.12. Being isomorphic to switch, their formalism, i.e. `PTC`, naturally concludes that both AP and BA causally influence the status of CD. A formalisation obtained by extending the one found in Example 4.1.12 with a variable that holds when the poison is neutralised `HP-05` does not declare BA to be a cause. However, on the same model, both `HP-05` and `HP-15` consider AF to be a cause (Joseph Y Halpern 2016a, p. 88). In (Bochman 2018a) they use use yet again a slightly different formalisation. That is, utilising `CT` they construct a theory expressing that AP and not BA causes CD; not AP causes not CD; A and B cause not CD. Using this, `BCI` concludes that only the absence of Alice poisoning the water is the cause of Carol's survival. In (Beckers and Vennekens 2018), they use their timing function to differentiate between between AP and BA. Therefore, if Alice's actions pre-date Bob's, then Alice's decision to refrain from poisoning the water is deemed to be the cause of Carol's survival by `BV-CM`. By contrast, if the order is reversed, the addition of the antidote would be classified as the cause. Additionally, if no timing is given this example is treated as a case of symmetric overdetermination. According to (Denecker, Bogaerts, and Vennekens 2019), adding an antidote does interrupt the mechanism activated by poisoning the water. However, as this never occurred the, it is impossible for BA to preempt an inactive mechanism. Hence, only the refusal of Alice to poison the water should be considered a cause of Carol's survival. Their discussed definition, i.e. `PCPS`, reflects this reasoning.
This example provides ample insights into the problem of omission in causal inference. XXXXX

4.1.8 Short Circuit

Variants of this example, often called "Careful Poisoning", can be found in (Baumgartner 2013; Joseph Y Halpern and Hitchcock 2015; Weslake 2015; Beckers and Vennekens 2018; Blanchard and Schaffer 2017).

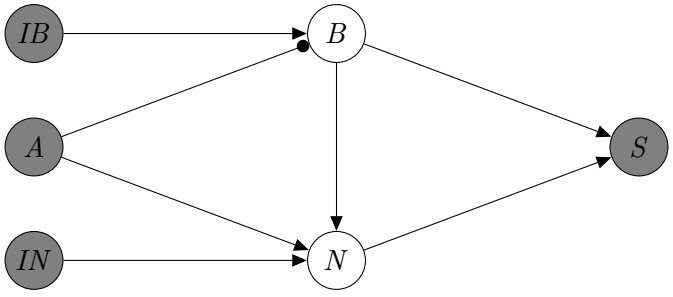
Example 4.1.14. (Careful Poisoning) Alice puts a harmless antidote in

Carol's water (A). Bob intended to not put poison into the water (IB). Seeing that the water contains an antidote Bob, adds the poison into the water (B - B holds if Bob does not administer the poison). This poison is countered by the antidote. Carol drinks the water and survives (S).



I was added to encode that per default Bob does not intend to poison the water. This is necessary due to the fact that neuron diagrams (in the common use) do not have an edge reserved for a negated relation.

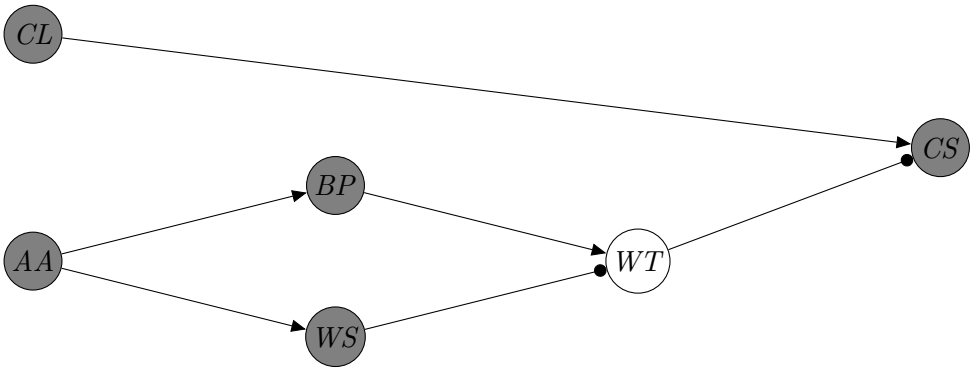
Similarly, to bogus preemption this formalisation of the presented story is isomorphic to the canonical early preemption case presented in Example 4.1.8. This would suggest that adding the antidote to the water caused the survival of Carol. While not entirely uncontested, intuition would dictate that neither A nor B should be considered a cause of S (Beckers and Vennekens 2018).
This example is the second one, to be referenced when talking about the limitations of structural equations and the necessity of extending causal models with some sense of normality ranking. Again this approach is criticised by (Blanchard and Schaffer 2017), citing the inaptitude of the model as source of the perceived similarities. Precisely as before, the argue that adding a variable tracking the lethality of the water is sufficient to resolve the issue of diverging intuitions on equivalent structures.



(Beckers and Vennekens 2018) takes a different approach. They argue that this issue has its origins in conflating early preemption and switching. That is, if Alice adds the antidote, then Bob will add poison to the the water, which promptly is neutralised allowing Carol to live on. Otherwise, Bob will

not add poison to the water and Carol will be unscathed. Hence, the actions of Alice are immaterial to the well-being of Carol. Therefore, this example should be modelled as an instance of switch. Thereby, realigning structure with intuition.
Given the formalisation found in Example 4.1.14 both `HP-05` and `HP-15` declare the addition of the antidote to be the cause of Carol's survival (Joseph Y Halpern 2016a, p. 90). The definition `PTC` arrives at the same conclusion (Weslake 2015). As already mentioned in (Beckers and Vennekens 2018) Example 4.1.14 is deemed to be a switch variant. Therefore, their formalism, i.e. `BV-CM`, is constructed in such a manner that adding the antidote is immaterial for Carol's survival.
In (Baumgartner 2013) another structure is classified under the umbrella of short circuit. Example 4.1.15 labels the structure taken from (Baumgartner 2013) to operate (roughly) within the narrative presented in Example 4.1.14.

Example 4.1.15. (Careful Poisoning) Carol lives (CL) Alice puts a harmless antidote in Carol's water (AA). Adding antidote to the water, protects it against poison (WS - "water save") If Alice puts the antidote into Carol's water, Bob will poison the water (BP) Adding poison to an unprotected water makes it toxic (WT). If Carol would drink toxic water she would die (i.e. inhibiting CS). Carol drinks her water and survives (CS).



4.1.9 Other Examples

Example 4.1.16. There are a left and a right window. Alice and Bob both order Carol to fire at the left window. Carol fires at the left window, shattering it. Commands from Alice always trump commands form Bob (e.g. if Bob would have ordered to fire at right window, Carol would still have fired at the left one.). Without a command Carol would not have fired at all. What caused the left window to shatter?

Example 4.1.16 is a reformulation of the canonical example found in (Joseph Y Halpern and Hitchcock 2011; Weslake 2015). It was constructed to demonstrate the phenomena of trumping causation, which is elaborated on in

XXXX. Here intuitions conflict whether one should consider Alice alone, both individually or Alice and Bob as a conjunction to be the cause of the left window shattering (Joseph Y Halpern and Hitchcock 2011; Weslake 2015).

Example 4.1.17 (Joseph Y Halpern and Hitchcock 2015)**.** If there is hot weather, flowers will die. Watering prevents the flowers to die in hot weather. The neighbour does not water the flowers. The flowers die. What caused the flowers to die?

Example 4.1.17 is commonly used to discuss whether omissions should be considered as causes or not. Here in particular, is the demise of the flowers the cause of the neighbour dying? While at first glance intuition would side for "yes", the discussion in XXXXX indicates that this question is not as straight forward as it may seem. Some articles discussing the topic of omission and/or presenting its canonical examples similar to Example 4.1.17 are (Glymour et al. 2010; Joseph Y Halpern and Hitchcock 2011; Joseph Y Halpern and Hitchcock 2015; Blanchard and Schaffer 2017)

Example 4.1.18 (Joseph Y Halpern and Hitchcock 2011)**.** Suppose that Billy is hospitalized with a mild illness on Monday; he is treated and recovers. In the obvious causal model, the doctor's treatment is a cause of Billy's recovery. Moreover, if the doctor does not treat Billy on Monday, then the doctor's omission to treat Billy is a cause of Billy's being sick on Tuesday. But now suppose that there are 100 doctors in the hospital. Although only doctor 1 is assigned to Billy (and he forgot to give medication), in principle, any of the other 99 doctors could have given Billy his medication. Is the nontreatment by doctors 2–100 also a cause of Billy's being sick on Tuesday?

For the sake of completion, Example 4.1.18 is another common example used for discussing causation. It is a slight extension of Example 4.1.17.

Example 4.1.19 (Joseph Y Halpern and Hitchcock 2015)**.** The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mailed them reminders that only administrators are allowed to take the pens. On Monday morning, one of the administrative assis- tants encounters professor Smith walking past the receptionists desk. Both take pens. Later, that day, the receptionist needs to take an important message...but she has a problem. There are no pens left on her desk.

Example 4.1.19, can be found in (Beckers and Vennekens 2016; Joseph Y Halpern and Hitchcock 2015), is an example that demonstrates how norms influence causal attribution. For a proper discussion about normality, norms

and causation see XXXX. In this example, intuition would dictate that professor Smith caused the absence of pens. This intuition was empirically tested and confirmed in (Knobe and Fraser 2008).

Example 4.1.20 (Joseph Y Halpern and Hitchcock 2015)**.** Consider a fire that is caused by a lit match. While the fire would not have occurred without the presence of oxygen in the atmosphere, the oxygen is deemed to be a background condition, rather than a cause.

Example 4.1.20 is presented to discuss whether and how one should distinguish between causes and background conditions, a discussion held in XXXX. Depending on the position taken in this discussion, one would either accept or deny the presence of oxygen the status of cause (Joseph Y Halpern and Hitchcock 2015).

Example 4.1.21 (Joseph Y Halpern and Hitchcock 2015)**.** A lit match aboard a ship caused a cask of rum to ignite, causing the ship to burn, which resulted in a large financial loss by Lloyd's insurance, leading to the suicide of a financially ruined insurance executive. The executive's widow sued for compensation, and it was ruled that the negligent lighting of the match was not a cause (in the legally relevant sense) of his death.

The answer whether the sailor dropping a lit match should be charged with being the cause of another persons death is yet again uncertain. Due to its long causal chain, the cause is so far removed from the effect that intuition would disagree with a complete causal attribution. Hence, Example4.1.21 is well suited for highlighting issues with transitivity and binary causal attribution (Joseph Y Halpern and Hitchcock 2015). For a more detailed discussion see XXXXX.

Example 4.1.22 (Glymour et al. 2010)**.** A boulder slides toward a hiker, who, seeing it, ducks. The boulder misses him and he survives. Did the boulder sliding cause his survival?

Clearly, intuition would dictate that the boulder is not the cause of the hikers survival. Hence, switching not the only example cited, when it comes to discussing the transitivity of causality. In particular, Example 4.1.22 is the second example used in (Hitchcock 2001) to demonstrate that causation is not transitive (see XXXX).

Example 4.1.23 (Blanchard and Schaffer 2017)**.** Consider a case where Doctor can administer no dose, one dose, or two doses of medicine to Patient. Patient will fail to recover if no dose is administered, but will recover if either one or two doses are administered. Let us suppose that Doctor in fact administers two doses, and Patient recovers.

In (Blanchard and Schaffer 2017) Example 4.1.25 was used to argue that causation should be considered as contrastive. This example is designed to provide different intuition about causation, depending on which possible scenario is used to counterfactually contrast the actual scenario against (Blanchard and Schaffer 2017). For further discussion see XXXX.

Example 4.1.24 (Weslake 2015)**.** A firing squad consists of shooters B and C. It is A's job to load B's gun, C loads and fires his own gun. On a given day, A loads B's gun. When the time comes, only C shoots the prisoner.

Example 4.1.24 is relative common, being discussed in (Weslake 2015; Chockler et al. 2015; Joseph Y Halpern 2016b; Bochman 2018a). It is of particular relevance, as it was this example that lead to a reformulation of `HP-01`. While `HP-01` struggled with this example, its successor `HP-05` was able to match the intuitive answer that C was the one causing the prisoners death. However, (Joseph Y Halpern 2016b) demonstrated that with proper modelling `HP-01` can achieve the same inferences as `HP-05` on this example.

Example 4.1.25 (Blanchard and Schaffer 2017)**.** Consider a case where Doctor can administer no dose, one dose, or two doses of medicine to Patient. Patient will fail to recover if no dose is administered, but will recover if either one or two doses are administered. Let us suppose that Doctor in fact administers two doses, and Patient recovers.

This example is used in (Blanchard and Schaffer 2017) to argue that causation is contrastive, i.e. the causal status of a variable depends on the variable it is contrasted against. Here in particular, giving the patient two doses rather than zero doses caused the patient to recover. However, administering two doses rather than one dose did not cause the patient to recover. XXXXX

Example 4.1.26. Alice plans to go camping in June (AC). If there is a forest fire in May (FF_m), Alice will not go camping. If Alice goes camping, she will cause a forest fire (FF_j).

Example 4.1.26, taken from (Joseph Y Halpern and Hitchcock 2011), stirs one directly into a meta-physical discussion about the nature of events. Meaning if the occurrence of an event is delayed, is it still the same event. Neglecting this distinction during the modelling process may result in undesirable results. That is, if one would not have explicitly distinguished between the forest fire in May and the forest fire in June by using separate variables, the model would contain circularities and would allow for counter-intuitive inferences, e.g. creating a forest fire in June causes Alice to go camping (Joseph Y Halpern and Hitchcock 2011). XXXX

Example 4.1.27 (Blanchard and Schaffer 2017)**.** A ranch has five individuals: Cowboy C, Ranger R, Wrangler W, and two Hands H_1, H_2. Everyone votes either for staying around the campfire (0), or for going on a round-up (1). A complicated rule is used to decide the outcome O:

1. if $C = R$, then $O = R$,
2. if R differs from the other four, then $O = R$,
3. otherwise, majority rules.

Suppose $C = R = 1$ and $W = H_1 = H_2 = 0$ (and so $O = 1$). Was $W = 0$ an actual cause of $O = 1$?

4.2 Properties of Causation

4.2.1 Redundant Causation

4.2.2 Transitive Causation

4.2.3 Normality, Norms and Causation

4.2.4 Contrastive Causation

4.2.5 Time and Causation

4.2.6 Certainty, Probability and Non-determinism

4.2.7 Trumping Causation

Trumping is discussed in (Joseph Y Halpern and Hitchcock 2011; Weslake 2015). According to (Weslake 2015), this example is rather contentions. Not only is there disagreements about what should be intuitively understood as a cause, but also the relationship of trumping causation with respect of other forms of causation, in particularly with symmetric overdetermination and preemption, is subject to a lively debate. As indicated by the name one speaks of trumping, if one causal process always dominates the other. That is, if the process A is active, process B is irrelevant for the status of C. However, if process A is inactive, the status of C depends on B.
This scenario presented in Example 4.1.16 shares clearly shares similarities with both symmetric overdetermination and late preemption. That is, similar to late preemption Bob's command starts a process that is "interrupted" by Alice's command. However, in this particular case the given commands are identical and are issues at the exact same moment, allowing one to draw a similarity to switching. (Hitchcock 2011) argues that this example has far reaching implications for the taxonomy of redundant causation. He claims that this example can not be classified as either symmetric overdetermination or preemption. Hence, contradicting the belief that redundant causation is characterised by this dichotomy.
Particularly interesting is that in order to relegate the causal attribution in Example 4.1.16 to Alice and Bob alone, one must accept that Carol is void of agency. Practically operating as a robot, as otherwise some responsibility must be attributed to Carol as well. Therefore, it is critical to include this

piece of information in Example 4.1.16 preventing hidden assumptions, in this case the agency of people, to tarnish the causal intuition. This highlights that using such stories as a foundation for constructing causation may be problematic.

4.2.8 Causation by Omission

Causation by omission, is the claim that the non occurrence of an event caused another event, e.g. $\neg A$ caused B. However, this is a contentions topic. Not only does (Blanchard and Schaffer 2017) claim that causation by omission is one of the open problems in determining actual causation. It is also the case, there remains disagreement within the literature of whether causation by omission should be considered when defining token causality. (Joseph Y Halpern and Hitchcock 2015) identified four established viewpoints within this debate. The first dismisses causation by omission, while the second completely embraces it. The third, is positioned somewhere in between, declaring omissions to have some kind of secondary status. The last argues that it is the normative status of an omission that determines its causal status, e.g. in Example 4.1.17 the inaction of the neighbour only caused the death of the flowers, if he had the obligation to do so.
When assessing Example 4.1.17 A person inclined to attribute the neighbours omission as the cause of the flower's demise, may have incorporated some implicit assumption into their reasoning process. Because, as described in the example the neighbour has no obligation to water the flowers. Meaning, without a statement such as "The neighbour was responsible to water the flowers", the neighbour is on equal footing with any other person in this world. One suggestion, would be to appeal to normality. That is, expecting the neighbour to water the flowers is considered less out of the ordinary that expecting some person on the other side of the world to do so. An additional benefit of this approach is that it provides sufficient flexibility to accommodate all the previously listed viewpoints. However, this flexibility can be problematic, as one has to rank scenarios based on their perceived normality. For example, what if the neighbour was gardener, but also is sworn to kill this particular kind of flowers. Is the scenario now more or less normal? To avoid the reliance on normality, one could also argue for the restriction of the models used. Limiting the variables used to only those relevant (seemingly) relevant for the scenario. For Example 4.1.17 this would imply that only the neighbour and no other person is relevant for the status of the flowers (Blanchard and Schaffer 2017).

Chapter 5

Token Causal Definitions: A comparison

5.1 Token Causal Definitions

5.1.1 Modified Halpern and Pearl Definition

5.1.2 Bochmans Causal Inference

5.1.3 Possible Causal Processes

5.2 Comparison

The important publications as contained in S_A (see Subsection **??**) rely on a variety of languages to encode causal relations. Table 5.1 provides insight into which publications discuss which language family. However, it must be remarked that some liberty with respect to aggregation was taken, e.g. causal models with and without defaults are considered part of the same language family. Here it is important to point out that by far the most discussed framework are causal models, with CP-Logic and Causal Logic tying for a distant second place.

Articles	Causal Models	CP-Logic	Situation Calculus	Non-Monotonic Causal Theory	FOL	Neuron Diagrams	Conditional Logic	$\mathcal{AL}$	SFCA	Abductive Causal Theory
Vennekens, Bruynooghe, and Denecker 2010		X								
Bex et al. 2010										X
Lee et al. 2010				X						
Lifschitz and F. Yang 2010				X						
Glymour et al. 2010	X					X				
Claassen and Heskes 2010	X									
Gerstenberg and Lagnado 2010	X									
Joseph Y Halpern and Hitchcock 2011	X									
Shulz 2011	X									
Briggs 2012	X									
Baumgartner 2013					X	X				
Hyttinen, Hoyer, et al. 2013	X									
Joseph Y Halpern and Hitchcock 2015	X									
Weslake 2015	X									
Chockler et al. 2015	X									
Beckers and Vennekens 2016		X				X				
Schaffer 2016	X									
Joseph Y Halpern 2016b	X									
Blanchard and Schaffer 2017	X									
Wright and Goldberg 2017										
T. F. Icard, Kominsky, and Knobe 2017										
Aleksandrowicz et al. 2017	X									
Fenton-Glynn 2017										
Lagnado and Gerstenberg 2017	X									
Bochman 2018a	X			X						
Ibeling and T. Icard 2018							X			
Beckers and Vennekens 2018	X									
Bochman 2018b	X			X						
Denecker, Bogaerts, and Vennekens 2018		X								
Batusov and Soutchanski 2018	X		X							
Denecker, Bogaerts, and Vennekens 2019	X	X								
Liepiņa, Sartor, and A. Wyner 2019									X	
LeBlanc, Balduccini, and Vennekens 2019								X		
Liepiņa, Sartor, and A. Wyner 2020	X								X	
Khan and Soutchanski n.d.	X		X							
Ibeling and T. Icard 2020	X						X			

Table 5.1: Depicts which publication discuss which languages families

Chapter 6

Appendix

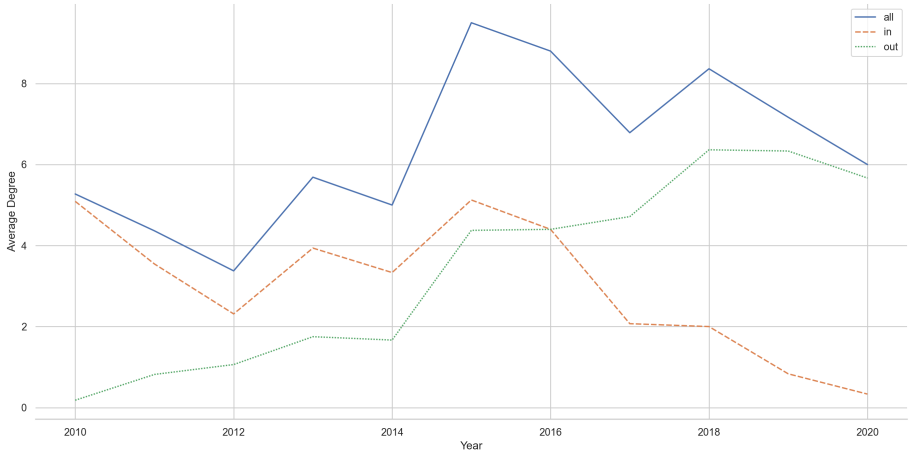


Figure 6.1: A line graph depicting the average in-degree, out-degree and overall degree of the publications in $\mathcal{G}_p$.

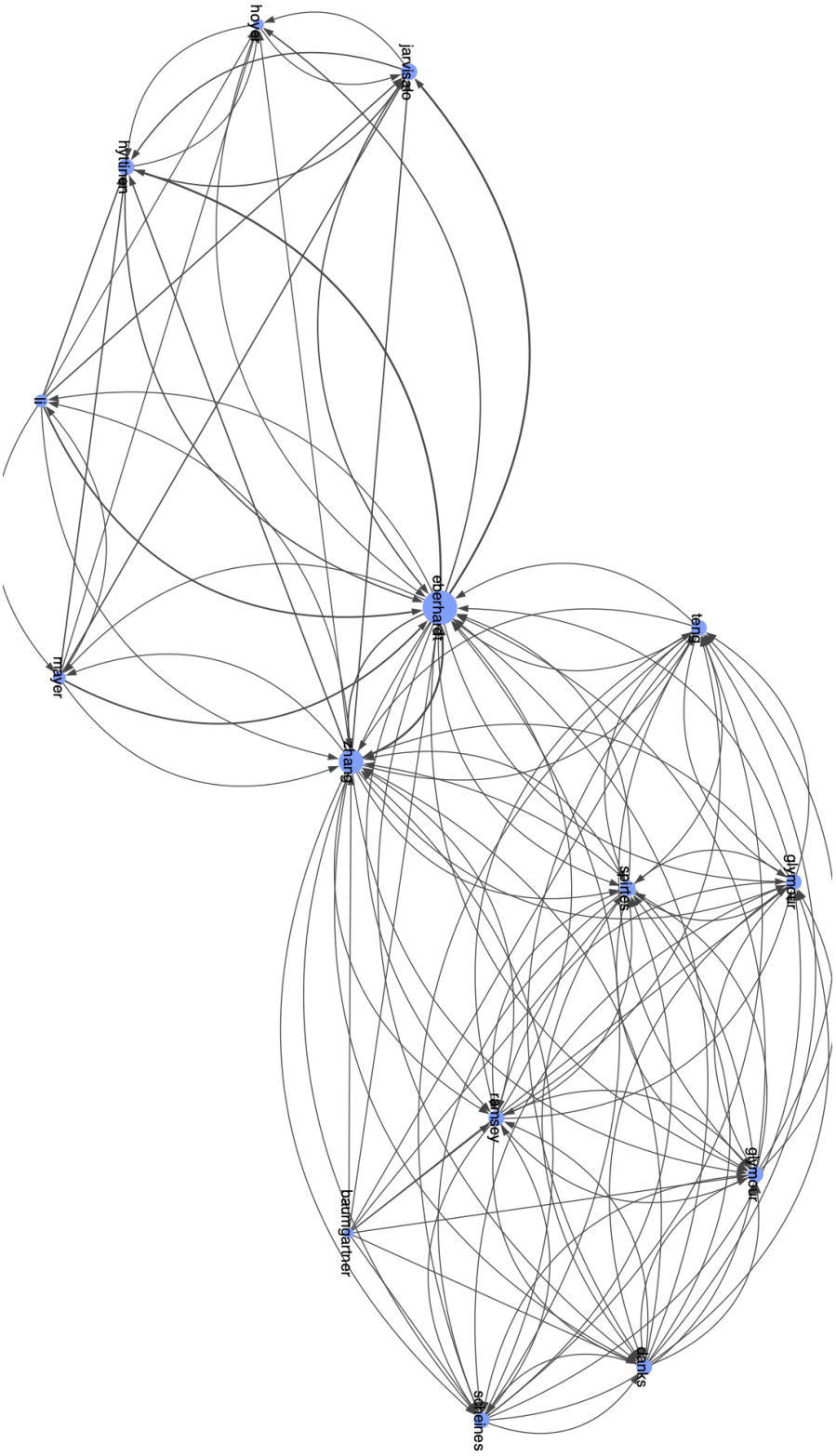


Figure 6.2: A subgraph of $\mathcal{G}_m$, depicting Group 1.

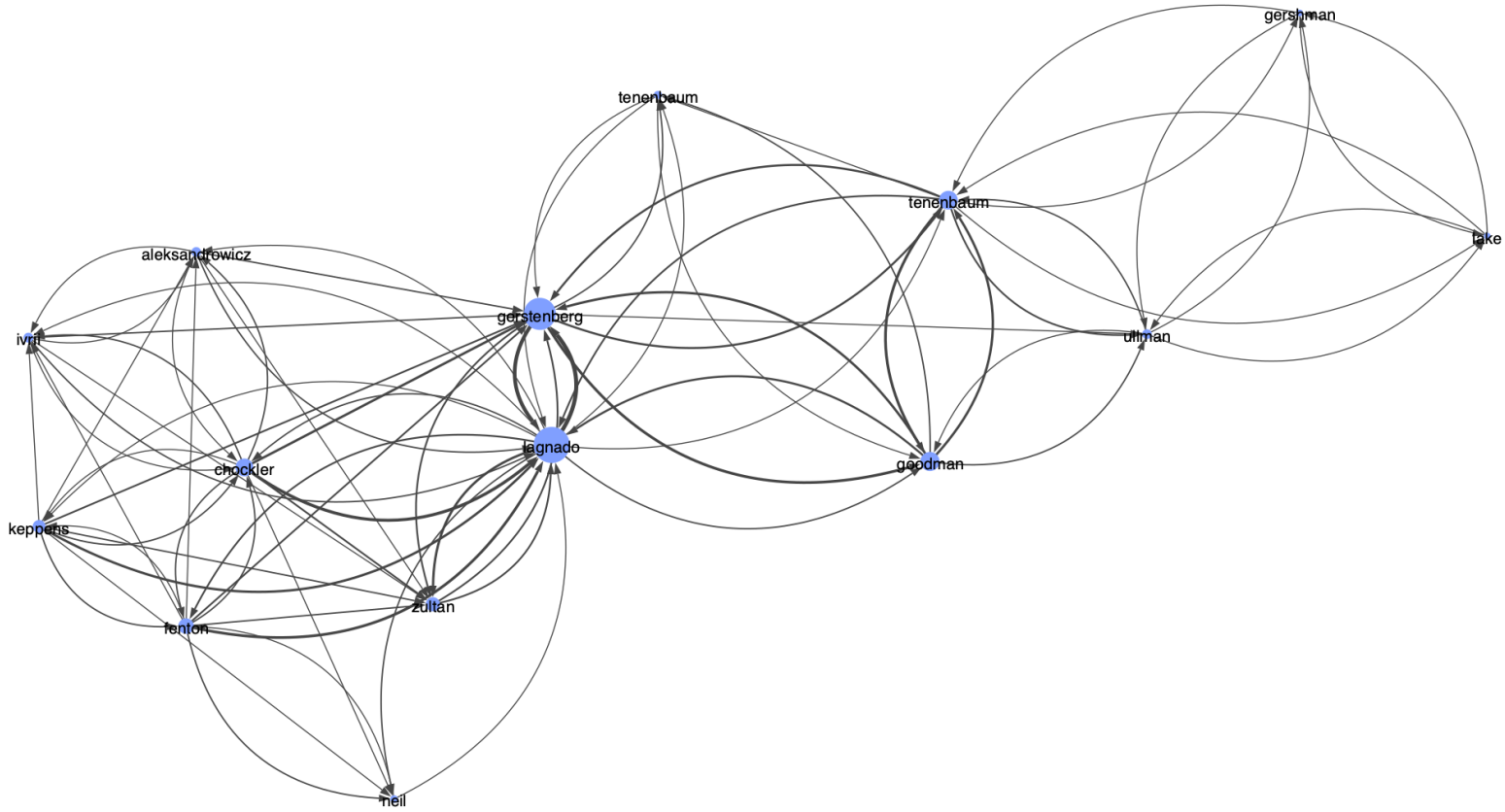


Figure 6.3: A subgraph of $\mathcal{G}_m$, depicting Group 2.

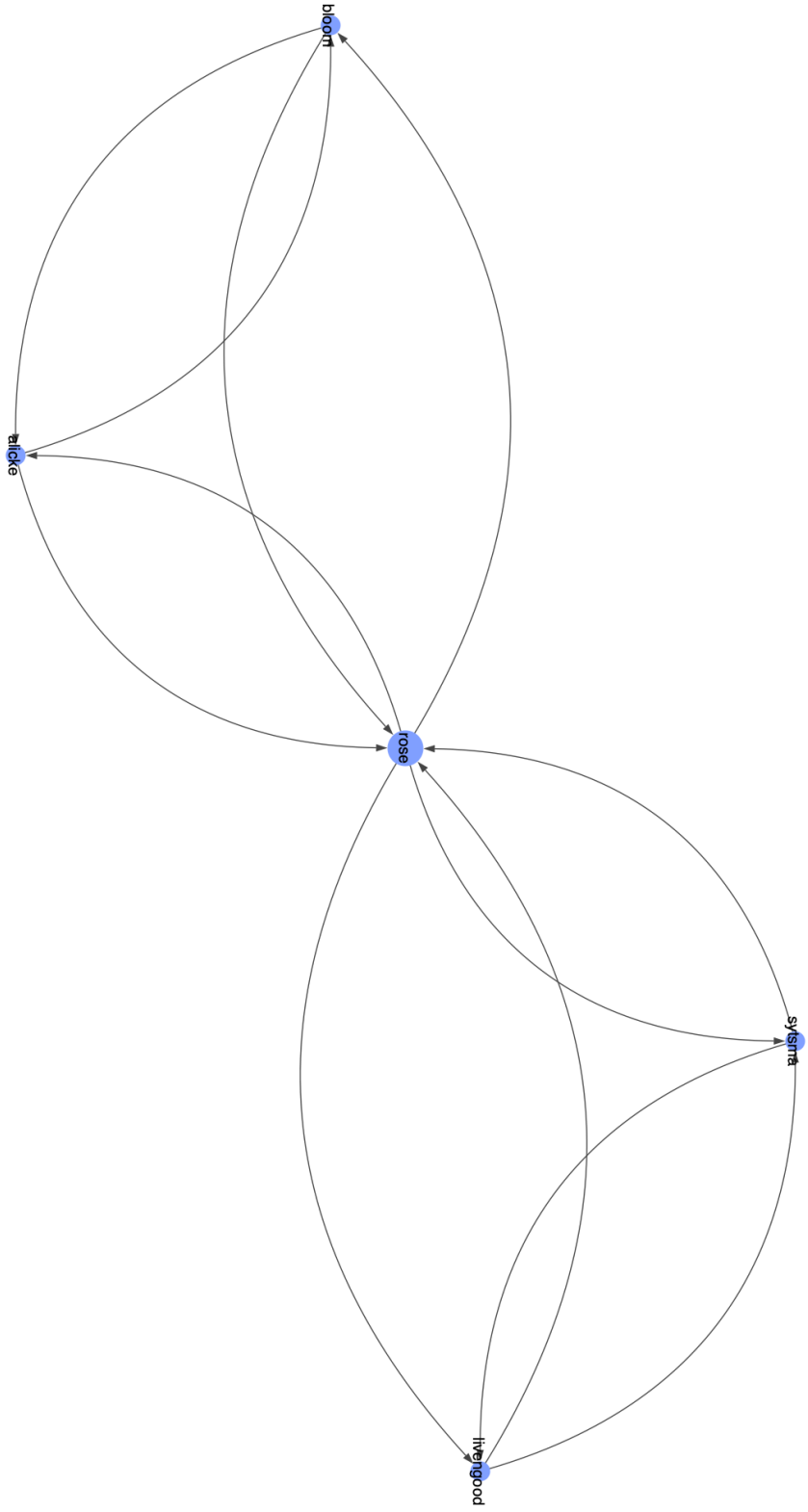


Figure 6.4: A subgraph of $\mathcal{G}_m$, depicting Group 3.

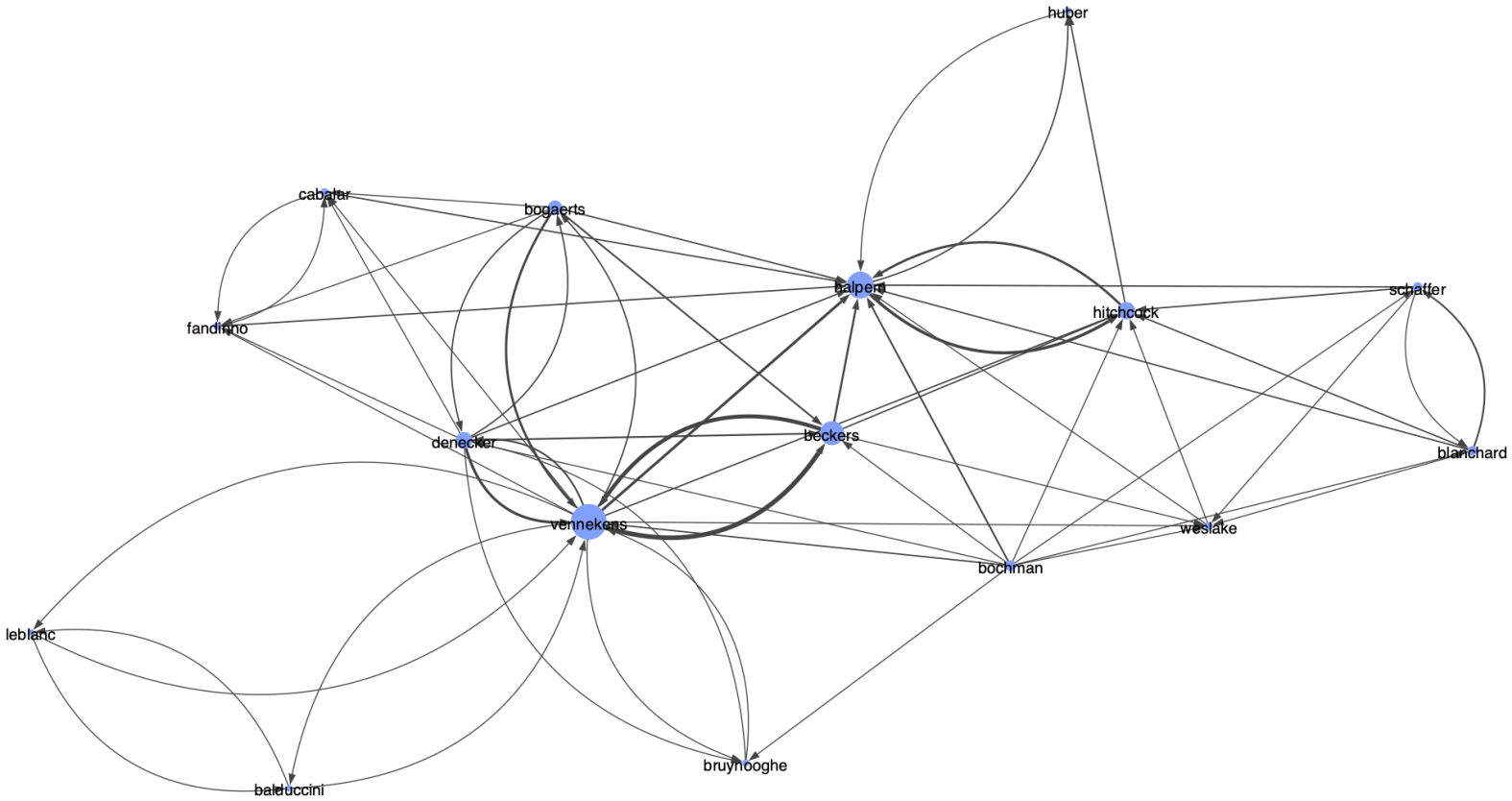


Figure 6.5: A subgraph of $\mathcal{G}_m$, depicting Group 4.

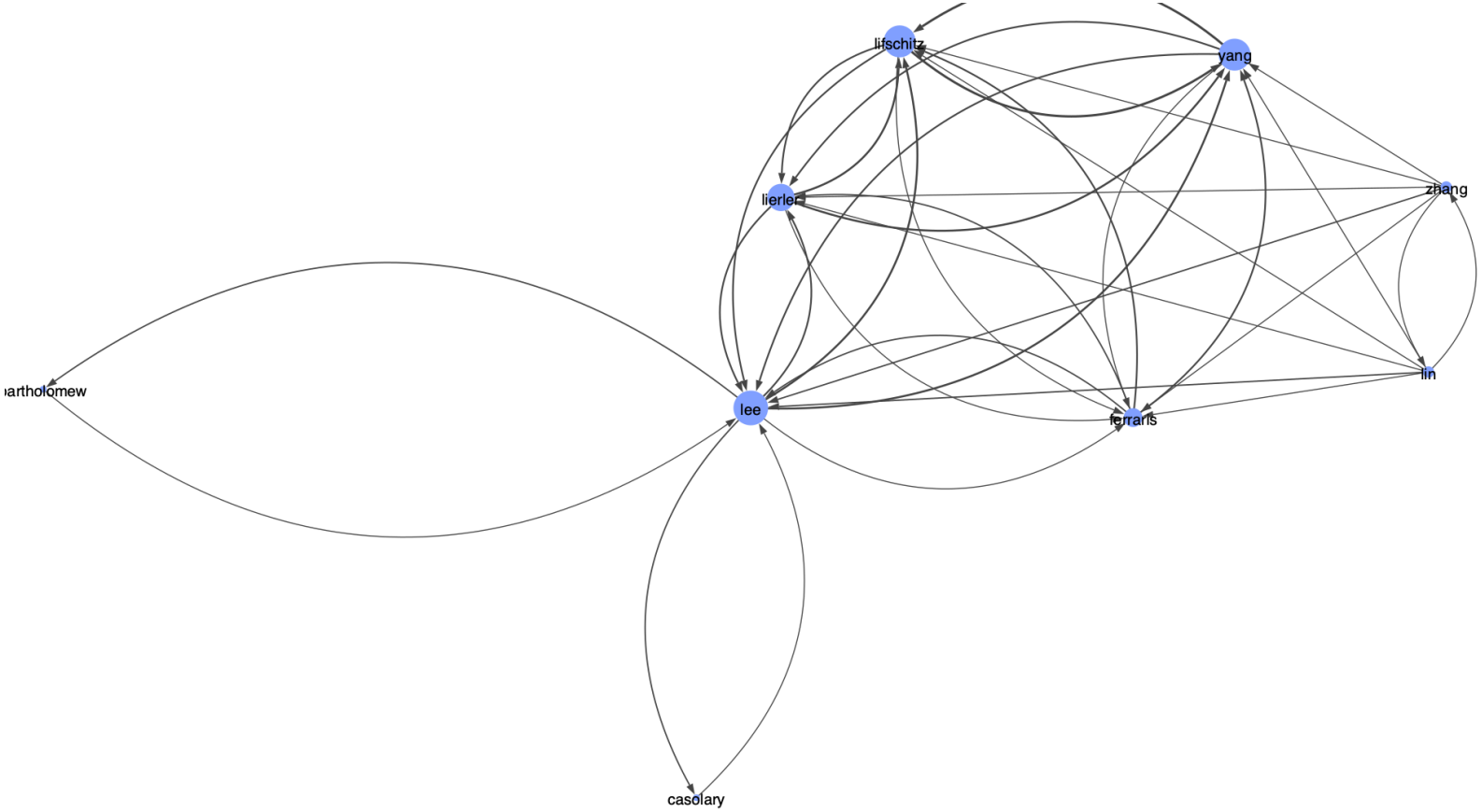


Figure 6.6: A subgraph of $\mathcal{G}_m$, depicting Group 5.

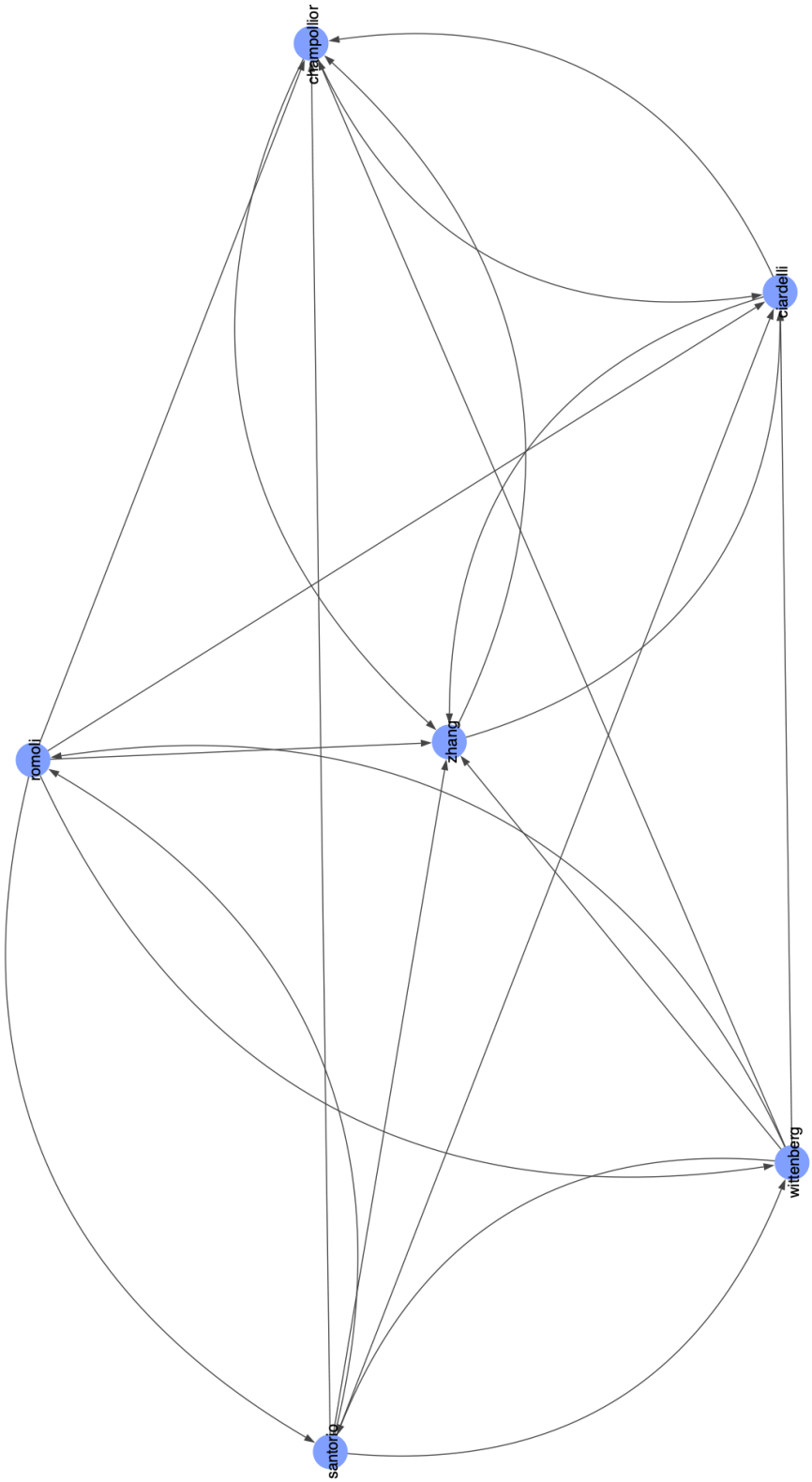


Figure 6.7: A subgraph of $\mathcal{G}_m$, depicting Group 6.

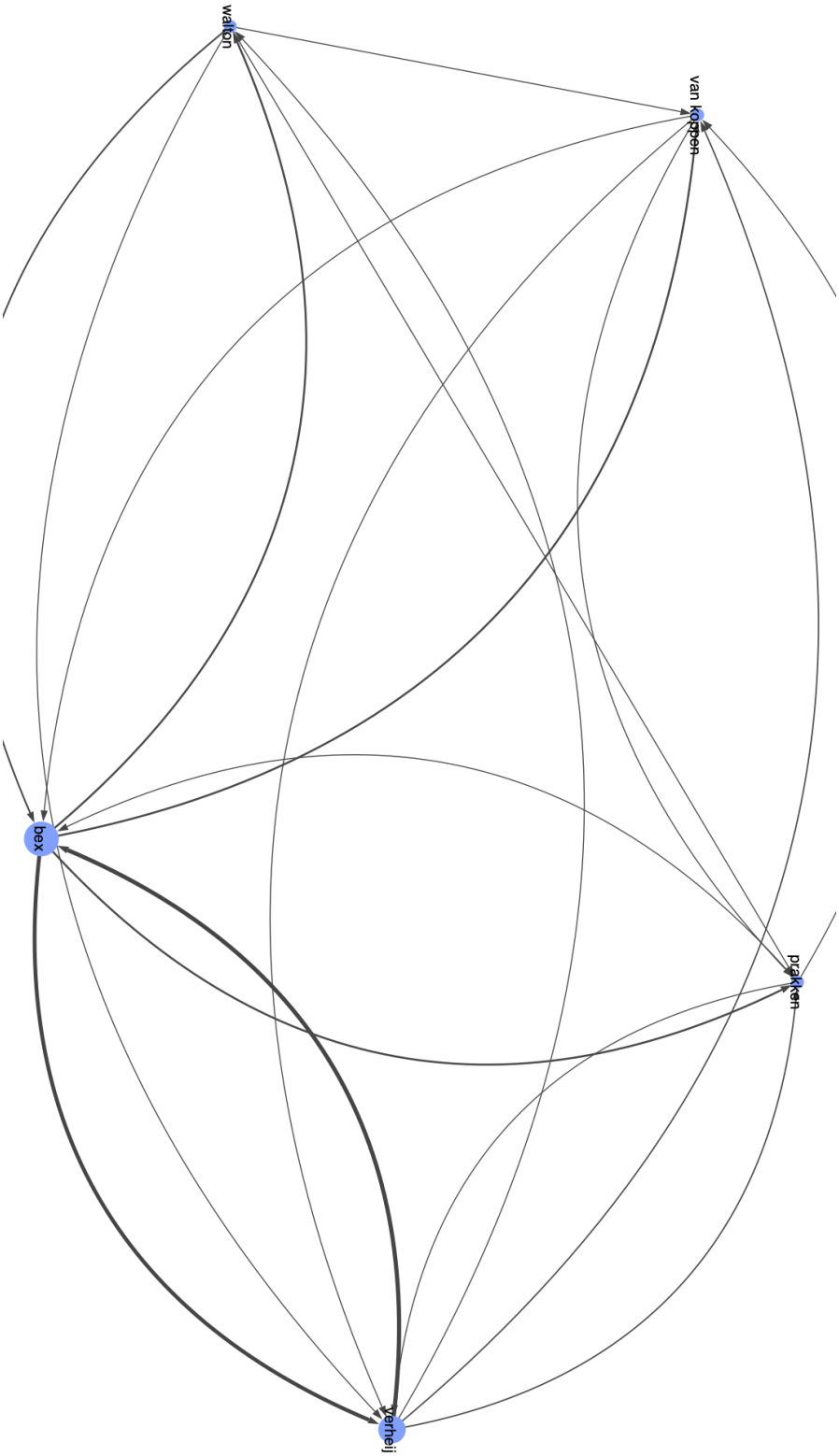


Figure 6.8: A subgraph of $\mathcal{G}_m$, depicting Group 7.

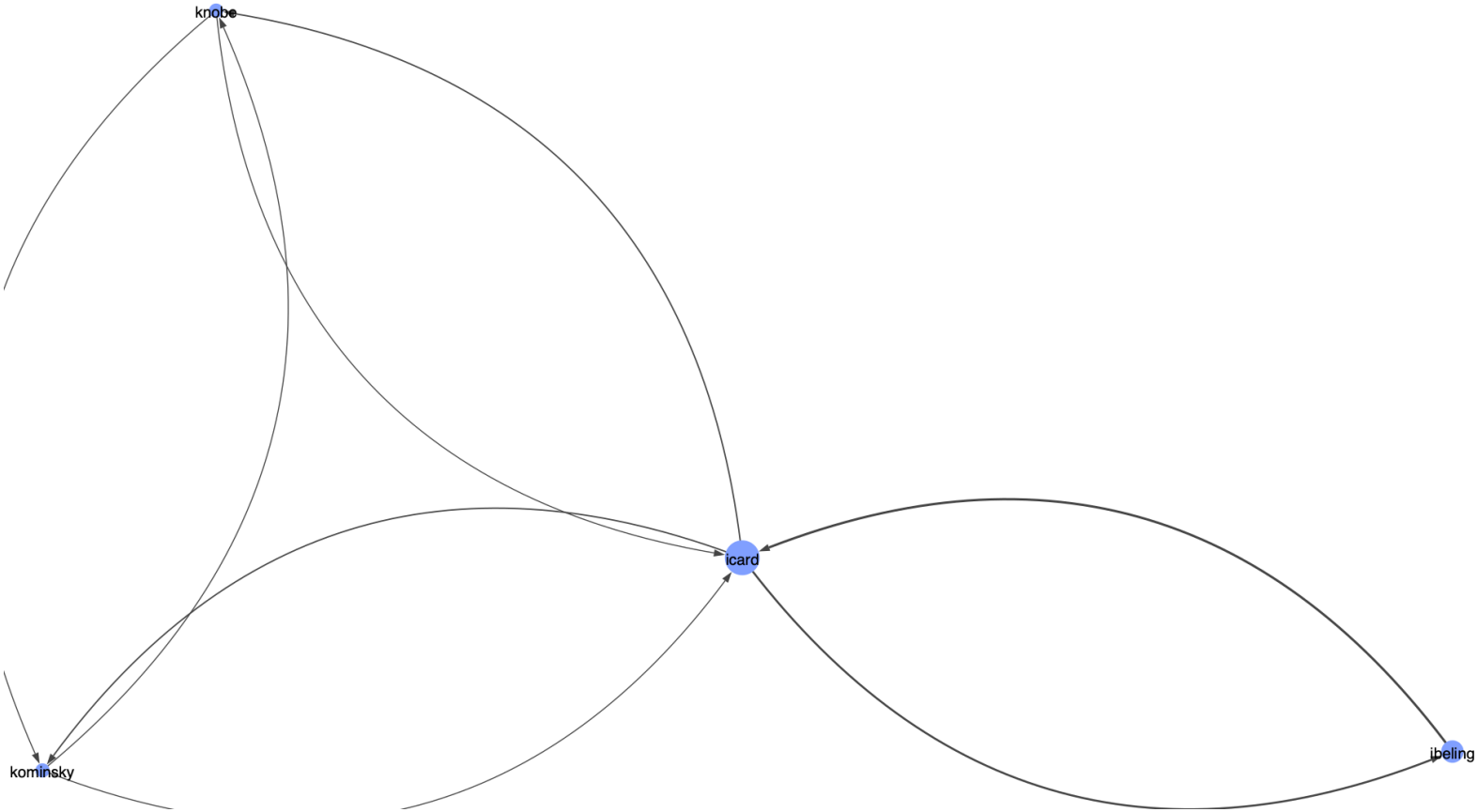


Figure 6.9: A subgraph of $\mathcal{G}_m$, depicting Group 8.

Bibliography

Hart, Herbert Lionel Adolphus and Tony Honoré (1959). *Causation in the Law*. OUP Oxford.

Mackie, John L (1965). "Causes and conditions". In: *American philosophical quarterly* 2.4, pp. 245–264.

Lewis, David (1974). "Causation". In: *The journal of philosophy* 70.17, pp. 556–567.

Chen, Peter Pin-Shan (1976). "The entity-relationship model—toward a unified view of data". In: *ACM transactions on database systems (TODS)* 1.1, pp. 9–36.

Smith, Linda C (1981). "Citation analysis". In:

Papadimitriou, Christos H and Mihalis Yannakakis (1982). "The complexity of facets (and some facets of complexity)". In: *Proceedings of the fourteenth annual ACM symposium on Theory of computing*, pp. 255–260.

Wright, Richard W (1987). "Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts". In: *Iowa L. Rev.* 73, p. 1001.

McDermott, Michael (1995). "Redundant causation". In: *British Journal for the Philosophy of Science*, pp. 523–544.

Pearl, Judea (1995). "Causal diagrams for empirical research". In: *Biometrika* 82.4, pp. 669–688.

McCain, Norman, Hudson Turner, et al. (1997). "Causal theories of action and change". In: *AAAI/IAAI*, pp. 460–465.

Pearl, Judea (1998). "On the definition of actual cause". In:

Page, Lawrence et al. (1999). *The pagerank citation ranking: Bringing order to the web.* Tech. rep. Stanford InfoLab.

Spirtes, Peter et al. (2000). *Causation, prediction, and search*. MIT press.

Bonacich, Phillip and Paulette Lloyd (2001). "Eigenvector-like measures of centrality for asymmetric relations". In: *Social networks* 23.3, pp. 191–201.

Halpern, Joseph Y. and Judea Pearl (2001). "Causes and Explanations: A Structural-Model Approach: Part i: Causes". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI'01. Seattle, Washington: Morgan Kaufmann Publishers Inc., pp. 194–202. ISBN: 1558608001.

Hitchcock, Christopher (2001). "The intransitivity of causation revealed in equations and graphs". In: *The Journal of Philosophy* 98.6, pp. 273–299.

Eiter, Thomas and Thomas Lukasiewicz (2002). "Complexity results for structure-based causality". In: *Artificial Intelligence* 142.1, pp. 53–89.

Hopkins, Mark and Judea Pearl (2003). "Clarifying the usage of structural models for commonsense causal reasoning". In: *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning.* AAAI Press Menlo Park, CA, pp. 83–89.

Bochman, Alexander (2004). "A causal approach to nonmonotonic reasoning". In: *Artificial intelligence* 160.1-2, pp. 105–143.

Giunchiglia, Enrico et al. (2004). "Nonmonotonic causal theories". In: *Artificial Intelligence* 153.1-2, pp. 49–104.

Hall, Ned (2004). "Two concepts of causation". In: *Causation and counterfactuals*, pp. 225–276.

Halpern, Joseph Y and Judea Pearl (2005). "Causes and explanations: A structural-model approach. Part I: Causes". In: *The British journal for the philosophy of science* 56.4, pp. 843–887.

Hausman, Daniel Murray (2005). "Causal relata: Tokens, types, or variables?" In: *Erkenntnis* 63.1, pp. 33–54.

Hiddleston, Eric (2005). "Causal powers". In: *The British journal for the philosophy of science* 56.1, pp. 27–59.

Woodward, James (2005). *Making things happen: A theory of causal explanation.* Oxford university press.

Chen, Peng et al. (2007). "Finding scientific gems with Google's PageRank algorithm". In: *Journal of Informetrics* 1.1, pp. 8–15.

Hall, Ned (2007). "Structural equations and causation". In: *Philosophical Studies* 132.1, pp. 109–136.

Hitchcock, Christopher (2007). "Prevention, preemption, and the principle of sufficient reason". In: *The Philosophical Review* 116.4, pp. 495–532.

Menzies, Peter et al. (2007). "Causation in context". In: *Causation, physics, and the constitution of reality*, pp. 191–220.

Halpern, Joseph Y (2008). "Defaults and Normality in Causal Structures." In: *KR*, pp. 198–208.

Knobe, Joshua and Ben Fraser (2008). "Causal judgment and moral judgment: Two experiments". In: *Moral psychology* 2, pp. 441–8.

Ma, Nan, Jiancheng Guan, and Yi Zhao (2008). "Bringing PageRank to the citation analysis". In: *Information Processing & Management* 44.2, pp. 800–810.

Maslov, Sergei and Sidney Redner (2008). "Promise and pitfalls of extending Google's PageRank algorithm to citation networks". In: *Journal of Neuroscience* 28.44, pp. 11103–11105.

Rosvall, Martin and Carl T Bergstrom (2008). "Maps of random walks on complex networks reveal community structure". In: *Proceedings of the National Academy of Sciences* 105.4, pp. 1118–1123.

Arora, Sanjeev and Boaz Barak (2009). *Computational complexity: a modern approach*. Cambridge University Press.

Beebee, Helen, Christopher Hitchcock, and Peter Menzies (2009). *The Oxford handbook of causation*. Oxford University Press.

Ding, Ying et al. (2009). "PageRank for ranking authors in co-citation networks". In: *Journal of the American Society for Information Science and Technology* 60.11, pp. 2229–2243.

Hitchcock, Christopher (2009). "Structural equations and causation: six counterexamples". In: *Philosophical Studies* 144.3, pp. 391–401.

Hitchcock, Christopher and Joshua Knobe (2009). "Cause and norm". In: *The Journal of Philosophy* 106.11, pp. 587–612.

Pearl, Judea (2009). *Causality*. Cambridge university press.

Vennekens, Joost, Marc Denecker, and Maurice Bruynooghe (2009). "CP-logic: A language of causal probabilistic events and its relation to logic programming". In: *Theory and practice of logic programming* 9.3, pp. 245–308.

Bex, Floris J et al. (2010). "A hybrid formal theory of arguments, stories and criminal evidence". In: *Artificial Intelligence and Law* 18.2, pp. 123–152.

Claassen, Tom and Tom Heskes (2010). "Causal discovery in multiple models from different experiments". In: *Advances in Neural Information Processing Systems*, pp. 415–423.

Erwig, Martin and Eric Walkingshaw (2010). "Causal reasoning with neuron diagrams". In: *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, pp. 101–108.

Gerstenberg, Tobias and David A Lagnado (2010). "Spreading the blame: The allocation of responsibility amongst multiple agents". In: *Cognition* 115.1, pp. 166–171.

Glymour, Clark et al. (2010). "Actual causation: a stone soup essay". In: *Synthese* 175.2, pp. 169–192.

Lee, Joohyung et al. (2010). "Representing synonymity in causal logic and in logic programming". In:

Lifschitz, Vladimir and Fangkai Yang (2010). "Translating first-order causal theories into answer set programming". In: *European Workshop on Logics in Artificial Intelligence*. Springer, pp. 247–259.

Vennekens, Joost, Maurice Bruynooghe, and Marc Denecker (2010). "Embracing events in causal modelling: Interventions and counterfactuals in CP-logic". In: *European Workshop on Logics in Artificial Intelligence*. Springer, pp. 313–325.

Halpern, Joseph Y and Christopher Hitchcock (2011). "Actual causation and the art of modeling". In: *arXiv preprint arXiv:1106.2652*.

Hitchcock, Christopher (2011). "Trumping and contrastive causation". In: *Synthese* 181.2, pp. 227–240.

Pozo, Mónica del et al. (2011). "Centrality in directed social networks. A game theoretic approach". In: *Social Networks* 33.3, pp. 191–200.

Shulz, Katrin (2011). ""If you'd wiggled A, then B would've changed": Causality and counterfactual conditionals."". In: *Synthese* 179, pp. 239–251.

Twardy, Charles R and Kevin B Korb (2011). "Actual causation by probabilistic active paths". In: *Philosophy of Science* 78.5, pp. 900–913.

Vennekens, Joost (2011). "Actual causation in cp-logic". In: *Theory and Practice of Logic Programming* 11.4-5, pp. 647–662.

Beckers, Sander and Joost Vennekens (2012). "Counterfactual dependency and actual causation in CP-logic and structural models: a comparison." In: *STAIRS*. Vol. 241, pp. 35–46.

Briggs, Rachael (2012). "Interventionist counterfactuals". In: *Philosophical studies* 160.1, pp. 139–166.

Runkler, Thomas A (2012). "Data Analytics". In: *Wiesbaden: Springer. doi* 10, pp. 978–3.

Spielman, Daniel (2012). "Spectral graph theory". In: *Combinatorial scientific computing*. 18. Citeseer.

Baumgartner, Michael (2013). "A regularity theoretic approach to actual causation". In: *Erkenntnis* 78.1, pp. 85–109.

Hyttinen, Antti, Patrik O Hoyer, et al. (2013). "Discovering cyclic causal models with latent variables: A general SAT-based procedure". In: *arXiv preprint arXiv:1309.6836*.

Lewis, David (2013). *Counterfactuals*. John Wiley & Sons.

Nykl, Michal et al. (2014). "PageRank variants in the evaluation of citation networks". In: *Journal of Informetrics* 8.3, pp. 683–692.

Wohlin, Claes (2014). "Guidelines for snowballing in systematic literature studies and a replication in software engineering". In: *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pp. 1–10.

Chockler, Hana et al. (2015). "Causal analysis for attributing responsibility in legal cases". In: *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pp. 33–42.

Halpern, Joseph (2015). "A modification of the Halpern-Pearl definition of causality". In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Halpern, Joseph Y (2015). "Cause, responsibility and blame: a structural-model approach". In: *Law, probability and risk* 14.2, pp. 91–118.

Halpern, Joseph Y and Christopher Hitchcock (2015). "Graded causation and defaults". In: *The British Journal for the Philosophy of Science* 66.2, pp. 413–457.

Segarra, Santiago and Alejandro Ribeiro (2015). "Stability and continuity of centrality measures in weighted graphs". In: *IEEE Transactions on Signal Processing* 64.3, pp. 543–555.

Weslake, Brad (2015). "A partial theory of actual causation". In:

Zhang, Haodi and Fangzhen Lin (2015). "Characterizing causal action theories and their implementations in answer set programming: Action lan-

guages b, c, and beyond". In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Zhao, Dangzhi and Andreas Strotmann (2015). "Analysis and visualization of citation networks". In: *Synthesis lectures on information concepts, retrieval, and services* 7.1, pp. 1–207.

Albrecht, Stefano V and Subramanian Ramamoorthy (2016). "Exploiting causality for selective belief filtering in dynamic Bayesian networks". In: *Journal of Artificial Intelligence Research* 55, pp. 1135–1178.

Beckers, Sander and Joost Vennekens (2016). "A general framework for defining and extending actual causation using CP-logic". In: *International Journal of Approximate Reasoning* 77, pp. 105–126.

Halpern, Joseph Y (2016a). *Actual causality*. MiT Press.

– (2016b). "Appropriate causal models and the stability of causation". In: *The Review of Symbolic Logic* 9.1, pp. 76–102.

Schaffer, Jonathan (2016). "Grounding in the image of causation". In: *Philosophical studies* 173.1, pp. 49–100.

Zhang, Lu, Yongkai Wu, and Xintao Wu (2016). "A causal framework for discovering and removing direct and indirect discrimination". In: *arXiv preprint arXiv:1611.07509*.

Aleksandrowicz, Gadi et al. (2017). "The computational complexity of structure-based causality". In: *Journal of Artificial Intelligence Research* 58, pp. 431–451.

Blanchard, Thomas and Jonathan Schaffer (2017). "Cause without default". In: *Making a difference*, pp. 175–214.

Constantinou, Anthony and Norman Fenton (2017). "Towards smart-data: Improving predictive accuracy in long-term football team performance". In: *Knowledge-Based Systems* 124, pp. 93–104.

Fenton-Glynn, Luke (2017). "A proposed probabilistic extension of the Halpern and Pearl definition of 'actual cause'". In: *The British journal for the philosophy of science* 68.4, pp. 1061–1124.

Hyttinen, Antti, Paul Saikko, and Matti Järvisalo (2017). "A Core-Guided Approach to Learning Optimal Causal Graphs." In: *IJCAI*, pp. 645–651.

Icard, Thomas F, Jonathan F Kominsky, and Joshua Knobe (2017). "Normality and actual causal strength". In: *Cognition* 161, pp. 80–93.

Lagnado, David A and Tobias Gerstenberg (2017). "Causation in legal and moral reasoning". In: *The Oxford Handbook of Causal Reasoning*, pp. 565–601.

Liang, Ruxia, Jianqiang Wang, and Hongyu Zhang (2017). "Evaluation of e-commerce websites: An integrated approach under a single-valued trapezoidal neutrosophic environment". In: *Knowledge-Based Systems* 135, pp. 44–59.

Liu, Quan et al. (2017). "Cause-Effect Knowledge Acquisition and Neural Association Model for Solving A Set of Winograd Schema Problems." In: *IJCAI*, pp. 2344–2350.

Summerville, Adam, Joseph Osborn, and Michael Mateas (2017). "Charda: Causal hybrid automata recovery via dynamic analysis". In: *arXiv preprint arXiv:1707.03336*.

Verheij, Bart (2017). "Proof with and without probabilities". In: *Artificial Intelligence and Law* 25.1, pp. 127–154.

Wright, Richard W and Richard Goldberg (2017). "The NESS account of natural causation: A response to criticisms". In: *Critical Essays on 'Causation and Responsibility*, pp. 13–66.

Zhang, Haodi and Fangzhen Lin (2017). "Characterizing causal action theories and their implementations in answer set programming". In: *Artificial Intelligence* 248, pp. 1–8.

Zhang, Junzhe and Elias Bareinboim (2017). "Transfer learning in multi-armed bandit: a causal approach". In: *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pp. 1778–1780.

Zhang, Kun et al. (2017). "Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination". In: *IJCAI: Proceedings of the Conference*. Vol. 2017. NIH Public Access, p. 1347.

Bäckström, Christer, Peter Jonsson, and Sebastian Ordyniak (2018). "Novel Structural Parameters for Acyclic Planning Using Tree Embeddings." In: *IJCAI*, pp. 4653–4659.

Batusov, Vitaliy and Mikhail Soutchanski (2018). "Situation calculus semantics for actual causality". In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

Beckers, Sander and Joost Vennekens (2018). "A principled approach to defining actual causation". In: *Synthese* 195.2, pp. 835–862.

Bochman, Alexander (2018a). "Actual Causality in a Logical Setting." In: *IJCAI*, pp. 1730–1736.

– (2018b). "On Laws and Counterfactuals in Causal Reasoning". In: *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning*.

Chai, Joyce Y et al. (2018). "Language to Action: Towards Interactive Task Learning with Physical Agents." In: *IJCAI*, pp. 2–9.

Chikahara, Yoichi and Akinori Fujino (2018). "Causal Inference in Time Series via Supervised Learning." In: *IJCAI*, pp. 2042–2048.

Denecker, Marc, Bart Bogaerts, and Joost Vennekens (2018). "Causal reasoning in a logic with possible causal process semantics". In: *17th INTERNATIONAL WORKSHOP ON NON-MONOTONIC REASONING NMR 2018*. AAAI Press 2018, pp. 90–98.

Ibeling, Duligur and Thomas Icard (2018). "On the Conditional Logic of Simulation Models". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, pp. 1868–1874. ISBN: 9780999241127.

Jaber, Amin, Jiji Zhang, and Elias Bareinboim (2018). "A Graphical Criterion for Effect Identification in Equivalence Classes of Causal Diagrams".

In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, pp. 5024–5030. ISBN: 9780999241127.

Laurent, Jonathan, Jean Yang, and Walter Fontana (2018). "Counterfactual Resimulation for Causal Analysis of Rule-Based Models." In: *IJCAI*, pp. 1882–1890.

Lu, Junli et al. (2018). "Mining strong symbiotic patterns hidden in spatial prevalent co-location patterns". In: *Knowledge-Based Systems* 146, pp. 190–202.

Mu, Kedian (2018). "Measuring inconsistency with constraints for propositional knowledge bases". In: *Artificial Intelligence* 259, pp. 52–90.

Sridhar, Dhanya, Jay Pujara, and Lise Getoor (2018). "Scalable Probabilistic Causal Structure Discovery." In: *IJCAI*, pp. 5112–5118.

Wenjuan, Wei, Feng Lu, and Liu Chunchen (2018). "Mixed causal structure discovery with application to prescriptive pricing". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. AAAI Press, pp. 5126–5134.

Wetzel, Linda (2018). "Types and Tokens". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2018. Accessed: 2020-03-30. Metaphysics Research Lab, Stanford University.

Zhang, Lu, Yongkai Wu, and Xintao Wu (2018). "Achieving Non-Discrimination in Prediction". In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. IJCAI'18. Stockholm, Sweden: AAAI Press, pp. 3097–3103. ISBN: 9780999241127.

Zhang, Zan et al. (2018). "Collective behavior learning by differentiating personal preference from peer influence". In: *Knowledge-Based Systems* 159, pp. 233–243.

Bloch, Francis, Matthew O Jackson, and Pietro Tebaldi (2019). "Centrality measures in networks". In: *Available at SSRN 2749124*.

Cai, Ruichu et al. (2019). "Causal Discovery with Cascade Nonlinear Additive Noise Models". In: *IJCAI*.

Chen, Daniel L (2019). "Judicial analytics and the great transformation of American Law". In: *Artificial Intelligence and Law* 27.1, pp. 15–42.

Denecker, Marc, Bart Bogaerts, and Joost Vennekens (2019). "Explaining actual causation in terms of possible causal processes". In: *European Conference on Logics in Artificial Intelligence*. Springer, pp. 214–230.

Hassanzadeh, Oktie et al. (2019). "Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts". In: *IJCAI19*.

Kronegger, Martin, Sebastian Ordyniak, and Andreas Pfandler (2019). "Backdoors to planning". In: *Artificial Intelligence* 269, pp. 49–75.

LeBlanc, Emily, Marcello Balduccini, and Joost Vennekens (2019). "Explaining actual causation via reasoning about actions and change". In: *European Conference on Logics in Artificial Intelligence*. Springer, pp. 231–246.

Li, Xiangju et al. (2019). "Context-aware emotion cause analysis with multi-attention-based neural network". In: *Knowledge-Based Systems* 174, pp. 205–218.

Liepiņa, Rūta, Giovanni Sartor, and Adam Wyner (2019). "Evaluation of Causal Arguments in Law: the Case of Overdetermination". In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pp. 214–218.

Liepina, Ruta, Giovanni Sartor, and Adam Z. Wyner (2019). "Arguing about causes in law: a semi-formal framework for causal arguments". In: *Artificial Intelligence and Law*, pp. 1–21.

Neil, Martin et al. (2019). "Modelling competing legal arguments using Bayesian model comparison and averaging". In: *Artificial Intelligence and Law* 27.4, pp. 403–430.

Shankar, Shiv et al. (2019). "Three-quarter sibling regression for denoising observational data". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5960–5966.

Sridhar, Dhanya and Lise Getoor (2019). "Estimating Causal Effects of Tone in Online Debates". In: *IJCAI*.

Xie, Zhipeng and Feiteng Mu (2019). "Boosting causal embeddings via potential verb-mediated causal patterns". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, pp. 1921–1927.

Xu, Depeng et al. (2019). "Achieving Causal Fairness through Generative Adversarial Networks". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by Sarit Kraus. ijcai.org, pp. 1452–1458. DOI: 10.24963/ijcai.2019/201. URL: https://doi.org/10.24963/ijcai.2019/201.

Zander, Benito van der, Maciej Liśkiewicz, and Johannes Textor (2019). "Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework". In: *Artificial Intelligence* 270, pp. 1–40.

Zhalama et al. (2019). "ASP-based Discovery of Semi-Markovian Causal Models under Weaker Assumptions". In: *IJCAI*.

Ibeling, Duligur and Thomas Icard (2020). "Probabilistic Reasoning across the Causal Hierarchy". In: *arXiv preprint arXiv:2001.02889*.

Liepiņa, Rūta, Giovanni Sartor, and Adam Wyner (2020). "Arguing about causes in law: a semi-formal framework for causal arguments". In: *Artificial intelligence and law* 28.1, pp. 69–89.

Khan, Shakil M and Mikhail Soutchanski (n.d.). "Necessary and Sufficient Conditions for Actual Root Causes". In: ().